Language WindOWing through COrpOra

Visualización del lenguaje a través de corpus

Isabel Moskowich
Begoña Crespo
Inés Lareo
Paula Lojo
Eds.

Language Windowing through Corpora.

Visualización del lenguaje a través de corpus

Part II

L-Z

Editors

Isabel Moskowich-Spiegel Fandiño

Begoña Crespo García

Inés Lareo Martín

Paula Lojo Sandino

Universidade da Coruña

A Coruña 2010

ISBN: 978-84-9749-401-4

Cover designed by Inés Lareo and Alejandro González

A Coruña 2010
Universidade da Coruña
Servizo de Publicacións

Language Windowing through Corpora.
Visualización del lenguaje a través de corpus
Part I. A - K

Editors:
Isabel Moskowich-Spiegel Fandiño
Begoña Crespo García
Inés Lareo Martín
Paula Lojo Sandino

Universidade da Coruña 2010
Servizo de Publicacións
www.udc.gal/publicacions

HANDLE: **http://hdl.handle.net/2183/28943**

Number of pages: V + 479
Index: pp. i - v

ISBN: 978-84-9749-401-4

Cover design by: Inés Lareo y Alejandro González

Contents/Contenidos Part II

Foreword

Though Corpus Linguistics, both as a methodology and as a branch of linguistics itself, has been among us for the last forty years, its development, mainly in the Anglophone world, has had a repercussion on the rest of the community of linguists. In Spain, for instance, the recently created association of Corpus Linguistics (AELINCO) testifies to this. The collection of essays we are presenting here are just a mere sample of the interest the topics relating to Corpus Linguistics have arisen everywhere.

Such different topics as those related to Computational Linguistics found in "Obtaining computational resources for languages with scarce resources from closely related computationally-developed languages. The Galician and Portuguese case" or "Corpus-Based Modelling of Lexical Changes in Manic Depression Disorders: The Case of Edgar Allan Poe" belonging to the field of Corpus and Literary Studies can be found in the ensuing pages. Almost all research areas can nowadays be investigated using Corpus Linguistics as a valid methodology. This is reason why *Language Windowing through Corpora* gathers papers dealing with discourse, variation and change, grammatical studies, lexicology and lexicography, corpus design, contrastive analyses, language acquisition and learning or translation.

This work's title aims at reflecting not only the great variety of topics gathered in it but also the worldwide interest awaken by the computer processing of language. In fact, researchers from many different institutions all over the world have contributed to this book. Apart from the twenty-two Spanish Universities, people from other Higher Education Institutions have authored and co-authored the essays contained here, namely, Russia, Venezuela, Brazil, UK, Finland, Portugal, Poland, Austria, Mexico, Thailand, Iran, the Netherlands, Belgium, Japan, Turkey, China, Italy, Malaysia, Romania and Sweden. All these essays have been alphabetically arranged, by the names of their authors, in two parts. Part 1 contains the papers by authors from A to K and Part 2, those of authors from L to Z.

Our special thanks to all the referees who carried out the selection of papers and to the contributors to this volume for giving us the opportunity to make a patchwork of different views and perspectives of what is being currently done in the field. Our thanks to Ms Agnieszka Kozera for her work as an assistant to the editors. We do hope the contents of these essays are illuminating for readers who may excuse all the mistakes and misprints that might remain after a hard editorial work. *The Editors*

In search of cross-linguistic anchor phenomena for translation quality assessment

BELÉN LABRADOR

Universidad de León

Abstract

The aim of this paper is to identify a new grammatical anchor phenomenon; anchor phenomena are "grammatical resources that are perceived as being cross-linguistically equivalent but that tend to and/ or convey partially divergent meanings" (Rabadán et al, 2007). These anchor phenomena will subsequently be used as criteria for application in Translation Quality Assessment and in Translator Training. This paper illustrates a contrastive corpus-based procedure for the identification of anchor phenomena with the case of neutral demonstrative pronouns; the methodology involves comparing data from the Spanish translations of English texts in P-ACTRES parallel corpus and from CREA (Corpus de Referencia del Español Actual by the Spanish Royal Academy) and spotting significant overuse and/ or underuse of some uses of certain grammatical resources in the translations as compared with non-translated target language.

Keywords: anchor phenomena, translationese, translation quality assessment, parallel corpora, monolingual corpora

Resumen

Esta investigación tiene como objetivo encontrar un nuevo fenómeno ancla; los fenómenos ancla son recursos gramaticales que se asumen como equivalentes interlingüísticos pero cuyos usos no siempre coinciden en cada lengua (Rabadán et al, 2007). Una vez identificados, se pretende que sean de utilidad en la evaluación de la calidad de las traducciones y en la enseñanza de la traducción. En el presente estudio ilustraré el procedimiento de búsqueda de fenónemos ancla con el caso de los pronombres demonstrativos neutrales; la metodología consiste en comparar datos extraídos de un corpus paralelo de traducciones en lengua española de textos ingleses (P-ACTRES) con los de un corpus monolingüe (CREA) y en identificar los casos en los que dicho recurso gramatical se ha utilizado en las traducciones con algún significado concreto demasiado o demasiado poco en comparación con la lengua no traducida.

Palabras clave: fenómenos ancla, translationese, evaluación de la calidad de las traducciones, corpus paralelos, corpus monolingües

1. INTRODUCTION

Although translation is a scientific discipline, it also undeniably imbued with artistic and creative features. That's why it is difficult to rate the quality of translations further than a mere impressionistic appraisal and the detection of interlinguistic and intralinguistic mistakes; the former mainly caused by misunderstanding the source language, thus producing a non-equivalent translation solution; the latter having to do with choosing a non-existent or not appropriate target language unit.

It is true that some translators are more innately talented to find the best words for the best patterns but it is also true that training is just as necessary and makes good translators

better. Learning translation strategies goes hand in hand with distinguishing what is better and what is worse in translation, i.e. with learning how to assess the quality of translations.

There are cases when we can *feel* that a translation is not too good but we cannot see any clear and unambiguous mistakes. "Assessing translation quality is generally seen as a difficult and elusive task because of a lack of conceptual clarity, and the inadequacy of the tools available" (Rabadán, Labrador and Ramón, 2009).

The ACTRES project claims that corpus-based research can offer translation evaluators, trainers and trainees objective and reliable criteria so that useful and usable tools can be built in order to improve evaluation methods (Rabadán, 2005-2008: 310).

We can use *anchor phenomena* as the objective and reliable criteria we need. In the alignment programmes used in parallel corpora, one of the most important matching parameters is an anchor word list, a sort of bilingual lexicon made up of statistically significant common words taken from the texts that constitute the corpus. They help get two texts aligned by pinning down or *anchoring* pairs of lexical words that are likely to be equivalent.

In a similar sense, the term *anchor phenomena* has been borrowed to mean language-specific associations between grammatical meaning and formal resource that can help find traces of *translationese* (i.e. linguistic features of translated language that can be attributed to the influence of the source language – Baker, 1996; Mauranen, 2004; Tirkkonen-Condit, 2002) and therefore, be applied in translation quality assessment. A translation which is close to non-translated usage of anchor phenomena will rate higher for quality than another which departs from it. *Anchor phenomena* help assess the quality of a translation by a) *anchoring* pairs of grammatical words in two languages that are likely to be equivalent but may differ in one or more meanings and b) checking if they are used with all their meanings to a similar extent in translated and non-translated language.

To date, a series of anchor phenomena have been already identified in previous descriptive studies –premodifying adjectives (Rabadán et al, 2009), adverbs ending in -mente (Ramón and Labrador, 2008), some uses of *poder*, some periphrastic uses of *estar* + *gerund* (Rabadán, 2005) and some uses of personal pronouns (Rabadán, Gutiérrez and Ramón 2007) – but more of them are needed so that they can be incorporated to a semi-automatic tool for the detection of deviance in Spanish translations from the common usage in non-translated original Spanish. A translation quality assessment application prototype has already been designed by the engineers in the ACTRES team but needs to be improved and amplified (http://actres.unileon.es/inicio. php?elementoID=53).

The present paper illustrates the process of finding another anchor phenomenon with the case of the neutral forms of Spanish demonstrative pronouns.

2. METHODOLOGY

2.1. The corpora

The two corpora used for this purpose are *CREA* http://corpus.rae.es/creanet.html), the large monolingual Spanish reference corpus by the Spanish Royal Academy and P-ACTRES (http://actres.unileon.es/inicio.php?elementoID=12), a custom-made English-Spanish parallel corpus built by the ACTRES research group at the Universities of León and Cantabria in collaboration with Knut Hofland and Øystein Reigem, from the University of Bergen (Norway).

P-ACTRES consists of 238 pairs of English original texts and their translations into Spanish published from 2000 to 2006. This amounts to approximately 2.5 million words (for more information see Izquierdo, Hofland and Reigem, 2008).

2.2. Selection criteria

The starting point for the identification of an anchor phenomenon is usually a hunch, or a previous contrastive study. In an earlier paper about English and Spanish demonstratives, significant overuse of these neutral forms was found (Labrador, in press). If there is significant overuse in the whole of their occurrences, there must be significant overuse in at least one of their uses. A subsequent analysis of their meanings along with their frequency rates follows to search for particular cases of deviance from non-translated Spanish usage.

For the purpose of this study, I have only resorted to the fictional subcorpus, as the final application of the findings is to rate the quality of novels.

In order to get comparable data, the selection criteria for the queries in *CREA* were:

- Diachronic: from 2000 onwards.
- Diatopic: Spain.
- Type of texts: books.
- Register/ subject matter: fiction.

Table 1 shows:

a) the number of occurrences of each neutral demonstrative pronoun in the part of *CREA* selected, which amounts to 2,487,141 words.
b) the number of occurrences of each neutral demonstrative pronoun in the subcorpus of P-ACTRES selected, which amounts to 421,065 words.

c) The occurrences per million words in *CREA*.
d) The occurrences per million words in P-ACTRES.
e) The p-value and the z-score, two indicators of significant difference[1]. A result is said to be statistically significant when p-value is less than 0.05 or z-score is less than 1.96 or more than 1.96, with a 95% confidence interval.
f) Finally, the interpretation of the last four columns: whether there is overuse or underuse of the pronoun and if it is statistically significant.

Table 1: Frequency rates, p-value, z-score and interpretation of neutral demonstrative pronouns.

	CREA	P-ACTRES	Occurrences per million words CREA	Occurrences per million words P-ACTRES	p-value	z-score	Interpretation
Esto	984	213	395.63	505.86	0.001	-3.26098	Significant overuse
Eso	3754	601	1509.36	1427.33	0.20	1.273009	Non-significant underuse
Aquello	336	109	135.09	258.86	0	-6.00485	Significant overuse

As can be seen, *esto* and *aquello* are significantly overused.

3. ANALYSIS OF SEMANTIC FUNCTIONS

The next step is to focus on the forms that have statistically significant overuse, that is, esto and aquello. I have analysed a representative sample of each form after using a statistical formula for simple random sampling. The formula is as follows: $n = N / ((N-1)E^2 + 1)$ where 'n' is the sample to be analysed, 'N' the population, i.e., the total number of occurrences yielded by the searches, and 'E' the estimative error (5%). Table 2 shows the total number of occurrences, which is the population in this case and the representative sample that has to be taken for analysis.

[1] P-value calculated with chi-square test (Preacher, 2001) and z-score calculated with Megastat add-in for Excel.

Table 2: Number of occurrences of neutral demonstrative pronouns and samples for analysis.

	Fiction CREA	Sample CREA	Fiction P-ACTRES	Sample P-ACTRES
esto	984	285	213	139
eso	3754	362	601	240
aquello	336	183	109	86

After obtaining as many lines of concordance as needed for each pronoun in each corpus, I proceeded to the analysis of their semantic functions by reading and classifying each occurrence in its co-text.

3.1. Analysis of esto

Table 3 shows each semantic function of *esto* and if there is significant deviance in the translations from non-translated Spanish.

Table 3: Frequency rates, p-value, z-score and interpretation of the semantic functions of *esto.*

	Occurrences in CREA	Occurrences in P-ACTRES	Occurrences per million words CREA	Occurrences per million words P-ACTRES	P-Value	Z-Score	Interpretation
anaphoric	199	90	80.01	213.74	0.64520432	1.05	Non-significant overuse
cataphoric	23	3	9.24	7.12	0.02366525	2.38	Significant underuse
deictic	47	43	18.89	102.12	0.0068993	-3.41	Significant overuse
Indefinite-contrastive	6	3	2.41	7.12	0.97477288	-0.04	Non-significant overuse
Explanatory	3	0	1.2	0	0.22724861	1.21	Non-significant underuse
Temporal	7	0	2.81	0	0.06571241	1.86	Non-significant underuse
Hesitation marker	0	2	0	4.74	0.04385853	-2.03	Significant overuse

The most literal meaning of demonstrative pronouns is deictic (example 1) but this type of exophoric reference is not as frequent as the anaphoric use (example 2), which overrides all other uses.

(1) *el vaso extendido hacia Pepita, insiste: - Bébete **esto**. Y se sienta junto a ella. - Bebe despacio.*

(2) *El profesor recitaba en lugar de hablar, y **esto**, combinado con el viejo y manoseado libro de texto, hacía que la clase pareciera una hora muerta cada dos días.*

Cataphoric usage – referring forwards to something mentioned afterwards in the text is not that common (example 3).

(3) "*me dijo poco más o menos **esto**: "Llegará un día en que usted eche de menos…"*

Finally, other uses include what I have called indefinite-contrastive, because it does not refer to some particular referent (indefinite) and is always used in contrast with something else (contrastive) (example 4); it appears in correlations with elements like *aquello, lo otro* or *esto*:

(4) *"Una llamada de teléfono de vez en cuando. Compra **esto**, vende aquello, haz lo que yo te diga…"*

Two uses only found in the original texts are an idiomatic expression, *esto es,* in the sense of *id est*, (example 5) and a temporal meaning in the expressions: *en esto* and *a todo esto* in the sense of *then* (example 6). These uses are missing in the translations.

(5) *"sin embargo, el proceso contrario no es tan sencillo, **esto** es, que para ser alguien hace falta creerse nadie"*

(6) *"En **esto**, las chicas llegaron a su parada, se bajaron, y aunque las seguí disimuladamente ya no pude sino coger fragmentos incomprensibles de su conversación"*

A minor use, only present in P-ACTRES, is as a hesitation marker to translate er, erm and similar devices (example 7):

(7) *"Bueno, **esto**, señora, estamos en una misión" from English: "Well, er, madam, we are on a quest."*

3.2.Analysis of aquello

Table 4. Frequency rates, p-value, z-score and interpretation of the semantic functions of *aquello.*

	Occurrences in *CREA*	Occurrences in P-ACTRES	Occurrences per million words *CREA*	Occurrences per million words P-ACTRES	P-Value	Z-Score	Interpretation
anaphoric	119	70	47.84	166.24	0.0062	-2.74	Significant overuse
cataphoric	52	8	20.90	18.99	0.003859	3.51	Significant underuse
deictic	9	5	3.61	11.87	0.76932529	-0.31	Non-significant overuse
Indefinite-contrastive	3	3	1.2	7.12	0.35040222	-0.96	Non-significant overuse

The same but fewer functions have been found for *aquello*: Anaphoric meaning, again the most frequent (example 8); cataphoric meaning, more frequent in CREA than the cataphoric meaning of *esto* (example 9); deictic meaning (example 10) and indefinite-contrastive (example 11):

(8) *Esta noche no me dejan -contestó el niño, como si* ***aquello*** *le alegrara.*

(9) *unas precisas reglas de alimentación (por* ***aquello*** *de que de lo que se come se cría)*

(10) ***Aquello*** *olía a pintura.*

(11) *Él y Wes se pasaron unos minutos charlando forzadamente de esto y* ***aquello****, aunque en más de un momento la conversación estuvo casi a punto de naufragar por falta de terreno común.*

4. Discussion of findings

The analysis has shown that there is a tendency to overuse two of the three neutral demonstrative pronouns, i.e. esto and aquello and to underuse the other, i.e. eso in translations from English into Spanish. This seems to be an instance of translationese, as there are only two demonstrative pronouns in English, this and that and it may be easier to associate this with esto and that with aquello, leaving eso aside more often, which corresponds to the unique items hypothesis ((Tirkkonen-Condit: 2002: 209) – translated texts feature lower frequencies of linguistic elements that are unique or specific of the target language because they do not have a similarly perceived equivalent. Although generally applied to lexical items, this theory has also been applied to grammatical features (Rabadán et al, 2009).

Both *esto* and *aquello* with cataphoric reference are significantly underused. Structures like *aquello* followed by a relative clause or preposition *de* (sometimes with a direct quotation afterwards), as in examples 12 and 13, are much more common in original language, which may be due to the choice of other referential resources like *lo* in translations:

(12) *la memoria de **aquello** que hemos visto con la imaginación*

(13) *es lo primero que nos enseñan en el colegio, **aquello** de "prímer, segun, tércer...".*

There is also lack of *esto* with explanatory and temporal meanings in the translations, where again, translators seem to prefer other expressions, maybe *o sea* or *es decir* instead of *esto es* and *entonces* instead of *a todo esto, en esto.*

On the other hand, *esto* with deictic meaning and as a hesitation marker and *aquello* with anaphoric reference are significantly overused, which means that translators tend to use these demonstrative pronouns with these particular meanings too much to the detriment of other resources in Spanish, which can be explained in terms of the *simplification hypothesis* (Baker 1993) in various possible senses a) that only a few from a number of resources may have been used to express certain functions, thus not fully exploiting the lexical or grammatical richness of the language, b) that the lexical density and the type-token ratio may be reduced and c) that too much grammatical anaphora (pronouns) may have been used to create cohesion; consequently less lexical anaphora, i.e. synonyms, superordinates and repetition of lexical items.

5. CONCLUSION

This paper has approached translation quality assessment from the perspective that it can benefit from applying objective criteria, anchor phenomena, found by corpus-based research. This study also proposes pronouns esto and aquello as a new anchor phenomenon in all those functions that have been proved to be significantly over/underused, e.g. the higher the frequencies of anaphoric aquello in a translation, the lower the translation rates for quality, and so on. These criteria can be summarized into: The more the translation resembles non-translated Spanish in the use of neutral demonstrative pronouns, the higher it rates for quality.

Before the advent of corpora, linguists only had their intuition and the usage (sometimes inevitably biased) of informants as sources of data. Nowadays linguists have both at their disposal – intuition and corpora. If we end up gathering a series of anchor phenomena

and we combine them into a tool that can do part of the work of checking the occurrences of these grammatical items automatically, then we will be able to offer translation critics and translator trainers something else to complement their intuition.

REFERENCES

Baker, M. (1993). Corpus Linguistics and Translation Studies. In M. Baker et al (Eds.) *Text and Technology. In Honour of John Sinclair* (pp. 233-250). Amsterdam/Philadelphia: John Benjamins.

Baker, M. (1996). Corpus-based translation studies: the challenges that lie ahead. In H. Somers (Ed.) *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager* (pp. 175-186). Amsterdam: John Benjamins.

Izquierdo, M., Hofland, K. and Reigem, Ø. (2008). The ACTRES Parallel Corpus: an English-Spanish Translation Corpus. *Corpora* 3 (1): 31-41.

Labrador, B. A corpus-based study of the use of Spanish demonstratives as translation equivalents of English demonstratives. *Perspectives: Studies in Translatology*. In press.

Mauranen, A. (2004). Corpora, universals and interference. In A. Mauranen and P. Kujamäki (Eds), *Translation Universals. Do they exist?* (pp. 65-82). Amsterdam-Philadelphia: John Benjamins.

Preacher, K. J. (2001, April). Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence [Computer software]. Available from http://www.quantpsy.org.

REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [online]. *Corpus de referencia del español actual.* <http://www.rae.es> [25-01-2010]

Rabadán, R. (2005). Proactive Description for Useful Applications: Researching Language Options for Better Translation Practice. *Meta 50*(4). CD-ROM.

Rabadán, R. (2005-2008). Tools for English-Spanish Cross-linguistic Applied Research. *Journal of English Studies*. 5-6: 309-324.

Rabadán, R., Gutiérrez, C. and Ramón, N. (2007). Exploring Translation Research Applicability: Description for Assessment (ACTRES/TRACE). Paper presented at the

5TH EST Conference—European Society for Translation Studies. Ljubljana, Slovenia, September, 3-5.

Rabadán, R., Labrador, B. and Ramón, N. (2009). Corpus-based contrastive studies and translation universals: A tool for translation quality assessment English-Spanish. *Babel 55*(4), 303-328.

Ramón, N. and Labrador, B. (2008). Translations of '-ly' adverbs of degree in an English-Spanish Parallel Corpus. *Target* 20 (2): 275-296.

Tirkkonen-Condit, S. (2002). Translationese - a myth or an empirical fact? A study into the linguistic identifiability of translated language. *Target* 14 (2): 207-220.

Toury, G. (1995). *Descriptive Translation Studies and Beyond.* Amsterdam/Philadelphia: John Benjamins.

Tel, comparativo atípico del francés : una gramática de usos

SARAH LEROY

SYLVAIN LOISEAU

Université Paris Ouest Nanterre La Défense

Resumen

En este artículo proponemos una descripción en corpus de una de las construcciones comparativas que utilizan el lexema tel *en francés. Este lexema plantea numerosas dificultades debidas a la variedad de sus empleos, de sus usos, y de sus formas. Mostramos que los corpora permiten de una parte delimitar y circunscribir un empleo estabilizado, y por otra parte describirlo. Para identificar esta construcción, mostramos su emergencia y su estabilización en una diacronía amplia, del siglo XV al siglo XX. Para describirla, movilizamos los métodos estadísticos de medida de las atracciones entre una construcción y los lexemas asociados, para describir la estructura gramatical a través de las regularidades léxicas que ella establece y que se perfila en el corpus.*

Palabras clace: Comparación, uso, corpora diacrónicas, francés

Abstract

In this paper we aim at describing a comparative construction with the lexeme tel *in French using a corpus. Used in many forms and many functions, this lexeme is challenging grammatical description. We show that corpora allow both for identifying a stabilized usage of this lexeme that is a pertinent level of description, and for characterizing and describing this usage. In order to identify this emerging usage, we use a diachronic corpus covering written texts from XV to XX c. In order to describe it, we observe the lexemes statistically attracted in the construction.*

Keywords: comparison, usage, diachronic corpora, French

1. INTRODUCCIÓN[1]

Debido a la diversidad de sus formas, de sus significados y de sus empleos, el lexema *tel* no es objeto de un tratamiento sistemático en las gramáticas francesas. Por una parte el lexema mismo no se describe en la diversidad de sus funciones gramaticales y por otra, se le considera de modo periférico (en forma de observación, o en los ejemplos) en el tratamiento de las diferentes funciones gramaticales donde puede ser utilizado (comparación, determinación, intensidad, etc.).

A veces sin embargo puede funcionar como instrumento de comparación; en numerosos contextos, encontramos "empleos comparativos" donde *tel (que)* forma una construcción comparativa y conmuta fácilmente con otro MIS, o "marcadores de identidad similativa"

[1] Muchas gracias a C. Romero por su ayuda.

(Pierrard *et al.,* 2006): *ainsi que*, *de même que*, *aussi bien que*, y sobre todo *comme*, que representa el número más grande de empleos.

Nos interesamos por uno de los empleos comparativos de *tel*, construcción gramatical que muestra bien las dificultades de su tratamiento (1).

(1) *Après, une heure après, on retourne dans la rue tout fier et tout flageolant tel un athlète hébété. (J.-M.-G. Le Clézio, Le Procès-verbal, 1963)*

En este empleo, *tel* (o, a veces, *tel que*) es una aposición y seguido por un grupo nominal generalmente indefinido que hace referencia a un miembro cualquiera de la clase que designa[2].

Proponemos una gramática de usos de esta construcción que se apoya en un análisis sistemático de corpus. La observación de los datos nos permite mostrar que varios fenómenos deben ser tomados en consideración, principalmente la forma del comparativo, el tipo de grupo nominal introducido por *tel*, y sobre todo el tipo semántico del lexema nominal del comparado.

2. UNA CONSTRUCCIÓN EMERGENTE

Para observar este fenómeno, nos apoyamos en los datos propuestos por la base Frantext[3]. Segmentamos los casi 4000 textos y 240 millones de palabras de la base en 19 períodos de volúmenes variables[4], en los cuales retuvimos en primer lugar la totalidad de los casos de tel (que), con el fin de medir, mediante concordancias y una comprobación manual, la importancia de su empleo comparativo.

Estas observaciones muestran, en primer lugar, que se trata de un empleo muy marginal de *tel (que)*: si es cierto que está presente en cada uno de los períodos (aunque sólo por 4

[2] Otros dos empleos comparativos de *tel (que)* pueden ser identificados: uno, donde *tel que* introduce una subordinada comparativa no elíptica (« Il faudrait un homme tel qu'il n'en existe plus de nos jours et qu'il n'en renaîtra de longtemps. » (Guéhenno J., *Jean-Jacques*, 1952)), otro, donde *tel que* introduce una subordinada comparativa elíptica anafórica, el elemento que precede y el que sucede tienen un referente directo comun (« Je n'en dis pas davantage à un homme tel que vous, et je vous envoie mon cordial serrement de main, et l'assurance de ma haute considération. » (Hugo, *Actes et paroles IV*, 1885)).

[3] Base de datos textuales de más de 3700 textos franceses comprendidos entre los siglos XVI y XX, pertenecientes a los dominios de la literatura (80 % del corpus), artes, ciencias (http://atilf.atilf.fr/).

[4] Un período para el siglo XVI, dos (1601-1650, 1651-1700) para el XVII y para XVIII (1701-1750, 1751-1800), seis períodos para el siglo XIX (1801-1830, 1831-1840, 1841-1850, 1851-1865, 1866-1880, 1881-1900), y ocho períodos para el XX y los primeros años del siglo XXI (1901-1920, 1921-1930, 1931-1940, 1941-1950, 1951-1960, 1961-1975, 1976-1990, 1991-2007).

casos, es decir el 0,06% de los casos de *tel (que)* para el período 1651-1700), su frecuencia relativa media es del 1,57% sobre la totalidad de la base.

Sin embargo, esta medida baja no es representativa de la evolución de este empleo comparativo de *tel (que)*, que adquiere poco a poco amplitud en el curso de las décadas. Si es casi inexistente en primer lugar, hasta el siglo XIX (su frecuencia relativa es del 0,23% del 1501 al 1880), sube hasta frecuencias todavía bajas, pero sin embargo en progresión, en el curso del siglo XX: *tel (que)* comparativo representa el 2,35% de los casos de *tel (que)* entre 1881 y 1975. En último lugar, en el curso de los treinta últimos años (1976-2007), el empleo realmente despega, pasando al 6,15 %, luego al 11,46 % de los casos.

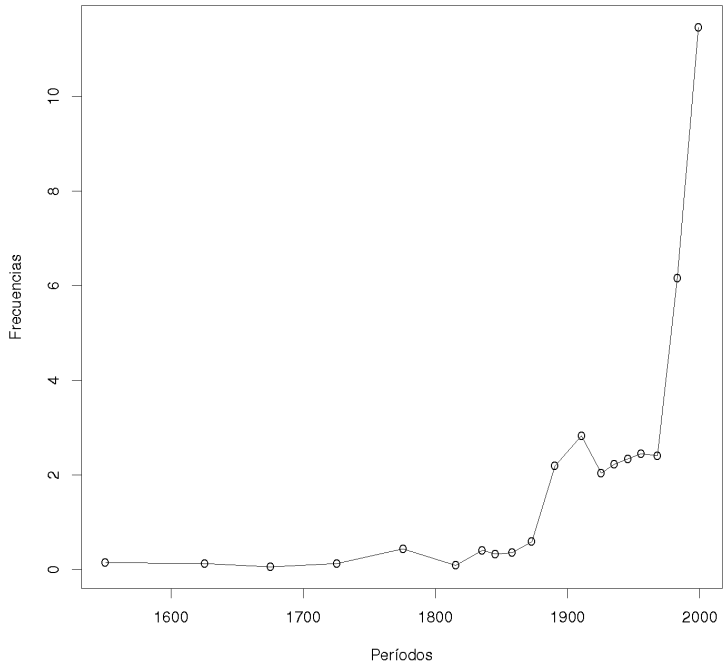


Figura 1: Evolución diacrónica

Parece pues que se trata de un fenómeno, si no completamente contemporáneo, por lo menos de una expansión reciente.

3. UNA CONSTRUCCIÓN POLIMORFA

No obstante, el examen de algunos ejemplos, a través de los siglos, muestra la diversidad de la construcción. Comprobamos en efecto, sobre los cuatro ejemplos más abajo, que la forma del comparativo mismo difiere (tel – (2) et (3) – o tel que – (4) et (5)), pero también la del elemento comparante, a la derecha del comparativo (nombre propio – (2) –, sintagma

nominal indefinido – ((2), (4) y (5) – o definido – (3)) ; por fin, la comparación misma puede efectuarse en una proposición (2), un nombre (3) o un verbo (4).

(2) *Tel qu'un Caesar, il fut grand en courage ; Tel qu'Adônis, il eut beau le visage. (P. de L'Estoile, Registre-journal du regne de Henri III, 1576-1578)*

(3) *Alban voyait le garçon devant lui, assis sur le seuil de sa cagna, avec la teinte grisâtre de son visage sculpté par la fatigue, la terre dans ses oreilles, ses cheveux qui lui descendaient sur le cou, rêches, incolores, tels que la toison d'une petite bête campagnarde. (H. de Montherlant, Le Songe, 1922)*

(4) *Le gaz pleure dans la brume, le gaz pleure, tel un œil. (J. Moréas, Les Cantilènes, 1886)*

(5) *Interdit, je restai un moment sans bouger, en regardant, de la barque, ce grand gars vêtu d'un pull étiré que le vent gonflait telle une courte robe de laine. (A. Makine, Le Testament français, 1995)*

3.1.Forma del comparativo

Así como se ve en los ejemplos citados arriba, las formas tel y tel que aparecen. Sin embargo, la forma tel es claramente más frecuente: representa el 88,76% de los casos, frente al 11,23% para tel que, hasta el punto que nos podemos interrogar sobre la importancia que debe tener, en la descripción de la construcción, la posibilidad de la forma comparativa tel que, muy utilizada para otros empleos (incluso comparativos), mientras que la forma tel, no seguida de que, es de un empleo adjetival más limitado.

3.2. Tipo sintáctico del comparante

El tipo de sintagma nominal que constituye el comparante, y está situado a la derecha de tel (que), indica de forma evidente una comparación de tipo estilístico, expresivo, que tiende a veces hacia el parangón. Encontramos mayoritariamente sintagmas indefinidos (66,24%), los sintagmas definidos llegan en la segunda posición (28,92%) y los nombres propios no determinados son los menos frecuentes (4,78%). Sin embargo, estas proporciones son más o menos las de las comparaciones en tel, mayoritarias[5]; las de las comparaciones en tel que, consideradas aisladamente, muestran una repartición más equilibrada entre sintagmas indefinidos (54,16%) y definidos (40,83 %)[6].

[5] Sintagmas indefinidos: el 67,76%, sintagmas definidos: el 27,42%, nombres propios sin determinación: el 4,8%.

[6] Los nombres propios representan el 5% de los casos.

3.3. Tipo lexical del comparante

Para explorar este aspecto, utilizamos un método estadístico que permite recolectar el léxico más "atraído" (Blumenthal, 2008) por la construcción comparativa, es decir significativamente sobrerrepresentado en los contextos de estas construcciones.

Este método probabilista, sólidamente definido desde hace muchos años (Lafon, 1980), recientemente conoció un repunte de atención en el marco de la lingüística cognitiva (Stefanowitsch & Gries, 2003).

Utilizamos aquí la ley hypergeométrica: una ventana de cinco palabras alrededor de los *tels (que)* es utilizada para oponer el subcorpus de los *tels (que)* comparativos a la totalidad de los *tels (que)*. Un indicio de asociación, logaritmo de la probabilidad de obtener por casualidad una frecuencia igual a la frecuencia observada en el subcorpus, es asociado con cada forma. Estas probabilidades son extremadamente bajas para las formas más atraídas, lo cual excluye que no haya asociación entre la construcción y estas formas.

Podemos así ver en las formas las más atraídas por los *tels (que)* comparativos:

des		Inf
le		Inf
qu		Inf
que		Inf
un		Inf
une		Inf
enfant		11.58
lune		10.13
elle		9.87
chien		9.73
tels		8.50
soleil		8.47
oiseaux		8.04
vieux		7.79
l		7.72
apparaît		7.63
balle		7.50
sur		7.48
qui		7.03
oiseau		6.51
phénix		6.45

es		6.41
conçoit		6.33
voyageur		6.24
qe		6.10
recouvrait		6.07
serpents		5.96
araignée		5.74
orphée		5.69
ombre		5.62
la		5.16
travers		5.16

Igualmente, las formas más rechazadas por los *tel (que)* comparativos son las siguientes:

luy		-5.16
estre		-5.27
estoit		-5.35
de		-5.35
mais		-5.42
moins		-5.70
mon		-5.77
si		-6.03
heure		-6.46
choses		-6.87
étoit		-6.94
maniere		-7.03
état		-7.37
fut		-7.88
cas		-8.10
bien		-8.31
manière		-9.13
n		-9.21
pas		-10.48
façon		-10.59
vous		-11.97

ne		-12.56
chose		-12.86
et		-15.42
point		-15.99
autre		-23.19
sont		-23.51
sorte		-27.32
est		-42.59
telle		-56.07
ou		-74.15)

Como podríamos esperarnos las comparaciones en *tel (que)* esencialmente introducen sintagmas nominales. El estudio de las especificidades muestra que, en el caso de los empleos comparativos estudiados, las palabras sobrerrepresentadas son un 65,62 % de los elementos susceptibles de figurar en un sintagma nominal: determinantes, nombres, adjetivo; a la inversa, otros empleos presentan sólo un 45,16 % de palabras sobrerrepresentadas susceptibles de figurar en un sintagma nominal, cuya repartición es por otro lado muy diferente: los nombres son mayoritarios, frente a solamente un adjetivo, un adverbio y dos grammemas.

Claramente vemos aparecer la naturaleza comparativa y expresiva, incluso orientada hacia el parangón, de la construcción estudiada, que contrasta con otros empleos de *tel*, si estudiamos los nombres sobrerrepresentados: esencialmente se trata de nombres concretos, mientras que los nombres subrepresentados, que se acercan a los *shell nouns* de Schmid (2000), son abstractos y poco referenciales, casi "huecos". Reconocemos, por otra parte, lexemas que componen con *tel* frases hechas (*sorte*) o diversos tipos de modismos (*façon*, *manière*).

Por otra parte, los numerosos lexemas atraídos por la construcción comparativa estudiada son unas lexicalizaciones sintéticas corrientes de semas intensivos y de topoï: *enfant* para la /juventud/ (6), *soleil* para /calor/ o / luminosidad/ (7), etc. Desde este punto de vista, podemos identificar pares de parangones antónimos: *lune* y *soleil*, *enfant* y *vieux*.

(6) *[…] tel qu'un jeune enfant, que poursuit la terreur, foible, il croiroit marcher environné d'horreur. (J.-A. Roucher, Les Mois, 1779)*

(7) *Et que dites-Vous de ces parterres éblouissants, tels que la Perse ne s'en est jamais glorifiée, tels que le soleil couchant n'en a jamais déballés sur la houle [...] (P. Claudel, Commentaires et exégèses. 4 : Le Cantique des cantiques, 1948)*

4. UNA CONSTRUCCIÓN EN PLENA MUTACIÓN

4.1. Evoluciones de la forma del comparativo

La repartición entre *tel* y *tel que* está claramente en relación con la evolución diacrónica; comprobamos en efecto una sustitución muy neta de tel que por tel a partir del período 1881-1900.

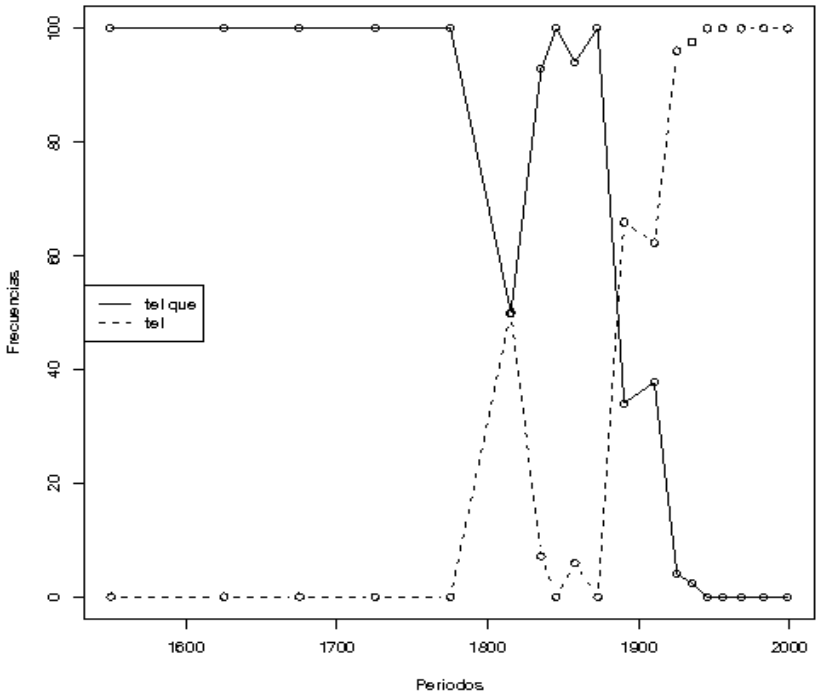


Figura 2: Reparto de *tel que* y *tel*

También comprobamos que este paso de *tel que* a *tel* coincide con el principio del aumento de las frecuencias de este empleo comparativo. Así, podemos pensar que es la aparición de una construcción comparativa en *tel* la que es notable, eliminando totalmente la construcción en *tel que* y progresando rápidamente.

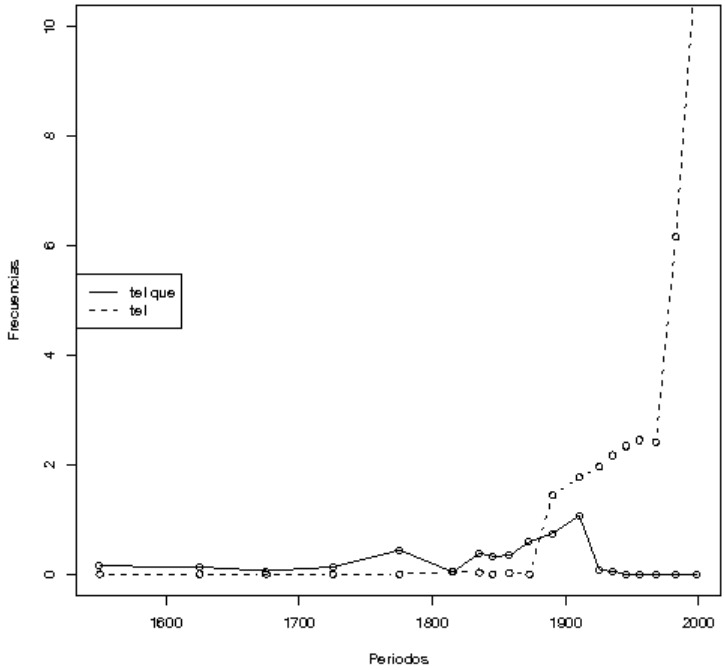


Figura 3: Frecuencias relativas

Más que de la evolución en el tiempo de una construcción comparativa, se trata de la emergencia de una construcción comparativa en *tel*, que se desarrolla real y rápidamente, mientras que la precedente, en *tel que*, se mantenía baja.

4.2. De tel que à tel: fin del cliché poético y nacimiento de una construcción comparativa

El estudio de las especificidades del léxico situado inmediatamente después del comparativo[7] ha sido llevado a cabo sobre un primer período, del 1501 al 1800, donde sólo *tel que* es representado, un segundo período, entre 1801 y 1940, donde tel que y tel son vecinos, y un tercero, del 1941 al 2007, donde sólo tel se mantiene.

En primer lugar, el tipo de palabras sobrerrepresentadas evoluciona fuertemente en el último período: aunque en las dos primeras los nombres respectivamente representan el 60 % y el 47 % de las formas sobrerrepresentadas, están totalmente ausentes en el tercero

Si se observa mejor estos nombres, vemos que los que se distinguen en el primer período corresponden a un léxico de la comparación clásica y literaria, incluso poético.

[7] Ventana de cinco palabras igualmente.

astre		4.47
milieu		4.08
adônis		3.37
brebis		3.37
caesar		3.37
nocher		3.37

Así, *nocher* aparece en una traducción del latín (8), *astre* en un cuento de estilo antiguo (9) y en un texto poético (10).

(8) *L'hiver, tel qu'un nocher qui, plein d'un doux transport, couronne ses vaisseaux triomphans dans le port, tranquille sous le chaume, à l'abri des tempêtes, l'heureux cultivateur donne ou reçoit des fêtes. (J. Delille, Les Géorgiques (trad.), 1770)*

(9) *Le temps approche où telle que l'astre du jour, lorsque du sein d'une nuée épaisse, il sort étincelant de lumière, la Messénie reparoîtra sur la scène du monde avec un nouvel éclat. (Abbé J.-J. Barthélemy, Voyage du jeune Anarchasis en Grèce dans le milieu du 4^{e} siècle avant l'ère vulgaire, 1788)*

(10) *Tel que l'astre aux flammes dorees Sortant des plaines azurees, Ce jeune Apollon reluira. (N. Vauquelin des Yveteaux, Œuvres poétiques, 1648)*

En el segundo período, volvemos a encontrar en parte este léxico y este tipo de comparación.

lune		6.10
travers		5.31
roses		3.12
bêtes		3.07
coryphée		3.07
judith		3.07
pâtre		3.07
rocher		3.07
tombes		3.07

Sin embargo, los términos son más corrientes y, si hay comparaciones muy literarias que se mantienen ((11), (12)), en otros, cierta trivialidad aparece, a propósito de la naturaleza de ambos elementos comparados como en (13), donde una temática sórdida va acompañada de una comparación elegante; en otros casos por fin, la elección de la comparación no carece

de ironía, como en (14), marcada por un desfase entre la situación descrita y el comparante escogido.

(11) *L'autre lune, croirais-tu, il a surpris un aigle ; il le traînait, et le sang de l'oiseau et le sang de l'enfant s'éparpillaient dans l'air en larges gouttes, telles que des roses emportées. (G. Flaubert, Salammbô, 1863)*

(12) *Quelle est celle-ci qui se tient debout en face de moi, plus douce que le vent l'été, telle que la lune à travers les jeunes feuillages? (P. Claudel, La Jeune fille Violaine,1892)*

(13) *La tête casquée de fleurs et de perles, vêtues d'amples blouses de soie et de larges pantalons brodés, immobiles et les mains sur les genoux, les prostituées, telles que des bêtes à la foire, attendent dans la rue, dans le pêle-mêle et la poussée des passants. (P. Claudel, Connaissance de l'Est,1907)*

(14) *Sur le trottoir qui fait face à la gare du Nord, Mme Cloche restait immobile et solitaire tel un rocher au milieu d'un torrent, contemplant les morceaux de son idéal, brisé par les Autres. (R. Queneau, Le Chiendent, 1933)*

Por fin, entre los elementos lexicales sobrerrepresentados en el tercer período, los constituyentes de un sintagma nominal, y en primer lugar el nombre, están ausentes.

En efecto, este tercer período es marcado por una "vulgarización" de la construcción. Ya no se limita a la expresión de una comparación estilísticamente marcada (aunque este tipo de empleo permanezca, como en (15)), sino que se presta a simples comparaciones de manera, sin ninguna dimensión poética (16).

(15) *Quand l'image cesse, la vitre gelée ne sort pas du champ de ma vision, telle une Bérénice d'alexandrins quittant un racinien Titus sur la scène du Théâtre-Français. (J. Roubaud J., La Boucle, 1993)*

(16) *Ils hantaient ces solitudes pour préparer du charbon de bois et vivaient dans des huttes, tels des sauvages. (C. Juliet, Accueils. Journal IV 1982-1988, 1994)*

Además, *tel*, ya simple morfema de comparación, puede directamente modificar adjetivos (17) o verbos (18), como en las comparaciones en *comme*, incluso a veces sin que la menor pausa sea marcada (19).

(17) *Une question précise, tel un bistouri, qui tranche l'abcès profond du secret. (B. Schreiber, Un silence d'environ une demi-heure, 1996)*

(18) *Domenica eût raison, tout en admirant les inflexions que prenait sa voix, et je craignais que celle-ci ne lui fît défaut lorsque, face à Marek, elle se dressa tel un cobra affolé à court de venin. (H. Bianciotti, Le Pas si lent de l'amour, 1995)*

(19) *Je suis arrivé au jour levant ma fille dormait dans son berceau il faisait obscur gris et calme tel un générique fin d'un film. (R. Morgiève, Ma vie folle, 2000)*

4.3. Tel en el paradigma de los comparativos

También hay que subrayar lo que la evolución de este comparativo tel tiene de original con relación a comparativos vecinos: con tel que, tenemos un MIS analítico (constituido por dos componentes); se sintetiza y producido tel, sinónimo de comme ; pero mientras que este último se deriva de la gramaticalización del segundo término del MIS analítico, el pronombre relativo, en el caso que nos ocupa es el primer término (adjetival) que es grammaticalizado. Entonces ninguno de otros MIS conoce esta evolución (18').

(18') **Elle se dressa, ainsi / de même / aussi bien un cobra affolé à court de venin.*

Por otro lado, esta emergencia de un comparado atípico todavía está en proceso. De una parte, la aparición de este nuevo morfema de comparación supone una modificación importante del sistema lingüístico, ya que esta nueva forma se inserta en un paradigma existente de comparativos. Por otra parte, esta nueva forma se añade a la forma existente de tel adjetivo; nos encontramos pues en un estado intermediario de coexistencia entre ambas formas, más precisamente en lo que Marchello-Nizia (2006: 258) nombra el « contexte de passage », en el cual « le nouveau sens permet au mot d'apparaître dans des contextes tout à fait nouveaux », entre los que se encuentra el empleo presentado aquí. La última etapa, donde la forma fuente y la forma meta pueden encontrarse en el mismo enunciado, si es posible, no parece estar atestiguada en nuestro cuerpo.

Podemos sin embargo evocar una de las causas posibles de esta gramaticalización de *tel* como comparativo. Al principio de un proceso de gramaticalización a menudo se encuentra un fenómeno semántico de subjectivación (Marchello-Nizia 2006: 26-28), por el cual « le locuteur rend son discours plus expressif pour l'allocutaire ». Propondremos ver en este fenómeno una de las causas mayores de la gramaticalización de *tel*. Si esta forma es sinónimo en efecto de *comme* en algunos de sus empleos, constituye una versión "más "fuerte", más intensiva y más orientada hacia un grado superior. El caso es que *comme* puede señalar la cuantificación de grado superior, pero no únicamente, mientras que *tel* se refiere

sistemáticamente a una forma más fuertemente evaluativa. La comparación con *tel* permite, en particular, una pausa prosódica que permite dar más expresividad al segmento comparativo conservando su carácter intensivo (20), mientras que *comme* en el mismo contexto es simplemente comparativo (20'), y necesita la supresión de la pausa para volverse intensivo (20'').

(20) *Cela commençait sur un rythme grave, tel qu'un chant d'église, puis, s'animant crescendo, multipliait les éclats sonores, s'apaisait tout à coup. (G. Flaubert, L'Education sentimentale, 1869)*

(20') *Cela commençait sur un rythme grave, comme un chant d'église, puis, s'animant crescendo, multipliait les éclats sonores, s'apaisait tout à coup.*

(20'') *Cela commençait sur un rythme grave comme un chant d'église, puis, s'animant crescendo, multipliait les éclats sonores, s'apaisait tout à coup.*

Este nuevo empleo de *tel* contribuiría pues a la renovación de estos morfemas que expresarían la intensidad y el grado superior, usos que suelen estar más expuestos al desgaste y al cambio.

5. CONCLUSIÓN

La construcción comparativa en tel estudiada en este artículo progresivamente emerge, desde un uso periférico permitido por este marcador de "comparación ficticia", hasta un uso estabilizado. En un plano gramatical, esta emergencia es asociada con el desarrollo de un empleo "desnudo" de tel, a partir de la construcción tel que. Desde un punto de vista semántico, este cambio se caracteriza por el paso de una "comparación virtual", a una comparación orientada hacia el parangón, luego a una comparación relativamente común, un paso que depende de un procedimiento clásico de debilitamiento semántico. La observación de los lexemas atraídos por la construcción permite caracterizar este cambio y esta dinámica semántica.

La utilización de corpus permitió describir el funcionamiento gramatical a través de las regularidades lexicales. También permitió asociar el análisis de la emergencia de la construcción en diacronía y el análisis de su funcionamiento. Desde este punto de vista, esta descripción se inscribe en la perspectiva de las gramáticas de los usos.

REFERENCIAS BIBLIOGRÁFICAS

Blumenthal, P. (2009). Éléments d'une théorie de la combinatoire des mots. *Cahiers de lexicologie*, 94: 11-29.

Blumenthal, P. (2008). Histoires de mots : affinités (s)électives. In *Actes du premier Congrès Mondial de Linguistique Française (CMLF 2008)*: 31-46.

Henry, A. (1991). *Tel* en français moderne. *Revue de Linguistique Romane*, 55: 339–426.

Kilgarriff, A. (2005). Language is never ever ever random. *Corpus Linguistics and Linguistic Theory*, 1:2: 263-276.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus, *Mots*, 1: 127-165.

Leroy, S. (2008). Changement et évolutions des emplois comparatifs de *tel (que)* (XVI^{e}-XX^{e} siècles). In Fagard B., Prévost S., Combettes B., Bertrand O. (éd.), *évolutions en français. Études de linguistique diachronique*. (pp. 231–348). Berne: Peter Lang.

Marchello-Nizia, C. (2006). *Grammaticalisation et Changement linguistique*. Bruxelles : De Boeck-Duculot.

Muller, C. (1990). *La Subordination en français. Le Schème corrélatif*. Paris: Armand Colin.

Pierrard, M., Van Raemdonck, D., Hadermann, P. (2006). Les marqueurs d'identité : subordonnants, coordonnants ou corrélateurs? In *Faits de Langue*, 28: 133-144.

Riegel, M. (1997). *Tel* adjectif. Grammaire d'une variable de caractérisation. In *Langue française*, 116: 81–89.

Schmid, H.J. (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. The Hague : Mouton de Gruyter.

Stefanowitsch, A. & Gries, S.-Th. (2003). Collostructions : Investigating the interaction of words and construction. In *International Journal of Corpus Linguistics*, 8/2: 209-243.

Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia: Benjamins.

The science of astronomy: passive constructions in eighteenth-century texts

PAULA LOJO SANDINO

Universidade da Coruña

Abstract

Scientists tend to use a specific language to show their theories and methods, and to prove their results. This language, or "jargon", is specific for each field of science, but it shares features which makes it identifiable as a language. However, languages change little by little, and therefore the period of time we are dealing with will determine the different characteristics we might face. The eighteenth century will be our context, and Astronomy our field of analysis. Once we are placed in this context, the aspect to be analysed is one of the more characteristic ones of the scientific language, the passive voice, and its use within this kind of texts.

Keywords: scientific language, astronomy, passive voice

Resumen

Los científicos suelen usar un lenguaje específico para mostrar sus teorías y métodos y para demostrar sus resultados. Este lenguaje, o "jerga", es específico para cada campo de la ciencia, pero muestra características que lo hacen identificable como lenguaje. Sin embargo, los lenguajes y lenguas cambian poco a poco y por lo tanto la época de la que estemos hablando determinará las diferentes características que podamos encontrar. Nuestro contexto será el siglo XVIII, y el campo de análisis la astronomía. Una vez ahí situados, el aspecto que vamos a analizar es uno de los más característicos del lenguaje científico, la voz pasiva y su uso en estos textos.

Palabras clave: lenguaje científico, astronomía, voz pasiva

1. INTRODUCTION[1]

Scientists tend to use a specific language to show their theories and methods, and to prove their results. This language, or "jargon", is specific for each field of science, but it shares features which makes it identifiable as a language. However, languages change little by little, and therefore the period of time we are dealing with will determine the different characteristics we might face. The eighteenth century will be our context, and Astronomy our field of analysis.

It is with the analysis of several texts that we can examine the use of the language and reach conclusions. The language in itself looses its simplicity when the text is written

[1] The research here reported on has been founded by Xunta de Galicia Programa de promoción xeral de investigación do Plan galego de investigación desenvolvemento e innovación tecnológica (INCITE), PGIDT07PXIB104160PR.

following the patterns of the scientific language, and some constructions are used in a different way than in everyday language.

The aim of the following analysis is to see the different, or similar, way in which passive constructions are used. We first need to see the context and corpus in the analysis, and their specific characteristics. The focus will be put mainly on the verb selection, the presence or absence of the *by*-phrase, and the reduced constructions. Other aspects such as the use of the construction *we + active verb* instead of the passive voice will be also considered in the analysis. All this will serve us to see what preferences these writers had regarding the use of the passive voice, always talking about the field of Astronomy, using for this purpose texts of that area of knowledge.

2. EIGHTEENTH CENTURY SCIENTIFIC WRITING IN ENGLISH

Before concentrating on the texts themselves we need to place them within a frame, to context the authors and their work; they all belong to the eighteenth century, which had a characteristic way of using the language to express ideas.

In the Enlightenment, "people spoke and wrote differently from the way in which they had spoken and written hitherto, and they would go on speaking and writing like that for some time to come." (Im Hof, 1994: 8) Those living in that time were under the influence of the scientific thinking, and their way of writing was changing, in part, due to the need of accommodation to science, since "science is an important candidate for promoting the growth of a Standard language because it uses a common set of methods and measurement-standards and is cumulative and self-referential", following Kaplan's (2001: 14) words.

The factor which most influenced the Enlightenment thinkers was the Scientific Revolution, which had taken place in the seventeenth century. These thinkers became fascinated by science, and because of that they went back in time to grab the essence of their thoughts; "it was the Enlightenment's *philosophers* who took up the science of the preceding age and helped to establish it as the dominant force in Western culture." (Henry, 2004: 10)

However, the concept of science people had during the Scientific Revolution, and even by the early 1700s, was slightly different; for them the term *concept* was related or linked to natural philosophy, and that is why philosophers were the ones closer to science. They tried to show and proved to everybody that science was interesting, but it took them more time than what they initially could have expected for the new doctrines to be completely accepted

in the new era, and "it was not until well into the eighteenth century, which makes a European 'scientific revolution' a thing of the Enlightenment." (French, 2003: 222)

Throughout the eighteenth century the new scientific ideas spread around the Western world; nevertheless, there were very few the institutions which were specialized on scientific issues. Although the Enlightenment helped these ideas to be known and studied, it was not until "about 1830 that scientific communities began to take meaningful shape and to function on a national and international scale." (Cahan, 2003: 11) Actually, it is in the course of the nineteenth century that all the previous scientific knowledge and ideas are brought to a successful conclusion, and are applied to real situations, in part due to the fact that "the nineteenth century witnessed the rise of modern industry." (Wengenroth, 2003: 221)

Thus, it is now that statements like the following will be taken as truth: "The substance of science comprises more than the discovery and recording of data; it extends crucially to include the act of interpretation." (Gopen and Swan, 2001) And the interpretation given by the scientist is expressed in the so called scientific language.

As Gopen and Swan (1990: 550) say, "The fundamental purpose of scientific discourse is not the mere presentation of information and thought, but rather its actual communication".

The term *scientific language* is a bit problematic in its definition, because there is not such a scientific language on its own; there are many different scientific languages depending on the branch of science we deal with. In the same way language is complex, science is as well, and as Gutiérrez Rodilla (1998:16) points out, "en cada una de sus ramas, se dan diferentes características y son cambiantes los recursos comunicativos."

The term *scientific language* encompasses all the different sublanguages used in the multiple branches of science, each one with its own characteristics. These characteristics, such as the linguistic structures used, are not chosen at random, but may have different purposes depending on the author.

These features include a monosemic language in order to avoid ambiguity, but different in each field of science. Syntax tends to be more fixed, although authors may use whatever structure they find more appropriate for their discourse. Among these structures we can find a prevalence of coordination and juxtaposition over subordination, although subordinates are frequent when adjectival and explicative. This is due to the abundant clarifications present in scientific texts, either in the form of apposition or between commas, dashes or parenthesis.

Other constructions used in scientific texts are those nominal over verbal ones, or impersonal constructions.

3. THE PASSIVE VOICE

As we can read in Banks (1996: 15), this scientific language, or scientific prose in this case, is said to be characterised by the use of the passive voice: "the passive form is a feature of scientific prose and that is so because of the 'impersonal' nature of scientific writing".

Linguistic users tend to resort to the passive voice to express or point out a specific aspect or relevant detail. Generally speaking, for a construction to be passive it must have the following characteristics: the subject of the passive is the direct object of its corresponding active; the subject of the active construction has to be expressed either in an agentive form or left unexpressed; and the verb has to be marked passive (Siewierska, 1984).

Passives may differ from actives not only in the pragmatic function of the patient and agent, but also syntactically, "in word order, case marking, verbal morphology and in the appearance of some additional word or particle." (Siewierska, 1984: 3) As Halliday (2004: 159) points out, "scientific texts are found to be difficult to read; and this is said to be because they are written in "scientific language", a "jargon" which has the effect of making the learner feel excluded and alienated from the subject-matter."

According to Palmer (1974: 81) "the passive is par excellence the grammatical structure that exemplifies the concept of transformation." The main use of the passive is to avoid naming the agent, that is, the subject of the active verb, because we do not know it or because we just do not want to say it. But, as we have already said above, we can also use the passive voice in order to emphasize part of the sentence.

Through the use of texts about Astronomy, we are going to see the different use of the passive voice in science, since, as Seoane Posse says (1999: 126), "the passive is specially significant in the description of scientific phenomena". Also our focus will be in what concerns to the presence or absence of the *by*-phrase and reduced constructions, as well as the verb used. And following Tarone et al.'s (1998) idea of the use of the pronoun *we* plus the active voice instead of the passive voice, some attention will be paid as well to this aspect. Their conclusion analysing scientific papers – astrophysics in this case- is "that the first person plural we + verb form occurs just about as often as the passive." (1998: 119)

Since we are going to analyse the use of the agent or *by*-phrase in our passive constructions, we could follow the distinction Keenan (1985) makes between *basic passives* and the rest – which could be divided as well into different types. Three are the characteristics Keenan (1985: 247) sees in this kind of passives, "no agent phrase is present,

the main verb (in its non-passive form) is transitive, and the main verb expresses an activity, taking agent subjects and patient objects."

According to this definition, part of our passive constructions, those with an agent represented by a *by*-phrase, would not be considered basic passives.[2]

We are going to follow the traditional requirements for a passive to be so; that is, verb *be* or *get* + past participle (with passive sense, not adjective), followed or not by a *by*-phrase denoting the agent.

4. CORPUS MATERIAL AND METHODOLOGY

My research will concentrate on eight texts dealing with astronomy and written in different decades of the eighteenth century. The texts have been extracted from the CETA (Corpus of English Text on Astronomy), which forms part of a bigger project, the *Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing*. Table 1 below shows the texts selected and the corresponding number of words:

Table 1: Corpus of analysis

Title, author	Publication date	Number of words
An Introduction to astronomy, geography navigation, and other mathematical sciences made easie by the description and uses of the cœlestial and terrestrial Globes. By Robert Morden.	1702	10,154
Astronomical Dialogues Between a Gentleman and a Lady. By John Harris.	1719	9,907
Practical astronomy, in the description and use of both globes, orrery and telescopes. By Samuel Fuller.	1732	10,232
Astronomy, in five books. By Roger Long.	1742	10,474
Astronomy explained upon Isaac Newton's principles and made easy to those who have not studied mathematics. By James Ferguson.	1756	10,519
The history of astronomy, with its application to geography, history, and chronology; occasionally exemplified by the globes. By George Costard.	1767	10,315
The universal system: or mechanical cause of all the appearances and movements of the visible heavens: shewing the true powers which move the earth and planets in the central and annual rotations. By John Lacy.	1779	5,908
A treatise on practical astronomy. By Samuel Vince.	1790	10,540

The selection of texts is not arbitrary, but it is done according to their publication date. In order to be representative of the whole 18th century, and following the theory that within

[2] Keenan justifies his division by saying that "they are the most widespread across the world's languages."(1985: 247)

30 years the language hardly changes (see Kytö *et al*.), these texts are not separated from each other enough to let the language change, but at the same time they cover all the century.

Table 1, as can be seen, contains samples of around 10,000 words, which makes a total of 78,049 on which my analysis will be based. One of the compilation principles observed is the number of 10,000 words per sample: "For each discipline we have selected two texts per decade, with each sample containing around 10,000 words, excluding tables, figures, formulas and graphs." (Parapar & Moskowich-Spiegel, 2007: 289-290).[3]

For the purpose of this analysis, we will take a close look at all the passive structures present in the selected texts using the *Coruña Corpus Tool.*[4] All forms – with apostrophe or wholly written – will be taken into account for the analysis, since they all fulfil the requirements established for a passive construction to be so.

5. ANALYSIS OF DATA

As we have already pointed out, the analysis will be focussed on the auxiliary verb, the lexical verb used, the by-phrases, and finally the use of the construction we + active verb instead of the passive voice

In these eight texts we can find 1920 passive voices. These constructions represent 2.4% of the total number of words. Scientific texts like these are likely to have a good number of passive constructions, because, as Goumovskaya (2007) points out, "The passive constructions are a helpful way of ensuring a smooth flow of ideas". The great majority of them appears together with an auxiliary verb, 77.4% of the total number – 1,487 constructions. (1) *and this Science which you are now inquiring into, is hence called the Doctrine of the Sphere* (Harris, 1719: 7); (2) *lower than if they were attracted in parallel lines* (Ferguson, 1756: 150). This figure includes those which do not appear next to the auxiliary verb, but which depend on it, even if there are words between them – adverbs, nouns, or whole sentences. (3) *the matter of the Moon was then created* (Costard, 1761: 289); (4) *These Material Spheres either in Wood or Brass, are now more exactly made and better contrived* (Morden, 1702: 7).

In what concerns the auxiliary verb used we know that the passive voice may be formed either with the verb *to be* or the verb *to get*. The main difference in use of these two verbs is

[3] There is an exception, the one from 1779 by John Lacy. But there is another text included in that decade which completes the number of words.
[4] The *Coruña Corpus Tool* is a concordance programme designed for specific searches. Parapar & Moskowich-Spiegel (2007).

their formality. The latter is more colloquial than the former, and thus, it is not present in these texts, not even once. This kind of writing is too high for a passive voice with the verb to get - unless we are reproducing textual oral language - since this feature is more typical of a spoken discourse rather than a written one. We also should take into account the fact that, following Denison (1993: 419), it is not until the mid-seventeenth-century that we find the first recorded passive with *get*, and here we are dealing with the 18th century. In all cases in these texts the auxiliary verb is *to be*, and in some occasions it is accompanied by a modal construction. (5) *such an instrument might be applied with advantage* (Vince, 1790: 13)

Those passive occurrences which do not have an auxiliary verb make 22.5%, with a total number of 433. (6) *it must be added to the angle shown by the index* (Vince, 1790: 12); (7) *according to the old Stile used by us here in England* (Morden, 1702: 11). These are a reduced form for a passive construction, since, as we have seen above, the passive voice in English requires an auxiliary verb. What happens here is that the auxiliary verb being omitted, it is implicitly understood due to the passive participle and the context of the sentence to which it belongs. Thus, the meaning of the verb is maintained and the passive construction identified.

In what concerns the verbs used by the scientists of the 18th century we have a great variety in these texts. We have found a total of 190 different verbs employed in the passive voice. Some of them occur more frequently than others. Table 2 below shows the verbs with the highest frequency – all of them over 2% of the total.

Table 2: Lexical verbs

Lexical / main verb	Number of occurrences
call	280
say	164
observe	94
make	86
see	71
set	66
draw	58
find	55
place	43
divide	42

In the English language, as in any other language, there are some verbs normally used in a field or register and which are not common in another one. From this list of verbs, we can observe that the most frequently used verbs are those typically found in expository prose of scientific kind (*call, find*) and others are characteristic of the astronomical discipline, including mathematical aspects (*divide, draw*).

In Table 3, there is a list of other verbs present in the texts which appear with a frequency of over 0.25%. Those present in less than 0.25% have been excluded because they are not relevant for the analysis.

Table 3: Other verbs and their frequency

Verb	Frequency	Percentage
Give	35	1.71
Suppose	31	1.51
Know	30	1.46
Show	28	1.37
Take / Turn	26	1.27
Attract	25	1.22
Measure / Fix	24	1.17
Add	20	0.97
Do	19	0.93
Carry	17	0.83
Express / Extend	16	0.78
Use	15	0.73
Mention / Move / Subtract	13	0.63
Eclipse	12	0.59
Affect / Contain / Show / Discover / Intercept / Determine / Explain	11	0.53
Conceive / Mark / Direct / Form / Project	10	0.48
Distinguish	9	0.44
Hide / Raise	8	0.39
Count / Fit / Require / Diminish	7	0.34
Cover / Hang / Note	6	0.29

The third aspect we are going to focus on is the *by*-phrase. When we talk about *by*-phrases in this context, we are making reference to those which introduce the agent of the passive construction. In our texts there are many *by*-phrases which cannot be considered as forming part of a passive construction; these are the prepositional complements, which belong to the verb used, not to the construction. This happens because of the type of discipline analysed, which includes mathematical structures, and verbs such as *divide* or *multiply*. These verbs are followed by the preposition *by*, but they, in most cases, only make reference to numbers, not to agents of the sentence. So we have to take special care with these verbs.

In all the passive constructions in the total of our eight texts, we have found 307 occurrences with agent, which make 17.5% of all the passives. If we take into account the fact that the passive voice is normally used in order to avoid the agent of the action, as Seoane Posse (2009: 366) says: "the primary motivation for the use of short passives is to background the agent/experiencer of the action by omitting it", this figure is quite big. The difference with any other situation where the passive voice might be used is that we are in front of scientific texts, where sometimes the agent of the action is even more important than the event in itself. Authors or peoples are important for the history, which is included in the writing of these texts. And if the passive voice is used instead of the active is simply because the author is describing a process, the emphasis is on the process. We are reading and analysing scientific texts, which require a specific writing which includes the use of the passive construction.

The final point to be dealt with is that of the use of the pronoun *we*. Out of a total of 78,049 words for our analysis, only 238 are this pronoun *we* + an active verb. According to Tarone *et al*. (1998) the frequency of this construction is comparable to that of the passive voice. However, in our analysis we have to disagree with them, since the number of passive occurrences is 1920, over eight times the number of the construction with *we*.

6. FINAL REMARKS

In order to recapitulate we need to remember that the use of the passive voice is something characteristic of the scientific writing, since facts can be explained more clearly without inferences. Contrary to what is expected with the use of the passive voice, in these particular texts, as we have seen, the presence of the agent is quite frequent, probably due to two different reasons. Firstly, these authors want to include in their discourses other colleagues, so that these may be identified as scientists as well – when they are not very famous at that time. The other reason would be the presence of classical references and authorities in the field, in order to reassert their own hypothesis or facts. If someone well-known in the field has already talked about that, it is handier to use that knowledge and work on it.

Another aspect of these texts is the auxiliary verb used in the passive voice. Either *be* or *get* are possible to form the passive voice, but *be-passives* are more much used in written language than *get-passives*, and that is why we do not find any example of passives with get in the texts.

Something similar happens with the verbs used in the passive occurrences. These verbs – the ones in Table 2 – are those normally employed when explaining something, and although they also appear in a different context, they are usually present in scientific prose. And others are the common ones for the discipline dealt with; in this case astronomy, which will include as well verbs from mathematics.

The fact that over 20% of the verbs in passive appear without an auxiliary verb gives us an idea of how languages tend to abbreviate, even in a scientific register, where repetitions are allowed in order to make understanding clear. Of course a scientific register as this one requires a complete context, that is to say, all sentences need to be fully formed in order to avoid ambiguities and make the meaning clear for the reader. In this sense verbs need their auxiliary, and that is why over 75% of these sentences have all their elements.

And in what makes reference to the use of the pronoun *we* followed by an active verb instead of a passive construction, as suggested by Tarone *et al*., the difference might be due, mainly not to the different fields of analysis – they deal with astrophysical papers – , but to the distance among them. Our texts belong to the 18th century, while those examined by Tarone *et al*. belong to the last quarter of the 20th century, which implies a quite important chronological distance, and probably involves a change of language use.

REFERENCES

Banks, D. (1996). The Passive and Metaphor in Scientific Writing. *Cuadernos de Filología Inglesa,* 5/2. 13-22.

Cahan, D. (2003). Looking at Nineteenth-Century Science: an Introduction. In D. Cahan (Ed.). *From Natural Philosophy to the Sciences. Writing the History of Nineteenth-Century Science* (pp. 3-15). Chicago: The University of Chicago Press.

Denison, D. (1993). *English Historical Syntax*. London: Longman.

French, R. (Ed.). (2003). *Medicine before Science. The Business of Medicine from the Middle Ages to the Enlightenment*. Cambridge: CUP.

Gopen, G.D. and Swan, J.A. (1990). The Science of Scientific Writing. *American Scientist,* vol. 78. 550-558.

Goumovskay, G. (2007). Scientific Prose Style. *English for Specific Purposes*. No. 3. Available at http://eng.1september.ru/articlef.php?ID=200701608

Gutiérrez Rodilla, B.M. (1998). *La ciencia empieza en la palabra. Análisis e historia del lenguaje científico*. Barcelona: Península.

Halliday, M.A.K. (2004). *The Language of Science*. Continuum: London.

Henry, J. (2004). Science and the coming of Enlightenment. In M. Fitzpatrick, P. Jones, C. Knellwolf & I. McCalman (Eds.). *The Enlightenment World* (pp. 10-26). London: Routledge.

Im Hof, U. (Ed.). (1994). *The Enlightenment*. Blackwell Publishers: Oxford.

Kaplan, R.B. (2001). English – the Accidental Language of Science? In A. Ulrich (Ed.). *The Dominance of Engloish as a Language of Science. Effects on Other Languages and Language Communities* (pp. 3-26). Berlin: Mouton de Gruyter.

Keenan, E.L. (1985). Passive in the world's languages. In T. Shopen (Ed.). *Language typology and syntactic description. Volume 1. Clause structure* (pp. 243-281). Cambridge: Cambridge University Press.

Kytö, M. *et al.* (2000). Building a bridge between the present and the past: A corpus of 19th-century English. *ICAME Journal* No. 24. 85-97.

Palmer, F.R. (1974). *The English Verb*. London: Longman.

Parapar, J. & Moskowich-Spiegel, I. (2007). The Coruña Corpus Tool. *Revista del Procesamiento de Lenguaje Natural* Vol 39. 289-290.

Seoane Posse, E. (1999). The consolidation of the indirect prepositional passive in Early Modern English: Evidence from the Helsinki Corpus. *Estudios Ingleses de la Universidad Complutense,* nº 7. 119-139.

Seoane Posse, E. (2009). Syntactic complexity, discourse status and animacy as determinants of grammatical variation in Modern English. *English Language and Linguistics* 13.3. 365-384.

Siewierska, A. (1984). *The Passive: A Comparative Linguistic Analysis*. Kent: Antony Rowe Ltd.

Tarone, E. *et al*. (1998). On the Use of the Passive and Active Voice in Astropysics Journal Papers: With Extensions to other Languages and other Fields. *English for Specific Purposes*. Vol. 17. No. 1. 113-132.

Wengeroth, U. (2003). Science, Technology, and Industry In D. Cahan (Ed.). *From Natural Philosophy to the Sciences. Writing the History of Nineteenth-Century Science* (pp. 221-242). Chicago: The University of Chicago Press.

El paradigma derivativo de los adverbios en el inglés antiguo

GEMA MAÍZ VILLALTA

Universidad de La Rioja

Resumen

El propósito de esta comunicación es examinar el paradigma derivativo de los adverbios en inglés antiguo. Se lleva a cabo un análisis exhaustivo de los derivados de los en torno a 21 adverbios básicos que proporciona la base de datos léxica del inglés antiguo Nerthus (www.nerthusproject.com), a partir de los cuales se organizan los correspondientes paradigmas derivativos, y se llega a la conclusión de que el paradigma adverbial tiene un peso específico destacable en la formación de palabras en inglés antiguo.

Palabras clave: morfología, formación de palabras, inglés antiguo, base de datos léxica.

Abstract

This paper deals with the derivational paradigm of Old English adverbs. An exhaustive analysis is carried out of derivatives from around 21 basic adverbs provided by the lexical database of Old English Nerthus (www.nerthusproject.com). Then, the derivational paradigms organized around these basic adverbs are compiled and the conclusion is reached that the adverbial class plays an important role in word-formation in Old English.

Keywords: morphology, word-formation, Old English, lexical database

1. INTRODUCCIÓN[1]

El objetivo de esta comunicación es analizar los paradigmas derivativos que se forman a partir de los adverbios que proporciona la base de datos léxica del inglés antiguo Nerthus (www.nerthusproject.com), que comprende alrededor de 30,000 predicados tomados princpalmente de A Concise Anglo-Saxon Dictionary de Clark Hall (1996), y en menor medida de *An Anglo-Saxon Dictionary* de Bosworth-Toller (1973) y *The Student's Dictionary of Anglo-Saxon* de Sweet (1976). Tras identificar estos paradigmas formados a partir de los adverbios básicos, analizaré las categorías gramaticales que surgen y el tipo de derivación.

Para esta comunicación, haré una distinción entre los siguientes procesos morfológicos: conversión es la extensión categorial de un predicado sin cambio formal; derivación cero es formación de palabras sin morfemas derivativos o por medio de morfemas flexivos; afijación es el uso de prefijos o sufijos para formar derivados; y composición, por último, es la combinación de dos formas libres. Sirvan como ilustración estos ejemplos:

(1) (Martín Arista, en prensa):

[1] Investigación financiada con cargo el proyecto FFI08-04448/FILO.

a. $[(\text{acan})_{V}\ (\text{ece})_{N}]_{N}$
derivación cero: *ece* ‘pain’

b. $[(\text{su:}\partial\text{erne 1})_{adj}\ (\text{su:}\partial\text{erne 2})_{adv}]_{adv}$
conversión: *su:ðerne 2* ‘southerly’

c. $[\text{æfter}]_{aff}[(\text{gengan})_{V}\ (\text{genga})_{N}]_{N}]_{N}$
prefijación: *æftergenga* ‘follower’

d. $[[\{\{[[\text{a:}]_{Aff}[\text{belgan}]_{V}]_{V}\}_{V}\ \{\text{a:bolgen}\}_{adj}\}_{adj}]_{adj}[\text{nes}]_{Aff}]_{N}$
sufijación: *a:bolgennes* ‘irritation’

e. $[[[\text{sweord}]_{N}\ [(\text{wyrcean})_{V}\ (\text{wyrtha})_{N}]_{N}]_{N}$
composición: *sweordwyrtha* ‘sword-maker’

En (1) los corchetes indican proceso morfológico y sus constituyentes mientras que los paréntesis indican relación dentro de un paradigma entre dos formas flexivas. Las cadenas derivativas ilustradas en (1) se caracterizan por dos propiedades morfológicas: recursividad y recategorización. En cuanto a la recursividad la forma prefijada *a:bolgen* ‘irritated’ en (1d) sirve como base al proceso de sufijación en *a:bolgennes* ‘irritation’, lo que constituye un proceso de afijación recursiva. Ejemplos de recategorización son los ejemplos de derivación cero *ece* (1a), conversión *su:ðerne 2* (1b) y *a:bolgennes* (1d) ya que cambian de categoría. Por otro lado, la categoría de la base del predicado no se modifica debido a los procesos morfológicos, como es el caso de *æftergenga* ‘follower’ (1c) y de *sweordwyrhta* ‘sword-maker’ (1e).

Junto con los procesos, la noción de paradigma derivativo es central para la discusión que se trae aquí. Un paradigma derivativo es el conjunto de predicados relacionados morfológica y léxicamente con un predicado básico, del que derivan. Un ejemplo de paradigma derivativo de verbo fuerte se ofrece a continuación:

(2) Paradigma derivativo del verbo fuerte *calan* ‘to grow cold or cool’:

> *a:calan* ‘to become frost-bitten’, *a:ce:lan* ‘to cool off’, *a:cealdian* ‘to become cold’, *a:co:lian* ‘to grow cold’, *ælceald* ‘altogether cold’, *brimceald* ‘ocean-cold’, *calan* ‘to grow cold or cool’, *ce:lan* ‘to cool, become cold’, *ce:ling* ‘cooling’, *ceald 1* ‘cool, cold’, *ceald 2* ‘coldness’, *cealde* ‘coldly’, *cealdheort* ‘cruel’, *cealdian* ‘to become cold’, *cealdnes* ‘coolness’, *ceolas* ‘cold winds’, *ciele* ‘coolness’, *cielegicel* ‘icicle’, cielewearte ‘goose-skin’, *co:l* ‘cool, cold’, *co:lnes* ‘coolness’, *edce:lnes* ‘refreshment’, *fæ:rcyle* ‘intense cold’, *forcilled* ‘chilled’, *(ge)ce:lnes* ‘coolness’, *(ge)co:lian* ‘to cool’, *gece:lan* ‘to quench (thirst)’, *hri:mceald* ‘icy cold’, *i:sceald* ‘icy cold’, *ofcalan* ‘to chill’, *ofercealdd* ‘excessively cold’, *sincald* ‘perpetually cold’, *sincaldu* ‘perpetual cold’, *sna:wceald* ‘icy cold’, *wælceald* ‘deadly cold’, *winterceald* ‘wintry-cold’.

Del paradigma de CALAN, el verbo fuerte *calan* es el único básico, hay 12 derivados cero, de los cuales 6 son verbos débiles (*a:ce:lan, a:cealdian, a:co:lian, cealdian*, *(ge)co:lian*), 4 son nombres (*ceald 2*, n; *ceolas*, mp; *ciele*, m; *sincaldu*, f) y dos adjetivos (*ceald 1* y *co:l*); 6 predicados sufijados, de los cuales 4 son nombres (*ce:ling*, f; *cealdnes*, f; *co:lnes*, f; *(ge)ce:lnes*, f), 1 adverbio (*cealde*) y un adjetivo (*forcilled*); 7 son predicados prefijados, de los cuales 3 son verbo (*a:calan*, fuerte; *gece:lan*, débil; *ofcalan*, fuerte), 1 es nombre (*edce:lnes*, f) y 3 adjetivos (*ælceald, ofercealdd, sincald*); y por último CALAN forma 10 compuestos, de los cuales, 3 son nombres (*cielegicel*, m; *cielewearte*, f; *fæ:rcyle*, m), y 7 son adjetivos (*brimceald, cealdheort, hri:mceald, i:sceald, sna:wceald*, *wælceald* y *winterceald*). En total, el paradigma derivativo de CALAN consta de 12 nombres, 13 adjetivos, 7 verbos débiles, 3 verbos fuertes y un adverbio, por lo tanto la categoría gramatical que forma con major frequencia es la de adjetivo, y el proceso derivativo el de derivación cero con 12 ejemplos. Las clases léxicas que no forma son las clases gramaticales de preposición, conjunción e interjección, y el proceso morfológico que tampoco contiene es la de conversión, pero por lo general, podemos encontrar al menos un ejemplo de cada categoría léxica major y de todos los procesos morfológicos.

Siguiendo a Hinderling (1967), se toma como punto de partida el paradigma derivativo de los verbos fuertes, ya que constituyen el núcleo de la derivación germánica. Sin embargo, esta comunicación pretende demostrar la existencia de paradigmas derivativos productivos creados a partir de de la categoría gramatical adverbio.

2. ANÁLISIS

A continuación ofrezco la lista de los paradigmas formados a partir de un adverbio básico y los predicados que hasta el momento hemos relacionado con dicho paradigma. Entre paréntesis aparece la categoría gramatical del predicado y su correspondiente proceso derivativo, esto es básico, convertido, derivado cero, prefijado, sufijado y compuesto. Aquellos predicados con el proceso llamado básico y cero derivado son dudosos por el momento (no queda claro si se engloban dentro de los básicos o son derivados cero). En total, hay 21 paradigmas, a los cuales pertenecen 290 predicados. Los paradigmas de adverbios son los siguientes:

(3) Æ:DRE 2, E:AC 1, ELLES, GE:O, GEOSTRA, LUNGRE, NI:∂ER 1, NU: 1, SAME, SAMEN, SAMOD, SELDAN, SIMBEL 1, SI:∂ 2, SUNDOR, ∂Æ:R 1, ∂ÆT 1, ∂E:AH 1, ∂ENDEN 1, ∂ON 1, WEST

Los predicados adscritos hasta el momento a estos paradigmas derivativos formados a partir de adverbios básicos y que constituyen el corpus de análisis de esta comunicación son los siguientes:

(4) Paradigmas a partir de adverbios básicos:

Æ:DRE 2 (1): *æ:dre 2* (adv; básico) 'at once, directly'

E:AC 1 (6): *e:ac 1* (adv; básico) 'also, likewise', *e:ac 2* (prep; convertido) 'together with, in addition to, besides' , *he:rto:e:acan* (adv; comp) 'besides', *∂æ:rto:e:acan* (adv; comp) 'besides, in addition to', *to:e:acan 1* (prep; conv) 'besides, moreover', *to:e:acan 2* (adv; prefijado) 'besides, moreover'.

ELLES (1): *elles* (adv; básico) 'in another manner, otherwise, else'.

GE:O (13): *ge:o* (adv; básico) 'once, formely', *ge:oabbod* (nombre,m; comp) `former abbot', *ge:odæ:d* (nombre,f; comp) 'deed of old, former deed', *ge:odæg* (nombre,m; comp) 'day of old', *ge:oge:ara* (adv; comp) 'of old', *ge:ohwi:lum* (adj; comp) 'of old', *ge:ole:an* (nombre,n; comp) 'reward for past deed', *ge:omagister* (nombre,m; comp) 'former teacher', *ge:omann* (nombre,m; comp) 'man of past times', *ge:ome:owle* (nombre,f; comp) 'aged wife', *ge:osceaft* (nombre,m; comp) 'destiny, fate', *ge:osceaftga:st* (nombre,m; comp) 'doomed spirit', *ge:owine* (nombre,m; comp) 'departed friend'.

GEOSTRA (2): *geostra* (adv; básico) 'yester', *geostran* (adv; sufijado) 'yester'.

LUNGRE (1): *lungre* (adv; básico) 'soon, forthwith, suddenly'.

NI∂ER 1 (34): *(ge)ni∂erian* (verbo débil; derivado cero) 'to depress, abase', *geni∂erigendlic* (adj; suf) 'deserving condemnation', *ni∂er 1* (adv; básico) 'below, beneath, down', *ni∂er 2* (adj; básico & derivado cero), *ni∂era:scu:fan* (str.verbo; comp) 'to push down', *ni∂era:settan* (verbo débil; comp) 'to set down', *ni∂era:sti:gan* (str.verbo; comp) 'to descend', *ni∂erbogen* (adj; comp) 'bent down', *ni∂erdæ:l* (nombre,m; comp) 'lower part', *ni∂ere* (adv; suf) 'below, down', *ni∂erecg* (nombre,f; suf) 'lower edge', *ni∂erflo:r* (nombre,f; comp) 'lower story', *ni∂erga:n* (irr.verbo; derivado cero) 'to descend', *ni∂ergang* (nombre,m; derivado cero) 'descent', *ni∂erheald* (adj; comp) 'bent downwards', *ni∂erhre:osende* (adj; comp) 'falling down', *ni∂erhryre* (nombre,m; comp) 'downfall', *ni∂erlæ:tan* (str.verbo; comp) 'to lose heart', *ni∂erlang* (adj; comp) 'stretching downward', *ni∂erlecgung* (nombre,f; comp) 'deposition, entombment', *ni∂erlic* (adj; suf) 'low, low-lying', *ni∂ernes* (nombre,f; suf) 'deepness, bottom', *ni∂eronwend* (adv; comp) 'downwards', *ni∂ersce:otende* (adj; comp) 'rushing downwards', *ni∂erscyfe* (nombre,m; comp) 'rushing downwards, descent', *ni∂ersettan* (verbo débil; comp) 'to set down', *ni∂ersige* (nombre,m; comp) 'going down', *ni∂ersti:gan* (str.verbo; comp) 'to descend', *ni∂ersti:gende* (adj; derivado cero) 'descending', *ni∂erstige* (nombre,m; derivado cero) 'descent', *ni∂ertorfian* (verbo débil; comp) 'to throw down', *ni∂erung* (nombre,f; suf) 'humiliation, abasement, downthrow, condemnation', *ni∂erweard* (adj; suf) 'directed downwards', *ni∂erweardes* (adv; suf) 'downwards, in a downward direction'.

NU: 1 (7): *efnenu:* (interj; comp) behold now', *nu: 1* (adv; básico) 'now, at present, at this time', *nu: 2* (conj; conv) 'now that, inasmuch as', *nu: 3* (interj; conv) 'lo!, behold!, come!',

nu:hwænne (adv; comp) ‘straightaway’, *nu:hwi:lum* (adv; comp) ‘now-a-days’, *nu:na* (adv; derivado cero) ‘now’.

SAME (14): *(ge)samhi:wan* (nombre,pl; pref) ‘members of the same household’, *sam* (conj; derivado cero) ‘whether, or’, *sam-* (prefix), *sam∂e* (conj; comp) ‘as well as’, *same* (adv; básico) ‘similarly, in the same way’, *samheort* (adj; pref) ‘unanimous’, *samhwylc* (pron; pref) ‘some’, *samli:ce* (adv; pref) ‘together, at the same time’, *sammæ:le* (adj; pref) ‘agreed, accordant, united’, *samra:d* (adj; pref) ‘harmonious, united’, *samræ:den* (nombre,f; pref) ‘married state’, *samswe:ge* (adj; pref) ‘sounding in unison’, *samwinnende* (adj; pref) ‘contending together’, *samwist* (nombre,f; pref) ‘living together, cohabitation’.

SAMEN (14): *a:samnian* (verbo débil; pref) ‘to gather something together’, *ætsamne* (adv; pref) ‘united, together, at once’, *bo:cgesamnung* (nombre,f; comp) ‘library’, *(ge)samnian* (verbo débil; derivado cero) ‘to assemble, meet, collect’, *(ge)samnung* (nombre,f; suf) ‘union, congregation’, *he:ahgesamnung* (nombre,f; comp) ‘chief synagogue’, *pre:ostgesamnung* (nombre,f; comp) community of priests’, *rihtgesamhi:wan* (nombre,pl; comp) ‘lawfully married persons’, *samen* (adv; básico) ‘together’, *samnunga* (adv; suf) ‘forthwith, immediately’, *samnungcwide* (adj; comp) ‘collect’, *samtinges* (adv) ‘all at once, immediately’, *samwræ:dnes* (nombre,f) ‘union, combination’, *unsamwræ:de* (adj; pref) ‘contrary, incongruous’.

SAMOD 1 (23): *gesamodlæ:can* (verbo débil; suf) ‘to bring together’, *samod 1* (adv; básico) ‘simultaneously, at the same time’, *samod 2* (prep; conv) ‘together with, at (of time)’, *samod∂yrlic* (adj; adj) ‘concordant’, *samodcumend* (adj; comp) ‘flocking together’, *samodeard* (nombre,m; comp) ‘common home’, *samodfæst* (adj; suf) ‘joined together’, *samodgang* (adj; comp) ‘continuous’, *samodgeflit* (nombre,n; comp) ‘strife’, *samodgesi:∂* (nombre,m; comp) ‘comrade’, *samodherian* (verbo débil; comp) ‘to praise together’, *samodherigendlic* (adj; comp) ‘engaged in worshipping’, *samodhering* (adj; comp) ‘praising’, *samodli:ce* (adv; suf) ‘together’, *samodrynelas* (nombre,mp; comp) ‘runners together’, *samodsi:∂ian* (verbo débil; derivado cero) ‘to accompany’, *samodspræ:c* (nombre,f; comp) ‘colloquy, conversation’, *samodswe:gende* (adj; comp) ‘consonantal’, *samodtang* (adj; comp) ‘continuous, succesive’, *samodwellung* (nombre,f; comp) ‘welding together (of substance in the birth of a bee)’, *samodwist* (nombre,f; suf) ‘a being one with’, *samodwunung* (nombre,f; comp) ‘common residence’, *samodwyrcende* (adj; comp) ‘co-operating’.

SELDAN (6): *seldan* (adv; básico) ‘seldom, rarely’, *seldcyme* (nombre,m; comp) ‘infrequent coming’, *sedli:ce* (adv; suf) ‘in a rare, strange manner’, *seldlic* (adj; suf) ‘rare, strange, wondrous’, *seldsi:ene* (adj; comp) ‘rare, extraordinary’, *seldum* (adv; suf) ‘seldom, rarely’.

SIMBEL 1 (7): *simbel 1* (adv; básico) ‘ever, always’, *simbelfarende* (adj; comp) ‘roving, nomanic’, *simbelgefe:ra* (nombre,m; comp) ‘constant companion’, *simble* (adv; derivado cero) ‘ever, for ever, always’, *simbles* (adv; suf) ‘ever, always’, *simblian 1* (verbo débil; derivado cero) ‘to frequent’, *simblunga* (adv; suf) ‘continually, constantly’.

SI:∂ 2 (6): *gesi:∂lic* (adj; suf) ‘intimate’, *si:∂ 2* (adv; básico) ‘late, afterwards’, *si:∂ 4* (conj; conv) ‘after, afterwards’, *si:∂boren* (adj; comp) ‘late-born’, *si:∂dagas* (nombre,mp; comp) ‘later times’, *si:∂li:ce* (adv; suf) ‘lately, after a time’.

SUNDOR (60): *a:synderlic* (adj; suf) ‘remote’, *a:syndrung* (nombre,f; suf) ‘division’, *ælsyndrig* (adj; pref) ‘quite apart, single’, *(ge)a:sundran* (verbo débil; pref) ‘to separate, divide’, *(ge)a:syndrian* (verbo débil; pref) ‘to separate, divide’, *(ge)syndrian* (verbo débil; derivado cero) ‘to sunder, separate’, *onsundran* (adv; pref) ‘singly, separately’, *onsundrum*

(adv; pref) ‘singly, separately’, *sunderanweald* (nombre,m; comp) ‘monarchy’, *sunderboren* (adj; comp) ‘reckoned apart, born of disparte parents’, *sunderfolgoð* (nombre,m; comp) ‘official teachership’, *sunderfre:odo:m* (nombre,m; comp) ‘privilege’, *sunderfre:ols* (nombre,m; comp) ‘privilege’, *sunderli:pes* (adv) ‘separately, specially’, *sundermæ:lum* (adv; suf) ‘separately, singly’, *sunderme:d* (nombre,f; comp) ‘private meadow’, *sundersto:w* (nombre,f; comp) ‘special place’, *sundor* (adv; básico) ‘asunder, apart’, *sundorcræft* (nombre,m; comp) ‘special power or capacity’, *sundorcræftigli:ce* (adv; suf) ‘with special skill’, *sundorcy:ðð̆u* (nombre,f; comp) ‘special knowledge’, *sundorfeoh* (nombre,n; comp) ‘private property’, *sundorgecynd* (nombre,n; comp) ‘special quality’, *sundorgenga* (nombre,m; comp) ‘solitary (animal)’, *sundorgere:fly* (nombre,n; comp) ‘ly reserved to the jurisdiction’, *sundorgifu* (nombre,f; comp) ‘special gift, privilege’, *sundorha:lga* (nombre,m; comp) ‘Pharisee’, *sundorly* (nombre,n; comp) ‘private property’, *sundorli:ce* (adv; suf) ‘apart’, *sundorli:f* (nombre,n; comp) ‘life in seclusion’, *sundorlic* (adj; suf) ‘special’, *sundormæsse* (nombre,f; comp) ‘separate mass’, *sundornotu* (nombre,f; comp) ‘special office’, *sundornytt* (nombre,f; comp) ‘special use, office or service’, *sundorriht* (nombre,n; comp) ‘special right, privilege’, *sundorseld* (nombre,n; comp) ‘special seat, throne’, *sundorsetl* (nombre,n; comp) ‘hermitage’, *sundorspræ:c* (nombre,f; comp) ‘private conversation’, *sundorweorðung* (nombre,f; comp) ‘special honour’, *sundorweormynt* (nombre,f; comp) ‘special honour’, *sundorwi:c* (nombre,n; comp) ‘separate dwelling’, *sundorwi:s* (adj; comp) ‘specially wise’, *sundorwine* (nombre,m; comp) ‘bosom friend’, *sundorwundor* (nombre,n; comp) ‘special wonder’, *sundoryrfe* (nombre,n; comp) ‘private inheritance’, *sundrum* (adv; suf) ‘singly, separately’, *synderæ:* (nombre,f; comp) ‘special law’, *synderli:ce* (adv; suf) ‘apart, in private’, *synderli:pe* (adj; comp) ‘peculiar, special’, *synderli:pes* (adv; suf) ‘singly’, *synderlic* (adj; suf) ‘singular, separate, special’, *synderlicnes* (nombre,f; suf) ‘separateness’, *synderly:pig* (adj; comp) ‘peculiar, special’, *syndrig* (adj; suf) ‘separate, single’, *syndrige* (adv; suf) ‘separately, specially’, *syndrigendlic* (adj; suf) ‘discretive’, *syndrigli:ce* (adv; suf) ‘specially, separately’, *syndriglic* (adj; suf) ‘special, peculiar’, *to:syndrian* (verbo débil; pref) ‘to sunder, separate’, *una:sundrodlic* (adj; pref) ‘inseparable’.

ðÆ:R 1 (31): *ðæ:r 1* (adv; básico) ‘there, thither, yonder, where’, *ðæ:ra:bu:tan* (adv; comp) ‘about the place’, *ðæ:ræfter* (adv; comp) ‘thereafter’, *ðæ:ræt* (adv; comp) ‘thereat’, *ðæ:rbig* (adv; comp) ‘thereby, thus’, *ðæ:rbinnan* (adv; comp) ‘therein’, *ðæ:rbufan* (adv) ‘besides that’, *ðæ:rin* (adv) ‘therein, wherein’, *ðæ:rinne* (adv) ‘therein’, *ðæ:rmid* (adv) ‘therewith’, *ðæ:rne:hst* (adv; comp) ‘next to that’, *ðæ:rof* (adv; comp) ‘thereof, of that’, *ðæ:rofer* (adv; comp) ‘over y above that’, *ðæ:ron* (adv; comp) ‘therein, thereon’, *ðæ:ronemn* (adv; comp) ‘alongside’, *ðæ:ronge:n* (adv; comp) ‘on the contrary’, *ðæ:ronuppan* (adv; comp) ‘thereupon’, *ðæ:rrihte* (adv; comp) ‘thereupon, forthwith’, *ðæ:rto:* (adv; comp) ‘thereto’, *ðæ:rto:ge:anes* (adv; comp) ‘on the contrary’, *ðæ:rðæ:r* (adv; comp) ‘wherever’, *ðæ:runder* (adv; comp) ‘beneath’, *ðæ:ruppan* (adv; comp) ‘thereon’, *ðæ:ru:t* (adv; comp) ‘outside, without’, *ðæ:rwið* (adv; comp) ‘against, in exchange for’, *ðæ:rymbe* (adv; comp) ‘thereabout, on that point’, *ðæ:rgemang* (adv; comp) ‘thereamong’, *ðæderlendisc* (adj; suf) ‘to that people, native’, *ðæ:rforan* (conj; comp) ‘before that’, *ðæ:r 2* (conj; conv) ‘there, thither’, *ðæ:ru:te* (adv) ‘outside, without’.

ðÆT 1 (5): *ðæt 1* (adv; básico) ‘that, so that, in order that’, *oððæt* (conj; comp) ‘until’, *ðæt 2* (conj; conv) ‘that, so that’, *ðæt 3* (demonstrative; conv) ‘that, so that’, *ðætte* (conj; conv) ‘that, so that’.

ðE:AH 1 (6): *ðe:ah 1* (adv; básico & derivado cero) ‘though, although, even if’, *ðe:ah 2* (conj; básico & derivado cero) ‘though, although’, *ðe:ahðe* (adv; comp) ‘although’, *ðe:ahhwæðere* (adv; comp) ‘yet, moreover’, *swa:ðe:ah* (adv; comp) ‘however, yet’, *swa:ðe:ahhwæðre* (adv) ‘however’.

∂ENDEN 1 (2): *∂enden 1* (adv; básico) ‘meanwhile’*, ∂enden 2* (conj; conv) ‘meanwhile’.

∂ON 1 (6), *∂on 1* (adv; básico) ‘then, now, thence’, *∂onne 1* (adv; suf) ‘then, therefore’, *∂onne 2* (conj; conv) ‘then, therefore’, *∂anon* (adv; derivado cero) ‘from that time or place’, *∂anonfor∂* (adv; comp) ‘after that, then’, *∂anonweard* (adv; comp) ‘departing hence’.

WEST (45): *bewestan 1* (prep; conv) ‘to the west of’, *bewestan 2* (adv; pref) ‘to the west of’, *nor∂anwestan* (adv; comp) ‘from the north-west’, *nor∂anwestanwind* (nombre,m; comp) ‘north-west wind’, *nor∂west* (adv; comp) ‘north-west’, *nor∂westende* (nombre,m; comp) ‘north-west end’, *nor∂westgemæ:re* (nombre,m; comp) ‘north-west boundary’, *su:∂anwestan* (adv; comp) ‘from the south-west’, *su:∂anwestanwind* (nombre,m; comp) ‘south-west wind’, *su:∂west 1* (nombre,m; derivado cero) ‘in the south-west’, *su:∂west 2* (adv; comp)’south-west’, *su:∂westerne* (adj; suf) ‘south-western’, *west* (adv; básico) ‘westwards’, *West-Centingas* (nombre,mp; comp) ‘people of West Kent’, *West-Dene* (nombre,mp; comp) ‘West Danes’, *West-Seaxe* (nombre,mp; comp) ‘West Saxons’, *West-We:alas* (nombre,mp; comp) ‘Cornishmen’, *westan* (adv; suf) ‘from the west’, *westane* (adv; suf) ‘from the west’, *westannor∂an* (adv; comp) ‘north-west’, *westansu:∂an* (adv; comp) ‘south-west’, *westansu:∂anwind* (nombre,m; comp) ‘south-west wind’, *westanweard* (adv; suf) ‘westward’, *westdæ:l* (nombre,m; comp) ‘west quarter’, *westende* (nombre,m; comp) ‘west end’, *westerne* (adj; suf) ‘western’, *westhealf* (nombre,f; comp) ‘west side’, *westlang* (adv; comp) ‘extending westwards’, *westmearc* (nombre,f; comp) ‘western boundary’, *westnor∂lang* (adv; comp) ‘extending north-westwards’, *westnor∂wind* (nombre,m; comp) ‘north-west wind’, *westri:ce* (nombre,n; comp) ‘western kingdom’, *westrihte* (adj; comp) ‘westward’, *westrihtes* (adv; suf) ‘due west, westwards’, *westrodor* (nombre,m; comp) ‘western sky’, *westsæ:* (nombre,f; comp) ‘western sea’, *westsu:∂ende* (nombre,m; comp) ‘south-west extremity’, *westsu:∂wind* (nombre,m; comp) ‘south-west wind’, *westweard 1* (adv; conv) ‘westwards’, *westweard 2* (adj; suf) ‘westerly’, *westweg* (nombre,m; comp) ‘western way’, *westwind* (nombre,m; comp) ‘west wind’, *wi∂westan* (adv; pref) to the west of’.

En total tenemos 21 paradigmas derivativos y 290 predicados. Las principales categorías gramaticales que surgen a partir de estos paradigmas son las categorías del nombre, adjetivo, adverbio, verbo, preposición, conjunción e interjección. A continuación hago un recuento del número total de cada categoría y del correspondiente proceso morfológico:

(5) categorías gramaticales y procesos derivativos a partir del paradigma derivativo de los adverbios:

nombre (98): 3 derivado cero, 3 prefijado, 6 sufijado, 85 compuestos.

adjectivo (48): 1 básico y derivado cero, 1 derivado cero, 8 prefijado, 16 sufijado, 22 compuestos.

adverbio (104): 20 básico, 1 básico y derivado cero, 3 derivado cero, 1 convertido, 7 prefijado, 23 sufijado, 45 compuestos.

verbo (19): 5 derivado cero, 4 prefijado, 1 sufijado, 9 compuestos.

preposición (4): 4 convertido.

conjunción (12): 1 básico y derivado cero, 1 derivado cero, 7 convertido, 3 compuestos.

interjección (2): 1 convertido, 1 compuestos.

En total, el número de cada proceso derivativo es el siguiente:

(6) número total de los procesos morfológicos:

básicos: 20

básicos y derivados cero: 3

derivados cero: 12

convertidos: 14

prefijados: 23

sufijados: 46

compuestos: 165

Hay seis predicados que no han sido analizados por no disponer de datos suficientes: *ðæ:ru:te* (adv) 'outside', *swa:ðe:ahhwæðre* (adv) 'however', *sam-* (prefix), *samtinges* (adv) 'all at once', *samræ:dnes* (nombre,f) 'union, combination' y *sunderli:pes* (adv) 'separately, specially'. Un posible análisis sería considerar ðæ:ru:te (adv) como compuesto, ya que en la base de datos léxica Nerthus existe tanto *ðæ:r 1* (adv) como *u:te* (adv) 'out' como predicados libres. En esta línea swa:ðe:ahhwæðre (adv) también se puede considerar como compuesto de *swa: 1* (adv) 'so as, consequently' y de *ðe:ahhwæðre* (adv) 'yet, moreover'. *sam-* (afijo) de momento se puede considerar como prefijado. samræ:dnes tambien sería un prefijado de *sam-* (afijo) mas *(ge)ræ:dnes* (nombre,f) 'agreement'. En cuanto a *sunderli:pes*, por analogía con otros predicados cuyo primer elemento es *sunder-*, podemos clasificarlo como compuesto de *sundor* (adv) 'asunder' y de un segundo elemento hipotético, ya que no he encontrado *li:pe** como predicado libre, sino siempre como segundo elemento de predicados tales como *a:nli:pe 2* (adv) 'alone'. El mismo problema aparece con el predicado *samtinges* (adv), qe podría considerar de momento como prefijado.

En resumen, los paradigmas de adverbio forman en su mayoría adverbios (104 de 290, es decir, un 35, 8% del total), seguido de la formación de nombres con 98, adjectivos con 48,

verbos con 19 predicados y en menor medida conjunciones con 12. En cuanto al proceso morfológico, el más productivo es la formación de compuestos (165 de 290 son compuestos, el 56,8% del total de predicados), empezyo por los nombres compuestos (85 de 98 nombres, es decir, el 86,7% del total de nombres), seguido de adverbios compuestos (45 predicados) y adjetivos compuestos con 22 predicados. Le sigue la formación de predicados mediante sufijación con 46 predicados y de prefijación con 23 predicados. En total, 69 predicados están formados mediante afijación. Particularmente, SUNDOR forma 32 nombres, de los cuales 30 son compuestos, seguido por WEST con 24 nombres, 23 compuestos. Por otro lado, *∂Æ*:R 1 es el paradigma que más adverbios crea, 28 adverbios, de los cuales 26 son compuestos.

Desde una perspectiva más explicativa, Kastovsky (1992:362) define los compuestos como *complex lexical items consisting of two or more lexemes, e.g. deofol-guld-hus 'heathen place'.* Este autor distingue entre compuestos nominales, adjetivales y verboales, este último grupo restringido a adverbios y preposiciones como primer elemento del compuesto. Kastovsky también menciona dos casos en cuanto a los compuestos adverbio+nombre: o bien el adverbio se combina con un primario independiente o un nombre derivado como en *oferealdorman* 'chief officer', o bien se combina con una forma compuesta derivativa como es el caso de *oferleones* 'transgression' < *oferleoran* 'transgress'. Kastovsky da una serie de advervios que aparecen como primer elemento de los compuestos de adverbio + nombre. Estos adverbios son los siguientes:

(7) Adverbios como primer elemento en los compuesto adverbio + nombre (Kastovsky 1992):

> *æt* 'at to, near', *an* 'single, alone, only; numeral one', *eft* 'again, anew', *fore* 'front, beforehy', *for∂* 'forth, forward, away, front', *in* 'within, inside', *innan* 'inside', *mid* 'together', *ofer* 'over, above; very much, in excess', *on* 'forward, onward', *ongean* 'again, against', *samod* 'simultaneous, together', *under* 'under; inferior, secondary', *wi∂er* 'against', *ymb* 'about, around'.

Sólo *samod* aparece como primer elemento. A esta lista habría que añadir los predicados expuestos en (3) que formen compuestos nominales a partir de un adverbio como primer elemento. En mi análisis de los paradigmas derivativos, he hallado los siguentes paradigmas que crean nombres compuestos: GE:O crea 10 compuestos nominales como *ge:o* como primer elemento, SAMOD crea 8 compuestos nominales con *samod* como primer elemento, SELDAN crea el compuesto nominal *seldcyme* (nombre, m) 'infrequent coming', SI:*∂* 2 crea *si:∂dagas* (nombre, mp) 'later times' como compuesto nominal, SIMBEL 1 crea *simbelgefe:ra* (nombre, m) 'constant companion', SUNDOR crea 23 nombres compuestos con *sundor* como primer elemento, y 6 nombres compuestos con *sunder* como primer

elemento y WEST crea 18 compuestos nominales con *west* como primer elemento. Por lo tanto, a la lista que da Kastovsky podríamos anñadir los adverbios básicos *ge:o*, *seldan*, *si:∂ 2*, *simbel 1*, *sundor* y *west* como primer elemento en la formación de compuestos nominales adverbio + nombre.

Para concluir con el análisis, me detendré en el caso especial del paradigma derivativo de NI:∂ER 1. A partir de las formas flexivas atestiguadas de comparativo y el superlativo, se crea el predicado hipotético *ni:∂er 2* (adj). Si *ni:∂er 2* (adj) fuera el predicado básico que diera pie al paradigma derivativo, tendríamos el paradigma derivativo adjetival NI:∂ER 2. En este caso, *ni:∂er 1* (adv) pasaría a ser un convertido de *ni:∂er 2* (adj) teniendo como premisa que el proceso derivativo sigue la dirección verbo fuerte > nombre > adjetivo > adverbio > verbo débil. Sin embargo, al no encontrar la forma atestiguada de *ni:∂er 2* (adj), tomamos el adverbio *ni:∂er 1* (adv) como punto de partida, y creamos el paradigma derivativo a partir del adverbio básico. Por lo tanto, nos hallamos ante un paradigma derivativo hipotético, por lo que en un futuro podríamos hablar de NI:∂ER como paradigma derivativo de adjetivo en vez de adverbio. De momento, tomaré este paradigma derivativo como paradigma adverbial.

3. Conclusión

Tras analizar los 21 paradigmas derivativos adverbiales hallados en Nerthus, podemos concluir que los paradigmas formados a partir de un adverbio básico son productivos, ya que encontramos 290 predicados englobados dentro de estos paradigmas. La categoría gramatical más productiva es la creación de adverbios a partir de otros adverbios (104), seguido por la creación de nombres (98), y el proceso derivativo más productivo es la composición, en especial, la creación de nombres compuestos (85 predicados). El paradigma derivativo SUNDOR (60) es el que más predicados forma hasta el momento, seguido por NI:∂ER 1 (34) y ∂Æ:R 1 (31).

Partiendo de la premisa de que los predicados en inglés antiguo se forman en torno a paradigmas derivativos de verbos fuertes, también existen predicados formados en torno a paradigmas derivativos de adjetivos, verbos débiles y, en el caso que nos ocupa, adverbios. Por lo tanto, la categoría adverbio es productiva dado que en torno a los paradigmas derivativos de adverbios básicos se crean no solo adverbios, sino también nombres, adjetivos, verbos, preposiciones, conjunciones e interjecciones.

Para la investigación futura queda añadir más predicados a estas categorías y revisar los predicados existentes, pues la agrupación de predicados en torno a los distintos paradigmas

derivativos está actualmente en proceso de análisis, ya que muchos de los predicados, y especialmente los compuestos, están agrupados en torno a dos o más paradigmas, por lo que se produce un solapamiento de paradigmas.

REFERENCIAS BILBIOGRÁFICAS

Bosworth, J. y Toller, T. N. (1973; 1898) *An Anglo-Saxon Dictionary*. Oxford: Oxford University Press.

Caballero González, L. et al. (2004-2005). Predicados Verbales primitivos y derivados en inglés antiguo. Implicaciones para la elaboración de una base de datos léxica. Vol. 17-18: 35-49.

Clark Hall, J. R. (1996; 1896) *A Concise Anglo-Saxon Dictionary*. Toronto: University of Toronto Press.

Hinderling, Robert. (1967). *Studien zu den starken Verboalabstrakta des Germanischen*. Berlin: Walter de Gruyter.

Kastovsky, D. (1992). Semantics y vocabulary. Hogg (ed.) *The cambridge History of the English Language I: The Beginnings to 1066.* Cambridge: Cambridge University Press. 290-224.

Kastovsky, D. (2006). Typological changes in derivational morphology. Kemenade and Los 2006. 151-176.

Sweet, H. 1976 (1896) *The Student's Dictionary of Anglo-Saxon*. Cambridge: Cambridge University Press.

Martín Arista, J. Building a lexical database of Old English: issues and landmarks. J. Considine, (ed.) *Current projects in historical lexicography*. Newcastle: Cambridge Scholars Publishing. En prensa.

Martín Arista, J. et al. (2009). *Nerthus: An Online Lexical database of Old English*. http://www.erthusproject.com.

Obtaining computational resources for languages with scarce resources from closely related computationally-developed languages. The Galician and Portuguese case.

PAULO MALVAR FERNÁNDEZ, JOSÉ RAMON PICHEL CAMPOS, OSCAR SENRA GÓMEZ,

*Area of Language Technology, **imaxin**|software, Santiago de Compostela*

PABLO GAMALLO OTERO,

Department of Spanish Language, Faculdade de Filologia, Universidade de Santiago de Compostela

ALBERTO GARCÍA,

Engineering department of Igalia, A Coruña

Abstract

*In order to build many statistically-driven NLP tools, it is essential to use a significantly large amount of data. To overcome the limitation of the scarcity of computational resources for languages such as Galician it is necessary to develop new strategies. In the case of Galician, well-known romanicists have theorized that Galician and Portuguese are two varieties of European Portuguese. From a pragmatic standpoint, this assumption could open up a new line of research to supply Galician with rich computational resources. Drawing from the ENGLISH-Portuguese Europarl parallel corpus, **imaxin**|software has compiled an English-Galician parallel corpus that we used to build an English-Galician Statistical Machine Translation prototype whose performance is comparable to Google Translate. We contend that this strategy can be implemented to develop a great variety of computational tools for languages like Galician that are closely related to languages for which there already exist great computational resources.*

Keywords: Parallel corpus, English, Galician, Portuguese, Statistical Machine Translation

Abstract

*Para desarrollar muchas herramientas estadísticas de Procesamiento del Lenguaje Natural resulta esencial utilizar grandes cantidades de datos. Para salvar la limitación de la escasez de recursos computacionales para lenguas, como el gallego, es necesario diseñar nuevas estrategias. En el caso del gallego, importantes romanistas han teorizado que gallego y portugués son dos variantes del portugués europeo. Desde un punto de vista pragmático, esta hipótesis podería abrir una nueva línea de investigación para proporcionar al gallego ricos recursos computacionales. Partiendo del corpus paralelo inglés-portugués Europarl, **imaxin**|software ha compilado un corpus paralelo inglés-gallego que hemos utilizado para crear un prototipo de traductor automático estadístico inglés-gallego, cuyo rendimiento es comparable a Google Translate. Sostenemos que es posible implementar esta estrategia para desarrollar una gran variedad de herramientas computacionales para lenguas, como el gallego, íntimamente relacionadas con lenguas que ya cuentan con un gran repertorio de recursos computacionales.*

Palabras clave: Corpus paralelo, Inglés, Gallego, Portugués, Traducción Automática Estadística

1. PREFACE

From the point of view of Hallidayan Functional-Systemic Linguistics (FSL) theory, language serves, according to Gee (1999: 1) "as both a tool for action and a scaffolding for 'human affiliation within cultures and social groups and institutions"'. In other words, language works as a tool not only for communication but also for negotiating the relationships and the structures of society itself. It is precisely through this social dimension that language manages to play an extremely crucial symbolic role.

In developing computational tools for particular languages, computational linguists, whether they are primarily computer scientists or linguists, have a responsibility to the language(s) they are working with. It is possible that for high-prestige languages this responsibility is not obvious. In these cases, decisions about which linguistic phenomena are to be studied and, more importantly from the point of view of this paper, which tools are to be developed may seem trivial because they seem to not imply any particular ideological position. However, for those scientists who work with and for minoritized languages, especially if they are speakers of those languages, their decisions are never innocuous.

It is with this responsibility as language researchers and as speakers firmly in mind that this work is has been undertaken.

2. INTRODUCTION

In 2008 and 2009, at **imaxin**|software we carried out a project, subsidized by the *Dirección Xeral de I+D* of the *Xunta de Galicia*, called "RecursOpentrad: Recursos lingüístico-computacionais para a tradución automática avanzada de código aberto para a integración europea da lingua galega"[2]. In this project, in addition to building an English-Galician Ruled-based Machine Translation (RBMT) system, we thought that, given the progress[3] achieved in the field of Statistical Machine Translation (SMT), it was an excellent moment for taking a step forward in the development of Natural Language Processing (NLP) tools for Galician.

When we decided to develop a prototype of a SMT system for English and Galician, we knew that "the larger the available training corpus, the better the performance of a translation system" (Popović & Ney, 2006: 25) we would achieve. However, while gathering the

[2] RecursOpentrad: Linguistic and computational resources for advanced open source machine translation for European integration of the Galician language.

[3] Serve, just as an example, how high-quality SMT has become popular with the implementation done by Google of their statistical machine translation system, Google Translated (freely available at http://translate.google.com/).

necessary resources for the development of such a prototype for the above-mentioned pair of languages we came to same conclusion with what Popović & Ney (2006) began their paper given at Language Resources and Evaluation (LREC) in 2006:

> Whereas the task of finding appropriate monolingual text for the language model is not considered as difficult, acquisition of a large high-quality parallel text for the desired domain and language pair requires a lot of time and effort, and for some languages is not even possible. (25)

It is worth noting that it is possible to find English-Galician parallel corpora[4] published under the General Public License (GPL) on the Internet. Xavier Guinovart's research group at the Universidade de Vigo's Facultade de Traducción e Interpretación has gathered a collection of parallel corpora[5] within which the pair English-Galician is represented with a subcorpus of approximately 9 million words, which is by all means insufficient for the purpose of building a SMT system.

At this point it was clear for us that we needed to take a different route in order to achieve our goal. It is well-known in the linguistic community that important romanicists, such as Coseriu (1987), Cunha & Cintra (2002) and Aracil (1985), have theorized that, from a linguistic point of view, Galician should be considered a variety of Portuguese together with European, Brazilian, Asiatic and African Portuguese varieties. This is exactly what Fernández Rei (1991), member of the *Real Academia Galega*[6], and Coseriu (1987), one of the most important romanicist of the XXth century, point out:

> Na actualidade, desde o punto de vista estrictamente lingüísitico, ás dúas marxes do Miño fálase o mesmo idioma, pois os dialectos miñotos e transmontanos son unha continuación dos falares galegos, cos que comparten trazos comúns que os diferencian dos do centro e sur de Portugal; pero no plano da lingua común, e desde unha perspectiva sociolingüísitca, hai no actual occidente peninsular dúas linguas modernas, con diferencias fonéticas, morfosintácticas e léxicas, que poden non impedi-la intercomprensión ó existir un bilingüismo inherente entre o galego e o portugués, semellante ó existente entre o catalán e o occitano, o danés e o noruegués, o eslovaco e o checo, o feroés e o islandés.[7] (Fernández Rei, 1991: 17-18)

---

[4] Thanks to the localization projects of open source tools and operating systems carried out by the Galician open source community it is possible to manually collect domain-specific English-Galician parallel corpora. However, these corpora lack uniformity and, once again, they are insufficient for the purpose of building a SMT system.

[5] This collection of parallel corpora can be consulted at http://sli.uvigo.es/CLUVI/.

[6] Royal Academy of Galician Language.

[7] Today, from a strictly linguistic point of view, on both sides of the Miño River the same language is spoken, since Miñoto and Transmontano dialects are a continuation of the Galician dialects with which they share common traits that make them different from the dialects of Midland and Southern Portugal. However, in terms of common language, and from a sociolinguistic perspective, currently in the west of the Iberian Peninsula there are two modern languages, with phonetic, morphosyntactic and lexical differences, which do not prevent mutual understanding because of the inherent bilingualism that exists between Galician and Portuguese, similar to the

> los romanistas e hispanistas están en general de acuerdo en que el gallego es una forma particular del conjunto dialectal gallego-portugués, en cuanto opuesto al conjunto dialectal español (no "castellano", sino: astur-leonés, castellano, en sus muchas formas, y navarro-aragonés) y al conjunto catalán (o catalán-valenciano)[8] (Cuseriu, 1987: 795)

Thus, drawing from the assumption that Galician and Portuguese are very closely related linguistic varieties and trying to take advantage of the privileged position of Portuguese as a computationally-developed language –that is, a language for which many NLP tools an resources have been developed–, we have investigated the possibility of using free English-Portuguese parallel corpora to create an English-Galician parallel corpus that we would use to develop an English-Galician SMT prototype.

3. CORPUS COMPILATION AND PROCESSING

3.1. The source corpus

Since our project was clearly guided by an open-sourced spirit, we wanted as many components of it as possible to be open source, or at least freely available for non-commercial use.

Because of its large size and liberal copyright license[9] we chose the English-Portuguese Europarl corpus as the source corpus for our project.

The Europarl corpus[10] is a parallel corpus extracted from the European Parliament Proceedings, dating back to 1996, that includes versions of its contents in eleven European languages: Romance (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish.

After an initial processing devoted to cleaning the XML tags that mark the discursive structure of the utterances contained in the corpus, we ended up with an English-Portuguese parallel corpus that contains 58 million words in total.

---

existing bilingualism between Catalan and Occitan, Danish and Norwegian, Slovak and Czech, Faroese and Icelandic.

[8] In general, romanicists and hispanists agree that Galician is a particular form of the Galician-Portuguese dialectological body, as opposed to the Spanish dialectological body (not "Castillian", but: Asturian-Leonese, Castillian, in its many forms, and Navarrese-Aragonese) and the Catalan body (or Catalan-Valencian).

[9] "The European Parliament web site states: "Except where otherwise indicated, reproduction is authorised, provided that the source is acknowledged."" (Koehn, 2005: 2).

[10] Freely available at http://www.statmt.org/europarl/.

3.2. From an english-portuguese to an english-galician parallel corpus

The conversion of the source parallel corpus into an English-Galician parallel corpus we designed at **imaxin**|software was a semi-automatized process that involved the use of two main pieces of software: a RBMT system and a spelling converter –that is, a transliteration engine[11].

Thus, the simplified workflow we designed was the following:

1. Machine translation of the Portuguese side of the source parallel corpus into Galician using EixOpentrad[12].
2. Identification of the unknown, therefore, untranslated words outputted by EixOpentrad.
3. Transliteration of these untranslated words into Galician spelling using a simple Portuguese-Galician transliteration Perl script called port2gal[13].
4. Manual correction of the transliteration errors.

This process took over three months to complete, which is insignificant in contrast to the time-consuming and cost-expensive effort that would have been associated with the manual compilation of a English-Galician corpus of this size.

3.3. The target corpus

After processing of the English-Portuguese Europarl parallel corpus we obtained an English-Galician tokenized corpus composed of 34,715,016 tokens in English and 34,688,010 tokens in Galician.

4. ENGLISH-GALICIAN SMT PROTOTYPE

In order to exemplify the utility of the English-Galician parallel corpus we built and also to demonstrate the validity of our strategy, we will show the quality of the translations we have

[11] Spelling converters are usually used to write the same language code in two different ways. Such converters do nothing more than replace string patterns of the source language into the corresponding string patterns of the target language. This strategy does not involve morphological, syntactic, or semantic information.

[12] EixOpenTrad is a further version of OpenTrad, a platform of open source Machine Translation services (www.opentrad.com). EixOpenTrad is a Galician-Portuguese and Portuguese-Galician MT prototype containing 8,500 words for both directions. This system is based on the Spanish-Portugues Apertium translation engine.

[13] Port2gal's first version was developed by Alberto García (Igalia Free software Company). It was later improved by Pablo Gamallo (Department of Spanish Language at Universidade de Santiago de Compostela). It simply converts European Portuguese spelling into current Galician spelling and vice versa. It is freely available at http://gramatica.usc.es/~gamallo/port2gal.htm.

achieved with the SMT prototype we trained[14] using that English-Galician corpus, in comparison to the quality of Google's own SMT service, which in 2008 incorporated Galician in its catalogue of linguistic tools.

The following example shows a sample automatic translation of the wikipedia *Art* entry[15] performed by our system:

> Arte é o proceso ou produto de arranxar deliberadamente elementos dunha forma que apela á sentidos ou emocións. Engloba un diversificado abano de actividades humanas, creacións e modos de expresión, inclusive da música, da literatura, filmes, escultura e pinturas. O significado de arte é explotada en un ramo da filosofía coñecida como aesthetics.

The next example shows the translation of the same wikipedia entry performed by Google Translate on March 2nd 2010:

> A arte é o proceso ou produto de deliberadamente organizar elementos de un modo que pide aos sentidos ou emocións. Engloba unha variada gama de actividades humanas, creacións, e modos de expresión, incluíndo a música, literatura, cine, escultura e pintura. O significado da arte é explotado desde unha rama da filosofía coñecido como estética. [16]

5. CONCLUSION

As shown in the previous section, we can confidently conclude that the strategy of creating NLP tools for Galician drawing from Portuguese resources is not only linguistically justifiable but, given the high quality of the results that can be achieved, is absolute legitimate.

---

[14] The English-Galician SMT prototype was built within the paradigm of what is known in the field of NLP as Phrase-based SMT. The two main pieces of software we used for this purpose were Moses and GIZA++, which can be respectively downloaded from http://www.statmt.org/moses/ and http://fjoch.com/GIZA++.html.

[15] The sample English sentence, located at http://en.wikipedia.org/wiki/Art, we used to test these two systems is: "Art is the process or product of deliberately arranging elements in a way that appeals to the senses or emotions. It encompasses a diverse range of human activities, creations, and modes of expression, including music, literature, film, sculpture, and paintings. The meaning of art is explored in a branch of philosophy known as aesthetics."

[16] Google Translate was trained using English-Portuguese parallel corpora partially converted into Galician spelling. Until recently, unlike **imaxin**|software's strategy, Google did not seem to use spelling converters. Thus, Portuguese words which were not in their dictionaries remained in their original spelling, as shown by a translation we performed with this service in April 2009: "A arte é o proceso ou produto de deliberadamente organizar elementos dun modo que apelido aos sentidos ou **emoções**. Engloba un conxunto diversificado de actividades humanas, **criações**, e modos de expresión, incluíndo a música ea literatura. O significado da arte é explorador no ramo da filosofía coñecido como estética."

It is, therefore, not adventurous to conclude that the use of resources from a closely related language, especially if this is a computationally-developed language, is extremely important for linguistic varieties, such as Galician, that lack NLP tools due to their minoritized position.

REFERENCES

Aracil, Ll. et al. (1985). *Lingüística e sócio-lingüística galaico-portuguesa: reintegracionismo e conflito lingüístico na Galiza.* Ourense: Asociaçom Socio-Pedagógica Galaico-Portuguesa.

Cunha, C. & Cintra, L. (2002). *Nova Gramática do Português Comtemporâneo*. Lisboa: Edições João Sá da Costa.

Coseriu, E. (1987). El gallego en la historia y en la actualidad. *Actas do II Congresso Internacional da Língua Galego-Portuguesa* (pp. 793-800). A Coruña: AGAL.

Fernández Rei, F. (1991). *Dialectoloxía da lingua galega*. (2nd ed.). Vigo: Edicións Xerais de Galicia.

Gamallo P. & Pichel, J.R. (2007). Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego. *Procesamiento del Lenguaje Natural*, 39: 241-248.

Gamallo P. & Pichel, J.R. (2008). Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary. *Lecture Notes in Computer Science* (Vol. 4919): 423-433.

Gee, J. P. (1999). *An Introduction to Discourse Analysis: Theory and Method.* London: Routledge.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. Paper presented at the MT *Summit* 2005. Pukhet, Thailand, September 12-16.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, M., Zens, R., Dyer, C., Bojar, O., Constantin, A. & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. Demonstration session at the *Annual Meeting of the Association for Computational Linguistics (ACL)*. Prague, Czech Republic, June, 23-30.

Malvar Fernández. P. (2008). *Improving Word-to-Word Alignment using Morphological Information* (Master Thesis). San Diego State University: San Diego, CA.

Och, F.J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1): 19-51.

Pichel, J.R. (2007). Falta de corpus. Galicia Hoxe. Available at (http://www.galicia-hoxe.com/index_2.php?idMenu=153&idNoticia=236722)

Pichel, J.R. (2009). "Estrategia google". Galicia Hoxe. Available at (http://www.galicia-hoxe.com/index_2.php?idMenu=149&idEdicion=1211&idNoticia=414218)

Popović, M. & Hey, H. Statistical Machine Translation with a Small Amount of Bilingual Training Data. Paper at *Language Resources and Evaluation (LREC): 5th SALTMIL Workshop on Minority Languages. Strategies for developing machine translation for minority language*. Genova, Italy, May 23rd, 25-29.

La importancia de la confección y el uso de un corpus para la investigación llevada a cabo en la tesis "sintaxis y semántica de la pasiva preposicional"

ANA ISABEL MARTÍN DOÑA
Universidad de Málaga

Resumen
El objeto de estudio basado en la confección de un corpus y su posterior análisis en la Tesis "Sintaxis y Semántica de la Pasiva Preposicional" fue la configuración de pasiva preposicional de la lengua inglesa. La configuración de pasiva preposicional (véanse He was looked at) sobresale desde un punto de vista interlingüístico por ser una estructura o construcción ausente en un gran número de lenguas. La aportación fundamental que se realizó en esta tesis doctoral consistió en un análisis descriptivo de las restricciones a las que se encuentra sometida la configuración de pasiva preposicional, las cuales son de índole léxico-semántica.
El hecho de que las fuentes utilizadas en la elaboración del corpus se identifiquen como distintas obras lexicográficas o diccionarios significa que dicha tarea – a saber, la elaboración o construcción del propio corpus – constituye un ejercicio de contraste lexicográfico. Esta dimensión del trabajo permitió corroborar la veracidad de la discrepancia idiolectal que caracteriza el dominio de la pasiva preposicional.
Palabras clave: pasiva preposicional, lexicografía

Abstract
The aim of study based on the creation of a corpus and analysis on the Doctoral dissertation "Sintaxis y Semántica de la Pasiva Preposicional" was the configuration of the prepositional passive in the English language. The configuration of the prepositional passive (see He was looked at) stands out from an interlinguistic point of view, since it is a structure or construction lacking in a great number of languages. The main contribution in this Doctoral dissertation was a descriptive analysis of the restrictions suffered by the configuration of the prepositional passive, which have a lexical-semantic character.
The fact that the sources used in the creation of the corpus are identified as different lexicographical works or dictionaries implies that this task – meaning the creation or construction of the corpus – is an exercise of lexicographical contrast. This dimension has enabled the corroboration of the truthfulness in the idiolectal discrepancy which characterises the huge presence of the prepositional passive.
Keywords: prepositional passive, lexicography

1. INTRODUCCIÓN: OBJETO DE ESTUDIO

El objeto de estudio basado en la confección de un corpus y su posterior análisis en la Tesis "Sintaxis y Semántica de la Pasiva Preposicional" fue la configuración de pasiva preposicional de la lengua inglesa, tanto desde una perspectiva sintáctica como, fundamentalmente, desde un punto de vista semántico. La configuración de pasiva preposicional (véanse *He was looked at*, *The new job will be applied for by all employees*,

That stool has been sat on) sobresale desde un punto de vista interlingüístico por ser una estructura o construcción ausente en un gran número de lenguas, tanto de la propia familia germánica como de familias tan extensamente conocidas como la romance. La configuración de pasiva preposicional, que se haya indisolublemente unida a elementos tan destacados de la gramática del inglés como el fenómeno de *preposition stranding* o el fenómeno del *verbo preposicional*,[1] plantea un importante interrogante desde la perspectiva interna de esta lengua, a saber, la falta de correspondencia existente entre estructuras (preposicionales) activas y estructuras (preposicionales) pasivas. Es decir, no todas las estructuras o secuencias cuyo Sintagma Verbal incluye un Sintagma Preposicional poseen contrapartida pasiva. La aportación fundamental que se realizó en esta tesis doctoral consistió en un análisis descriptivo de las restricciones a las que se encuentra sometida la configuración de pasiva preposicional, las cuales son de índole semántica o, más propiamente, léxico-semántica.[2]

Con objeto de acometer el estudio léxico-semántico de las secuencias pasivas preposicionales, se partió de una clasificación o taxonomía de estructuras preposicionales activas según el valor loco-direccional o no loco-direccional de los núcleos o predicados verbales. La elección de este criterio clasificatorio se fundamenta en el valor original loco-direccional de las preposiciones de la lengua inglesa. Los diferentes Tipos de estructuras preposicionales que resultan de la aplicación del mencionado valor locativo o loco-direccional son sometidos a análisis según la incidencia que las nociones de actividad/agentividad y de objeto afectado tienen sobre los mismos.

No obstante, al margen de la necesidad de especificar o explicar la forma o modo en que las mencionadas nociones de *actividad* y *objeto afectado* rigen la configuración de pasiva preposicional, es obligado prestar atención al llamado fenómeno de la *lexicalización*. Según se ha puesto de manifiesto en la bibliografía lingüística sobre el tema, la capacidad o incapacidad de un número indeterminado de estructuras de verbo y preposición, o verbo y Sintagma Preposicional, no parece obedecer a reglas o enunciados semánticos – los cuales girarían, como he mencionado anteriormente, en torno a los conceptos de *actividad/experiencia* y *objeto afectado* – sino antes bien a la pura *idiosincrasia léxica* de

---

[1] Como es ampliamente conocido, el término *preposition stranding* se identifica con el mecanismo mediante el cual el movimiento del objeto de una preposición *no* es acompañado del movimiento o desplazamiento de la propia preposición En lo que se refiere a la identidad de *verbo preposicional*, éste se puede definir como la combinación de verbo y preposición que convierte a ambos elementos en una unidad semántica y asimismo sintáctica.

[2] En las primeras décadas de la teoría generativo-transformacional, si bien desde el modelo no ortodoxo de la *Teoría Léxica Extendida* (*Extended Lexical Theory*), Bresnan (1978) pone de manifiesto que un análisis de la

estos elementos. La meta fundamental de describir las restricciones semánticas a las que se encuentra sometida la configuración de pasiva preposicional es complementada en la investigación por un estudio cuantitativo y cualitativo del fenómeno de la lexicalización, y ambos – es decir, el análisis semántico y el estudio de los procesos de lexicalización – son llevados a término a través de la construcción y del posterior análisis de un *corpus* de pasiva preposicional. Este trabajo comparte, pues, un rasgo muy característico de la investigación que se lleva a cabo en el dominio de la lingüística actual, a saber, el cotejo o comprobación del análisis de las reglas y procesos que subyacen a las lenguas naturales mediante material lingüístico ajeno a la introspección del propio investigador.

Se acometió el estudio de la semántica de la pasiva preposicional a partir de la implementación de enunciados regidos por los conceptos de *actividad/experiencia* y *objeto afectado*, si bien el elemento primordial que rige una investigación como ésta, en torno a secuencias preposicionales, es el valor original *locativo* o *loco-direccional* de las preposiciones (de la lengua inglesa). El modo cómo la referencia o valor *locativo* o *loco-direccional* y las nociones de *actividad* y *objeto afectado* quedan vinculados en este trabajo es mediante el argumento de que los *espacios* o *lugares* – los cuales son especificados de forma canónica por las preposiciones – tienen la habilidad de ser considerados *objetos afectados*. El valor locativo se identifica asimismo en este trabajo como el principal criterio clasificador del corpus, si bien en el caso del corpus dicho valor o significado locativo se aplica a los núcleos verbales, y no a las preposiciones. Se distinguirán, así, estructuras con valor locativo y estructuras con valor no locativo dependiendo del tipo de predicado que sea el núcleo de las mismas.

La metodología que se aplicó en la doble vertiente sintáctica y semántica de este trabajo es, por tanto, igualmente doble o diversa: por un lado, los principios y reglas del aparato generativo-transformacional, y por otro lado, referentes claves de la gramática tradicional – a saber, la referencia *loco-direccional* y los conceptos de *actividad/experiencia* y *objeto afectado*. Adicionalmente al aparato minimalista y a conceptos básicos semánticos o interpretativos de la gramática tradicional, la construcción del corpus de pasiva preposicional que se llevó a cabo en esta investigación constituye el componente más destacado de la metodología empleada.

---

pasiva preposicional en términos estrictamente estructurales no es suficiente, y que es necesario tener en cuenta las conexiones semánticas existentes entre verbo y Sintagma Preposicional.

2. CARACTERIZACIÓN Y DESCRIPCIÓN DEL CORPUS UTILIZADO

El corpus de pasiva preposicional que se confeccionó se caracteriza fundamentalmente por estar construido a partir de otros tantos corpus o corpora propiamente dichos: no se trata, pues, de material lingüístico extraído de obras ajenas al estudio del lenguaje, o al dominio específico de la pasiva preposicional, ni tampoco de información recogida directamente de los hablantes a modo de trabajo de campo, sino de material procedente de las obras o fuentes lexicográficas que son diccionarios contemporáneos de verbos frasales de las casas editoriales Oxford (1993), Longman (2000), y Cobuild (2002) y asimismo del corpus contenido en una obra teórica-descriptiva sobre pasiva preposicional. Esta última es Couper-Kuhlen (1979), obra que, junto con Vestergaard (1977), es utilizada con profusión en este trabajo por constituir ambas hitos fundamentales en la bibliografía lingüística sobre pasiva preposicional.

El hecho de que las fuentes utilizadas en la elaboración del corpus se identifiquen como distintas obras lexicográficas o diccionarios significa que dicha tarea – a saber, la elaboración o construcción del propio corpus – constituye un ejercicio de contraste lexicográfico. Esta dimensión del trabajo permitió corroborar la veracidad de la discrepancia idiolectal que caracteriza el dominio de la pasiva preposicional. Como es sabido, innumerables o, mejor dicho, un número infinito de construcciones sintácticas de las lenguas naturales se prestan a diversos enjuiciamientos de gramaticalidad por parte de los hablantes o usuarios de las mismas, lo que significa que, entre los polos de la gramaticalidad y la agramaticalidad, existe un espacio más o menos extenso donde tienen cabida las dudas y vacilaciones de los hablantes en relación a las secuencias lingüísticas construidas por ellos mismos según las reglas y principios que conocen de forma implícita. Estas dudas o vacilaciones y, en definitiva, la discrepancia entre unos hablantes y otros en relación a los posibles juicios de gramaticalidad, se intensifican de forma especial en el dominio de las construcciones pasivas, hecho que debe imputarse al carácter relativo o relativizable que pueden poseer las nociones de actividad/experiencia y objeto afectado mencionadas con anterioridad, o dicho de otro modo, a la circunstancia de que el concepto de transitividad no es un concepto que atienda a una dicotomía u oposición binaria, sino que se trata de una noción basada en una escala gradual de valores.

El desacuerdo o discrepancia existente entre los diccionarios utilizados en la construcción del corpus de pasiva preposicional pone de manifiesto de manera contundente la naturaleza relativa o inestable del fenómeno pasivo. No cabe duda de que la información

contenida en uno o más diccionarios puede ser errónea, si bien no resulta posible confirmar este extremo. Teniendo presente, no obstante, que el error o la falta de acierto por parte de las obras lexicográficas en cuestión en la caracterización de una estructura pasiva es teóricamente posible, querría insistir en que el desacuerdo o discrepancia que se ha constatado entre las mismas pone de relieve de manera fundamental el carácter relativizable o inestable de la configuración de pasiva preposicional, más aún cuando se trata de obras escritas, y no sólo eso, sino de obras de naturaleza metalingüística. Quiero decir con esto que el desacuerdo sería quizás menos destacable si se manifestara en un tipo de material procedente de entrevistas o cuestionarios realizados a hablantes, ya que éste podría imputarse a circunstancias presentes en los actos de habla pero ajenas a la naturaleza abstracta reglada o estructural de las lenguas naturales. Por el contrario, el empleo de diccionarios como fuentes garantiza en principio el rigor y la precisión del análisis y la información contenidos en los mismos.

3. CODIFICACIÓN EN EL CORPUS DE LA INFORMACIÓN QUE PROPORCIONAN LAS FUENTES LEXICOGRÁFICAS SOBRE LA (A)GRAMATICALIDAD DE LA CONFIGURACIÓN PASIVA DE LAS ESTRUCTURAS PREPOSICIONALES

Couper-Kuhlen (1979) utiliza los códigos *p*, *np*, y *?*, que coloca inmediatamente después de cada entrada, para indicar la *gramaticalidad*, *agramaticalidad*, o *dudosa gramaticalidad* de la construcción de pasiva preposicional. La propia entrada no consiste en la explicitación del significado del verbo o estructura preposicional, sino en la ilustración de uno o más casos en pasiva: concretamente, las combinaciones marcadas previamente *p* no contienen ninguna notación adicional en la ilustración correspondiente, mientras que las marcadas *np* llevan un asterisco (*) en el ejemplo o ilustración, y las marcadas *?* llevan el mismo símbolo (*?*) en el ejemplo o ilustración. En contraposición a Oxford, y en realidad a los tres diccionarios de verbos frasales, CK ofrece como norma una o varias ilustraciones en pasiva de las estructuras preposicionales correspondientes: de hecho, tal y como se ha señalado anteriormente, es el único tipo de información que ofrece, además de la del código gramatical relativo a la (a)gramaticalidad de la configuración pasiva. Ello se debe a que el objetivo de CK es ofrecer un corpus de pasiva preposicional, mientras que Oxford, Longman, y Cobuild son diccionarios de verbos preposicionales (o frasales).

En cuanto al método utilizado por Oxford, en esta obra se incluye la notación *pass* como parte de la codificación gramatical en caso de que el verbo preposicional en cuestión

admita configuración pasiva; por contraposición, la ausencia de dicha notación implica la agramaticalidad o ilegitimidad de la construcción pasiva. Adicionalmente, algún ejemplo o ilustración de las que el diccionario incluye tras la definición del significado y el resto de la información gramatical puede aparecer aleatoriamente en pasiva.

En lo que respecta a Cobuild, se hace uso de la notación *has passive* con objeto de indicar la legitimidad de la configuración pasiva de la correspondiente estructura preposicional; al igual que es el caso en Oxford, la ausencia de la mencionada notación significa que la construcción de pasiva preposicional es caracterizada por los editores como ilegítima o agramatical. Adicionalmente, Cobuild señala con el código *usually passive* aquellas estructuras de verbo y preposición que se emplean comúnmente en voz pasiva.

Finalmente Longman ha resultado de menor utilidad en la construcción o elaboración del presente corpus de pasiva preposicional, debido a que esta obra distingue varios tipos de estructuras en relación al fenómeno pasivo, pero en ningún caso especifica las estructuras preposicionales que admiten la configuración pasiva. Concretamente, Longman distingue tres tipos de construcciones: (a) las que nunca aparecen en voz pasiva, las cuales marca *not passive*; (b) las que aparecen frecuentemente en pasiva, las cuales se identifican como *usually passive*; finalmente, (c) las que normalmente únicamente aparecen en pasiva, las cuales se registran ya en la cabeza de la entrada, o en una división de la entrada, en forma pasiva (por ejemplo, *something is spoken for*).

Según lo señalado anteriormente, el punto crucial de contraste entre Longman por un lado, y Oxford y Cobuild (y asimismo CK) por otro, reside en que Longman no indica qué verbos preposicionales poseen una contrapartida pasiva: esto es, la ausencia del código *not passive no* significa en el sistema de codificación empleado por Longman que la estructura en cuestión posea contrapartida pasiva. Por otra parte, hay ocasiones en las que se ofrece alguna ilustración de una estructura de verbo preposicional en forma pasiva, pero esta práctica no es sistemática, con lo que de nuevo no se proporciona información precisa sobre qué verbos preposicionales poseen contrapartida pasiva, y cuáles no. Esta falta de precisión se refleja en el corpus, ya que un elevado número de casillas correspondientes al espacio dedicado a Longman aparece marcado con un símbolo que indica o significa la citada ausencia de exactitud o precisión (concretamente, se trata del símbolo o notación *!*): sólo un porcentaje relativamente bajo de las casillas de Longman aparecen marcadas con los mismos símbolos o notaciones que las casillas de las restantes fuentes, y no con el mencionado símbolo *!*.

Los símbolos utilizados son los que a continuación se especifican:

ˇ : significa legitimidad o gramaticalidad de configuración de pasiva

preposicional

* : significa ilegitimidad o agramaticalidad de configuración de pasiva preposicional

--: significa ausencia o no inclusión de una estructura en la obra lexicográfica de que se trate

! : significa falta de explicitación (en Longman) sobre la legitimidad o gramaticalidad de la configuración de pasiva preposicional

!(٧): significa falta de explicitación (en Longman) sobre la legitimidad o gramaticalidad de la configuración de pasiva preposicional, aunque al menos una de las ilustraciones que se proporcionan incluye una estructura de pasiva preposicional.

A continuación aparecen dos ejemplos extraídos del corpus ya confeccionado de estructuras de Verbo+Preposición en la que se aprecia la discrepancia o la falta de consenso entre las diferentes fuentes lexicográficas a la hora de establecer si esa estructura posee contrapartida pasiva o no:

VERBO + PREPOSICIÓN	COUPER-KUHLEN (1979)	OXFORD (1993)	LONGMAN (2000)	COBUILD (2002)
plump for (sent. met[3].: 'choose with decision and confidence')	٧	٧	*	*

Ejemplo de Couper-Kuhlen (1979):

(1) *The liberal candidate was plumped for by the majority of the delegates.*

Ejemplos de Oxford (1993):

(2) *I wanted the red car, but Mary plumped for the blue one with the grey seats.*

[3] También el significado literal de la estructura preposicional en cuestión se ha diferenciado del significado no literal de las estructuras preposicionales que conforman el corpus. A modo de resumen, se ofrece un listado de los códigos relativos al significado de las estructuras preposicionales que componen el corpus:

a. ausencia de código: el significado de la estructura es literal

b. *sigdo. no lit.:* el significado de la estructura no es literal; tras el código aparece la correspondiente definición

c. *sent. met.:* el significado del objeto preposicional es abstracto; tras el código aparece ocasionalmente la correspondiente definición; únicamente se emplea este código si el significado o referencia abstracta del objeto preposicional afecta la (a) gramaticalidad de la configuración pasiva.

(3) *He offered André Gide English lessons for five francs an hour, but Gide plumped for the Berlitz School instead.*

Ejemplos de Longman (2000):

(4) *In the end we plumbed for a bottle of Chateau Musar.*

(5) *Faced with a choice between Bob Dole and Bill Clinton, most voters plumped for Clinton.*

Ejemplos de Cobuild (2002):

(6) *In the event, they plumbed for a three-stage conference.*

(7) *The oil industry had plumped for unmanned underwater robots instead.*

(8) *Few gentlemen would now care to plump for an army career.*

VERBO + PREPOSICIÓN	COUPER-KUHLEN (1979)	OXFORD (1993)	LONGMAN (2000)	COBUILD (2002)
run into (sigdo. no lit.[4]: 'strike, collide with sth')	√	*	!	√

Ejemplo de Couper-Kuhlen (1979):

(9) *John was run into downtown yesterday.*

Ejemplos de Oxford (1993):

(10) *The prow of the boat ran into a bank of soft mud and stuck fast.*

(11) *Russian counter-attacks during the afternoon had run into the full strength of the 4th Panzer Army.*

Ejemplos de Cobuild (2002):

(12) *The bus ran into a car.*

(13) *He ran into the back of a van at a zebra crossing.*

Las estructuras pasivas preposicionales que fueron objeto de esta investigación son aquellas estructuras que se corresponden con secuencias activas cuyo Sintagma Verbal consta únicamente de (i) un verbo y (ii) un Sintagma Preposicional transitivo[5]. La acotación que se realizó deja fuera del alcance de este trabajo estructuras tan destacadas como las pasivas correspondientes a (i) las llamadas secuencias con *verbos frasales preposicionales*, o *adverbiales preposicionales* (*He must put up with their selfishness*), (ii) las secuencias o estructuras de *movimiento dativo* (*Sheila taught the children Mathematics*), o (iii) las secuencias que constituyen *frases idiomáticas* (*beat about the bush*).

Asimismo se desestimó en la investigación los aspectos discursivos supraoracionales presentes en el uso real y efectivo de cualquier estructura sintáctica. Por otra parte, y en lo relativo al corpus, es absolutamente oportuno indicar que las ilustraciones extraídas de las fuentes, las cuales constituyen oraciones completas propiamente dichas, se localizan en un contexto supraoracional determinado. No cabe duda de que el material recogido o confeccionado por el o los editores/autores de los diccionarios se inserta idealmente en un discurso; a pesar de ello, el estudio se ciñe al análisis de las nociones o conceptos de *actividad/experiencia* y *objeto afectado* dentro de los límites de la oración propiamente dicha.

Según la propuesta defendida en esta investigación, las secuencias que denotan una *actividad*, o asimismo un evento que se podría describir como *experiencia*, poseen *generalmente* contrapartida pasiva, independientemente de que su objeto preposicional se pueda considerar o no entidad afectada. Por el contrario, las secuencias o estructuras que *no* describen actividad o experiencia, carecen *en líneas generales* de la capacidad de pasivizar, al margen de que su objeto preposicional sea o no objeto afectado. Así las estructuras que componen el corpus se clasifican las siguientes Tipos léxico-semánticos:

Tipo A.1: estructuras que no indican actividad puramente física

Tipo A.1.a: estructuras con agente/experimentante

Tipo A.1.b: estructuras sin agente/experimentante (= estructuras predicativas)

Tipo A.2: estructuras que indican actividad fundamentalmente física

Tipo B.1: estructuras que no implican movimiento

Tipo B.2: estructuras que implican movimiento

---

[4] Véase nota 3 a pie de página.

[5] Esto es, un Sintagma Preposicional que consta del núcleo preposicional y de un Sintagma Determinante objeto o complemento.

El total de estructuras contabilizadas asciende a 3.137, las cuales se agrupan en los distintos Tipos de la forma como se indica a continuación. El Tipo más numeroso es el locativo B.2, mientras que el más escaso es el Tipo predicativo A.1.b.

Tipo A.1.a: 577 ítems o estructuras (18,39%)
Tipo A.1.b: 78 ítems o estructuras (2,48%)
Tipo A.2: 644 ítems o estructuras (20,52%)
Tipo B.1: 546 ítems o estructuras (17,40%)
Tipo B.2: 1292 ítems o estructuras (41,18%)

Las preposiciones o núcleos preposicionales que concurren en las estructuras que conforman el corpus son las siguientes:

> *about, above, across, after, against, along, among, around, as, at, before, behind, beside, between, beyond, by, down, for, from, in , into, inside, like,near, of, off, on, onto, out of, over, past, round, through, to, toward(s), under,up, upon, with, within, without*

4. CONCLUSIÓN

La elección de fuentes metalingüísticas en la conformación del corpus se ha debido a la necesidad de que el análisis semántico o léxico-semántico de la pasiva preposicional posea una base sólida, y asimismo al deseo de dar un testimonio lo más certero y objetivo posible sobre la *discrepancia* que caracteriza el dominio de la pasiva preposicional. En relación a este último aspecto, quisiera resaltar que, en las secciones correspondientes del capítulo 5 de la tesis [Martín (2007)], se muestran la totalidad de ilustraciones ofrecidas en cada una de las obras lexicográficas con el propósito de que sea posible cotejar de forma directa la discrepancia o ausencia de acuerdo entre las mismas. Con la salvedad de un pequeño porcentaje de falta de acuerdo o consenso que podría ser razonado o justificado atendiendo a los factores que se mencionan en el citado capítulo 5, el desacuerdo o falta de consenso existente entre las obras consultadas asciende a una proporción elevada de casos, lo que pone de manifiesto, como he mencionado más arriba, el dominio gramatical altamente relativizable o inestable que constituye la configuración de pasiva preposicional. Específicamente, es el

Tipo locativo B.2 el que se caracteriza por el porcentaje más alto de desacuerdo o falta de consenso.[6]

Considero que la elaboración de un corpus de pasiva preposicional de estas características supone una tarea de investigación innovadora en el panorama de la lingüística actual. Concretamente, en lo que al fenómeno de la lexicalización se refiere, desconozco la existencia de otro posible estudio donde se ofrezcan datos precisos sobre la idiosincrasia léxica de un material más o menos extenso como el que se pretendió mostrar en este trabajo. Por otra parte, en lo que respecta al análisis de las propiedades semánticas de las estructuras preposicionales – esto es, en lo que se refiere a la medición de las conceptos de actividad/agentividad y objeto afectado creo acertar igualmente si afirmo que la clasificación y caracterización del material que conforma el corpus es exhaustiva, al menos desde el punto de vista del criterio utilizado, que es el que atiende a la tipología verbal.

La discusión que se planteó en este trabajo se fundamenta sobre la base de múltiples bifurcaciones o contrastes binarios, como por otra parte es casi inevitable que suceda en una investigación lingüística. Así, se habló de *preposition stranding vs. pied piping*[7], de objeto afectado *vs.* objeto no afectado, de complemento *vs.* adjunto, de pasiva preposicional *vs.* pasiva común, etc. Efectivamente varias de estas oposiciones o caracterizaciones binarias forman parte del entramado que sostiene el dominio de la pasiva preposicional de una manera fundamental.

El doble propósito que se persiguió en esta investigación se identifica con el título mismo del trabajo: *sintaxis* y *semántica* de la pasiva preposicional. Si bien el análisis semántico (o mejor dicho, léxico-semántico) de las restricciones de la configuración de pasiva preposicional ha significado la meta prioritaria de la investigación, un destacado segundo lugar corresponde al estudio del ordenamiento estructural de proyecciones funcionales responsables de una secuencia pasiva preposicional.

El esfuerzo y el trabajo contenidos en esta investigación basada en la confección de un corpus y su posterior análisis y contraste tienen como intención contribuir a avanzar eficazmente en el estudio de la pasiva preposicional.

Para terminar, las que menciono a continuación serían algunas de las posibles vías futuras de estudio que me ha sido posible visualizar a través de este trabajo:

---

[6] En efecto, a B.2, que es adicionalmente el Tipo léxico-semántico que agrupa un mayor número de estructuras, le corresponde la Taxonomía (3) con el más alto porcentaje de desacuerdo, específicamente 8,41% [véase 5.4, capítulo 5 en Martín (2007)].

[7] Etiqueta anglosajona que se traduce al español como *atracción* o *arrastre* de la preposición.

- clasificación del corpus en torno a una taxonomía de predicados verbales más atomista o pormenorizada
- caracterización semántica individualizada del conjunto total de Sintagmas Preposicionales que concurren en las estructuras que conforman el corpus
- estudio diacrónico del desarrollo experimentado por las preposiciones contenidas en el corpus según el predicado verbal con que se asocien
- estudio del contraste existente entre las fuentes lexicográficas utilizadas en la construcción del corpus en relación al alcance de la categoría *verbo preposicional*
- ...

REFERENCIAS BIBLIOGRÁFICAS

Bresnan, J. (1978). A realistic transformational grammar. En M. Halle et al. (eds.) *Linguistic Theory and Psychological Reality*. Cambridge, Mass.: MIT Press.

Collins Cobuild (2002). *Dictionary of Phrasal Verbs*. Glasgow, Great Britain: Harper Collins Publishers.

Couper-Kuhlen (1979). *The Prepositional Passive in English. A Semantic-Syntactic Analysis, with a Lexicon of Prepositional Verbs*. Tübingen: Max Niemeyer Verlag.

Cowie, A.P. and R. Mackin (1993). *Oxford Dictionary of Phrasal Verbs*. Oxford: Oxford University Press.

Dixon, R.M.W. (1991). *A New Approach to English Grammar, on Semantic Principles*. Oxford: Clarendon Press.

Hornby, A.S. et al. (1948/1963). *Advanced Learner's Dictionary of Current English*. London: Oxford University Press.

Levin, B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago and London: The University of Chicago Press.

Levin, B. and M. Rappaport Hovav (1995). *Unaccusativity at the Syntax–Lexical Semantics Interface*. Cambridge, Mass.: MIT Press.

Longman (2000). *Phrasal Verbs Dictionary*. Essex, England: Longman.

Martín, A. (2007). *Sintaxis y Semántica de la Pasiva Preposicional.* Tesis Doctoral: Universidad de Málaga (UMA).

Vestergaard, T. (1977). *Prepositional Phrases and Prepositional Verbs*. The Hague: Mouton Publishers.

Reaction Object constructions in English. A corpus-based study

MONTSERRAT MARTÍNEZ VÁZQUEZ

Universidad Pablo de Olavide

Abstract
Little attention has been paid to Reaction Objects (RO) constructions (Levin, 1993), which involve the use of an intransitive manner of speaking verb or a verb of signs with a non-subcategorized object (Pauline smiled her thanks). Besides, the researchers who touch on this construction do not usually include empirical evidence. The examples supplied are limited, which leads to a general tendency to overcite the examples supplied by Levin (1993). In this talk I will attempt a more fine-grained analysis of ROs in English based on the analysis of examples extracted from the Corpus of Contemporary American English (COCA). The corpus findings suggest that the construction is richer and more varied than it has been assumed.
keywords: reaction objects, Levin, COCA

Resumen
Las construcciones de objeto de reacción (Levin 1993), formadas por un verbo intransitivo de manera de hablar o un verbo de signos con un objeto no subcategorizado (Pauline smiled her thanks) han recibido poca atención. Además, los investigadores que las tratan no suelen aportar datos empíricos. Los ejemplos que aparecen en estos estudios son limitados, con una tendencia general a repetir los ejemplos proporcionados por Levin (1993). En esta comunicación intentaré llevar a cabo un estudio más pormenorizado de las construcciones de objeto de reacción en inglés basándome en el análisis de ejemplos extraídos del Corpus of Contemporary American English (COCA). Los resultados sugieren que la construcción es más rica y variada de lo que hasta ahora se ha asumido.
Palabras Clave: Objetos de reacción, Levin, COCA

1. INTRODUCTION[1]

Levin (1993) analyses under the rubric of "reaction object construction" an alternation involving the transitive use of an intransitive manner of speaking verb or a verb of signs with an extended sense of "express a reaction by V-ing":

Certain intransitive verbs—particularly verbs of manner of speaking and verbs of gestures and signs—take nonsubcategorized objects that express a reaction (an emotion or disposition); possible objects include: *approval, disapproval, assent, admiration, disgust, yes, no.* When these verbs take such objects they take on an extended sense which might be paraphrased "express (a reaction) by V-ing," where "V" is the basic sense of the verb. For instance, *She mumbled her adoration* can be paraphrased as "She expressed/signalled her

[1] This study is part of a project funded by the Spanish Ministry of Science and Innovation (FI2008-04234/FILO).

adoration by mumbling." Most of the verbs that allow such reaction objects name activities that are associated with particular emotions, and the action they name is performed to express the associated emotion. (1993: 98)

The meaning of this construction is not predictable form the verb's semantics. The verb undergoes a conceptual subordination process, which turns it into secondary information, while the object conveys the main action. Little attention has been paid to this construction. However, empirical data suggest that the construction is much more productive and extensive than is usually believed.

In this talk I will present an analysis of Reaction Object (RO) constructions in English based on the examination of more than 500 examples extracted from the *Corpus of Contemporary American English* (COCA).

2. Previous approaches to the RO construction

RO constructions have not received much attention. English Grammars like Quirk et al. (1985) do not mention them, even though they discuss resultant, cognate and eventive objects (1985: 749-52). Huddleston & Pullum (2002: 305) present the examples reproduced in (1) under the rubric of "object of conveyed reaction".

(1)

a. He grinned his appreciation.

b. I nodded my agreement.

c. He roared his thanks.

They make a distinction between two types: constructions with verbs of "non-verbal communication" (*grin, laugh, nod, sigh, smile,* and *wave*), and manner of speaking verbs (*mumble, roar, scream,* and *whisper*). In the first group, the object does not express an argument of the verb, as in (2a), so the passive construction is ungrammatical, (2b). Constructions with manner of speaking verbs, on the other hand, take a wider range of objects, (2c), and may appear in passive constructions, (2d).

(2)

a. She smiled her assent.

b. *Her assent was smiled.

c. He roared the command.

d. On the parade ground commands must be roared, not whispered.

Martínez Vázquez (1998) analyses ROs as cases of non-subcategorized effected objects. It is argued that their production is pragmatically restricted; in order to be integrated in the construction the non-subcategorized object has to be compatible with the verb's semantics/pragmatics (Goldberg, 1995). Thus, the act of nodding, a cultural sign of affirmation, is easily taken for an act of transmission of this affirmation:

(3)

Baron Zginski nodded his assent and departed. (COCA)

Felser & Wanner (2001), analyse ROs in relation to Cognate Objects. The latter require modification and do not appear in passive constructions, (4a). ROs may not take modification and may passivize (4b). Both Cognate Objects and ROs are complements, since they do not allow a manner adverb separating them from the verb (4d).

(4)

a. *A smile was smiled (Felser & Wanner, 2001: #5a)

b. Warm thanks were smiled at the audience (Felser & Wanner, 2001: #6b).

c. *He sighed wearily a (heavy) sigh. (Felser & Wanner, 2001: #8a)

d. *She nodded gracefully her approval. (Felser & Wanner, 2001: #8b)

From a semantic point of view, Felser & Wanner view ROs as "abstract nouns, expressing an attitude which can be made visible by the action denoted by the verb" (2001: 11). Their meaning needs to be in accordance with the verb's meaning; more precisely, since they argue that the intransitive verbs in these constructions take an empty category (a null pronominal) as their internal argument, the RO has to be "compatible with the production of the indefinite object of the verb" (2001: 11). They claim that even though the act of approving does not necessarily imply a smile, it could easily do, so *smile one's approval* is

acceptable. By contrast, an act of disapproval is difficult to associate to a smile, hence, *smile one's disapproval* would be highly improbable.

Mirto (2007) underlines the predicative nature of the object in RO constructions. Thus, the canonically cited example, *She nodded her approval,* paraphrased as 'she approved (of something) by nodding', entails both 'she approved' and 'she nodded'. The object as a predicative, shares the clausal subject with the verb.

The studies on ROs reviewed above have all evolved out of Levin (1993) and they all share a vision of ROs as nonsubcategorized objects expressing an emotion or disposition closely associated to the action denoted the verb. The construction conveys a communicative act (oral or visual). However, the above-mentioned studies do not supply enough empirical evidence to present a comprehensive description of the construction. I will attempt at completing the analysis of ROs in English after careful analysis of a sample of more than 500 examples of ROs extracted from the COCA corpus.

3. A CORPUS-BASED STUDY OF THE RO CONSTRUCTION

The studies on RO constructions mentioned above seem to agree in the distinction of two groups of verbs taking part in the construction: manner of speaking verbs, and verbs of signs and gestures; i.e. expression through words and expression through nonverbal signs. However, the interpretation as one or the other is highly dependent on the constructional meaning. For example, a verb like *bark* expresses non-verbal communication in a RO construction like (5a). But the same sound may be used metonymically to denote an aggressive way of speaking when a human being emits the sound, as in (5b). In this case, the construction encapsulates a verbal communicative act.

(5)

a. They ran flat out, while the dog barked his warning. (COCA)

b. He barked his sentences military style. (COCA)

A classification of verbs into verbal or non-verbal communication may not fit the semantics of the construction, since the constructional meaning generally involves a figurative mapping. Only after a constructional reading will we get the precise meaning of the expressive act implied in the RO construction.

Besides, the class of verbs appearing in the construction is not closed. Many RO constructions reflect metaphorical mappings, which increases substantially the potential list of verbs that may take part in the construction. For example, we have noted that sound emission verbs related to animals are frequently taken to the human domain, as in (5b). This metonymic mapping is so conventional that it is even listed in dictionaries. But other more creative mappings are less predictable. Thus, we have extracted from the COCA corpus examples of other sound emission verbs inserted in RO constructions not enumerated in Levin (1993). For example, in (6a) a sound emission verb, *jingle*, emitted by an inanimate being, *the door*, is used metaphorically in a RO construction. This association is cognitively much more elaborated than the metonymic uses listed under (5). It would, therefore, be difficult to predict a class of inanimate sound emission verbs that would fit in the RO construction. Likewise, other less conventional sound emission verbs have been attested in the RO construction, as in (6b-c).

(6)

a. The butcher store was crowded, and as we stepped inside the door *jingled a welcome.* (COCA)

b. I simply do not have time to go back in there. I *honk good-bye* to Charles and Faye. (COCA)

c. I have great respect for the work you are doing. Seed *tut-tuts his thanks*. (COCA)

Other verbs of gestures have been attested in the RO construction yet they are not itemized in Levin (1993):

(7)

a. I *gestured hello*, a kind of jutting out of one elbow while shaking my head side to side. (COCA)

b. 'Howdy, Miss Emily,' I said as I creaked open the screen door, my hand springing up, my fingers *fluttering hello*. (COCA)

c. We *bowed farewell* to the last man. (COCA)

The object in the RO construction has been described as a noun expressing a feeling or disposition compatible with the act denoted by the verb. This semantically compatible relation, however, varies significantly in the examples mentioned in the literature. Some

objects are almost redundant with the verb's meaning (*I nodded my assent*), while some are not so close the verb's semantics (*He roared his thanks*). For example, a specific gesture like, *frown*, only yielded 5 examples of RO constructions in the COCA, all of which, in accordance with the verb's meaning, share a negative component (*their displeasure, his disapproval, her confusion, her annoyance, my disapproval*). On the other hand, the verb *smile* shows a richer productivity, and does not necessarily imply a positive outcome (*her surprise, her courage to always surrender, encouragement, reassurance, etc.*). The act of whistling is even less circumscribed to a particular type of message and may even convey opposite communicative functions, as in (8a-b).

(8)

a. Artoo whistled his assurances. (COCA)

b. Gerry whistles his disapproval. (COCA)

Even a verb like *nod*, which is conventionally restricted to affirmative scenarios, shows certain variation in our corpus; for example, the objects in sentences (9a-b) do not refer to affirmative acts directly, their assertive meaning derives from the pervious linguistic context. The objects in (9c-d) also differ from the canonical *nod-agreement* construction.

(9)

a. Are you all *familiar* with gangster rap? "McPherson asked. We were, despite the fact that, besides me, all of the students were white and mostly middle- to upper-class. While we each nodded *our familiarity with the genre*, McPherson reached into a shopping bag he'd brought and removed a magazine. (COCA)

b. What can we do, Paul? … The two girls nodded *their wishes to help* too. (COCA)

c. The other man nodded his goodbye. (COCA)

d. Captain Campbell nodded his forgiveness and closed his eyes. (COCA)

If we take a closer look at the examples of ROs provided by Levin (1993: 95) – *approval, disapproval, assent, admiration, disgust, yes, no, thanks, welcome*– we find different types of messages.

Expressions like *yes, no, thanks* and *welcome*, are classified as *formulae* by Quirk et al. (1985: 852) and they are regarded as "nonsentences". As such, they may appear in isolation: "Most formulae used for stereotyped communication situations are grammatically irregular.

Only in a very limited way can they be analysed into clause elements." (1985: 852). However, when they enter into syntactic and semantic relations with other clause elements, as in the RO construction, they lose part of their original formulaic sense, exhibiting a variation from direct statements to signs, depending on the other elements of the construction. Their integration in the clause is also formally marked: they may appear as direct quotes, which may be nominalized by inserting a determiner, (10a-b), or by direct insertion after a manner of speaking verb, as in (10c).

(10)

a. Be prepared for random recordings of children hollering *a* welcome. (COCA)

b. Aringarosa grumbled *his* hello. (COCA)

c. Hannah murmured good-bye. (COCA)

When the nominalization is not formally marked, as in (10c), the construction stands closer to ordinary communicative constructions. The presence of a possessive determiner in (10b) brings it closer to a RO construction. When these *formulae* appear after a verb of non-verbal expression they are no longer messages but signs carrying a message:

(11)

a. She watched her knight bow good-bye to the Witch. (COCA)

b. Mariah waved hello and went into the kitchen. (COCA)

c. He shrugs hello to the band. (COCA)

d. Most of the artists nod hello to Fred. (COCA)

Nouns like *approval, disapproval*, or *assent* constitute another different group of ROs. These are also reaction signals, but in contrast with the previous group, they cannot stand in isolation as communicative formulae; they need to be integrated as clause elements.

(12)

a. Barnett smiles his approval.

b. Moshe snorted his disapproval.

c. All the women nod their assent.

Finally, ROs like *admiration* or *disgust* have an expressive rather than a reactive meaning:

(13)

a. The women of the family murmured admiration. (COCA)

b. The two Jacksons whisper their disgust at 'all the injustice' they see on TV. (COCA)

Although ROs lie on a continuum and the boundaries between different classes are fuzzy we may distinguish three varieties: formulae (11), reaction nouns (12) and expressive nouns (13).

Besides the verb and its non-subcategorized object, RO constructions may also include other optional arguments. Since they evoke a communicative scenario, they may integrate a receiver –either a listener or a viewer– introduced by the prepositions *to* (14a-c) or *at* (14-e).

(14)

a. Roscoe waved farewell *to* the boys. (COCA)

b. He shrugs hello *to* the band. (COCA)

c. Abisel poised nearby, humming a welcome *at* Hollis. (COCA)

d. But for the most part people looked disapproving, frowning their displeasure *at* him and making comments to their neighbours. (COCA)

In most RO constructions the receiver is not explicitly mentioned, although it is implied in the context. However, some RO constructions convey situations where the speaker is alone releasing a strong or repressed emotion. In these constructions a goal may be included, emulating thus a transfer event, as in (15), but the transfer is only the cathartic liberation of a feeling.

(15)

Greta Marie threw her head back and howled her misery *to the skies*. (COCA)

Another argument optionally inserted in a RO construction is the stimulus that causes the reaction, introduced by the preposition *at*. This element marks a difference with the communicative construction, which does not add this argument.

(16)

Max chuckled his delight *at her ingenuity*.

ROs may appear in ditransitive patterns. The first object is both the affected participant of the main verb and the recipient of the communicative act implied in the construction after the insertion of the RO:

(17)

a. Will you let her kiss you hello when you see her? (COCA)

b. Paul kissed her goodbye (COCA)

c. Their first date she had kissed him goodnight (COCA)

Other ditransitive RO constructions have been attested with the verbs *hug* and *wave,* (18a-b). Notice, however, that the recipient object in (18b) is not subcategorized by the verb.

(18)

a. Irma handled the waiting room, hugging patients hello, pushing homemade sweets on them, balancing the books. (COCA)

b. Their two children were still on the verandah, waving her good-bye.

Since greetings are usually reciprocal actions, they may also emerge in an intransitive reciprocal pattern:

(19)

a. As we hugged hello, I casually put his hand on my butt.

b. […] holding their son's body and sobbing their final good-byes.

4. CONCLUDING REMARKS

Our corpus analysis of RO has proved that the construction is richer than is usually assumed. Constructions fall on a continuum extending from those with manner of speaking verbs,

which could also qualify as regular communicative constructions, to those with verbs of non-verbal communication. The objects also oscillate from those expressing a verbal exchange, with formulae like *yes/no/hello* and other reactive nouns like *agreement* or *assent*, to nouns expressing feelings. The object does not necessarily imply a "reaction"; it may denote an expressive act not caused by an external stimulus. Sometimes the RO construction is used to describe a cathartic situation: the release of a strong emotion. This "cathartic" function relates to the meaning underlying all RO constructions: the emotive release of an expressive act. Since the expression is not transmitted through the usual means of communication, ie. a verb of saying, the means is naturally highlighted.

REFERENCES

COCA: Davis, M. (2008). *Corpus of Contemporary American English*. Brigham Young University.

CREA: Real Academia Española (2009). Banco de datos (CREA) [online].

Felser, C. & Wanner, A. (2001). The Syntax of Cognate and Other Unselected Objects. In N. Dehé & A. Wanner (Eds.), *Structural Aspects of Semantically Complex Verbs.* (pp. 105-130). Frankfurt, Bern & New York: Peter Lang. Available at http://privatewww.essex.ac.uk/~felsec/research/

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University Press.

Huddleston, R. & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge: University Press.

Levin, B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago: University Press.

Martínez Vázquez, M. (1998). Effected Objects in English and Spanish. *Languages in Contrast,* 1, 245-264.

Martínez Vázquez, M. (2005). Communicative Constructions in English and Spanish. In C. Butler, Mª A. Gómez-González & S.M. Doval-Suárez (Eds.), *The Dynamics of Language Use: Functional and Contrastive Perspectives* (pp. 79-109). Amsterdam: John Benjamins.

Mirto, I. M. (2007). Dream a little dream of me: Cognate Predicates in English. Paper presented at the *26th conference on Lexis and Grammar*. Bonifacio, October 2-6.

Quirk, R., Greenbaum S., Leech, G., & Jan Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Corpus of Interpreting Discourse = Speech Corpus + Parallel Corpus?

MIKHAIL MIKHAILOV

University of Tampere

Abstract

The need for corpora of interpreting discourse in translation studies is gradually increasing. The main reason of the lack in such resources is the difficulty of obtaining data and the inevitability of manual data input. An interpreting corpus would be a collection of transcripts of speech in two or more languages with part of the transcripts aligned. The structure of the corpus should reflect the polyphonic nature of the data. Thus, markup becomes extremely important in this type of corpora. The research presented in this paper deals with corpora of Finnish-Russian interpreting discourse. The software package developed for processing of the corpora includes routines specially written for studying speech transcripts rather than written text. For example, speaker statistics function calculates number of words, number of pauses, their duration, average speech tempo of a certain speaker.

Keywords: interpreting corpus, corpus software, corpus markup, Russian language, Finnish language

Resumen

La necesidad de utilizar corpora del discurso interpretativo está aumentando gradualmente en los estudios de traducción. La razón principal de la escasez de este tipo de recursos es la dificultad de obtener los datos y la inevitabilidad de introducir los datos manualmente. Un corpus de interpretación es una colección de transcripciones del habla en dos o más lenguas alineadas. La estructura del corpus debe reflejar la naturaleza polifónica de los datos. Por eso, el marcado es extremadamente importante en este tipo de corpus. La investigación presentada en este artículo trata de un corpus del discurso interpretativo finlandés-ruso. El paquete de software desarrollado para el procesamiento del corpus incluye los procedimientos específicamente desarrollados para los estudios de transcripciones del habla, y no del lenguaje escrito. Por ejemplo, la función que calcula las estadísticas del hablante obtiene el número de palabras, el número de pausas, su duración y el promedio de la velocidad del habla de un hablante.

Palabras clave: Corpus interpretativo, software para corpus, marcación de corpus, lengua rusa, lengua finlandesa

1. INTRODUCTION

Compiling written text corpora has become a relatively easy task in our modern information society. Some published texts are ready available in digital form, other can be digitized with the help of OCR software. Plenty of texts of different genres written in all imaginable languages are being accumulated on the web. Therefore, text corpora exceeding 100 million running words in size are quite common today.

At the same time, compiling spoken corpora remains hard, time-consuming, expensive and extremely slow work. Not many transcripts of spoken language are available (e.g. speeches of politicians, TV interviews, etc.), and most of them are adaptations of oral speech into written form and need manual checking. In contrast to written resources, transcripts of oral speech are not subject to amateurish collecting. Recording and transcribing oral speech

remain scholars' activity. Although the quality of sound recording and possibilities for data storage have remarkably improved during the last decades, speech recognition technologies are still under development. As a result, the transcribing is to be done manually for the time being.

English language resources dominate in spoken corpora as they do in text corpora, which is quite predictable. It will be enough to mention Cambridge International Corpus (CANCODE, http://www.cambridge.org/elt/corpus/cancode.htm), Diachronic Corpus of Present-Day Spoken English (DCPSE, http://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm), and Michigan Corpus of Academic Spoken English (MICASE, http://quod.lib.umich.edu/m/micase/), a considerable list of spoken corpora can be found at http://corpus-linguistics.de/html/corp/corp_spoken.html. Non-English spoken corpora are much less common. The research presented in this paper deals with two languages, Finnish and Russian, which are no exception. There is a spoken subcorpus of about 6 million running words in the Russian National Corpus (www.ruscorpora.ru). As regards transcripts of Finnish speech, they are available from the Finnish Broadcast Corpus (http://www.csc.fi/english/research/software/fbc). Most of other existing collections of transcripts of prepared or spontaneous speech are of modest size and with only basic search interface or no search interface at all.

Today, a lot of research in the field of interpreting is being conducted, however, interpreting corpora are still quite a new kind of language resources and thus far not much empirical and quantitative data can be obtained. Let us mention the European Parliament Interpreting Corpus (EPIC) as one of very few examples. It is composed of speeches at the European Parliament interpreted into English, Italian, and Spanish, and is arranged as a parallel corpus (http://sslmitdev-online.sslmit.unibo.it/corpora/ corporaproject.php?path=E.P.I.C.). The lack of language resources makes it difficult to get extensive research data. Therefore, a huge demand exists for more electronic data to be employed in interpreter training and in research on interpreting.

2. THE CORPORA

The research on the data from the interpreting corpora should not be confined to examining corresponding passages in the original speech and in the speech of the interpreter. The researcher should study the interpreting discourse as a whole: communication between the

participants and the interpreter (including pauses and interruptions), message transmission via interpreting, communicative failures during interpreting, the timing, etc.

Two Finnish-Russian interpreting corpora are currently being collected at the School of Modern Languages and Translation Studies of the University of Tampere. They are the Corpus of Court Interpreting (CIC, collected by Nina Isolahti) and the Corpus of Learners' Interpreting (CLI).The structure of the database is established and a pilot version of the search engine has been developed. The data is currently stored on the server of the Section of Russian Translation Studies (https://mustikka.uta.fi/spoken/, access restricted to the members of the research team).

In institutional discourse, part of communication often takes place in one language without the help of the interpreter, who gives assistance only when needed. The speakers often interrupt each other, forget that the other party does not understand their language, and as a result the interpreter works under pressure. The interpreting discourse is thus a blend of verbal and non-verbal communication, part of which is mediated by the interpreter (see e.g. González et al., 1991; Hale, 2004; Välikoski, 2004).

A corpus of interpreting discourse can be arranged as a hybrid of a bilingual corpus and a parallel one. It would be a collection of transcripts of speech in two (or even more) languages with part of the transcripts aligned (Mikhailov & Isolahti, 2008). Audio and visual components would be in many cases extremely useful additions to the corpus data. Unfortunately, it is not always possible to include them due to problems of ethical, copyright, and technical nature. However, remarks and comments are seriously considered as a part of corpus structure.

The transcripts are annotated using xml markup. The transcription is broad, however, speech is not smoothed up to written language as it happens in many projects, which do not directly contribute to linguistic research. We mark pauses and their lengths as well as some prosodic features (logical accent, rising/falling pitch, etc.). No punctuation marks are used in the transcripts but question and exclamation marks, which make reading easier. The features relevant from the point of view of translation process are also subject to markup, these are deletions, additions, changes, etc.

Nonetheless, the xml document is not the final representation of the corpus, which is stored in a database format. The reason why transcripts are not fed into the database directly is the relative ease of markup in xml, which can be done in any word processor. It is also quite a simple task to check the consistency of the markup. So, the data extracted from xml files are uploaded to Postgresql databases (http://www.postgresql.org). The database handles

many different routines like data maintenance, search, corpus users, sessions, etc. The most important for the search engine database tables are the following:

Transcripts. Each running word, pause, tag is stored in a separate record. All transcripts are loaded into the same table. This makes it possible to build concordances, word lists, collect statistics with the help of SQL queries.

Phrases. The start and end of each phrase is marked in Transcripts table with special tags. However, the data on the phrases (speaker, start time, end time, language, etc.) are stored in a separate table. The Phrases table is linked to the Transcripts table.

Lemmas. The lemmas of the word tokens are kept in separate tables (a table per language) linked to Transcripts table. This makes the Transcripts table more transparent, saves space on disk, and makes it possible to generate lemmatized frequency lists directly. Lemmatization is done after tokenization with external software. English and Finnish tokens are lemmatized with Connexor machinese phrase taggers (http://www.connexor.eu/technology/machinese/machinese phrasetagger/), Russian lemmatization is performed with Rmorph (http://www.cic.ipn.mx/~sidorov/ rmorph/index.html, Gelbukh & Sidorov, 2003).

Library. The information on each item of the corpus (e.g. an interview, a hearing at the Court, a class in interpreting, etc.) is stored in a separate record. The data

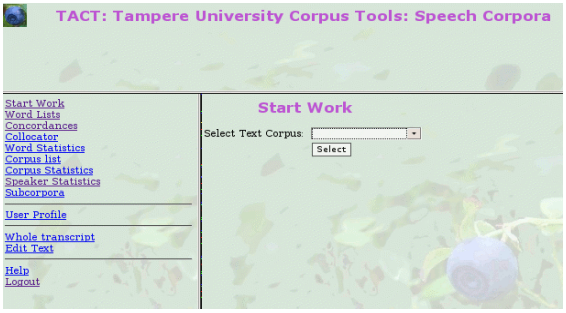


Figure 1: TACT: User interface.

available is text code in the corpus, title, author (for written text), date of issue, as well as text statistics (number of characters, number of running words, etc.).

3. CORPUS TOOLS

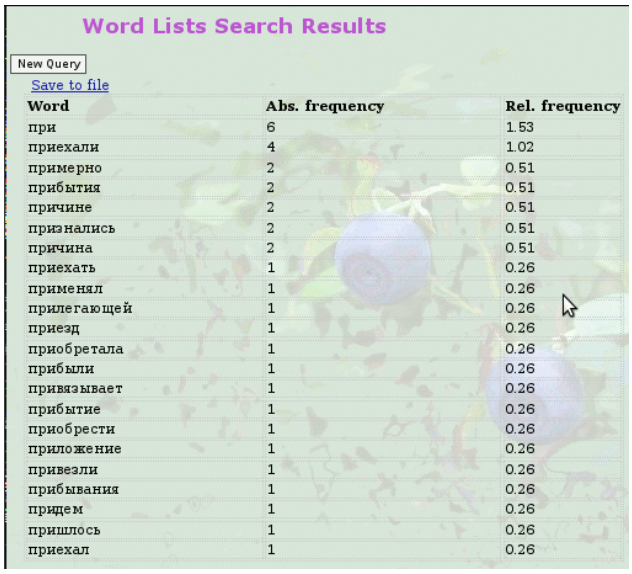
Maintenance of the corpus database (tokenization, lemmatization, updating statistics, etc.) is performed by running php-scripts via ssh protocol, administrator's privileges are required.

The most important and frequently used search routines are included into the TACT web interface (Tampere University Corpus Tools, developed by Mikhail Mikhailov). Not surprisingly, a written-language bias in the tools and methodology of spoken corpus research is quite obvious. I mean that same tools are used for processing spoken corpora as for the written ones. Some of the spoken corpora do not even use transcribing conventions (e.g. MICASE). The TACT package also includes routines, which can be used for processing written texts as well. However, certain functions were developed specially for spoken corpora. The software package is being constantly modified, and new functions are added to meet the requirements of the research team. Most of the functions work both with the whole corpus or with subcorpora (i.e. groups of texts defined by the user). The following research tools are currently available:

Frequency lists: tokens, lemmas, other elements;

Concordances: usage examples matching the search query;

Collocation lists: statistics on the tokens or lemmas frequently occurring in the close context from the search item;



Word	Abs. frequency	Rel. frequency
при	6	1.53
приехали	4	1.02
примерно	2	0.51
прибытия	2	0.51
причине	2	0.51
признались	2	0.51
причина	2	0.51
приехать	1	0.26
применял	1	0.26
прилегающей	1	0.26
приезд	1	0.26
приобретала	1	0.26
прибыли	1	0.26
привязывает	1	0.26
прибытие	1	0.26
приобрести	1	0.26
приложение	1	0.26
привезли	1	0.26
прибывания	1	0.26
придем	1	0.26
пришлось	1	0.26
приехал	1	0.26

Figure 3: Frequency lists: search result.

Corpus statistics: size of the whole corpus (in running words, phrases, characters), sizes of subcorpora;

Speaker statistics: pauses, speech tempo, etc.

3.1. Frequency lists

This tool is more flexible than standard applications for building word lists. The user can generate frequency lists of lemmas, running words, grammar tags, or even

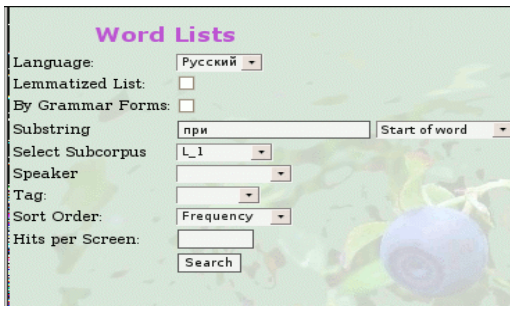


Figure 2: Word Lists. Dialog

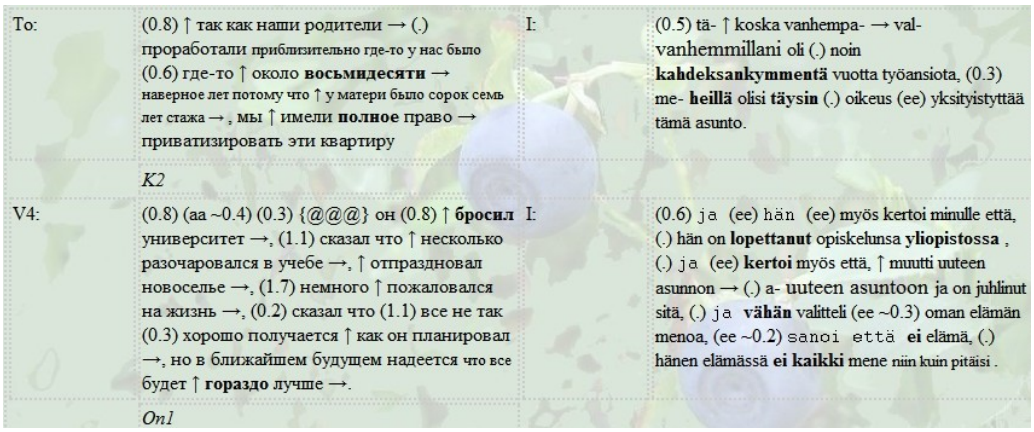
frequency lists of elements marked by certain tags, which makes it possible to create lists of terms, idioms, omissions, etc. The utility creates frequency lists for the whole corpus or for a subcorpus. Sometimes it might be quite useful to obtain a frequency list for a particular speaker. There is no need to waste time on generating the whole list if the researcher is interested only in the most frequent words, or in the words following a certain pattern. On Fig. 2 the user is requesting a Russian unlemmatized frequency list of tokens starting with *pri* from subcorpus *L_1* ordered by frequency. The resulting frequency list is displayed in Fig. 3 with absolute frequencies and relative frequencies per 1000 words.

3.2. Concordances

Looking up usage examples for words is a most common routine for the text corpora. Spoken corpora are a different resource compiled for purposes other than collections of written texts. Due to modest sizes of spoken corpora, only high frequency words can be looked up. Therefore, concordancing for specific words is not so typical. The researcher of spoken corpora might be more interested in other things, e.g. in prosody or certain pronunciation features.

The user of a corpus of interpreting may be looking for the phrases which were never interpreted, for the contexts where an interpreter speaks slowly, or for the contexts with the speaker and the interpreter speaking in chorus. So, the TACT concordance query makes it possible to use markup in the search and it is not obligatory for the user to supply a word to look for. As a result, the user can construct queries like the following one: "Find the dialogue units where the speech tempo is X, the speaker is Y, and the language is Z". The concordance in the Fig. 4 is the result of the query "Interpreter speaks loudly".

It is much more difficult to present a readable concordance from a speech transcript



To:	(0.8) ↑ так как наши родители → (.) проработали приблизительно где-то у нас было (0.6) где-то ↑ около **восьмидесяти** → наверное лет потому что ↑ у матери было сорок семь лет стажа →, мы ↑ имели **полное** право → приватизировать эти квартиру	I:	(0.5) tä- ↑ koska vanhempa- → val- vanhemmillani oli (.) noin **kahdeksankymmentä** vuotta työansiota, (0.3) me- **heillä** olisi **täysin** (.) oikeus (ee) yksityistyttää tämä asunto.
	K2		
V4:	(0.8) (aa ~0.4) (0.3) {@@@} он (0.8) ↑ **бросил** университет →, (1.1) сказал что ↑ несколько разочаровался в учебе →, ↑ отпраздновал новоселье →, (1.7) немного ↑ пожаловался на жизнь →, (0.2) сказал что (1.1) все не так (0.3) хорошо получается ↑ как он планировал →, но в ближайшем будущем надеется что все будет ↑ **гораздо** лучше →.	I:	(0.6) ja (ee) hän (ee) myös kertoi minulle että, (.) hän on **lopettanut** opiskelunsa **yliopistossa** , (.) ja (ee) **kertoi** myös että, ↑ muutti uuteen asunnon → (.) a- uuteen asuntoon ja on juhlinut sitä, (.) ja **vähän** valitteli (ee ~0.3) oman elämän menoa, (ee ~0.2) sanoi että **ei** elämä, (.) hänen elämässä **ei kaikki** mene niin kuin pitäisi .
	On1		

Figure 4: Concordance

than from a written text. Moreover, the interpreter's speech must be linked to the source speech. The solution we suggest is to use a two-column presentation with source speech in the left column and interpreter's speech in the right one (see Fig. 4). The rise and fall of the tone, emphasis and other prosodic features are also visualised. The problem of presenting speech overlapping remains unsolved; overlaps are currently marked with brackets, which is not very user-friendly.

3.3. Speaker statistics

The most interesting research tool of the TACT application is the utility presenting speaker statistics. It calculates speaker's speech tempo, number of pauses, total duration of pauses and other parameters. The script calculates statistics separately for all languages the person speaks in the recording. This feature is particularly important for the interpreter, the person, who is supposed to speak different languages during the meeting. It becomes possible to measure the fluency of speech, to compare the parameters of the translator's speech to that of the other speakers.

Speaker Statistics

Save to file

Speaker	Number of Words	Number of Phrases	Pauses: number / duration	Time	Average speech tempo
K1					
EAo, fi	34	4	21 / 35.7 s.	00:00:57	0.60/1.60
EV, fi	495	32	95 / 39.8 s.	00:03:26	2.40/2.98
I, fi	1312	119	459 / 159.04 s.	00:12:50	1.70/2.15
I, ru	1351	130	354 / 109.3 s.	00:16:19	1.38/1.55
S, fi	775	91	265 / 285.5 s.	00:11:22	1.14/1.95
T, fi	157	23	39 / 40.9 s.	00:01:42	1.54/2.57
V, ru	1591	128	315 / 204.8 s.	00:15:32	1.71/2.19
XX, fi	25	6	4 / 3.6 s.	00:00:13	1.92/2.66
K2					
Com, fi	34	1	4 / 1 s.	00:00:10	3.40/3.78
EAo, fi	36	2	14 / 6.9 s.	00:00:19	1.89/2.98
EV, fi	355	28	80 / 42.1 s.	00:02:53	2.05/2.71
I, fi	638	26	259 / 87.4 s.	00:06:49	1.56/1.98
I, ru	99	12	33 / 14.2 s.	00:00:56	1.77/2.37
S, fi	46	6	27 / 19.6 s.	00:00:38	1.21/2.50
T, fi	51	7	11 / 9.7 s.	00:00:29	1.76/2.64
To, fi	183	31	35 / 16.6 s.	00:01:29	2.06/2.53
To, ru	847	33	108 / 47.6 s.	00:04:55	2.87/3.42
XX, fi	4	3	3 / 4.2 s.	00:00:06	0.67/2.22

Figure 5: Speakers statistics

It is clear from the Fig. 5 that the interpreter (I, fi & I, ru) in the meeting K1 spoke Finnish faster than Russian, even though the number of pauses in Finnish speech was greater. The meeting K2 shows the opposite tendency, although both interpreters are native speakers of Russian. After obtaining more data, we would probably be ready to answer the question whether it is always easier to interpret into the mother tongue.

4. THE UNBIASED PICTURE OF REAL INTERPRETING

The only way to learn how interpreting happens is to collect the transcripts of the source speech and interpreter's speech, to do the markup and to align them. Even interpreters have a very vague idea of what amount of information was lost during interpretation, what kind of distortions took place, and if any misunderstanding between the parties happened. Studying the paper version of transcripts is extremely difficult and in most cases result in usage examples with little generalization. In contrast, using electronic corpora gives an opportunity not only to easily find usage examples but also to summarize the findings and to collect statistics.

I believe that the results of the research would be applied both directly and indirectly in Interpreting and Translation Studies. They would be indispensable in theoretical descriptions of the structure of multilingual and multimedia discourse. The corpora can also be used as a practical tool for training interpreters. The corpus-based research would also displace many myths and misapprehensions about interpreting and give the unbiased picture of this activity.

REFERENCES

González, R., Vásquez, V.F., Mikkelson, H. (1991). *Fundamentals of Court Interpretation. Theory, Policy, and Practice*. Durham, North Carolina: Carolina Academic Press.

Hale, S. (2004). The Discourse of Court Interpreting. Discourse practices of the law, the witness and the interpreter. Amsterdam/Philadelphia: John Benjamins.

Välikoski, T.-R. (2004). *The Criminal Trial as a Speech Communication Situation.* Tampere: Tampere University Press, 2004

Mikhailov, M. & Isolahti, N. (2008). Korpus ustnyx perevodov kak novyj tip korpusa tekstov (= The corpus of interpreting as a new type of text corpora, in Russian). *Dialog-2008 International Conference*, June 4th–8th in Moscow, http://www.dialog-21.ru/dialog2008/ materials/html/58.htm, accessed on 21.2.09.

Gelbukh, A. & Sidorov, G. (2003). Approach to construction of automatic morphological analysis systems for inflective languages with little effort. *Computational Linguistics and Intelligent Text Processing (CICLing-2003), Lecture Notes in Computer Science*, N 2588, Springer-Verlag, 2003, 215–220; www.cic.ipn.mx/~sidorov/ GelbukhSidorovMorphCICLING2003.ps.

Paradigmas derivativos del inglés antiguo organizados en torno a adjetivos básicos

CARMEN NOVO URRACA

Universidad de La Rioja

Resumen

Esta comunicación tiene como objetivo mostrar algunos paradigmas derivativos del léxico del inglés antiguo organizados en torno a adjetivos básicos. Se analizan distintos procesos de formación de palabras y predicados pertenecientes a todas las categorías. Se discuten también las implicaciones del análisis para la elaboración de la base de datos léxica del inglés antiguo Nerthus. Se llega a la conclusión de que el verbo fuerte no es el único que puede ser básico y dar lugar a paradigmas derivativos en inglés antiguo y de que en la derivación a partir de adjetivos básicos produce derivados de varias categorias y por medio de varios procesos distintos.

Palabras clave: inglés antiguo, base de datos léxica, formación de palabras, adjetivos básicos, paradigmas derivativos

Abstract

The aim of this paper is to analyse some derivational paradigms of Old English organized around basic adjectives. The analysis comprises different word-formation processses as well as members from all lexical categories. The implications are also discussed for the compilation of the lexical database of Old English Nerthus. Conclusions insist on the fact that, firstly, the strong verb is not the only one category that can give rise to derivational paradigms in Old English, and, secondly, basic adjectives turn out derived predicates of different categories throughout several processes.

Keywords: Old English, lexical database, word-formation, basic adjectives, derivational paradigms

1. INTRODUCCIÓN[1]

Esta comunicación muestra algunos paradigmas derivativos del inglés antiguo organizados en torno a adjetivos básicos. Un paradigma derivativo contiene todos aquellos predicados relacionados entre sí por forma y significado, de tal manera que todos ellos se crean a partir de uno básico. Los procesos de formación de palabras nuevas a partir de un predicado básico incluyen la composición, derivación cero, conversión y afijación. Estos procesos se ilustran con relación al paradigma que puede verse en (1):

[1] Investigación financiada con cargo el proyecto FFI08-04448/FILO.

(1) Adjetivo básico: ∂icce 1_{ADJ} ‘thick, viscous, solid; dense, stiff; numerous, abundant; hazy, gloomy; deep’

Compuesto: bri:w∂icce$_{ADJ}$ ‘as thick as pottage’

Derivado cero: ∂iccian$_{V}$ ‘to thicken; crowd together’

Convertido: ∂icce 2_{ADV} ‘thickly, closely; often, frequently’

Prefijado: in∂icce$_{ADJ}$ ‘crass, thick’

Sufijado: ∂iccet$_{N}$ ‘thinck bushes, thicket’

Este paradigma derivativo nos sirve para ilustrar todos los procesos de formación de palabras en inglés antiguo. Consideramos básicos todos aquellos predicados que no se pueden descomponer morfológicamente. Tanto en (1) como en el resto de los ejemplos que se discuten en esta comunicación, se trata de los adjetivos en torno a los cuales se organizan distintos paradigmas derivativos. Son compuestos, como se admite en general en el campo de estudio, los elementos resultantes de la combinación de dos formas libres. En (1) hay un único predicado compuesto, *bri:w∂icce*, compuesto por *bri:w* ‘pottage, porridge’ y *∂icce 1* ‘thick, viscous, solid; dense, stiff; numerous, abundant; hazy, gloomy deep’. De acuerdo con Martín Arista (en prensa-b), la derivación cero es la derivación sin morfema derivativo y la derivación por medio de morfema flexivo, mientras que la conversión es la extensión categorial sin cambio formal. El adjetivo *∂icce 1* cuenta con un único derivado cero en su paradigma derivativo, *∂iccian*. En cuanto a la conversión, ésta aparece representada por el adverbio *∂icce 2*, que ha sufrido extensión categorial respecto al adjetivo básico. El último proceso a tener en cuenta es la afijación, que consta en sus vertientes tradicionales de la prefijación y la sufijación. En el ejemplo (1) encontramos como prefijado *in∂icce* y como sufijado *∂iccet*.

En general, existe consenso en la disciplina respecto a que el verbo fuerte con sus formas de presente, pretérito y participio perfecto constituye el punto de partida de la derivación léxica en las lenguas germánicas antiguas. Kastovksy (1992) siguiendo a Hinderling (1967) es de esta opinión. En efecto, la investigación llevada a cabo para compilar la base de datos léxica del inglés antiguo Nerthus (www.nerthusproject.com) ha demostrado que alrededor del 40% de las palabras derivadas proceden directa o indirectamente de los verbos fuertes (Martín Arista, en prensa-a).

La categoría adjetivo también organiza paradigmas derivativos. Nos hemos basado en el trabajo de Heidermanns (1993) y de Orel (2003) para proponer unos 200 paradigmas basados en esta clase léxica. En especial, Heidermanns nos muestra los adjetivos básicos del

inglés antiguo entre otras lenguas germánicas antiguas y añade predicados que pertenecen al paradigma derivativo de estos adjetivos. Además utiliza varias formas para clasificarlos, entre ellas, crea un sistema de letras que añade al lado de cada entrada, que indican si la motivación de ese adjetivo es germánica o anterior. En esta comunicación he seleccionado cinco familias de adjetivos básicos a los que Heidermanns añade la letra *P*, es decir, adjetivos primarios de los que se deriva toda una familia indicando la *P* que no hay ningún tipo de motivación germánica. A partir de estos cinco ejemplos, veremos las implicaciones para el análisis llevado a cabo hasta ahora en la base de datos léxica *Nerthus*. Los datos aportados por Heidermanns nos han llevado a modificar algunos paradigmas derivativos de dos maneras: bien extrayendo el predicado de un paradigma para organizar uno nuevo en torno a dicho predicado o creando directamente un paradigma nuevo.

Los datos para esta comunicación pertenecen a la ya citada base de datos léxica Nerthus, compuesta por unas 30.000 predicados en su mayoría extraídos de Clark Hall (1996) *A Concise Anglo-Saxon Dictionary,* pero también se incluyen datos extraídos de Bosworth Toller (1973) *An Anglo-Saxon Dictionary* y de Sweet (1976) *The student´s dictionary of Anglo- Saxon.* Cabe mencionar que ahora mismo está en proceso la comparación completa de los datos ya incluidos en Nerthus con el diccionario de Bosworth-Toller, ya que anteriormente esta fuente sólo se utilizó para resolver dudas y completar datos que faltaban en Clark Hall. Ahora, como resultado de la comparación, preevemos que aumentará el número de predicados en torno al 20%, ya que Bosworth Toller hace una compilación más amplia del inglés antiguo que la de Clark Hall.

2. PARADIGMAS ADJETIVALES

En primer lugar, nos centraremos en paradigmas derivativos que han sido modificados, es decir, la situación que se produce cuando es necesario extraer uno predicado o varios predicados de un paradigma derivativo para formar otro paradigma derivativo nuevo.

El primer paradigma se organiza en torno al adjetivo básico *gle:aw*. En un principio *Nerthus* contemplaba este predicado como un derivado cero del verbo fuerte *glo:wan*, ya que por la forma podría perfectamente formar parte del paradigma derivativo de este verbo, pero el significado plantea dudas. El verbo *glo:wan* significa ‘to glow’, mientras que el adjetivo *gle:aw* significa ‘penetreting, keen, prudent, wise, skilful; good’, con lo que los significados de ambos no están relacionados. Heidermanns nos propone *gle:aw* como un adjetivo básico con los siguientes predicados:

(2) Adjetivo básico: *gle:aw*

1 adjetivo: *ungle:aw*

2 adverbios: *gle:awe, gle:awli:ce*

2 nombres: *gle:awnes, gle:awscipe*

Vistos los predicados ofrecidos por Heidermanns, hemos trasladado los datos a *Nerthus*, para comprobar si la base de datos incluye más predicados que pertenezcan al paradigma, y hemos encontrado 26 entradas, que se muestran en (3):

(3) Adjetivo básico: *gle:aw* 'penetrating, keen, prudent, wise, skilful; good'

16 adjetivos: *æ:gle:aw* 'learned in the law'
cræftgle:aw 'skilful, wise'
ferhðgle:aw 'wise, prudent'
foregle:aw 'forseeing, provident, wise, prudent'
fre:agle:aw'very wise'
gle:awferhð 'prudent'
gle:awhy:dig'thoughtful, wise, prudent'
gle:awhycgende'thoughtful, wise, prudent'
gle:awlic 'wise, prudent, skilful, diligent'
gle:awmo:d 'wise, sagacious'
hreðergle:aw'wise, prudent'
hyrgegle:aw 'prudent in mind'
mo:dgle:aw 'wise'
steorgle:aw 'clever at astronomy'
ungle:aw 'ignorant, foolish, unwise'
wordgle:aw 'skilful in words'

4 adverbios: *foregle:awli:ce* 'providently, prudently'
gle:awe'wisely, prudently, well'
gle:awli:ce 'prudently, wisely, clearly, well'
ungle:awli:ce 'without understanding, without sagacity, unwisely, imprudently'

5 nombres: *gereordgle:awnes* 'skill in singing'
gle:awnes 'wisdom, prudence, skill, penetration;

diligence; sign, token'

gle:awscipe 'wisdom, thoughtfulness, diligence; proof, indication, test'

ungle:awnes'want of discernment, folly, ignorance'

ungle:awscipe 'folly'

Todos los predicados listados en el ejemplo (2) y (3) se crean a partir de *gle:aw* mediante distintos procesos de formación de palabras, mencionados en la introducción (composición, derivación, derivación cero, conversión y afijación). Cabe mencionar, que no todos las palabras se forman directamente desde el adjetivo básico, sino que hay distintas generaciones o niveles de derivación, y aunque este tema no se va a tratar en esta comunicación, valga como ejemplo el dado en (4), en el que el adverbio *gle:awli:ce* deriva directamente del adjetivo *gle:aw* y el adverbio negativo *ungle:awli:ce*, a su vez del adverbio *gle:awli:ce*.

(4) *gle:aw* > *gle:awli:ce* > *ungle:awli:ce*

↓ sufijación ↓ prefijación

Este fenómeno se puede observar en todas las familias que forman parte de esta comunicación pero también en prácticamente todos los paradigmas ya incluidos en *Nerthus*.

El segundo paradigma que hemos seleccionado requiere las mismas modificaciones que el primero, es decir, se trata de predicados que estaban incluidos dentro de otro paradigma, y que han sido extraídos para agruparlos dentro del paradigma derivativo de un adjetivo básico. En este caso ese adjetivo es *teart*, para el que Heidermanns propone cuatro predicados, incluido el propio adjetivo. Esto se muestra en (5):

(5) Adjetivo básico: *teart*

1 adjetivo: *teartlic*

1 adverbio: *teartli:ce*

1 nombre: *teartnes*

En *Nerthus* encontramos un total de 6 predicados que hemos extraído del verbo fuerte *teran*, y que hemos pasado a organizar dentro del adjetivo *teart*, como puede verse en (6):

(6) Adjetivo básico: *teart* ‘sharp, rouge, severe’
2 adjetivos: *teartlic* ‘sharp, rough’
teartnumol ‘biting, effectual’
1 adverbio: *teartli:ce* ‘sharply, sverely’
1 nombre: *teartnes* ‘sharpness, roughness, hardness’

Igual que en el ejemplo (3), el adjetivo *teart* que estaba analizado como un derivado cero del verbo fuerte *teran*, pasa ahora a convertirse en un básico dando lugar al paradigma derivativo ya mostrado.

A continuación vamos a ver otros tres paradigmas que se han creado a partir de predicados que estaban sin analizar ni clasificar. El primer paradigma surge a partir del adjetivo *smo:ð*. En este caso, Heidermanns propone un total de tres palabras a incluir en esta familia, incluido el adjetivo básico, que se da en (7):

(7) Adjetivo básico: *smo:ð*
2 adjetivo: *sme:ðe, unsmo:ð*

Al trasladar estos datos a *Nerthus*, y realizar una búsqueda de las posibles entradas contenidas en la base de datos que pueden formar parte del paradigma derivativo de este adjetivo, hemos encontrado un total de 8 predicados, que hemos agrupado en torno a *smo:ð*. El listado aparece en (8):

(8) Adjetivo básico: *smo:ð*’smooth, serene, calm’
3 adjetivos: *sme:ðe* ‘smooth, poilished, soft; suave, agreeable; not harsh (of the voice); lenitive’
unsme:ðe’not smooth, uneven, rouge, scabby’
unsmo:ð’not smooth, uneven, rouge, scabby’
2 nombres: *sme:ðnes*’smoothness, smooth place’
unsme:ðnes ‘roughness’
2 verbos débiles: *(ge)sme:ðan* ‘to smooth, soften, polish; appease, soothe’
(ge)smeðian ‘to smoothen; become smooth’

En la mayoría de los casos, los procesos de formación de palabras en los que ha sido incluido cada predicado no varían, independientemente de que antes no formasen parte de un paradimga y ahora sí. Solamente aquellas palabras sobre las que quedaba decidir si eran básicas o derivadas varían. Así por ejemplo, el predicado *sme:∂e* pasa a clasificarse como derivado cero.

El siguiente paradigma se organiza en torno al adjetivo básico *wlaco*. Al igual que en el ejemplo anterior, los predicados que forman el paradigma derivativo de este adjetivo no estaban incluidos en ninguna familia en *Nerthus*. Heidermanns acompaña este adjetivo de 5 predicados, como puede verse en (9):

(9) Adjetivo básico: *wlaco*
1 adverbio: *wlæcli:ce*
2 nombres: *wlæce, wlæcnes,*
2 verbos: *(ge)wleccan, (ge)wlacian*

Tras realizar la búsqueda oportuna en *Nerthus*, ésta arroja un total de 7 predicados incluido el adjetivo básico, que deben incluirse en este paradigma, dado en (10):

(10) Adjetivo básico: *wlaco* 'tepid, lukewarm, cool'
1 adjetivos: *wlæclic* 'lukewarm'
2 nombres: *wlæce* 'tepidity'
wlænes 'lukewarness'
3 verbos: *a:wlacian* 'to ve or become lukewarm'
(ge)wlacian 'to become lukewarm, be tepid'
(ge)wleccan 'to make tepid'

Podemos observar que Heidermanns propone un predicado *wlæcli:ce* que *Nerthus* no contempla. Sin embargo, esta palabra sí representa una entrada en Bosworth Toller; es un adverbio cuyo significado es 'lukewarmly'. Como ya ha sido mencionado al principio, ahora mismo está en proceso la comparación entre *Nerthus* y Bosworth Toller, y este predicado forma parte del aproximadamente 20% de entradas nuevas pendientes de creación en *Nerthus*, a raíz del cual clasificaremos este predicado dentro del paradigma derivativo del adjetivo básico *wlaco*.

El último paradigma que voy a mostrar no estaba organizado anteriormente en *Nerthus*. Se trata del adjetivo básico *wlanc*, al que en Heidermanns acompañan otros dos predicados, dados en (11):

(11) Adjetivo básico: *wlanc*
1 adjetivo: *gewlenced*
1 nombre: *wlenc*

Por su parte, *Nerthus* devuelve muchos más predicados que deben ser incluidos. De hecho, encontramos 18 palabras, que anteriormente no formaban parte de ningún paradigma derivativo y aparecen en (12):

(12) Adjetivo básico: *wlanc* 'stately, splendid, lofty, magnificient, rich; boastful, arrogant, proud'
6 adjetivos: *felawlonc* 'bery stately'
goldwlanc 'brave with gold, richly, adorned'
hygewlanc 'haughty, proud'
mo:dwlanc 'stout-hearted; haughty'
symbelwlonc 'elated with feasting'
wlanclic 'proud, arrogant'
1 adverbio: *wlancli:ce* 'proudly, arrogantly'
4 nombres: *goldwlencu* 'gold ornament'
oferwlencu 'excessive riches'
wlenc 'pride, arrogante, haughtiness; glory, promp, splendour, bravado; prosperity, riches, wealth'
woruldwlenco 'magnificence, ostentation'
6 verbos débiles: *a:wlancian* 'to exult, to be proud'
a:wlencan 'to make proud, enrich'
forwlencan 'to fill with pride, puff up'
(ge)wlencan 'to enrich, exalt'
oferwlencan 'to be bery wealthy'
wlancian 'to become proud or boastful, exult'

A propósito de este paradigma, y en concreto del verbo *gewlencan* 'to make proud, rich, to exalt' debe tenerse en cuenta que no he encontrado la forma flexiva *gewlenced* en ninguno de los diccionarios citados antes, ni en el Dictionary of Old English Corpus (di Paolo Healey, 2004).

Como en casos ya mencionados, este paradigma incluye tres predicados, *wlanc, wlancian* y *wlenc*, sobre los que se planteaban dudas respecto a si eran básicos o derivados cero. Indiscutiblemente, ahora *wlanc* es el único que permanece como básico, ya que es la palabra a partir de la cual se forman todas las incluidas en la familia, y *wlancian* y *wlenc* pasan a clasificarse definitivamente como derivados cero de este adjetivo.

Finalmente, y con respecto al análisis, cabe destacar el hecho de que los predicados compuestos pueden, y en la mayoría de los casos así sucede, pertenecer al menos a dos paradigmas, el adjunto pertenece a un paradigma y la base a otro. Sirvan como ejemplo, uno de los compuestos de los paradigmas que hemos discutido, repetido en (13):

(13) *cræftgle:aw* 'skilful, wise' < *cræft* + *gle:aw*
cræft 'physical strength, might, courage; science skill art, ability, talent, virtue, excellence; trade, handicraft, calling; work or product of art; trick, fraud, deceit, machina, instrument; in pl. great numbers, hosts?' < *crafian*
gle:aw 'penetrating, keen, prudent, wise, skilful; good' < *gle:aw*

Como se puede ver en (13), las dos palabras que forman este predicado compuesto pertenecen a paradigmas derivativos distintos. De manera que este predicado aparecerá incluido en ambos paradigmas, en el de *crafian* por *cræft* y en la de *gle:aw* por el propio *gle:aw*, de manera que habrá que estudiarlo desde ambas perspectivas para tener una visión completa de la formación de la palabra.

3. CONCLUSIÓN

Esta comunicación sólo muestra algunos ejemplos representativos del tipo de análisis y las fuentes que estamos consultando en la elaboración de una base de datos léxica del inglés antiguo. En lo que respecta a los adjetivos, la mayor parte de los adjetivos básicos y los paradigmas que se que agrupan en torno a ellos coinciden completamente con el análisis de Heidermanns. Hay algunos casos, como los mostrados en esta comunicación, en los que la

propuesta de Heidermanns es más coherente y facilita la tarea de completar el analisis de la parte correspondiente a la formación de palabras de la base de datos léxica. Por otra parte, *Nerthus* incluye muchos más predicados que Heidermanns, que solo nos da una pequeña referencia de cada paradigma, y que nosotros hemos ampliado, en algunos casos considerablemente, gracias a la información proporcionada por *Nerthus*.

En general, con el análisis y la información obtenidos hasta el momento podemos concluir que evidentemente no sólo los verbos fuertes constituyen paradigmas derivativos en inglés antiguo, sino que cualquier predicado perteneciente, al menos, a las categorías léxicas mayores (nombre, adjetivo, verbo), puede ser base de derivación de distintas palabras. Además, en el caso de los adjetivos, éstos son base de predicados pertenecientes a todas las categorías: adjetivos, adverbios, nombres y verbos débiles; y por medio cualquier proceso de formación de palabras: composición, derivación, derivación cero, conversión y afijación.

Finalmente, como investigación pendiente cabe destacar la necesidad de completar la formación de paradigmas derivativos de base adjetival, para lo cual es necesario, como paso previo, revisar el inventario de registros de la base de datos.

REFERENCIAS BIBLIOGRÁFICAS

Bosworth, J. y Toller, T. N. (1973). *An Anglo-Saxon Dictionary.* Oxford: Oxford University Press.

Clark Hall, J.R. (1996). *A Concise Anglo-Saxon Dictionary.* Toronto: University of Toronto Press.

diPaolo Healey, A. *et al*. (2004). *The Dictionary of Old English Corpus in Electronic Form*, (Toronto: DOE Project 2004, on CD-ROM).

Heidermanns, F. (1993). *Etymologisches Wörterbuch der germanischen Primäradjective*. Berlin – Nueva York: de Gruyter.

Martín Arista, J. Building a lexical database of Old English: issues and landmarks. In Considine, J. ed. *Current projects in historical lexicography*. Newcastle: Cambridge Scholars Publishing. En prensa-a.

Martín Arista, J. Morphological relatedness and zero alternation in Old English. In C. Butler and P. Guerrero (eds.) *Morphosyntactic Alternations in English*. London: Equinox. En prensa-b.

Orel, V. (2003). A Handbook of Germanic Etymology. Leiden: Brill.

Sweet, H. (1976). *The Student's Dictionary of Anglo-Saxon.* Cambridge: Cambridge University Press.

Lexical evidential verbs in English computing scientific articles

IVALLA ORTEGA BARRERA

MARGARITA ESTHER SÁNCHEZ CUERVO

CeTIC (Centro Tecnológico para la Innovación en las Comunicaciones)

Universidad de Las Palmas de Gran Canaria

Abstract

Different studies have analysed the linguistic devices used by writers to transmit their personal feelings and evaluation. The writer can be subjective or objective depending on his attitude towards the topic. The present paper forms part of the research project "Evidentiality in a multidisciplinary corpus of research papers in English" at the University of Las Palmas of Gran Canaria. The aim of this paper is to analyse the presence and use of lexical evidential verbs in a corpus of English computing scientific articles as an instrument to state the degree of certainty and the source of knowledge used by the author. The analysis performed shows the choice of verbal forms that reinforce the notion of objectivity and confidence in the hypothesis design, and the soundness of the final outcomes.

Keywords: evidentiality, modality, scientific, lexical verbs

Resumen

Diversas investigaciones han analizado los mecanismos lingüísticos utilizados por los escritores a la hora de transmitir sus sentimientos y evaluaciones. El autor puede ser subjetivo u objetivo, dependiendo de su actitud hacia el tema que expone. El presente trabajo se enmarca en el proyecto de investigación "Evidencialidad en un corpus multidisciplinar de artículos científico-técnicos en lengua inglesa", de la Universidad de Las Palmas de Gran Canaria. El objetivo de este trabajo es analizar la presencia y el uso de los verbos léxicos evidenciales que aparecen en un corpus de artículos de investigación informáticos, como instrumento para manifestar el grado de certeza y la fuente de información utilizada por el autor. El análisis realizado demuestra la elección de formas verbales que refuerzan la noción de objetividad y seguridad en el planteamiento de la hipótesis, y en la firmeza de los resultados obtenidos.

Palabras clave: evidencialidad, modalidad, científico, verbos léxicos

1. INTRODUCTION

Numerous studies have examined the linguistic devices used by writers to transmit their personal feelings and evaluations. Charles, 2003; Hyland, 1994; 1996a; 1996b; Meyer, 1997, for instance, focus on the author's attitude. This attitude of the author towards the topic can be either emotional/expressive (subjective) or rational/evaluative (objective). The latter is related to evidentiality since the author's intention is expressed through the use of different markers to state the degree of certainty and the source of knowledge. From a lexical perspective, the markers used to express evidentiality are nouns, adjectives, and verbs.

In this paper we will focus on evidentiality through the presence of lexical verbs in a corpus of English computing scientific articles. Our main aim is to analyse the lexical way of

referring to the source of information included in each section of the computing scientific articles, except for the title and the reference sections.

The structure of this paper is as follows: first, we describe the corpus used for this paper. After that, we explain the meaning of evidentiality to give a general overview of the concept and the background. The following section touches upon the lexical verbs found in our corpus and their analysis as far as evidentiality is concerned. In the last section we offer the results and conclusion of our research.

2. CORPUS

The corpus used for the analysis has been excerpted from the Corpus of Specialized Research Papers in English. The creation of this corpus is part of a project called *Evidentiality in a multidisciplinary corpus of research papers in English*, currently undergone in the University of Las Palmas de Gran Canaria by the TeLL research group ("Tecnologías Emergentes aplicadas a la Lengua y la Literatura"). This multidisciplinary corpus is characterised by the inclusion of three different register domains: legal, medical and computing. The criteria for selection concern the impact index, year of publication, and sociological aspects. The texts for compilation must have a high impact index level, they cover a time span of 10 years, from 1998 to 2008, and they all are available in online databases through Las Palmas de Gran Canaria University Library hired databases. There are no more than two articles per year from one journal, covering between four and eight articles per year, amounting to a total of 32000 words per year for each register subdomain. All articles are written in English and at least one of the authors has to be a native speaker. We have currently compiled and stored the texts awaiting for XML tagging following TEI2 recommendations.

To carry out our analysis, we have selected eight computing articles from the corpus above mentioned. We have taken two articles per year from different academic journals published between 2005 and 2008.

3. ANALYSIS

Scientific research articles are considered as an independent genre which can be divided into different stages: title, abstract (itself an independent genre), introduction, materials and methods, results, discussion, conclusion and references. Each section has a main goal: the title expresses what the article is about; the abstract is a brief summary of the whole article; the introduction establishes the context of the research and depicts the exact subject matter of

the paper; the materials and methods describe the research, including the locations and times of sample collection, what equipment will be employed, and the techniques that are used; the results describe the outcomes of the research article including a summary of those facts, figures and statistical tests used to obtain the final data. In the discussion the writers interpret the results obtained; the conclusion sets up the research contribution to the area of study, and the references is the list of the sources used in the article.

According to Grossman and Wirth (2007: 202), "evidentiality covers all marks signalling what testifies to the validity of the information stated by a speaker or writer". Following this definition, different authors have studied the presence of evidentiality or evaluative expressions in research articles. Some of them (Crompton, 1997; Ferrari, 2006) have focused on some sections of the research articles and other authors have dealt with the research paper as a whole (López Ferrero, 2005; Chafe, 1986; Grossmann and Wirth, 2007). In the case of studies on verbs in research articles, several authors have underlined verbs that express evaluation or stance (Hunston, 1995), as reporting verbs (Thompson and Ye, 1991) or appearance verbs (Gisborne and Holmes, 2007).

Following Biber, Johansson, Leech, Conrad, and Finegan's (2000) classification, we can divide verbs into three major classes: lexical verbs, primary verbs, and modal verbs. Primary verbs are those that can function "as either main verbs or auxiliary verbs" (Biber *et al.*, 2000: 358), and modal verbs can function only as auxiliary verbs. In our case, we have decided to analyse lexical evidential verbs as they "comprise an open class of words that function only as main verbs" (Biber *et al*. 2000: 358). However, as in an academic article the information that is aimed at being proved is supposedly tested by means of different tests and experiments, we have decided to incorporate the primary verb *be.* While reading the considerable amount of utterances that contain this verb, either in the form of definitions or assertions, we assume that the information comes from a very reliable source. In this specific occasion, no assumptions should be made about how to interpret these utterances:

(1) Another case *is* when system failure results from an unanticipated change […].

(2) One way of increasing a system's availability *is* to make it fault-tolerant.

The certainty that is conveyed by the use of the verb *be* in these two examples provides that the information presented is true. There is no reason to think about the possibility of being wrong; the claims are presented as facts. If knowing the source of information gives a

cue to the validity of information, evidential meaning constitutes a procedure that may contribute to this process (Boye and Harder, 2009: 34).

In the abstract section of the papers revised, the most used verbs are *show*, *derive, construct* and *present*, as can be seen in the following examples:

> (3) We *show* that fault recovery done by intrinsic reconfiguration has some restrictions […].
>
> (4) We *present* an algorithm for calculating hypervolume exactly […].
>
> (5) The paper *derives* and *constructs* a qualitative framework […].

These verbs confirm evidentiality in the sense that they all introduce the main objective of the research, and imply the author's responsibility for the assertion. The verb *show* marks the evidence offered by the data. In the case of the verb *present*, it implies certainty and achievement, that is, the author brings in the research without judging or evaluating it. The verb *construct* implies the idea of process, and the verb *derive* can be also seen as part of a procedure. All these verbs involve an assumed knowledge. Thus, in (3) and (4) the verbs *show* and *present* imply the author's responsibility for the assertion by using the subject pronoun *we*, whereas in the last example (number 5) the sentence is depersonalised, so the own text is the source of knowledge.

In the introduction, the lexical evidential verbs used are *describe*, *present*, *provide*, *show* and *use*:

> (6) The empirical study *provides* a body of evidence for use in understanding the behaviour of the metrics.
>
> (7) We *describe* a simple distributed algorithm that permits terminals to schedule collision-free transmissions with the only requirement.
>
> (8) This work *uses* the term "visualization" in the latter sense […].

All these examples indicate an objective and clear presentation of what the reader is going to encounter in his findings of the article contents. In (6) and (8), the text is the source of the judgement (*the empirical study, this work*) so there is depersonalisation, whereas in (7), the authors get involved in the discourse by using the personal subject *we*. By employing this pronoun, the authors leave the claim open to the reader's judgement.

In the materials and methods section, the most utilised evidential verbs are *apply*, *present*, *describe*, *refer,* and *represent,* as shown in the samples:

(9) Their analysis does not *apply* to CGP because all CGP genotypes are bounded.

(10) In the following sections, parenthetical references refer to the leftmost column in Table 3 […].

(11) These complexity metrics are likely to represent worst-case rather than typical costs.

In (9) and (10), the verbs *apply* and *refer* denote a precise construction that can be related to some other information that should clarify the author's exposition. The last example can be considered as a synonym of *show*, in the sense that it illustrates some evidence presented in the information given.

The stages devoted to the results and the discussion will be analysed together since in some papers they are not clearly divided into exact sections. Those evidential verbs that mostly occur here are *show*, *perform*, and *calculate*:

(12) The net effect of changes was also calculated by summing the values of version-to-version changes […].

(13) Clearly, the constant power allocation *performs* very poorly.

These two instances express the idea of process that has been accomplished through different mathematical operations. Here the reader should verify the results offered, conferring certainty on the propositions. This kind of judgement derives from inferential reasoning or theoretical calculation Thus, in (12) the presence of the passive voice gives prominence to the object and marks the author's position (Baratta, 2009), that is, the writer avoids commitment, showing some distance between the author and his proposition.

In the conclusion the most recurrent verbs are *describe*, *present* and *require*. *Describe* and *present* have the same value as in the analysis exposed above. As to *require*, we perceive the need for a recommendation or urge in the author's final remarks, as seen in the following example:

(14) A standard casual model requires the specification of a set of structural equations and a mapping from equations to variables.

In this instance, although the need for some recommendation to validate the scientific knowledge seems to be implied in this sentence, the presence of *a standard casual model* as the subject of the sentence clarifies that the text is the source of knowledge and, hence, the evidential nature of the verb is proved.

After this brief analysis, we consider, furthermore, a possible classification that may convey the source of information, that is to say, how this knowledge is created or acquired. For this purpose, we offer Willet's (1988) conception of evidentiality that is based on a triple system: information that can be attested, reported and inferred. The last two systems represent indirect ways of presenting information. Attested information is acquired through the senses; information is reported when obtained from hearsay or folklore and, as regards the inferential information, it can be marked as relating to observable evidence (results) or mental constructs (logic, intuition, or dreams).

In the articles under scrutiny, the authors get their data after testing different experiments, hypotheses and formulae, being the results generally numerically proved. In accordance with the analysis presented, with a majority of lexical evidential verbs that express the writers' full commitment to the truth-value of the proposition, the source of information would be inferential. They would be achieved in this occasion by the results and findings which are prototypical of scientific communities, motivated by epistemological reasons (Koutsantoni, 2004: 172-173). Furthermore, these expressions of certainty impose views on readers, control their inferences, do not consent for room for disagreement or negotiations, and regard them as passive recipient or ideas, unable to make their own evaluations and judgements (Hyland, 2000). This idea has been explained throughout the analysis, with instances such as *show*, *provide*, *represent*, *present*, *refer*, and *perform*, among others.

4. CONCLUSION

This paper has considered some most usual occurrences of evidential lexical verbs in computing scientific articles. Bearing in mind the different sections that an academic article is generally comprised of, we have encountered several coincidences in our corpus that allows us to establish the authors' confidence. This assurance has been conveyed by using specific verbs from the beginning of each article, wherein the notion of an assumed knowledge that is to be verified all through the contents is being demonstrated in each subsequent section, up to the closing remarks. This certainty of ideas posed at the beginning of each article is in

relation to the source of information transmitted. The inferential information is the most characteristic, in this study, to offer the validity of those results that have been previously described in detail to the reader. Therefore, evidentiality proves in this particular case the writers' stated authority in their findings and soundness of their outcomes.

REFERENCES

Baratta, A. M. (2009). Revealing stance through passive voice. *Journal of pragmatics*, *7*, 1406-1421.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2000). *Longman Grammar of Spoken and Written English*. Longman: London.

Boye, K. and P. Harder (2009). Evidentiality. Linguistic Categories and Grammaticalization. *Functions of Language*, *16.1*, 9-43.

Chafe, W. (1986). Evidentiality in English conversation and academia writing. In Wallace Chafe and Johanna Nichols (eds.), *Evidentiality: The linguistic coding of epistemiology* (pp. 261-272). Norwood, NJ: Abblex.

Charles, M. (2003). `This mystery...´: A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes,* 2, 313-326.

Crompton, P. (1997). Hedging in academic writing: Some theoretical problems. *English for Specific Purposes, 16*, 271-287.

Ferrari, L. (2006). Modalidad y evaluación en las conclusiones de artículos de investigación. *Proceedings from 33rd International Systemic Functional Congress*, 514-530.

Gisborne, N. and Holmes, J. (2007). A history of English evidential verbs of appearance. *English language and linguistics,* 11.1, 1-29.

Grossmann, F. and Wirth, F. (2007). Marking evidentiality in scientific papers: the case of expectation markers. In K. Flottum (ed.), *Language and discipline perspectives on academic discourse* (pp. 202-218). Cambridge scholars publishing, Newcastle, UK.

Hunston, S. (1995). A corpus study of some English verbs of attribution. *Functions of Language, 2,* 133-158.

Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for specific purposes,* 13, 239-256.

Hyland, K. (1996a). Talking to the academy: Forms of hedging in science research articles. *Applied linguistics,* 17, 251-281.

Hyland, K. (1996b). Writing without conviction? Hedging in science research articles. *Applied linguistics,* 17, 433-454.

Hyland, K. (2000). Disciplinary Discourses: Social Interactions in Academic Writing. London, Longman.

Koutsantoni, D. (2004). Attitude, Certainty and Allusions to Common Knowledge in Scientific Research Articles. *Journal of English for Academic Purposes*, 3, 163-182.

López Ferrero, C. (2005). Funciones retóricas en la comunicación académica: formas léxicas de modalidad y evidencialidad. *Signo y Seña*, 14, 115-139.

Meyer, P. G. (1997). Hedging strategies in written academic discourse: Strengthening the argument by weakening the claim. In R. Markkanen, & H. Schroder (eds.), *Hedging and Discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts* (pp. 21-41). Berlin: Walter de Gruyter & Co.

Thompson, G. and Ye, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied linguistics,* 12, 365-382.

Willet, T. (1988). A Cross-linguistic Survey of Grammaticalization of Evidentiality. *Studies in Language*, 12, 51-97.

Pride – Stolz – orgullo: A corpus-based approach to the expression of emotion concepts in a foreign language

ULRIKE OSTER

Universitat Jaume I

Abstract
The expression of emotions in a foreign language is difficult even for advanced learners because, even if there is a shared physiological and psychological basis for the feelings themselves, there can be considerable differences among languages in how an emotion is expressed linguistically. This paper presents an example of how the use of corpora can contribute to raising learner awareness of these differences. It describes a task that combines different types of corpus data (information from corpus-based dictionaries, lists of collocations and co-occurrences provided by on-line corpora, and corpus searches as well as analyses of sample concordance lines performed by the learners themselves). Working on the task requires students to process the foreign language data actively and to consciously trace the way an emotion concept (pride) is conceptualised and expressed in the foreign language in contrast to their mother tongue.
Keywords: data-driven learning, conceptual metaphor theory, figurative language, corpus-based approaches to language learning, emotion concepts

Resumen
En este trabajo se aborda la utilización de corpus como herramienta para el aprendizaje de lenguas, concretamente en el ámbito de la expresión de conceptos de emoción y el lenguaje figurado. Se presenta una actividad que combina la utilización de diferentes recursos basados en corpus (desde información de diccionarios basados en corpus hasta búsquedas independientes en diferentes corpus en línea realizadas por los alumnos). En dicho análisis, se tienen en cuenta las metáforas y metonimias conceptuales, la proximidad conceptual así como aspectos descriptivos y evaluativos (prosodia semántica). A través de este trabajo consciente con datos lingüísticos, el aprendiz explora cómo los hablantes expresan los conceptos de emoción (en este caso el del orgullo) en diferentes lenguas (inglés, español, alemán).
Palabras clave: Aprendizaje de lenguas con ayuda de corpus, metáforas conceptuales, lenguaje figurado, conceptos de emoción

1. INTRODUCTION[1]

Figurative speech is one of the most fascinating aspects of language. To the foreign language learner, the idiomatic use of figurative expressions is a challenging and tricky area. Because of the creativity involved in its use and its close relationship to cultural issues, figurative language is a field in which interlinguistic differences abound. This makes it an important source of interference from the learner's mother tongue.

As happens with other aspects of foreign language competence, there are two important complementary strategies for acquiring idiomaticity in this field: Exposure to extensive input

[1] Funding for this work was provided by the research project "Compilación y análisis traductológico de un corpus de textos literarios valencianos no traducidos comparable con COVALT" (Fundació Caixa Castelló – Universitat Jaume I).

(e.g. extensive reading practice) on the one hand, alongside intensive and conscious exploration of the regularities and irregularities of the foreign language, on the other. Over the last decades, there have been promising attempts to use corpora and classroom concordancing as a tool for the latter of the two strategies.

This paper presents a contrastive study of the metaphorical expression of the emotion concept "Stolz"[2] in German, and its Spanish and English equivalents ("orgullo" and "pride") as an example of such a classroom analysis and its possible results. The approach provides learners with the basics of concordancing and co-occurrence analysis as well as with the relevant theoretical knowledge (e.g. conceptual metaphor theory) in order to enable them to carry out their own corpus-based contrastive enquiries about figurative language use and, as a result, to become aware of the similarities and differences in the metaphorical structuring of equivalent emotion concepts in different languages.

2. CORPUS-BASED LANGUAGE PEDAGOGY

Although many corpus-based pedagogical applications are concerned primarily with grammatical topics or with the development of communicative skills, co-occurrence analysis has also proved useful for dealing with lexical and semantic issues. Some of the aspects that have been discussed for direct corpus applications (in the sense of Römer, 2008) in foreign language or translation teaching are:

[2] For the sake of clarity, emotion concepts will be marked by inverted commas ("pride"), lexical units that are used as search words by single inverted commas ('pride'), and co-occurrences from the corpus by italics (*chest*).

- collocation and phrase patterns (Legenhausen, 1997; Partington, 1998; Tognini-Bonelli, 2001),
- synonymy (Legenhausen, 1997; Partington, 1998, 2001),
- true and false friends (Oster, 2009; Partington, 1998; Zanettin, 2001),
- semantic prosody and/or semantic preference (Hoey, 2000; Kenny, 1998, 2001; Oster, 2007; Partington, 1998, 2001; Stewart, 2009),
- delexicalisation (Kenny, 2004; Tognini-Bonelli, 2001),
- phraseology (Coxhead, 2008; Oster and van Lawick, 2008; Wible, 2008)
- idiom and metaphor (Partington, 1998, 2001),
- ideology (Tognini-Bonelli, 2001),
- cultural connotations (Bertaccini and Aston, 2001).

Methodologically, we find a cline that goes from teacher-controlled, data-driven learning (cf. especially Johns and King, 1991; Johns, 1994, 2002; King, 2003; Legenhausen, 1997) to autonomy-oriented approaches like those of Bernardini (e.g. 2000, 2002), Gavioli (2009) or Seidlhofer (2002).

3. A CLASSROOM CO-OCCURRENCE ANALYSIS OF THE EMOTION CONCEPTS "PRIDE", "STOLZ" AND "ORGULLO"

3.1. Setting

The task has been developed for Spanish intermediate learners of German (B1). Another characteristic of the group is that German is a second foreign language after English so that a high intermediate / advanced knowledge of English can be counted on. Although the German language is the main focus of the exercise, carrying out the task in three languages not only enriches the contrastive results, but also offers additional help to the students as categorising is easier in the mother tongue or in a more advanced foreign language.

3.2. Aims

The most important aim of the task is getting acquainted with figurative language (especially metaphorical expressions) used to express emotions (in this case "pride") in the foreign language(s). Furthermore, the contrastive analysis is designed to make students aware of the differences among languages in the conceptualisation of emotions and consequently the

linguistic expressions associated with them. Finally, becoming familiar with the basics of concordancing and co-occurrence analysis is also an aim in itself because this will help them be more autonomous in their learning process, enabling them to carry out their own corpus-based contrastive enquiries about figurative language use.

3.3. Resources

There are considerable differences in size, accessibility, search facilities and other aspects, among the corpora that are available in the three languages concerned. For English, there are several freely accessible very large online corpora. The one selected for this task is the *Corpus of Contemporary American English (COCA)*[3], which contains approximately 400 million words and is especially easy to use. For German, the *DWDS*[4] is employed, a corpus consisting of 120 million words from 20th century publications. In order to obtain Spanish data from a similar time range, the 20th century part of the *Corpus del Español (CDE)*[5] (20 million words) is used. With the English corpus being considerably larger than the German and Spanish ones, a much wider range of expressions can be expected from it. This difference is compensated for by providing the students with the following additional corpus-based resources for German and Spanish:

- The lists of co-occurrences and collocations generated by the corpora *Wortschatz Deutsch*[6] and *Wortschatz Spanisch*[7]
- The information on relevant collocations from the corpus-based dictionaries *DWDS*[8] and *Redes* (Bosque, 2004)
- A random sample of concordance lines from the corpora *Wortschatz Deutsch* and *CREA*[9] (for manual analysis)

3.4. Procedure

In the first place, students are provided with some theoretical tools, which consist in very basic notions of conceptual metaphor theory (as exemplified in the work of Lakoff and

[3] Davies, 2008. Available online at http://www.americancorpus.org/.
[4] Digitales Wörterbuch der deutschen Sprache des 20. Jhs, compiled by the Berlin-Brandenburgische Akademie der Wissenschaften. Available online at http://www.dwds.de/.
[5] Davies, 2002. Available online at http://www.corpusdelespanol.org/.
[6] A corpus compiled by the Universität Leipzig. Available online at http://wortschatz.uni-leipzig.de/.
[7] Available online at http://corpora.informatik.uni-leipzig.de/?dict=es.
[8] Available online at http://www.dwds.de/?woerterbuch=1&qu=Stolz.
[9] Corpus de Referencia del Español Actual, compiled by Real Academia Española. http://corpus.rae.es/creanet.html.

Johnson, 1980), metaphor and emotion (especially the approach proposed by Kövecses, 1986, 2000, etc.) and semantic preference and prosody (e.g. Hunston, 2007; Louw, 1993; Sinclair 1987, 1996). Before and during the task, the learners receive methodological guidance on how to perform a simple search for relevant co-occurrences in an online corpus and how to analyse the results. For the task itself, they are given the following concrete aims:

1. Analyse the co-occurring concepts in searching relevant information on:

a) What conceptual metaphors structure the emotion concepts "pride", "Stolz" and "orgullo"?

b) Can you find evidence for physical or behavioural reactions or side-effects of the emotion?

c) What are the causes of the emotion and who experiences it?

d) How is pride described or evaluated? (Look at adjectives!)

e) What other emotions or attitudes are mentioned alongside "pride", "Stolz" and "orgullo"?

2. Drawing your conclusions:

Students work in small groups, performing complementary searches in the corpora. Each group carries out a number of small tasks, at least one for each language, distributed in a way that ensures every group has to handle different types of data.

Table 1: Example of task distribution

	English data	German data	Spanish data
Group 1	*COCA* nouns (part 1)	*DWDS*	*Redes* dictionary
Group 2	*COCA* verbs	*Wortschatz Deutsch* collocations and co-occurrences	sample of concordance lines *CREA*
Group 3	*COCA* adjectives	*DWDS* dictionary	*CDE*
Group 4	*COCA* nouns (part 2)	sample of concordance lines *Wortschatz Deutsch*	*Wortschatz Spanisch* collocations and co-occurrences

In the first phase, the groups focus on the information from dictionaries, which is easier to process than the other types of data. After sharing their results with the other groups and putting them together, they move on to the more complicated resources: the lists of collocations and co-occurrences provided directly by *Wortschatz Spanisch* and *Wortschatz Deutsch*, the corpus search performed by themselves, and the analysis of sample concordance

lines. Depending on the language proficiency of the learners, their familiarity with (or interest in) linguistic enquiries of this type and the time available, the classroom analysis can be structured and guided to a greater or lesser degree by the teacher. In this case, the learners are provided with tables like the ones in appendices 1 to 5, which they use to write down their results.

4. RESULTS OF THE ANALYSIS

4.1. Conceptual metaphors

In a thorough linguistic study, quantitative data on co-occurrence frequency would make it possible to draw a much more accurate picture of the emotion concept "pride" in the three languages. However, the number of different expressions that can be found for every metaphor type allows us to draw some (more limited) conclusions on the conceptual metaphors that are used to express this emotion (appendix 1 shows the results in detail).[10]

- The metaphor PRIDE IS SOMETHING INSIDE THE BODY is used in a similar way in all three languages. In English there are many examples for the subtype PRIDE IS SOMETHING THAT TENDS TO GO UP AND OUT WHEN IT BECOMES STRONGER (e.g. *swallow, rise, brim with*), whereas in German and Spanish there are more expressions relating to the idea that PRIDE IS SOMETHING THAT MAKES THE BODY SWELL (e.g. *schwellen, platzen, aufgeblasen, henchir, reventar de, no caber en sí)*.
- Within the metaphor PRIDE IS AN ANTAGONIST, English shows a remarkable variety of expressions relating to the subtype PRIDE IS SOMETHING THAT DOMINATES (e.g. *force, dominate, overwhelm, forbid*). In the Spanish corpus, on the other hand, fewer examples are found, but they cover several subtypes not present in the other languages: PRIDE IS AN ATTACKER (*atacar*), PRIDE IS SOMETHING THE SELF FIGHTS BACK AGAINST (*vencer*), and PRIDE IS SOMETHING THAT INFLICTS PAIN (*doler*).
- The metaphor PRIDE IS AN AUTONOMOUS FORCE is not very well developed. "Pride" as a wild animal (*fierce, creep, unbändig, desatado*) is present in all three languages, whereas the conceptualisation of "pride" as FIRE (*spark, fire, burn,*

[10] The classification of metaphor types and subtypes that is used is based on Oster, in press a) and b).

inflamado) has only been found in the English and the Spanish but not the German corpus.

- The conceptualisation of pride as AN AUTONOMOUS BEING, THOUGH STILL PART OF THE PERSON is one of the most important metaphors in this domain. This "person inside a person" can be seen as a passive being that is attacked from the outside (*hurt, bruise*) or is protected against aggression (*salvage, soothe*). In German and Spanish, examples in which "pride" plays an active role (*sich empören, sublevarse*) can also be found.
- With respect to the metaphor PRIDE IS AN OBJECT, one must note that it is much more frequent in English and Spanish than in German, especially as regards the subtype PRIDE IS A PHYSICAL OBJECT (*exhibit, palpable, damaged, battered, bedraggled, crush, mostrar, intacto, dañar*). PRIDE IS A POSSESSION is equally present in all three languages (*give, lose, shed, bewahren, verlieren, nehmen, besitzen, tener, heredar, perder*).

4.2. Physical or behavioural manifestations of pride (conceptual metonymy)

Although there are many coincidences in the physical or behavioural manifestations of pride expressed in the three languages, we can also find noticeable differences when we look at them separately (cf. appendix 2):

- In English the largest group of expressions is that of PRIDE SHOWS IN THE FACE OR IN THE EYES (e.g. *beam, radiate, blush*) and there are many other manifestations, some of which could not be found in the other languages, such as PRIDE CAUSES AGITATION (*twinge, flicker, trepidation*) or PRIDE CAUSES BODY TEMPERATURE TO RISE (*glow, burn, hot*).
- In Spanish there is much less variety, with a strong preference for expressions relating to the swelling of the breast (*e.g. pecho hinchado, henchirse de*).
- This is also the case in German, but there is an additional physical reaction consisting in the adoption of an upright posture (*e.g. erheben, recken, aufrichten*).

4.3. Related concepts: Who experiences pride and why?

The most frequent causes of pride are achievements, especially those related to one's work (*e.g. accomplishment, reputation, ability*) and family members, mainly children. In English there are also frequent references to personal possessions (*home, house, garden*) and heritage shared with others (*traditions, history, culture*). Pride is often experienced by a family member or (in English and in German) by a child or a professional group of some kind

(*student, worker, soldier*). It is also interesting to analyse those cases in which the experiencer and the cause of pride coincide (in what we might call group pride). In all three languages, the largest group is the one related to a country or nation (and more so in German). In English, there are also many references to smaller units (*neighborhood, county*), educational institutions (*school, university*) or society in general (*community, citizen*). Expressions related to race and homosexuality were found in English and Spanish, and to sports teams in German and Spanish (cf. appendix 3).

4.4. Description and evaluation of the emotion

The descriptions that corpus co-occurrences provide of the emotion "pride" are very similar across the three languages (cf. appendix 4). Pride is most frequently described as big or strong, especially in Spanish and English, where we find many adjectives that convey the idea that pride is excessively strong (*exaggerated, inordinate, huge, demasiado, exacerbado, desmedido*). In German, on the other hand, there is a balance between adjectives that describe the emotion as "big" and those that convey it as being "weak" or "small" (*bisschen, gewiss, Anflug von, verhalten*).

In the field of evaluation, there is also a difference between German on the one hand and Spanish and English on the other. While in all three languages "pride" can be perceived both positively and negatively, in Spanish and English we find long lists of negative adjectives used to qualify the emotion (e.g. *perverse, foolish, ridiculous, estúpido, detestable, temerario*), whereas in German the only negative evaluation is *dumm* (stupid). Additionally, there is a hint of sinfulness in the Spanish and English corpora, which is not the case in the German texts (*sin, pardonable, pecaminoso, confesar, pecar de*). By contrast, in German there are more expressions that characterise "Stolz" as justified and a number of adjectives that describe it as childlike (*naiv, kindlich, trotzig*).

4.5. Emotions, attitudes and attributes (the way one feels, behaves or is)

The words 'pride', 'Stolz' and 'orgullo' co-occur frequently with a very wide range of other emotions and attitudes. Looking more closely though, we find some striking differences among the three languages. Whereas in German and English the number of co-occurrences with a positive load is roughly double that of the negative ones, in Spanish there are more negative expressions, including a very large number of negative attitudes and attributes (*vanidad, presunción, soberbia, prepotencia, egoísmo, hipocresía, prurito, petulancia, autosuficiencia, engaño, impiedad, falsa modestia, autoalabanza, codicia, ostentación, altanero, altivo, avaricia*). English and German, on the other hand, seem to relate "pride" and

"Stolz" more strongly to positive feelings (like *joy, hope, enthusiasm*) and positive attitudes (*self-esteem, integrity*). Additionally, in English we find many expressions relating to the concepts of "endurance" (*patience, stoicism, persistence, unwavering*) and "loyalty" (*solidarity, loyalty, commitment*).

5. CONCLUSION

One of the most important benefits of using corpora in the foreign language classroom is the raising of learner awareness with respect to the way language works. The task presented in this paper combines different types of corpus-based data (information from corpus-based dictionaries, lists of collocations and co-occurrences provided by on-line corpora, and corpus searches as well as analyses of sample concordance lines performed by the students themselves). These are used in a progressive manner, to allow the gradual introduction of a complex tool that is not easily handled in a foreign language. Working on the task requires learners to process foreign language data actively and to consciously trace the way an emotion concept is conceptualised and expressed in the language. Apart from training these more general competences, by comparing results in several languages, the learners arrive at concrete contrastive conclusions about the emotion concepts involved.

Regarding the specific case of the metaphorical expression of the concept "pride", a number of points have become clear. In German it is not as conventional as in Spanish to conceptualise the emotion as a physical object. By contrast, in German it is very common to refer to erect posture as a physical manifestation of the emotion "Stolz", which does not seem to be the case in either Spanish or English. On the other hand, the results concerning evaluation and co-occurring emotions and attitudes, if taken together, reveal that Spanish "orgullo" is an emotion judged more negatively than German "Stolz", the former being associated more frequently with concepts like vanity, presumptuousness and arrogance. "Stolz" and "pride", however, seem to be closer to positive emotions and attitudes like self-respect and dignity.

Appendix 1: Linguistic expressions relating to metaphor subtypes in English, German and Spanish

	PRIDE IS	English	German	Spanish
PRIDE IS SOMETHING INSIDE THE BODY	*Something that is unspecifically inside the body*	full	voll, voller	lleno de,
	Something that is located in or affects specific body parts	chest, throat, breast	Brust,	pecho, llenarse la boca
	Something that affects the soul/heart	heart, heartfelt	Herz, Seele	alma,
	Something that comes from the outside	fill, instil(l), inspire, infuse	erfüllen, stolzerfüllt	llenar, colmar de
	Something that is deep inside when it is strong	deep, profound	tief	profundo
	Something that tends to go up and out when it becomes stronger (like a liquid in a container)	swallow, rise, eat one's pride, brim with, well up, surge, a swell, flush,		tragarse, sacar el orgullo,
	Something that makes the body swell	swell, burst, puffed up	schwellen, platzen, aufgeblasen geschwellt, stolzgeschwellt, stolzgeschwollen	henchir, reventar de, henchido de, no caber en sí, no cabe de orgullo,
	Something that emanates from the body and is thus perceptible	exude		
PRIDE IS AN ANTAGONIST	*An attacker*			sucumbir, a ataque de
	Something that dominates	pride forces, grip, dominate, overwhelm, forbid, not permit, prevent from, allow, let, overwhelming, keep from	verbieten, frei von Stolz	no permitir
	A burden	carry		
	Something the self fights back against			vencer
	Something that inflicts pain			doler
PRIDE IS AN AUTONOMOUS FORCE	*A living being in general*		erwachen, wachsen	regresa, creció
	A human being			orgullo ciego
	A plant	blossom		
	A wild animal	fierce, creep	unbändig	fiero, desatado
	Fire	spark, fire, burn, flash		inflamar, inflamado

PRIDE IS AN AUTONOMOUS BEING, THOUGH STILL PART OF THE PERSON	*A passive "person inside"*	blow to one's pride, hurt, wound, salvage, injure, foster, pride suffers, sting one's pride, preserve, salve, soothe, rescue, bruise, dent, intact, bruised	verletzen, verletzt, gekränkt, beleidigt, beleidigen, appellieren an, leidet, vertreiben, Angriff auf unseren	defender, pisotear, humillar, lastimar, herido, herir, lesionar, ultrajar, ofender, mortificar su orgullo, aguijonear, matarte el orgullo, dar puñaladas a, ofendido, desafío a, humillado en su orgullo, pisoteado, orgullo vencido, defender, defensa de,
	An active "person inside"		aufbäumen, sich empören, unbeugsamer	se sublevaba, sublevación, no lo admitiría tu orgullo, le salvó el orgullo
PRIDE IS AN OBJECT	*A physical object*	display, restore, build, put (aside, behind us, etc.), destroy, exhibit, palpable, damaged, battered, bedraggled, crush	zeigen, brechen, ungebrochen	mostrar, intacto, dañar, reponer, derrochar, mancillar, pisotear, pasear, devolver, tocado en su orgullo, depositar, llevar, reducir a escombros, pedazo de orgullo, pisoteado,
	A possession	have pride, give, bring, lose, lost, new(-)found, retain, shed, maintain, keep, abandon, strip of	bewahren, verlieren, Stolz benehmen, nehmen, besitzen	tener, heredar, perder, recuperar, perdido,

Appendix 2: Physical or behavioural manifestations of pride/Stolz/orgullo

	English	**German**	**Spanish**
PRIDE SHOWS IN THE FACE OR IN THE EYES	face, beam, expression, grin, smile, radiate, blush, face alight with pride, eye, look of pride, (eyes) shine, (eyes/smile) flash, shine in one's eyes,	Gesicht, blicken, Auge, leuchten, stolz-strahlend	sonrisa, mirada
PRIDE MAKES THE BODY/BREAST SWELL	swell, burst, puffed up	geschwellt, aufgeblasen, schwellen, brüsten, in die Brust, platzen, stolzgeschwellt, stolzgeschwollen	pecho hinchado, henchirse de, henchido, inflar el pecho, desbordar el pecho,
PRIDE MAKES THE PERSON ADOPT AN UPRIGHT POSTURE	bearing, posture, carry oneself with pride, stiff-necked	erheben, recken, erhoben, empor, aufrichtend, in die Höhe gereckt, hochfahrend, aufrichten	
PRIDE CAUSES AGITATION	twinge, trepidation, flicker, thrill		
PRIDE CAUSES SCREAMING OR CRYING	tear		
PRIDE AFFECTS THE VOICE	voice, thick voice	Stimme	
PRIDE DISTURBS BREATHING	choke		
PRIDE CAUSES BODY	glow, flush, burn, to		

TEMPERATURE TO RISE	warm, hot		
PRIDE INTERFERES WITH PERCEPTION	blinds		

Appendix 3: Syntagmatic relations: Cause, experiencer

	English	German	Spanish
CAUSES	work, accomplishment, achievement, job, ability, project, success, skill, victory, reputation, career, prowess, workmanship, professionalism, profession	Leistung, Erfolg, Werk, Arbeit, Sammlung, Geleistetes, Erreichtes, Entdecker, Sieger	profesional, formación, tecnología, tecnológico
	child, son, daughter, brother, sister, husband, baby	Sohn, Kind, Filius	hijo, hija
	ownership, home, house, garden, owner	besitzen	
	tradition, heritage, history, culture, roots, ancestry	Kunst	
	identity, role, body		
		Name	linaje, casta

EXPERIENCER	family, father, parent, mother, Dad, Mom	mütterlich, Eltern, Mutter, Familie, Vater, väterlich	madre
	child, boy, girl, kid	Kind, Mädchen	
	student, worker, soldier, athlete	Soldat, Bauer, Siedler, Hausfrau	artista, uniforme, militar
	man, guy, woman	Mann, Mensch, Leute	hombre, viril, mujer, vieja
		König	

GROUP PRIDE (EXPERIENCER AND/OR CAUSE OF PRIDE	country, nation, American, yankee, patriotism	national, deutsch, schwarz-weiß-rot, Deutsche, Volk, Nation, Deutschland, Flagge, Vaterland, Reich, Land, Staat, amerikanisch, England,	pueblo , patria , nacional, país , alemán
	school, university, student		
	neighborhood, resident, town, county, hometown, company,	Stadt	ciudad
	community, citizen, society, group,	Bürger, bürgerlich	
	gay, lesbian,		gay, homosexual
	race, black		raza
	team	Mannschaft, Team	club
	church		

Appendix 4: Description and evaluation

Description			
Intensity/Quantity			
	English	**German**	**Spanish**
big	big, greatest, great, exaggerated, inordinate, enormous, huge, immense, considerable, boundless, tremendous	groß, besonders, unbändig, nicht gering, größter	desmedido, exacerbado, exaltado, demasiado, mayor, mucho, gran, máximo grande, tan, enorme, especial,
strong	intense, strong,	hoch, höher, höchst	fuerte,
large quantity	a lot	viel	
weak, small	bit, certain, small, little, modest	bisschen, gewiss, Anflug von, verhalten, leise	cierto, un poco de, un poquito de
Quality:			
pure	real, genuine, pure, true, hard		puro,
visible		sichtlich, sichtbar, hörbar, offensichtlich, erkennbar, unverkennbar, unüberhörbar, spürbar	claro,
***Evaluation*:**			
justified	understandable, justified, justifiable, well-earned	berechtigt, gerecht, ehrlich, natürlich, zu Recht, mit Recht, mit Fug und Recht	justificado, legítimo, justo, santo
unjustified		falsch	injustificado, falso,
negative	perverse, false, stubborn, foolish, overweening, misplaced, smug, undue, ridiculous, stupid, misguided, dumb, childish, absurd	dumm	estúpido, altanero, altivo, detestable, imbécil, malsano, masoquista, tonto, temerario, sanguinario
childlike		naiv, kindlich, trotzig,	
positive	endearing	edel, ehrlich	noble
shameful	hide, conceal, unabashed, secret,	heimlich, verbergen, unverhohlener, verhohlener, verhehlen, klammheimlich, unterschwellig,	indisimulado, disimular, ocultar, esconder, disfrazar, dignificar, disimulación,
sinful	sin, pardonable,		pecaminoso, confesar, pecar de, pecado

Appendix 5: Paradigmatic relations: Emotions, attitudes and attributes co-occurring with ‘pride’, ‘Stolz’ and ‘orgullo’

		English	**German**	**Spanish**
Negative emotions, attitudes and attributes				
	fear	fear, anxiety	Furcht	miedo, temor, ansiedad
	hate	hatred		
	anger	anger, wrath, resentment, angry, bitter, prickly, outraged	Zorn, Wut	cólera, rabia, furia

	neg. emotions oriented towards oneself	shame, embarrassment, humiliation	Scham	culpa, bochorno,
	sadness	sorrow, regret, nostalgia,	Trauer	
	pain	pain	Schmerz	dolor, tormento
	neg. emotions oriented towards others	envy, jealousy, gleeful, jealous,	Neid, Verachtung,	
	negative attitudes	prejudice, arrogance, ego, greed, vanity, ambition, hubris, self-sufficiency, selfish, haughty, arrogant	Habsucht, Arroganz, Egoismus, kalt, Eitelkeit, Ehrgeiz, Hochmut, Überheblichkeit	vanidad, presunción, soberbia, prepotencia, egoísmo, hipocresía, prurito, petulancia, autosuficiencia, engaño, impiedad, falsa modestia, autoalabanza, codicia, ostentación, altanero, altivo, avaricia, autosuficiencia,
	neg. attributes	stubbornness	Dummheit	
Positive emotions	*positive emotions oriented towards others*	love, passion, affection, gratitude, respect, loyalty, friendship, admiration, mercy, sympathy, devotion	Dankbarkeit, Liebe, Vertrauen, Sehnsucht, Leidenschaft	amor, piedad, aprecio, pasión, confianza,
	positive feelings as a reaction to good things in the present or future	joy, pleasure, hope, satisfaction, excitement, relief, delight, happiness, optimism, anticipation, buoyant, enthusiasm	Freude, freudig, Glück, Genugtuung, Hoffnung, Befriedigung, Rührung, Zufriedenheit, Erleichterung, Zuversicht	satisfacción, alegría, esperanza, felicidad,
	positive attitudes or attributes	confidence, self-respect, self-esteem, self-knowledge, self-reliance	Selbstbewusstsein, Selbstachtung, Selbstgefühl	confianza en uno mismo, amor propio
		dignity, honesty, integrity, honor, values	Würde, ehrlich, Ehre, Anstand, Ehrgefühl	dignidad, honor, caballerosidad
		humility, humble	Bescheidenheit, Demut, bescheiden	modestia, humildad
		courage, bravery	Mut	coraje, valor, garra, hombría y pundonor
		power, glory	Ruhm, Macht	poder
		determination, will, faith, strength, sincere	Kraft, Wille	
		enduring, discipline, independence, patience, stoicism, persistence, endurance, unwavering		
		solidarity, generosity, loyalty, commitment, responsibility		
			Ernst, ruhig, Sicherheit	sobrio
Ambivalent or neutral emotions		lust, desire, awe, defiant, defiance, bravado, shy	trotzig, Verletzlichkeit, Wehmut,	deseo

REFERENCES

Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard, & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective. Papers from the Third International Conference on Teaching and Language Corpora*, (pp. 225-234). Frankfurt a. M.: Peter Lang.

Bernardini, S. (2002). Exploring New Directions for Discovery Learning. In B. Kettemann, & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Analysis*, (pp. 165-182). Amsterdam/New York: Rodopi.

Bertaccini, F., & Aston, G. (2001). Going to the Clochemerle: exploring cultural connotations. In A. Partington (Ed.), *Learning with corpora*, (pp. 198-219). Houston: Athelstan.

Bosque Muñoz, I. d. (2004). *Redes: Diccionario combinatorio del español contemporáneo*. Madrid: Ediciones SM.

Coxhead, A. (2008). Phraseology and English for Academic Purposes: Challenges and Opportunities. In F. Meunier, & S. Granger (Eds.), *Phraseology in language learning and teaching,* (pp. 149-161). Amsterdam: John Benjamins.

Davies, M. (2002-). *Corpus del Español* (100 million words, 1200s-1900s). Available online at http://www.corpusdelespanol.org.

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA)*. Available online at: http://www.americancorpus.org/.

Gavioli, L. (2009). Corpus Analysis and the Achievement of Learner Autonomy in Interaction. In L. Lombardo (Ed.), *Using Corpora to Learn about Language and Discourse,* (pp. 39-72). Bern: Peter Lang.

Hoey, M. (2000). A world beyond collocation: new perspectives on vocabulary teaching. In M. Lewis (Ed.), *Teaching Collocation. Further Developments in the Lexical Approach,* (pp. 224-243). Hove: LTP.

Hunston, S. (2007). Semantic prosody revisited. *International Journal of Corpus Linguistics*. 12(2), 249-268.

Johns, T. (1994). From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-driven Learning. In T. Odlin (Ed.), *Perspectives on Pedagogic Grammar,* (pp. 293-313). Cambridge: CUP.

Johns, T., & King, P., Eds. 1991. Classroom Concordancing. (*English Language Research Journal* 4). Birmingham: Birmingham University.

Kenny, D. (1998). Creatures of habit? What translators usually do with words. *Meta*. XLIII(4), 515-523.

Kenny, D. (2001). *Lexis and creativity in translation. A corpus-based study*. Manchester: St. Jerome.

Kenny, D. (2004). Die Übersetzung von usuellen und nicht usuellen Wortverbindungen vom Deutschen ins Englische. Eine korpusgestützte Analyse. In K. Steyer (Ed.), *Wortverbindungen - mehr oder weniger fest*, (pp. 335-347). Berlin / New York: de Gruyter.

King, P. (2003). Parallel concordancing and its applications. In S. Granger, J. Lerot, & S. Petch-Tyson (Eds.), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, (pp. 157-167). Amsterdam/New York: Rodopi.

Kövecses, Z. (1986). *Metaphors of Anger, Pride, and Love: A Lexical Approach to the Structure of Concepts.* Amsterdam: John Benjamins.

Kövecses, Z. (2000). *Metaphor and Emotion: Language, Culture, and Body in Human Feeling*. Cambridge: Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Legenhausen, L. (1997). Grammatikunterricht einmal anders. Das Arbeiten mit Konkordanzen im Fremdsprachenunterricht. In D. Kranz (Ed.), *Multimedia Internet Lernsoftware,* (pp. 37-44). Münster: agenda.

Louw, B. (1993). Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair*, (pp. 240-251). Amsterdam, Philadelphia: Benjamins.

Oster, U. (2009). *La adquisición de vocabulario en una lengua extranjera: De la teoría a la aplicación didáctica.* Porta Linguarum. 11, 33-50.

Oster, U. (in press a). "Angst" and "fear" in contrast: A corpus-based analysis of emotion concepts. Cognitive Linguistics between Universality and Variation, Dubrovnik.

Oster, U. (in press b). El análisis de corpus como herramienta de la enseñanza de la traducción: Metáforas conceptuales y emociones. In M. Emsel, & A. Endruschat (Eds.), *La metáfora en la traducción*: Martin Meidenbauer.

Oster, U., & van Lawick, H. (2008). Semantic preference and semantic prosody: A corpus-based analysis of translation-relevant aspects of the meaning of phraseological units. In M. Thelen, & B. Lewandowska-Tomaszczyk (Eds.), *Translation and Meaning.* Part

8, (pp. 333-344). Maastricht: Hogeschool Zuyd, Maastricht School of Translation and Interpreting.

Partington, A. (1998). *Using Corpora for English Language Research and Teaching.* Amsterdam/Philadelphia: Benjamins.

Partington, A. (2001). Corpus-based description in teaching and learning. In A. Partington (Ed.), *Learning with Corpora,* (pp. 63-84). Houston: Athelstan.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (volume 1), (pp. 112-130). Berlin: Mouton de Gruyter.

Seidlhofer, B. (2002). Pedagogy and local learner corpora. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, (pp. 213-234). Amsterdam/Philadelphia: John Benjamins.

Sinclair, J. (1987). Collocation: a progress report. In R. Steele, & T. Threadgold (Eds.), *Language Topics. Essays in Honour of Michael Halliday*, (pp. 319-332). Amsterdam / Philadelphia: Benjamins.

Sinclair, J. (1996). The Search for Units of Meaning. *TEXTUS*. IX(1), 75-106.

Stewart, D. (2009). Safeguarding the lexicogrammatical environment: Translating semantic prosody. In A. Beeby, P. I. Rodríguez, & P. Sánchez-Gijón (Eds.), *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate,* (pp. 29-46). Amsterdam: Benjamins.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam: John Benjamins.

Wible, D. (2008). Multiword expressions and the digital turn. In F. Meunier, & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching*, (pp. 163-181). Amsterdam/Philadelphia: Benjamins.

Zanettin, F. (2001). Swimming in words. Corpora, translation and language. In A. Partington (Ed.), *Learning with corpora*, (pp. 177-197). Houston: Athelstan.

Variations in the use of "I" in casually spoken English

MICHAEL PACE-SIGGE

University of Liverpool

Abstract

Although the most frequently occurring word of reference is I, and I is also the most frequent item occurring overall in spoken corpora, little research has been undertaken in how it occurs. While a lot of corpus linguistic research has been looking at I in written academic texts, this paper will focus on the general patterns in which I usage can be found in a selection of casual spoken English corpora (including the BNC, Bloomsbury-Macmillan subcorpora, WSC and specialist corpora of spoken English varieties) and compare its occurrence pattern with a corpus of radio-transcripts (prepared speech). High frequency of an item means that it can be found in a number of different corpora, and it also means that a good insight into its uses can be obtained even from smaller corpora.

The use of personal pronouns (I, you etc.) is necessarily entwined with cultural practice. (Hanks: 1990) Hanks makes the point that the "schematic knowledge (...) that practise presupposes, is also produced in the practise". In other words, speakers not only know when to use the item I, but also, by using it in a particular context (but less so in another context) speakers define the meaning of the item. In the context of language use, this seems to be congruent with the theory of lexical priming. Words like I or you do not exist outside the social context. Comparing the use of I in spoken to written contexts, furthermore, assists teachers of English to demonstrate appropriate use.

Beyond occurrence patterns which can be observed to be broadly similar in all corpora, the collocations and nestings of I can vary in different corpora however. There will be uses that are specific to a group of users, reflecting cultural conventions or preferences. Where these uses are found to be employed not just by individuals (where they would have to be classed as features of idiolects) but by a number of people in a speech community, a group-specific pattern surfaces. This paper aims to describe these observations and provide an explanation for this occurrence of use.

Keywords: Spoken Corpora, Lexical Priming, Collocation, Colligation, Self-reference

Resumen

Aunque "yo" es la palabra de referencia más frecuente y es también el ítem de mayor presencia en los córpora hablados, pocas investigaciones se han llevado a cabo en relación a cómo se da este hecho. Mientras que gran parte de la investigación en lingüística corpus se ha centrado en los textos escritos académicos, este ponencia lo hará en los patrones generales que se dan en una selección de inglés hablado (incluido el BNC, el Bloomsbury-Macmillan, WSC y córpora especializados de variedades del inglés hablado), y comparará el patrón de ocurrencia con un corpus de transcripciones radiofónicas (discurso preparado). La alta frecuencia de un ítem cualquiera significa que se puede encontrar en un número de córpora diferentes; también supone que se puede obtener incluso de córpora pequeños conclusiones relevantes sobre sus usos.

El uso de los pronombres personales (yo, tú, etc.) está íntimamente relacionado con la práctica cultural (Hanks 1990). Hanks señala que "el conocimiento esquemático (...) que presupone la práctica, se produce en la práctica". ("schematic knowledge (...) that practise presupposes, is also produced in the practise"). En otras palabras, los hablantes no sólo saben cuándo usar el "yo"; sin que también, mediante su uso en un contexto particular (pero no tanto en otro), éstos definen el significado del mismo pronombre. En el contexto del uso del lenguaje, esto parece ser congruente con la teoría del priming léxico. Palabras como "yo" o "tú" no existen fuera del contexto social. La comparación del uso de "yo" en textos escritos y hablados, por otra parte, ayuda a los profesores de inglés a demostrar cuál es el uso más apropiado.

Más allá de los patrones de ocurrencia que se observa que son muy parecidos en todos los córpora, las colocaciones y los anidamientos de "yo" varían en éstos, sin embargo. Habrá usos que son específicos de un grupo de usuarios, lo que refleja las convenciones culturales o sus preferencias. Cuando se haya que estos usos los emplean no sólo los individuos (en lo que tendría que clasificarse como características de idiolectos diversos), sino también un número de personas en una comunidad de hablantes, distinguimos un patrón específico de un grupo. Este trabajo tiene como objetivo describir estas observaciones y proporcionar una explicación para este hecho.

Palabras clave: Corpus hablados, priming léxico, colocación, coligación, autorreferencia

1. INTRODUCTION

> Deictic reference is a communicative practise based on a figure-ground structure joining a socially defined indexical ground, emergent in the process of interaction, and a referential focus articulated through culturally constituted schematic knowledge. The horizon of schematic knowledge (…) that practise presupposes , is also produced in the practise. (Hanks, 1990: 515)[1]

The use personal pronouns (*I, you* etc.), is for Hanks (who writes with reference to the language of the Maya) necessarily entwined with cultural practice. The interesting point here is that the "schematic knowledge (…) that practise presupposes, is also produced in the practise". This can be read as knowledge gained through practice. In the context of language use, this seems to link to the propositions of lexical priming.

There appears to be not as much research on the first person singular pronoun available as might be expected. There is wide-spread reference to the *academic I* (or the lack of it). More literature on the first person singular use appears to be in psychological and cultural research than in language studies:

> A conception of a person is also coded in the use of person-indexing pronouns, or *deixis*, such as "I" and "you" in English. Deixis are used to indicate extralinguistic entities in discourse (sic): Personal deictic pronouns index the speaker and the addressee within the specific social context. (…)
> Hanks[2] argued that deictic systems evolve, to a large extent, through culturally specific, situated practices. Specific uses of personal deixis in everyday discourse may require users to pay close attention to (…) personal relationships. (Kashima & Kashima, 1998: 464)

Words like *I* or *You* therefore do not exist outside the social context, meaning they tend to be found in less abstract texts such as casual conversation. The reference to Hanks is of particular interest in the context of this thesis: "culturally specific, situated practices" are, after all, what human beings, in the course of their socialization, are primed to follow. As this paper looks at priming in spoken language, the highest occurring deictic, *I*, can be expected to reveal culturally specific usage.

---

[1] American English spelling used in the original.

Indeed, Fasulo and Zucchermaglio (2002) claim, based on a sample taken from 10 informants, that utterances with *I* have four discursive functions:

> Four basic classes were identified on the basis of their semantic and pragmatic meaning: *Epistemics, Decisionals, Operatives,* and *Impersonals*. (…)*Epistemic* IMU [I-marked utterances] refers to the speaker's state of knowledge. The range of Epistemics found in the corpus include parentheticals … probability such as *I think,* parentheticals of necessity (mostly of the negative form, such as *I am not convinced*), verbs of perception used in a metaphorical fashion such as *I see*, references to cognitive states such as *I remember*, and expressions of one's inclination for a certain possible line of action, such as *I am in favor* or *I agree*. (…)
> *Decisional* utterances are those in which the speaker defines his stance toward a given line of action by proposing it to the interlocutors or committing himself to it. … These are modals such as *I shall, I can, I want, I say, I go* (…)
> *Operatives* … are utterances directly concerned with practical operations; they can be reports of things done, in the past tense, of simple announcements of next actions, in the present tense. E.g. *I came here, I begin to* (...)
> *Impersonal IMUs* are those where the agent is not the speaker, but a generic person doing the action in question. E.g. *If I click, When I'm doing* ... (Fasulo & Zucchermaglio, 2002: 1125ff.)[3]

Fasulo and Zucchermaglio also note that there is also strong use for *I* as the first word when interrupting a speaker (they refer to them as "cutoffs"). This might be an area worthy of further investigation. This paper, due to constraints of space, will give a first overview how *"I"* occurs in a selection of different spoken corpora and focus on a few specific features rather than providing a full analysis for the occurrence patterns found in each one of these corpora.

1.1. "I" – making the implicit explicit[4]

"I" in almost every set of spoken utterances plays an important role and can be found in about every corpus of spoken English as one of the three highest-occurring words:

> Conversation is interactive as a form of personal communication. It is not surprising, then, that conversation shows a frequent use of the first-person *I* and *we* and the second-person pronoun *you*. (Biber *et al.,* 2002: 5)

As such, the pronoun is a potentially valuable pointer to differences of use between speech communities. If *I* is highly frequent, it does not automatically follow that its nearest collocates and clusters are similar in their frequency in any two corpora.

---

[2] See Hanks, W. F. (1990: 514)

[3] For their Italian speakers, the authors found that *operative I* is the most commonly used (over 1/3 of all occurrences). As Italian is a pro-drop language, which means one can drop the subject ("I" included). This happens especially in spoken Italian. –Thanks to Pierfranca Forchini for clarifying this point.

[4] This title refers to the following statement: An advantage of the original first person singular present indicative active form (…) is that this implicit feature of the speech-situation is made explicit. – (Austin, J.L., 1962 [2001]: 61).

Table 1 below shows the distribution of *I* in the various corpora:

Table 1: *I* use in different spoken corpora

Item	Relation	Total "I"	Total Corpus (Tokens)
I (MAC)	1.13%	37,127	3,300,000
I (BNC/C)	3.28%	132,397	4,022,428
I (COLT)	2.90%	14,868	511,834
I (LTT)	2.60%	323	12,406
I (SCO)	2.26%	2,693	119,079
I (WSC)	2.27%	27,691	1,218,957
I (SEC rec.)	2.42%	1,343	55,561

Table 1 also highlights the importance of *I* in spoken English. The *Relation* column shows the relative frequency of the target word *"I"* within the whole corpus[5]. MAC (Macmillian Casual Spoken English) and BNC/C (Conversation subcorpus of the BNC) are used as a standard of *I* use found across Britain. This is compared to a number of specialist corpora which are based on regional variants: the *Bergen Corpus Of London Teenage Language* (COLT) and the smaller and more recent *London Teenage Talk* (LTT). SCO accounts for speakers from Liverpool, WSC is the *Wellington Spoken Corpus* based on recordings of New Zealand English. Finally, the SEC is based on transcripts of radio recordings, *i.e.* is non-casual spoken English. The SEC therefore provides a record of prepared and or / more careful speech. Overall, *I* occurs with proportionally similar frequencies in all corpora.

[5] All these frequencies are based on percentages calculated in Mike Scott's *Wordsmith 4.0* wordlists.

2. *"I"* COLLOCATES

Looking at collocates within 5 words either to the left or the right of *I* (in Tables 2-4[6]), we find that *I* collocates with a wide variety of words. It tends to be found mostly in short (2w or 3w) clusters while longer clusters are found to be of relatively low frequency.

Table 2: 10 most frequent collocates of *I* in MAC and BNC/C

MAC Col.	%	total	BNC/C Col.	%	total
IT	21.0	7,627	TO	18.1	23,984
YOU	18.0	6,684	IT	17.5	23,231
AND	15.8	5,860	THE	17.4	23,099
TO	14.1	5,224	AND	17.3	22,895
KNOW	13.5	4,987	YOU	17.1	22,660
THE	13.2	4,882	KNOW	15.4	20,386
THAT	12.3	4,578	A	13.9	18,389
DON'T	11.9	4,399	DON'T	12.9	17,130
THINK	11.4	4,215	THAT	12.5	16,488
A	10.9	4,042	THINK	12.3	16.262

Table 2 compares the MAC, which contains material recorded up to 2001 with the BNC/C which has as its most recent material recordings that are about 10 years older. Though there are slight differences in the percentages of occurrences and in the ranking, all top ten *I* collocates can be found in both corpora, indicating a high degree of congruence.

[6] Collocate calculations were done with *Wordsmith 4.0*. Near collocates of *"I"* will be discussed as *2w clusters* below.

Table 3: 10 most frequent collocates of *I* in two London Teen corpora

COLT Word	Total	LTT Word	Total
KNOW	2870	LIKE	89
YOU	2597	WAS	83
DON'T	2224	AND	79
IT	2154	DON'T	53
AND	2113	KNOW	51
TO	2017	TO	43
THE	1785	THE	38
WAS	1561	A	34
YEAH	1527	THINK	33
THAT	1498	JUST	29

The COLT and LTT corpora use recordings about 30 years apart. Furthermore, COLT consists of roughly half-a-million words of spontaneous conversations between 13- to 17-year old boys and girls from socially different school districts[7], while LTT was recorded at Dartford Grammar School for Girls and Gravesend Grammar School for Boys in England in 2008[8]. Even without going into further detail, Table 3 already indicates that the most common collocates for those two groups of London teenagers differ to a large degree. While *don't, to* and *the* co-occur with *I* within the same ranking, among the top ten collocates we find four collocates in each corpus that are not among the ten most frequent *I* collocates in the other. Apart form the fact that COLT is meant to reflect a general, mixed group and LTT a specific, limited group, it may also seen to reflect both the distance of time and the fact that teenage talk is not fixed and subject to strong change over time (cf. Ito & Tagliamonte, 2003). Yet, no final conclusion can be drawn based on such different groups of speakers in a linguistically diverse city like London.

[7] More details under http://www.hf.uib.no/i/Engelsk/colt/COLTinfo.html. Last accessed 28/I/2010.
[8] More details under http://www.uni-saarland.de/fak4/norrick/scose.html. Last accessed 28/I/2010.

Table 4: 10 most frequent collocates of I occurring in Liverpool (SCO), New Zealand (WSC) and in radio interviews (SEC).

SCO Word	Total	WSC Word	Total	SEC Word	Total
KNOW	480	AND	6928	THE	331
AND	478	THE	6374	IN	242
TO	453	TO	5674	TO	238
THE	451	IT	4917	AND	230
IT	444	THAT	4830	WAS	187
YOU	405	WAS	4527	A	177
A	381	KNOW	4347	YOU	149
THAT	344	THINK	4327	THINK	142
WAS	312	YEAH	4280	OF	134
DON'T	308	YOU	4276	THAT	124

The three corpora in Table 4, by contrast, reflect *I* collocates that are closer to those found in MAC and BNC/C. There are, however, *I* collocates that hint at use that is specific to one place: *Know* is only the highest occurring collocate in SCO and COLT; likewise, *yeah* only occurs in COLT and WSC, while *don't* is not amongst the ten most frequent collocates in either the Wellington corpus or in the radio interview transcripts. Indeed, the radio transcripts are characterised by a high amount of prepositions as *I* collocates. While *to* (which can function, yet must not necessarily be a preposition) is commonly used in all corpora, *in* and *of* occur only with *I* in SEC.

Overall we may be surprised to find *the* being a frequent collocate of *I* in all corpora. Though widely spread, it is mostly occurrent in R3 position, and then mostly in connection with the function word *when*: *when **I** lived in **the** States*, *when **I** talked to **the** baby* (when+*I*+verb+preposition+*the*).

3. "I" 2-WORD CLUSTERS

3.1. Areas of convergence

O'Keefe *et al.* (2007) give a valuable overview of the top 20 two-word chunks of their 5-million-word CANCODE spoken corpus, and I give here an excerpt:

Table 5: Chunks with *"I"* amongst CANCODE top 20 2w chunks (top 5 "I" 2w clusters).

rank	item	frequency
2	***I*** mean	17,158
3	***I*** think	14,048
6	***I*** don't	11,975
8	and ***I***	9,722
11	***I*** was	8,174

This gives a good indication what to look for in the other corpora.

Table 6: Most frequent 2w clusters (chunks) with *I* in SCO, MAC and BNC/C.

rank	SCO	freq.	MAC	freq.	BNC/C	freq.
1	I DON'T	282	I DON'T	3,811	I DON'T	15,982
2	I MEAN	249	I MEAN	3,663	I MEAN	15,258
3	AND I	225	I THINK	3,326	I THINK	14,228
4	I WAS	205	I , I	3,302	AND I	10,704
5	I THINK	197	I KNOW	2,728	I I*	9,846
6	I KNOW	148	(* no comma in BNC/C concordance)			
6	I-I	148				

Tables 5 and 6 show that *I* most frequent 2w clusters are found both in SCO and CANCODE, and the degree of convergence with MAC and BNC/C is also very high. This means, these 4 corpora highlight that *I* use found in 2w clusters is mainly similar and shows only a very low degree of user-specific variation.

3.2. "I" 2w clusters: areas of divergent use

Table 7: Top 5 2w clusters in COLT, LTT, WSC and SEC.

COLT	Freq.	LTT	Freq.	WSC	Freq.	SEC	Freq.
I DON'T	1995	I WAS	75	I THINK	3348	I WAS	113
I KNOW	1578	AND I	54	I DON'T	3185	I THINK	81
I MEAN	1157	I DON'T	43	I MEAN	2958	I SUPPOSE	62
I WAS	1021	I THINK	26	AND I	2900	I DON'T	60
YEAH I	996	I KNOW	25	I I	2493	AND I	60

Comparing Table 7 with Tables 5 and 6 we see that there is a strong agreement between all the corpora. *I don't* appears in all of them, *I think* in seven and *I mean* or *and I* in six out of eight corpora. Even the marker of hesitancy – repeat *I* – is accounted for in more half of the corpora's five most frequent 2w clusters.

There are, however, important differences to be found. *I know* appears in MAC, SCO, LTT and COLT. These are all, with the exception of COLT, fairly recent British English corpora. The absence of *I know* in BNC/C, which has earlier recording dates, while it is found in COLT which reflects young people's speech, leads to the assumption that this 2w cluster has been used more strongly amongst younger people for some time now.

COLT also stands out as the only corpus to have *Yeah I* as a high-frequency 2w *I* cluster. (The closest similarity is WSC, where *Yeah I* is used as the eighth-most frequent 2w *I* cluster). Similarly, only SEC uses *I suppose* as a frequent 2w *I* cluster. *I suppose* can be seen as a marker of more careful, less casual wording – something to be found in a radio interview, yet less so in casual conversation.

4. A TEST CASE: LONG CLUSTERS WITH THE NEGATIONS *I'M NOT* AND *I CAN'T* COMPARED IN MAC AND SCO.

In this section we explore how far the usage of **I'm not** and **I can't** differs in the two corpora.

Table 8: Occurrence distribution of *I can't* and *I'm not* amongst *I* use in SCO and MAC

item	SCO tot.	SCO %	MAC tot.	MAC %
I CAN'T	51	1.6	902	2.4
I'M NOT	78	2.4	867	2.2

Table 8 above shows the proportional frequencies of the 2w clusters *I can't* and *I'm not* are similar both in relation to each other and in the two corpora.

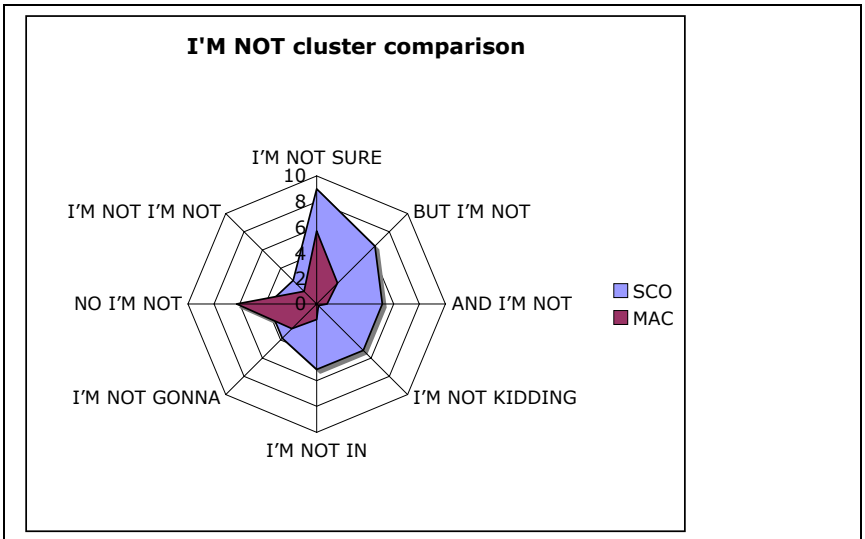


Figure 4: SCO vs MAC differences of use in I'M NOT clusters made visible

Yet when we look at longer, 3-word, clusters, shown in Figure 4, we find that there is only one cluster where *I'm not* is incorporated with about the same proportional frequency: *I'm not gonna*.

However, the 3w cluster *no I'm not* (a very finite statement) is the only one of the clusters incorporating the 2w cluster *I'm not* that is used markedly more often in MAC than SCO. The respective proportional figures are 6.2% in MAC compared to 3.8% in SCO. Other *I'm not* clusters found in MAC are *well I'm not* (45 occ.) and *I'm not going* (44 occ.), where the former is not recorded in SCO and the latter appears only twice in SCO.

Conversely, the hedge *I'm not sure* is noticeably more widely employed in SCO (9.0%) than in MAC (5.7%). However, *I'm not sure* is nearly as frequent in BNC/C (8.3%) as in SCO. This is one of the rare occasions where MAC is the outlier.

All of the listed clusters bar *no I'm not* are more frequent in SCO, but the clusters *and I'm not* and *I'm not kidding* in particular barely occur in the larger MAC. *And I'm not* is, proportionally, over six times more frequent in SCO than MAC. More striking still is the occurrence of the cluster *I'm not kidding*, which appears proportionally 25 times more often in SCO than in MAC. *I'm not kidding* appears to be likely to be a SCO-specific phrase. This indicates how a particular form of negation with *I* may have a different field of semantic association for SCO speakers when compared to MAC speakers.

5. LONGEST AVAILABLE CLUSTERS

5.1. 3w "I" clusters

Table 9 below shows that *I don't know* can be seen to be the most common longer *I* cluster, yet exceptions can still be found in specific corpora.

LTT, for one, shows the use of *I was like* (47 occurrences) is being used $2^{1/2}$ times more often than *I don't know*. I will come back to this in section 5.4

Table 9: *I* 3w-4w clusters compared in COLT, WSC and SEC.

MAC	**Occ.**	**COLT**	**Occ.**	**WSC**	**Occ.**	**SEC**	**Occ.**
I DON'T KNOW	1412	I DON'T KNOW	674	I DON'T KNOW	1571	I THINK THE	22
I MEAN I	1238	YEAH I KNOW	324	I DON'T THINK	459	I WANT TO	22
I I I	946	I DON'T THINK	240	YOU KNOW I	357	I DON'T THINK	18
I THINK I	742	NO I DON'T	160	DON'T KNOW I	333	WHEN I WAS	13
I KNOW I	614	YOU KNOW I	160	I DON'T KNOW I	319	I DON'T KNOW	12

Similarly, people tend to avoid saying *I don't know* during radio interviews, too: In SEC, *I don't know* appears 12 times, the less specific, less certain *I don't think*[9] 18 times, while the most-used 3w *I* cluster in SEC is *I think the* (22 times).

Furthermore, Table 9 highlights another fact: While there is strong agreement in the clusters appearing in MAC, COLT and WSC, the SEC corpus records some very different clusters. Rather than showing opinion (*I mean I, I don't think*), we find chunks that appear to belong to more elaborate phrases like *I want to* and *when I was*[10].

5.2. Differences in use: The case of I DON'T REALLY

I don't really is an interesting case of an *I* 3w cluster when compared in a number of corpora. It should simply be a medium-high frequent phrase with *I* in any conversation. Looking at Table 11, however, divergence of use becomes clear:

[9] Amongst the corpora, *I don't think* occurs nearly as often as *I don't know*

[10] The cluster *I want to get through* appears 6 times in SEC, but is rare in the, far larger, casual English MAC.

Table 10: Occurrence of *I don't really* in 7 corpora. (% of total *I* occ.)

Corpus	Freq.	%
BNC/C	213	0.16
MAC	21	0.06
COLT	30	0.20
LTT	5	1.55
SCO	6	0.20
WSC	90	0.33
SEC	0	0.0

First of all, Table 10 appears to show that *I don't really* is a feature of casual speech, as it appears not at all in SEC. Apart from that, it appears around the 0.2% to 0.3% mark of all uses of *I* in most corpora. Looking at the usage more closely, however, differences become clearer. MAC (0.06% of all uses of *I*) and LTT (1.55% of all uses of *I*) are both outliers, marking opposite extremes. While the string of words is marginal in its use in MAC, it appears with strong preference of use in LTT. Apart from this, *I don't really* appears with proportionally more frequent use amongst New Zealand speakers compared to BNC/C informants, Scousers and COLT London Teenagers.

5.3. Longest meaningful clusters

Observations similar to what we have seen with *I don't really* can be made we look at the longest available clusters available in the corpora.

Table 11: Top 3 occurring *I* 5w cluster in BNC/C, MAC, COLT, WSC and SEC.

Corpus	5w cluster	Freq.	%
BNC/C	YOU KNOW WHAT I MEAN	326	0.25
	DO YOU KNOW WHAT I	101	0.08
	YOU SEE WHAT I MEAN	45	0.03
MAC	YOU KNOW WHAT I MEAN	153	0.43
	I DON T KNOW I DON'T	100	0.27
	MEAN I SOMETIMES I MEAN	64	0.17
COLT	YOU KNOW WHAT I MEAN	81	0.54
	I KNOW I KNOW I	42	0.28
	KNOW I KNOW I KNOW	33	0.22
WSC	I DON'T KNOW I DON'T	83	0.3
	DON'T KNOW I DON'T KNOW	48	0.17
	YOU KNOW WHAT I MEAN	45	0.16
SEC	I WANT TO GET THROUGH I	6	0.45
	I KNOW I THINK YOU'D	6	0.45
	EVEN WORSE WHILE I WAS	6	0.45

Table 11 shows that long clusters with I are rare. It demonstrates that SEC, displaying features of prepared speech, has not a single cluster that is found in the casually spoken corpora. Amongst the latter, You know what I mean appears to the most common meaningful cluster, while other longer clusters seem to incorporate I know in all corpora bar SEC. Yet even the use of You know what I mean differs widely in their respective proportional frequencies. In Table 11, it is most often found in COLT (0.54% of all uses of I) and least often in the Wellington Corpus (0.16%). Comparing the WSC to the other casual spoken corpora in Table 11, this may indicate a geographical difference, where You know what I mean is less used in New Zealand than in Britain.

5.4. Long clusters incorporating frequent constituent parts

One important issue that must also be discussed is that quite often we find that highly frequent 2w and/or 3w clusters are constituent parts of highly frequent 5w/6w clusters. This has been demonstrated by O'Donnell (2009). There is no space to investigate the use of the shorter clusters where they are not part of longer clusters in detail. However, looking at the LTT and SCO clusters below, we can see how individual parts are sections of longer clusters that are most frequent.

Table 12: appearance of constituent parts in longest meaningful clusters in LTT and SCO.

LTT	Frq. / %	SCO	Frq. / %
I WAS LIKE	47 / 14.6	WHAT I MEAN	61 / 2.4
AND I WAS	31	KNOW WHAT I	51
AND I WAS LIKE	27	YOU KNOW WHAT	49
WAS LIKE OH	6	YOU KNOW WHAT I	48
LIKE OH MY GOD	5	KNOW WHAT I MEAN	47
AND I WAS LIKE OH MY GOD	5 / 1.5	YOU KNOW WHAT I MEAN	46 / 1.7

And I was like oh my god is the most frequent long cluster in LTT, while *You know what I mean* is the longest such cluster in SCO[11].

Table 12 shows constituent parts of the longer clusters that are, in themselves, fairly frequent. Both long phrases are used with about the same proportional frequency of all uses of *I*. Although structurally similar, we can see that in SCO the 3w clusters *what I mean* etc. are nearly always part of the 5w cluster *You know what I mean* while *I was like* appears ten times more often than *I was like oh my god*. It is possible, therefore, that *You know what I mean* is, in SCO, generally a fixed phrase, whereas other long clusters found in other corpora are less so.

6. CONCLUSION

Given the limits of space, this discussion of the occurrence of the item *I* in a variety of corpora must be limited to a broad overview with the occasional deeper insight where only two selected corpora are directly compared.

Although the corpora count *I* occurrence for roughly similar total percentages of the whole of the corpus, and although a great number of *I* collocates co-occurs within similar probability (rank) in all seven corpora discussed, we are able to see a number of clear divergences.

Amongst the collocates, the biggest difference was found between two corpora that have material from the same age-group and geographical area (London). Being based on recordings 30 years apart and two different social group selections, the comparison between COLT and LTT gives an idea how corpora, who on the surface appear to represent similar groups, reflect the usage of very dissimilar groups when their occurrence patterns are

[11] Looking at SCO, we see that *You know what I mean* is used far more frequently here than in any of the corpora shown in Table 11.

compared - and these occurrence patterns reflect a differing socio-economic background and a long time-interval between the recording of the respective corpora. We also saw that each corpus has at least one of the ten most frequent collocates with *I* that is specific in use only in one corpus.

An investigation of longer clusters gives an idea that casual spoken English, regardless of geographical roots, has large areas of convergence, while there is a marked divergence to prepared speech (as seen in SEC) when we look at *I* use.

This finding is supported by the fact that the phrase *I don't know* is the highest occurring 3w cluster in all corpora bar SEC and LTT.

All this indicates the potential value of comparing the same item in corpora that differ by only one or two major criteria. That *I don't know* is not found to be the most frequent cluster in every single one of the corpora points to another discovery in this paper: Specific social groups of networks of speakers appear to have key phrases that are unique to them. Using the principle of weighing in how far frequent clusters are a constituent part of other, longer, frequent clusters, we find the clusters *I was like (oh my god)* and *I don't really* are proportionally far more frequent in LTT than in the other corpora, while *I'm not kidding* and *You know what I mean* is proportionally far more frequent in SCO than in the other corpora. It can be said that each community's self-reinforcing use of certain phrases can be seen as imprints of *lexical priming* (Hoey, 2005) found in use in individual speech communities.

REFERENCES

Austin, JL (1962) [2001]. *How to do things with words*. Oxford, OUP.

Biber D., Johansson S., Leech G., Conrad, S. and Finegan, E. (2000). *Longman Grammar of Spoken and Written English*.

Biber, D; Conrad, S; Leech, G. (2002). *Student Grammar of Spoken and Written English.* Harlow, Essex: Longman.

Fasulo, A & Zucchermaglio, C. (2002). My selves and I: identity marker in work meeting talk. In: *Journal of Pragmatics* 34, (1119 – 1144).

Halliday, M.A.K. (2004). The Spoken Language Corpus: a foundation for grammatical theory. In: *Language and Computers*. Vol. 49, no. 1.

Hanks, W. F. (1990). *Referential practise. Language, and lived space amongst the Maya*. Chicago: University of Chicago Press.

Hoey, Michael (2005). *Lexical Priming.* A new theory of words and language. London: Routledge.

Hunston, Susan & Francis, Gill. (2000). *Pattern Grammar.* A corpus-driven approach to the lexical grammar of English. Amsterdam: John Benjamins Publishing Co.

Ito, Rika & Tagliamonte, Sali (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. In: *Language in Society* 32, 257-279.

Kashima, E. & Kashima, Y. (1998). Culture and Language: The Case of Cultural Dimensions and Personal Pronoun Use. In: *Journal of Cross-Cultural Psychology* 29 461-488.

Knowles, Gerald O. (1973). *Scouse: The urban dialect of Liverpool*. PhD Thesis, unpublished. University of Leeds.

Leech, G & Svartvik, J. (1975). *A communicative Grammar of English.* London: Longman.

O'Donnell, Matthew Brook (2009). The *Adjusted Frequency* List: Evaluating a method for producing cluster-sensitive frequency counts. Presentation, AACL Edmonton, Canada – 10 October 2009.

O'Keefe, A., McCarthy, M., Carter, R. (2007). *From Corpus to Classroom.* Cambridge: Cambridge University Press.

Quirk, R.; Greenbaum, S.; Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language.* Harlow, Essex: Longman.

Pace-Sigge, M. (2010) *Evidence of Lexical Priming in Spoken Liverpool English*. PhD Thesis, unpublished. University of Liverpool. (forthcoming).

Renouf, A. & Sinclair, J.M. (1991). Collocational frameworks in English. In: Aijmer, Karin and Altenberg, Bengt: *English Corpus Linguistics.* London: Longman.

Rundell, Michael (Editor-in-Chief); Hoey, Michael (Chief Adviser) (2002). *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan.

Scott, M. (2004). WordSmith Tools version 4, Oxford: Oxford University Press. http://www.lexically.net/wordsmith/index.html (last accessed 02/10)

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford: OUP.

Stubbs, Michael (2001b). *Words and Phrases*. Oxford: Basil Blackwell.

Wolfram Alpha. Computational Knowledge Engine. At: http://www.wolframalpha.com/ (last accessed 10/09).

Diseño y técnicas de explotación de un corpus oral para el análisis de parámetros de calidad en interpretación

JOSÉ MANUEL PAZOS BRETAÑA
OLALLA GARCÍA BECERRA
RAFAEL BARRANCO-DROEGE
Universidad de Granada

Resumen
La investigación en el campo de la evaluación de la calidad en interpretación simultánea puede enriquecerse significativamente con las ventajas que aporta el uso de un corpus estructurado. En esta contribución presentamos las razones que subyacen a la creación del corpus oral ECIS, sus características y técnicas de aplicación que bien pudieran resultar útiles para otros campos.
Palabras clave: corpus oral, interpretación simultánea, evaluación de la calidad

Abstract
Research in simultaneous interpreting quality assessment could draw significant benefits from a structured corpus-assisted approach. In this paper we discuss the reasons behind the creation of ECIS - an oral corpus-, its characteristics and some implementation techniques that might well prove useful in other fields.
Keywords: oral corpus, simultaneous interpreting, quality assessment

1. ANTECEDENTES Y GÉNESIS DEL CORPUS[1]

El grupo *Investigación pedagógica y social sobre interpretación de conferencias y comunitaria* (HUM-560, Junta de Andalucía) desarrolla su trabajo en la Universidad de Granada desde el año 1995. Su directora, Ángela Collados Aís, inició, en 1998, una nueva línea de investigación en el ámbito de la calidad de la interpretación de conferencias con un estudio experimental en el que se analizaba la incidencia de la entonación monótona en la evaluación de la calidad que realizaban los usuarios (Collados Aís, 1998).

El primer estudio empírico de la calidad en interpretación simultánea (IS) se remonta a Bühler (1986), quien propuso dieciséis parámetros para sistematizar las expectativas de los usuarios respecto a la calidad. Entre los parámetros sugeridos por esta investigadora (op. cit.) destacan los siguientes: acento, agradabilidad de la voz, fluidez, cohesión lógica, transmisión

[1] Este trabajo de investigación se presenta en el marco de los proyectos de investigación HUM2007-62434/FILO (Ministerio de Ciencia y Tecnología, España) y P07-HUM-02730 (Junta de Andalucía, España).

correcta del sentido, transmisión completa del discurso, gramaticalidad, terminología y estilo. Este catálogo de parámetros ha sido utilizado, con ligeras modificaciones, en la mayoría de los estudios posteriores, tanto de expectativas (por ejemplo, Kurz, 1989, 1993), como de evaluación de la IS por los usuarios (entre otros, Gile, 1990).

Desde 1998, el grupo de investigación ha ido completando una serie de estudios, que se han centrado en once parámetros de calidad, añadiendo dicción y entonación a los mencionados (Collados Aís, 2001; Pradas Macías, 2003, 2004, 2006; Collados Aís & Pradas Macías & Stévaux & García Becerra, 2007; Barranco-Droege, 2009; Stévaux, en preparación).

Los resultados de estos estudios pusieron de manifiesto la necesidad de realizar un trabajo de carácter observacional en el que se analizasen los distintos parámetros en una situación real de interpretación. Para ello, en el marco del Proyecto de Investigación del Ministerio de Ciencia y Tecnología *Evaluación de la calidad en interpretación simultánea. Parámetros de incidencia* (BFF2002-00579), se llevó a cabo la recopilación de material para la elaboración de un corpus oral de interpretaciones reales de una sesión plenaria del Parlamento Europeo. Este material constituye el núcleo sobre el que se construye ECIS. Durante la realización de dicho proyecto, se estudiaron por primera vez en su contexto real los once parámetros de calidad, incorporando, además, dos aspectos nuevos: la influencia de los llamados *problem triggers*[2] de la interpretación y las estrategias de resolución de los intérpretes a partir de tres combinaciones lingüísticas (alemán-español, francés-español e inglés-español). Se analizaron, también, otros posibles factores que pudiesen afectar a la la calidad y su evaluación, como pueden ser las primeras impresiones (García Becerra, 2006; 2008).

En la etapa actual[3] se están efectuando un análisis y una cuantificación pormenorizados de los parámetros, factores y funciones relevantes del proceso de la IS, maximizadores y minimizadores de la calidad, a través de estudios empíricos con intérpretes y usuarios que, desde metodologías y enfoques complementarios, contribuyan a una definición de la calidad de la interpretación de conferencias a partir de las combinaciones lingüísticas ya mencionadas. Esta línea de investigación hizo necesario emprender la creación de un corpus

[2] Elementos que pueden desencadenar problemas en la interpretación (Gile, 1995).
[3] Una vez finalizada la primera etapa, el grupo recibió, en 2007, la concesión de dos proyectos: un proyecto I+D financiado por el Ministerio de Ciencia y Tecnología (HUM2007-62434/FILO) y un Proyecto de Excelencia de la Junta de Andalucía (P07-HUM-02730).

oral que permitiera la extracción de todos los datos necesarios y que, además, presentase una organización sistemática de esta información.

2. PREPARACIÓN TÉCNICA DEL MATERIAL

El corpus ECIS se ha elaborado partiendo de la grabación en vídeo de una sesión plenaria del Parlamento Europeo celebrada en Bruselas los días 10, 11 y 12 de marzo de 2003. De esta sesión se han obtenido un total de 44 discursos originales, presentados en inglés, francés o alemán. La duración de los discursos varía entre 1 y 8 minutos aproximadamente. A cada discurso original (DO) se asocian las IS realizadas a las distintas lenguas de trabajo. Dado que el interés principal del proyecto es, en primera instancia, el estudio de la incidencia de los parámetros que influyen en la calidad de la interpretación al español, ésta se ha incluido en todos los casos. Además se han incorporado al corpus –atendiendo a las futuras necesidades de los miembros del equipo– las interpretaciones realizadas al alemán y al francés de aquellos discursos originales realizados en lengua inglesa.

En resumen, el material en bruto del corpus consta de un total de 44 discursos originales (15 en inglés, 14 en alemán y 15 en francés) y de 44 interpretaciones al español, 15 al alemán y 15 al francés.

Tras la grabación de los discursos se ha procedido a aislar los segmentos relevantes y extraer el canal de audio correspondiente a los DO y a las IS, generándose ficheros individuales para cada uno. Debido a que las características técnicas de los DO y las IS no coincidían, ha sido preciso realizar un resampleado de los DO. Tanto el demultiplexado como el resampleado se han realizado mediante el programa Audacity[4], obteniéndose así el conjunto de ficheros normalizados[5].

El último paso de la preparación de los ficheros ha consistido en la sincronización de las IS y los DO. En este proceso, se han editado cada DO y sus correspondientes IS para sincronizar los inicios con el fin de obtener el contexto de "simultaneidad" de las interpretaciones.

---

[4] Audacity. Versión 1.2.6. [Software para OS X y otros sistemas] s.l. s.n. http://audacity.sourceforge.net.
[5] Contenedor *wav,* códec *araw* y pista *mono* a 22.050 Hz. 16 bits.

3. ANÁLISIS DEL MATERIAL SONORO Y EXTRACCIÓN DE INFORMACIÓN

3.1. Evaluación de las interpretaciones

Una tarea fundamental, puesto que sus resultados se utilizarán después como *tertium comparationis*, es la evaluación de las interpretaciones. Para llevar a cabo esta tarea, los miembros del grupo de investigación han utilizado los cuestionarios empleados en trabajos anteriores (por ejemplo, Collados Aís, 1998). Tras la escucha de las grabaciones de las IS, se han valorado, en una escala de 1 a 5, los 11 parámetros de calidad. Además, se han introducido las variables profesionalidad y fiabilidad, así como una valoración global de la actuación del intérprete.

3.2. Transcripción, segmentación y anotación

Cada fichero DO y sus IS asociadas se han transcrito ortográficamente, utilizando el programa Praat[6]. Esta transcripción se ha llevado a cabo segmentando los ficheros sonoros, de tal manera que es posible, mediante Praat (Fig. 1), acceder a cada fragmento y obtener, dado el caso, toda la información fonética para cada segmento en cuestión, además de para la totalidad del discurso. Durante la segmentación se han marcado, además, las pausas discursivas y silencios (vid. 3.5.).

[6] Boersma, Paul and Weenink, David (1992–2010). Praat: A system for doing phonetics by computer http://www.praat.org.

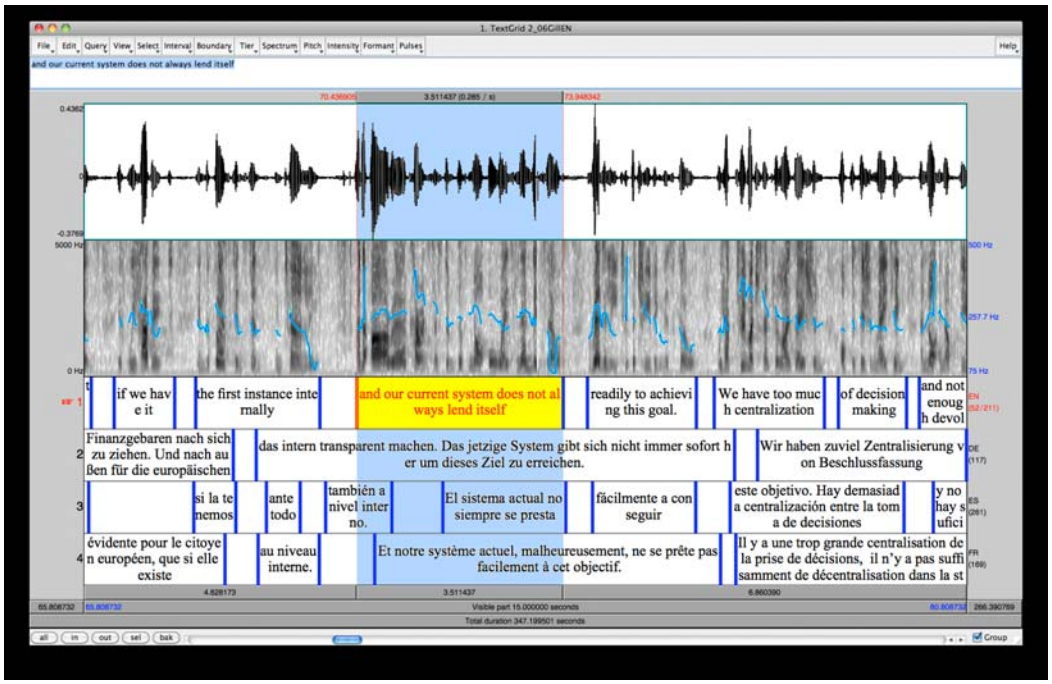


Figura 1: Instantánea de una ventana de Praat en la que se muestran el oscilograma, el espectrograma y la curva melódica correspondientes a un discurso (en este caso un DO en inglés) y la transcripción segmentada del mismo y de sus interpretaciones al alemán, español y francés.

3.3. Análisis cuantitativo

Hasta ahora se han descrito procesos cualitativos como la evaluación y la anotación. En la siguiente etapa del análisis del corpus para su explotación, se procederá a extraer del mismo aquella información cuantitativa relevante para la naturaleza de las investigaciones que se van a realizar. Esta información procederá, fundamentalmente, del análisis fonético acústico, del recuento de las palabras y sílabas articuladas y del etiquetado *POS*.

3.4. Análisis fonético acústico

Un parámetros acústico cuya cuantificación puede mostrar la relación entre la calidad de la interpretación y la entonación es la frecuencia fundamental. Una de las hipótesis de trabajo (Collados Aís, 2008) es que la entonación monótona se percibe, a partir de cierto umbral, como un elemento que influye de manera negativa en la recepción de una interpretación.

La *frecuencia fundamental*[7], en concreto su desviación típica, nos permite obtener un índice para medir este fenómeno. Praat nos proporciona un listado de los valores de la frecuencia fundamental de cada DO e IS, así como el valor de su desviación típica.

[7] La frecuencia fundamental (F_0) es la frecuencia más baja del espectro armónico tal que las frecuencias del resto de armónicos pueden expresarse como múltiplos de esta frecuencia fundamental.

3.5. Recuento de sílabas y palabras

La velocidad articulatoria o cadencia es un parámetro al que se le atribuye en algunos estudios (por ejemplo, Pio, 2003) una influencia determinante en la interpretación. Ello, además, desde una doble vertiente: por un lado, la velocidad de articulación del ponente en el DO y, por otro, la del intérprete (que puede estar, o no, en relación con la del DO).

Este parámetro se calcula de dos maneras distintas: medido en palabras por minuto y en sílabas por minuto. El hecho de calcularlo en sílabas por minuto permite obviar cierto desequilibrio que podría producirse según la combinación de lenguas DO/IS, y que se debe a la propia idiosincrasia tipológico-morfológica de cada lengua (Luque Durán, 2001: 51) –por ejemplo, el número de palabras de un discurso en alemán puede diferir del de su interpretación al español, simplemente por el hecho de ser una lengua más sintética (alemán) que analítica (español)–. El cálculo se realiza de manera sencilla haciendo un recuento del número de palabras y sílabas[8] de cada DO e IS dividiendo estas cantidades por la duración.

El hecho de haber tenido en cuenta las pausas y silencios durante la segmentación nos permite obtener lo que denominamos cadencias efectivas. Entendemos por tales el número de palabras (cadencia de palabras efectiva: CW_{ef}) y de sílabas (cadencia silábica efectiva: CS_{ef}) articuladas por segundo durante el discurso, sin tener en cuenta la duración de las pausas relevantes y silencios. El cálculo de las cadencias efectivas se realiza de manera similar al de la cadencia normal, pero restando de la duración los intervalos silentes que obtenemos utilizando Praat.

3.6. Etiquetado POS

Otro nivel que puede tener una influencia sobre la calidad de la interpretación es el morfológico-sintáctico. La existencia de una concentración de elementos léxicos pertenecientes a ciertas categorías gramaticales (por ejemplo, textos predominantemente nominales, verbales, etc.) bien pudiera reflejarse sobre los parámetros evaluados y cuantificados.

Para poder medir y evaluar esta hipotética influencia se ha realizado una anotación POS de la transcripción de los DO e IS. Se han utilizado para ello, según las lenguas, las aplicaciones *Freeling* (Carreras et al., 2004) y *Morphy* (Lezius, 2000). Los *tagsets* de ambos programas difieren un poco entre sí; el utilizado por Freeling está basado más consistentemente en EAGLES que el de Morphy. Sin embargo, para nuestro propósito,

ambos son más que adecuados y no se produce ningún conflicto, ya que sólo necesitamos para nuestro análisis la categoría gramatical general a la que pertenece cada elemento léxico. En el estudio, se considera principalmente la influencia de cuatro categorías: nombres, verbos, adjetivos y adverbios, aunque la disponibilidad inmediata de la información correspondiente al resto de categorías nos ha hecho considerar la pertinencia de incluir éstas también. Asimismo, se está contemplando la posibilidad de realizar análisis complementarios similares a los descritos en Pazos Bretaña & Pamies Bertrán (2008).

4. EXPLOTACIÓN DEL CORPUS Y ESTUDIOS DE LOS PARÁMETROS

Toda la información recopilada en las diferentes fases de elaboración y análisis del corpus se han recogido en una base de datos relacional convencional (MS Access) que permite exportar la información a hojas de cálculo cuando las circunstancias y necesidades lo requieran, bien para la realización de análisis estadísticos suplementarios, bien para la generación de tablas y gráficos[9].

Uno de los mayores problemas con los que nos encontramos a la hora de realizar un análisis estadístico de la influencia de las distintas variables, es que al existir un gran número de éstas y no poder conocer de antemano las relaciones que pueden existir entre algunas –si es que efectivamente existen– el volumen ingente de cálculos necesarios y datos resultantes enseguida alcanza una complejidad tal que no sólo puede ser difícil de mostrar, sino que incluso ciertas relaciones bien pudieran pasar desapercibidas.

Idealmente, sería necesario poder observar la variación e interacción de una manera gráfica e intuitiva en la que se muestren las variaciones e interacciones de los distintos parámetros, tanto aisladamente como en combinación con otros. Ello nos permitiría obtener y observar patrones relacionales y de comportamiento que de otra manera no serían detectados o requerirían de un esfuerzo mucho mayor para serlo.

Afortunadamente, existe una herramienta que cumple con estos requisitos. Se trata de *Trendalyzer,* desarrollado por la fundación *Gapminder* en un proyecto dirigido por el matemático y médico sueco Hans Rosling. Este software fue creado para permitir la visualización animada de la evolución de diversos conjuntos de datos a lo largo del tiempo.

---

[8] Existen varias herramientas para realizar el recuento silábico. Por ejemplo, *Morphy* y *Contador de sílabas* de Lexiquetos. A ambas puede accederse en http://wolfganglezius.de/doku.php?id=cl:morphy y http://lexiquetos.org/silio/ respectivamente.

[9] Véase apéndice 2.

En 2007, Google Inc. adquirió los derechos sobre este software, renombrándolo *Motion Chart* y lo puso a disposición del público, integrándolo en su *API*[10] de visualización.

Si bien la aplicación es una herramienta para analizar la evolución de variables en un marco temporal, es posible, realizando ciertos ajustes en la estructuración de los datos, utilizarla para el análisis de variables "estáticas", i.e. sin dependencia temporal.

Hemos volcado el conjunto de datos obtenidos en las fases de evaluación y análisis en una hoja de cálculo para poder estudiarlos con Trendalyzer (Fig. 2), y los resultados han sido muy satisfactorios. Al no existir una variación temporal de las variables, el efecto que se consigue es el equivalente a una animación en la que cada fotograma representa el gráfico de la combinación de los parámetros elegidos para una IS (o DO). La animación muestra la evolución de la combinación de variables elegidas en todo el conjunto, pudiéndose reconocer a simple vista la existencia o no de patrones de regularidad.

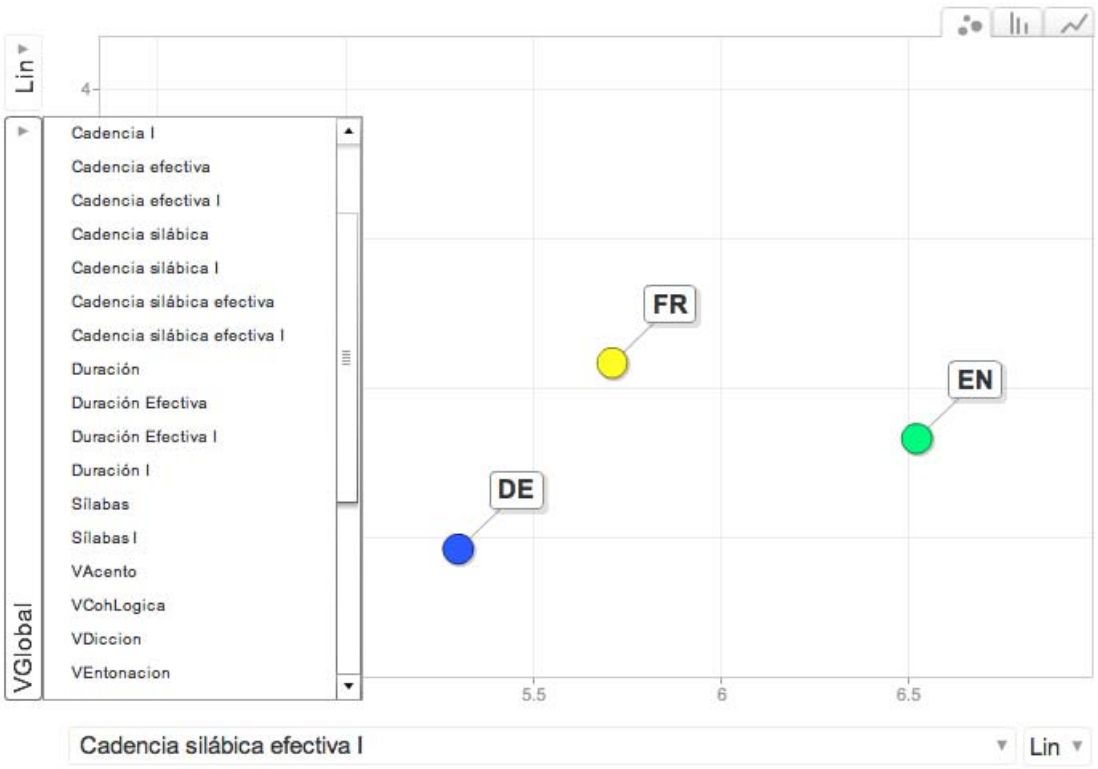


Figura 2: Vista de un momento de la animación del conjunto de datos correspondiente a la Valoración Global de las interpretaciones función de su CS_{ef} (la etiqueta muestra la lengua del DO). Se muestra asimismo parte del conjunto de variables disponibles en el eje de ordenadas. El mismo conjunto de variables está también disponible en el eje de abscisas.

[10] Application programming interface.

La aplicación permite mostrar la interacción de cualquier par de variables, es decir, en la agrupación que hemos realizado en este primer experimento (47 variables[11]), resultan un total de 1.081 combinaciones posibles. Si a ello se añade la posibilidad de comparar tres parámetros simultáneamente, representándose la tercera mediante una variación en el diámetro de los puntos, resultan en total 16.215 combinaciones.

5. CONCLUSIÓN

Hemos presentado un ejemplo de elaboración, tratamiento y aplicación de un corpus oral para el estudio de parámetros de calidad en IS sobre la base del desarrollo de ECIS.

Este corpus se distingue de otros *corpora* actualmente en uso en el campo de la interpretación[12] no sólo en el enfoque metodológico, sino en que ha sido diseñado atendiendo a los requisitos específicos de la investigación de la calidad en IS. Por primera vez en los estudios de interpretación asistidos por corpus, se aplica un enfoque comprehensivo que introduce la combinación de técnicas de análisis fonético-acústico, morfológico y estadístico. Otra importante innovación metodológica es la inclusión de una herramienta de análisis intuitivo, *Trendalyzer,* que es de una excepcional utilidad para un análisis de grandes volúmenes de datos que permite poner de manifiesto relaciones entre las distintas variables.

Si bien ECIS ha sido creado a la medida de las necesidades de nuestro grupo de trabajo, centradas en el estudio de los denominados parámetros "verbales[13]" y "no verbales[14]", creemos que, con las variaciones y adaptaciones específicas necesarias, la metodología puede ser aplicada a las investigaciones en otros campos del lenguaje.

[11] Véase apéndice 1.
[12] Monti et al. (2005), Meyer (2008) y Bendazzoli & Sandrelli (2009).
[13] Cohesión lógica, transmisión correcta, transmisión completa, gramaticalidad, terminología y estilo.
[14] Acento, agradabilidad de la voz, fluidez, entonación y dicción.

APÉNDICE 1

Tabla A1: Listado de los parámetros medidos y contabilizados en los discursos originales (DO) y las interpretaciones simultáneas (IS)

Lengua	Lengua del DO
Sexo	Sexo del orador
Tipo de Exposición	Espontánea, leída o semi-espontánea
Duración	Duración del DO en segundos
Duración Efectiva	Duración del DO en s. sin contabilizar intervalos silentes
Palabras	Número de palabras del DO
Cadencia	Nº de palabras articuladas por minuto en el DO
Cadencia efectiva	Nº de palabras articuladas por minuto en el DO sin contabilizar la duración de los intervalos silentes
Sílabas	Nº de sílabas del DO
Cadencia silábica	Nº de sílabas articuladas por minuto en el DO
Cadencia silábica efectiva	Nº de sílabas articuladas por minuto en el DO sin contabilizar la duración de los intervalos silentes
Frecuencia Fundamental	Media de la frecuencia fundamental del DO
Desviación Típica F0	Desviación estándar típica de la frecuencia fundamental del DO
Sustantivos	Nº de sustantivos del DO
Verbos	Nº de verbos del DO
Adjetivos	Nº de adjetivos del DO
Adverbios	Nº de Adverbios del DO
Lengua I	Lengua de la IS
Sexo I	Sexo del intérprete
Duración I	Duración de la IS en segundos
Duración Efectiva I	Duración del IS en s. sin contabilizar intervalos silentes
Palabras	Número de palabras de la IS
Cadencia I	Nº de palabras articuladas por minuto en la IS
Cadencia efectiva I	Nº de palabras articuladas por minuto en el IS sin contabilizar la duración de los intervalos silentes
Sílabas I	Nº de sílabas de la IS
Cadencia silábica I	Nº de sílabas articuladas por minuto en la IS
Cadencia silábica efectiva I	Nº de sílabas articuladas por minuto en la IS sin contabilizar la duración de los intervalos silentes
Frecuencia Fundamental I	Media de la frecuencia fundamental de la IS
Desviación Típica F0 I	Desviación estándar típica de la frecuencia fundamental de la IS
Sustantivos I	Nº de sustantivos de la IS
Verbos I	Nº de verbos de la IS
Adjetivos I	Nº de adjetivos de la IS
Adverbios I	Nº de Adverbios de la IS
VGlobal	Valoración global de la IS
VAcento	Valoración del acento del intérprete en la IS
VVoz	Valoración de la voz del intérprete en la IS
VFluidez	Valoración de la fluidez del intérprete en la IS
VCohLogica	Valoración de la coherencia lógica de la IS
VTransCorrecta	Valoración de la transmisión correcta del sentido en la IS
VTransCompleta	Valoración de la transmisión completa del sentido en la IS
VTerminologia	Valoración de terminología usada en la IS
VEstilo	Valoración del estilo de la IS

VEntonacion	Valoración de la entonación en la IS
VDiccion	Valoración de la dicción en la IS
VGramaticalidad	Valoración de la gramaticalidad de la IS
VProfesionalidad	Valoración de la profesionalidad del intérprete
VFiabilidad	Valoración de la fiabilidad de la IS

APÉNDICE 2.

Tabla B1: Ejemplo de la información extraída de un discurso original (DO)

ID	1
Código	208
Fecha	11/03/03
Sesión	Mañana
Lengua	DE
Orador	Ceyhun
Sexo	H
Temática	Inmigración
Tipo de Exposición	Espontánea
Duración	132,83
Duración Efectiva	110,08
Palabras	288
Cadencia	130,1
Cadencia efectiva	157
Sílabas	602
Cadencia silábica	271,9
Cadencia silábica efectiva	328,1
Frecuencia Fundamental	181,6
Desviación Típica F0	29,52
Sustantivos	82
Verbos	33
Adjetivos	24
Adverbios	37
Comentarios	

Tabla B2: Ejemplo de la información extraída de la interpretación simultánea (IS) del DO de la tabla A1

ID	1
Lengua I	ES
Sexo I	H
Duración I	134,47
Duración Efectiva I	102,75
Palabras	265
Cadencia I	118,2
Cadencia efectiva I	154,7
Sílabas I	545
Cadencia silábica I	243,2
Cadencia silábica efectiva I	318,3
Frecuencia Fundamental I	103,47
Desviación Típica F0 I	33,99
Sustantivos I	69
Verbos I	36
Adjetivos I	28
Adverbios I	15
VGlobal	2,46
VAcento	4,86
VVoz	3,21
VFluidez	2
VCohLogica	2,71
VTransCorrecta	2,57
VTransCompleta	2,57
VTerminologia	3
VEstilo	2,29
VEntonacion	2,21
VDiccion	2,5
VGramaticalidad	2,96
VProfesionalidad	2,57
VFiabilidad	2,5
Comentarios I	

REFERENCIAS BIBLIOGRÁFICAS

Barranco-Droege, R. (2009). *La transmisión correcta del sentido como parámetro de calidad en interpretación simultánea*. [Trabajo de investigación tutelada] Granada: Universidad de Granada.

Bendazzoli, C. & Sandrelli, A. (2009). Corpus-based interpreting studies: Early work and future prospects. *Tradumàtica 7.*

Consultado en http://webs2002.uab.es/tradumatica/revista/num7/articles/08/08art.htm

Bühler, H. (1986). Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters. *Multilingua 5:4*, 231-235.

Carreras, X. & Chao, I. & Padró, L. & Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisboa, Portugal

Collados Aís, A. (1998). La evaluación de la calidad en interpretación simultánea. La importancia de la comunicación no verbal. Granada: Comares Interlingua.

Collados Aís, A. (2001). The Evaluation of Quality in Simultaneous Interpreting. The importance of Nonverbal Communication. En F. Pöchhacker & M. Shlesinger (Eds), *The Interpreting Studies Reader,* (pp. 236-336). Londres: Routledge.

Collados Aís, A.; Pradas Macías, E. M.; Stévaux, E. & García Becerra, O. (Eds.) (2007). *Evaluación de la calidad en interpretación simultánea: parámetros de incidencia*. Granada: Comares Interlingua.

Collados Aís, A. (2008). Evaluación de la calidad en interpretación simultánea: contrastes de exposición e inferencias emocionales. Evaluación de la evaluación. En G. Hansen; A. Chesterman & H. Gezymisch-Arbogast (Eds.), *Efforts and Models in Interpreting and Translation Research* (pp. 193-214) Amsterdam/Philadelphia: John Benjamins.

García Becerra, O. (2006). La incidencia de las primeras impresiones en la evaluación de la calidad en interpretación simultánea: estudio piloto. [Trabajo de investigación para la obtención del DEA] Granada: Universidad de Granada.

García Becerra, O. (2008). La incidencia de las primeras impresiones en la evaluación de la calidad en interpretación simultánea: estudio piloto. En M. M. Fernández Sánchez & R. Muñoz Marín (Eds.), *Aproximaciones cognitivas al estudio de la traducción y la interpretación*. (pp. 301-326) Granada: Comares Interlingua.

Gile, D. (1990). L'évaluation de la qualité de l'interprétation par les délégués: une étude de cas, *The Interpreters' Newsletter* 3, 66-71.

Gile, D. (1995). Basic concepts and models of for interpreter and translator training. Ámsterdam: John Benjamins.

Kurz, I. (1989). Conference interpreting - user expectations. En D. Hammond (Ed.), *Coming of Age. Proceedings of the 30th Conference of the ATA* (pp. 143-148). Medford, N.J.: Learned Information Inc.

Kurz, I. (1993). Conference interpretation: expectations of different user groups. *The Interpreters' Newsletter* 5, 13-21.

Lezius, W. (2000). Morphy - German Morphology, Part-of-Speech Tagging and Applications. En U. Heid; S. Evert; E. Lehmann & C. Rohrer (Eds.), *Proceedings of the 9th EURALEX International Congress* (pp. 619-623). Stuttgart, Alemania.

Luque Durán, J. de D. (2001). Aspectos universales y particulares de las lenguas del mundo. Granada: Granada Lingvistica.

Meyer, B. (2008). Interpreting proper names: Different interventions in simultaneous and consecutive interpreting? *Trans-kom 1:1*.

Consultado en http://www.trans-kom.eu/ihv_01_01_2008.html

Monti, C.; Bendazzoli, C.; Sandrelli, A. & Russo, M. (2005). Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta* 50:4.

Consultado en http://www.erudit.org/revue/meta/2005/v50/n4/

Pazos Bretaña, J. M. & Pamies Bertrán, A. (2008). Combined statistical and grammatical criteria for the retrieval of phraseological units in an electronic corpus. En S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 391-406). Ámsterdam/Philadelphia: John Benjamins

Pio, S. (2003). "The relation between st delivery rate and quality in simultaneous interpretation" *The Interpreters' Newsletter* 12, 69-100.

Pradas Macías, E. M. (2003). Repercusión del intraparámetro pausas silenciosas en la fluidez: Influencia en las expectativas y en la evaluación de la calidad en interpretación simultánea. [Tesis doctoral sin publicar] Universidad de Granada. Granada.

Pradas Macías, E. M. (2004). La fluidez y sus pausas: enfoque desde la interpretación de conferencias. Granada: Comares.

Pradas Macías, E. M. (2006). Probing quality criteria in simultaneous interpreting: The role of silent pauses in fluency. *Interpreting* 8:1, 25-43.

Stévaux, É. (en prep.). *El acento como parámetro de calidad en interpretación simultánea*. [Tesis doctoral] Granada: Universidad de Granada.

El uso de aunque y pero por hablantes nativos y aprendices suecos

AYMÉ PINO RODRÍGUEZ
University of Gothenburg

Resumen
Se analizan variaciones del uso de aunque y pero en 42 textos producidos por suecos aprendices de E/LE (Corpus SAELE), comparado con el Corpus del español (nativos de lengua española), y el Corpus GP04 (nativos de lengua sueca). Los usos de estos conectores por estudiantes suecos corresponde en funciones al de hablantes nativos de español, pero se emplea por los primeros más frecuentemente. Ambos conectores se utilizan en proporción muy superior entre los suecos aprendices de E/LE que en los Corpus de referencia, probablemente como recurso idiomático que aún en niveles iniciales ya ha sido adquirido. El conector pero (o sus equivalentes idiomáticos), es el que con mayor frecuencia se emplea en los tres Corpus, pero evidentemente mucho más frecuentemente en lengua sueca que en español.
Palabras claves: Lingüística de Corpus, análisis contrastivo, conectores, estudiantes suecos, español como lengua extranjera

Abstract
The use of aunque and pero were analyzed in 42 texts produced by Swedish learners of E/LE (Corpus SAELE) compared with the respective use in the Corpus del español (natives Spanish), and the Corpus GP04 (natives Swedish). The use of these connectors in natives Spanish and Swedish learners roughly in terms of functions, but not in frequency. Both connectors are used in higher proportion among Swedish learners of E/LE than in the reference Corpus, probably as a language resource that even at initial levels has already been acquired. The connector pero *(or its idiomatic equivalent) is the most frequently used in the three corpora, but clearly more frequently in Swedish than in Spanish.*
Keywords: Corpus linguistics, contrastive analysis, connectors, Swedish learners, Spanish as a foreign language

1. INTRODUCCIÓN

El análisis contrastivo (AC) y la lingüística de Corpus (LC) se han empleado anteriormente para estudiar la interacción interlingüística en el proceso de adquisición de segundas lenguas, lenguas extranjeras y en el proceso de traducción (Altenberg, 2007; Butler, 2008; Johansson, 2007; Siepman, 2005; Tizón Couto, 2009). El análisis contrastivo compara sistemáticamente dos o más lenguas, para tratar de detectar y describir eventuales similitudes y diferencias (Johansson, 2007: 23). La lingüística de Corpus es una metodología basada en el uso de colecciones electrónicas para analizar diferentes aspectos de la lengua natural (Granger, 2002: 4).

El presente estudio se propone desarrollar un análisis contrastivo de las variaciones en el uso de los conectores contra-argumentativos *aunque* y *pero* en textos producidos en un

curso de Destrezas impartido para estudiantes suecos de E/LE, en dos universidades en Suecia durante el curso lectivo 2008-2009. El análisis no ventila aspectos de la traducción, sino de la interlengua para determinar si la función en contexto y proporción de *aunque* y *pero* puede considerarse idéntica o parecida en los textos producidos por estudiantes suecos de español como lengua extranjera, comparado con el Corpus del español (nativos de lengua española) y el Corpus GP04 (nativos de lengua sueca) .

1.1. Aunque y pero en español

Para algunos autores los conectores incluyen a las conjunciones, locuciones conjuntivas, adverbios y sintagmas preposicionales (Martínez, 1997: 19). Otros consideran que el comportamiento sintáctico de los conectores es diferente al de las conjunciones típicas (Alcaraz Varó & Martínez Linarez, 1997: 133). Conector puede definirse como la serie de unidades que enlazan enunciados del discurso para hacer explícita relaciones semánticas de adición, equivalencia, contraste o causa-consecuencia (Alcaraz Varó & Martínez Linares, 1997: 132).

El uso de estos elementos en el discurso ha sido anteriormente descrito como un problema a que el estudiante de E/LE se enfrenta en fases superiores en el aprendizaje de la lengua y no en fases iniciales (Patricio Da Silva, C. A., 2004: 39). El análisis de mensajes electrónicos elaborados entre estudiantes de E/LE en Viena y Tel Aviv con estudiantes de Hispánicas en Barcelona, reportó el uso del conector *pero* en frecuencia exactamente coincidente con las registradas en el CREA (Corpus de Referencia del Español Actual), y de *aunque* en proporción muy cercana al uso por los nativos de lengua española (Górska, 2006: 15-16).

Los conectores contra-argumentativos de mayor frecuencia de uso son *aunque*, *pero* y *mientras que* (Montolío, 2001; Portolés, 2001); pero el primero se considera como el transpositor más frecuentemente empleado en expresiones concesivas y no siempre puede alternarse en su función por *pero*, que a su vez es el más utilizado en las conjunciones adversativas (García, González & Polanco, 2007: 56). Ambos se utilizan en oraciones concesivas (Alarcos, 2005: 321, 373). Sin embargo, *pero* como conjunción no puede combinarse con otras ni tiene movilidad posicional ya que ha de preceder siempre al segundo de los miembros enlazados (Alcaraz Varó & Martínez Linarez, 1997: 133).

Como sustitutos del conector aunque pueden aparecer a pesar de que, pese a que, y si bien. Sustitutos de pero podrían ser sin embargo, no obstante, ahora bien, con todo, aun así y mas (Teberoski, 2007: 38).

1.2. Equivalentes de aunque y pero en sueco

En sueco *men* (*pero, mas*) y *fast* (*aunque, si bien, a pesar de que, pese a que*), son consideradas conjunciones adversativas y la palabra une usualmente enunciados y frases que tienen básicamente una función descriptiva (Teleman, Hellberg & Andersson, 1999: 729-730). Conectores como *men* y *trots* indican comparación u oposición (Strömquist, 2001: 119).

Entre las conjunciones concesivas están *fast* (*aunque*), *fastän* (*aunque*), *ehuru* (*aunque*, *aun*, *si bien*, en lengua escrita y formal), *trots att* (*a pesar de* [*que*], *pese a* [*que*], *oaktat* [*att*] (*aunque*, en lengua escrita), *även om* (*aunque*, *aun cuando*), *om* [*och*] *så* (en ese caso)*, om än* (*sino*). Estas conjunciones indican que la afirmación antecesora es un obstáculo insuficiente o una falsa premisa en el contexto en que aparecen y que la afirmación de la oración subordinada se considera verdadera (Teleman y otros, 1999: 737).

En la Tabla 1. aparece la correspondencia español-sueco, sueco-español de los conectores contra-argumentativos *aunque* y *pero* (Benson, Strandvik & Santos, 1999).

Tabla 1: Correspondencia de pero y aunque

Conectores en español	Conectores en sueco
Pero, mas	Men, dock, fast
Aunque, si bien, a pesar de que, pese a que	även om, fastän, trots (att), fast

2. MATERIAL Y MÉTODO

En el presente trabajo se ha creado un Corpus de suecos aprendices de español lengua extranjera (en lo sucesivo denominado SAELE), constituido por 42 textos argumentativos (2 textos por sujeto), con una extensión de 150-300 palabras elaborados por 21 aprendices de primer año de español como lengua extranjera (E/LE), con un nivel básico situado entre el A2 y el B1 de acuerdo a la escala del Marco común europeo de referencia para las lenguas[1] (MCER), donde se presupone que los aprendices : saben describir en términos sencillos aspectos de su pasado y su entorno así como cuestiones relacionadas con sus necesidades

[1] Véase en: http://cvc.cervantes.es/obref/marco/cap_03_01.htm. 13-02-10

inmediatas (A2) [...] y han comenzado a producir textos sencillos y coherentes sobre temas que les son familiares o en los que tiene un interés personal, pudiendo también describir experiencias, acontecimientos, deseos y aspiraciones, así como justificar brevemente sus opiniones o explicar sus planes (B1).

El trabajo se llevó a cabo en el Departamento de lenguas y literatura de la Universidad de Gotemburgo y en la Facultad de educación y comunicación de la Universidad de Jonköping, en Suecia. El Corpus SAELE comprende textos (lengua escrita) producidos por aprendices cuya lengua materna es el sueco y participaban en un curso de Destrezas. El trabajo se limita sincrónicamente al curso lectivo 2008-09. La edad de los sujetos oscilaba entre 20-38 años (16 del sexo femenino y 5 del masculino). Todos dieron su consentimiento anticipado e informado de participación.

Para la conformación del Corpus SAELE se empleó como herramienta el programa WordSmith tools. Primeramente se codificaron los textos eliminando todo tipo de dato personal y anotando en la copia sin formato un código uniforme para cada texto.

Por ejemplo: 08guh2ka que refiere el año (08), la universidad (gu= Universidad de Gotemburgo), el sexo (h= hembra o v= varón), el número de texto (2= el número 2) y el nombre ficticio (ka= Katarina). Posteriormente con la ayuda de la herramienta *Wordlist* se creó una lista de SAELE con 42 textos, se precisó el número total de palabras y tipos de palabras; y se identificó la posición y frecuencia de uso de las palabras objeto de análisis. Con la herramienta *Concord* se determinó el número de concordancias u ocurrencias, y las colocaciones más frecuentes de las palabras objeto de estudio.

Los Corpus de control empleados fueron el Corpus GP04 y el Corpus del Español.

El Corpus GP04 se encuentra en el banco de datos denominado *Språkbank* de la Universidad de Gotemburgo. Está formado por artículos del periódico sueco *Göteborg Posten*, recopilados desde el año 1987. Su elección como referencia se basó en que tanto su contenido como sincronicidad histórica es similar al Corpus del Español, y permite además extraer y reconocer diferentes significados en contexto de palabras y sus combinaciones a través de la búsqueda de concordancias. El GP04 tiene un total de 19 406 813 palabras y 597 056 tipos de palabras[2].

El Corpus del Español (Davies, 2002) está conformado por 100 millones de palabras y abarca el período histórico comprendido entre 1200 al 1900. Para el presente trabajo se utilizó exclusivamente la parte del período 1900 en el registro de noticias, conformada por

[2] Búsqueda realizada el 13-2-2010.

5 144 631 palabras[3] y 6 810 artículos de ABC Cultural 1991-95; Noticias - Argentina - La Prensa; Noticias – Argentina - El Cronista; Noticias - Bolivia – ERBOL; Noticias - Perú – Caretas; Noticias - Colombia –Semana y Noticias - Cuba/EEUU –CubaNet. La limitación a este período histórico pretendió homogenizar la muestra con el otro Corpus de control; el emplear la parte de noticias permitió por su parte armonizar en contenido y obtener registros lingüísticos relativamente similares a los efectos de la comparación.

3. RESULTADOS Y DISCUSIÓN

El Corpus SAELE contiene un total de 12 414 palabras y 2 530 tipos de palabras. El conector *pero* ocupa el número 18 en la lista de palabras, con ocurrencia de 111 (0,87%), apareciendo en 36 de los 42 textos (85,71%), mientras que *aunque* ocupa el número 155 en la lista de palabras, con una ocurrencia de 10 (0,08 %) apareciendo en 10 de los 42 textos (23,81%).

El Corpus GP04 contiene un total de 19 406 813 de palabras y 597 056 tipos de palabras. La Tabla 2 muestra que las expresiones correspondientes a *pero*: *men* (109 823), *dock* (9 716) y *fast* (5 821), aparecen con ocurrencia total de 125 360 (0.64%). Los conectores equivalentes en sueco a la palabra *aunque*: *även om* (5 124), *trots att* (3 775) y *fastän* (147), aparecen con ocurrencia total de 9 046 (0.04%).

La parte del Corpus del Español empleada contiene 5 144 631 de palabras. El conector *pero* tiene una ocurrencia de 12 786 (0.24%), y *aunque* de 3 487 (0.06%).

La Tabla 2 muestra la frecuencia superior de aparición del conector *pero* en el Corpus SAELE y en los de control. En el Corpus GP04 (nativos suecos) ocurren los equivalentes a la palabra *pero* 0.64%, y el total de los equivalentes al conector *aunque* ocurren en frecuencia del 0.04%.

[3] Búsqueda realizada en febrero de 2010.

Tabla 2: El Corpus SAELE, el Corpus del español y el Corpus GP04

Corpus	Corpus SAELE	Corpus del español Período histórico: 1900 Registro: News	Corpus GP04 Período:1987-- Registro: Nyheter
Total de palabras	12 514	5 144 631	19 406 813
Total de ocurrencias pero/men,dock, fast	Pero: 110 0,87%	Pero: 12 786 0,24%	Men: 109 823 0,56% *Dock*: 9 716 0,05% *Fast*: 5 821 0.02%
Total de ocurrencias aunque/Även om, fastän, trost att, fast	Aunque: 10 0,08%	Aunque: 3487 0,06%	Även om: 5 124 0,027% Fastän:147 7.57% Trots (att): 3 775 0.01% *Fast*: 5 821 0.02%

El conector *pero* (o sus equivalentes idiomáticos), aparece como el de mayor frecuencia de uso en el Corpus SAELE y los de control; pero entre hispanohablantes 4 veces más frecuentemente que *aunque* y 16 veces más frecuentemente sus respectivos equivalentes entre suecos del Corpus GP04. La proporción entre ambos conectores es en el Corpus SAELE intermedia: 10 veces más frecuentemente *pero* que *aunque*. La comparación revela que el uso de *pero* y sus equivalentes triplican en frecuencia su uso en el Corpus GP04 (nativos suecos) comparado con el Corpus del español (nativos hispanohablantes).

El conector *aunque* aparece 1,5 veces más usado entre nativos hispanohablantes.

En el Corpus SAELE aparece *pero* utilizado aproximadamente 1,5 veces más frecuentemente que en el Corpus GP04 y alrededor de 3 veces más frecuentemente que en el Corpus del español. Igualmente aparece *aunque* dos veces más frecuentemente usado en el Corpus SAELE, que sus equivalentes en el Corpus GP04 y aproximadamente 1,5 veces más frecuentemente que en el Corpus del español.

La Figura 1 ofrece las proporciones de usos de los conectores objeto de estudio, en el presente Corpus en relación con los de control.

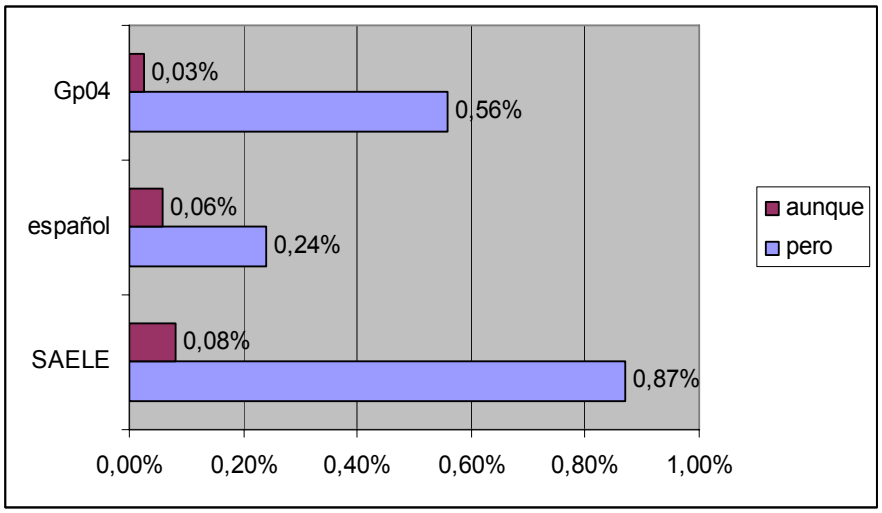


Figura 1: El uso de *aunque* y *pero* en SAELE, Corpus del español y GP04

Estos resultados contrastan con los de Górska (2006), quien reportó similitud en la proporción de uso de *pero* y *aunque* entre aprendices de E/LE comparado con nativos del Corpus CREA. Es prematuro decidir si la diferencia podría explicarse a partir de las diferentes lenguas maternas de los sujetos que integraron las muestras respectivas, o si simplemente indica diferencias metodológicas decisivas entre ambos estudios, por ejemplo la muestra del estudio de Górska aplica criterios de inclusión/exclusión que no permiten ver claramente el nivel de conocimiento de los sujetos que participan en el estudio; además de otros factores a tener en cuenta como: tipo de textos, número de participantes y elección del Corpus de control.

El uso en las concordancias en SAELE (Figura 2), muestra que los aprendices suecos utilizan adecuadamente *aunque* en las líneas 2, 3, 4, 6, 8 y 10; pero en ejemplos como los de las líneas 7 y 9, no aparece en expresiones gramaticalmente correctas con usos menos afortunados de preposiciones asociadas, omisión de acento en palabra esdrújula, tiempos verbales alterados, o su utilización en lugar de *a pesar de que*, en este caso semánticamente más correcto. En la línea 1, aparece un uso completamente agramatical.

En la Figura 2 aparecen en las líneas 8 y 9, casos en los cuales *aunque* podría sustituirse por *pero*, o sea en los cuales el conector cumple una función adversativa.

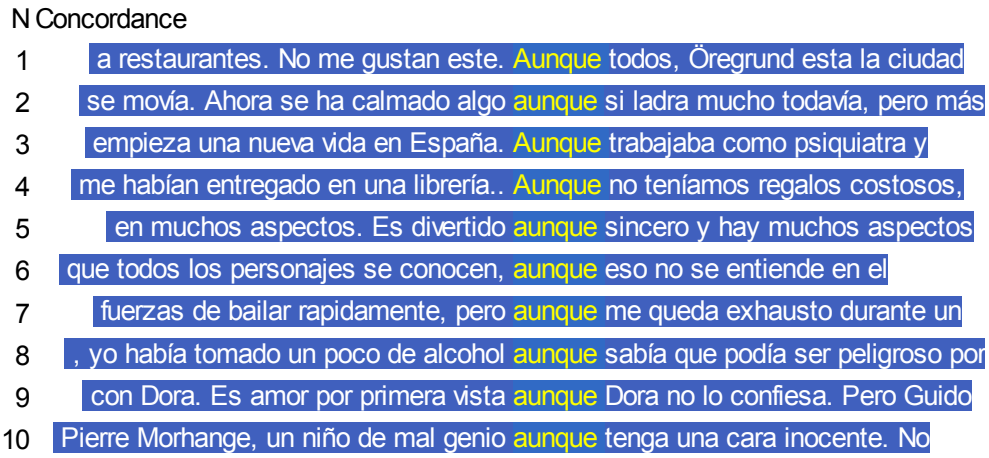


Figura 2: Lista con la totalidad de concordancias con *aunque*

En las líneas 3, 4, 5, 7 del Corpus SAELE (Figura 3) aparecen ejemplos del uso de *pero* como conector contra-argumentativo (de carácter adversativo), que se corresponde con la función de su equivalente en sueco *men*, en el ejemplo extraído del Corpus GP04: {Det är roligt, *men* absolut inget nytt}/Es divertido, *pero* de ninguna manera algo nuevo. En la línea 9: {Tal como temaba a los gatos tema a los coches...}, se aprecia un error en la conjugación verbal que suelen cometer los aprendices de este nivel A2 en camino hacia el nivel B1 según la escala de el MCER. En la línea 5:{casi es como una habitación pero sin paredas, y...}, se aprecia una errada construcción del plural de pared.

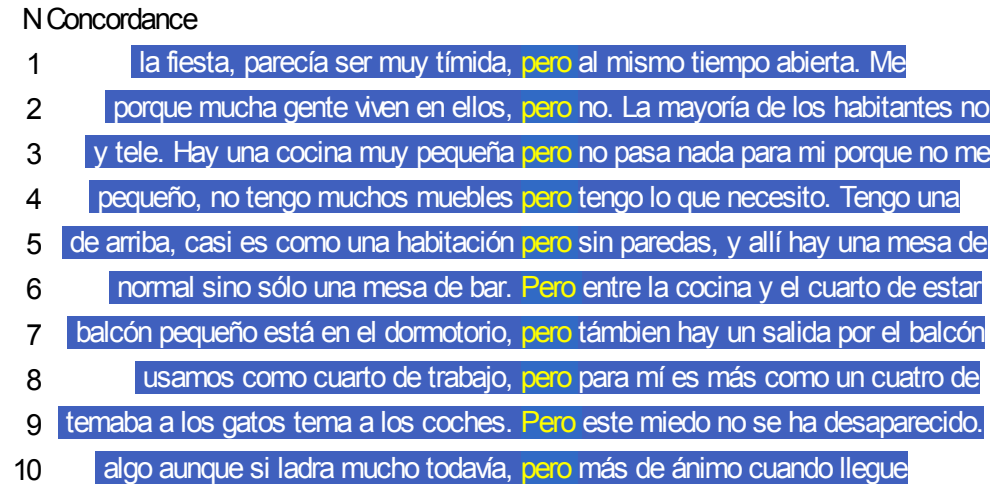


Figura 3: Lista con las diez primeras concordancias con *pero* (total=110)

A pesar de estos ejemplos, el uso de ambos conectores aparece en los textos de los aprendices suecos de E/LE en formas correctas en la mayoría de los casos.

4. Conclusión

Los usos de los conectores analizados se corresponden aproximadamente en hablantes nativos del español y estudiantes suecos en cuanto a funciones, pero sus frecuencias de usos divergen. Ambos se utilizan en proporción muy superior entre los suecos aprendices de E/LE que en los Corpus de referencia. Los aprendices suecos de E/LE parecen haber comenzado a utilizar ambos conectores en contexto como concesivos y opositores, a pesar de encontrarse entre el nivel A2 y en camino al B1, según la escala del MCRE.

Indiscutiblemente es *pero* (o sus equivalentes), el conector con mayor frecuencia de uso en el Corpus SAELE y en los Corpus de referencia. Apareció en el presente estudio 10 veces más frecuentemente que *aunque* y más frecuentemente que en los Corpus de referencia.

Referencias bibliográficas

Alarcos Llorach, E. (2005). *Gramática de la lengua española*. Madrid: ESPASA.

Alcaráz Varó, E. & Martínez Linares, M. A. (1997). *Diccionario de lingüística moderna*. Barcelona: Ariel.

Altenberg, B. (2007). The Correspondence of resultive Connectors in English and Swedish. *Nordic Journal of English Studies*, 6(1), 1-26.

Benson, K., Strandvik, I. & Santos M., M.E. (1999). *Norstedts stora spanska ordbok : spansk-svensk, svensk-spansk*. Stockholm : Norstedt.

Butler, C. S. (2008). Three English adverbs and their formal equivalents in Romance languages, A Corpus-based collocation study. *Languages in Contrast* 8(1), 107–124. doi: 10.1075/lic.8.1.10butissn 1387–6759 / e-issn 1569–9897.

Davies, M. (2002). *Corpus del Español* (100 million words, 1200s-1900s). Disponible en: http://www.Corpusdelespanol.org.

El *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación* (2001). Estrasburgo: Consejo de Europa. Departamento de política lingüística. Centro Virtual Cervantes. Disponible en: http://cvc.cervantes.es/obref/marco/cap_03_01.htm.

García A., M. A., González A., M. V. & Polanco M. F. (2007). *Lengua española: aspectos normativos y descriptivos en usos orales y escritos*. Barcelona: Universidad de Barcelona.

Górska, W. (2006). La enseñanza y el aprendizaje de los conectores contraargumentativos en español como lengua extranjera. El uso del CORPUS de alumno. *Investigaciones linguisticae*. Volumen XIII, 1-19.

Granger S. (2002). A Bird's-eye View of Computer Learner Corpus Research. En S. Granger, J. Hung y S. Petch-Tyson (Eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Language Learning and Language Teaching* 6, (pp. 3-33). Amsterdam & Philadelphia: Benjamin.

Johansson, S. (2007). *Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam: John Benjamin Publishing Company. Disponible en: http://site.ebrary.com.bibl.proxy.hj.se/lib/jonhh/docDetail.action?docID=10161067.

Martínez, R. (1997). *Conectando textos. Guía para el uso efectivo de elementos conectores en castellano.* Barcelona: OCTAEDRO.

Montolío, E. (2001). *Conectores de la Lengua Escrita*. Barcelona: Ariel Practicum.

Patricio Da Silva, C. (2004). Ordenación de los marcadores discursivos para la E/LE. En M. Martí (Eds), *Estudios de pragmagramática para la enseñanza del español como lengua extranjera,* (pp. 39-136). Madrid: Editorial Edinumen.

Portolés, J. (2001). *Marcadores del discurso*. Barcelona: Ariel Practicum.

Scott, M. *WordSmith Tools*. Versión 4. Licencia, Universidad de Gotemburgo. Suecia. http://www.lexically.net/wordsmith/.

Siepman, Dick. (2005). *A contrastive study of second-level discourse markers in native and non-native text with implications for general and pedagogic lexicography*. London & New York: Routledge, Taylor & Francys Group.

Språkbanken vid Göteborgs universitet, <http://spraakbanken.gu.se> Sökning utförd kl. 10.43 den 13 februari 2010 en GP04.

Strömquist, S. (2001). *Konsten att tala och skriva*. Malmö: Gleerups Utbilning AB.

Teberoski, A. (2007). El texto académico. En M. Castelló (Eds), *Escribir y comunicarse en contextos científicos y académicos. Conocimientos y estrategias,* (pp. 17-46). Barcelona: Editorial Graó, de IRIF, S.L.

Teleman, U., Hellberg, S. & Andersson, E. (1999). *Svenska Akademiens grammatik*. Stockholm: Svenska Akademien.

Tizón Couto, B. (2009). Un estudio basado en Corpus de la complementación clausal dependiente de verbos en el inglés oral de aprendices y nativos: comparando VICOLSE, LINDSEI Spa y LOCNEC. En: *XVII Congreso internacional de AESLA. Modos y formas de comunicación* (p.282). Ciudad Real: Universidad de Castilla- La Mancha.

Fragmentation of parallel sentences

SERGEY B. POTEMKIN

Moscow State University

Abstract

This paper presents novel approaches to phrase alignment for example-based machine translation (EBMT). We use matching of delimiters instead of word matching while determining fragment borders. Than we construct the critical path within two-dimensional bilingual space using dynamic programming. We follow a monotonic machine translation approach, for which we develop a flexible partial reordering with constraints. Experimental results show the significant improvement of translation quality.

Keywords: parallel sentences, translation, critical path, local inversion

Resumen

En el artículo se presentan nuevos modos de tratar la alineación de las oraciones para fines de la traducción automática sobre la base de ejemplos (EBMT). Al ejecutar la comparación de los límites de fragmentos de las oraciones se utilizan los separadores entre las palabras y no las palabras mismas. Luego, mediante métodos de programación dinámica, se constituye la vía crítica en el espacio bidimensional de bilingüismo con la utilización de un diccionario bilingüe. Se supone una coincidencia aproximativa del orden de las palabras en la oración inicial y en la de destino especial, pero el algoritmo es capaz de tratar las inversiones locales. Los resultados experimentales demuestran una mejora considerable de la calidad de la traducción.

Palabras clave: oraciones paralelas, traducción, vía crítica, inversión local

1. INTRODUCTION

Perfect fragmentation of the parallel bilingual texts represents an essential and first of all step in solving the problem of Example Based Machine Translation (EBMT). Fragmentation of bilingual corpora can be considered with various degrees of details - from the super-sentence unity (paragraph, chapter) to the word-level matching. We considered fragmentation of previously aligned sentences of parallel texts. The steady parallel fragments of sentences occur in the parallel texts rather frequently so the consistent selection of parallel fragments in two sentences one of which is "the perfect translation" of the other one (made by a qualified human translator, and probably strictly verified) is an important phase of EBMT implementation. This paper presents a new approach to fragmentation of sentences based on lexical and structural comparison of fragments of the source sentence (SS) and the translated sentence (TS). In the contrary to the known techniques, we use intervals bounded by delimiters (blank) between the words, not the words itself as the columns and rows of the adjacency matrix. It enables matching word combination contained in the translation lexicon, not singular words only. Then we process those fragments of the source sentence which can

be the inverted to match the translated fragment. Dynamic programming algorithm then is used for searching the best fragmentation of the sentence. Quality of fragmentation is defined by experts and could be increased by tuning of the weight factors. Also the algorithm of fragmentation without morphology markup is presented. We conclude the paper with presenting and discussing some matching improvements.

2. BILINGUAL SPACE

Each parallel text corpus defines a rectangular bilingual space (Melamed, 1999), as illustrated in Fig. 1. The upper left corner of the rectangle is the origin of the bilingual space and represents the two texts' beginnings. The lower right corner is the terminus and represents the texts' ends. Each bilingual space is spanned by a pair of axes. The lengths of the axes are the lengths of the two component texts. In our case the axes of a bilingual space are measured in words or tokens. Each bilingual space contains a number of true points of correspondence (TPCs), other than the origin and the terminus. If a token at position p on the x-axis and a token at position q on the y-axis are translations of each other, then the coordinate (p, q) in the bilingual space is a TPC. Similarly, TPCs arise from corresponding boundaries between paragraphs, chapters, list items, etc. A complete set of TPCs for a particular bilingual text is the true bilingual map (TBM). The purpose of a bilingual mapping algorithm is to produce bilingual maps that are the best possible approximations of each bilingual's TBM. The order of sentences in the source and the target texts coincide, therefore map is monotone. When we proceed to mapping separate sentences the situation is changed. For pairs of rather different languages the order of words of SS and TS is different.

3. ADJACENCY MATRIX

To compare the sentences their words are usually arranged by rows and columns of a rectangular table (matrix). This matrix gives an image for evaluating proximity of the texts (Kedrova & Potemkin, 2005). Let us consider sentences taken from "Crime and Punishment" by F.M. Dostoevsky, translated (Garnett, sentence #2 from Part 1 Chapter 2: *«Но теперь его вдруг что-то потянуло к людям»* to a target sentence (human translation): *"But now all at once he felt a desire to be with other people"*

The words of the SS could be placed on by X-axes and the words of TS - by Y-axes. Each word is set at the coordinate according to its number in the sentence. For example, *к* is the seventh word in the source text, therefore it is located in column x = 7. We use translation

lexicon to match words of SS and TS. Translation lexicon L can be represented as a sequence of entries, where each entry is a pair of equivalent words: L ~ *{(x1,y1) . . . . . (xt, yt)}*. If a mesh lays on intersection of two words fixed in the translation lexicon as a pair of equivalents, it contains 1 otherwise 0. For instance mesh (2,2) designates a pair of words {*теперь, now*}. Generally the value in a matrix mesh lies between 0 and 1 and depends on the "measure of proximity" of equivalents, to be defined forth.

4. DELIMITERS AS COORDINATES

In the contrary to the described representation we consider delimiters between the words as coordinate references of the bilingual space, not words or tokens. So mapping of SS word on a TS word is a segment with coordinates {(x1, y1); (x2, y2)} where x1, x2 - beginning and ending of the source word, and y1, y2 - beginning and ending of the target word. If we compare words one by one x2 = x1 + 1, y2 = y1 + 1. The “slope” of this segment is 45 degrees in the direction from 0 point downwards right. It is interesting, that if we compare sentences with different writing direction (e.g., English and Arabian) the slope of such segment will be -45 degrees. Let us try now to compare not only word-to-word equivalents found in the translation lexicon, but also such equivalents as word to word-combination, word-combination to word and word-combination to word-combination. It could be made easily within our approach. Fig. 3 shows mapping of a SS on a TS with word-combinations from the translation lexicon, (*вдруг == all at once*).

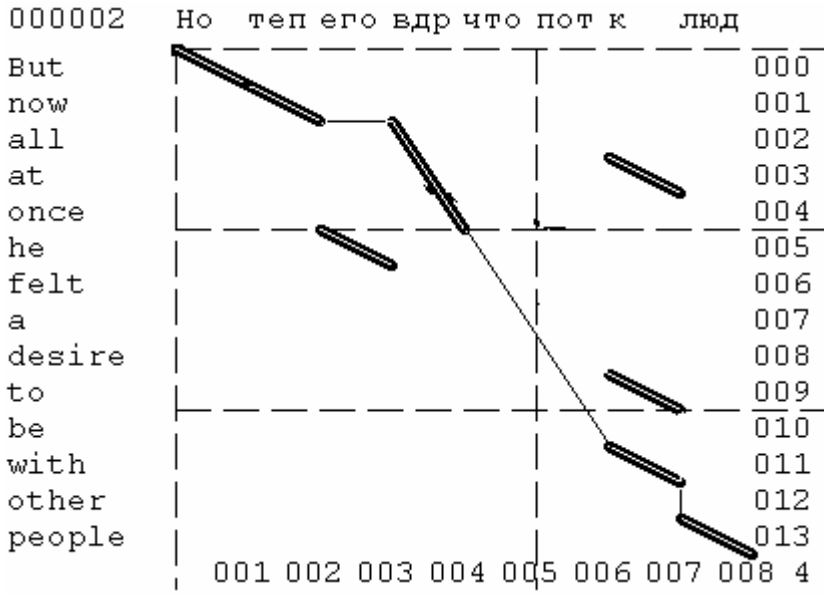


Figure 1: TPCs (double lines) and the critical path (thin lines)

Note that the fragments, which we are going to detect, could be included into translation lexicon and used as word-combinations. Thus we made the first generalization of the earlier used paradigm - we have changed comparison of words to comparison of intervals and at once we obtained an opportunity to compare word-combinations. The adjacency matrix was transformed to SS - TS mapping including word-combinations. If some mapping segments are superimposed on a horizontal or on a vertical axis we'll have a collision (i.e. not one-to-one mapping). We can see that word of the source text *"к"* is mapped on three words of the target text "*at", "to", "with"*. Collision is frequently produced by syntactic words and by punctuation marks. Collision resolving, i.e. elimination of all except one words in collision is necessary for perfect fragmentation.

5. SEGMENT WEIGHT

It was mentioned that the measure of (semantic) proximity between words is a normalized variable with value between 0 and 1. For two words being the most statistically probable translation of each other, (*люди, people*) in our example this measure is higher, and for the rare equivalents (*к, with*) - is lower. We use Lexical Database (LDB) with the superimposed semantic metric (Kedrova & Potemkin, 2004), (Potemkin, 2004) for evaluation of semantic proximity between the words. The essence of our method for proximity evaluation consists in calculation of normalized scalar product of two words taken as vectors in the space of the bilingual translation lexicon. When we change coordinates represented by tokens to coordinates represented by delimiters we have to change the proximity concept to the concept of segment weight. For a segment mapping one word onto another, its weight is equal to a measure of semantic proximity of these two words. For segments composing a continuous line *(но теперь ⇔ but now)* it is necessary to take into account cumulative effect of merging because such mapping should be more trustable than when the same words are mapped separately and, hence, we should attach to such combined segments the weight higher, than the sum of their components. Even better when a SS word-combination is mapped onto a TS word-combination.

6. FRAGMENTATION

After attaching weights to all corresponding segments it is possible to deal with fragmentation itself: to map SS intervals onto TS intervals between already matched

segments, i.e. to make interpolation between the matched segments. Then it is necessary to choose the best one among all virtual fragmentations according to some criterion such as:

- Maximize weight of all matched segments
- Minimize total length of interpolation segments
- Maximize number of fragments … etc.

The words of two sentences can be matched in different sequence. The same sentence can be translated in a direct or in an inverse order of words and both translations were right. A more common case - when one group of words is translated in a direct, and another - in an inverse direction these groups are matched chaotically. However if we consider only monotone mappings (i.e. we assume that order of words of source and target sentence coincide), the sequence of matched segments in a fragmentation could be considered as a path from point 0 to terminus. The path with maximal weight corresponds to the best fragmentation, each sub-path of the critical path also is a critical path and the problem permits solution within dynamic programming (DP).

This DP algorithm constructs the interpolating segments in the bilingual space as shown in Fig. 1 by thin line. The algorithm implicitly eliminates collisions as only one word in a collision is chosen. Note that we permit violation of one-to-one mapping, i.e. the construction of segments parallel to axes X and Y are admissible. Such segments reflect the fact that some SS words not correspond any TS word and vice versa. So, the fragment *be with other* does not correspond any SS fragment. Fragment *его*, though has the correspondence *he*, does not enter into the critical path and is considered as missed in translation.

7. LOCAL INVERSION

As a rule, the source sentence and its translation while having coinciding order of words in general, contain some fragments with inversion. Let's consider sentence #34 from Chapter 2 of the above-mentioned novel by F.M. Dostoevski: *«Он был хмелен, но говорил речисто и бойко, изредка только сбиваясь немного и затягивая речь»* and its translation: *«He was drunk but spoke fluently and boldly, only occasionally losing the tread of his sentence and drawling his words».* This pair of sentences contains an inverse fragment: *{изредка только* (only occasionally}. Such local inversion should be included in the critical path, but the aforesaid algorithm does not accept it. Having maps изредка (occasionally and только (only we are going to generate a map between word-combinations изредка только (only occasionally, as shown in Fig. 2 by solid line {(10,10); (12,12)}. Such virtual segments can

be generated for any pair of words bounding "inverse segments", but the number of such "fragments" for each sentence is rather great that's why the upper threshold of length of fragments, say 4 words, and also the lower threshold of their length ratio, say 0.5 is imposed. After generating all mapping segments for inverse fragments they are included in the set of mapping segments and participate in the DP algorithm.

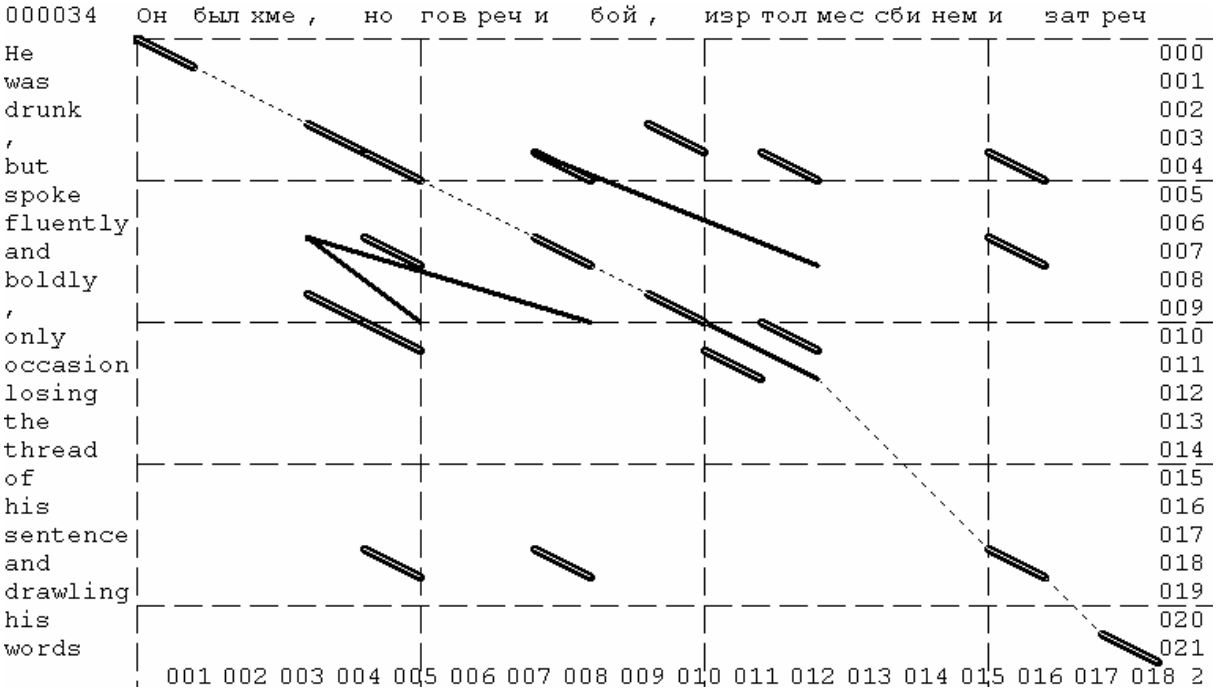


Figure 2: Inverse fragments are shown as solid lines

8. FRAGMENTS MERGING

Now let us again consider sentence #2. The critical path breaks the SS, TS pair into the following fragments:

1. Но теперь (But now

2. теперь его вдруг (now all at once

3. вдруг что-то потянуло к (all at once he felt a desire to be with

4. к людям (with other people

Fragment 1 is perfect. Fragment 2 contains pronoun его in SS, which is missed in the TS. Fragment 3 contains pronoun he in TS, which is missed in the SS. If we merge fragments 2 and 3, the outcome will be more meaningful. Fragments' merging is governed by two criteria: a) ratio of lengths of source and target fragments should be near 1 and b) ratio of the

weight of the matched segments in the critical path to the gross weight of the matched segments, should not be too small. We intentionally use the fuzzy definitions, because the values of the said threshold should be tuned experimentally. Just according to b) criterion fragments 2 and 3 were merged by the program.

9. Conclusion

This paper presents the strategy for fragmentation of sentences of parallel texts. The delimiters as coordinate references of the bilingual space was introduces, not the words themselves. It enables to extend boundaries of fragments using word-combinations from the translation lexicon. Semantic evaluation of the mapping segment weight is offered. Local inversion is effectively processed by inclusion of virtual segments in the critical path. Obtained fragmentation is evaluated according to structural and semantic criteria. If either one is violated two subsequent fragments are merged (in extreme case all fragments are merged and form the initial pair of sentences). Our experiments show, that the method improves sentence-level fragmentation of bilingual texts. Implementation of the procedure enables to build an automatic dictionary of fragments for use in Example Based Machine Translation - EBMT (Brown, 1996).

References

Brown, R. (1996). Example-Based Machine Translation in the Pangloss System. *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING-96)*. (pp. 169–174 (vol 1)), Copenhagen, Denmark.

Garnett, C. (1994) Crime and Punishment by F. Dostoevsky. translated by Garnett, Constance Black, Publisher: Random House Inc; New edition 1994

Kedrova, G. & Potemkin, S. (2005). Automatic evaluation of machine translation based on semantic metrics. *Bulletin of the Lugansk NPGU 15(95).* (pp. 35-41) (in Russian)

Kedrova, G. & Potemkin, S. (2004). Semantic homograph disambiguation using translation lexicon and dictionary of synonyms. *Proceedings of the II International congress "Russian language: Historical faith and contemporary"* Moscow, Russia. (in Russian)

Melamed, I. (1999). Bitext Maps and Alignment via Pattern Recognition, *Computational Linguistics 25(1),* 107-130

Potemkin, S. (2004) Lexical database with superimposed semantic metrics. *Proceedings of the II International congress "Russian language: Historical faith and contemporary"*. Moscow, Russia. (in Russian)

How to work with smaller corpora of indigenous languages

REGINA PUSTET
University of Munich

Abstract
Current corpus linguistics rarely takes data from indigenous languages into account. Sooner or later, however, the theoretical generalizations and methods produced by this approach need to be tested against data from "exotic", i.e. non-Indo-European languages. It cannot be assumed that the facts extracted from corpora of Indo-European languages alone represent the totality of corpus-related phenomena to be discovered in the languages spoken on this planet. From the perspective of standard corpus linguistics, however, working with corpora from indigenous languages is fraught with difficulties, mainly because such corpora either are not available at all, or they are much smaller than average corpora from Indo-European languages. On the basis of data from the Native American language Lakota, this paper demonstrates that for many types of theoretically significant corpus analyses, especially those which target grammatical items, smaller corpora (20 000+ words) are sufficient.
Keywords: indigenous languages, Lakota, discourse frequency, usage-based models, grammatical items

Resumen
La lingüística de corpus actual rara vez tiene en cuenta datos de lenguas indígenas. Sin embargo, tarde o temprano, las generalizaciones y métodos teóricos que se extraen de este enfoque necesitan contrastarse con datos "exóticos", es decir, de las lenguas no indoeuropeas. No se puede asumir que los hechos extraídos únicamente de corpus de lenguas indoeuropeas representan la totalidad de los fenómenos relacionados con corpus que se encuentran en las lenguas habladas en este planeta. Sin embargo, desde la perspectiva de la lingüística de corpus estándar, trabajar con corpus de lenguas indígenas está plagado de dificultades, principalmente porque dichos corpus no están disponibles, o porque son mucho más pequeños que los existentes de lenguas indoeuropeas. Basándonos en datos de la lengua nativa americana Lakota, esta ponencia demuestra que corpus más pequeños (20.000+ palabras) son suficientes para muchos tipos de análisis de corpus con significado teórico, especialmente aquellos centrados en elementos gramaticales.
Palabras clave: lenguas indígenas, Lakota, frecuencia de discurso, modelos basados en uso, elementos gramaticales

1. INTRODUCTION

Corpus-based linguistic research has produced a multitude of approaches to and insights into the structure of language, and it is more than likely that this relatively recent offspring of linguistics will continue to expand in the future.

However, anywhere in science, novel frameworks must be tested against a wide variety of diverse data to ensure the validity of the methods used and the resulting theoretical conclusions, and corpus linguistics is no exception. In this context, a crucial "playground" for the further extension of corpus studies is indigenous languages. To this point, the impressive achievements of current corpus linguistics are predominantly derived from Indo-European

input, since in addition to just a handful of globally influential non-Indo-European languages such as Chinese and Japanese, it is Indo-European languages that provide the databases for corpus studies.

The challenges to generative language theory posited by phenomena found in non-Indo-European languages (e.g. ergativity, split intransitivity), and which have stimulated substantial revisions of the original Chomskyan framework, have shown that the potential contributions of non-Indo-European languages to general linguistic theory are absolutely essential. Thus, the novel framework of corpus linguistics will, sooner or later, be faced with the structural diversity which inevitably emerges in the systematic study of genetically unrelated languages. To take just one random example: discourse structure in the languages of the world varies considerably, for instance, with regard to sentence length; in many Papuan languages in which particular clause chaining mechanism are used, a sentence can be coextensive with a whole narrative, while in Blackfoot (Algonquian language family, North America), sentences tend to be very short. Indo-European languages would fall in between these extremes, but without taking data from non-Indo-European languages into account, significant parameters along which discourse structures in the languages of the world differ, and which are highly relevant for corpus linguistics, would remain undetected.

2. Theoretical potential of research based on smaller corpora

The reasons why indigenous languages do, so far, not figure very prominently in corpus research include

(a) Availability: many indigenous languages are exclusively spoken languages, i.e. they have never been documented in ways that produce corpora. When such languages are being described, the compilation of grammars and dictionaries usually takes precedence over the publication of corpora.

(b) Corpus complexity: if the documentation of an indigenous language includes corpora, these tend to be much less complex than those that mainstream corpus linguistics works with.

It is obvious that small corpus size imposes limitations on certain types of analysis. For instance, the semantic range of a lexical item cannot be effectively investigated if the latter occurs in the corpus only once or twice. However, many types of corpus analysis do in fact not require large corpora, especially those which target grammatical elements. Obviously, in the discourse of any language, grammatical items occur more frequently than lexical items. In

smaller corpora (less than 100 000 words), individual grammatical items, such as person/tense/aspect/modality markers, or case markers, are frequent enough to conduct Bybee-style studies of language change -- despite the fact that within the framework of usage-based models, including Bybee's research, large corpora are the norm. For instance, by means of a 70 000-word corpus, Pustet (2008) shows that in the Native American language Lakota (Siouan language family, Central North America), case marking affixes are developing out of postpositions. In this case, as predicted by Jurafsky et al. (2001) on the basis of English corpus data, the discourse frequency of individual collocations, i.e. postposition-plus-noun syntagms, is the decisive trigger for the development from postposition to affix. For instance, the postposition *mahél* 'inside' shows a significantly higher percentage of affixal uses in the high-frequency collocation with the noun *thí* 'house' than in the lower-frequency collocation with *makhá* 'earth'. The general conclusions drawn on the basis of this study are that at first glance, the likelihood of transformation of a postposition into an affix depends on the discourse frequency of that postposition in and of itself: at least in Lakota, it is mostly the higher-frequency postpositions that show affixal alternants at all. An in-depth analysis, however, reveals that raw frequency of postpositions is only of secondary importance since numerous postpositions in the high-frequency range never figure as affixes. The crucial difference between affixally used postpositions and "pure" postpositions lies in that fact that the former occur in high-frequency collocations such as *thí* 'house' plus *mahél* 'inside' (= 'inside the house'), whereas the latter are mostly found as components of low-frequency collocations. Affixal uses of postpositions are observed in high-frequency collocations only. As predicted by usage-based models, the Lakota data thus identify high-frequency collocations as the primary locus of language change. As time passes, lower-frequency collocations will be affected by the change as well, until the change is fully implemented. At some point in time (if Lakota survives a few more centuries), all of today's postpositions might end up as affixes. A comparison between two Lakota corpora which are about 70 years apart indicates an increased use of affixal postpositions in the more recent corpus. Similarly, Pustet (forthcoming), a detailed study that specifically deals with the phonetic reduction of the Lakota postposition *etąhą* 'from' to the affixes *-tąhą* 'from' and finally, *-tą* 'from', demonstrates that this change is sensitive to the frequency of the collocations in which the element in question occurs. A comparison of a recent Lakota language corpus with an older corpus clearly documents a gradual change which increasingly favors the reduced forms. For instance, there are only two elements which occur with the maximally shortened form *-tą* in both corpora: *hé* 'that' and *lé* 'this'. From the older to the

recent corpus, the percentage of *-tą*-forms in collocations with the demonstratives *hé* 'that' and *lé* 'this' goes up from 28.7% and 33.3%, respectively, to 100% and 88.9%. Again, the sizes of the corpora used for the study Pustet (forthcoming) range between 20 000 and 70 000 words.

As a further illustration of what is said above about the fruitfulness of work with small corpora, the present paper is, in particular, concerned with the mechanisms underlying the complex changes observed in the context of the structural reduction of the Lakota plural marker *-pi*, which exhibits various additional phonetic shapes, such as *-p*, *-b*, *-m*, *-w, -o*, and *-Ø*. Note that although these alternants are not dialectal, idiolectal, or otherwise limited in their distribution, they are still written as *-pi* in the standard transcription of Lakota. When presented with *-pi*-forms by linguists, Lakota native speakers often classify them as "not natural" because in the spoken language, shortening of *-pi* is so pervasive today. Needless to say, the data used for the present study are taken from an exact phonetic transcription of a modern spoken Lakota language corpus. Examples:

(1) *-pi* 'plural':

*lé ąpétu ki tókhel Lakhóta ki wóyute kąġa-**pi** na ...*
this day DEF how Lakota DEF food make-**PL** and
'today I will talk about how the Lakota made food and …'

(2) *-p* 'plural':

*cha thokáheya thaló pus-yá **- p** na pápa kąga-**p** na'íš ...*
so first meat dry-CAUS-**PL** and jerky(dried meat) make-**PL** or
'so first they dried meat and made jerky or …'

(3) *-b* 'plural':

*hél wíyaka núm a-íyakaška-**b** cha hél waphégnaka eyá-**b** .*
there feather two L-tie **-PL** so there headgear say-**PL**
'they tied two feathers there, and so they called it headgear'

(4) *-m* 'plural':

*iyáya-pi nahą thípi wą ékigle-**m** na'íš ...*
go-PL and tipi IDF put up-**PL** or
'they went and put up a tipi or…'

(5) -*w* 'plural':

lé *wagmú* *kihą* *waslésleca-**w*** *na* ...

this squash DEF cut into small strips-**PL** and

'they cut the squash into small strips and…'

(6) -*o* 'plural':

*yužáža-**o*** *na* *yuškíca-**o*** *nahą* ...

wash-**PL** and wring out-**PL** and

'they wash it and wring it out and…'

The occurrence of these plural marking alternants is not random, but rather, depends (a) on discourse frequency, (b) on morphosyntactic factors, and (c) on phonological factors. The findings of this study are in line with Bybee's (2007:ch.11) recent claim that reductive change is, in certain cases, probabilistically determined by both discourse frequency and phonological factors. A 22000-word Lakota corpus is sufficient to arrive at statistical figures that prove this point. Since Bybee's (2007:ch.11) results are exclusively derived from data from English and Spanish, the Lakota data can be regarded as a useful test for the applicability of usage-based models to non-Indo-European languages.

The variation in the phonetic shapes of the plural marker *-pi* is not caused by regular phonological processes in Lakota, i.e. it is not the result of the phonological rules which operate in this language. Lack of conditioning by general phonological rules, in turn, is a widespread characteristic of linguistic items undergoing reductive change. An example is the phonologically unpredictable transformation of the English future marker *going to* into *gonna*. However, the respective allomorphs of *-pi* show strong, though not exclusive, preferences for specific morphosyntactic contexts. For instance, almost all occurrences of the variant *-o* 'plural' are in contexts in which the future marker *-ktA* follows directly; the variant *-p* 'plural' occurs in the majority of contexts for plural markers found in the corpus, but in contexts in which the plural marking slot directly precedes the elements *cha* 'and so', *chąkhé* 'and so', *na'íš* 'and, or', and *-šni* 'not', the variant *-p* 'plural' is far more likely to be used than any of the other variants. These cases have been analyzed as instances of morphosyntactic conditioning. The hypothesis that they might be due to phonological conditioning has been taken into consideration but could not be verified on the basis of the statistical data.

However, as in the English and Spanish data given in Bybee (2007:ch.11), reductive change may be sensitive to phonological factors in that specific phonetic environments speed

up the rate of change: "Words that more frequently occur in the context favoring a change undergo the change at a faster rate than those that occur less frequently in the appropriate context" (Bybee 2007:250). Word-final /t/ or /d/ is often lost in American English, but the probability of loss depends on the phonetic environment of these sounds:

> "…with a word such as *different*, the /t/ always follows the /n/, which presents a favorable condition for deletion, but sometimes *different* precedes a vowel, as in *different one*, and other times it precedes a consonant, as in *different story*… the preceding environment was the factor with the strongest effect on deletion, conditioning more strongly than the following environment…For example, the *t* in words ending in *-nt* actually deletes more often before vowels than before consonants" (Bybee, 2007: 247).

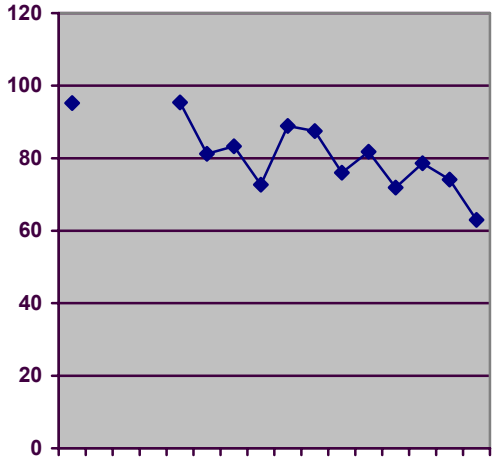
Phonetic environments have a probabilistic impact on the choice of Lakota plural marking allomorphs as well: the percentage of occurrence of the full form *-pi* 'plural' decreases significantly in favor of one of the reduced forms *-p*, *-b*, *-m*, *-w,* and *-o* when the plural marking slot is followed by an element with an initial alveolar affricate (symbolized by /*c*/). Thus, while the overall frequency of the full form *-pi* in the corpus is 4.1% when an element with initial /*c*/, such as *cha* 'and so' or *chąšna* 'whenever' follows, the overall frequency of *-pi* is 24.9% in all other phonological contexts that follow this marker in the corpus. A chi-square test reveals that this difference is statistically significant at the $p < 0.05$ level.

Of particular interest in the Lakota case is the fact that the locus of the phonetic environment which impacts the rate of reductive change differs in the Lakota case and in the data on English t/d deletion discussed by Bybee. According to statistical tests conducted for the present study, the phonetic environment preceding the plural marker is not relevant for reductive change, while the environment following it (/*c*/ vs. non-/*c*/) is highly significant. This is the reversal of the English case, where the preceding environment triggers reductive change more strongly than the following environment does.

However, phonology aside, the decisive factor that stimulates the appearance of reduced variants rather than the full form of the Lakota plural marker is discourse frequency, as can be expected on the basis of research such as Bybee (2001), Jurafsky et al. (2001), or Pustet (2008). One of the widely accepted recent findings in the theoretical context of usage-based approaches to languages is that reductive change, i.e. the shortening of the phonetic representations of linguistic items, proceeds at different rates depending on the discourse frequency of the collocations in which individual linguistic items occur. For instance, a grammatical marker that is used with verbs, and which has both a full and a reduced form,

should, in any given corpus, exhibit a higher percentage of reduced forms in combination with high-frequency verbs than with low-frequency verbs. This prediction is borne out by the data on the corpus distribution of the alternants of the Lakota plural marker. The highest-frequency verb in the corpus found in collocation with the plural marker, *káġa* 'to make', exhibits 95.2% reduced plural forms (*-p* 64.3%, *-b* 9.5%, *-w* 16.7%, *-o* 4.8%) vs. only 4.8% occurrences with the full form *-pi*. The overall analysis of the data is done on the basis of frequency groups which subsume elements in collocation with plural markers according to their corpus frequency: elements with frequencies between 1 and 9 constitute a group, as do elements with frequencies from 10 to 19, 20 to 29, 30 to 39, and so on. The average percentage of *-pi* vs. its reduced forms per group shows a skewed distribution -- on the whole, the likelihood of reductive change in plural markers increases as the discourse frequency of collocations containing a plural marker increases, cf. diagram (1). The average reduction rate of the lowest-frequency group, i.e. elements with a corpus frequency below 10, is 63.0%.

Diagram 1: Correlation of frequency and reduction of plural marker in Lakota



horizontal axis: frequency (not presented proportionally for graphical reasons) vertical axis: % reduction of plural marker

Regarding the locus of the context which interferes with the rate of reductive change, it is interesting to note that while the preceding context is highly significant, as diagram (1) indicates, there is no statistical correlation with reduction rates and the properties of the linguistic items following the Lakota plural marker. Thus, the collocational component which precedes the plural marking slot influences the structural realization of this element, while the component following it does not.

3. Conclusion

The Lakota data bear on an important issue that has emerged in the discussion of language change in progress over the past couple of years. According to Labov (1994), sound change gradually spreads from word to word, or from grammeme to grammeme, under the influence of phonological factors. Thus, there are phonological environments that probabilistically facilitate and foster a specific sound change, and there are phonological environments that delay this sound change. In the works of Bybee (e.g. 2001), sound change is, likewise, portrayed as proceeding on a word-to-word, grammeme-to-grammeme basis, but the rate of change and the locus of change are mainly determined by discourse frequency, rather than by phonological factors. However, Bybee's model leaves some room for the possibility that phonological conditioning may interact with frequency to produce different probabilities of change in the linguistic items which are targeted by a specific sound change (Bybee, 2007: ch.11). The issue to be resolved by future research is the status of frequency vs. phonological conditioning in sound change. Are both factors always equally active in sound change? If not, what determines the relative importance of each factor in individual cases? The Lakota data presented in this study do, of course, not provide answers to these profound and complex questions, but they document a new case of an interplay of frequency with other factors, i.e. morphosyntax and phonology, as the driving forces behind language change.

Abbreviations

CAUS = causative, DEF = definite, IDF = indefinite, L = locative prefix, PL = plural.

REFERENCES

Bybee, J. L. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. L. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Jurafsky, D., A. Bell, M. Gregory, and W. D. Raymond. (2001). "Probabilistic relations between words: evidence from reduction in lexical production." In J. L. Bybee and P. Hopper (eds.), *Frequency and the emergence of linguistic structure*, (pp. 229-254). Amsterdam: Benjamins,

Labov, W. (1994). Principles of linguistic change. Volume 1: Internal factors. Oxford: Blackwell.

Pustet, R. (forthcoming). Discourse frequency and language change: the case of Lakota *etą(hą)/-tą(hą)* 'from'. in: Cuyckens, H.; W. de Mulder; M. Goyens; T. Mortelmans (eds.), *Variation and change in adpositions of movement*. Amsterdam: Benjamins [SLCS].

Pustet, R. 2008. Discourse frequency and the collapse of the adposition vs. affix distinction in Lakota. in: Seoane, E. & M. J. López-Couso (eds.), *Theoretical and empirical issues in grammaticalization*, (pp. 269-292). Amsterdam: Benjamins [TSL].

The appraisal of lexical content in ESP coursebooks against corpus-driven and frequency vocabulary lists

CAMINO REA RIZZO

Universidad Politécnica de Cartagena/Murcia

Abstract

Vocabulary competence is an essential component of the subject-specific literacy that new graduated engineers are requested to acquire during their training period. Therefore, teaching materials must be endowed with the lexical content which provides the most beneficial results and assures the development of communicative skills: academic and technical vocabulary of the domain. However, there seems to be a general lack of adequate ESP coursebooks capable of meeting students' needs.

The goal of this work is to appraise the lexical content of three coursebooks which have ever been used in teaching English for telecommunication engineering. First, the books will be assessed to evaluate to what extent they reflect the language deployed by the discourse community as drawn from corpus-based research and against a corpus-driven telecommunication word list. Second, the lexical input will be examined in terms of word frequency bands and against the Academic Word List. The results of the analysis reveal some deficiencies in the presentation of vocabulary which show the unsuitability of the manuals examined in several respects: lexical choice, quantity, repetition and balance.

Keywords: ESP coursebooks, specialized and academic vocabulary, word frequency bands

Resumen

Entre las competencias que los nuevos graduados en ingeniería deben adquirir durante su formación se encuentra la capacidad de comunicarse en una lengua extrajera. El uso idiomático del vocabulario especializado característico es un componente esencial. Por lo tanto, los libros de texto destinados a la enseñanza del inglés especializado deben aportar el contenido léxico que proporcione los resultados más beneficiosos y asegure el desarrollo de las destrezas comunicativas. Sin embargo, tales libros no suelen cubrir las necesidades de los alumnos.

En este trabajo se evalúa el contenido léxico de tres libros de texto utilizados en la enseñanza de inglés de telecomunicaciones. Primero, se comprobará en qué medida reflejan el lenguaje utilizado por la comunidad discursiva de acuerdo con resultados derivados de investigaciones basadas en corpus y frente al repertorio léxico especializado de telecomunicaciones extraído mediante análisis de corpus. Después, el vocabulario se analizará en relación a diferentes bandas de frecuencia léxica y frente a la lista de vocabulario académico. Las deficiencias halladas en la presentación del vocabulario muestran que los manuales analizados no son apropiados en diferentes aspectos: selección léxica, cantidad, repetición y equilibrio.

Palabras clave: libros de inglés para fines específicos, vocabulario especializado y académico, bandas de frecuencia léxica.

1. INTRODUCTION

The need for ESP teaching is felt everywhere. The European Union (2003) has recently established the bases for the teaching and learning of foreign languages through its communication for the promotion of foreign language learning, as an answer to the demands

of nowadays' society. Strong pressure from the industry, together with the insistence of employers and institutions, on the importance of English and other foreign languages, resulted in the revision of the accreditation criteria for technical universities in the late 90s, making European universities introduce training in social and communicative skills in their curricula (Räisänen and Fortanet, 2008). As a consequence, the majority of science and technology programmes currently include some kind of language courses, namely technical writing and oral presentation.

Teaching methodologies are also being revised and changed so as to meet the requirements established by the new Higher Education system within the framework of the Bologna Process. Multidisciplinarity is strongly promoted, and there is an increasing awareness of the need to integrate content and language in the domain-specific courses. However, according to a survey on the ways in which the Bologna Reform has been implemented and its effects on ESP teaching (Räisänen and Fortanet, 2008), most universities offer ESP courses as a separate subject independently of the content courses. Furthermore, most ESP teachers claim to base their teaching on self-designed materials drawn and adapted from up-to-date publications, since there is a general lack of adequate ESP coursebooks capable of meeting students' needs.

In the realms of technology and telecommunication, the scarcity of ESP textbooks is singularly surprising regarding the leading role that English performs today as a lingua franca, and the number of universities which offer a degree in telecommunication engineering. On the contrary, it may not seem so surprising if we consider the accelerated development of the field, where there is a constant creation of new concepts and the subsequent coinage of new terms that might become obsolescent overnight. In addition, the domain might seem too specific to raise publishing companies' interest. All in all, a representative common core vocabulary of the subdisciplinary fields in telecommunications, properly sequenced in a coursebook, would be appreciated and rewarding for teaching, learning and testing. Corpus Linguistics tools and analyses can provide better quality information about how words are used in specific contexts on the basis of frequency data required for measures of vocabulary size.

Once the demand of learning specialized English and the lack of appropriate teaching materials have been pointed out, the aim of the this work is to appraise the lexical content of three coursebooks which have ever been used in teaching English for telecommunication engineering at Tertiary Education. First, the books will be assessed to evaluate to what extent they reflect the language deployed by the discourse community as drawn from corpus-based

research and against the corpus-driven telecommunication engineering word list. The second part of the analysis will examine the lexical input in terms of word frequency bands and against the Academic Word List (Coxhead, 2000).

2. FOCUS ON VOCABULARY

After finishing Secondary Education in Spain, students are assumed to have attained an intermediate level of English, corresponding to the B1 on the scale of the Common European Framework of Reference for Languages (CEF). Those early independent users of the language are characterized by the ability to maintain interaction and deal flexibly with topics focusing on daily activities and events (Council of Europe, 2001). The type of texts they can understand comprises mainly high frequency everyday language whose lexical content is consistent with the most frequent words in the general language. Therefore, B1 students have presumably gained an overall command of the 2000 most frequent words registered in the *General Service List of English Words* (West, 1953) or in a general list based on corpus frequency data like Nation's frequency lists drawn from the British National Corpus (Nation, 2004). In fact, 2000 words seems to be the most commonly cited initial goal for second language learners (Criado and Sánchez, 2009; Nation, 2001; Schmitt, 2001; Schonell, Meddleton and Shaw, 1956; West, 1953).

The successive goal set to gain a better command in vocabulary may be directed towards the next frequency band of general vocabulary or, for learners who are going to move on to special purposes study, it is advisable to concentrate on specialized vocabulary (Nation and Hwang, 1995). The rationale behind introduces the variable of text coverage as an index of utility in addition to the thrust that some words are particularly useful in a specific topic area. A foundation of higher-frequency general service vocabulary of close to 2000 words is seen as providing a solid basis for moving into more advanced specific scenarios, where the study of specialized vocabulary is better rewarded since it offers a greater text coverage than the words in the next frequency band (Nation and Hwang, 1995; Nation and Waring, 1997; Schmitt, 2001).

Research has proved that different vocabulary sizes are needed to understand different types of text depending on several factors such as genre, register, text length, topic, and number of authors. According to Laufer (1989), it is necessary to understand 95% of the words in a text in order to get a fair comprehension. This rate means that the reader will encounter one unknown word in every 20 running words, that is, around one word per two

lines. In the case of novels written for teenagers, Hirsh and Nation (1992) found out that a vocabulary size of 2600 word families is required to gain 96% coverage. In turn, a range from 2000 to 4000 word families plus proper nouns and marginal words is demanded to get 95% coverage of different television programs (Webb and Rodgers, 2009). Regarding academic discourse, a 4000 family word size vocabulary could reach 95%. This vocabulary should be made of 2000 high-frequency general service words, about 570 general academic words and 1000 or more technical words, proper nouns and low-frequency words (Nation, 2001). Independently, each kind of vocabulary covers a distinct fraction. The first 1000 high-frequency words do the bulk of the work reaching some 77%; the second frequency band adds around 5%; next, the academic vocabulary covers about 9% and finally, technical vocabulary renders another 5%. The remaining uncovered fraction of the text may be filled by a massive amount of low frequency words pertaining to virtually any specific domain, including proper names and general words seldom deployed in ordinary language. Unlike the other types of vocabulary, low frequency words are relatively unpredictable.

In Tertiary Education, where students need to cope with academic English and become competent language users in their discourse community, and on the assumption that general service words have been assimilated, teaching materials should introduce both academic vocabulary and technical vocabulary of the domain. On the one hand, academic vocabulary will facilitate the acquisition and understanding of the subject content as conveyed by the university community. On the other hand, it is arguable that terminology learning is a process running in parallel to subject learning. Still, content subject lectures are delivered in Spanish so that explicit teaching of technical vocabulary is worthwhile, especially how they operate and are idiomatically used in the register.

Vocabulary competence is an essential component of the subject-specific literacy that new graduated engineers are requested to acquire during their training period. Therefore, teaching materials must be endowed with the lexical content which provides the most beneficial results and assures the development of communicative skills. In this work, such relevant lexical content is equated with the most frequent and representative words in the domain which manage to cover the corresponding 5% of specialized vocabulary in technical texts.

Regarding technical vocabulary, it is important to revise the concept, since it might be considered as a current controversial issue due to the implications it may carry. In this respect, technical vocabulary or terminology alludes to the collection of terms which designate concepts and notions specific to a particular discipline or field of human activity.

Within specialized languages, there are lexical units whose use is restricted to the discipline, units from the general language or other registers which activate a different meaning in the domain, and even units with little or no specialization whose meaning is accessible through their use outside the field. Those facts lead us to incline towards the conception of specialized vocabulary very much in line with Nation's, who asserts that there are different categories or *"degrees of 'technicalness' depending on how restricted a word is to a particular area"* (Nation, 2001:198). Moreover, Hyland and Tse (2007) firmly reject the division between academic and technical vocabulary in favour of regarding vocabulary in different disciplinary contexts as *"a cline of technically loaded or specialized words ranging from terms which are only used in a particular discipline to those which share some features of meaning and use with words in other fields"* (Hyland and Tse, 2007: 249).

In keeping with this vocabulary approach, a word list of the specialized vocabulary of telecommunication engineering English was made (Rea, 2008) to the purpose of designing a suitable syllabus with the appropriate lexical input. The list succeeds in covering the corresponding 5% of the words in technical texts, so that it could be a potential learning target that teaching materials should incorporate.

All in all, the following research questions are formulated:

1. Are there significant differences between the specialized vocabulary of telecommunication engineering and the lexical content provided in ESP coursebooks?
2. What kind of vocabulary is found in ESP coursebooks?

3. METHODOLOGY

3.1. Materials

3.1.1. The lexical repertoire of telecommunication engineering

The telecommunication engineering word list (TEWL) (Rea, 2008) operates as the reference list against which the lexical content in coursebooks is compared. TEWL is a corpus-driven list generated adopting a corpus-comparison approach and applying statistical tests which quantify occurrence probability and representativiness of lexical units. Additionally, the words comply with the quantitative conditions which determine a technical term as defined by Chung (2003): a lexical unit must be at least 50 times more frequent in the specialized register than in general language.

The words pertaining to the list correspond to the most salient, central and typical specialized lexical units in telecommunications, in accordance with the notion of specialized

vocabulary stated above. Therefore, the list comprises both words whose use is restricted to the subject-domain, and those which activate a specialized meaning in the discipline even though they may be also used in other fields or in general language.

TEWL holds 402 specialized families plus 1,017 individual specialized forms that amount to 2,747 types altogether. They are all found within the range of the 1000 most statistically significant word families as drawn by the comparison of the general language corpus LACELL (20 million words) and the specific corpus TEC (of 5.5 million words). The Telecommunication Engineering English corpus is a fairly representative sample of the professional and academic written language, which covers the main divisions of the realm (Electronics; Computing Architecture and Technology; Telematic Engineering; Communication and Signal Theory; Materials Science; Business Management; and System Engineering) and two branches of expertise (Communication Networks and Systems; and Communication Planning and Management). The language samples originated in native and non-native English and were extracted from a wide gamut of sources in professional and academic settings.

The specialized list was subjected to a validation process through a test consisting in the filtering of academic and professional texts with the available general service and academic word lists, in addition to the new specialized vocabulary list. The process was carried out in two stages: first with a series of texts belonging to TEC and second, with a series of texts alien to the corpus. In both cases the samples were selected at random but collecting texts of different size.

The combination of the three lists managed to cover an average of 95.6% of the samples from TEC, that is, 4.4% uncovered text, being 89% the minimum coverage rate and reaching a maximum of 98.8% in the best case. Similarly, for the second series, the average coverage increased slightly to 95.8% that is equivalent to 4.2% of uncovered text, being 97.7% and 92.2% the maximum and minimum coverage rates respectively.

As a result, the outcome of the validation analyses proved that TEWL accounted for a successful percentage of the running words in the specialized texts both from and alien to the corpus. Consequently, evidence suggests that the list could be set as a guide for planning the vocabulary component in an ESP course design, as a target for learning purposes and as a means of assessing vocabulary competence in the register.

3.1.2. ESP coursebooks

No coursebook designed for the general market will be entirely suitable for a specific group of learners, but the aim is to find the best possible fit. So three books related to

telecommunication domain or any of its constituents were selected for lexical analysis: *English for Electronics* (Álvarez et al., 1990), *English for the Telecommunications Engineering Industry* (Comfort et al., 1998) first edited in 1986, and *Information Technology* (Glendinning and McEwan, 2002).

All of them are designed for an intermediate level of English, aim at non-native speakers of the language who are working or being trained for subsequent careers in the field, and place a major focus on reading comprehension skill. The thematic content varies considerably as they are intended for distinctive branches but closely connected with telecommunications.

English for Electronics (hereafter B.1_EFE) is divided into four blocks of five units, being the last one devoted to revision. Every unit introduces two short specialized texts followed by reading comprehension activities. The topics are electronic components, electronic materials, circuit diagrams, security norms, integrated and printed circuits, measurement devices, communication history, television, phone, computing systems, office automation and troubleshooting.

English for the Telecommunications Engineering Industry (hereafter B.2_TEI) consists of 15 main units structured in three blocks of 5 units (Networks, Transmission, Switching, Computer communications and Radio communications) plus a revision unit. Lexically, the blocks are sequenced to first introduce or revise telecommunication concepts and terminology, then exploit the technical information and finally use the technical input in management-related contexts.

Information Technology (hereafter B.3_IT) is divided into 25 topic-based units each covering a key area of IT: computer users, architecture, applications and peripherals; operating systems, multimedia and application programs; networks, internet and www; communication systems and data security; software engineering, people in computing, and developments and the future of IT. Authentic short texts introduce new content and longer ones are intended to stretch the students.

3.2. Analysis

3.2.1. Text analysis against TEWL

Texts from coursebooks have been digitalized and processed by using WordSmith Program (Scott, 1998) so that the numerical data obtained allow to perform statistical calculations and contrast results. The processing has been performed both individually for every single book and all together, considering the three samples as a whole. The corresponding basic statistical

information is displayed in table 1. Likewise, four frequency lists are retrieved to be used in two stages: (i) the extraction of the most significant words, that is, to obtain keywords, and (ii) the comparison with the TEWL in an Excel spread sheet to check to what extent they coincide.

Table 1: Basic statistical information

	Book 1: EFE Álvarez, 1990	**Book 2: TEI** Comfort, 1998	**Book 3: IT** Glendinning, 2002	**Together**
Tokens	5,383	11,777	11,650	28,810
Types	1,442	2,403	2,418	4,332
Type/Token Ratio	26.79	20.4	20.76	15.04
Standardised Type/Token	42.8	43.98	41.41	42.71

3.2.2. RANGE-based analysis: General Service and Academic word lists.

Range software (Nation and Heatley, 2002) is the program used to classify the type of vocabulary presented in the coursebooks. Range can list the words occurring in a text according to their frequency against different frequency bands based on the British National Corpus or against the General Service list and the Academic Word List.

RangeGSL/AWL was preferred to Range/BNC because in Tertiary Education, students are expected to be in a B1 on the scale of CEF which means that they should be ready to learn the vocabulary needed to communicate in an academic setting and the most significant specialized vocabulary of the discipline. By running the texts through Range, we can find out how many words come from the most frequent first and second thousand words of the GSL, those from the AWL, and how many words do not come from any of these lists. In doing so, the academic and new vocabulary introduced in the books can be detected. Conversely, if Range/BNC was chosen for the analysis, the academic vocabulary could not be highlighted because the BNC first 2000 words contain many from the AWL (Nation, 2004).

4. RESULTS AND DISCUSSION

4.1. The specialized repertoire in ESP coursebooks

Once a specialized vocabulary word list is available, ESP textbooks are analysed to assess whether they actually incorporate the specialized repertoire and to what extent they reflect the

words from the list. Table 2 illustrates the procedure carried out for contrasting. It shows the word frequency in both corpora (F:TEC and F:LACELL), the keyness index, and how every word distributes throughout the three books, that is, word range (B.1_EFE, B.2_TEI and B.3_IT). The keyness index marks how significantly higher the frequency of a word is in the specific corpus as compared to the general one, yielding the most representative and relevant words of the domain on statistical basis – Log likelihood test has been applied. Given that TEWL is not arranged in word families but according to the keyness index scored by each form individually, the most relevant member of the family is highlighted in bold.

Table 2: Specialized vocabulary in ESP coursebooks

N	Specialized vocabulary	F: TEC	F: LACELL	Keyness	B1. EFE	B2. TEI	B3. IT
1	**NETWORK**	16.649	1.686	41.784	1	2	3
2	**DATA**	14.613	2.787	31.852	1	2	3
3	**SIGNAL**	7.022	641	17.922	1	2	3
4	**SYSTEMS**	9.479	3.000	17.377	1	2	3
5	**IP**	5.239	20	16.182	-	-	3
6	SYSTEM	12.624	8.707	14.831	1	2	3
7	**PROTOCOL**	4.742	139	13.677	-	-	3
8	**ROUTER**	3.910	25	11.974	-	-	-
9	**WIRELESS**	4.083	171	11.454	1	-	3
10	**FREQUENCY**	4.551	455	11.439	1	2	3
11	ROUTING	3.542	40	10.690	1	-	3
12	**LAYER**	4.425	569	10.604	1	-	3
13	**MODEL**	5.895	2.290	9.860	1	2	3
14	**INTERNET**	4.504	910	9.651	-	-	3
15	INTERFACE	3.526	207	9.557	1	2	3
16	**BANDWIDTH**	3.119	20	9.551	1	2	3
17	**PACKET**	3.577	251	9.485	-	2	3
18	**CIRCUIT**	3.932	525	9.348	1	2	-
19	**SERVER**	3.574	362	8.963	-	-	3
20	**SOFTWARE**	4.575	1.412	8.470	1	2	3
21	**SIMULATION**	2.817	73	8.189	-	-	-
22	**VOLTAGE**	2.945	220	7.743	1	-	-
23	**OPTICAL**	2.822	164	7.658	1	2	3
24	**TRAFFIC**	4.345	1.615	7.420	1	2	3
25	**ALGORITHM**	2.799	229	7.264	-	-	3
26	**NODE**	2.822	294	7.042	-	-	3
27	**FILTER**	2.627	247	6.671	-	-	-
28	PACKETS	2.308	166	6.099	-	-	3
29	NODES	2.361	232	5.951	-	-	3
30	ROUTERS	1.891	4	5.876	-	-	-
31	**WEB**	2.978	791	5.838	-	-	3
32	PROTOCOLS	1.996	52	5.800	-	-	3
33	CIRCUITS	2.122	131	5.717	1	2	3
34	SERVICE	7.085	7.291	5.707	1	2	3
35	**FIBER**	1.869	19	5.658	1	-	-
36	**ETHERNET**	1.897	37	5.601	-	-	3
37	SERVICES	6.161	5.742	5.524	1	2	3
38	**TCP**	1.717	12	5.248	-	-	3
39	**CONFIGURATION**	1.885	106	5.134	1	-	-
40	HARDWARE	2.257	423	4.939	1	2	3
41	ALGORITHMS	1.777	102	4.828	-	-	3
42	**ATM**	1.639	35	4.817	-	-	-
43	**TYPE**	4.612	3.750	4.716	1	2	3
44	**LAN**	1.481	27	4.387	1	2	3
45	**ARCHITECTURE**	2.581	998	4.325	-	2	-
46	**PATH**	2.700	1.210	4.196	1	2	3
47	**OSPF**	1.284	0	4.027	-	-	-
48	**WAVELENGTH**	1.352	24	4.010	-	2	-
49	**EXAMPLE**	6.379	8.009	3.992	1	2	3
50	**JAVA**	1.344	34	3.912	-	-	-

On initial consideration, it is observed that not all the words from the specialized repertoire occur in the three books. It is particularly remarkable that such relevant units as *protocol, internet, encryption, wavelength, antenna, switch, buffer, carrier, telecommunication, Bluetooth, multiplexing*, etc… appear in only one book or in none of them like *router, simulation, filter, ATM, broadband*, *JAVA, cell, VHDL, OSPF, multicast, WLAN, impedance, fuzzy,* etc. With respect to range, only 52 words from the list occur in the

three manuals, 67 are present in two books and, out of the 236 which occur in one, 63 appear in B.1_EFE, 58 in B.2_TEI and 115 in B.3_IT.

Of the 2,747 individual forms constituting the specialized repertoire, the first book includes 154, the second one 158 and the third one 214. Those data can be interpreted in two ways: (i) in relation to the number of content words and the sheer number of tokens in each book (Figure 1) or (ii) with respect to the total set of forms on the specialized list (Figure 2).

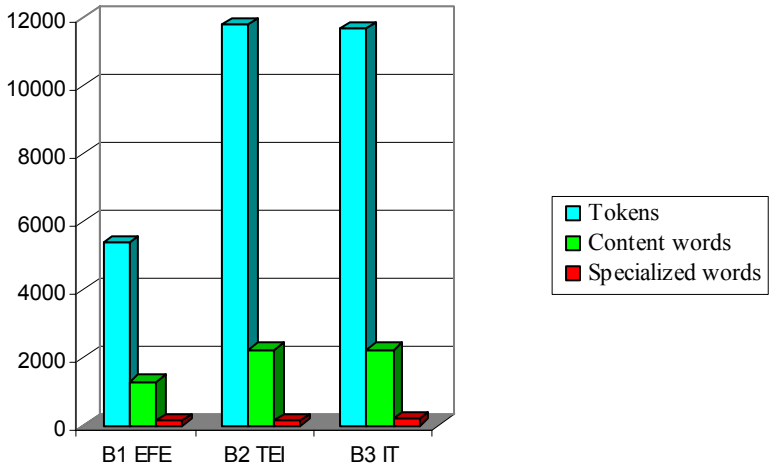


Figure 1: Tokens - content words - specialized forms in books

Figure 1 depicts the relationship among tokens, content words and specialized forms in every book. The language samples taken from textbooks were filtered by removing all functional words, so that the deriving frequency lists only contain notional words (B.1_EFE: 1,315, B.2_EI: 2,231 and B.3_IT: 2,253). The proportions of content words which coincide with specialized forms from TEWL are 11.7%, 7% and 9.4% for B.1_EFE, B.2_TEI and B.3_IT respectively. The percentages substantially decrease to 2.8%, 1.3% and 1.8% by computing the proportion of specialized forms in the materials with regard to the sheer size of tokens. Thus, taking into account the relationship existing between the different word lists and the stretch of text they cover, the results reveal a remarkable underrepresentation of specialized vocabulary in the analysed books. The percentage of specialized forms should be purposefully increased in books up to 5% so that learners get exposed to a broader variety of the typical vocabulary.

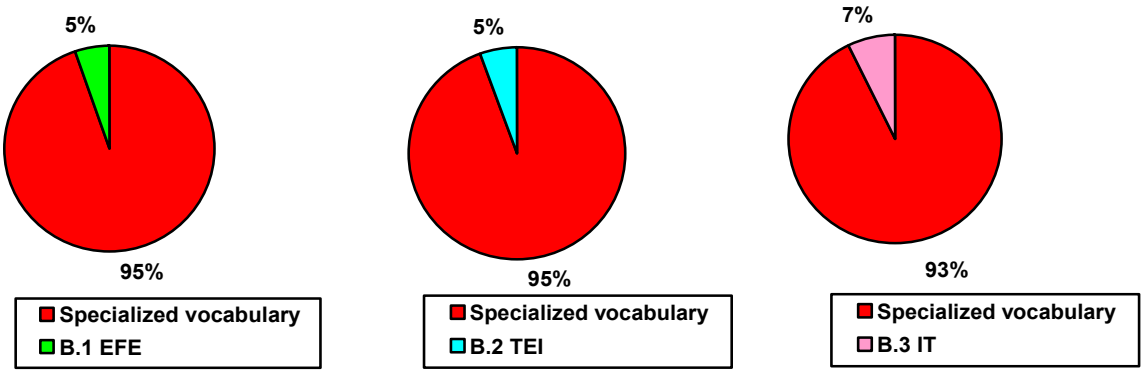


Figure 2: Specialized vocabulary in coursebooks

Indeed, the pie charts in figure 2 illustrate the small fraction of specialized vocabulary that coursebooks reflect. They just supply with a 5.6%, 5.7% and 7.7% (B.1_EFE, B.2_TEI and B.3_IT respectively) of the lexical content that learners should master to achieve basic understanding of technical texts. Even so, students could not cope with technical texts to the extent of inferring efficiently the meaning of unknown words from the context. Students need to know about 98% or more of the words in the text (1 new word in 50) before successful guessing can take place (Hu and Nation, 2000; Waring, 2002).

4.2. Word ranges in ESP coursebooks

Allowing for the fact that specialized vocabulary does not seem to partake very much in textbooks, the goal of the second part of the analysis is to find out what kind of lexical content is actually presented, in terms of frequency bands and academic character. The different frequency ranges are connected to three basewords available in Range. Baseword one includes the first 998 word families of the GSL (4,119 types), baseword two comprises the following 988 word families of the GSL (3,708 types), and baseword three embraces the 570 word families of the AWL (3,107 types). The immediate information retrieved from Range is exhibited in table 3a, 3b, 3c for the three books.

Table 3a: Lists coverage in B.1_EFE

Word List	Tokens / %	Types / %	Families
one	3532 / 67.47	648 / 44.29	422
two	357 / 6.82	166 / 11.35	117
AWL	616 / 11.77	255 / 17.43	176
off-lists	730 / 13.94	394 / 26.93	(not given)
Total	5235	1463	715

Table 3b: Lists coverage in B.2_TEI

Word List	Tokens / %	Types / %	Families
one	8232 / 72.13	1064 / 43.38	623
two	575 / 5.04	237 / 9.66	164
AWL	1101 / 9.65	426 / 17.37	271
off-lists	1505 / 13.19	726 / 29.60	(not given)
Total	11413	2453	1058

Table 3c: Lists coverage in B.3_IT

Word List	Tokens / %	Types / %	Families
one	8059 / 70.27	1002 / 40.47	586
two	672 / 5.86	286 / 11.55	198
AWL	1140 / 9.94	439 / 17.73	261
off-lists	1597 / 13.93	749 / 30.25	(not given)
Total	11468	2476	1045

Data show that texts in books slightly reduce the use of the first range words (904 types on average, out of 4,119), that word selection from the second range is even narrower (229 types on average, out of 3,708), probably due to the topic, but they altogether reach a reasonable coverage of the running words (75.86% on average). Perhaps, the remaining 24.14% seems to be somewhat high considering the lexical burden it may imply: either the recurrence of the same types or the occurrence of new low frequency words.

On the one hand, the books favour the teaching of a particular set of academic vocabulary (B.1: 176, B.2: 271 and B.3: 261 word families) which covers the usual percentage (9%) and sometimes exceeds it (B.1_EFE 11.77%). Nevertheless, of the types falling in AWL, 58 occur twice in B.1, 31 in B.2 and 70 in B.3. Moreover, it is striking to

find 108, 350 and 267 types with frequency 1 in B.1, B.2 and B.3 respectively. There are a few types whose frequency is higher than 10. It means that the real target set at academic vocabulary is not so challenging and focus basically on no more than 18 words in the best case (see table 4) or, conversely, the vocabulary goal is extremely ambitious if students are intended to learn that number of words with only one encounter. Research indicates that it takes between 5 and 16 meetings to cement a word in memory (Nation, 2002).

Table 4: Higher-frequency academic types

	Academic type and frequency.
B.1 EFE	*components 22, computer 20, devices 17, computers 16, integrated 14, communications 13, device 13, area 12, process 11, elements 9, functions 9*
B.2 TEI	*network 53, data 41, area 21, technology 13, range 9, available 8, period 8*
B.3 IT	*data 102, computers 23, network 21, computer 20, access 19, layer 18, devices 17, media 16, available 14, file 14, protocol 14, technology 13, code 12, intelligent 11, files 10, tape 10, text 10, computing 9, document 9*

On the other hand, as far as off-list words are concerned, their representation is similar to academic ones in the sense that low frequency words greatly outnumber higher frequency words. Among off-list words, there ought to be specialized vocabulary and low frequency words in equitable proportions, but there is an average of 68% of types whose frequency is 1. In general, the coursebooks do not allow for enough repetition of specialized words to facilitate learning, as only 24, 29 and 74 types in B.1, B.2 and B.3 respectively, occur more than 4 times. Thus the vocabulary target of the different books concentrates basically on the words in table 5:

Table 5: Higher-frequency off-list types.

	Off-list type and frequency.
B.1 EFE	*circuit 27, circuits 24, electronic 18, transistors 15, resistors 9, digital 8, electronics 8, PCB 8, satellites 8*
B.2 TEI	*subscriber 27, switching 26, traffic 26, telecom 25, telecommunications 23, Swedish 22, digital 21, satellites 20, Prestel 18, circuits 16, Indian 14, subscribers 14, cable 13, India 13, electronic 12, PABS 12, satellite 12, CCS 11, LAN 10, Telex 10, circuit 9, mobile 9, MTX 9, TV 9*
B.3 IT	*software 42, storage 38, cache 31, internet 21, server 19, XML 18, web 17, digital 14, disk 13, html 12, Linux 12, TCP 12, bandwidth 11, email 11, Mac 11, IP 10, offline 10, online 10, virus 10, packets 9*

Finally, if the distinction between general, academic and technical is set aside, and vocabulary is instead considered as a cline of words technically loaded whose frequency in the subject domain is significantly higher than in general language, the evaluation of lexical content in books would improve. After generating and comparing the corresponding keyword lists, a matching of 35% among the first 100 most representative words in TEWL takes place (*data, network(s), circuit(s), digital, transmission, device(s), system(s), user, communications, bandwidth, components, information, TCP, services, LAN, IP, optical, server, layer, traffic, protocol, frame, signals, processing, frequency, internet, technology, applications, software*). Although the specialized vocabulary input in coursebooks is not satisfactory, at least the lexical units introduced are found among the most relevant ones in telecommunications.

5. CONCLUSION

The outcome of the analysis carried out in this paper reinforces ESP teachers' general sense of disappointment about the coursebooks available for teaching the specialized register. The deficiencies observed in the presentation of vocabulary show the unsuitability of the manuals examined in several respects: lexical choice, quantity, repetition and balance.

Results indicate that there are significant differences between the lexical selection of the specific discourse community and the lexical input introduced in textbooks, where specialized vocabulary is actually underrepresented covering only a small fraction of the most significant words. In addition, the recurrence of the specialized units is not high enough to

facilitate learning. Furthermore, an excess of low frequency words has been detected, leading to an unbalance in the desired distribution of vocabulary which allows inferring meaning from context. In sum, the lexical content in the textbooks is not satisfactory enough to meet students' needs.

Many universities which are still adapting to Bologna directives are also actively involved in educational projects and workshops with the aim to develop teaching materials in accordance with the renewed methodology. This is quite an occasion for appealing to publishing companies to intervene and cater for adequate textbooks which agree with the new demands and linguistic needs as defined by European standards.

REFERENCES

Alvarez, I., Lerchundi, Mª. y Moreno P. (1990). *English for Electronics*. Madrid: McGraw Hill.

Comfort, J., Revell, R., Simpson, I. Stott, T. and Utley, D. (1986). *English for the Telecommunications Engineering Industry*. Oxford: Oxford University Press.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. URL http://www.coe.int/T/DG4/Linguistic/CADRE_EN.asp

Chung, T. (2003). "A corpus comparison approach for terminology extraction". *Terminology, 9,* 2.

Coxhead, A. (2000). "A New Academic Word List". *TESOL Quarterly 34*(2), 213-238.

Criado, R. y Sánchez, A. (2009). Vocabulary in EFL Textbooks. A Contrastive Analysis against Three Corpus-Based Word Ranges. En A. Sánchez y P. Cantos (Eds.), *A Survey on Corpus-based Research/Panorama de investigaciones basadas en corpus,* (pp. 862-875). Murcia: Editum.

European Union. (2003). Promoting Language Learning and Linguistic Diversity: An Action Plan 2004-2006. Brussels, URL http//europa.eu.int/comm/education/doc/oficial/keydoc/actlang/act_lang_en.pdf.

Glendinning, E & McEwan, J. (2002). *Information technology*. Oxford: Oxford University Press.

Hirsh, D. & Nation, P. (1992). 'What vocabulary size is needed to read unsimplified texts for pleasure?' *Reading in a Foreign Language,* 8*,* 689-696.

Hwang, K. & Nation, P. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System,* 23(1), 35-41.

Hyland, K. & P. Tse (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly,* 41(2), 235-253.

Hu, M. & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language,* 13(1), 403-430.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Normand (Eds.), *Special Language: From Humans Thinking to Thinking Machines,* (pp. 316-323). Clevedon: Multilingual Matters.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P. & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts. URL http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Nation, P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing,* (pp. 3–13). Amsterdam: Benjamins.

Räisänen, C. & Fortanet, I. (2008). The state of ESP teaching and learning in Western European higher education after Bologna. In *ESP in European Higher Education. Integrating Language and Content.* John Benjamins.

Rea, Camino (2008). *El inglés de las telecomunicaciones: estudio léxico basado en un corpus específico*. Tesis doctoral. Murcia: Servicio de Publicaciones de la Universidad de Murcia. URL http://www.tesisenred.net/TDR-0611109-134048/index_cs.html.

Schonell, F., Meddleton, I. & Shaw, B. (1956). *A study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.

Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Scott, M. (1998). *WordSmith Tools Manual version 3.0.* Oxford: Oxford University Press.

Waring, R. (2002). Basic Principles and Practice in Vocabulary Instruction. *The Language Teacher Online.* URL http://www.jalt-publications.org.

Webb, S. & Rodgers, M. (2009). Vocabulary demands of television programs. *Language Learning,* 59(2), 335-366.

West, M. (1953). A General Service List of English Words. London: Longman.

Does valency theory offer a holistic approach to teaching language?

RENATE REICHARDT

University of Birmingham

Abstract
Valency theory, the property of a word to combine with or demand a certain number of elements in forming larger units, offers an approach to investigate the interface of local grammar and lexis, which works for monolingual, as well as bilingual analysis. For exemplification, a bilingual (English and German) corpus study of the verb CONSIDER is undertaken to investigate whether valency patterns relate to word meaning, i.e. translation equivalents. The findings will be useful in first and second language teaching in that awareness of local grammar will contribute to the understanding of sentences, as well as the production of sentences.
Keywords: valency, local grammar, syntax, contrastive linguistics, parallel corpus, meaning interpretation, language teaching

Resumen
La teoría de valencia, es decir, la propiedad de una palabra para combinarse con cierto número de elementos con el fin de formar unidades mas grandes, ofrece un método para investigar la interfaz del léxico y gramática local, lo cual funciona para análisis monolingües y bilingües. Esta investigación analiza el verbo CONSIDERAR y aplica un estudio bilingüe (Ingles y Alemán) para investigar si los patrones de Valencia hacen relación al significado de las palabras, por ejemplo, equivalentes de traducción. Los resultados serian útiles para la enseñanza del primer y segundo lenguaje, ya que la gramática local contribuiría al entendimiento de oraciones, así como también a la producción de ellas.
Palabras clave: valencia, gramática local, sintaxis, lingüísticas en contraste, corpus paralelo, interpretación de significado, enseñanza de lenguaje

1. AIMS AND OUTLINE

Over the last 30 years the distinction between lexis and grammar as separate areas of study has moved towards a theory which emphasises the strong interaction between the two (Singleton, 2000: 17). Concurrently, there has been an increased emphasis on the study of phraseology and phrase patterns, at the expense of sentence grammar, in the English language classroom (Granger, 2009). However, if we accept that

- learners will always relate a new language to previous knowledge of language, mainly their native language (Lightbown and Spada, 1999: 45; Nunan, 1999: 40),
- word meaning depends to a great degree on the surrounding words (Firth, 1957: 11; Sinclair, 1991: 110) and
- language use is a creative process which requires negotiation amongst its users (Teubert, 2004: 98)

Then, as a result, it appears to be important to provide language learners with a more holistic approach to address the lexis-syntax continuum.

The argument pursued in this paper it that word meaning, expressed as a translation equivalent or synonym, depends on the syntactic environment, *i.e*. the local grammar of words, and valency analysis raises awareness of local grammar. The findings have relevance to first and second language teaching as awareness of local grammar aids both the understanding and the production of sentences.

In section two valency theory, its application, and its interface with meaning interpretation is discussed. Valency theory is concerned with the property of words to combine with or demand a certain number of elements in forming larger units (Emons, 1974: 34), and therefore offers an approach to investigate the interface of local grammar and lexis, which works for monolingual, as well as bilingual, analysis.

In section three issues regarding the use of corpora and frequency analysis for linguistic study are discussed. It is shown that purely computational frequency searches can be misleading and further investigation into their reliability is required. Using a corpus linguistic approach for grammatical analysis inevitably highlights the difficulties that are often faced by students and scholars when working with authentic texts, where the analysis is often more varied and difficult than textbooks imply (Hoey, 2005: 46).

The final section draws valency theory and corpus investigation together. For exemplification the verb CONSIDER was chosen. In a contrastive corpus study (English and German) the interface of valency patterns and word meaning, i.e. translation equivalents, was investigated. It will emerge that although the valency patterns for CONSIDER to some extent show preference for translation equivalents, there is also a great degree of freedom in the translations (see also Kenny, 2005: 162; Altenberg and Granger, 2002: 21; Aitchison, 1994: 97).

2. Benefits of Valency Theory in Language Analysis

Valency theory is based on dependency relations. Words do not occur randomly in a sentence but form ‘connexions’, *i.e*. words are in relationships with other words syntactically and semantically. Probably because of this dual aspect, Sinclair (2004: 18) envisaged valency grammar to experience “an upsurge of interest in the next few years”.

Valency is a special form of dependency grammar, stating that some dependents, the complements, are specific to individual words, i.e. subclasses of words, while others, the adjuncts, are aspecific, i.e. they can occur with any word. The verb takes a special role as its dependents form a grammatically correct and meaningful sentence (Tesnière, 1980: 26). These hierarchical connexions can be visualised in a valency tree, also called stemma, as shown in figures 1 and 2.

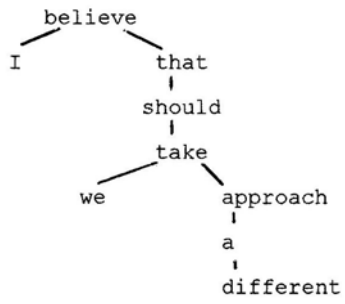


Figure 1: Valency Analysis - English complex clause

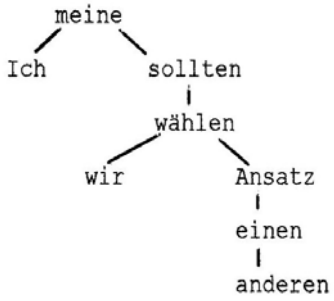


Figure 2: Valency Analysis – German complex clause

The remit of valency theory is to provide an account of the local grammar of verbs, nouns and adjectives, focussing on those features for which general grammar cannot account. It offers an approach to investigate the interface of local grammar and lexis, which works for monolingual, as well as bilingual, analysis.

For example, English distinguishes between subclasses of verbs typically followed by an '-ing' or a 'to-infinitve' form, while in German a division is made between verbs followed by nouns in the accusative form and those typically followed by a dative (table I).

Table I: Examples of subclasses of web valency in German and English

	English	*Example verb*	*Example sentence*
(1)	-ing form	consider	… The Commission therefore does not consider contributing to the FAO … [EuroParl]
(2)	to-infinitive form	need	… Pupils need to aquire facts and skills, but … [BoE]
	German	*Example verb*	*Example sentence*
(3)	Akkusativergänzung	halten (consider)	… und ich persönlich halte ihn für eine sehr gefährliche Person … [EuroParl]
(4)	Dativergänzung	helfen (help)	Er hilft dem Kind. [Helbig and Schenkel 1982:178]

Tab I: Examples of subclasses of verb valency in German and English

The difference between the descriptions of grammatical structures in English and German already becomes noticeable.

However, for a contrastive analysis between languages it is important to use the same 'labels' to identify similarities and differences. Valency complements are suitable for most languages. Table II shows the 11 verb complements identified by Engel (2009: 134) for German, which can also be applied to English (Fischer, 1997: 94-150). We have now established a basis for comparison of verbs and their respective translation equivalents.

Table II: Valency complements in German and English

	German	**English**
Case complements	Subject **sub**	subject **sub**
	Akkusativergänzung **acc**	direct complement **drt**
	Genitivergänzung **gen**	--
	Dativergänzung **dat**	indirect complement **ind**
Prepositional complements	Präpositivergänzung **prp**	prepositional complement **prp**
Adverbial complements	Situativergänzung **sit**	situational complement **sit**
	Direktivergänzung **dir**	directional complement **dir**
	Expansivergänzung **exp**	expansive complement **exp**
Predicative complements	Nominalergänzung **nom**	nominal complement **nom**
	Adjectivalergänzung **adj**	adjectival complement **adj**
	--	modificational complement **mod**
Verbal complements	Verbativergänzung **vrb**	verbal complement **vrb**

The example sentences in table I would then be analysed as follows (verb complements are in square brackets [ ], verb phrases are underlined):

(1) … [The Commission] therefore does not <u>consider</u> [contributing to the FAO] …
<sub vrb$_{ing}$>

(2) … [Pupils] <u>need</u> [to acquire facts and skills], … **<sub vrb$_{to\text{-}inf}$>**

(3) … nd [ich persönlich] <u>halte</u> [ihn] [für eine sehr gefährliche Person] … **<sub acc prp$_{für}$>**

(4) [Er] <u>hilft</u> [dem Kind]. **<sub dat>**

Negations and adverbs, as in example (1), or modal and auxiliary verbs, as in example (5), which can occur with any verb are classified as adjuncts and excluded from the valency pattern. It has to be noted though, that the categorisation of complements and adjuncts, central to valency theory, is probably the most contentious issue in its study (Welke, 1988: 2).

2.1. Complementation Patterns vs. Valency Patterns

The use of complementation patterns in order to identify meaning and to close the gap between syntax and lexis is not unique to valency theory. Two dictionaries, which refer to syntactic complementation patterns for the differentiation of word meanings, ought to be mentioned: the Collins Cobuild English Dictionary (CCED) and the Valency Dictionary of English (VDE). Table III compares the entries for CONSIDER in the CCED and the VDE. The CCED distinguishes three meanings represented by seven patterns, with two meanings sharing the same pattern; the VDE distinguishes two meanings and 13 patterns. The similarities in the representation of the complementation patterns between the CCED and the VDE are striking.

Table III: Comparison *CCDE*, *VDE* and valency complements

CCDE *(Sinclair 1995:345-346)*		**VDE** *(Herbst et al 2004:175-176)*		**Valency complements** *(Fischer 1997:151, 191)*
--	--	think	$[N]_A$ / [by N]	*(passive)*
think	V n to-inf	regard	T: + N_P + to-INF	sub drt $vrb_{to\text{-}inf}$
think	V n n / adj	regard	T: + N_P + N / it + N-$pattern_P$	sub drt ind
			T: + N_P + ADJ / it_P + ADJ-pattern	sub drt adj
think	V n as adj / n	regard	T: + N_P + as ADJ / V-ing / it + as ADJ-pattern	sub drt adj_{as} sub drt nom_{as}
		think regard	T: + N + as N / it + as N-pattern	
think	V that	think regard	D: + (that)-$CL_{P:it}$	sub vrb_{that}
think about (carefully) think about (intention)	V n	think	D: + N_P	sub drt
think about (carefully)	V wh	think	D: + wh-$CL_{P:it}$ D: + wh to-$ING_{P:it}$	sub vrb_{wh}
think about (intention)	V –ing	think	D: + V-ing_P	sub $vrb_{\text{-}ing}$
--	--	think	T: + N_P + for N	sub drt prp_{for}
--	--	think	D: $SENTENCE_{P:it}$	--

Tab. III: Comparison CCDE, VDE and valency complements

In my opinion, both the CCED and the VDE suppress important syntactic features of clause structure in favour of a description of the surface structure of clause elements based on word classes and their realisation forms. For comparison, the third column shows valency complement types, which are used for the comparative analysis in this investigation. The valency complements represent a 'building kit' for clause construction, which can be further sub-categorised (subscript) if wished.

In a contrastive analysis between languages the differences in clause structure become apparent when using valency theory as it establishes the relationship between the syntactic

context of clause structure based on specific lexical choices. This is a feature which is especially important in language teaching as it enables students to explore the differences between a second language and their native language. Knowledge of verb valency patterns thus contributes to learners' understanding of grammatically correct sentence formation.

3. The Use of Corpora in Linguistic Study and the Language Classroom

Different corpora are likely to show different results due to register variation and different sampling criteria for texts. This is certainly a restriction of which anybody undertaking linguistic study using a corpus needs to be aware of.

The corpus used for this investigation is EuroParl, consisting of European Parliament Proceedings published in the official languages of the European Union (www.statmt.org/europarl). It thus represents a specific language domain. A disadvantage of EuroParl is that it does not identify from which language a text was translated. However, since there is always a great degree of freedom in translations (Kenny, 2005: 162) I feel that not knowing the source language should not have a significant impact on the investigation.

		BoE	EuroParl
CONSIDER	*per million*	**253.64**	**549.51**
consider	*per million*	89.76	300.64
	%	35.39%	54.71%
considered	*per million*	109.75	136.53
	%	43.27%	24.85%
considering	*per million*	43.07	60.07
	%	16.98%	10.93%
considers	*per million*	11.06	52.27
	%	4.36%	9.51%

Table IV: CONSIDER and its word-forms in two corpora

A comparison (table IV) of EuroParl with the Bank of English (BoE)[1], a general corpus with various sources and genres, shows that CONSIDER is twice as frequent per million words in EuroParl than in the BoE. The individual word-forms of the lemma also show a different distribution. Whilst in the BoE the perfect is most frequent, indicating either past tense or passive use, in EuroParl the present forms, indicating present tense or infinitive use, are notably more frequent. However, these differences do not affect the purpose of this investigation, which is a description of the different syntactic structures of the verb CONSIDER and its German translation equivalents.

3.1 How to Explain Irregularities in the Corpus

The distinction of whether a verb can be followed by an '-ing' or 'to-inf' verb

		BoE	EuroParl
CONSIDER	*per million*	**253.64**	**549.51**
-ing	*per million*	25.34	25.30
	%	9.99%	4.60%
to-inf	*per million*	9.66	35.97
	%	3.81%	6.55%
that-cl	*per million*	7.11	87.27
	%	2.80%	15.88%

Table V: 'Patterns' of CONSIDER in two corpora

[1] Corpus at the University of Birmingham.

structure features prominently in English language teaching (Parrott, 2000: 136-150; Sinclair, 2005: 184-193). I therefore undertook a computational search for these structures. Additionally, since CONSIDER is often used as a reporting verb (Sinclair, 2005: 321), a search for a following 'that'-clause was also undertaken (table V). 'That-cl' complementation is by far the preferred pattern in EuroParl, indicating a preferred reporting function of CONSIDER, while in the BoE '-ing' complementation is more typical. According to Parrott (2000: 141) '-ing'-forms provide additional information on the verb or clause.

However, these 'mechanical' frequency searches can be misleading and it is important to look at the concordance lines. For example, from the language classroom we know that 'some verbs are always followed by a present participle clause', i.e. an '-ing'-form, and CONSIDER is one of them (Sinclair, 2005: 186). So, how are the occurrences of 'to-inf' structures directly following CONSIDER to be explained?

An extract of the concordance lines (Figure 3) shows that in these cases we are not dealing with a canonical clause, but a complex clause where CONSIDER is either the main

```
01 ... ld like to congratulate Mr Berenguer Fuster on what I [[consider]] to be an excellent report . I can identify with it very  ...
02 ... opean Free Alliance - the reasoning in whose speech I [[consider]] to contain a slight contradiction . On the one hand , M  ...
03 ... ent Denmark in Parliament , made contributions that I [[consider]] to be part of the discussion of and campaign for domesti ...
04 ... , just as they reject elderly men and women whom they [[consider]] to be a burden . Instead of this culture of death , let  ...
05 ... ort and I would like to emphasise one aspect which we [[consider]] to be essential . It is not enough merely to state that  ...
06 ... ll take into account a whole raft of subjects that we [[consider]] to be crucial . The document still , however , falls sho ...
07 ... ng the agreement on intellectual property , which you [[consider]] to be necessary , I do not think that a thorough revisio ...
08 ... , since such judges will only be able to handle cases [[considered]] to be minor ones , following an investigation of the c ...
09 ... prove the common position , rapid implementation was [[considered]] to be the most important thing . We are well aware of  ...
10 ... in relation to 1990 , as at the time this measure was [[considered]] to be exceptional . At the present time , it is becomi ...
```

Figure 3: Concordance lines 'CONSIDER+to+verb(inf)' in EuroParl and BoE

verb of a relative clause (01 –08), or in the passive (09, 10). A transformation in a simple clause would be:

01' I consider it to be an excellent report.

05' We consider this aspect to be crucial.

09' Someone considered the rapid implementation to be the most important thing.

Therefore we are not dealing with a sentence pattern in its own right, but with a transformation pattern or rule. The valency pattern for the canonical clause is therefore <sub drt $vrb_{to\text{-}inf}$> for the examples 01 to 10. In the example sentence 08 the relative pronoun is omitted:

09' ... will only be able to handle cases (which are) considered to be minor ones ...

4. CONSIDER – TWO VALENCY PATTERNS AND THEIR TRANSLATION EQUIVALENTS

In the following section I look at 10 randomly chosen EuroParl files in greater depth to find all the possible translation equivalents and their patterns. Due to the relatively small text sample (594,461 words in English) this study is not to be seen as representative, but requires further investigation. However, it gives an indication of how frequency analysis can be applied in order to identify preferred translation equivalents and their patterns. This knowledge can be used in the language classroom or in the compilation of dictionaries and grammars. In general, it was noticed that a wide variety of translation equivalents can be utilised equally well. Tables VI and VII below show specimen dictionary entries based on the presented investigation.

4.1 Valency Pattern <sub vrb$_{that}$>

This pattern is very frequent with CONSIDER (Herbst et al, 2004: 176; Sinclair, 1995: 345), which is confirmed in the analysis by 51 occurrences[2]. The translation equivalents are:

5x der Auffassung sein / der Meinung sein / der Ansicht sein

4x nach Auffassung / halten (für)

3x nach Ansicht / denken / meinen / no translation

2x sehen / berücksichtigen

[2] Occurrences with 'that' as determiner were excluded.

1x nach Meinung / zu der Ansicht gelangen / die Meinung vertreten / die Auffassung vertreten / vor Augen halten / ausgehen (von) / überlegen / urteilen / feststellen / betrachten / hinweisen / signalisieren / aussprechen (für)

consider **<sub vrb$_{that}$>**	sein, der Ansicht/Meinung/Auffassung **<sub vrb$_{dass}$>**
However , [the Commission] considered [that the political compromise in the Council was the best] ...	[Die Kommission] war jedoch der Ansicht , [dass der im Rat erzielte politische Kompromiß der beste war] ...
	nach Ansicht/Auffassung/Meinung ***adjunct***
[The Commission] considers [that family reunification is an essential element ...]	Nach Auffassung der Kommission ist [die Familienzusammenführung] [ein wesentliches Element ...]
	halten **<sub acc prp$_{für}$>**
[It] considers [that the proper legal basis for this regulation is Article 62 (3) of the EC Treaty] ...	[Sie] hält [Artikel 62 Absatz 3 des EG-Vertrags] [für die geeignete Grundlage dieser Verordnung] ...
	denken/meinen **<sub vrb$_{dass}$>**
[I] do not consider [that any precedent is being set] ...	[Ich] denke nicht , [daß es hier einen Präzedenzfall gibt] ...

Table VI: Suggested specimen dictionary entry based on the most frequent translation equivalents for CONSIDER <sub vrb$_{that}$>

Similar to English, the most frequent translation equivalents occur with a <vrb$_{dass}$> complement, expressing or reporting someone's opinion (Table VI). It is notable, that the most frequent translations are multi-word units, verb-noun combinations, functioning as verbs. However, as 'nach Ansicht' shows, the translation equivalent does not need to occur in the same syntactic function as the original. Despite being quite frequent, this translation is not found in bilingual dictionaries. Dictionaries generally only show the translation equivalents with the same syntactic function as the original.

4.2 Valency Patterns <sub drt nom>, <sub drt adj> and <sub drt vrb$_{to\text{-}inf}$>

These patterns are combined as the key translation equivalents are the same. This might not be surprising as nominal and adjectival complements can be expanded by a verbal complement into a verb phrase without changing the meaning (Fischer, 1997: 151). For example: (EuroParl) ... Some consider the 20th century a century of war ...

(Transformation)... Some consider the 20th century to be a century of war ...

The translation equivalents for the 50 occurrences are:

18x halten (für)

8x betrachten (als) / no translation

2x der Ansicht sein / der Meinung sein / (an)sehen (als) / finden

1x beimessen / prüfen / erörtern / glauben / verstehen (als) / empfinden (als) / bezeichnen (als) / herausstellen (als) / gelten (als) / bewerten (als) / ansehen (als) / anstreben / Erachten (für) / angeblich

It stands out that with these valency patterns of CONSIDER the correlate 'it' in the main clause is more frequent than with the other patterns (Table VII). Correlates occur in the main clause of a sentence and refer to an extension clause (Engel, 1988: 252). The correlate structure is generally maintained in the German translation equivalents. This might be the reason for 'HALTEN für' being the most frequent translation equivalent.

consider **<sub drt vrb$_{to\text{-}inf}$> <sub drt nom> <sub drt adj>**	halten **<sub acc prp$_{für}$>**
At the present time , [I] would consider [a statement of any kind] [to be premature .]	Zum jetzigen Zeitpunkt würde [ich] [jede Form einer Äußerung an dieser Stelle] [für verfrüht] halten .
Often with correlate 'it' for drt extension!	***Often with correlate 'es' for accusative extension!***
At the same time , however , [I] do not consider [***it***] [wrong for us] [***to link ...***]	Gleichzeitig halte [ich] [***es***] aber nicht [für falsch] , [***dass wir ...***]
	betrachten **<sub acc prp$_{als}$>**
... [the European Parliament] considered [this action programme] [to be of undoubted benefit ...]	... [das Europäische Parlament] hat [dieses Aktionsprogramm] sicher [als nützlich] betrachtet und ...
	finden **<sub acc adj>**
... but [I] do not consider [the substance] entirely [convincing] either ...	... [ich] finde auch [den Inhalt] nicht ganz [überzeugend] ...

Table VII: Suggested specimen dictionary entry based on the most frequent translation equivalents for CONSIDER <sub drt vrb$_{to\text{-}inf}$> <sub drt nom> <sub drt adj>

5. SUMMARY

In this investigation valency theory, the property of a word to combine with or demand a certain number of elements in forming larger units is utilised in the exploration of the

grammar-lexis continuum. For exemplification the polysemous verb CONSIDER and its German translation equivalents were studied to investigate whether valency complements relate to word meaning. A corpus linguistic approach was chosen to identify patterns and translation equivalents.

Two corpora were used, the BoE and EuroParl. Frequency analysis has shown that syntactic patterns for CONSIDER differ between the two corpora. This is not surprising since different corpora will generally show different results, and it is often difficult to replicate corpus studies.

Frequency analysis was seen to be useful in order to identify preferred, though varied, translation equivalents and their valency complements and this knowledge can be useful in the language classroom or in the compilation of dictionaries and grammars. The contrastive analysis has shown that there is interplay between meaning, i.e. the chosen translation equivalent, and valency patterns. All the patterns investigated show some preference for certain translations. Valency patterns are thus a useful indicator of likely meaning and are therefore valuable for language learners or in translation training. However, whilst frequency analysis has its uses in the identification of differences between registers for word and pattern distribution, it has its limitations in the interpretation of irregularities and in the investigation of meaning.

It has also been shown that current dictionaries which include syntactic word patterns neglect important features of clause structure, and thus do not enhance the understanding and production of sentences.

Overall, it has been shown that valency theory links syntactic clause structure with specific lexical choices. A chosen translation equivalent will have its own valency pattern and impose this on the clause.

REFERENCES

Aitchison, J. (1994). *Words in the mind: An introduction to the mental lexicon*. 2nd edition Oxford and New York: Basil Blackwell.

Altenberg, B., Granger, S. (2002). *Lexis in Contrast – Corpus-Based Approaches.* Amsterdam: John Benjamins B.V.

Emons, R. (1974). *Valenzen englischer Prädikatsverben.* Tübingen: Max Niemeyer Verlag.

Engel, U. (1988). *Deutsche Grammatik.* Heidelberg: Julius Gross Verlag.

Engel, U. (2009). *Syntax der deutschen Gegenwartssprache*. 4th edition Berlin: Erich Schmidt Verlag.

Firth, J.R. (1957). *Papers in Linguistics 1934-1951.* London: Oxford University Press.

Fischer, K. (1997). *German-English Verb Valency – A Contrastive Analysis.* Tübingen: Gunter Narr Verlag.

Granger, S. (2009). *More Lexis, less grammar? What does the learner corpus say?* Paper presented at the conference Grammar and Corpora 3. Mannheim (Germany), 22-24 September.

Herbst, Th., Heath, D., Roe, I.F., Götz, D. (2004). *A Valency Dictionary of English – A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Berlin: Walter de Gruyter.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. Oxon: Routledge.

Kenny, D. (2005). 'Parallel Corpora and Translation Studies' in Barnbrook, G., Danielsson, P., Mahlberg, M. (eds) *Meaningful Texts – The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. London: Continuum.

Lightbown P.M., Spada, N. (1999). *How languages are learned.* Oxford: Oxford University Press.

Nunan, D. (1999). *Second Language Teaching & Learning.* Boston, Massachusetts: Heinle & Heinle Publishers.

Parrott, M. (2000). *Grammar for English Language Teachers.* Cambridge: Cambridge University Press.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. (ed) (1995). *Collins Cobuild English Dictionary.* London: Harper Collins Publishers.

Sinclair, J. (2004). *Trust the Text – Language, Corpus and Discourse*. London: Routledge.

Sinclair, J. (ed) (2005). *Collins Cobuild English Grammar.* Glasgow: Harper Collins Publishers.

Singleton, D. (2000). *Language and the Lexicon: An Introduction*. London: Arnold.

Tesnière, L. (1980). *Grundzüge der strukturalen Syntax.* German translation by Engel, U., Stuttgart: Klett-Cotta.

Teubert, W. (2004). ‘Language and Corpus Linguistics’ in Halliday, M.A.K.; Teubert, W.; Yallop, C.; Cermakova, A. *Lexicology and Corpus Linguistics: An Introduction.* London: Continuum.

Welke, K. (1988). *Einführung in die Valenz- und Kasustheorie.* Leipzig: Bibliographisches Institut.

Concordance Programmes and Corpora

Bank of English; University of Birmingham

EuroParl; http://www.statmt.org/europarl

ParaConc629; Michael Barlow

Sintaxis de un tipo de cláusula interrogativa a través de datos de corpus

IRIA DEL RÍO GAYO

Universidad de Santiago de Compostela

Resumen

Presentamos un estudio sintáctico de un tipo de cláusula interrogativa del español (interrogativa directa parcial con un objeto) basado fundamentalmente en datos de corpus.

El trabajo se enmarca en una investigación que ha tenido como objetivo la formalización de este tipo de cláusula para la construcción de un módulo que se integra en la gramática formal de propósito general Avalon.

Tres son los aspectos de la sintaxis de estas cláusulas de los que nos ocuparemos: el orden de constituyentes, la partícula interrogativa y el número de argumentos. La elección de estos tres aspectos responde a motivaciones diferentes: los dos primeros por ser los rasgos más característicos de la sintaxis de estas cláusulas; el tercero, número de argumentos, por ser un perfecto ejemplo del tipo de dato lingüístico que nos puede ofrecer el corpus y no la teoría.

Palabras clave: cláusula interrogativa, corpus, orden de constituyentes, partícula interrogativa, número de argumentos

Abstract

We present a syntactic study of a particular type of Spanish interrogative clause, mainly based in corpus data.

This work is part of a research project whose main goal is the formalization of these interrogative clauses for the construction of the module that gives account for them in the general purpose formal grammar Avalon.

We analyze three syntactic aspects of these clauses: order of constituents, interrogative particle and number of arguments. We have chosen these three aspects for different reasons: both first ones because they are the syntactic aspects most typical of this type of clauses; the last one, number of arguments, because it is the perfect case of linguistic data that we can obtain from a corpus and no from the theory.

Keywords: interrogative clause, corpus, order of constituents, interrogative particle, number of arguments

1. INTRODUCCIÓN

En el presente artículo abordaremos el estudio sintáctico de un tipo de cláusula (Rojo, 1978; Rojo y Jiménez Juliá, 1989) interrogativa del español basado fundamentalmente en datos de corpus.

El trabajo se enmarca en una investigación que ha tenido como principal objetivo la formalización de este tipo de cláusula para la construcción de un módulo que se integra en la gramática formal de propósito general Avalon (Álvarez, 1998). Avalon es una gramática formal de corte constitutivo-funcional, escrita en lenguaje *AGFL*[1] y diseñada para el análisis de corpus y la creación de bases de datos lingüísticos. Debido a su orientación, posee un nivel

[1] http://www.agfl.cs.ru.nl/

de descripción lingüística muy detallado. La gramática describe dos grandes unidades: la frase y la cláusula (Álvarez, 1998: 136). La descripción de ambos tipos de unidades se lleva a cabo en distintos niveles, de manera que tenemos determinados módulos que dan cuenta de los distintos tipos de frases y determinados módulos que dan cuenta de los distintos tipos de cláusulas (entre estos módulos clausales se encuentra el de nuestras interrogativas). Las reglas sintácticas escritas en *AGFL* que forman los módulos se construyen con dos tipos de unidades: determinadas *funciones argumentales* (sujeto, complemento directo, complemento indirecto, complemento predicativo, complemento preposicional y agente) y *el verbo* (con una determinada *voz y esquema* (Santalla, 2002: 46-52); así mismo, esas reglas se estructuran en el módulo, como principio general, por número de argumentos: de cuatro a cero (verbo solo; verbo más clíticos y/o interrogativo argumental).

El objetivo práctico de este estudio (la formalización de las interrogativas) ha definido tanto la forma de la investigación como el tipo de información sintáctica que nos interesaba obtener. No obstante, *nuestro interés no se centrará aquí en el proceso de formalización, sino en el estudio de determinados aspectos sintácticos de nuestra cláusula interrogativa.*

Tres son los aspectos sintácticos de los que nos ocuparemos: el orden de constituyentes, la partícula interrogativa y el número de argumentos. Las tres características sirven para definir el patrón sintáctico de este tipo de cláusula, pero su elección responde a motivaciones diferentes: los dos primeros, orden y partícula interrogativa, han sido seleccionados por ser considerados los dos rasgos más distintivos de la sintaxis de estas interrogativas (Escandell, 1999: 3951); el número de argumentos, sin embargo, es importante para la formalización (como veíamos más arriba, se utiliza como parámetro de ordenación de las reglas en el módulo), y no, *a priori*, un rasgo sintáctico diferenciador como los dos anteriores. Su selección viene determinada, además de por su relevancia para la formalización, porque consideramos que ejemplifica perfectamente el tipo de datos que no puede aportarnos la teoría y sí el trabajo con corpus. Además, creemos que en conjunto con los otros dos rasgos, nos sirve para delimitar la estructura prototípica de nuestras cláusulas interrogativas.

El corpus utilizado ha sido la Base de datos sintácticos del español actual (en adelante BDS[2]) (Rojo, 2001). La elección del corpus viene determinada por el tipo de información que necesitamos, recordemos, en nuestro caso: información muy detallada sobre la sintaxis de un tipo concreto de cláusula del español.

[2] Una versión de la base de datos más restringida que la utilizada para el presente trabajo puede consultarse en: http://www.bds.usc.es/.

La estructura del artículo es la que sigue: en primer lugar presentaremos el tipo de interrogativa objeto del estudio así como los aspectos de su sintaxis que analizaremos aquí. A continuación, trataremos la información más relevante que la teoría nos ofrece sobre esos aspectos. Finalmente, en lo que supone la parte central del artículo, presentaremos los datos que el corpus nos ofrece.

2. INTERROGATIVAS DIRECTAS PARCIALES CON UN OBJETO: LA TEORÍA

La estructura lingüística objeto de nuestro estudio es, como ya se ha adelantado, un tipo de cláusula interrogativa: interrogativa directa parcial con un objeto[3]. Las interrogativas se han dividido tradicionalmente en dos grandes grupos: *directas* (ejemplo 1) e *indirectas* (ejemplo 2). Según la RAE las primeras *constituyen enunciados interrogativos* (RAE, 2009: 3152), mientras que las segundas *son una variedad de las oraciones subordinadas sustantivas*.

(1) […] *¿has encontrado algo?* [ENCONTRAR, LABERINTO: 93, 30] [4]

(2) […] *nunca supe si las balas le alcanzaron o no.* [ALCANZAR, SUR: 67, 14]

Dentro de las *interrogativas directas*, se distinguen a su vez dos grandes grupos: *totales* (ejemplo 3) y *parciales* (ejemplo 4). Las primeras presentan una disyunción, es decir, dos o más alternativas entre las que el oyente debe elegir (las prototípicas se responden con *sí* o *no*); las segundas, sin embargo, se caracterizan porque presentan siempre una partícula interrogativa (*qué*) o constituyente interrogado (*qué hombre*) que supone la incógnita que debe ser resuelta con la respuesta.

(3) *¿Ha descubierto usted una estrella nueva?* [DESCUBRIR, LABERINTO: 237, 1]

(4) […] *¿por qué ponen huevos las gallinas?* [PONER, LABERINTO: 257, 33]

[3] Existen muchas clasificaciones de las oraciones interrogativas. En nuestro trabajo nos hemos inclinado por la clasificación de la gramática tradicional española que utiliza por ejemplo la RAE (2009: 3152 y ss.).
[4] La referencia final permite localizar el ejemplo en la BDS, al indicar: verbo ('poner'), obra ('Laberinto'), página (257) y línea (33). Este será el formato que utilizaremos para todos los ejemplos extraídos de la base de datos.

Finalmente, las directas parciales se dividen en dos tipos según el número de partículas interrogativas o constituyentes interrogados que presenten: las simples o con un objeto (ejemplo 5) presentan solo uno, y las múltiples o complejas (ejemplo 6) presentan varios que tienen diferentes funciones sintácticas en la cláusula.

(5) *¿Quién va a hacerte muy feliz?* [HACER, HOMBRE: 39, 22]

(6) *¿Cómo te ha sonado qué?* [SONAR, OCHENTA: 33, 23]

La cláusula interrogativa directa parcial posee, según los trabajos que se han ocupado de su estudio[5], una serie de características sintácticas que la singularizan no solo dentro del grupo de las oraciones interrogativas, sino entre los restantes tipos de oraciones del español. Entre estas características, destacan: el orden rígido de sus constituyentes; la presencia de una partícula o constituyente interrogativo encabezando la estructura; la especialización funcional de las partículas interrogativas. Analizaremos a continuación cada una de estas cuestiones con detalle, para después, cuando tratemos el trabajo con corpus, poder establecer una comparación entre teoría y datos de uso real[6].

2.1. El orden de constituyentes en las interrogativas directas parciales con un objeto

Como indicábamos más arriba, las cláusulas de las que nos ocupamos en este estudio presentan un comportamiento bastante peculiar en lo que se refiere a este parámetro: al contrario de lo que ocurre con la mayoría de las cláusulas del español, que se caracterizan por su orden libre, las interrogativas directas parciales se definen, en general, por presentar un *orden de constituyentes rígido* (Escandell, 1999: 3951; Contreras, 1999: 1939-1940). Ese orden, sería el siguiente:

Constituyente Interrogativo + Verbo + Argumentos (en orden libre)

(7) *¿Quién te lo ha contado a ti?* [CONTAR, OCHENTA: 14, 27]

[5] La bibliografía que se centra en la descripción de la sintaxis de nuestras interrogativas, o en aspectos más concretos de su estructura, no es muy abundante. Existe una amplísima bibliografía (sobre todo para el inglés) sobre el concepto de pregunta, la relación pregunta-respuesta, etc, pero no sobre la estructura sintáctica que define las cláusulas interrogativas. En el apartado bibliográfico de este artículo pueden encontrarse los principales trabajos que se han manejado para el estudio teórico.

[6] Como hemos adelantado, la teoría no nos ofrece información sobre el número de argumentos. Este es un aspecto que podemos trabajar solo con los datos de corpus.

Denominaremos a esta ordenación, por ser, *a priori*, la característica de nuestras interrogativas, *orden prototípico*.

Dos son los rasgos peculiares del orden prototípico: al inicio de la estructura se coloca siempre una partícula interrogativa y se produce inversión sujeto/verbo.

Esta disposición rígida de los elementos parece responder a motivaciones de tipo semántico: el interrogativo funciona como foco (Escandell, 1999: 3934-3935), siendo el elemento que define la incógnita, la información que se desea obtener con el enunciado. Por ejemplo, en

(8) *¿Cuándo se la llevaron?* [LLEVAR, DIEGO: 66, 25]

la incógnita, la variable que entra en juego y que se pretende resolver es el momento en que 'se la llevaron'. De hecho, lo que aparece tras el interrogativo constituye información conocida (en este caso, el hecho de que 'alguien se la ha llevado'). Tenemos por tanto una incógnita en primer plano (representada por el interrogativo) y, en segundo plano, información conocida que forma la presuposición (lo que el hablante y el oyente comparten). Esto es lo que explica además que las interrogativas parciales permitan paráfrasis como

(9.) *¿Quién les provee de alimentos y ropas?* [PROVEER, LABERINTO: 229, 3]

(10) *Alguien les provee de alimentos y ropas. ¿Quién?*

Pero no es solo la posición del interrogativo lo que permite identificarlo como foco, está también la especificad léxica que lo caracteriza, como veremos más adelante, y también su prominencia prosódica (el interrogativo constituye el pico de la curva entonativa de las parciales, que es descendente o en 'cadencia'). Parece, por otro lado, que la colocación en posición inicial para marcar al interrogativo como foco es la que desencadena, a su vez, la inversión sujeto/verbo, como ocurre también, por ejemplo, con los constituyentes focalizados antepuestos (Escandell, 1999: 3935). De hecho, si los interrogativos no van colocados en posición inicial, es decir, en su posición no marcada, se marcan mediante una prominencia fonológica especial, como en

(11) *¿Que has visto A QUIÉN?*

Debemos indicar, no obstante, que la inversión sujeto/verbo no se da siempre en las interrogativas parciales. Con determinados interrogativos (*por qué*, *cómo*, *dónde*, etc) uno de

los argumentos (el sujeto, por ejemplo) puede ir colocado entre el interrogativo y el verbo, de manera que tendríamos la siguiente estructura:

Constituyente Interrogado (en función de adjunto) + Argumento+ Verbo + Argumentos (en orden libre)

(12) *¿Por qué Sepúlveda no quiso que lo operaran de nuevo?* [QUERER, HISTORIAS: 74, 30]

Denominaremos este tipo de ordenación Anteposición tipo B. Este nombre tiene sentido por oposición al orden prototípico: en casos como este, se 'antepone' un argumento al verbo, siendo colocado a continuación del interrogativo[7].

Este tipo de estructura parece darse tan solo con constituyentes interrogativos que funcionan como adjuntos. De hecho, varios son los autores que recogen esta peculiaridad: según Goodall (2004), los 'interrogativos-adjuntos' permiten con mucha más facilidad la colocación del sujeto antes del verbo; según Torrego (1984), la inversión de sujeto no es obligatoria con *cómo*, *cuándo*, *por qué*, *en qué medida* y *si*.

2.2. Los constituyentes interrogativos: particularidades

Como se ha visto, una de las características sintácticas que definen nuestras cláusulas es la colocación, al inicio de la estructura, de una partícula interrogativa: *qué*, *quién*, *cuándo*, *dónde*, *cómo*, *por qué*, etc. Este tipo de unidades presentan a su vez una serie de particularidades que las convierten en un conjunto lingüístico muy interesante.

En primer lugar, las partículas pueden funcionar como adjuntos o como argumentos. El *grado de especialización léxico-funcional* es bastante grande: tenemos un grupo de partículas que funcionan prototípicamente como adjuntos (*cuándo*, *dónde*, *por qué, cómo*) y un conjunto de partículas que funcionan prototípicamente como argumentos (*quién*, *qué*, *cuál*, *cuánto*).

Otra de las características a tener en cuenta de las partículas interrogativas, además de su especificidad funcional, es el hecho de que algunas de ellas pueden introducir antes del verbo elementos con los que forman un constituyente sintáctico. Veamos un ejemplo:

(13) ¿*Qué* quieres leer?

[7] No podemos entrar aquí en los porqués de esta estructura; sin embargo, vale la pena apuntar que la anteposición tipo B presenta una serie de rasgos definidos y característicos que también han sido analizados.

(14) ¿*Qué libro* quieres leer?

En 13 la función de complemento directo interrogado está desempeñada por la partícula interrogativa; en 14, por el contrario, la misma función está desempeñada por la partícula interrogativa más un sustantivo, 'libro'; 'qué' y 'libro' forman aquí un único constituyente sintáctico. La especialización de las partículas interrogativas vuelve a entrar en juego también en este caso: solo determinadas partículas admiten determinados elementos con los que forman una unidad sintáctica. Las partículas que pueden admitir esos elementos son *qué*, *quién*, *cuál* y *cuánto*. El tipo de unidad depende del interrogativo: a *qué* y *cuánto* los puede seguir una frase nominal; a *quién*, *cuál* y *cuánto* una frase preposicional. Como vemos, *cuál* y *cuánto* son los únicos que pueden ir seguidos de ambos tipos de frases. El grupo de interrogativos seguidos de una frase preposicional presenta además otra particularidad: la frase preposicional puede ir situada tanto antes como después del verbo o del verbo y el argumento que le sigue:

(15) ¿*Cuántos de esos* has leído?

(16) ¿*Cuántos* has leído *de esos*?

Esto no es posible, sin embargo, con la frase nominal:

(17) ¿*Qué libro* quieres?

(18) *¿*Qué* quieres *libro*?

Las funciones que desempeñan este tipo de constituyentes interrogativos complejos siempre son argumentales, fundamentalmente: complemento directo, sujeto o atributo.

3. INTERROGATIVAS DIRECTAS PARCIALES CON UN OBJETO: LOS DATOS DE CORPUS[8]

En los apartados siguientes nos ocuparemos del análisis de los datos que la BDS nos ofrece sobre los dos aspectos ya presentados, orden de constituyentes y partículas interrogativas, además de abordar una nueva característica: el número de argumentos.

[8] No nos detendremos, pues la extensión del artículo no lo permite, en la exposición del proceso previo de extracción y análisis de la información que nos ofrece la BDS sobre nuestras interrogativas.

3.1. El orden de constituyentes

Recordemos que, según la teoría, existen dos posibilidades de ordenación de los constituyentes en nuestras interrogativas: un orden que hemos denominado *prototípico*

Constituyente Interrogativo + Verbo + Argumentos (en orden libre)

y el que hemos denominado Anteposición tipo B

Constituyente Interrogado (en función circunstancial) + Argumento+ Verbo + Argumentos (en orden libre)

Pues bien, el análisis de los datos que nos ofrece la BDS *corrobora la hipótesis teórica*: el orden de las interrogativas es en general rígido y responde al patrón establecido, con el interrogativo en primer lugar seguido del verbo y a continuación una serie de argumentos en orden libre[9]. Además de este orden prototípico, *el corpus documenta también el orden alternativo con anteposición.*

Pero además, y este es un ejemplo de la importancia de trabajar con datos de uso real para estudiar tendencias lingüísticas, la BDS nos ofrece otra nueva posibilidad de ordenación no contemplada por la teoría:

Argumento + Constituyente Interrogativo + Verbo + Argumentos(en orden libre)

(19) *¿Tú qué opinas de él?* [OPINAR, MADRID: 298, 28]

Como podemos observar, en este orden también se produce una anteposición, pero esta es, digamos, más 'agresiva' que la ya conocida puesto que el argumento antepuesto se sitúa antes del interrogativo y, como sabemos, la posición del interrogativo al inicio de la cláusula es una de las características más definitorias de las interrogativas directas parciales.

Hemos denominado a este tipo de ordenación *Anteposición tipo A*[10].

En cuanto a la incidencia de las tres posibilidades en la BDS, tenemos los siguientes datos:

---

[9] El análisis de los argumentos posverbales ha revelado también preferencias de ordenación, pero estas no son tan claras y definidas como las que rigen la ordenación de los elementos preverbales, de ahí que prefiramos seguir hablando de 'orden libre'.

[10] No nos podemos detener ahora en esto, pero en la anteposición tipo A encontramos también una serie de características sintácticas aún más definidas que las de la tipo B.

Tabla 1: Orden prototípico vs Anteposición.

ORDENACIÓN	**Nº CASOS**	**%**[11]
ORDEN PROTOTÍPICO	2726	97,5
ANTEPOSICIÓN A	43	1,8
ANTEPOSICIÓN B	21	0,75

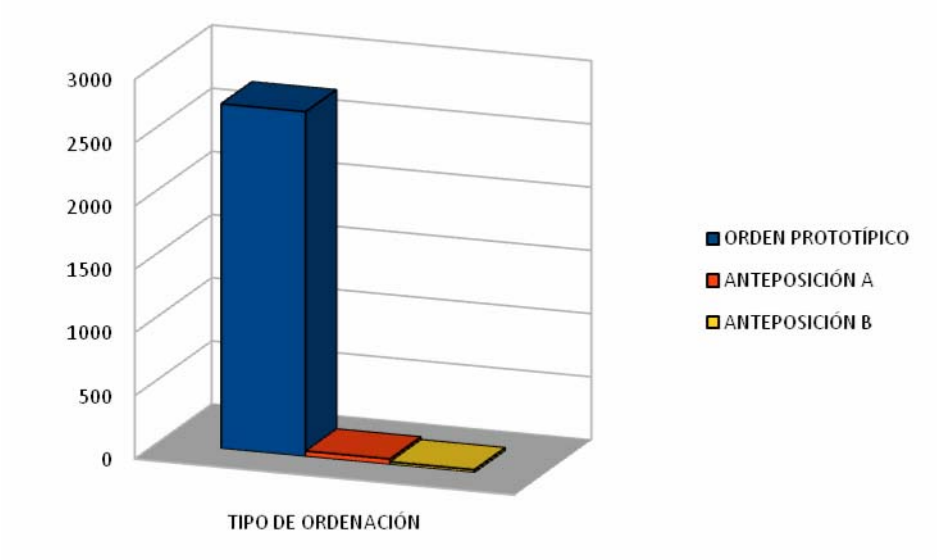


Figura 1: Orden prototípico vs Anteposición.

Como vemos, el orden prototípico bien merece esta etiqueta, pues aparece casi en el *98%* de los casos. De los dos tipos de anteposición, sorprende que es bastante más común (si tenemos en cuenta que, en conjunto, la anteposición solo supone el 2,5% de los casos) la tipo A que la tipo B, que no llega al 1%.

3.2. Tipo de interrogativo

Nos ocuparemos ahora de los datos que nos ofrece la BDS sobre la partícula interrogativa en nuestras cláusulas.

Analizaremos en primer lugar la frecuencia del tipo de interrogativo según su función en la cláusula, distinguiendo entre función argumental o función como adjunto. Presentaremos, por tanto, la alternancia entre dos patrones

- Interrogativo como Adjunto (argumento)[12] + Verbo + Argumentos (en orden libre)
- (Argumento)[13] Interrogativo como Argumento+ Verbo + Argumentos (en orden libre)

[11] Frecuencias totales.
[12] Con el argumento entre paréntesis tendríamos anteposición tipo B.

A continuación, nos centraremos en el análisis de los interrogativos en función argumental, para determinar qué funciones son las más comunes.

En lo que se refiere al reparto funcional de los constituyentes interrogativos, tenemos los siguientes datos:

Tabla 2: Frecuencia del tipo de interrogativo.

TIPO INTERROGATIVO	Nº DE CASOS	%
FUNCIÓN DE ADJUNTO	1021	37
FUNCIÓN ARGUMENTAL	1769	63
TOTAL	2790	100

La frecuencia de casos donde tenemos un interrogativo funcionando como argumento de la cláusula es, como podemos ver, muy superior a la de un interrogativo funcionando como adjunto. Este dato nos da una nueva pista sobre la estructura de las interrogativas: lo más frecuente es, cuando nos encontramos ante una cláusula interrogativa, que el constituyente interrogado esté desempeñando una función argumental.

Si ahora atendemos a la función del constituyente interrogado, nos encontramos lo siguiente:

Tabla 3: Función del constituyente interrogado.

	S[14]	DO	IO	PC	PR
TOTAL	367 (21%)*	616 (35%)	13 (0.7%)	293 (16%)	480 (27%)

* % calculado sobre el total de casos con interrogativo en función argumental (1769)

Si observamos en primer lugar los datos totales, vemos que el argumento que con más frecuencia tiende a aparecer interrogado en nuestro tipo de cláusulas es el DO; le siguen el PR y el S con frecuencias de aparición superiores al 20%; el PC es también bastante frecuente, mientras que encontrarnos un IO interrogado es, a la vista de los datos, muy extraño. La alta frecuencia del DO (y también el S) es bastante lógica si tenemos en cuenta que el esquema verbal *SDO* es el más frecuente (con diferencia) en nuestro corpus.

Si presentamos los mismos datos pero organizados por número de argumentos[15], tendríamos el siguiente resultado:

---

[13] Lo mismo: teniendo este argumento en cuenta tendríamos anteposición tipo A.

[14] Las abreviaturas hacen referencia a las siguientes funciones argumentales: sujeto, complemento directo, complemento indirecto, complemento preposicional y complemento predicativo.

Tabla 4: Función del constituyente interrogado por número de argumentos.

	S	DO	IO	PC	PR
4 ARG	1	0	0	0	0
3 ARG	6	33	0	0	22
2 ARG	118	226	5	146	178
1 ARG	242	357	8	147	280
TOTAL	367 (21%)[16]	616 (35%)	13 (0.7%)	293 (16%)	480 (27%)

Comprobamos que el DO sigue siendo el más común sea cual sea el número de argumentos de la cláusula, seguido, como en el caso de los datos respectivos al total, siempre del PR y el S; el PC y el IO no aparecen hasta los dos argumentos, siendo el IO, obviamente, muy poco frecuente.

3.3. Número de argumentos

Nos ocuparemos ahora del número de argumentos presentes en nuestras cláusulas.

Siguiendo los parámetros de Avalon, el número máximo de argumentos que se tiene en cuenta es cuatro, y el mínimo, cero (verbo solo; verbo más clíticos).

Analizaremos en primer lugar los datos totales referidos al reparto de casos por número de argumentos:

Tabla 5: Número de argumentos.

4 ARG	3 ARG	2 ARG	1 ARG	0 ARG	TOTAL CASOS
1 (0.03%)	69 (2%)	799 (28%)	1648 (59%)	273 (10%)	2790

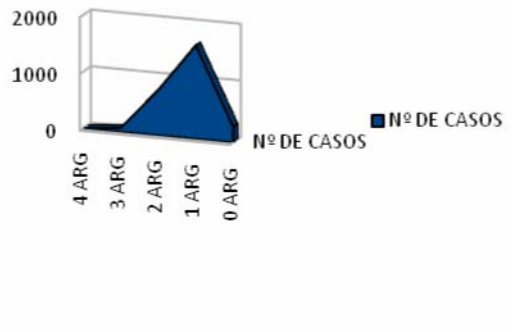


Figura 2: Número de argumentos.

[15] En Avalon el número máximo de argumentos es cuatro y el mínimo cero (solo verbo; verbo más clíticos y/o interrogativo argumental.

[16] % calculado sobre el total de casos de Interrogativo argumental (1769).

Como vemos, lo más común es encontrarnos cláusulas con un argumento pleno[17] (59%); le siguen, con porcentajes bastante más bajos, los casos con dos (28%) y cero argumentos (10%). Estas tres posibilidades, suman, en total, más del 97% del total de casos. Las cláusulas con más de dos argumentos, como podemos ver en la tabla, son muy, muy escasas: con tres argumentos un 2% y con cuatro 0.03% (un solo caso en toda la BDS).

Parece lógico deducir que, siempre según los datos de la BDS, *este tipo de cláusulas se caracterizan por presentar pocos argumentos plenos, a lo sumo, dos* (contando, recordemos, el interrogativo en función argumental).

4. Conclusión

Hemos analizado tres rasgos de la sintaxis de un tipo concreto de cláusula interrogativa (interrogativa directa parcial con un objeto): orden de constituyentes, partícula interrogativa y número de argumentos. Para los dos primeros nos hemos valido de datos teóricos y datos de uso real (BDS); para el tercero, que es un rasgo de corte más cuantitativo que cualitativo, hemos utilizado los datos que nos aporta la BDS. En conjunto, las tres características analizadas nos ofrecen el siguiente patrón sintáctico prototípico para nuestras interrogativas:

- orden de constituyentes rígido (97,5% de los casos): *Constituyente Interrogativo + Verbo + Argumentos (en orden libre)*

Constituyente interrogativo en función argumental (63% de los casos), más concretamente en función de DO, PR o S (las tres posibilidades por encima del 20%).

Número de argumentos reducido: como máximo dos (28% de los casos), aunque lo más común es uno (59%).

Referencias bibliográficas

Álvarez Lebredo, C. et al. (1998). AVALON, una gramática formal basada en corpus, *Procesamiento del lenguaje natural* 23, 132-139.

Contreras, H. (1999). Relaciones entre las construcciones interrogativas, exclamativas y relativas, en I. Bosque y V. Demonte (Eds.), *Gramática descriptiva de la lengua española,* T. 2 (pp. 1931-1963). Madrid: Espasa Calpe.

[17] Estamos contando el argumento interrogado como argumento pleno.

Escandell Vidal, M. V. (1988). *La interrogación en español: semántica y pragmática*, Madrid: Editorial de la Universidad Complutense.

Escandell Vidal, M. V. (1999). "Los enunciados interrogativos. Aspectos semánticos y pragmáticos", en I. Bosque y V. Demonte (Eds.), *Gramática descriptiva de la lengua española,* T. 3 (pp. 3929-3991). Madrid: Espasa Calpe.

Goodall, G. (2004). On the Syntax and Processing of Wh-Questions in Spanish. In *Proceedings of the 23rd West Coast Conference on Formal Linguistics,* (pp. 101-114). California: Davis.

Real Academia Española (2009) *Nueva gramática de la lengua española*, Vol. II, Espasa Libros, Madrid.

Rojo, G. (1978). *Cláusulas y oraciones (*Anexo 14 de *Verba).* Santiago de Compostela: Universidad de Santiago de Compostela.

Rojo, G. y Jiménez Juliá, T. (1989). *Fundamentos del análisis sintáctico funciona*l, Lalia 2. Santiago de Compostela: Universidad de Santiago de Compostela.

Rojo, G. (2001). La explotación de la Base de datos sintácticos del español actual (BDS). En J. De Kock (Ed.), *Apuntes metodológicos: lingüística con corpus: Catorce aplicaciones sobre el español (*pp. 255-286). Salamanca*:* Ediciones Universidad de Salamanca.

Torrego, E. (1984). On Inversion in Spanish and some of its Effects, *Linguistic Inquiry 15,* 103-129.

The problem of *false friends* in learner language: Evidence from two learner corpora

MARÍA LUISA ROCA VARELA
University of Santiago de Compostela

Abstract
False friends are a real problem for language learners. Even so, little research has been done on the identification of the difficulties learners have when it comes to the use of these words. The aim of this paper is to analyze false friends in the interlanguage of Spanish learners of English. Two learner corpora: the Santiago University Learner of English Corpus (SULEC) and the International Corpus of Learner English (ICLE), representative of the students´ interlanguage, will allow us to identify the extent of the problem. The conclusions of this study will provide new insights into the linguistic and the communication problems derived from a misuse of these lexical items.
Keywords: false friends, learner corpora, language learning

Abstract
Los falsos amigos constituyen un problema real para los estudiantes de lenguas. Aún así, se ha invertido poco tiempo en la identificación de las dificultades que los estudiantes presentan al utilizar estas palabras. Esta comunicación tiene como objetivo analizar los falsos amigos en la interlengua de los estudiantes españoles de inglés. Dos corpus de aprendices: el Santiago University Learner of English Corpus (SULEC) y el International Corpus of Learner English (ICLE), representativos de la interlengua de los estudiantes, nos permitirán identificar la dimensión del problema. Las conclusiones extraídas nos ayudarán a entender los problemas lingüísticos y de comunicación derivados de un mal uso de estos elementos léxicos.
Palabras clave: falsos amigos, corpus de aprendices, aprendizaje de lenguas

1. INTRODUCTION[1]

False friends have become a real problem for language learners. This paper focuses on false friends and on their role in the interlanguage of Spanish learners of English. Two learner corpora have been used to determine the difficulties these units produce, from two different viewpoints: from the learners´ standpoint, that is to say, the linguistic problems students should face to achieve a complete command of the English lexicon and from the point of view of the recipients, that is, the communication problems that may arise from the misuse of false friends on the hearer´s/reader´s side. The present study will hopefully provide teachers with some evidence of the student´s use of these lexical items, which may be helpful for their actual teaching in the classroom.

Before going into the study and its results, I will look at the concept of false friends and their classification.

2. THE CONCEPT OF FALSE FRIENDS

False friends have been extensively studied in different language areas: translation studies, language teaching, lexicography or contrastive linguistics. This expression goes back to 1928, when Koessler and Derocquigny used the term *faux amis* in their well-known book *Les Faux Amis, ou les Trahisons du Vocabulaire Anglais*. Here is the origin of this metaphor which is widely used in language teaching all over the world.

From an EFL context, a false friend could be defined as an L2 word that is formally similar to an L1 word in spelling and/or pronunciation but whose meanings are totally or partially different in both languages. In this case, the students´ L1 or mother tongue is Spanish/Galician and the foreign language English.

3. GENERIC CLASSIFICATION OF FALSE FRIENDS

This section deals with the generally accepted classification of false friends: the semantic classification. This categorization focuses on the semantic differences existing between two similar word pairs in two different languages. According to this, false friends can be divided into two types: total and partial.

Total false friends imply a conspicuous semantic difference between the L2 and the L1: English and Spanish in this case (e.g. English *vase* vs. Spanish *vaso,* English *avocado* vs. Spanish *abogado,* English *robe* vs. Spanish *robo*). This group could be represented in two separate circles which lay emphasis on the semantic divergence existing in both languages:

[1] The research reported in this paper was supported by the Spanish Ministry of Education (grant reference number AP2007-04477) and by the Galician Ministry of Innovation and Industry (INCITE grant number 08PXIB204033PRC-TT-206). These grants are hereby gratefully acknowledged.

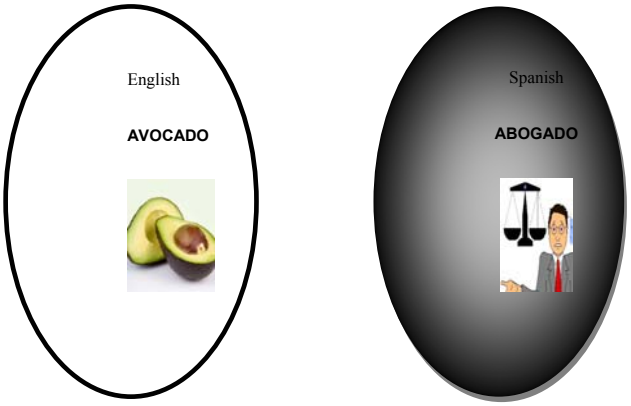


Figure 1: Semantic Divergence/Contrast

As regards *partial false friends*, they show a certain semantic *overlap* (figure 2, below). This semantic overlap occurs when two similar words have at least one shared meaning and at least one different meaning. One factor which triggers off this type of false friends is the polysemic nature of words.

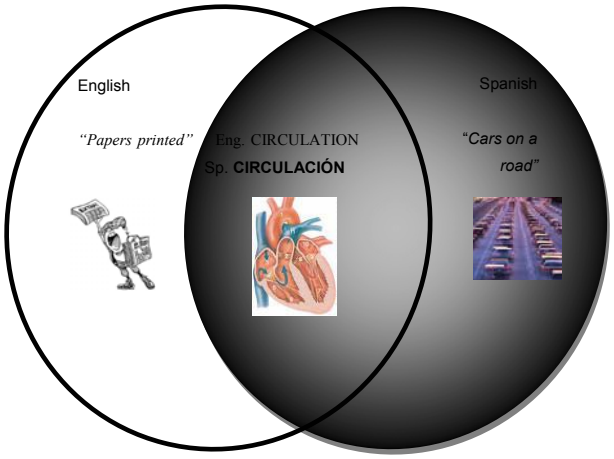


Figure 2: Semantic overlap

The present study examines word pairs belonging to both categories and shows the challenges these different types of false friends pose to language learners.

4. THE STUDY

4.1. Motivation

Surveys on the false friendship phenomenon are relatively scarce (Chacón, 2005; Holmes and Guerra Ramos, 1993) and corpus-based studies on this very same issue are even less common. This study presents a corpus-based approach to the analysis of false friends in the interlanguage of Spanish students. It explores the use Spanish students of English make of false friends as revealed in two learner corpora, the Santiago University Learner of English Corpus (SULEC) and the International Corpus of Learner English (ICLE). As shown by these corpora, it is not unusual to find sentences like *Smoking in public places should be illegal because those people that smoke in these places* ***molest*** *the rest.* (SULEC-WP-IL-DOC 1149) in the interlanguage of our students[2]. From the point of view of language teaching, the misuse of false friends has two types of "side effects". On the one hand, it reveals the students´ incomplete lexical competence and, on the other hand, it leads to the communication of unintentional meanings (see example above). These reasons clearly justify the need for discussing and identifying these problems in order to alleviate them.

4.2. Aims of the study

The aims of this study might be expressed in the following research questions:

- Do students have real difficulties with false friends?
- What type of problems can be identified from the data in the learner corpora?
- Are there any possible solutions to these problems?

4.3. Methodology

As stated in the introduction, I propose a corpus-based approach for the identification of these problems. The actual procedure involved three main stages:

a) *First stage: item-selection.* As a starting point, I made a pre-selection of FFs to limit the number of items under investigation. Four main sources played a key role in this stage:

- *False Friends and Semantic Shifts* by Samuel Walsh (2005).
- Marcial Prado's dictionary on English/Spanish false friends (2001)

[2] The examples from the learner corpora are in bold type, and they are followed by a code in brackets. Documents extracted from SULEC have the following notation: the source (SULEC) + the text type : written (WP) or oral (SP) + the students' level of English: intermediate (IL) or advanced (AL) + the document number. E.g. (SULEC-WP-IL-DOCUMENT 814). The code used with examples from ICLE makes reference to the name of the corpus (ICLE) followed by the students' native language (Spanish), the place where the corpus was

- The glossary in *Os falsos amigos da Traducción* by Álvarez Lugrís (1997).
- *The Cambridge International Dictionary of English* (Procter, 1995: 435).

All these sources include an extensive record of *false friends* which allowed me to determine those high frequency items which were worthy of attention.

b) Second stage: decision on two learner corpora. Two learner corpora: SULEC and ICLE. Both of them contain samples of Spanish students of English. These two corpora were selected because they complement each other very well. ICLE mainly focuses on written samples of advanced students while SULEC adds two interesting elements: it incorporates spoken data and language produced by students of two different levels of linguistic competence: intermediate and advanced.

*c) Third stage: qualitative data analysis***.** The aims of this study require a qualitative analysis since it seeks to understand students' problems in the use of FF. Data were processed with the SULEC lexical search query and with the concordance program ConCApp3 for ICLE**.** On some occasions, an interpretation was necessary to understand learners´ use of certain items. Some monolingual dictionaries, such as the Collins English Dictionary and native speakers were consulted in order to confirm the natural use of some false friends and to compare it with native language models.

4.4. Qualitative analysis: Results

In the list of selected false friends, well-known instances of FF between English and Spanish such as *actual, career* or *attend* are included. In this section- and for reasons of space-, I will discuss some of the most problematic items for Spanish EFL students and examples which could be enlightening for language teachers.

Actual: The first false friend to be analysed is one of the most well-known examples of false friends: *actual*. This adjectival false friend is mostly used meaning "current" under the influence of the Spanish term *actual*. In fact, typical commonplace Spanish phrases are adopted and literally transferred into English. Thus, "collocations" such as *in the *actual* society, *in the *actual* world, *the *actual* government, *the *actual* law, **actual* life, *the *actual* moment, *the *actual* social situation and *the *actual* society instead of using *present-day* or *current.* Examples 1 and 2 illustrate this.

1) (…) our *actual government* is trying to modify the law to make homosexual marriage possible. (SULEC-WP-IL-DOCUMENT 671)

---

collected (UCM=Universidad Complutense de Madrid) and finally, the document number. E.g. <ICLE-SP-UCM-0017.5>

2) *Actual society* is extremely violent, television has undoubtedly influenced in this increase of violence.<ICLE-SP-UCM-0005.5>

Advertise: The verb *advertise* is a false friend with Spanish *advertir* "warn" due to their noticeable formal resemblance. Data shows us that learners are not aware of this distinction.

3) Smoke is bad for the health and many organizations and a lot of doctors *advertise* this problem. (SULEC-WP-IL-DOCUMENT 651)

By contrast, an accurate use of the word is found in example 4. The linguistic form *advertise* has the same connotation as a native speaker of English would give to it ("announce a product in order to induce people to buy it"):

4) The clothes: If you want to be in fashion you will dress the clothes which are *advertised* on television; there are fashion shows every day. <ICLE-SP-UCM-0018.7>

Despite the fact that there is not a high incidence of this word, evidence suggests that even advanced students of English show some difficulties with this verb.

Attend: This verb comes out in conventional English collocations produced by Spanish students, such as *attend university, classes, conferences* and *schools.*

5) Some people prefer to do short courses, *attend conferences* etcetera instead of doing a degree. (SULEC-WP-IL-DOCUMENT 1341)

However, there are instances in which Spanish learners get the message of the English verb but with some traces of Spanish transfer since *attend* is frequently followed by *to*.

6) University degrees are theoretical, they doesn't prepare people for the real world. You go to the university, *attend to classes* but you don't learn anything about real world. <ICLE-SP-UCM-0001.4>

Language transfer can also be seen in example 7. In this case, the student uses the word *attend* instead of *treat.* This is the result of the influence of the Spanish collocation "atender a un paciente".

7) It is very expensive to *attend* the smokers in the hospitals, hundreds of people are *attended* every year in the hospital for this bad habit, and lots of people die for smoke (SULEC-WP-IL-DOCUMENT 986)

Career: English *career* implies "having an occupation." However, this noun is used for a "university degree" in 99% of the cases. The most prevalent uses in learners' interlanguage are illustrated in examples 8 and 9:

8) When a person decides to go to university to study a *career*. (SULEC-WP-AL-DOCUMENT)

9) When you choose to study a university *career*, you expect you may get a job within the branch you have chosen; but in the majority of the cases, that is not so. <ICLE-SP-UCM-0030.4>

There are few examples that show a correct use of the word *career*.

10) Men and women [can] develop a military *career*. <ICLE-SP-UCM-0046.3>

Comprehensive: Under the influence of Spanish *comprensivo*, students use the English adjective *comprehensive* with the meaning of "understanding." Students seem to ignore the actual sense of the English item ("thorough").

11) Smokers must be more *comprehensive* and they have to understand that other people who are in the same restaurant or in the same pub with them may feel uncomfortable breathing the smoke of a cigarette. (SULEC-WP-IL-DOCUMENT 593)

Library: This item seems to be *one of the most frequently mentioned instances of false friends* in the classroom which apparently shows no traces of the Spanish influence. Consider the following:

12) Today each prisoner have access to a gym, to a video and television room, to a *library*... <ICLE-SP-UCM-0057.4>

Molest: English *molest* differs from the Spanish term *molestar* quite considerably in meaning. The Spanish word does not have any connotations of sexual abuse. The Spanish idea of "molestar" rarely involves violence and can be translated into English as *bother.* Conversely, English *molest* means "attack someone with the intention of assaulting this person sexually," it implies the idea of "sexual harassment." Spanish learners might be seriously misunderstood when they resort to *molest* to express the Spanish idea of *molestar*

"bother or disturb." Thus the use of *molest* in the examples below would produce serious misinterpretations to such an extent that any English person would understand that these speakers are considering smokers as rapists.

13) If anyone is smoking in a public place he should try don't *molest* around him. (SULEC-WP-IL-DOCUMENT 906)

14) For example would be a good idea separated in others places persons who smoke for that way they don't *molest* persons who don't like smoke. (SULEC-WP-IL-DOCUMENT 913)

These texts have been produced by intermediate students and we might assert that these problems have their origin in language transfer since these learners have the Spanish verb *molestar* in mind.

Pretend: Examples of the verb *pretend* and its related forms: *pretends* for the present and *pretended* for the past, are found in both corpora. English *pretend* whose meaning is "feign" has nothing to do with Spanish *pretender "*try to get something" and even "woo." However, when analysing the examples where this item occurs, we notice that the Spanish meaning is transferred to the English word. There is evidence for this in both corpora: SULEC and ICLE.

15) I don't *pretend* that everybody stop to smoke, but I *pretend* that they do it when they were alone xx or with others smokers. (SULEC-WP-IL-DOCUMENT 889)

16) Since people live together, there has always been someone who *pretends* to dominate the others, so I guess that society means unequality. <ICLE-SP-UCM-0013.5>

I don't pretend that and *I pretend that* ....are word-for-word translations of the Spanish expressions *no pretendo que* y *pretendo que*...

Sensible: English *sensible* and the adjective *rational* can be considered as synonyms. However, *sensible* is identical to the Spanish adjective *sensible* "emotionally responsive." The conspicuous coincidences in spelling and in word class (both are adjectives) between these items often bring about problems leading to misunderstandings. In fact, in most cases, students draw a direct connection between the English term *sensible* and the Spanish word *sensible*.

17) The dictatorial period imposed by Franco is not very far; we must be *sensible* and try to be in their feet. (SULEC-WP-IL-DOCUMENT 249)

18) Uncultivated people that are more *sensible* and accessible to external influences;….<ICLE-SP-UCM-0007.4>

Thus, the English adjective *sensible* poses serious problems for students because they wrongly assume that this term is the translation equivalent of Spanish *sensible*, and they use it as such.

Tremendous: Unlike Spanish *tremendo,* the English item *tremendous* has positive connotations: "marvellous, wonderful" (e.g He *had a tremendous time at the theatre last night*). The example below shows the use of *tremendous* with a negative sense "terrible, horrific"; it collocates with the adjective *hard,* also denoting a rather negative quality of something.

19) From my point of view the prison and the Justice System are outdated. But we should not rehabilitate criminals at all, because they are the scum of the humanity. They must be punished in a *tremendous* harder way. <ICLE-SP-UCM-0025.7>

Thus, an influence of the Spanish similar sounding and looking word *tremendo,* meaning "terrible," on the English term *tremendous*, meaning "wonderful," is clearly observed in the example from ICLE.

5. CONCLUSION

On the basis of the data provided by the two learner corpora, a number of conclusions can be drawn. The purpose of this section is to give an answer to the initial research questions presented above:

▪ Do students have real difficulties with false friends?

Yes, they do. The problem of false friends is evident from the learner data considered. This well-known language learning problem is not an invention on the teachers' part but a real problem for EFL students.

▪ What type of problems can be identified from the learner corpora?

Linguistically speaking, language learners are victims of three main problems: semantic transfer, syntactic transfer or problems of usage.

Concerning semantic transfer, Spanish students use some English terms as translation equivalents for some Spanish items as in *actual, advertise, career* or *pretend*. The influence of the L1 is, therefore, perceived and it could be highly reduced if teachers presented false friends periodically so that students can fully interiorize the semantic properties of these lexical items. The lack of problems in the word *library* suggests that students have been presented to this item in the classroom. Thus, an introduction to meaning differences between the L1 and the L2 could be effective.

Regarding syntactic transfer and problems of usage, it is clear that although students may know the meaning of a given lexical item, they are not always familiar with its particular uses. The preposition *to* in *attend to classes "regularly"* is an example of the application of L1 syntactic patterns to English even when students are acquainted with the meaning and the collocation of English *attend*. Typical Spanish expressions involving false friends, such as *no pretendo que* are literally transferred into English I don't *pretend* that (example 15).

Pragmatically speaking, the misuse of these words can easily cause serious communication problems as exemplified by *molest.* The use of *molest* in the example given would produce misinterpretations. Any native English speaker would understand the use of *molest* as "sexual harass". This is one of the main reasons why we must be careful. There are other words that can be really confusing, and even funny: the misuse of the word *preservative* or *constipation* with the meaning of the Spanish counterparts *preservativo* "prophylactic" or *constipado* "have a cold" could make us laugh in some situations.

▪ Are there any possible solutions to these problems?

All these problems should be mitigated by the teachers' action in the classroom. The next section offers some clues for the teaching of false friends.

5.1. Implications for the EFL classroom

As suggested by the present study, students´ problems with false friends could be greatly reduced if teachers paid more attention to a meaningful teaching of these lexical items. One way of doing this is by teaching false friends in context. The use of audiovisual materials

(pictures, videos, cartoons) in the classroom might be useful and could promote students' reflection on the potential misunderstandings caused by those problematic words in naturally occurring situations.

Figure 3: Teaching false friends through pictures

Apart from using audiovisual materials, teachers might encourage students to use of ICTs so that they can expand their knowledge and obtain a better command of the English language.

To summarize, this study shows that there is room for teachers' action concerning false friends. EFL learners have serious problems when using these lexical items and teachers should deal with this issue so that learners´ lexical competence expands and potential misunderstandings can be avoided.

REFERENCES

Álvarez Lugrís, A. (1997). *Os Falsos Amigos da Traducción: Criterios de Estudio e Clasificación*. Vigo: Universidade de Vigo.

Chacón Beltrán, R. (2006). Towards a Typological Classification of False Friends (Spanish-English). *Revista Española de Lingüística Aplicada* 19, 29-39.

Holmes, J., & Guerra Ramos, R. (1993). False Friends and Reckless Guessers: Observing Cognate Recognition Strategies. In T. Huckin, M. Haynes & J. Coady (Eds.), *Second Language Reading and Vocabulary Learning,* (pp. 86–107). Norwood, NJ: Ablex.

Koessler, M., & Derocquigny, J. (1928). *Les Faux Amis ou les Trahisons du Vocabulaire Anglais*. Paris: Librairie Vuibert.

Procter, P. (Eds.). (1995). *Cambridge International Dictionary of English.* Cambridge: C.U.P.

Selinker, L. (1992). *Rediscovering Interlanguage*. London: Longman.

Walsh, A**. (**2005). *False Friends and Semantic Shifts.* Granada: Universidad de Granada.

Corpora used:

Granger, Sylviane. (2002). The International Corpus of Learner English. Louvain: Université Catholique de Louvain.

Palacios Martínez, I. M. (2005). The Santiago University Corpus of Learner English. Santiago: University of Santiago de Compostela. Available at http://sulec.cesga.es/

The discourse of Americans in Brazilian cookbooks: a proposal for an analysis based on Corpus Linguistics

ROZANE RODRIGUES REBECHI

University of São Paulo

Abstract

Corpus Linguistics has played an important role in terminological and lexicographical research. However, this methodology has been underused in Discourse Analysis, for example. In accordance with criteria underlying Corpus Linguistics, the present study aims to propose a methodology for an analysis of American discourse regarding Brazil and its people, having Brazilian culinary books — originally written in English by North-Americans — as its study corpus. Results show that semi-automatic methods of data retrieval are faster and more reliable in identifying patterns than methods based on sequential reading.

Keywords: Corpus Linguistics, Discourse Analysis, Americans, Brazilian cooking

Resumen

Linguística de Corpus se ha desempeñado un papel importante en la investigación lexicográfica y terminológica. Sin embargo, esta metodología ha sido poco utilizada en el análisis del discurso, por ejemplo. De conformidad con criterios subyacentes a la Linguística de Corpus, el presente estudio tiene por objeto proponer una metodología de análisis del discurso del americano sobre Brasil y su gente, con los libros culinarios de Brasil — originalmente escritos en inglés por los norteamericanos — en su corpus de estudio. Los resultados muestran que los métodos semiautomáticos de recuperación de datos son más rápidos y más fiables en la identificación de patrones que los métodos basados en la lectura secuencial.

Palabras clave: Linguística de Corpus, análisis del discurso, americanos, cocina brasileña

1. INTRODUCTION

If we consider the number of .com domain sites dedicated to Brazilian cooking, it could be claimed that Brazilian culinary appeals a lot to Americans. A search with the expressions 'Brazilian cooking', 'Brazilian culinary' and 'Brazilian cuisine' results in more than 93.000 findings[1], only considering American sites. But a simple investigation of these findings also reveals several mistakes involving the translation and/or definition of vocabulary related to Brazilian cooking written in English: misspellings, lack of standardization and confusion between Portuguese and Spanish are some of them. As a matter of fact, lack of standardization regarding culinary issues is not limited to translation. Rather, it affects culinary in general (cf. Tagnin & Teixeira, 2004).

The strategies involved in the translation of Brazilian culinary terms into English have been addressed by Costa (2006: xiii), who calls attention to the effects translation strategies

[1] According to searches carried out in February 2010.

have on the culture being translated and, in relation to Brazilian cultural references, concludes that the image conveyed has been that of "amateurism" and "inefficiency".

This study aims to find out what Americans *say* about Brazilians in cookbooks and what image of Brazilian people is created by these beliefs. Are the misunderstandings limited to vocabulary translation/definition, or are they also present in the image Americans have of Brazil and Brazilian people, at least when we consider eating habits?

2. CORPUS LINGUISTICS AND DISCOURSE ANALYSIS

The methods underlying Corpus Linguistics (hereafter CL) are responsible for the collection and examination of large amounts of texts (Sinclair, 1991). Semi-automatic processing of language enables the access to data that could remain unobserved through an intuitive analysis of fragments of language: "Language should be studied in actual, attested, authentic instances of use, not as intuitive, invented, isolated sentences" (Stubbs, 1996).

CL has played a very important role in terminological and lexicographical studies, as can be seen in Sinclair (1991 and 2004), Mahlberg and O'Donnell (2008) and Stubbs (2002), among many others. Nevertheless, this methodology has not been widely explored in Discourse Analysis (hereafter DA), although studies such as Orpin (2005) and Berber-Sardinha and Barbara (2008) are beginning to change this picture.

At first sight, discourse analysis and Corpus Linguistics seem to have little in common, but Sinclair (2004: 10) understands them as "twin pillars of language research", since modern technology helps find and manipulate evidence, which is essential for formulating new hypotheses. Other than merely demonstrating patterns previously predicted, computational tools can show linguistic evidence that can serve as the starting point for further investigations by the analyst.

By using the concepts inherent to Corpus Linguistics, this study proposes a methodology of analysis of the American discourse (represented, here, by the authors of the recipe books that make up the corpus) towards Brazil and Brazilians, by means of the introductory texts found in Brazilian culinary books written in English by Americans.

3. THE CORPUS

In order to reveal data worth analyzing, a corpus must be compiled according to well-established criteria:

> The beginning of any corpus study is the creation of the corpus itself. The decisions that are taken about what is to be in the corpus, and how the selection is to be organized, control almost everything that happens subsequently. The results are only as good as the corpus (Sinclair, 1991: 13).

When we consider the translation/definition mistakes mentioned in the introduction, it could be argued that anything can be found in the internet, including misspellings, untrue statements, texts whose authorship and origin are unknown etc. Therefore, in order to investigate what Americans *say* about Brazilian cooking, it was decided to compile a study corpus in which only published books would be included, because, in relation to published texts, it is expected that:

a. the author should have researched the subject in depth;
b. before being published, the book must have been carefully revised;
c. books serve as reference for correct use of language.

The study corpus is constituted of eight cookbooks of Brazilian recipes written in English which were available for sale at *Amazon.com* between 2007 and 2008. A quick analysis of these cookbooks showed that problems regarding the translation of Brazilian cooking into English are not restricted to internet sites. The books also contain problems such as incorrect definitions, substitutions of ingredients that result in erasure of cultural references, translation of a term by a word whose spelling cannot be found in dictionaries in the target language, misspelling and confusion between terms in Portuguese and in Spanish.

These findings show that even the authors of these books, people who, in general, lived in Brazil and wanted to share their experience with their fellow countrymen, sometimes are unaware of aspects they are willing to talk about, and end up spreading wrong information about Brazil and its culinary:

> The study of recurrent wordings is […] of central importance in the study of language and ideology, and can provide empirical evidence of how the culture is expressed in lexical patterns. The cultural assumptions connoted by such patterns, especially when they are repeated and become habits, are an important component of socialization (Stubbs, 1996: 169).

It is well known that the availability of bilingual (English-Portuguese and Portuguese-English) dictionaries in the area of culinary is scarce, which makes things very difficult for the translator and/or the writer who deal with this subject (cf. Teixeira, 2004). Nevertheless, a

simple search using the internet would help avoid some of the problems mentioned above. Besides, why would a writer, willing to spread Brazilian cooking, substitute ingredients that are essentially Brazilian?

Those findings in relation to Brazilian recipes written in English led us to analyze what these authors *say* about Brazil and Brazilian people in the introductions to the books and/or to the recipes, and find out what these comments have in common. In these texts the writers include: curiosities about Brazil and Brazilian people, their experience in the country, historic facts, curiosities about the recipe they are about to describe etc. So, much more than simply providing recipes, these books intend to provide their countrymen with *facts* of the Brazilian culture. And it is these texts, introductions to the books and recipes, which are used in this analysis.

The texts were scanned and saved in TXT format to be explored semi-automatically with the help of the computational tool *WordSmith Tools 5.0* (hereafter WST5) (Scott, 2007). The word lists of the texts of each book were generated and saved as follows:

Table 1: Number of words of the texts in the corpus.

WordList	**Cookbook**	**Nr of tokens**
WLatbINF	*A Taste of Brazil* (Neeleman & Neeleman, 2007)	8.339
WLbcINF	*Brazilian Cooking* (Leroux, 1980)	1.475
WLbcjINF	*Brazil: A Culinary Journey* (Hamilton, 2005)	7.390
WLbctINF	*Brazil: A Cook's Tour* (Idone, 1995)	32.732
WLctbwINF	*Cooking the Brazilian Way* (Behnke & Duro, 2004)	5.879
WLdbcINF	*Delightful Brazilian Cooking* (Ang, 1993)	650
WLesbINF	*Eat Smart in Brazil* (Peterson & Peterson, 1995)	16.309
WLtcobINF	*The Cooking of Brazil* (Locricchio, 2005)	3.884
Total words		**76.657**

As can be seen, the number of words of the introductory texts varies a lot from book to book and the total number of words (tokens) would be considered insufficient, for example, if this study had terminological or lexicographical aims. But the criterion of size must be directly linked to the criterion of representativeness of the corpus (Aijmer & Altenberg,

1991). This study aims to find what the eight books have in common in relation to what is mentioned about Brazil and Brazilian people. That is why the corpus is considered representative for this study.

4. ANALYSIS OF THE CORPUS

In order to find out what is characteristic of the corpus studied, the keywords, defined by Hunston (2002: 68) as "words which are significantly more frequent in one corpus than another", were generated for each word list, using as reference the word list of the BNC[2].

Indeed, the list of keywords reveals many content words of the corpus. Among words related to Brazilian culinary and culture there are words that are specific to each book, though. And since the aim of this study is to analyze what is recurrent in different books, written by different authors, in order to, possibly, identify characteristics that are common to most of them, the lists of keywords were compared among themselves by using the function *database* of WST5, resulting in key-keywords, that is, words that are key in a certain number of files (for this study, it was set that for it to be considered a key-keyword, the word should be key in at least five keyword lists:

[2] The British National Corpus (BNC) comprises texts of both written and spoken general English. It can be consulted from http://www.natcorp.ox.ac.uk/.

Table 2: First 40 key-keywords in descending order of key-keyness.

N	**KW**	**Texts**	**%**	**Overall Freq.**	N	**KW**	**Texts**	**%**	**Overall Freq.**
1	BRAZIL	8	100.00	314	21	MEAL	6	75.00	99
2	BRAZILIAN	8	100.00	342	22	MEAT	6	75.00	125
3	COOKING	8	100.00	142	23	MEATS	6	75.00	39
4	AND	7	87.00	3,113	24	NATIVE	6	75.00	60
5	DISHES	7	87.00	177	25	OIL	6	75.00	121
6	FOOD	7	87.00	204	26	PORTUGUESE	6	75.00	173
7	RECIPES	7	87.00	91	27	RICE	6	75.00	85
8	AFRICAN	6	75.00	99	28	SHRIMP	6	75.00	90
9	AMAZON	6	75.00	55	29	BAHIA	5	62.00	59
10	BEANS	6	75.00	79	30	COCONUT	5	62.00	83
11	BEEF	6	75.00	57	31	COOKED	5	62.00	40
12	BRAZILIANS	6	75.00	76	32	CORN	5	62.00	64
13	BRAZIL'S	6	75.00	52	33	CUISINE	5	62.00	88
14	CHEESE	6	75.00	51	34	DE	5	62.00	174
15	CHICKEN	6	75.00	57	35	DELICIOUS	5	62.00	73
16	CULINARY	6	75.00	46	36	DESSERTS	5	62.00	33
17	DISH	6	75.00	100	37	DRIED	5	62.00	78
18	FISH	6	75.00	179	38	FAVORITE	5	62.00	33
19	FRUITS	6	75.00	103	39	FEIJOADA	5	62.00	31
20	INGREDIENTS	6	75.00	68	40	FLAVOR	5	62.00	24

Despite the fact that the words *Brazil*, *Brazilian* and *cooking* have the highest keyness, as they are key in the eight books analyzed, if we consider that it is possible to refer to *brasileiros* (Brazilians) as *the Brazilian people*, *the Brazilian*, besides *the Brazilians*, it could be argued that *Brazilians* is a very meaningful word in the list. Therefore, it was chosen to exemplify this methodology of analysis.

The following step of this study was to run concordance lines for the word *Brazilians*. By using the function *Concord* of *WST5*, it is possible to analyze the word in the context it appears:

N	Concordance		
1	many believe originated in Minas Gerais.	Brazilians	call their chicken soup, canja, and
2	a festive atmosphere is it any wonder that	Brazilians	call their hometown Cidade
3	It is the state of Bahia, however, that many	Brazilians	associate with outstanding
4	and once you taste it you'll know why	Brazilians	love it. As you travel across the
5	types can be an overwhelming experience.	Brazilians	use fruits in many ways. They are
6	we know as guava is called goiaba by the	Brazilians,	and in the form of a sweet paste
7	from the dorsal hump of the zebu steer.	Brazilians	have been known to walk out of a
8	City, which has the largest population of	Brazilians	in the United States. Stores and
9	$20 and include a monthly newsletter, The	Brazilians,	which contains articles about
10	will be plenty of items to identify. While	Brazilians	eat a light breakfast, the
11	served from 7-11 PM. In metropolitan areas	Brazilians	dine late. If you arrive much
12	they quickly learned from the native	Brazilians	how to make manioc meal and
13	mandioca, which is as basic to the diet of	Brazilians	today as it was to the early
14	diet. To this day the diet of many	Brazilians	includes few leafy green
15	became the beverage of choice for most	Brazilians,	replacing long-established drinks
16	that a recent guide book says that	Brazilians	do not do very much with meat,
17	(p. 65). The vendors, and many of the other	Brazilians	around you, will be happy to
18	are associated with particular celebrations,	Brazilians	also enjoy them throughout the
19	Ever since the arrival of the Portuguese,	Brazilians	have loved sweets. The colonists'
20	are popular on hot summer days.	Brazilians	enjoy thick fruit shakes and drinks

Figure 3: First twenty concordance lines with search word *Brazilians*.

Isolated words serve as starting points, but collocations create denotations (Stubbs, 2005). Therefore, from the concordance lines, the following step of this study was to identify the main collocates of the search word, that is, words that occur in the neighborhood of the search word (Brazilians):

Table 3: Collocates with search word *Brazilians*.

N	Word	With	Relation	Texts	Total
1	BRAZILIANS	brazilians	0	25	77
2	THE	brazilians	0	21	47
3	OF	brazilians	0	14	18
4	AND	brazilians	0	11	16
5	TO	brazilians	0	8	14
6	A	brazilians	0	10	14
7	IN	brazilians	0	9	11
8	IS	brazilians	0	8	10
9	THAT	brazilians	0	7	9
10	FOR	brazilians	0	7	8
11	MANY	brazilians	0	7	8
12	IT	brazilians	0	6	8
13	ALL	brazilians	0	5	7
14	HAVE	brazilians	0	6	7
15	ENJOY	brazilians	0	5	6
16	THEIR	brazilians	0	4	6
17	AS	brazilians	0	5	6
18	NOT	brazilians	0	5	5
19	WHO	brazilians	0	4	5
20	DO	brazilians	0	4	5
21	LOVE	brazilians	0	3	4
22	ARE	brazilians	0	4	4
23	USE	brazilians	0	4	4
24	WITH	brazilians	0	4	4
25	BY	brazilians	0	4	4
26	CALL	brazilians	0	3	4
27	WERE	brazilians	0	3	3
28	WHAT	brazilians	0	3	3
29	BASIC	brazilians	0	3	3
30	BUT	brazilians	0	3	3
31	ALSO	brazilians	0	3	3
32	ALWAYS	brazilians	0	3	3
33	DAY	brazilians	0	3	3
34	FROM	brazilians	0	3	3
35	LIKE	brazilians	0	3	3
36	DIET	brazilians	0	2	3
37	EAT	brazilians	0	3	3

This search considers only the frequency in which a certain word occurs near a search word, but not the relation strength between the search word and its collocate (fourth column). Table 3 above shows, for example, *the* as the most frequent collocate of the search word *Brazilians*. Nevertheless, grammatical items, such as articles, prepositions etc., do not always provide meaningful information when the main interest is finding characteristics that are more associated to Brazilians. So, this search underwent changes.

In order to generate a list of meaningful collocates with the search word *Brazilians*, the settings of the concordance lines were rearranged. It was established that a collocate should occur at least three times with the search word, within a window of four words to its right and four to its left. The measure of significance chosen was Mutual Information (MI) score, which "[…] measures the amount of non-randomness present when two words co-occur" (Hunston, 2002: 71):

Table 4: Collocates with search word *Brazilians*, using MI-score to measure the strength relation.

N	Word	With	Relation	Texts	Total
1	BRAZILIANS	brazilians	9,959	25	77
2	ENJOY	brazilians	7,9	5	6
3	BASIC	brazilians	7,844	3	3
4	LOVE	brazilians	7,204	3	4
5	CALL	brazilians	6,915	3	4
6	DIET	brazilians	6,335	2	3
7	USE	brazilians	6,178	4	4
8	ALWAYS	brazilians	6,052	3	3
9	EAT	brazilians	6,021	3	3
10	ALL	brazilians	5,587	5	7
11	DO	brazilians	5,513	4	5
12	MANY	brazilians	5,337	7	8
13	WHAT	brazilians	5,315	3	3
14	WHO	brazilians	5,237	4	5
15	DAY	brazilians	5,118	3	3
16	HAVE	brazilians	5,108	6	7
17	NOT	brazilians	5,052	5	5
18	THEIR	brazilians	4,425	4	6
19	BUT	brazilians	4,415	3	3
20	LIKE	brazilians	4,364	3	3
21	THAT	brazilians	4,358	7	9
22	IT	brazilians	4,235	6	8
23	WERE	brazilians	4,101	3	3
24	ALSO	brazilians	4,085	3	3
25	FOR	brazilians	3,716	7	8
26	BY	brazilians	3,66	4	4
27	AS	brazilians	3,446	5	6
28	IS	brazilians	3,267	8	10
29	THE	brazilians	3,213	21	47
30	TO	brazilians	3,168	8	14
31	IN	brazilians	2,888	9	11
32	A	brazilians	2,818	10	14
33	ARE	brazilians	2,767	4	4
34	OF	brazilians	2,759	14	18
35	FROM	brazilians	2,693	3	3
36	AND	brazilians	2,34	11	16
37	WITH	brazilians	2,252	4	4

Table 4 shows the collocates of the search word *Brazilians* in descending order in relation to the strength of their association (fourth column). Church and Hanks (1990) suggest a minimum MI of 3 to identify meaningful lexical patterns. This study does not aim to identify lexical patterns, although we believe that lexical items provide more relevant data for the continuity of the analysis. If we establish 3 as cut-off point, the grammatical words *in*, *a*, *of*, *from*, *and* and *with*, besides the verb *are*, will be eliminated from the list of collocates of *Brazilians*. However, the word *are* can be a lexical or a grammatical item. It is considered a grammatical item when it is an auxiliary verb, but it is lexical when it functions as main verb. In this case, it can provide interesting data for the analysis. This choice is intuitive, but, as Stubbs (1995: 19) reminds us, "[…] no procedures can ever be entirely automatic. We always start with intuitions about what is interesting to study". Therefore, we decided to include *are* in the analysis of collocates.

Among the 26 collocates with an MI score higher than three, besides *are*, ten (38,4%) are verbs (*enjoy*, *love*, *call*, *use*, *eat*, *do*, *have*, *like, were* and *are*). So, this study followed with the retrieval of the lines in which *Brazilians* occurs with each of those verbs:

N	Concordance
1	ssociated with particular celebrations, Brazilians also enjoy them throughout th
2	verages are popular on hot summer days. Brazilians enjoy thick fruit shakes and
3	s. With a day off from work and school, Brazilians enjoy picnics during pleasant
4	nd dinner. Nowadays, the upper class Brazilians enjoy going out to dinner in
5	y continue late into the evening as the Brazilians enjoy their tasty food and li
6	il and vinegar and chopped cilantro. Brazilians traditionally enjoy beef tong

Figure 4: Concordance lines in which the search word *Brazilians* occurs with the collocate *enjoy*.

Next, the whole context in which the search word *Brazilians* occurs with each of the collocates (the verbs) was analyzed. This analysis enabled us to group some characteristics attributed to Brazilian people in the corpus (due to limitation of size of this paper, only two examples of each characteristic are shown):

→ Party-lovers:

- *All **Brazilians love** a party. If no reason can be found for a celebration, Brazilians will manage to invent one.*
- *With a day off from work and school, **Brazilians enjoy** picnics during pleasant weather, or they may join friends and family for meals at restaurants.*

→ Unaware of healthy eating habits:

- *Generally, **Brazilians are** not great salad eaters, and this display of simply prepared salads is a nice surprise.*

- *The **Brazilians**, who **have** never heard of caffeine-free coffee and who still manage to sleep eight hours per night, drink numerous small cups of coffee during the day, either in the office or in the* botequim *on the corner:* "Um cafezinho por favor!"

→ Mystic:

- ***Brazilians are** above all people with a belief in the supernatural; every so often one feels the touch of strange influences, and I occasionally hear stories of apparitions, ghosts, and other enchantments.*

→ Amateurs, in relation to culinary issues:

- *I learned to cook by feel like the **Brazilians do***

.

- *It is incredible that a recent guide book says that **Brazilians** do not **do** very much with meat, other than just cooking it!*

Regardless of having lived in Brazil for some time, and willing to share their experience with their fellow countrymen, the cookbook writers convey an image of Brazilians that does not differ a lot from that found in other genres of writing. The stereotype of happy, party lovers, mystic etc. is recurrent. Faria (2005) analyzed the discourse of Americans in textbooks of Portuguese as a foreign language published in the United States and concluded that the most recurring themes are those related to soccer, beach, parties etc.

Identity and difference exist through representation (Silva, 2000). The one who is in a privileged position usually takes for himself the normal identity and classifies the *Other* as *different*. By means of the statements retrieved from the books, it could be inferred that north-Americans have healthy eating habits, have more important things to do than going out etc.

As we have stated before, a deep analysis of American discourse in relation to Brazilians is not the focus of this study. Rather, we wanted to demonstrate a methodology through which data can be retrieved and used as a starting point for this type of analysis. Moreover, the analysis of a single key-keyword is not conclusive to make generalizations about what is said about Brazilians, but we believe that the methodology demonstrated here can serve as a starting point for the analyst.

5. FINAL REMARKS

With the help of the computational tool WST5, the retrieval of the contexts in which a key-keyword and its main collocates appear in the corpus enabled us to group some characteristics attributed to Brazilian people by Americans in eight Brazilian cookbooks originally written in English. In spite of not being conclusive, the methodology proposed in this study can serve as a *kick-off* for a deeper analysis of discourse.

The traditional methodology of discourse analysis may lead the researcher to seek evidence in the text to prove his/her previous hypotheses. Intuition is present in various moments of a search, but semi-automatic methods help with the identification of patterns in a more reliable way than that based on an intuitive reading of individual texts.

Analyses based on Corpus Linguistics demonstrate that individual texts can only be explained when compared to what is expected in relation to the general use of language. Computational tools provide several techniques to investigate characteristics of texts and corpora (Stubbs, 2005), and we do not need to know an author's intention in order to interpret his/her text. The text is autonomous and the author is irrelevant for its interpretation (Stubbs, 1996).

REFERENCES

Aijmer, K. & Altenberg, B. (Eds.). (1991). *English Corpus Linguistics: Studies in honour of Jan Svartvik*. London: Longman.

Ang, E. T. (1993). *Delightful Brazilian cooking*. Seattle: Ambrosia.

Behnke, A. & Duro, K. L. (2004). *Cooking the Brazilian way*. Minneapolis: Lerner Publications Company.

Berber-Sardinha, T. & Barbara, L. (2008). Linguística de *corpus* e análise do discurso. In *Desvendando discursos: conceitos básicos*. Florianópolis: UFSC.

Church, K. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, 16(1), 22-29.

Costa, A. T. P. (2006). *Brasil mostrando a sua cara: estratégias de tradução no material de divulgação cultural – um estudo baseado em corpus*. Dissertação (Mestrado em Linguística Aplicada) – Departamento de Línguas Estrangeiras e Tradução da Universidade de Brasília. Brasilia.

Faria, A. P. (2005). *A identidade brasileira nos livros de português para estrangeiros publicados nos Estados Unidos*. Dissertação (Mestrado em Estudos Linguísticos e Literários em Inglês) – Departamento de Letras Modernas da Universidade de São Paulo: São Paulo, 2005.

Hamilton, C. (2005). *Brazil: a culinary journey*. New York: Hippocrene Books.

Hunston, S. (2002). Methods in corpus linguistics: Beyond the concordance line. In Hunston, S. Corpora in Applied Linguistics, (pp. 67-95). Cambridge: Cambridge University Press.

Idone, C. (1995). *Brazil: a cook's tour*. New York: Clarkson/Potter.

Leroux, G. (1980). *Brazilian cooking*. Les Éditions du Pacifique.

Locricchio, M. (2005). *The cooking of Brazil*. Tarrytown: Benchmark.

Mahlberg, M. & O'donnell, M. B. (2008). A fresh view of the structure of hard news stories. In *The 19th European systemic functional linguistics conference and workshop*. Retrieved from http://scidok.sulb.uni-saarland.de/volltexte/2008/1700/

Neeleman, G. & Neeleman, R. (2007). *A taste of Brazil*. São Paulo: Marco.

ORPIN, D. (2005). *Corpus* linguistics and critical discourse analysis: examining the ideology of sleaze. *International journal of corpus linguistics,* (pp. 37-61). University of Wolverhampton: John Benjamins.

Peterson, J. & Peterson, D. (1995). Eat smart in Brazil: how to decipher the menu, know the market foods & embarking on a tasting adventure. Wisconsin: Ginkgo.

Scott, M. (2007) *WordSmith tools, version 5*. Oxford: Oxford University. Retrieved from http://www.lexically.net/wordsmith/

Silva, T. T. A produção social da identidade e da diferença. In Silva, T. T. da (Ed.). *Identidade e diferença – a perspectiva dos Estudos Culturais,* (pp. 73-102). Petrópolis: Vozes, 2000.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (2004). *Trust the text: language, corpus and discourse*. London & New York: Routledge.

Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language 2*, *1*. Retrieved from http://www. uni-trier.de/fileadmin/fb2/ANG/Linguistik/Stubbs/stubbs-1995-cause-trouble.pdf

Stubbs, M. (1996). Text and corpus analysis: computer-assisted studies of language and culture. Oxford: Blackwell.

Stubbs, M. (2002). Words and phrases: corpus studies of lexical semantics. Oxford: Blackwell.

Stubbs, M. (2005). Conrad in the computer: examples of quantitative stylistic methods. Language and Literature, 14, 1 (pp. 5-24). Retrieved from http://www.uni-trier.de /uni/fb2/anglistik/Projekte/stubbs/conrad.htm

Tagnin, S. E. O. & Teixeira, E. D. (2004). British vs. American English, Brazilian vs. European Portuguese: how close or how far apart? – a *corpus*-driven study (pp. 193-208). Frankfurt am Main: Lodz Studies in Language 9.

Teixeira, E. D. (2004). *Receita qualquer um traduz. Será? – a Culinária como área técnica de tradução*. Dissertação (Mestrado em Estudos Linguísticos e Literários em Inglês) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2v.

For example and *for instance* as markers of exemplification in Present-day English: A corpus-based study

PAULA RODRIGUEZ ABRUÑEIRAS

Universidade de Santiago de Compostela

Abstract

Exemplification is a kind of hypotactic relationship between two units where one of those elements, which is more specific, provides an example of the other, which is more general. The present paper is a corpus-based study about exemplifying constructions with the markers for example and for instance in Present-day British English. The use of both markers will be compared in order to find out whether there is any significant difference between them. The analysis will cover several issues, such as the position of the markers in the exemplifying construction, the kind of syntactic units which they link or the text-types where they appear. The potential combination of for example and for instance with other markers of exemplification will also be taken into account.

Palabras clave:Exemplification, exemplifying marker, General Element (GE), Exemplifying Element (EE)

Resumen

La ejemplificación es un tipo de relación hipotáctica entre dos unidades en la que una de esas unidades, que es más específica, proporciona un ejemplo de la otra, que es más genérica. Esta comunicación ofrece un estudio de corpus sobre las construcciones ejemplificativas que contienen for example y for instance como marcadores en inglés británico contemporáneo. Se comparará el uso de ambos marcadores para averiguar si existe alguna diferencia significativa entre ellos. El análisis cubrirá diversos aspectos, tales como la posición de los marcadores dentro de la construcción ejemplificativa, el tipo de unidades sintácticas que unen o el tipo de texto en el que aparecen. También se tendrá en cuenta la posible combinación de for example y for instance con otros marcadores de ejemplificación.

Palabras clave: Ejemplificación, marcador ejemplificativo, Elemento Genérico (EG), Elemento Ejemplificativo (EE)

1. INTRODUCTION[1]

Exemplification consists in the illustration or explanation of something by providing an example of it. Some of the most common markers of exemplification in Present-day English (PDE) are *including, included, such as, say, like, eg, for example* and *for instance*. In its most basic form, an exemplifying construction entails a relation between two units where the second element (the Exemplifying Element or EE) refers back to the first element (the

[1] For generous financial support, I am grateful to the following institutions: Spanish Ministry for Education (grant AP2007-04514), Spanish Ministry for Science and Innovation and European Regional Development Fund (grant HUM2007-60706), Autonomous Government of Galicia (grants 2008-047 and INCITE-08PXIB204016PR). I would also like to thank my supervisor, María José López-Couso, for helpful discussion and valuable feedback on earlier versions of this paper. All remaining errors are mine.

General Element or GE). The referent of the EE is more specific, and is included in the referent of the GE, which is more general (see Quirk *et al.,* 1985 or Meyer, 1992). Therefore, exemplification denotes a relation of dependence where the GE is the head and the EE is the dependent. The present paper offers a corpus-based study of *for example* and *for instance* as markers of exemplification, as in (1) and (2) below, in contemporary British English. In example (1), the GE is *newer skills*, whereas *modern hair colouring* is the EE. In (2), *similar measuring instruments* and *clocks* are the GE and the EE respectively.

(1) There are newer skills and there will be even more. Modern hair colouring, **for instance**, is comparatively new. (LOB E34: 60)

(2) This principle presupposes that the observers associated with such frames of reference employ similar measuring instruments, **for example** clocks. (LOB J51: 70)

However, the two markers under analysis also occur in exemplifying constructions in which this division into two units is not so simple. Let us illustrate this with an example.

(3) Don't write numbers (ie figures) into your speech. Write them out in full. **For example**, 1,797,021 is much more easier to say if you write one million, seven hundred and ninety-seven thousand and twenty-one. (FLOB F03: 169)

The components of example (3) are more difficult to identify due to their complex syntactic form: they are not noun phrases (NPs), as in (1) and (2) above, but rather whole sentences. *1,797,021 is much more easier to say if you write one million, seven hundred and ninety-seven thousand and twenty-one* is the EE which provides an example of the GE *[d]on't write numbers (ie figures) into your speech. Write them out in full*. On other occasions the GE is not overtly expressed, but can easily be deduced from the context. For instance, in (4) below the GE is not explicitly mentioned, but we can derive from the sentence a meaning like "one thinks, *among other things*, of the great eighteenth century battles as Marlborough's Blenheim, Ramillies, Oudenarde and Malplaquet." In examples of this kind, the GE is omitted because it is redundant.

(4) All modern wars are People's Wars, in the sense that wars are no longer fought between professional armies meeting each other in pitched battles, the names of which are later enshrined in the textbooks - one thinks, **for example**,

> of the great eighteenth century battles such as Marlborough's Blenheim, Ramillies, Oudenarde and Malplaquet. (*FLOB* F24: 7)

In these cases, the idea of inclusion of the EE within the GE can only be understood in a general and abstract way.

2. DESCRIPTION OF THE CORPORA USED: *LOB* AND *FLOB* AS A SOURCE OF DATA

For the analysis of the behaviour of *for example* and *for instance* in the contemporary language, the *LOB* (*Lancaster-Oslo/Bergen Corpus of British English*) and *FLOB* (*Freiburg-Lob Corpus of British English*)[2] corpora have been selected. *LOB* was compiled by researchers at the universities of Lancaster, Oslo and Bergen, whereas the compilation of *FLOB* was an initiative of Professor Christian Mair from Freiburg University. These two corpora are especially valuable due to their structural similarities. Both are one-million-word databases consisting of 500 samples where each text is 2,000 words long. The texts which they contain belong to fifteen different genres:

A: Press: Reportage
B: Press: Editorial
C: Press: Review
D: Religion
E: Skills, trades and hobbies
F: Popular lore
G: Belles lettres, biographies, essays
H: Miscellaneous
J: Science
K: General fiction
L: Mystery and detective fiction
M: Science fiction
N: Adventure and western
P: Romance and love story
R: Humour

[2] *LOB* and *FLOB* are available in the *ICAME CD-Rom*.

The main difference between these two corpora is the fact that *LOB* portrays British English from the 1960s, whereas *FLOB* depicts the same variety of English from the 1990s. Since these corpora show a parallel structure, a contrastive analysis between them can easily be carried out. Hence, the use of exemplifying constructions with *for example* and *for instance* as markers can be studied making a comparison between both decades, thus trying to find out whether any significant change has taken place in the latter part of the twentieth century.

These two corpora are available in electronic format, which allows the use of text-analysis tools. For my purposes, I used WordSmith Tools, which is a computer program for the analysis of words in texts. By using it we can make three main actions: create a wordlist, create a concordance and find the keywords of a text. Electronic corpora and text-analysis tools make the work of the linguist much easier and less time-consuming. However, the use of these tools does not completely eclipse the work of the linguist. In this piece of research, I had to search for all the occurrences of the forms *example* and *instance*, and then rule out all those cases where these two items do not function as markers of exemplification in the prepositional phrases *for example* and *for instance*. Table 1 below shows the total number of tokens containing the forms *example* and *instance* in the corpora, as well as the number of cases where they occur in combination with the preposition *for* and function as markers of exemplification. This final group of instances represents the core of the present piece of research.

Table 1: *Example* and *instance* in the corpora

	EXAMPLE		***INSTANCE***	
	NUMBER OF TOKENS	***FOR EXAMPLE***	**NUMBER OF TOKENS**	**FOR INSTANCE**
LOB	242	140 (57.85%)	113	92 (81.42%)
FLOB	405	270 (66.67%)	100	83 (83%)
TOTAL	647	410 (63.37%)	213	175 (82.16%)

3. ANALYSIS OF THE DATA

In what follows, I will first analyse the main features of the exemplifying markers *for example* and *for instance* considering aspects such as frequency, position or distribution according to genre. Then, I will comment on the exemplifying constructions where these

markers appear. The kind of relationship between GE and EE or the syntactic forms usually adopted by these elements will be taken into account.

3.1. For example vs. for instance as markers of exemplification

According to the *Oxford English Dictionary* (*OED*), *for example* and *for instance* are semantically equivalent in PDE. In fact, the definition of each of these forms given in the *OED* refers to the definition of the other:

> *OED* s.v. *example, n.*1. A typical instance; a fact, incident, quotation, etc. that illustrates, or forms a particular case of, a general principle, rule, state of things, etc.; a person or thing that may be taken as an illustration of a certain quality.
>
> *OED* s.v. *instance, n.* III.6.a. A fact or example brought forward in support of a general assertion or an argument, or in illustration of a general truth. Hence, any thing, person, or circumstance, illustrating or exemplifying something of a more general character; a case, an illustrative example. Also, in broader sense, a case occurring, a recurring occasion.

Data from the corpora prove that *for example* is more frequent than *for instance* as a marker of exemplification in contemporary British English. In fact, there is a slight decrease in the use of *for instance* from the 1960s (92 examples) to the 1990s (83 examples), whereas *for example* sharply increases in frequency in the course of time (140 examples in *LOB* vs. 270 examples in *FLOB*) (cf. Table 1 above). Figure 1 below illustrates this development in a graphic way:

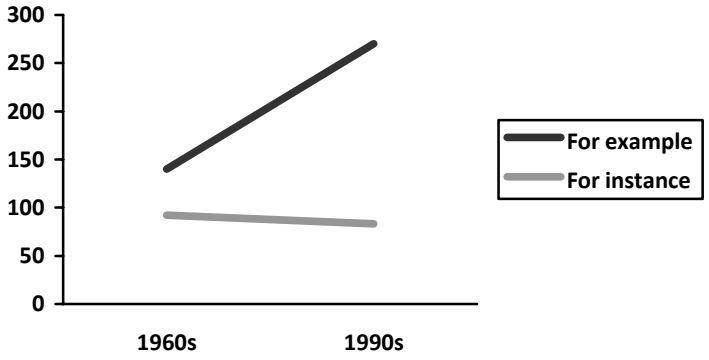


Figure 1: Development of *for example* and *for instance* as markers of exemplification from the 1960s to the 1990s

3.2. Position of the exemplifying marker with respect to the EE

Markers of exemplification may occupy different positions in exemplifying constructions. Nonetheless, each marker usually has a fixed position. Thus, *including, say, like, e.g.* or *such as* appear before the EE, whereas *included* comes after it. By contrast, the markers at issue in this paper show greater positional variability, since they may occur before, after or in the middle of the EE.

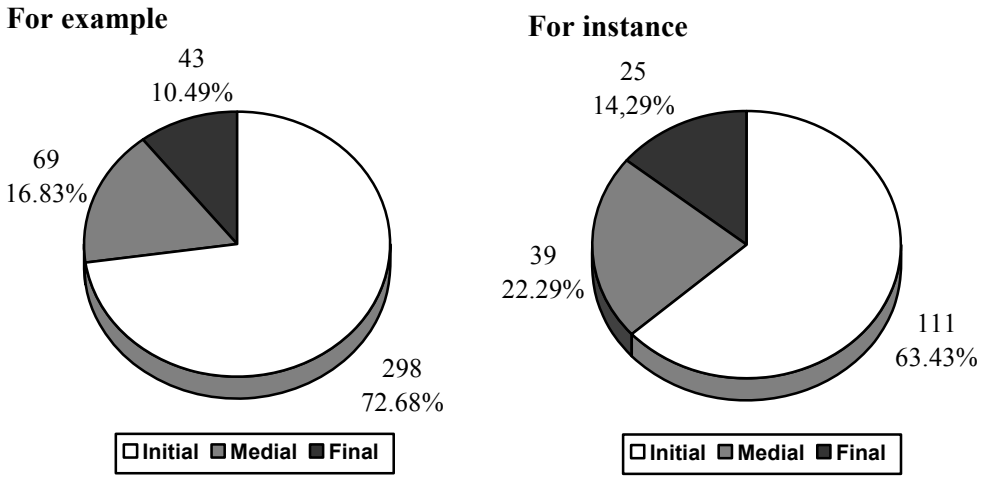


Figure 2: Position of for example and for instance with respect to the EE

As shown in Figure 2, both markers present a very similar distribution. There is a clear tendency for them to appear at the beginning of the EE (see example (2) above). On the contrary, final position (cf. example (5) below) is very rare, probably because it may be potentially ambiguous: when the marker appears at the end of the construction, we only get to know that it is an exemplifying construction when we reach the end of the sequence. In turn, medial position seems to be a strategy used to focus on a given part of the EE. For instance, in (6) below *China* is split from the rest of the EE by means of the exemplifying marker, thus becoming more prominent within this constituent.

(5) Of course, there were many sources of tension and deprivation – race relations and high-rise housing **for instance**. (*FLOB* F01: 162)

(6) It is said that there is nothing new under the sun, but regarding foodstuffs the traveller occasionally encounters a certain measure of novelty. In China, **for instance**, dried rats are esteemed a delicacy. (*LOB* F07: 91)

3.3. Combination of for example and for instance with other markers of exemplification

On some occasions the two markers at issue combine in the same sentence with other markers of exemplification, as illustrated in the following examples:

(7) Furthermore, it can give the projects it supports credibility in the eyes of bigger and wealthier grant-giving charities **like**, **for instance**, the Pilgrim Trust. (*FLOB* F40: 108)

(8) Applicants should possess a degree in statistics or mathematics, and should if possible be able to show evidence of an interest in some specialized aspect of the subject **such as**, **for example**, decision theory, information theory or stochastic processes. (*LOB* J18: 55)

(9) And the damage then will be worse than ever. Nor would it take much rain to reduce it once more to the porridge stage. Elsewhere, however, **as** in Leicester, **for instance**, the land really has dried out, and the arable was mostly in tilth by the middle of March. (*LOB* E15: 87)

Table 2 details the number of corpus instances where *for example* and *for instance* combine with other markers of exemplification. These 26 examples (4.44% of the total) show a similar distribution in both corpora.

Table 2: Combination of *for example* and *for instance* with other markers of exemplification

OTHER MARKERS	*FOR EXAMPLE*			*FOR INSTANCE*			TOTAL
	LOB	*FLOB*	TOTAL	*LOB*	*FLOB*	TOTAL	
AS	9	3	12	3	3	6	18
SUCH AS	1	2	3	0	0	0	3
LIKE	0	1	1	2	2	4	5
TOTAL	10	6	16	5	5	10	26

The combination of two exemplifying markers goes against the principle of economy in language (see Quirk *et al.,* 1985: 1350). However, it seems that two markers can co-occur as long as there is some difference between them. *As, such as* and *like* have their origin in the comparative domain, which makes them clearly different from our two markers. However, at present I cannot offer an explanation for the occurrence of such combinations and further research needs to be carried out. Another aspect to take into account is the fact that *like, as* and *such as* precede our markers in all the examples analysed. Consequently, *for example* and

for instance are extra markers which come close to the category of discourse markers used to "indicate [...] how the utterance that contains them is a response to, or a continuation of, some portion of the prior discourse" (Levinson, 1983: 197-8). In other words, they are used to signal "a sequential discourse relationship" where it is clear "how the speaker intends the basic message that follows to relate to the prior discourse" (Brinton, 1996: 30). In some cases, *for example* and *for instance* appear next to the other exemplifying markers, as in examples (7) and (8) above, whereas in other cases the EE brings the two markers apart (cf. example (9)).

3.4. Distribution of for example and for instance according to genre

Table 3 provides a breakdown of the distribution of the two markers at issue according to text-type. As can be seen, there is a clear distinction between categories A-J, which are more formal, and categories K-R, which are more informal.

Table 3: Distribution of *for example* and *for instance* according to genre[3]

GENRE	*FOR EXAMPLE*	*FOR INSTANCE*	TOTAL
A	8	4	12
B	13	6	19
C	12	6	18
D	13	9	22
E	19	22	41
F	43	24	67
G	52	42	94
H	56	6	62
J	187	34	221
K	3	7	10
L	2	4	6
M	0	3	3
N	1	2	3
P	0	0	0
R	1	6	7
TOTAL	410	175	585

[3] Genres in Table 3 are listed following the nomenclature offered in section 2.

As the data in Table 3 show, formal texts favour the use of exemplifying constructions with *for example* and *for instance* to a greater extent than informal or colloquial texts, where their use is more marginal. These markers are especially prolific in Science (category J), where 221 examples (representing 37.78% of the total) were found. This is not surprising if we take into account that scientific texts are usually complex and difficult to follow. Therefore, any comment or additional remark made for clarification may be very useful. As Duque-García (1999: 23) states, "la característica más importante, aunque no la única, de la prosa científico técnica se encuentra en el esfuerzo del autor para transmitir un único e inequívoco significado que debe ser claro y preciso" given that "si el lector puede interpretar más de un significado la escritura técnica no será efectiva". Therefore, clarity and accuracy may be some of the reasons why so many exemplifying constructions were found in scientific texts. In (10) below, the EE is used to clarify how large a synchrotron may be.

(10) Such devices are called synchrotrons and can be physically very large. **For example**, the so-called Super Proton Synchrotron (SPS) at CERN (Geneva) has a circumference around 6 km and can produce protons with energies up to around 450GeH. (*FLOB,* J01: 78)

After Science, Belles lettres, biographies and essays (category G) is the second genre showing a frequent use of these two markers (94 examples, which represent 16.07% of the total). Again, formality explains the proliferation of both markers in this genre, since Belles lettres correspond to an elegant and carefully-created kind of literature. On the opposite side are fictional texts (categories K-R), where exemplification is scarce. None of the fictional subgenres contains more than ten examples with the markers under analysis. In Romance or love story (category P), for instance, no examples were detected.

Another interesting feature of fictional texts is the fact that, contrary to the general trend described above (cf. section 3.1), *for instance* is the preferred marker. The *OED* suggests that *for instance* may be (or at least it may have been in origin) more formal than *for example* because it appears after a generic entry which reads "In Scholastic Logic, and derived senses" (*OED* s.v. *instance* III). According to this, we should expect more examples of *for instance* in formal text-types, but my data show that it is exactly the opposite situation that holds.

3.5. Elements of the exemplifying construction

As already mentioned in Section 1, exemplification entails a relation of inclusion and dependence between two units. In some cases, these units are simple elements like NPs (cf. examples (1) and (2) above) and hence the relationship of inclusion between them is clear. However, only 16.58% of the corpus examples consist of simple syntactic units. The idea of inclusion is more difficult to appreciate in the remainder instances, where inclusion has to be understood in a more general and abstract way. In some of these instances, the units in exemplification are more complex elements, especially whole sentences, as in example (3) above. In other cases, the GE is not even present, as in example (4).

In constructions with *for example* and *for instance* as markers, the GE and the EE have a certain degree of autonomy. In fact, in most of our examples (53.16%) the GE and the EE belong to different sentences. Such an autonomous character is not common in other exemplifying constructions with markers like *including* or *included*, where both GE and EE belong to the same sentence (cf. example (11) below). Moreover, in 49.52% of these examples, the exemplifying marker is right at the beginning of the sentence (after a full stop), as shown in our earlier example (3). Initial position after a full stop is not very common among the other markers of exemplification.

(11) The whole theatre area, **including** a new lighting board, was damaged by smoke. (*FLOB* A34: 83)

4. CONCLUSION

The foregoing discussion has shown that the exemplifying markers *for example* and *for instance* show a very similar use and distribution in contemporary British English. However, *for example* (410 occurrences) is much more common than *for instance* (175 tokens), and its use sharply increases in the course of time, whereas *for instance* decreases in frequency. Both markers tend to appear at the beginning of the EE because this way they clearly establish the beginning of the EE, whereas final and medial positions are less frequent, probably because they are also potentially more ambiguous. Moreover, the units linked by these two items tend to be syntactically complex, which means that the inclusion of the EE within the GE can only be understood in most cases in abstract terms. This implies a certain degree of autonomy and independence of the EE from the GE, which is the head of the construction. There are even examples where the GE is not explicitly expressed, but whose meaning can easily be deduced

from the context. In spite of this, the idea of inclusion is still present because it is inherent in the nouns *example* and *instance* themselves. As regards genre, both markers are very common in formal texts, particularly in Science, where exemplifying constructions contribute to clarity by providing examples, whereas their use is more marginal in Fiction. Moreover, it is only in fictional texts where *for instance* is more commonly used than *for example*, although according to the *OED* we should expect a higher frequency of this marker in more formal texts. Finally, there are also some cases where *for example* and *for instance* co-occur with other markers of exemplification, namely *as* (18 examples), *like* (5 examples) and *such as* (3 examples). The reasons for the occurrence of such combinations must be left, however, for future research.

REFERENCES

Brinton, Laurel J. (1996). *Pragmatic Markers in English: Grammaticalization and discourse functions.* Berlin: Mouton de Gruyter.

Duque-García, María Mar (2000). *Manual de Estilo: El Arte de Escribir en Inglés Científico-Técnico*. Editorial Paraninfo.

FLOB=*Freiburg-Lob Corpus of British English* (1991). Compiled by Christian Mair. Albert-Ludwigs-Universität Freiburg.

Levinson, Stephen C. (1983). *Pragmatics*. Cambridge: CUP.

LOB=*Lancaster-Oslo/Bergen Corpus* (1961). Compiled by Geoffrey Leech et al. Lancaster University, University of Oslo and University of Bergen.

Meyer, Charles F. (1992). *Apposition in Contemporary English. Cambridge: CUP.*

OED=*The Oxford English Dictionary*. 2nd ed.1989. OED online. Oxford: Oxford University Press. 4 Apr. 2000. < http://oed.com/>

Quirk, Randolph, Sidney Greenbaum, Greoffrey Leech and Jan Svartvik (1985). *A Comprehensive Grammar of the English Language.* London: Longman.

Corpus léxico de onomatopeyas españolas

JORGE RODRÍGUEZ GUZMÁN

Universidad de Salamanca

Resumen

En esta comunicación describimos la elaboración en curso de un corpus de onomatopeyas españolas. El objetivo general de dicho corpus es recopilar las distintas ocurrencias de las onomatopeyas en todo tipo de textos con el fin de caracterizarlas y clasificarlas. Nuestro corpus evidencia las restricciones de uso que este tipo de enunciados presenta en español: pertenencia a un estilo informal sin apenas fijación escrita; registro fundamental en la oralidad; y renovación constante. Estas particularidades de las onomatopeyas junto a una falta de corpus sobre estas unidades implican un procedimientos metodológicos y unos criterios de selección específicos a la hora de la confección del corpus y de la interpretación de los datos: búsqueda y diversificación de fuentes, delimitación de variedades diacrónicas y diatópicas, organización y etiquetado de variantes gráficas y fonéticas y transcripciones, asignación de distintos tipos de significado, etc.

Palabras clave: onomatopeyas, corpus de onomatopeyas, restricciones de uso, estilo informal, etiquetado de variantes

Abstract

In this paper, we describe the current development of a corpus of Spanish onomatopoeias. The main aim of this corpus is to compile the different collections of onomatopoeias throughout texts in order to characterize and classify them. Our corpus demonstrates the constraints of usage that this kind of sentences presents in Spanish: they belong to an informal style without hardly written fixation; they are recorded mainly in spoken language; and they undergo constant change. These peculiarities of onomatopoeias in addition to a lack of corpora on these units involve some methodological procedures and specific selective criterions when constructing the corpus and when annotating the data: retrieval and coverage of a wide range of sources, limitation of diachronic and diatopic varieties, structuring and labelling graphic and phonetic variants and transcriptions, assigning different senses, etc.

Keywords: onomatopoeias, corpus of onomatopoeias, constraints of usage, informal style, variants labelling procedures

1. OBJETIVOS

La confección de nuestro corpus tiene como finalidad principal documentar, clasificar y caracterizar las onomatopeyas de la lengua española. Esta tarea no ha sido realizada hasta ahora con la salvedad del excelente y copioso *Diccionario de Voces Naturales* de García de Diego, que, sin embargo, adopta una perspectiva etimologista y refleja un período cronológico alejado de nuestros propósitos. De todos modos, la vigencia de las palabras de García de Diego, (1968: IX), es, aún hoy, una razón fundamental para emprender este trabajo: "Los escasos frutos logrados no pueden obedecer más que a no haberse hecho una investigación fundamental y metódica de los elementos de la onomatopeya, tan complejos y tan levemente estudiados". Indudablemente, disponer de un corpus de onomatopeyas es un

trabajo indispensable para atestiguar y confirmar el uso de este tipo de enunciados, generalmente obviado de toda descripción lingüística. Por eso, como comenta Santos Río, (2003: 8), en su *Diccionario de Partículas*, nuestra intención no "tiene sólo como función importante la de mostrar los significados de los significantes sino también la de acreditar la existencia de los propios significantes (los cuales, por ser significantes, no podrán no tener significado)".

Ahora bien, antes de emprender esta labor, fue necesario plantearnos uno de los problemas teóricos más importantes en la realización del corpus, a saber, ¿qué era una onomatopeya?, lo cual exigió determinar y precisar previamente el objeto de búsqueda.

2. OBJETO DE ESTUDIO: LAS ONOMATOPEYAS

A la vez que fuimos recopilando material para nuestro corpus, advertimos la necesidad de establecer una caracterización lingüística de la onomatopeya desde los distintos niveles de la lengua (Rodríguez Guzmán, 2009). El trabajo llevado a cabo nos permitió afinar nuestro objeto de estudio en tres sentidos:

a) Por una parte, delimitamos lingüísticamente la onomatopeya respecto a la interjección (de hecho, clasificamos en un continuo las *onomatopeyas*, las *onomatopeyas interjectivas*, las *interjecciones onomatopéyicas* y las *interjecciones*) y también respecto a otro tipo de unidades: las palabras expresivas, los hipocorísticos, los fonestemas, o las aliteraciones.

b) Por otra parte, dividimos las onomatopeyas en distintas clases:

– *onomatopeyas propias*; son los datos centrales del corpus. Representan un contenido percibido o mostrado que ha ocurrido con absoluta certeza: "A las ocho, como siempre, el cohete hizo PUM; la puerta de los corrales, PAM, y... a correr. Lo que se encontraron delante los bureles de Cáceres... de traca (PUM, PAM y CATACLOC, todo junto)".

– *onomatopeyas fáticas*. Utilizada como marca en la cooperación comunicativa: pausa, asentimiento, etc.: "Claro tío, JA, JA, otro güisquito".

– *onomatopeyas discursivas*. Sirve para estructurar el discurso: como cierre, marca de apertura, etc.: "Resulta que la Editorial Y, la Editorial Z, han sacado ya sus programaciones, entonces es muy fácil transcribirlo, pasarlo al ordenador y CHIMPUN".

– *onomatopeyas lúdicas*. Sin apenas contenido representativo, es utilizada en canciones, generalmente infantiles, para marcar el ritmo o para repetir una secuencia de sonidos: Un día

en las carreras, CHIM PÚN / rompimos un cristal / TA, TA, TARATA, TÁ. / Y al ruido que produjo, / CHIM PÚN / vino un municipal / PA, PA, PARAPA, PÁ.

– *onomatopeyas ideogramáticas*. Utilizadas en las obras ilustradas como símbolo: ZZZ, `símbolo del dormir´.

c) Por último, acotamos las restricciones de uso de este tipo de enunciados: su pertenencia a un estilo informal sin apenas fijación escrita (García de Diego, 1968); su registro principalmente en la oralidad (Karcevski, 1941; Alvar López, 1999); y su renovación constante (Grammont, 1901).

3. CRITERIOS DE DISEÑO Y ELABORACIÓN DEL CORPUS

3.1. Límites temporales

Nuestro corpus de onomatopeyas abarca desde el año 1970 hasta 2010. El estudio sincrónico de las onomatopeyas es defendido por Malkiel, (1990: 9), en aras de una claridad interpretativa. En efecto, el desgaste expresivo de este tipo de enunciados junto a su poca fijación textual nos obligaron a establecer un período acotado para la búsqueda de datos. Aún con todo, constatando la dificultad de localizar ocurrencias de onomatopeyas, recopilamos también, ya sin límite temporal y secundariamente, todos aquellos casos en los que atestiguamos su uso, y los incorporamos al corpus por separado. Por tanto, el corpus, si bien se centra en los sentidos actuales de la onomatopeya, también pretende reflejar las onomatopeyas utilizadas en otras épocas.

3.2. Límites geográficos

El corpus en curso se centra en el español de España. Atiende fundamentalmente, por ello, a las obras editadas en España. Desde el comienzo de nuestro trabajo, comprobamos que resultaba muy complicado incorporar las variedades dialectales de la lengua española y las variedades americanas del español[1]. De todas maneras, debido a la escasez de muestras de onomatopeyas (ya comentada en el apartado anterior) hemos ido almacenando también estas otras variedades del español, marcándolas convenientemente.

[1] Matamala Ripoll, (2004: 583), en su estudio sobre interjecciones catalanas, tuvo que analizar la variedad del catalán central, ya que no disponía de documentos suficientes para completar las otras variedades dialectales. Tampoco los diccionarios contrastivos del español de Cuba y del español de Argentina (Haensch y Werner, 2000a y 2000b) registran muchas onomatopeyas: en el primero encontramos seis, y una tan sólo en el segundo.

3.3. El nivel de lengua

Las onomatopeyas son enunciados informales (Borrego Nieto, 2002) que se realizan en situaciones de inmediatez comunicativa (López Serena, 2007). Estos parámetros determinan los tipos de texto en los que podemos localizar las onomatopeyas. Por ejemplo, en obras informales de carácter divulgativo, en periódicos (sobre todo en textos en los que predomine el estilo propio de un autor: crónicas o columnas), en obras destinadas a un público infantil..., pero también en textos en los que se desarrollen secuencias narrativas (Rodríguez Guzmán, 2010). En cambio, no aparecerán ni en textos orales ni en textos escritos si el estilo es neutro y distante: libros de texto, exposiciones científicas, clases magistrales, definiciones lexicográficas, artículos enciclopédicos, sermones, todo tipo de obras informativas que versen sobre temas "serios" o técnicos (diseño, arte moderno, historia, cirugía...).

3.4. Tipos y proporciones del material

Uno de los mayores retos planteados en la confección de este corpus consistió en establecer criterios pertinentes de selección textual para obtener resultados representativos. Por el momento, creemos que hemos logrado establecer unas proporciones adecuadas a la presencia de onomatopeyas en los textos. De todos modos, esta selección no es fija y está siempre abierta a nuevos cambios según vamos obteniendo datos. Evidentemente, el carácter periférico de este tipo de unidades lingüísticas no permite recoger materiales homogéneamente y, por tanto, aunque nuestro enfoque de trabajo aspira a la exhaustividad, somos conscientes de sus limitaciones.

En un principio, atendimos a distintos tipos de textos orales: obras de teatro, dibujos animados, cuentacuentos y corpus orales. Sin embargo, pronto observamos que el número de datos obtenidos no era considerable y que, además, debíamos realizar la tarea de transcribir las onomatopeyas a una forma escrita o bien registrarlas y presentarlas en una forma oral. De ahí que conviniéramos en dedicarnos mayoritariamente a los textos escritos y rellenar de esta forma un vacío existente en cuanto a repertorios escritos de onomatopeyas.

El corpus en curso pretende abarcar unas 3000 obras. En su estado actual, está formado por tres bloques: uno fundamental que recoge obras alfaicónicas (un 50%); otro también muy importante, pues representa un 47%, que recopila obras destinadas a un público infantil (literatura infantil, revistas infantiles, canciones y juegos...), obras de teatro (también guiones de cine), obras de ficción (novelas y cuentos) y periódicos y revistas; el último bloque solo supone un 3% y está formado por textos orales (corpus orales, dibujos animados), repertorios lexicográficos (vocabularios, diccionarios, gramáticas, manuales de enseñanza...) y un grupo

variado de miscelánea (folletos publicitarios, poesía, canciones). En el siguiente cuadro quedan reflejadas las distintas proporciones estimadas:

Tabla 1: Selección textual

TIPOS DE TEXTO	%	NÚMERO DE OBRAS
Obras alfaicónicas	50%	1500
Obras destinadas a un público infantil	15%	450
Obras de teatro	15%	450
Obras de ficción	10%	300
Periódicos y revistas	7%	210
Textos orales	1%	30
Repertorios lexicográficos	1%	30
Miscelánea	1%	30

Hemos ido precisando estas proporciones en el transcurso de la elaboración del corpus, porque resulta adecuada a la proporción de documentos en los que podemos recopilar onomatopeyas[2]. Debemos tener en cuenta la disparidad de presencia de las onomatopeyas en géneros discursivos iguales. Así, damos importancia a los periódicos y a las revistas, pero es necesario fijar de antemano los contextos más propicios para su localización. De igual manera, aunque pretendemos seleccionar siempre las obras con más repercusión social (las más vendidas), somos conscientes de que el estilo personal de cada autor influye en el uso o no de la onomatopeya. Por otro lado, puede parecer que el número de textos orales, de repertorios lexicográficos o de miscelánea es reducido, sin embargo, creemos que el material del que podemos servirnos en estos tipos de texto no va a sobrepasar en mucho estas estimaciones. En cuanto a las traducciones, pese a que al principio descartamos utilizarlas, ahora hemos establecido el criterio de no recoger más de un 5% de obras traducidas por cada tipo de texto.

[2] Por ejemplo, Swiatkowska (2000), utiliza para su corpus de interjecciones dos tipos de registros: uno formal, las gramáticas y los diccionarios; y otro no formal, las obras de teatro y las obras alfaicónicas. Por su parte, el *Diccionario de onomatopeyas del cómic* de Gasca y Gubern, (2008), se basa específicamente en obras alfaicónicas.

4. ALMACENAMIENTO Y RECUPERACIÓN DE LA INFORMACIÓN

El procesamiento de la información se realiza mediante bases de datos relacionadas. Esta herramienta informática nos permite de una forma sencilla y rápida vincular datos, acceder y consultar campos preestablecidos, o editar registros específicos. No descartamos, en un futuro, cuando el corpus sea más amplio, utilizar gestores de bases de datos más complejos como *TermStar*, o *WordSmith* para análisis textuales. Por el momento, diseñamos dos tablas básicas de almacenamiento de datos a las que se han relacionado otras secundarias según la búsqueda de la información necesitada:

1) Las fichas bibliográficas de las obras vaciadas. En ellas, se registran de forma alfabética, el nombre del autor, el título, el año, el tipo de materia o subgénero en el que se clasifica, el país, si se trata de una obra traducida, etc.

2) La base de datos principal que comprende las distintas ocurrencias de las onomatopeyas ordenadas alfabéticamente. Dividimos esta tabla en los siguientes campos de información: el código de la obra en que se localiza la onomatopeya; el lema; si se trata, en principio, de una variante de otro lema; la ocurrencia de la onomatopeya; el gesto con que se enuncia o con que se describe; la fuente (del sonido) de la que emana; la descripción de la acción que conlleva; una anotación lingüística (si es un caso de sustantivación, una locución, si posee complementos...); y un comentario de tipo pragmático (si no es preciso su significado, si puede tratarse de una variedad diastrática, si hay alguna referencia determinada al contexto...).

5. CLASIFICACIÓN Y DESCRIPCIÓN DE DATOS

Los datos de las onomatopeyas obtenidos en función de su lema se clasifican cuando llegan a un total de 50 ocurrencias en otra base de datos que describe, de forma ya individual, si se trata o no de una variante, la clase de onomatopeya a la que pertenece y los tipos y subtipos de significado que la caracterizan. De hecho, un corpus es un instrumento valiosísimo para asignar distintos sentidos a las unidades léxicas (Sinclair, [1991] 1993: 65). Por ejemplo, en nuestro corpus, con las más de 150 ocurrencias que hallamos de la onomatopeya *¡Brrr!*, hemos establecido las siguientes caracterizaciones: 1. onomatopeya que indica la acción de temblar; 1.1 onomatopeya interjectiva que indica frío o miedo; 2. interjección onomatopéyica que indica rechazo, protesta, enfado o aburrimiento; 3. onomatopeya que indica la acción de bramar o gruñir; 4. onomatopeya que indica la acción de vomitar; 5. onomatopeya que indica la acción de un motor en marcha; y 6. onomatopeya que indica la acción de ventosear.

Evidentemente, para obtener estas descripciones ha sido necesario resolver, en primer lugar, una serie de problemas que concernían a la lematización, al tratamiento de las variantes o a los criterios de representatividad (v. *infra*).

6. Estado actual y futuro del corpus léxico de onomatopeyas

Hasta el momento, la fase de recogida y almacenamiento de datos alcanza el 17% del corpus previsto. El número recopilado de onomatopeyas (incluidas las variantes) suma unos 600 registros diferentes. A pesar de que la realización del trabajo es lenta, creemos que en un corto plazo podremos publicar y difundir este corpus de forma escrita (las divisiones y caracterizaciones de los lemas) y de forma digital (las ocurrencias localizadas). A largo plazo, desearíamos que el corpus estuviera en línea, que fuera posible realizar distintos tipos de búsqueda, y que fuera creciendo a manera de un corpus monitor de onomatopeyas.

7. Problemas afrontados

Un corpus léxico de onomatopeyas es una obra singular que requiere unas decisiones metodológicas particulares. En primer lugar, tuvimos que delimitar la caracterización lingüística de la onomatopeya con respecto a otros tipos de unidades colindantes; por ejemplo, ¿cómo debíamos tratar los casos de aliteraciones?

Otra de las cuestiones más difíciles fue la de elaborar un corpus de vaciado de obras que fuera capaz de juntar el número mayor de datos posibles abarcando de una forma exhaustiva y homogénea todo tipo de textos. En este sentido, todavía estamos abiertos a modificar las proporciones de los textos seleccionados, si comprobamos que no hemos reparado en algún subgénero textual en el que pudiéramos localizar onomatopeyas.

En cuanto a la interpretación de los datos, al ser un corpus que trataba textos completos, los resultados obtenidos podrían reflejar el estilo personal de un autor (Sinclair, [1991] 1993: 19)[3]. Para dotar al corpus de una representatividad más objetiva, decidimos que las onomatopeyas clasificadas y caracterizadas debían estar localizadas preferiblemente en 3 autores diferentes. Por otra parte, constatamos la dificultad a la hora de lematizar la distinta escritura de las onomatopeyas: utilización de grafías anómalas o vacilaciones en la grafía

[3] Este es el motivo por el que Fernández Ramírez, ([1951] 1986: 122-ss.), rechaza ciertas onomatopeyas utilizadas por autores literarios. Sin embargo, en Rodríguez Guzmán (2009: 56-57), comprobamos que era necesario poseer unos amplios corpus y contar con la poca fijación escrita de las onomatopeyas antes de caracterizarlas como creaciones individuales.

(entre signos o no de exclamación, con inicial mayúscula o con minúscula, sin hache o con hache, unidas o separadas, repetidas o no, etc.). Por otra parte, tuvimos que resolver las ocurrencias de diferentes variantes gráficas y fonéticas. En principio, partimos del presupuesto de García de Diego, (1968: 29), de considerar cada alfabetización de la onomatopeya como una unidad independiente. En una etapa posterior, en la ficha individual de cada onomatopeya, determinamos las distintas variantes mediante una frecuencia relativa que tomaba como término absoluto de comparación la variable más frecuente. Establecimos cuatro variantes: a) una variante poco usada con una frecuencia relativa entre el 0,2 y el 0,4; b) una variante usada (entre el 0,4 y el 0,6); c) una variante bastante usada (entre el 0,6 y el 0,8); y d) una variante muy usada (más del 0,8). Por último, en el corpus de onomatopeyas debemos prestar mucha atención a las obras traducidas, ya que, en muchas ocasiones, nos vamos a encontrar con calcos de otras lenguas (sobre todo en las obras alfaicónicas en las que predomina la influencia del inglés).

7.1. La transcripción de las onomatopeyas en los corpus

No querríamos concluir esta comunicación sin mencionar un problema que nos atañe de una forma indirecta. Debemos estar satisfechos puesto que, en general, en los corpus, se transcriben las onomatopeyas de igual forma que otras palabras. Sin embargo, todavía muchas veces se prefiere su descripción como ruidos extratextuales: *(tos)*, *(risas)*, *(eructo)*, *(bostezo)*, *(respiración)*. Este procedimiento de recurrir al paréntesis no puede justificarse por dificultades en la transcripción, ya que, en los corpus en los que se utilizan estas descripciones, sí que se transcriben otras onomatopeyas e interjecciones más complejas: *tch tch* `negación´ (Pino Moreno y Sánchez Sánchez, 1999); *ff* `soplido que indica rechazo´, *eeaee* `imitación onomatopéyica del mareo´ (Briz y grupo Val.Es.Co., 2002); o *brrr*, *nnn* (Fernández Juncal, 2005). Desde luego, en español, "el inventario lingüístico está falto de una rigurosa clasificación de interjecciones primarias", (Ávila Muñoz, 1999: 203), pero no por ello debemos dejar sin transcribir estos sonidos cuando el objetivo de los corpus es "reproducir lo más fielmente la conversación" (Briz y otros, 2002: 38). Al introducir estas descripciones, se obra de forma semejante a la escritura y función de las acotaciones de los textos teatrales. Éstas son indicaciones con una clara finalidad escénica que desaparecen en el escenario "al menos en su condición de signos lingüísticos" (Pérez Bowie, 2010: 5). Efectivamente, con las descripciones por medio de los paréntesis que realizan los corpus, se rompe la fluidez del diálogo de la conversación, se mezclan dos niveles enunciativos diferentes (el del personaje y el del acotador), y por último, se prescinde de una importante

información lingüística que nos obliga a los lectores a "construir imaginariamente una escena", (Bobes Naves, 1998: 814). Evidentemente, este no es el propósito de los corpus. Por otro lado, en muchas ocasiones, este tipo de onomatopeyas o interjecciones omitidas ya cuenta con una larga tradición lexicográfica y literaria: "A.a.a interjección del que ríe. [...] A.a.a voz confusa de mudos", (Nebrija, [1495] 1951: *s/v a.a.a*), por eso, no creemos que exista motivo alguno para no seguir una "fidelidad a lo hablado" o una "legibilidad de su puesta por escrito" en este tipo de palabras (Blanche-Benveniste y Jeanjean, 1987: 115, *apud* López Serena, 2007: 208)[4]. Defendemos de esta manera la propuesta de Bajo Pérez, (2007: 507), quien apuesta por transliterar o transcribir todas las onomatopeyas e interjecciones onomatopéyicas, a pesar de las dificultades que entrañe esta realización:

> Los inconvenientes no deberían justificar la preterición de las onomatopeyas: antes bien, las onomatopeyas interjectivas (o interjecciones onomatopéyicas) siempre deberían transliterarse (o en su caso, transcribirse), en lugar de ser descritas entre corchetes.

8. Conclusión

El corpus léxico de onomatopeyas españolas intenta cubrir un vacío en la lingüística de corpus y en los repertorios lexicográficos de la lengua española. Esta obra nos sirve para constatar y descubrir nuevos aspectos del lenguaje que hasta ahora no se han venido teniendo en cuenta. En la elaboración de dicho corpus, hemos perfilado directrices y hemos propuesto soluciones teóricas y metodológicas que pueden delimitar de forma congruente y precisa el objeto de estudio tratado. Esperamos que la realización de esta tarea concluya en breve y obtengamos una amplia base de datos para el análisis de este tipo de enunciados.

Referencias bibliográficas

Alvar López, M. (1999). Acerca de la interjección. En E. Forasteri Braschi (coord.) y J. Cardona, H. López Morales y A. Morales de Walters (eds.), *Estudios de literatura hispánica: homenaje a María Vaquero,* (pp. 22-55). San Juan: Universidad de Puerto Rico.

[4] En el caso de que se quieran considerar estas voces como ruidos ajenos al hablante por ser involuntarios (por ejemplo un tic, o el hipo), bastaría con señalarlo en una nota a pie de página como, de hecho, se realiza con cualquier anotación pragmática.

Ávila Muñoz, A. M. (1999). El valor pragmático de las interjecciones. En J. Fernández González, C. Fernández Juncal, M. M. Marcos Sánchez, L. Santos Río y E. Prieto de los Mozos (eds.), *Lingüística para el siglo XXI: III Congreso organizado por el Departamento de Lengua Española, 1999,* (pp. 201-208). Salamanca: Universidad de Salamanca.

Bajo Pérez, E. (2007). Carmen Fernández Juncal: Corpus de habla culta de Salamanca. Segovia: Fundación Instituto Castellano y Leonés de la lengua, 2005. *Moenia 13*, 503-509.

Bobes Naves, M. C. (1998). El discurso de la obra dramática: diálogo, acotaciones y didascalias. En J. C. de Torres Martínez y C. García Antón (coords.), *Estudios de literatura española de los siglos XIX y XX: homenaje a Juan María Díez Taboada* (pp. 812-820). Madrid: Consejo Superior de Investigaciones Científicas.

Borrego Nieto, J. (2002). Niveles de lengua y diccionarios. En J. L. Blas Arroyo, M. Casanova, S. Fortuño y M. Porcar (eds.), *Estudios sobre lengua y sociedad* (Vol. 9), (pp. 105-151). Castelló de la Plana: Universidad Jaume I.

Briz, A. y Grupo Val.Es.Co. (2002). *Corpus de conversaciones coloquiales*. Madrid: Arco Libros.

Fernández Juncal, C. (2005). *Corpus de habla culta de Salamanca*. Segovia: Fundación Instituto Castellano y Leonés de la Lengua.

Fernández Ramírez, S. ([1951] 1986). *Gramática Española*. Madrid: Arco Libros.

García de Diego, V. (1968). Estudio de las voces naturales. *Diccionario de voces naturales*. Madrid: Aguilar.

Gasca, L. y Gubern, R. (2008). *Diccionario de onomatopeyas del cómic*. Madrid: Cátedra.

Grammont, M. (1901). Onomatopées et mots expressifs. *Revue des Langues Romanes*, 97-158.

Haensch, G. y Werner, R. (dirs.). (2000a): *Diccionario español de Cuba*. Gredos.

Haensch, G. y Werner, R. (dirs.). (2000b): *Diccionario del español de Argentina*. Gredos.

López Serena, A. (2007). *Oralidad y escrituralidad en la recreación literaria del español coloquial*. Madrid: Gredos.

Karcevski, S. (1941). Introduction à l´étude de l´interjection. *Cahiers Ferdinand de Saussure 1*, 57-75.

Malkiel, Y. (1990). *Diachronic problems in phonosymbolism*. Amsterdam, Philadelphia: John Benjamins.

Matamala Ripoll, A. (2004). *Les interjeccions en un corpus audiovisual: descripció i representació lexicográfica*. Tesis. Barcelona: Pompeu Fabra. Disponible en http://www.tdx.cat/TDX-1003105-130347 [noviembre 2008]

Nebrija, E. A. ([1495] 1951). *Vocabulario Español-Latino*. Madrid: RAE.

Pérez Bowie, J.A. (2010). En torno a las disdascalias. Algunas aportaciones teóricas recientes. *En prensa*.

Pino Moreno, M. y Sánchez Sánchez, M. (1999). El subcorpus oral del banco de datos CREA-CORDE (Real Academia Española): Procedimientos de transcripción y codificación. *Oralia 2*, 83-138.

Rodríguez Guzmán, J. (2009). *Onomatopeyas. Lexicografía y metalexicografía*. Grado de Salamanca. Salamanca: Universidad de Salamaca.

Rodríguez Guzmán, J. (2010). La función textual de las onomatopeyas. *En prensa*.

Santos Río, L. (2003). *Diccionario de partículas*. Salamanca: Hispanolusa de Ediciones.

Sinclair, J. ([1991] 1993). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Świątkowska, M. (2000). *Entre dire et faire. De l´interjection*. Kraków: Wydawnictwo Uniwersytetu Jagiellonskiego.

Lancashire English in diachronic perspective: evidence from the Salamanca Corpus

JAVIER RUANO-GARCÍA
Universidad de Salamanca

"They'n done enough for me, hannot they?" he co'd eaut, dhroppin into dialect,
as o Lancashire chaps are olez ready to do, shuz heaw mony sorts o' talk they'n larn't.
The Works of James Trafford Clegg (1895)

Abstract
This paper looks at some of the Lancashire texts hitherto included in the Salamaca Corpus, with a view to illustrating how this might contribute to bridging some of the gaps still existing in the field of English diachronic dialectology. For this purpose, the paper provides a preliminary diachronic insight into morphological issues related to the use of was/were in the county of Lancashire.
Keywords: Lancashire English, was/were levelling, the Salamanca Corpus, diachronic dialectology

Resumen
Este trabajo pretende analizar parte del material del condado de Lancashire que se ha incluido hasta la fecha en el Corpus de Salamanca con el propósito de ilustrar el modo en que esta compilación puede contribuir a llenar algunos vacíos existentes en el campo de la dialectología inglesa diacrónica. Para ello, proporcionamos una primera aproximación diacrónica al uso de was y were en el condado de Lancashire.
Palabras clave: Inglés de Lancashire, regularización de was/were, el Corpus de Salamanca, dialectología inglesa diacrónica

1. INTRODUCTION

As is true of other regional varieties, our knowledge about early Lancashire speech is characteristically scarce. The linguistic history of English dialects is still distinguished by a relative lack of diachronic data representative of the period that extends from early modern English up to modern times (1500-1900). Whilst the increasing availability of textual corpora has enabled successful diachronic research into the history of standard English, variation in regional English dialects remains virtually unexplored. In fact, no diachronic compilations have hitherto been available to fill the *lacunae* still present in the field. For this reason, the Salamanca Corpus (henceforth SC) has been conceived as a repository of diachronic dialect material which might bridge some of the gaps still existing in the field.

This paper looks at some of the Lancashire texts included in the SC. My primary aim is to illustrate the validity of the corpus for dialect research, arguing that this compilation may serve as a missing link to expand the database of English diachronic dialectology. For this purpose, I will examine issues related to *was*/*were* levelling in the county. In particular, I will pay attention to the the spread of the plural stem throughout the paradigm (e.g. *I were*, *he were*), and to the chronology of the loss of the verbal *-n* in the past plural forms (i.e. *they weren* > *they were*).

2. THE SALAMANCA CORPUS: LANCASHIRE DATA

The SC is a long-term ongoing project which is currently been undertaken at the University of Salamanca. The lack of linguistic histories of English dialects is the main reason behind the compilation of the present corpus. It has been conceived as an electronic tool consisting of literary artefacts representative of the different pre-1974 regional varieties of (English) English. Its primary purpose is to illustrate linguistic features of English dialects, and place the texts at the disposal of scholars who may wish to get a more refined diachronic insight into provincial speech. Whilst they are not meant to detract from the linguistic validity of other text types, literary data may prove particularly useful for historical periods when regional materials are rather scarce. In addition, this kind of data may contribute to the analysis of how and why specific varieties became enregistered throughout history, this being a recent field of investigation in sociolinguistics and dialectology (Beal, 2009).

The compilation of texts has followed specific criteria. Chronologically, documents written between 1500 and 1900 have been considered. They have been arranged into distinct sections corresponding with three broad time periods – 1500-1700, 1700-1800 and 1800-1900 –, which will allow for the comparative study of certain features across time. From a diatopic perspective, the corpus aims to present documents representative of pre-1974 dialects. Although a balanced number of texts from the different counties would be expected, areas such as Essex or Buckinghamshire suffer from a relative lack of vernacular literature, making it complex to retrieve historical data about these dialects. Given this, the diatopic information now provided by the SC is somewhat unbalanced, although texts representative of counties such as those mentioned have also been found. Typologically, the corpus is literary restricted. Different genres have been considered irrespective of their literary value. Needless to say, a distinction has been made between examples of literary dialect and dialect literature (see Shorrocks, 1996: 386 for a clear distinction between literary dialects and

dialect literature). Again, it is impossible to offer a balanced number of localised documents, since the amount of regionally-anchored material dating to early periods is significantly scarce. The examples representative of each text type have not been therefore selected randomly, but according to their availability, which is in turn dependant on the literary practices of each time period (see further García-Bermejo, Sánchez-García and Ruano-García, 2009; García-Bermejo, 2010).

As for the Lancashire materials, it comes as no surprise that early evidence about Lancashire English is relatively scant. Along with the sparse data which can be gleaned from glossaries, pronunciation dictionaries or survey-books, literary texts are fruitful early sources of information. Documents such as the hitherto unpublished 'A Lancashire tale' (c.1690-1730) and the different editions of John Collier's *A View of the Lancashire Dialect* are outstanding early witnesses to the dialect of the county (Haworth, 1920; Ruano-García, 2010; Wagner, 1999; Whitehall, 1929). Unfortunately, specimens purporting to reproduce the dialect of the area are not particularly abundant during the early modern period and the eighteenth century. In fact, Shorrocks (1999a) explains that it was during the nineteenth century that the strongest tradition of dialect literature developed in the county.

Given this, the earliest localised records which have been compiled in the SC correspond with the two mentioned above along with an unpublished anonymous ballad that begins 'Robin an's Gonny' (c.1690-1730). So far, additional documents from the eighteenth century are not numerous. For example, some verse dialogues written by John Byrom (1773), Henry Clarke's *The School Candidates* (1788) or Robert Walker's *Plebeian Politics* (1798). As for the nineteenth century, literary texts have been easier to find. The available digitised material comprises works of major Lancashire figures such as Benjamin Brierley, Edwin Waugh or Oliver Ormerod. Minor writers have also been considered: Roper Robinson or James Bowker.

It must be acknowledged that evidence from the first half of the nineteenth century has also been hard to find. Shorrocks (1999a: 89-90) avers that "By 1860, dialect literature was appearing in truly large quantities, and continued to do so for the rest of the century and the first quarter or third of the twentieth century. Its burgeoning between 1850 and 1860 coincided with a marked improvement in material prosperity". As such, the hitherto available Lancashire texts from the first half of the nineteenth century are still scarce. In particular, only seven out of the seventy-two nineteenth-century documents compiled belong to the first half of the 1800s. As can be seen in Figure 1, there are certain periods which are still

underrepresented in the corpus (see also Table 1 in section 3.1. below). Current research aims to bridge the present gaps.

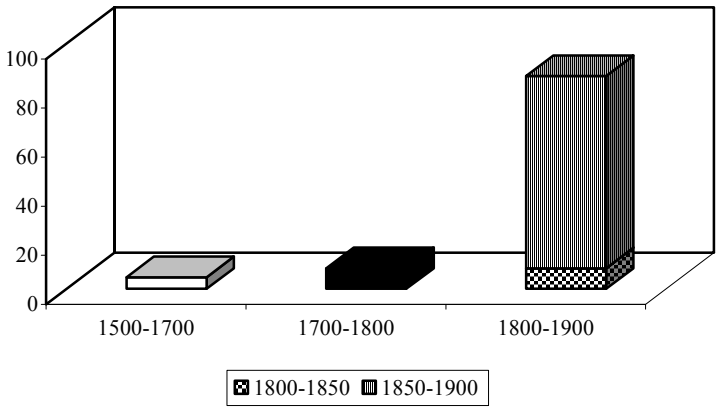


Figure 1: Chronological distribution of the Lancashire texts so far included in the SC.

3. *WAS* / *WERE* IN LANCASHIRE ENGLISH: A BRIEF DIACHRONIC SURVEY

3.1. Primary data

As indicated above, the primary aim of this paper is to illustrate the linguistic potential of the SC by means of a diachronic approach to *was*/*were* levelling in Lancashire English. For this purpose, a selection of the material has been made. Firstly, I have found it necessary to consider cases of dialect literature only, mainly because these texts are written entirely in dialect and, therefore, occurrences of *was* and *were*, as used in the county, are expected to be higher. Consequently, preference has been given to those documents written by natives to the area. As dialect speakers, the writers are supposed to have a first-hand knowledge of the variety represented, their texts bearing thus faithful witness to the dialect.[1] Next, only prose texts and dialogues, some of them written in verse, have been taken into account.[2] Finally, I have tried to provide a balanced sample of the material, although this has not been totally possible given the little number of texts from the early modern period and the 1700s. In fact, the seventeenth-century texts collected are examples of literary dialect, which has made this preliminary approach concentrate on the period 1700-1900.[3] In like manner, only a few

[1] Needless to say, the anonymous pieces which are clear examples of dialect literature have likewise been taken into account, although their authorship still remains unknown.
[2] This is not meant to suggest that dialect drama is not valid for this analysis. Yet, the compilation of the corpus has been thus far mainly concentrated on prose and verse. Drama remains to be investigated.
[3] See that the two regional specimens dated to c.1690-1730 have been included in the eighteenth-century section for obvious reasons.

nineteenth-century documents have been taken into consideration, so as to avoid an excessive nineteenth-century bias in the results. They have been chosen in a manner that attempts to prevent idiosyncratic traits and individual practices. In other words, only works written by different authors have been examined. It is worth noting that the nineteenth-century documents compiled which are representative of dialect literature date to the second half of the 1800s. In total, the material selected comprises thirteen texts which amount to c.100,000 words (see Appendix 1). Although this represents a considerably small sample, it seems to be statistically significant for the present purpose. Table 1 shows the number of texts and words for each time period considered.[4]

Table 1: Selected Lancashire documents examined: number of texts and words analysed.

Time span	N texts	N words
1700-1800	9	17,154
1700-1750	3	5,007
1750-1800	6	12,147
1800-1900	6	80,956
1850-1875	2	15,497
1875-1900	4	65,459
Total	13	98,110

3.2. Theoretical background

In her illustrative chapter on the morphology and syntax of northern English dialects, Joan Beal (2005: 122) comments that "Accounts of the traditional dialects of Yorkshire and Lancashire (Wright 1892; Ellis 1869-1889) suggest that the typical pattern in these areas was one in which *were* occurred with all subjects, singular and plural". From this, one could assume that the past plural form of *to be* was levelled throughout the paradigm in Lancashire speech since early times, which would shed diachronic light upon Shorrock's (1999b: 168-169) parallel modern findings for the Bolton area.

As is well known, there exists a plethora of studies approaching *was*/*were* variation in British and overseas Englishes. Whilst most of the hitherto undertaken research has been focused on the present-day setting (Anderwald, 2001; Britain, 2007; Cheshire and Fox, 2009; Hollman and Siewierska, 2006: 25-26; Pietsch, 2005; Tagliamonte 1998; among others), the available historical information is not very abundant (Kytö, Grund and Walker, 2007; Nevalainen, 2006; or Visser, 1963). Suffice it to say that *was*/*were* alternation in English is documented as far back as the Middle English (ME) period in view of the large number of

[4] See that different subperiods have been considered for each century given the lack of material from the first

alternative forms attested in the record (Smith and Tagliamonte, 1998: 106-108). As for the North, Pietsch (2005: 149-150) holds that *was*/*were* variation had an analogous development to that of the Northern Subject Rule, *was* being used with all subjects except *I*, *you*, *we* and *they* in verb adjacent position. The following example taken from Kytö *et al*. (2007: n.p.) may suffice to illustrate this:

> (1) *The words was* spoken, *they were* married, but *they*, he understood, *was* married [my italics][5]

However, this alternation presents different patterns across the North in view of the data documented in the traditional dialects covered by the *Survey of English Dialects* (1962-1971) (henceforth *SED*). Actually, drawing on the *SED* findings, Pietsch (2005: 150-151) distinguishes different areas. Firstly, the Central North, covering Cumberland, Northumberland, Westmorland and Durham, where the use of *was* and *were* is similar to that of *is* and *are*, *was* being used throughout the singular, *were* being used throughout the plural. Yet, *was* is also licensed in plural contexts by the Northern Subject Rule (see also Beal, 2005: 122). Secondly, an area including southern Lancashire, southwestern Yorkshire and Derbyshire where *were* is strongly preponderant for all subjects singular and plural (see also Britain, 2007: 91). Finally, a transitional zone between the above mentioned areas which comprises northern Lancashire and the northeastern half of Yorkshire where the past plural *were* is likewise used in singular environments, although on a less frequent basis. Clearly, the six northern counties show differences with regard to the use of *was*/*were*. Either a tendency towards levelling in the paradigm or alternation are in the main observed.[6]

From what has been said thus far, it is clear that accounts both of traditional and modern Lancashire – Ellis (1869-1889), the *SED*, Shorrocks (1999b) – agree that some areas were distinguished by *was*/*were* levelling, this being realised in the generalisation of *were* to the singular, thus neutralising the singular-plural distinction characteristic of the standard. The question raised in this paper is whether the past tense of *to be* has always been distinguished by *were*-levelling and, thus, neutralisation of number distinction in Lancashire.

---

half of the 1800s. See further note 9 below.

[5] In what follows, cases of *was* and *were* in the examples given will be italicised. Given the difficulty of some of the passages written in dialect, literal translations have been provided in brackets under each example.

[6] This is no doubt an oversimplification of the use of *was*/*were* in the North. As Tagliamonte (1998: 156) puts it, variation between *was* and *were* seems not to have been random, but to have followed "identifiable and consistent pattern[s]". A detailed account of the patterns of *was*/*were* variation is beyond the scope of this paper. See further Tagliamonte (1998), Beal (2005: 122-123), Pietsch (2005: 149-161) or Nevalainen (2006: 359-367).

3.3. Survey and discussion of the data

As pointed out above, levelling of *were* seems to have been characteristic of southern Lancashire. In fact, Shorrocks (1999b: 168-169) reports that *were* is used for both singular and plural contexts in the Bolton area, in the South-East of the county, now part of Greater Manchester. The texts selected for this analysis present renderings of the dialect of the South of the county, so as to test whether the singular-plural distinction neutralised by *were*-forms was a distinctiveness of the area before it has been so far recorded.

In her analysis of Collier's Lancashire dialect, Wagner (1999: 201) claims that it was during the early modern period that "generalisations of the past plural stem to the singular" took place. This might have been so, since a summary look at the Linguistic Profiles (LPs) of Lancashire included in the LALME evince that in late ME *was* and *were* conformed to a very great extent to present-day patterns. In fact, a 100% of the LPs in which both singular and plural forms are recorded show that *was* and *were* were used for the singular and plural, respectively.[7] The documentary *lacunae* existing from late ME to Collier's work render it complex to know whether Wagner's contention is historically accurate. Indeed, Kytö *et al*. (2007) have not recorded Lancashire data from this time span, but from the 1700s. Similarly, the SC does little to bridge this gap, for, as stated above, the early modern Lancashire material consists of literary dialects written by non-natives to the area. However, taking the evidence supplied by the literary records dated circa 1690-1730, one could think that Wagner's statement seems not to be totally exact. Actually, data extracted from 'A Lancashire tale' and 'Robin an's Gonny' suggest that *was* was relatively predominant over *were* in singular contexts in the late seventeenth and early eighteenth century. In particular, 14 tokens of *was* against 4 of *were* have been found. As shown below, *was* and *were* are attested with NPs and personal subject pronouns:

(2) Th' Monn's Cwote *wur* a Grey (c.1690-1730, 'A Lancashire Tale')
[The man's coat was grey]

(3) When he *wus* awar o three men come or th'… (c.1690-1730, 'Robin an's Gonny')
[When he was aware of three men [who] came over the…]

[7] The rest of the LPs either record cases of *was* or of *were*. These do also manifest that *was* appeared in singular contexts, whilst *were* was used for plural environments. In what follows, *was* and *were* are used to refer to the singular and plural forms of *to be*, although the Lancashire data analysed show different spellings. A preference for <u>-spellings is attested in the corpus: *wus*, *wur*, *wuren*, *wur'n*, etc.

Interestingly, the data indicate that *were* in singular environments was more frequently used when functioning as an auxiliary in passive constructions: three of the four occurrences recorded testify to this fact. Consider the following:

(4) His bond *wur* teed wuth a Congerton-Pwoint (c1690-1730, 'A Lancashire Tale')[8]
[His band was tied with a Congorton point]

Suffice it to say that this does not provide enough evidence so as to make generalisations of any kind. Still, as early testimonies to the language of the county, these data should be taken into account.

Table 2 below shows, however, that preference for *were* in singular contexts seems to have been the rule in the eighteenth century, which tallies with Wagner's argument. Although the existence of *was*-forms could be taken as indicative of *were* not having been fully generalised by then, it is worth remarking that most of them have been attested in one single document: Byrom's Lancashire dialogue (1773). An individual bias might probably explain this.

Table 2: Distribution of *was*/*were* in singular contexts in the 1700s: raw figures, normalised frequencies (per 1,000 words) and total percentages[9]

	1700-1750			1750-1800		
	was	*were*		*was*	*were*	
		were	contracted		*were*	contracted
NP	4	31		9	43	
	(0.79)	(6.19)		(0.74)	(3.53)	
I		10	44	3	6	13
		(1.99)	(8.78)	(0.24)	(0.49)	(1.07)
you				1	2	
				(0.08)	(0.16)	
he, she, it	9	24	13	3	33	16
	(1.79)	(4.79)	(2.59)	(0.24)	(2.71)	(1.31)
existential	1	3			16	
there	(0.19)	(0.59)			(1.31)	
relative pn.		4			4	
		(0.79)			(0.32)	
pronoun		16		2	12	
		(3.19)		(0.16)	(0.98)	
Total	14	145		18	145	
%	8,8	91,2		11,04	88,9	

[8] See that no examples of passive constructions have been found in 'Robin an's Gonny'. In this, all examples of the singular of the past tense of *be* are represented by *was* forms.

[9] As shown in Tables 2 and 3, the corpus evidence has been divided into subperiods, so as to illustrate more clearly the prevalence of *were*-forms. In like manner, Tables 4 and 5 follow the same pattern, with a view to clarifying the decline of the verbal *-n* in past plural environments. As data from the first half of the 1800s has not been compiled yet, the nineteenth-century figures have been arranged into two sections of twenty-five years each. See that normalised frequencies are given in brackets.

Clearly, there is a strong preponderance of *were* for the singular in the eighteenth-century material, suggesting that all the types of subjects considered favoured the plural stem. Sometimes, adjacent subject pronouns invite contraction, as in (6):

(5) o Butcher *wur* ith' Eleheause (1748, *A View of the Lancashire Dialect*)
[a butcher was in the ale house]

(6) When I'*re* a book-keeper (1788, *The School Candidates*)
[When I was a book-keeper]

(7) for ther' *wur* oz little onnisty at furst, oz ther'*wur* onnor at th' last (1798, *Plebeian Politics*)
[for there was as little honesty at first, as there was honour at (the) last]

As for the nineteenth century, the data recorded in Table 3 suggest an identical tendency to use *were* for singular subjects and singular existentials. In fact, a 100% of *were*-forms have been attested, likewise indicating that *were* was frequently used with (non-) adjacent subject pronouns, NPs, relative pronouns with singular antecedents, etc. By way of illustration:

(8) th' furst thing us aw did *wor* to set… (1856, *O full tru un pertikler okeawnt...*)
[the first thing that I did was to set…]

(9) I ne'er thowt uv o'th' misery ther *wur* at whoam (1865, *Jim Wilson's Resolve*)
[I never thought of all the misery there was at home]

(10) Th' day as they should o come back *wir* varra misty (1883, *Goblin Tales*)
[The day that they should have come back was very misty]

Table 3: Distribution of *was*/*were* in singular contexts in the 1800s: raw figures, normalised frequencies (per 1,000 words) and total percentages

	1850-1875			1850-1875		
	was	*were*		*was*	*were*	
		were	contracted		*were*	contracted
NP		121			107	
		(7.8)			(1.63)	
I		169	17		28	1
		(10.9)	(1.09)		(0.42)	(0.01)
he, she, it		192	17		110	
		(12.38)	(1.09)		(1.68)	
existential		30			29	
there		(1.93)			(0.44)	
relative		37	1		3	
pn.		(2.38)	(0.06)		(0.04)	
pronoun		23			20	
		(1.48)			(0.3)	
Total	0	607		0	298	
%	0	100		0	100	

Evidence pertaining to plural contexts shows that only *were*-forms were used in the county both in the eighteenth and nineteenth centuries, irrespective of the type of subject examined, as in (11) and (12):

(11) His Principles *wurn* of another Mack (1762, *Lancashire Dialogue*)
[His principles were of another kind]

(12) Aw did just happen to know which they *wur* then (1869, *Ab-o'the-yate...*)
[I did just happen to know who they were then]

Yet, as these examples and tables 4 and 5 below indicate, there is a difference between the periods considered in that the eighteenth-century data manifest a strong preference for forms marked for plurality, whilst the 1800s data point in a somewhat different direction.

John Collier asserted in his prefatory remarks to the *Miscellaneous Works of Tim Bobbin* (1775) that one of the most salient Lancashire features was the plural *-(e)n* inflection: "The Saxon Termination *en* is generally retained but mute; *hal'n*, *lov'n*, (...)" (fols.A2-A2v). As is well known, the verbal *-n* for the present indicative plural dates back to ME, being characteristic of the West Midlands, and has been preserved in some areas of Derbyshire, Cheshire or Staffordshire (Wright, 1905: §435; Upton, Parry and Widdowson, 1994: 492). The verbal *-n* was likewise used for the past plural of *to be* in Lancashire or western Yorkshire, still persisting in a small relic area of the North-West Midlands (Orton, Sanderson and Widdowson, 1978: Maps 21-23). The Lancashire texts analysed back this fact, as plural forms such as *wuren*, *wurn*, *wur'n*, etc. are documented, especially during the 1700s.

Table 4: Distribution of *were* / *weren* in plural contexts in the 1700s: raw figures, normalised frequencies (per 1,000 words) and total percentages

	1700-1750			1750-1800		
	weren		*were*	*weren*		*were*
	-n, -'rn, -r'n	contracted		*-n, -'rn, -r'n*	contracted	
pl. NP	3 (0.59)			17 (1.39)		1 (0.08)
sing. NP + sing. NP	1 (0.19)			2 (0.16)		1 (0.08)
we						
you	2 (0.39)	1 (0.19)				
they	1 (0.19)	4 (0.79)	1 (0.19)	6 (0.49)	12 (0.98)	1 (0.08)
exist. *there*						7
relative pn.				11 (0.9)		1 (0.08)
pronoun			1 (0.19)			1 (0.08)
Total	7	5	2	36	12	12
%	85,7		14,3	80		20

Table 5: Distribution of *were / weren* in plural contexts in the 1800s: raw figures, normalised frequencies (per 1,000 words) and total percentages

	1850-1875			1875-1900		
	weren		*were*	*weren*		*were*
	-n, -'rn, -r'n	contracted		*-n, -'rn, -r'n*	contracted	
pl. NP			49 (3.16)			40 (0.61)
sing. NP + sing. NP			4 (0.25)			7 (0.1)
we		14 (0.9)	4 (0.25)		1 (0.01)	7 (0.1)
you		1 (0.06)				
they	7 (0.45)	36 (2.32)	6 (0.38)		2 (0.03)	21 (0.32)
exist. *there*			10 (0.64)			6 (0.09)
relative pn.			21 (1.35)			
pronoun			4 (0.25)			2 (0.03)
Total	7	51	98	0	3	83
%	37,2		62,8	3,5		96,5

Clearly, the type of subject does not seem to have constrained the use of either plural or zero marked forms. Rather, as indicated in Table 6 and illustrated in Figure 2, the marked differences observable between the eighteenth and the nineteenth century may be due to the gradual loss of verbal *-n* during the 1800s. It is worth stressing, however, that an important number of the forms marked for plurality during the 1800s were those contracted when preceded by subject pronouns, as in (13). It appears likely that contracted forms may have favoured the preservation of verbal *-n* so as to distinguish the present from the past indicative, as in (14). This seems to have been in decline during the last quarter of the nineteenth century too.

(13) an' *we'rn* gettin' eaut o'th' seet o' lond (1869, *Ab-o'the-yate...*)
[and we were getting out of the sight of [the] land]

(14) *We're* gettin' very nee to th' fur end (1869, *Ab-o'the-yate...*)
[We are getting very near to the far end]

Table 6: Decline of the verbal *-n* for the past plural: raw figures and percentages

	1700-1750		1750-1800		1800-1850	1850-1875		1875-1900	
	N	%	N	%		N	%	N	%
pl. *-n*	12	85,7	48	80	?	58	37,2	3	3,5
pl. *-ø*	2	14,3	12	20	?	98	62,8	83	96,5
Total	14	100	60	100	?	156	100	86	100

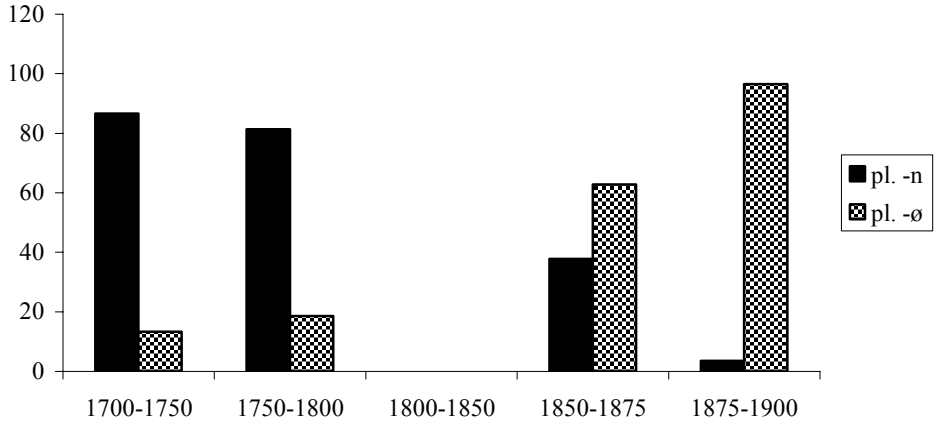


Figure 2: Decline of the verbal *-n* for the past plural: percentages.

In view of this, it appears thus likely that the gradual decline of verbal *-n* triggered the full levelling of the plural form throughout the paradigm. In other words, it seems that the generalisation of *were*-forms to the past tense of *to be* was not fully realised until the second half of the nineteenth century. The higher frequency of marked plurals during the eighteenth

century suggests that the singular-plural distinction was not neutralised, although the *were* stem had spread to the singular.

4. CONCLUDING REMARKS

This paper has endeavoured to illustrate the linguistic potential of dialect data as preserved in old literary documents. In particular, it has been my concern to exemplify the diachronic linguistic possibilities of the SC for shedding light upon issues which have been insufficiently documented. For this purpose, an examination of some of the corpus texts has been made to further our understanding of *was*/*were* levelling in Lancashire. As such, the data, though unbalanced at some points, suggest that the plural stem was generalised to the singular since, at least, the early eighteenth century. In fact, a high predominance of *were*-forms has been observed in singular environments both in the 1700s and the 1800s, to the extent that *was* appears not to have been used during the nineteenth century at all. As expected, *were* was used in plural contexts too. Yet, a difference has been detected as regards the existence of forms with *-n* or zero marking in the corpus. The former, reflecting a West Midlands typical verb plural inflexion, were more frequently used during the eighteenth century. By contrast, zero marked predominated over *-n* forms in the second half of the 1800s. This suggests that it was not until verbal *-n* declined that the levelling of the past plural form was completed, thus engendering the neutralisation of the number distinction which had been retained by the *-n* plural. Thus, Lancashire seems not to have always been defined by a pattern in which *were* was used for both singular and plural contexts.

In sum, I hope to have cast some light upon the diachrony of Lancashire English, arguing that views contending that the area has traditionally been distinguished by *were*-levelling for singular and plural subjects need further historical consideration. As Tagliamonte (1998: 156) holds, "any attempt to explain the current state of the *was*/*were* alternation in English is dependent on a diachronic perspective in order to contextualize contemporary patterns of variation". It remains a question for future undertaking to elucidate the rationale behind the spread of *were* forms and the loss of verbal *-n*. There is hope that further data from other text types and other periods which the SC has not covered yet will help towards clarifying the chronology of the levelling and the speakers who motivated the regularisation.

APPENDIX 1. SELECTED LANCASHIRE TEXTS: NUMBER OF WORDS

1700-1800

Anon. (c.1690-1730). 'A Lancashire tale'. 995 words.

Anon. (c.1690-1730). 'Robin an's Gonny...'. 322 words.

Byrom, J. (1773). *Miscellaneous Poems* [selected material]. 4,496 words.

Clarke, H. (1788). *The School Candidates* [selected material]. 483 words.

Collier, J. (1748). *A View of the Lancashire Dialect*. 3,690 words.

Collier, J. (1763). *Tim Bobbin's Toy-Shop Open'd or, his Whimsical Amusements* [selected material]. 619 words.

Walker, R. (1798), *Plebeian Politics*. 6,549 words.

1800-1900

Bowker, J. (1883). *Goblin Tales of Lancashire*. 29,115 words.

Brierley, B. (1869). *Ab-o'the-yate at the Isle of Man*. 5,423 words.

Clegg, J. T. (1895). *The Works of John Trafford Clegg* [selected material]. 27,779 words.

Fergusson, Ch. (1865). *Jim Wilson's Resolve and What Came Out of It: A Lancashire Temperance Tale*. 3,142 words.

Ormerod, O. (1856). *O full, tru, un Pertikler Okeawnt...* 13,513 words.

Waugh, E. (1855). *Sketches of Lancashire Life and Localities* [selected material]. 1,984 words.

APPENDIX 2. NUMERICAL INFORMATION

Time span	N texts	%
1500-1700	4	4,8
1700-1800	7	8,4
1800-1900	72	86,7
1800-1850	7	8,4
1850-1900	65	78,3
Total	83	99,9

Chronological distribution of the Lancashire texts so far included in the SC

REFERENCES

Anderwald, L. (2001). *Was*/*Were* Variation in Non-Standard British English Today. *English World-Wide*, *22* (1), 1-21.

Beal, J. C. (2005). English Dialects in the North of England: Morphology and Syntax. In E. Schneider *et al*. (Eds.), *A Handbook of Varieties of English. Vol. 2 Morphology and Syntax,* (pp. 114-141). Berlin, MY: Mouton.

Beal, J. C. (2009). Enregisterment, Commodification, and Historical Context: "Geordie" versus "Sheffieldish". *American Speech*, *84* (2), 138-156.

Britain, D. (2007). Grammatical Variation in England. In D. Britain (Ed.), *Language in the British Isles* (pp.75-104). Cambridge: Cambridge UP.

Cheshire, J. & Fox, S. (2009). *Was*/*Were* Variation: A Perspective from London. *Language Variation and Change*, *21*, 1-38.

Collier, J. (1775). *The Miscellaneous Works of Tim Bobbin*. Manchester: Printed for the Author.

Ellis, A. J. (1869-1889). *On Early English Pronunciation*. Oxford: Basil Blackwell.

García-Bermejo, M. F. (2010 fc.). Towards a History of English Literary Dialects and Dialect Literature in the 18th and 19th Centuries: The Salamanca Corpus. In B. Heselwood & C. Upton (Eds.), *Papers from Methods XIII*. Frankfurt am Main: Peter Lang.

García-Bermejo, M. F., Sánchez-García, P. & Ruano-García, J. (2009). The Literary Representation of Vernacular Speech from Early Modern English to the Early 20th Century: The Salamanca Corpus as a Source for Diachronic Dialectology. Paper presented at the *Workshop on Studying the Representation of Dialect in Literature: How? and Why?* Sheffield, September, 25-26.

Haworth, P. (1920). The Language of Tim Bobbin. *Manchester Quarterly*, 113-126.

Hollman, W. & Siewierska, A. (2006). Corpora and (the Need for) Other Methods in a Study of Lancashire Dialect. *Zeitschrift für Anglistik und Amerikanistik*, *54* (2), 203-216.

Kytö, M., Grund, P. & Walker, T. (2007). Regional Variation and the Language of English Witness Depositions 1560-1760: Constructing a 'Linguistic' Edition in Electronic Form. In P. Pahta *et al*. (Eds.), *Studies in Variation, Contact and Change in English 2: Towards Multimedia in Corpus Studies*. Helsinki: VARIENG. Retrieved from <http://www.helsinki.fi/ varieng/journal/Volumes/02/kyto_et_ al/>.

LALME = A Linguistic Atlas of Late Mediaeval English. McIntosh, A. *et al*. (Eds.) (Vols.1-4). Aberdeen: Aberdeen UP.

Nevalainen, T. (2006). Vernacular Universals? The Case of Plural *Was* in Early Modern English. In T. Nevalainen *et al*. (Eds.), *Types of Variation: Diachronic, Dialectal and Typological Interfaces* (pp.351-369). Amsterdam: Benjamins.

Orton, H., Dieth, E., Halliday, W., Barry, M. V., Tilling, P. M. & Wakelin, M. F. (Eds.). (1962-1971). *Survey of English Dialects (B) The Basic Material* (Vols. 1-13). Leeds: E.J. Arnold & Son Limited.

Orton, H., Sanderson, S. & Widdowson, J. (1978). *The Linguistic Atlas of England*. London: Routledge.

Pietsch, L. (2005). "*Some do and some doesn't*": Verbal Concord Variation in the North of the British Isles. In B. Kortmann *et al*. (Eds.), *A Comparative Grammar of British English Dialects: Agreement, Gender, Relative Clauses* (pp.125-209). Berlin, NY: Mouton.

Ruano-García, J. (2010 fc.). *I'll tell o how Gilbert Scott sowd is mere berry*: 'A Lancashire Tale' as a Source for Lancashire Speech in the Late Seventeenth and Early Eighteenth Century. In B. Heselwood & C. Upton (Eds.), *Papers from Methods XIII*. Frankfurt am Main: Peter Lang.

Shorrocks, G. (1996). Non-Standard Dialect and Popular Culture. In J. Klemola *et al*. (Eds.), *Speech Past and Present: Studies in English Dialectology in memory of Ossi Ihalainen* (pp. 385-411). Frankfurt am Main: Peter Lang.

Shorrocks, G. (1999a). Working-Class Literature in Working-Class Language: The North of England. In T. Hoenselaars and M. Buning (Eds.), *English Literature and the Other Languages* (pp. 87-96). Amsterdam: Rodopi.

Shorrocks, G. (1999b). A Grammar of the Dialect of the Bolton Area. Part II: Morphology and Syntax. Frankfurt am Main: Peter Lang.

Smith, J. & Tagliamonte, S. (1998). 'We *were* all thegither...I think we *was* all thegither': *Was* Regularization in Buckie English. *World Englishes*, *17* (2), 105-126.

Tagliamonte, S. (1998). *Was*/*Were* Variation Across the Generations: View from the City of York. *Language Variation and Change*, *10*, 153-191.

Upton, C., Parry, D., Widdowson, J. A. (Eds.). (1994). *Survey of English Dialects: The Dictionary and Grammar*. London: Routledge.

Visser, F. Th. (1963). An Historical Syntax of the English Language. Part I: Syntactical Units with One Verb. Leiden: Brill.

Wagner, T. (1999). John Collier's 'Tummus and Meary'. Distinguishing Features of 18th-Century Southeast Lancashire Dialect – Morphology. *Transactions of the Philological Society*, *2* (C), 191-205.

Whitehall, H. (1929). Tim Bobbin Again. *Philological Quarterly*, 8, 395-405.

Wright, J. (1905). *English Dialect Grammar*. Oxford: Henry Frowde.

Achieving representativeness through the parameters of spoken language and discursive features: the case of the Spoken Turkish Corpus

ŞÜKRIYE RUHI[1]
HALE IŞIK-GÜLER
ÇILER HATIPOĞLU
BETIL ERÖZ-TUĞA
DERYA ÇOKAL KARADAŞ
Middle East Technical University

Abstract
In this paper we overview the ongoing debate on achieving representativeness in general spoken corpora with the purpose of proposing a model for spoken corpora design and construction workflows. The proposal is illustrated in the context of an ongoing implementation for the Spoken Turkish Corpus, a corpus that will consist of one million words of present-day Turkish spoken in Turkey in its initial stage. The paper proposes a cyclic workflow and design scheme that is based on the principles of an "agile" corpus design and annotation system (Voorman and Gut, 2008), and argues that a three-pronged set of feature criteria, namely, demographic, contextual, and discursive features can be fruitfully combined to monitor and achieve representativeness. The paper discusses the underlying principles in the design scheme and outlines the metadata features of the web-based corpus management system, which utilizes and complements EXMARaLDA tools (Schmidt, 2004) in corpus construction and monitoring.
Keywords: spoken corpus design criteria, representativeness, metadata, discursive features, web-based corpus management

Resumen
En el presente trabajo se examina el debate en curso sobre adquirir representatividad en corpus generales de lengua hablada con el objetivo de proponer un modelo para el diseño y trabajo de construcción de corpus orales. La propuesta está enmarcada dentro de un trabajo en curso que se encuentra en su fase inicial, la lengua implementación del Spoken Turkish Corpus, un corpus que constará con un millón de palabras de la lengua turca actual hablada en Turquía. Este trabajo propone un método de trabajo cíclico y un esquema de diseño basado en unos criterios fundados en un conjunto de tres características, a saber, demográficas, contextuales y discursivas, pueden estar perfectamente combinadas para observar y conseguir representatividad. Este trabajo trata de los principios subyacentes en el esquema de diseño y esboza los rasgos de los metadatos del sistema de gestión de corpus basado en la web, que utilizan y complementan las herramientas EXMARaLDA (Schmidt, 2004) en la construcción y seguimiento del corpus.
Palabras clave: diseño de corpus hablado criterios, representatividad, metadatos, rasgos discursivos, gestión de corpus basado en la web

[1] Corresponding author. Contact: Dept. of Foreign Language Education, Faculty of Education, Middle East Technical University, İnönü Blvd., 06531 Ankara, Turkey, e-mail: sukruh@metu.edu.tr, sukriyeruhi @gmail.com

1. INTRODUCTION[2]

Achieving representativeness, balance and comparability in corpus construction are three requirements that have engaged and are still engaging scholars in debate as to how best to approach these issues in terms of theory, methodology, and the dire practicalities of corpus compilation, especially since Biber's (1993) seminal article on representativeness (see, e.g., Leech, 2007; Váradi, 2001). Two central points of this debate concern approaches to sampling (proportional vs. stratified) and the conceptualization of frequency of communication types. Underlying the various debates is the fundamental question: What is it that one expects to achieve with corpus construction? Is it to produce a resource that lays open a maximal view on language variation (Biber, 1993), or is it to produce a resource in the standard statistical sense of representing language, based on demographic criteria (Váradi, 2001). While the two appear to be in opposition, both goals translate themselves within corpus linguistics into the expectation that one should be able to use the resource to make generalizations about the language (Leech, 2007).

Whilst the dust has certainly not settled, a less frequently broached issue is how to mesh features emerging from demographic, contextual, topical (Crowdy, 1993), and the more newly introduced "situation-governed" categories (Čermák, 2009: 116) within a framework that is responsive to the demands of representativeness. After a very brief overview of proposals in this regard, this paper argues in favor of a three-pronged set of features to achieve representativeness, and illustrates its implementation within the context of the Spoken Turkish Corpus (STC).

2. YARDSTICKS IN SAMPLING

The following sets of criteria and sampling procedures have been proposed and used in corpus compilation (Crowdy, 1993; Čermák, 2009):

1. Demographic
2. Contextual features
3. Topical
4. Situation-governed

[2] This study and STC is financed by a research grant from the Turkish Scientific and Technological Research Institution (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu, TÜBİTAK, Grant No. 108K283). We are deeply grateful to Dr. Thomas Schmidt and Dr. Kai Wörner for their support in the construction of STC.

The spoken component of BNC, for example, is based on the first three criteria, and the coding of texts according to the second and third criteria is reflected as genres. While the first set of criteria is geared toward representing geographical variation, the second is geared to capture register variation. If one works with factors in the first set, one runs the risk of producing a "skewed" compilation while the second set of criteria would allow for heterogeneity (Leech, 2007: 138). There is thus a certain tug-of-war between the two sets. Although he admits that it is more of an ideal rather than something that can be directly implemented, Leech states that the unit for sampling is the "initiator-text-receiver nexus", which he refers to as an "ATOMIC COMMUNICATIVE EVENT" (Leech, 2007: 138). Thus, whether one applies proportional or stratified sampling, one needs to consider frequency of reception.

Čermák (2009) introduces another model that is based partly on contextual features in the sense of setting variables, and features that are characteristic of the spoken form of language as opposed to the written form of language. He argues that "a (proto)typical spoken corpus is […] made up of data where specific spoken features, that are not to be found in written corpora, predominate, or are sometimes even exclusively present […]" (Čermák, 2009: 114). Along with the parameter of "awareness" during recording (p. 117), he thus suggests that a prototypical spoken text would have plus values for all twelve parameters:

1. +spoken	6. +informal	11. +casual
2. +dialogue	7. +interactive	12. +not aware
3. +proximity	8. +present	
4. +equality	9. +non-multiple	
5. +private	10. +spontaneous	

(from Čermák, 2007: 118)

This way of approaching spoken corpus design is parallel to the nature of data that has typically formed the empirical bases of research in conversation analysis and discourse analysis, and meets the demands of corpus-based pragmatics, which go beyond what "traditional" corpus linguistics caters for in terms of data structures (see Teubert, 2005; Schmidt & Wörner, 2009).

How is the model to be implemented and monitored, though, in a manner that also takes into consideration both the demographic and the topical dimensions of spoken discourse? In the following, we dwell on the corpus design and corpus management features of STC, which will be the product of a project that started in October 2008 with the aim of producing a

general corpus of one million words of present-day Turkish spoken discourse in its initial stage.[3]

3. FEATURES OF THE SPOKEN TURKISH CORPUS

To briefly describe the features of the technical aspects of STC, let us note that it employs EXMARaLDA (Schmidt, 2004), which is an open source software for corpus production that allows for online access to multimodal files. A detailed description of the technological infrastructure of STC is provided in Ruhi, Eröz-Tuğa, Hatipoğlu, Işık-Güler, Acar, Eryılmaz, Can, Karakaş and Çokal Karadaş (2010).

3.1. Corpus design: metadata and annotation

Independent of Čermák's study, STC was designed along the above-mentioned parameters. It attempts to monitor and address representativeness through demographic statistical measures, and enhances the monitoring of register variability through a close tracking of topics and speech acts.

Besides constructing a metadata system for domain, interactional goal and speaker features (e.g. age, education and language proficiencies), we maintain that the inclusion of speech acts (Searle, 1973) and conversational topics provides a crucial tool in monitoring the samples according to the tenor and affective tone of communicative events. While enabling future use of the corpus for a variety of research purposes ranging from discourse-level annotation to corpus-based and/or corpus-driven emotion research, these discursive dimensions are significant in tracing what may be the 'hidden' dimensions of the communicative events, which would not be available for the monitoring of the corpus compilation if sampling were based only on contextual and sociopragmatic variables. Naturally, the annotation of speech acts is but one scheme that would serve these purposes, but it renders granularity to the sampling beyond what can be achieved with domain and setting categorization.

Viewed from another perspective, spoken texts are slippery resources of language in terms of domain and setting categorization such that they are spatio-temporally characterized by shifts in interactional goals. A service encounter on a public transportation vehicle or at a shop, for example, can easily turn into a chat. Thus, if a communicative event were to be classified only for its domain of interaction, one would risk the chance of tracing subtle

[3] A DEMO version is available for browsing and research purposes via http://std.metu.edu.tr/en/ .

differences within the same domain, and hence, lose track of variability along the formality-informality dimension. In this regard, the simultaneous annotation of topics and speech acts addresses the concern for achieving maximal variability in register.

Other than a proportional sampling approach that controls the demographic dimension, the sampling of recordings is based on the identification of domains of discourse, for which the physical space of the interaction, the social relationships between the participants, the main thrust of the communication (e.g., chatting, transactional, educational, etc.), and the medium of communication are taken into consideration. Table 1 below reflects the design along these dimensions.

Table 1: Major interactional samples in STC (from Çokal Karadaş and Ruhi 2009: 317)

	TALK TYPE	PARTICIPATION FORMATS AND SETTINGS
Topic of conversation:	Personal/Impersonal	
Participation type:	1) Monologue	2) Dialogue a. 2 -5 persons b. 6 -10 persons c. More than 10
Medium:	1) face-to-face	2) Mediated: a) Telephone b) Broadcasts
Face-to-face:	A. Chats	1) In the family; family with guests (e.g., at dinner) 2) Educational locations (e.g., chats during lunch or coffee) 3) Chats in business locations
	B. Institutional or semi-institutional	5) In hospitals/medical centers: (e.g.: doctor-patient encounters) 6) Rituals (e.g., engagements; festivities in business locations; condolences) 7) On public transportation (e.g. inter-city bus, taxi, on the *dolmuş*[4]) 8) Service encounters (e.g., making an appointment, malls, bazaar) 9) Business settings (e.g., meetings, talk in the secretary's office; job interviews) 10) Educational settings: meetings 11) Classroom discourse: Lectures; group activities
Telephone:	1) Institutional	2) Between family members and friends
Mass media:	1) TV and radio talk that is close to spontaneous talk (e.g., talk shows)	2) Scripted (e.g., excerpts from series) 3) Text reading (e.g., news)

[4] *dolmuş*: a minibus used for public transportation

Taking this layout as a starting point, what we have tried to achieve in STC is a "balanced" corpus. We take Leech's (2007) definition: "a corpus is 'balanced' when the size of its subcorpora (representing particular genres or registers) is proportional to the relative

frequency of occurrence of those genres in the language's textual universe as a whole. In other words, balancedness equates with proportionality" (p. 4). There have been few attempts, however, to explain what this requirement means, and no serious attempt was ever made to ensure that the genres, in the Brown Corpus or the BNC, for example, were proportional in this sense (ibid.). Balancedness is very difficult to demonstrate, even for very carefully constructed corpora.

For the development of STC, 8 major domains were identified (see Table 2). As will be observed, the major categories are based on social role relationships and the sub-categories are a mixture of topics, goals of interaction and conversational topics.

Table 2

MAJOR DOMAINS	MAIN INTERACTIONAL GOAL & MEDIUM
1. FAMILY MEMBERS & RELATIVES (13 hours)	*chats, cultural events, narratives, telephone conversation, educational interaction, trips with the family*
2. FRIENDS AND FAMILY (3 hours)	*(same as in 1)*
3. FRIENDS (7 hours)	*(same as in 1)*
4. WORKPLACE COMMUNICATION (10 hours)	*meeting, shopping, workplace chats, telephone conversations, cultural events, work-related dinners interviews, appointments*
5. EDUCATION (7 hours)	*lecture in the social sciences, lecture in science, lecture in skills courses, seminars, conferences, panels student conferencing, parent-teacher meeting educational panel, interviews for educational programs school trips*
6. SERVICE ENCOUNTERS (2,5 hours)	institutional, shopping, service encounter on public transport
7. BROADCASTS (7,5 hours)	*news, news commentary, debate, series & films, sports educational, documentary, entertainment, competition culinary, health, children's programs*
8. OTHER (2 hours)	*brief encounter, religious discourse (sermons), legal discourse (e.g. court cases) political speech, political meeting, other public speeches, other public meeting, research*
9. UNCLASSIFIED	

The relative weightings of these domains were computed according to the results gained by small-scale data collection on "what Turkish people do and what type of

interactions they hold in a regular day" as well as by consulting available demographic statistics.

Participants were asked to record everything they did, and how many hours in a number of (a) weekdays and (b) weekends they spend conversing in these domains (e.g. with friends, with colleagues, on the phone in the workplace, etc.) or are a recipient of such conversations (i.e. for broadcast sub-types). Considering the daily engagements of the working population, stay-at home, retired people and students, and researcher intuitions, representative 24-hour breakdown scenarios were created. Based on these average values, the projected weightings of each of the conversational domains/events in terms of hours in the 1 million spoken words in STC were calculated. Using the grid system, the breakdown was also projected on to the seven geographic regions of Turkey, in line with the ratio of the population in the regions. This gave the team slots to be filled according to domain>region>interaction types. Secondary level delimiters on these slots were gender and age.

Initially starting opportunistically, the STC had now reached 86 spoken data collection volunteers around Turkey who have submitted recordings for the corpus. The team closely guides the volunteers according to the grid system on the types of future interactions that need to be recorded.

Due to the nature of spoken discourse, not much value can be arrived at by controlling the length of each sample from a specific interaction sub-type, as written corpora compilers often do. Spoken corpora would lose from its linguistic and socio-pragmatic value if communication types are screened for equaling length and cut for that purpose. For instance, the length of workplace meetings in the Marmara region may be conventionally different than those held in the northern region (Black Sea) owing to socio-cultural traits and values (e.g. longer phatic talk before decision-making). Thus, for STC, no cutting or altering of individual samples collected is implemented beyond that of maintaining the privacy of sensitive information in the name of ensuring proportionality. This procedure will thus make the resource valuable for pragmatics research, which would require that communicative events be recorded in full rather than cut off to maintain proportionality.

The three-pronged scheme in STC is also enhanced by the design of the transcription and annotation scheme. STC takes within its purview a number of features that interactional sociolinguistics (see, e.g., Goffman, 1971) and the field of discourse analysis reveal as being significant in interaction. To keep track of the tenor of the communicative events, STC thus prioritizes the annotation of following pragmatic features:

a. Overlaps, filled and unfilled pauses, repairs

b. Discursive, formulaic expressions (e.g., thanking formulae)

c. (Im)politeness markers (e.g. address forms and T/V forms)

d. Non-prosodic features (e.g. laughing)

e. Gestures[4]

3.2. Corpus management

The STC corpus management system enhances EXMARaLDA with a web-based system interface and a relational (MySQL) database for metadata, which has been developed for making the management of corpus production and presentation flexible enough for use by non-experts. In this manner, experts and non-experts can submit annotation on conversational topics and speech acts, and edit them at any stage of the workflow to attain a finer-grained description of the sample. The system thus enables continuous monitoring of the corpus design parameters, with loops at each stage to the upper levels:

1. Annotation scheme of metadata for the samples
2. Entry of samples into the system, along with domain and speaker metadata
3. Transcription and annotation of recording, conversational topics and speech acts

In other words, the system implements an "agile" (Voorman and Gut, 2008) workflow in monitoring representativeness. As the system allows for the construction of sub-corpora at any stage in the workflow, it is possible to produce intra-corpus comparable data using any one of the design features. This enables issues concerning the practicalities of introducing 'missing' samples to be handled by using standard statistical measures.

Figure 1 illustrates the main page of the STC DEMO Version, where the three lists include each sample, the speech acts in the corpus and the speaker IDs. By clicking any of the items in the list, one arrives at its metadata. For example, clicking on one of the speech acts allows one to see links to all the samples containing tokens of the speech act.

[4] To be implemented in later stages of the development of the corpus.

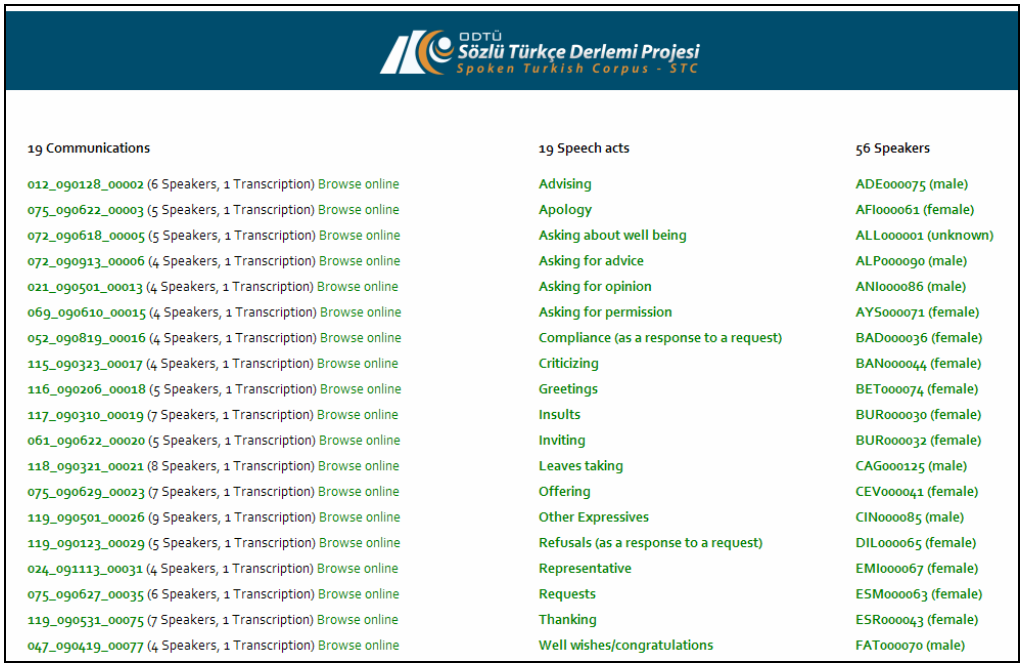


Figure 1: Main page of STC DEMO Version

Figure 2 illustrates the metadata data for one sample.[5]

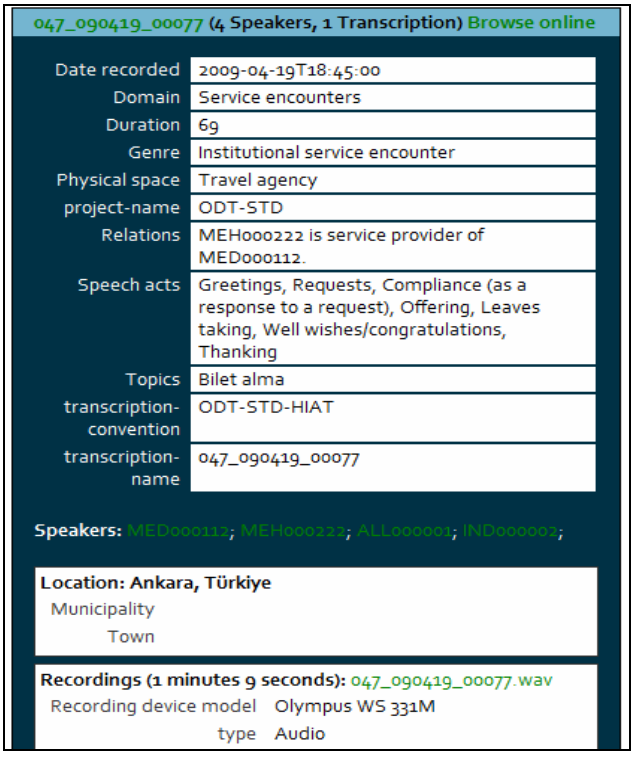


Figure 2: Part of a communication metadata

[5] Currently, conversational topics in STC are annotated in Turkish for ease of use by the transcribers. In the final version, metadata and annotation will be retrievable in Turkish and English.

4. CONCLUDING REMARKS

Our experience during the construction of STC is that, owing to the mobility in the population, spoken corpora for Turkish on a much larger scale require one to keep close track of place of birth and length of stay in various locations in order to achieve representativeness in accent and dialects. A further issue is that speakers in the modern world may be diglossic and multilingual. In this regard, education appears to be a more reliable feature in terms of keeping track of expectations in sampling in the Turkish context.

Arguably, conversational topics can be searched to keep track of register and genre variation through word searches, and indeed this could have been an option for STC. But the added value of this annotation has been that one can observe the accumulation of topics even without interim sub-corpora constructions. This has the added value of overviewing the topical range at any time in the workflow. Since the size of the corpus is extremely small at its current stage of development, it remains to be seen whether doing speech act annotation will eventually produce better representativeness. Our experience is that they are rich data for tracing the sociopragmatically significant aspects of language use. Thus speech act metadata are functioning as a higher-order variable in monitoring the heterogeneity of the samples with respect to the interactional parameters. While topic and speech act annotation obviously is time-consuming, the pay-off is considerable, especially in regard to its potential to maintain a corpus-driven approach to corpus construction itself.

REFERENCES

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics, 19* (2), 219-241.

BNC www.natcorp.ox.ac.uk

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing, 8*(4), 259-265.

Čermák, F. (2009). Spoken corpora design: Their constitutive parameters. *International Journal of Corpus Linguistics 14* (1), 113-123.

Çokal Karadaş, D., & Ruhi, Ş. (2009). Features for an internet accessible corpus of spoken Turkish discourse. *Working Papers in Corpus-based Linguistics and Language Education 3*, 311-320.

EXMARaLDA. http://exmaralda.org/

Goffman, E. (1971). *Frame Analysis*. New York: Harper and Row.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp.133-149). Amsterdam: Rodopi.

Ruhi, Ş. Eröz-Tuğa, B., Hatipoğlu, Ç., Işık-Güler, H., Acar, G. C., Eryılmaz, K., Can, H., Karakaş, Ö., Çokal Karadaş, D. (2010). Sustaining a Corpus for Spoken Turkish Discourse: Accessibility and Corpus Management Issues. Paper to be presented at *LREC 2010*. Malta, 17-23 May.

Schmidt, T. (2004). Transcribing and annotating spoken language with EXMARaLDA. *Proceedings of the LREC-Workshop on XML Based Richly Annotated Corpora*, *Lisbon 2004.* Retrieved from http://www1.uni-hamburg.de/exmaralda/files/Paper_LREC.pdf

Schmidt, T. & Wörner, K. (2009). EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics, 19* (4), 565-582.

Searle, J. (1976). A classification of illocutionary acts. *Language and Society,* 5, 1-23.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics* 1, 1-13.

Váradi, T. (2001). The linguistic relevance of corpus linguistics. In P. Rayson, A. Wilson, T. McEnery, A. Hardie & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference.* UCREL Technical Papers (Vol. 13) (pp.587-593). Lancaster: UCREL.

Voormann, H., & Gut, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory 4* (2), 235-251.

Los elementos de prolongación copulativos en textos científicos ingleses del siglo XVIII

ESTEFANÍA SÁNCHEZ BARREIRO

Universidade da Coruña

Resumen

Existen ciertos elementos lingüísticos que denominaremos elementos de prolongación o prolongadores, que podríamos definir como aquellas frases del tipo or whatever, and so on and so forth, and things, que aparecen normalmente al final de una lista enumerativa y cuya función primordial es alargar esa lista de algún modo. Nuestro principal objetivo en esta investigación consiste en estudiar el uso de estos elementos dentro de un contexto científico-histórico. Se trata pues de observar e identificar aquéllos que los autores científicos utilizaban con mayor o menor frecuencia en el siglo XVIII, aún a pesar de tratarse de elementos lingüísticos que, de alguna manera, se caracterizan por su vaguedad y contravienen, por esto mismo, la esencia de la escritura científica que ha de caracterizarse por su rigurosidad.

Palabras clave: prolongadores, lingüística histórica, lingüística de corpus

Abstract

There are some linguistic elements that we will called extenders (elementos de prolongación o prolongadores for this work) which could be defined as those phrases like or whatever, and so on and so forth, and things, which normally show at the end of an enumeration and whose main function consists of extending that list. Our main aim in this paper is to study the use of these elements within a scientific and historical context in order to observe and identify those ones that scientific authors would use in the XVIII century to a greater or lesser degree. And this, in spite of being linguistic elements characterised by their vagueness what contravenes the very essence of scientific writing.

Keywords: extenders, historical linguistics, corpus linguistics

1. INTRODUCCIÓN[1]

Ya en los períodos del inglés medio y moderno temprano y hasta nuestros días, un número significativo de listas terminan con construcciones del tipo *and so on*, *or such like*, *&c*. En el presente trabajo hemos identificado dichas formas con el nombre de elementos de prolongación o prolongadores, cuyas estructuras y significados han sido objeto de diversos estudios del análisis sobre discurso, aunque basados éstos especialmente en muestras orales.

Nuestro principal objetivo en esta investigación consiste en estudiar el uso de estos elementos, pero en muestras escritas y dentro de un contexto científico-histórico. Para ello tomaremos fragmentos de textos pertenecientes al siglo XVIII, a través de los que trataremos de observar e identificar aquellos prolongadores utilizados con mayor o menor frecuencia. De

[1] Este trabajo de investigación ha sido financiado por la Xunta de Galicia. Programa de promoción xeral de investigación do Plan galego de investigación, desenvolvemento e innovación tecnolóxica (Incite), PGIDIT07PXIB104160PR.

esta manera, podremos perfilar el comportamiento, uso y evolución de estas formas en el llamado registro científico.

Escogimos para el estudio veinte textos de inglés científico del citado siglo tomados todos ellos del *Coruña Corpus*[2]. Las muestras pertenecen en su totalidad a la misma disciplina científica, la que denominamos Life Sciences (*CELiST, Corpus of English Life Sciences Texts*).

2. LOS ELEMENTOS DE PROLONGACIÓN

Los elementos de prolongación son, en pocas palabras, aquellas frases del tipo *or whatever*, *and so on and so forth*, *and things*, que aparecen normalmente al final de una lista enumerativa y cuya función primordial es alargarla de algún modo. Pueden tomar diversas formas de manera que su extensión semántica aparece representada por distintas estructuras también con diversas funciones sintácticas, aunque la función pragmática parece ser siempre la misma. Así, podemos encontrar tanto frases muy simples (*and stuff, and others*) como estructuras mucho más largas y complejas (*or anything of the kind*, *and things of that nature*).

Sin embargo, a pesar de ser un recurso lingüístico extendido, estos elementos no se tratan en las gramáticas de carácter general al uso. Por ejemplo, tanto en la gramática de Huddleston (1984) como en la de Quirk and Greenbaum (1985) no hemos encontrado ninguna referencia para este tipo de estructuras. Ni siquiera obras como el *Oxford English Dictionary* o el *Middle English Dictionary* contienen descripción o mención alguna. Es necesario acudir a otros reconocidos autores para hallar algún tipo de definición.

Siguiendo a Ruth Carroll (2007), los elementos de prolongación serían frases formadas por una conjunción y una frase nominal no específica cuya función básica es la de extender una lista o enumeración y son, por lo tanto, elementos que aportan mucho más de lo que pudiera parecer a la carga semántica del mensaje que se transmite. En general, este tipo de construcciones parecen identificarse mediante el uso de conjunciones copulativas o disyuntivas como *and* u *or*, aunque éstas no aparecen de forma exclusiva ya que podemos encontrar estructuras de este tipo encabezadas por la preposición *with*.

[2] El *Coruña Corpus of English Scientific Writing* es un proyecto aún en proceso de desarrollo llevado a cabo en la Universidade da Coruña con colaboración de investigadores de otras instituciones. Dado que el subcorpus constituido por los textos de la disciplina Life Sciences está aún en revisión, es posible que algún dato de los que aquí se aportan se pueda ver alterado en trabajos futuros.

Como confirma Overstreet (1999: 3), aparecen con frecuencia en posición final de la cláusula y su forma básica responde a conjunción más frase nominal. Tanto Overstreet[3] (1999) como Carroll (2007) optaron por el nombre de *extenders* para hacer referencia a esas expresiones, pero lo cierto es que hay diversas denominaciones preferidas por otros autores, quienes han utilizado términos como: *set-marking tags* (Dines, 1980; Ward and Birner, 1993), *list completers* (Jefferson, 1990; Lerner, 1994), *extension particles* (Dubois, 1992) e incluso *vague category identifiers* (Channell, 1994).

Una de las fuentes que hemos usado referidas exclusivamente al español para comparar el tratamiento de estas estructuras con lo que se ha estudiado para el inglés nos llevó al tratamiento de series enumerativas en el discurso oral en español (Cortés, 2006). Cortés llama a estas construcciones *elementos de final de serie enumerativa*. El término que hemos decidido utilizar aquí, *elementos de prolongación o prolongadores*, constituye una amalgama de esta forma y la inglesa *extender* propuesta por Overstreet (1999) o Carroll (2007) que ya hemos utilizado en trabajos anteriores (Sánchez Barreiro: en prensa). Entendemos que la denominación *elementos de prolongación* explicita un poco más, si cabe, el significado real y la función de lo que definimos que el nombre sugerido por Cortés (2006) y que, de esta manera, se deja entrever la idea de extensión o de prolongación del significado que estas construcciones encierran.

En resumidas cuentas, los elementos de prolongación son un recurso lingüístico cómodo para el emisor cuando necesita ampliar de una manera rápida y eficaz cualquier lista de elementos. Su función principal sería la de dilatar el campo semántico al que pertenecen las palabras a las que precede sin necesidad de acudir a un gran número de términos de ese mismo campo semántico que pudiera hacer la lista interminable. De este modo, también se ve minimizado el esfuerzo lingüístico (economía del lenguaje) que ha de realizar el hablante.

2.1. Caracterización semántica y formal

Overstreet (1999: 3) establece una división básica de los elementos de prolongación: aquéllos que empiezan por *and* llamados copulativos, dentro de los que incluiríamos, por ejemplo, *and that, and so on* y aquéllos que comienzan por *or* conocidos también como disyuntivos, incluyendo éstos estructuras del estilo de *or what, or something*. No obstante, en el presente artículo sólo nos ocuparemos de analizar los primeros.

Según Quirk (1985: 930) *and* es la conjunción coordinante con el significado y uso más generales. En el caso de nuestro objeto de estudio, aparecerá siempre al término de una

[3] Overstreet ya había usado el término general *extenders* en un trabajo anterior. Ver Overstreet and Yule (1997).

enumeración con el fin de completar una lista. Por lo tanto, sólo trataremos el tema de la coordinación a nivel de sustantivos, de frases nominales y de cláusulas en casos como *and things, and things of that kind* y *and what have you*, respectivamente. Por otro lado, debido a su composición y manera de extender el significado dentro de la enumeración, los elementos de prolongación pueden clasificarse en dos tipos. Su taxonomía atiende, en primer lugar, a criterios formales y está basada en los estudios realizados por Overstreet (1999) en los que ya distingue entre prolongadores (para ella, *extenders*) generales y específicos. El primer tipo contiene formas no concretas como *and the rest of things*. En el caso de los prolongadores específicos podríamos incluir ejemplos como *and many things that we had borrowed* cuyo uso, en este caso, de la cláusula de relativo los hace más concretos y precisos al delimitar su referente.

Al contrario de lo que pueda parecer, no siempre resulta sencillo establecer diferencias entre los elementos de prolongación generales y los específicos. Podríamos afirmar que construcciones como *and stuff like that* o *and the rest* son buenos ejemplos de prolongadores generales. Éstos han de aparecer en posición final de la lista enumerativa dispuesto de manera que es el prolongador el que cierra la enumeración.

Por otro lado, podríamos clasificar *and any other problem of the body* como un buen ejemplo de prolongador específico puesto que el campo semántico al que se refiere el prolongador es más determinado y, como consecuencia, más reducido. Por eso, llamamos a estos prolongadores específicos, ya que con ellos disponemos de un material léxico mucho más restringido que con los generales.

3. ALGUNAS CONSIDERACIONES PREVIAS

En su trabajo, Köhnen (2007) comenta que existen dos problemas que surgen a raíz del estudio del acto de habla. Uno es el de conseguir un inventario completo de formas que lo expresen y otro, lo que él ha dado en llamar *hidden manifestations* (2007: 139). Así, expresa que es virtualmente imposible confeccionar un catálogo completo de todas y cada una de las manifestaciones de un acto de habla en la historia de la lengua inglesa. Un segundo problema se presenta con las denominadas *hidden manifestations*. No podemos hacer afirmaciones claras sobre el desarrollo del acto de habla porque normalmente no se pueden cubrir todas sus diferentes manifestaciones a lo largo del tiempo (2007: 140). Así pues, debemos asumir una serie de formas típicas o comunes a más textos para tener al menos un punto del que partir.

Teniendo en cuenta todo lo mencionado por Köhnen (2007), y que nuestro principal objetivo será identificar construcciones copulativas nos dispusimos a elaborar una lista en la que se especificase cada una de las estructuras de este tipo con aquellas formas que consideramos que era posible encontrar. Dichas formas fueron, en su mayoría, identificadas por otros autores en estudios anteriores: Overstreet (1999), Aijmer (2002) y Carroll (2007). De ellas, se consideró conveniente incluir en este inventario inicial una selección lo más representativa posible tanto para el tipo de estudio como para el tipo de textos.

Tabla 1: Lista de construcciones copulativas para su posterior búsqueda en las muestras.

and (all) (this/that) (sort/kind/type) of (thing/stuff)	and (all) the rest
and all (that)	and so forth
and c. (&c.)	and so on
and everything (like that)	and whatnot
and other	and others
and so	and so on and so on
and that	and the like
and the whole thing	and other things
and things like (this/that)	and so on and so forth
and things of (this/that) (kind/nature/sort)	and stuff (like that)
and this and that	and such (or any)
And whatever	

En el momento de realizar el estudio, hubimos de rechazar aquellos elementos que, aunque por su estructura podrían ser considerados prolongadores, una vez comprobada la sintaxis y significado demostraban no ser válidos en muchos de sus usos. Tal fue el caso de la frase *and that.* Después de un minucioso análisis resultó estar compuesta en la mayoría de las ocasiones por un *that* relativo y no por un pronombre.

Por otro lado, para este estudio hemos considerado que una serie enumerativa no es únicamente aquella que consta de un mínimo de tres elementos tal y como afirman algunos autores (Jefferson, 1990; Lerner, 1994) para quienes las listas tienden a estar formadas por tres partes o segmentos. En nuestro caso hemos optado por considerar también aquellos ejemplos en los que se emplean prolongadores acompañando a un único referente previo.

4. MATERIAL DE ANÁLISIS

Como principal fuente de datos hemos tomado los textos del *Coruña Corpus of English Scientific Writing,* un proyecto que se está llevando a cabo actualmente en la Universidade da Coruña y que comprende una serie de muestras para el estudio histórico de la escritura científica en inglés. Se incluyen en éste, entre otras disciplinas, textos que tocan diferentes ramas de las ciencias de la vida, y son las muestras pertenecientes a esa parte del corpus las empleadas en este estudio. Suman un total de veinte y todas ellas corresponden al siglo XVIII, concretamente entre los años 1707 y 1794. El número de palabras por texto se sitúa alrededor de 10.000, y de ahí que nuestro corpus esté constituido por un total de 200.828 palabras.

Nuestro estudio también ha formado parte un importante instrumento sin el cuál la búsqueda habría sido más costosa. La *Coruña Corpus Tool* (*CCT*) es la herramienta que se usa para recuperar la información de los datos recopilados en el *Coruña Corpus*. Se trata de una aplicación informática que todavía estaba en fase beta en el momento de redactar este trabajo y que está siendo desarrollada como ayuda para que los lingüistas puedan extraer y condensar toda la información necesaria sobre el corpus para sus posteriores estudios. Este programa de concordancia está siendo desarrollado en colaboración directa con el IRLab (Information Retrieval Laboratory) también de la Universidade de A Coruña.

5. ANÁLISIS DE DATOS: LOS PROLONGADORES COPULATIVOS

Los resultados obtenidos fueron clasificados según la taxonomía presentada en la sección 2 teniendo en cuenta que la enumeración comenzase por la conjunción *and*, y que la construcción pudiese clasificarse en general o específica desde un punto de vista semántico. Desafortunadamente, si bien hemos identificado un número amplio de elementos de prolongación generales, la variedad de estructuras no es grande. Este hecho podría venir condicionado por la idea predominante de que el nivel formal de este tipo de textos necesita de precisión en sus afirmaciones y que, por lo tanto, no deja lugar a palabras o expresiones vagas.

De las 200.828 palabras que constituyen nuestro corpus de datos, hemos encontrado que en 302 ocasiones se usaron estructuras que podemos denominar prolongadores copulativos. La tabla 2 expuesta a continuación muestra con mayor detalle la relación entre el inventario inicial y las formas correspondientes a todos los ejemplos copulativos que se han identificado en los textos. Pretendemos con ella mostrar y especificar el resultado

correspondiente a cada una de las estructuras en concreto e indicar el número de ocasiones en las que hemos identificado algún ejemplo independientemente del autor.

En muchas ocasiones encontraremos que la cifra 0 representa una gran mayoría de los ejemplos. Éste hecho, además de corroborar que no hemos encontrado ninguna ocurrencia para esa expresión, permite hacernos una idea más clara de la diferencia que existe entre la expresión oral y escrita en un contexto como el que trabajamos y en el que además se incluyen cambios temporales y de registro. Recordemos que el citado inventario inicial se construyó a partir de formas procedentes de varias fuentes y que, lamentablemente, todavía no existen estudios suficientes que ofrezcan resultados para nuestra misma contextualización. Parece confirmarse, pues, que el rango de prolongadores esperables en el registro oral es más amplio que el que, de hecho, encontramos en el escrito.

Tabla 2: Lista de prolongadores copulativos con su correspondiente número de ocurrencias.

PROLONGADORES COPULATIVOS	GENERALES	ESPECÍFICOS	TOTAL	%
And (all) the rest	0	1	1	0,33%
And all (that)	0	7	7	2,31%
And c. (&c.)	189	0	189	62,58%
And everything (like that)	0	0	0	0%
And other	2	72	74	24,50%
And other things	2	0	2	0,66%
And others	3	7	10	3,31%
And so	0	0	0	0%
And so forth	0	0	0	0%
And so on	0	0	0	0%
And so on and so forth	0	0	0	0%
And so on and so on	0	0	0	0%
And stuff (like that)	0	0	0	0%
And such (or any)	0	4	4	1,32%
And that	0	0	0	0%
And the like	3	0	3	0,99%
And the whole thing	0	0	0	0%
And things like (this/that)	0	0	0	0%
And things of (this/that) (kind/nature/sort)	0	0	0	0%
And this and that	0	0	0	0%
And (this/that) (sort/kind/type) of (thing/stuff)	0	11	11	3,64%
And whatever	0	1	1	0,33%
And whatnot	0	0	0	0%
TOTAL	199	103	302	100%

A la vista de los primeros resultados convendría aclarar que, aunque la más productiva de las construcciones se corresponde con la de prolongadores copulativos generales donde se han contabilizado 199 casos, la mayor parte de los ejemplos coinciden con la forma *&c* (forma abreviada de *et coetera*). Exceptuando unos pocos casos como *and others, and other*

things, and other y *and the like*, el resto de los prolongadores copulativos generales identificados correspondieron a la mencionada forma, lo cual parece decir bien poco a favor de la riqueza del léxico usado por los autores seleccionados.

Lo que quizás llame más nuestra atención aquí es su uso generalizado, sobre todo entre algunos autores. La forma *&c.*, que representa a *etcetera*, aparece en su forma abreviada en todas las muestras analizadas. Es curioso que, aunque existan otros modelos ortográficos en la época para esta forma latina de acuerdo con el *Oxford English Dictionary*, nosotros sólo hayamos localizado ejemplos correspondientes a la abreviatura. Formas como *&co.*, *et cetera*, *et caetera* y *etc.* bien podrían haber sido otras posibilidades para la expresión latina cuyo significado se ajustaría a "and other things" o "and so forth". Hoy en día también es utilizada con el significado de "and so on" o "and more". La frase *&c.* se usaba a menudo para referirse a la continuación en cualquier serie de descripciones. Esta es, al menos, la función principal que representa en todas las muestras examinadas. Otros prolongadores como *so forth* o *and so on* podrían haber sido buenos sustitutos. No obstante, no se han encontrado ejemplos de ninguna de esas dos expresiones en todo el corpus examinado, mientras que *&c.* aparecía ampliamente representada. Así lo ilustra el ejemplo (1) que se presenta a continuación:

(1) *It is nouriſh'd with Veins, Arteries,* ***&c****. and is alſo porous […]* (Gibson, 1720: 5)

Lo más llamativo de este resultado puede ser el uso generalizado, y suponemos que aceptado, en textos científicos. La expresión *&c*. aparece en enumeraciones compuestas desde por un único elemento hasta por un conjunto de cinco aunque por norma general se dan en listas de más de dos elementos. Asimismo, nos gustaría destacar una particularidad que sólo se da en uno de los textos. Para ser más exactos, se trata de la muestra de Bolton perteneciente a finales del siglo XVIII. En dicho texto se presenta la abreviatura anteriormente citada por partida doble, es decir, *&c. &c.* lo cual parece reflejar, directamente, la forma en que alguien se expresaría al hablar más que al escribir hoy en día. Así lo mostramos con el siguiente ejemplo.

(2) *in imitation of an hand, a flower, the horn of a rein deer,* ***&c. &c****.* (Bolton, 1789: 115)

El fenómeno llega a destacarse hasta tres veces en todo el escrito.

Según el *Oxford English Dictionary*, esta forma doble estaba bastante arraigada y era empleada con asiduidad en el pasado. El diccionario indica que ahora apenas se utiliza excepto en determinados asuntos burocráticos y para las cartas como sustituto de títulos nobiliarios o de altos cargos vinculada al nombre de la persona a la que es dirigida la misiva.

Aparte de esto, la frecuencia de estos prolongadores copulativos generales es relativamente baja en la mayoría de las muestras. En general, podemos encontrar cinco casos, como mucho, en los textos que hemos analizado, aunque también hay un texto (Boreman, 1730) en el que hemos contabilizado hasta 33 usos. De este modo, *&c.* se convierte, con diferencia, en el elemento de prolongación copulativo general más frecuente de nuestras muestras.

La abreviatura *&c*. se ha contabilizado un total de 189 veces, mientras que el resto de los elementos de prolongación generales se han computado sólo 10 veces. Una de estas ocasiones se da en un texto de 1737 donde encontramos la expresión: *and the like*. Dicha forma aparece tres veces, y todas en esa misma muestra.

(3) *they likewiſe ſpin their Hair into Garters, Girdles, Saſhes,* ***and the like,*** (Brickell, 1737: 108)

And other things es también otro de los prolongadores generales identificados durante el estudio. Se trata quizás de una estructura que expresa más incertidumbre y resulta más informal que el resto. Aparece en dos ocasiones, una de ellas modificada por el adjetivo *many*.

(4) *deſtroying Corn, Fruit,* ***and many other things****.* (Brickell, 1737: 130)

En lo que se refiere a los prolongadores copulativos específicos también éstos constituyen un grupo algo más diverso, aunque no mucho, en comparación con los prolongadores copulativos generales. En esta ocasión, se han contabilizado 103 ejemplos distribuidos en siete tipos diferentes: *and the rest, and all, and other, and such, and others, and sort/kind of things/stuff* y *and whatever*. Se podría concluir entonces que tampoco existen diferencias extremas entre un tipo de prolongadores copulativos y otro en este sentido. Recordemos que en el caso de los generales se habían localizado cinco tipos de expresiones distintas. Por lo tanto, las cifras considerablemente altas que se llegan a manejar, en un caso 199 (prolongadores copulativos generales) y en otro 103 (prolongadores copulativos

específicos), tendrían más que ver con la repetición continuada de algunas expresiones y no tanto con la variedad de las formas utilizadas.

Con respecto a estos prolongadores, las colocaciones más frecuentes se corresponden con la conjunción *and* acompañada de la palabra *other* (72 apariciones, es decir, el 23,84% de los casos). En una cantidad tan elevada de ocurrencias también hemos encontrado que la frase contiene adjetivos en múltiples ocasiones. Aparte de los modificadores más corrientes con los que se combina la forma *other* como son *many, several, some* y *all*, también cabe destacar varios ejemplos con algunos un tanto inusuales como los que aparecen en los ejemplos (5) y (6).

(5) *To this we owe the number, variety, and excellence of our cattle, the richneſs of our dairies,* ***and innumerable other*** *advantages.* (Pennant, 1766: 7)

(6) *the Glands, for the Secretion of Juices; the Ventricle and Inteſtines, for digeſting their Nouriſhment;* ***and numberleſs other*** *Parts which are neceſſary to form an organic Body.* (Hughes, 1750: 62)

Asimismo, hemos analizado tanto la forma *and other* como *and others* en plural como prolongadores específicos diferentes. De cualquier modo, la frecuencia es mayor en el caso de *and other* y, a menudo, ambos se presentan modificados por los adjetivos señalados arriba.

Por ultimo, no podemos dejar de referirnos a la forma *and whatever*, una expresión que aparece en una única ocasión usada como elemento de prolongación dentro de una cláusula mayor.

(7) *[...] that are ſold and eaten almoſt as ſoon as caught,* ***and whatever*** *ſome may imagine, [...]* (Dodd, 1752: 60)

Para finalizar, hemos considerado interesante mencionar un último caso cuya presencia tampoco se podría considerar impactante, pero dentro de los resultados obtenidos con nuestro corpus sí tiene un peso significativo puesto que se trataría del tercer prolongador copulativo más numeroso siguiendo a las formas *&c.* (189 ejemplos) y *and other* (74 ejemplos), aunque a gran distancia de éstas. Estos casos suman un total de 11 apariciones y todos ellos representan elementos de prolongación copulativos específicos. Su estructura básica parte como siempre de una conjunción copulativa *and* a la que se suma un nombre con cualquier premodificador. Las posibilidades incluyen alguno de estos tres sustantivos: *sort, kind, type* (bien en singular o bien en plural). A ellos, se suma una frase preposicional introducida por *of* y acompañada ésta o bien por las palabras *thing* o *stuff* en caso de tratarse de prolongadores

generales (no se han hallado casos), o bien por otro sustantivo si dicho prolongador fuera específico. De esta manera, encontramos en las muestras ejemplos del siguiente tipo:

(8) *Nuts, Corn,* ***and feveral forts of*** *Fruits.* (Brickell, 1737: 128)

(9) *Epilepsy, Apoplexy,* ***and all kinds of*** *Convulsions and nervous Affections, […]* (Blackwell, 1737: 17)

En ambos ejemplos, los sustantivos principales de nuestras expresiones *sorts* y *kinds* se hacen acompañar de sendos premodificadores *several* y *all* y, en las dos ocasiones encontramos prolongadores específicos.

En la tabla que se adjunta podemos ver con mayor claridad la citada estructura así como algunas posibilidades que se nos han presentado en los textos con respecto al uso de modificadores, pues todos los ejemplos que se incluyen en este sentido han sido extraídos de nuestro propio corpus. Las palabras que aparecen en negrita representan aquéllas inalterables, que siempre deben aparecer en alguna de sus formas y sin las cuales la estructura estaría incompleta.

Tabla 3: Estructura de prolongador copulativo general con las posibles formas.

CONJUNCIÓN COPULATIVA	MODIFICADOR	SUSTANTIVO	PREPOSICIÓN	(MODIFICADOR)	SUSTANTIVO
and	several all different all other	**kind(s)** **sort(s)** **type(s)**[4]	**of**		stuff things fevers fluxes

Aquellos módulos que aparecen entre paréntesis (modificador) indican que éstos son opcionales y que no necesariamente han de aparecer cubiertos en todas las ocasiones. Hemos de señalar que en el caso de toparnos con un prolongador copulativo específico la parte final se podría ver alterada como en los ejemplos (8) o (9) donde aparecen diferentes sustantivos comunes o cláusulas.

[4] Nótese que, aunque la palabra *type(s)* haya sido incluida como posible sustantivo en la formación de la estructura no hemos encontrado ejemplo alguno construido con dicha forma.

6. Conclusión

Los elementos de prolongación desempeñan una función que varía de acuerdo con los contextos de uso. Hasta donde sabemos, no hay trabajos publicados sobre textos científicos y prolongadores en la lengua inglesa. Hubimos de hacer un inventario previo a la búsqueda ya que se trataba de unidades que no son evidentes ya que, pueden existir en un texto pero a priori no lo sabemos. Es por eso que, a pesar de disponer de un corpus de textos electrónicos y el software necesario para explotarlo, hay que seguir una metodología que implica leer el texto y hacer una lista de los elementos que creemos que podemos encontrar.

En el presente estudio se han examinado los elementos de prolongación copulativos más frecuentes en varios textos científicos de ciencias de la vida (Life Sciences) del siglo XVIII obteniendo como resultado que la forma de este tipo más empleada corresponde a la abreviatura *&c.*

Debemos tener en cuenta que el número de usos encontrados en las muestras no corresponde necesariamente a todos los prolongadores puesto que hemos limitado nuestra búsqueda a una lista particular. Esto supone que debemos continuar nuestras investigaciones por este camino basándonos quizás en listas mayores y/o en otras disciplinas científicas para observar su uso en función del nivel técnico de los textos y de la inclusión de las disciplinas en el ámbito de las Humanidades o de las Ciencias Exactas.

Apéndice

Lista final de prolongadores contenidos en las muestras analizadas

Prolongadores copulativos generales utilizados por los autores analizados

And c. (&c)
And other
And other things
And others
And the like

Prolongadores copulativos específicos utilizados por los autores analizados

And (all) the rest

And all (that)

And other

And others

And such (or any)

And (this/that) (sort/kind/type) of (thing/stuff)

And whatever

REFERENCIAS BIBLIOGRÁFICAS

Fuentes primarias

Bancroft, Edward. (1769). *An essay on the natural history of Guiana, in South America*. London: printed for Becket and P. A. Hondt,

Blackwell, Elizabeth. (1737). A Curious Herbal, containing five hundred cuts of the most useful cuts which are now used in the practice of physick. Vol.I. London: printed for Samuel Harding.

Blair, Patrick. (1723). Pharmaco-botanologia: or, an alphabetical and classical dissertation on all the British indigenous and garden plants of the new London. London: printed for G. Strahan; W. and J. Innys; and W. Mears.

Bolton, James. (1789). An History of Fungusses growing about Halifax. Wherein their varieties, and various appearances in the different stages of growth, are faithfully exhibited […]. Vol III. Huddersfield: printed by J. Brook for the author.

Boreman, Thomas. (1730). A Description of Three Hundred Animals; viz. Beasts, Birds, Fishes, Serpents, and Insects. With a Particular Account of the Whale-fishery. London: Printed by Rich Ware; T. Game.

Borlase, William. (1758). The natural history of Cornwall. The air, climate, waters, rivers, lakes, sea and tides. Oxford : printed for the author by W. Jackson.

Brickell, John. (1737). The Natural History of North-Carolina: with an account of the trade manners and customs of the Christian and Indian inhabitants. Dublin: James Carson.

Dodd, James Solas. (1752). *An Essay towards a Natural History of the Herring*. London: Printed for T. Vincent.

Donovan, Edward. (1794). Instructions for Collecting and Preserving Various Subjects of Natural History: As Animals, Birds, Reptiles, Shells, Corals Plants, &c. Together with

a treatise on the management of insects in their several states. London: printed for the author.

Douglas, James. (1707). Myographiæ comparatæ specimen: or, a comparative description of all the muscles in a man and in a quadruped. London: printed by W. B. for G. Strahan.

Edwards, George. (1743). A natural history of uncommon birds. Most of which have not been figur'd or describ'd, and others very little known from obscure or too brief descriptions without figures, or from figures very ill design'd. London: printed for the author, at the College of Physicians in Warwick-Lane.

Gibson, William. (1720). The farriers new guide: containing first, the anatomy of a horse, […]. London: printed for William Taylor.

Goldsmith, Oliver. (1774). *An history of the earth, and animated nature: In eight volumes*. Vol. 8. London: printed for J. Nourse.

Hughes, Griffith. (1750). *The natural history of the Island of Barbados*. London: printed for the author.

Keill, James. (1717). Essays on several parts of the animal oeconomy. Essay IV: Of Animal Secretion. London: printed for George Strahan.

Pennant, Thomas. (1766). *British Zoology*. London: printed for Benjamin White.

Sloane, Hans. (1707). A Voyage to the islands Madera, Barbadoes, Nieves St Christophers and Jamaica; with the Natural History of the Herbs and trees, four footed beasts, fishes, birds […]. London: printed for the author.

Smith, Sir James Edward. (1793). *English Botany.* Vol. II. London: printed for the author by J. Davis.

Speechly, William. (1786). A treatise on the culture of the pine apple and the management of the hot-house. Together with a description of every species of insect that infest hot-houses, with effectual methods of destroying them. Dublin: printed for Luke White.

Withering, William. (1776). A botanical arrangement of all the vegetables, naturally growing in Great Britain, with the descriptions of the genera and species. Birmingham: printed by M. Swinney.

Fuentes secundarias

Aijmer, K. (2002). *English Discourse Particles: Evidence from a Corpus*. Amsterdam: John Benjamins.

Carroll, R. (2007). Lists in Letters: NP-lists and general extenders in Early English correspondence. En I. Moskowich y B. Crespo (Eds.), *Bells Chiming from the past,* (pp. 37-53). The Netherlands: Rodopi.

Channell, J. (1994). *Vague Language.* Oxford: Oxford University Press.

Cortés, L. (2006). Los elementos de final de serie enumerativa del tipo y todo eso, o cosas así, y tal, etc. Perspectiva interactiva. *Boletín de Lingüística,* 26(18), 102-129.

Dines, E. (1980). Variation in discourse – 'and stuff like that'. *Language in Society*, 9, 13–33.

Dubois, S. (1992). Extension particles etc. *Language Variation and Change*, 4, 179-204.

Huddleston, R. (1984). I*ntroduction to the Grammar of English*. Cambridge: CUP.

Jefferson, G. (1990). List-Construction as a Task and Resource. En G. Psathas (Ed.), *Interaction Competence,* (pp. 63-92). Washington, DC: The International Institute for Ethnomethodology and Conversation Analysis and University Press of America.

Köhnen, T. (2007). Text types and the methodology of diachronic speech act analysis. En S. M. Fitzmaurice y I. Taavitsainen (Eds.), *Methods in Historical Pragmatics*, (pp. 139-166). Mouton de Gruyter: Berlin.

Lerner, G. (1994). Responsive List Construction. A conversational resource for accomplishing multifaceted social action. *Journal of Language and Social Psychology*, 13, 20-33.

Middle English Dictionary. (2001). University of Michigan. http://quod.lib.umich.edu/m/med/.

Moskowich-Spiegel, I. et al. *Corpus of English Life Sciences Texts.* (forthcoming).

Sánchez Barreiro, E. Adjunctive and disjunctive lists in Modern English scientific discourse. En M. L. Gea-Valor, I. García Izquierdo y M. J. Esteve (Eds.), *Linguistic and Translation Studies in Scientific Communication*, 86 (en prensa)

Overstreet, M. & Yule, G. (1997). The metapragmatics of "and everything". *Journal of Pragmatics,* 34, 785-794.

Overstreet, M. (1999). *Whales, Candlelight, and Stuff like That: General Extenders in English Discourse*. New York: Oxford University Press.

Oxford English Dictionary Online. (2ª ed.). (2009). Oxford: Oxford University Press. http://dictionary.oed.com/.

Quirk, R. & Greenbaum, S. (1985). *A Comprehensive Grammar of the English Language*. New York: Longman.

Ward, G. & Birner, B. (1993). The semantics and pragmatics of "and everything". *Journal of Pragmatics*, 19, 205-214.

I just come in Hong Kong by myself: Tense in spoken Hong Kong English

ELENA SEOANE

(University of Santiago de Compostela)

CRISTINA SUÁREZ-GÓMEZ

(University of The Balearic Islands)

Abstract

Perfect meaning, that is, the expression of actions within "a time span beginning in the past and extending up to now" (Huddleston & Pullum, 2002: 143; cf. also Quirk et al, 1985: 192-195; Biber et al, 1999: 467), is traditionally ascribed to the analytic construction have + past participle exclusively. However, alternative constructions are also attested, in particular the synthetic preterite, which competes with the analytic construction in some registers and some geographical varieties of English. The aim of the study is to analyse the expression of the perfect in Hong Kong English. First, we will revise the forms available for the expression of perfect meaning; then, we will study their use in spoken language; finally, we will compare whether the variation observed for Hong-Kong English matches the variation described for American, Scottish and Irish Englishes.

Keywords: perfect meaning, new Englishes, analytic construction, synthetic construction

Resumen

El perfecto, es decir, la expresión de acciones que tienen lugar durante un periodo de tiempo que comienza en el pasado y se extiende hasta el presente (Huddleston & Pullum, 2002: 143; cf. Quirk et al, 1985: 192-195; Biber et al, 1999: 467), se expresa tradicionalmente en lengua inglesa a través de la perífrasis verbal have + participio pasado. Sin embargo, existen construcciones alternativas, en particular el pasado simple, en competición directa con la construcción perifrástica en algunos registros y variedades geográficas de la lengua inglesa. El objetivo principal de este estudio es analizar la expresión del perfecto en la variedad de inglés hablada en Hong Kong. En primer lugar, revisaremos las formas que existen para expresar perfecto; a continuación, analizaremos su uso en la lengua oral; por último, compararemos si la variación existente en el inglés de Hong Kong coincide con la variación existente en las variedades habladas en Estados Unidos, Escocia e Irlanda.

Palabras clave: perfecto, nuevos ingleses, construcción analítica, construcción sintética

1. INTRODUCTION[1]

Perfect meaning, that is, the expression of actions within "a time span beginning in the past and extending up to now" (Huddleston and Pullum, 2002: 143; cf. also Biber, Johansson, Leech, Conrad and Finegan, 1999: 467; Quirk, Greenbaum, Leech and Svartvik, 1985: 192-195), is traditionally ascribed to the analytic construction *have* + past participle exclusively.

[1] For generous financial support thanks are due to the Autonomous Government of Galicia (INCITE grant 08PXIB204016PR) and the Spanish Ministry for Science and Innovation and the European Regional Development Fund (grant HUM2007–60706).

However, alternative constructions are also attested, in particular the synthetic preterite,[2] which competes with the analytic construction in some registers (e.g. spontaneous informal English, cf. Miller, 2000, 2004) and some geographical varieties of English, such as American English (Elsness, 1997, 2009), Scottish English (Miller, 2004), Irish English (Kirk, 2009) and in many of the varieties known as *New Englishes* (Kortmann and Schneider, 2004; Kortmann and Szmrecsanyi, 2004) or *Postcolonial Englishes* (Schneider, 2007). More marginal ways of expressing this meaning are also possible, such as preterite + deictic *there* in Scottish English (Miller, 2000: 338-339; 2004: 237) and the present tense in Irish English (Kirk, 2009). The traditional analysis of the expression of perfect meaning no longer holds for Present-day English, and therefore a more adequate account must be sought.

While in certain varieties of English this issue has been examined in detail (see above), in different New Englishes further research is still needed. Accordingly, in this paper we undertake the analysis of the expression of the perfect in one variety in particular, that of Hong Kong English (HKE), with the intention of extending the analysis to other varieties in future research. The ultimate goal of this research project is to discover to what extent the variation observed in the expression of perfect meaning is a vernacular universal (in Chambers's 2004 terms), that is, a non-standard pattern attested in most nonstandard dialects around the world, and whether it can be explained in terms of historical diffusion from British English or as an independent development in each variety, as well as its relationship to the variation found in the history of English.

We concur with Schneider (2007: 16) that, in the study of New Englishes, "[d]etailed linguistic descriptions should constitute a prerequisite for generalizations and applications of all kinds". In this paper we intend to offer one such detailed description of the expression of perfect meaning in HKE. We will focus on the spoken language, given that it is the medium in which the greatest deal of variation has been registered, and is generally considered the most vernacular type of language and therefore the most likely locus of change (Miller, 2006: 689). The aim of the study is to answer the following questions: (i) What forms are available for the expression of perfect meaning? (ii) How are they used in spoken language (meaning, type of verb, grammatical environment, etc.)? (iii) Does the variation observed match the variation described for American, Scottish and Irish Englishes?

---

[2] The preterite is traditionally attributed a different meaning, since it "describes a situation that no longer exists or an event that took place at a particular time in the past." (Biber et al, 1999: 467).

The outline of this paper is as follows: Section 2 elaborates on the variation found in the expression of perfect meaning in English. Section 3 describes the corpus used and offers the analysis of the corpus data. Finally, Section 4 presents conclusions.

2. The Expression of Perfect Meaning

The present perfect as described in grammars of English (Biber et al, 1999: 467; Quirk et al, 1985: 192-195) seems to be quite stable in standard formal written English but far less so in spontaneous spoken English, in non-standard English, and in various geographical varieties of English (Kortmann, 2006: 607-608). In these varieties there is a lack of functional distribution between the periphrastic and the synthetic forms, as shown in Elsness (1997, 2009) and Miller (2000: 323; 2004), among others. Map 1 below, from Kortmann and Schneider (2004), shows the levelling between present perfect and synthetic preterite in English worldwide (red spots indicate the relevant feature).

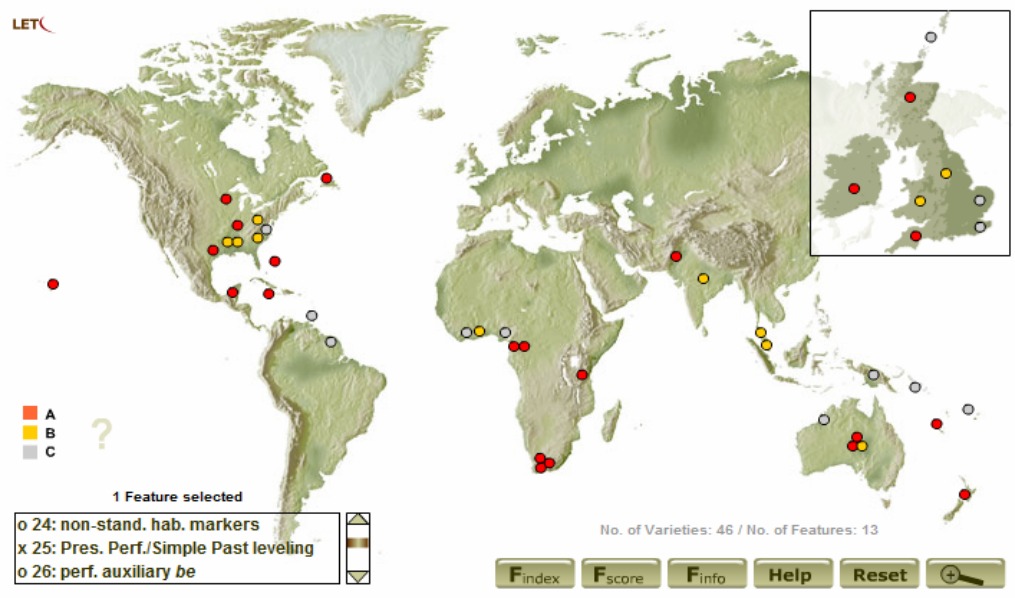


Figure 1: Levelling between present perfect and synthetic preterite in English (Kortmann and Schneider 2004)

In addition to the analytic *have* + past participle and the synthetic preterite, other forms are also possible to express perfect meaning. This variation has been attributed to temporal indefiniteness, which leaves room for individual interpretations (Elsness, 2009: 228); in fact, the analytic construction seems to be favoured in contexts where the time is made explicit by an adverb, to the extent that some authors believe that, in spoken English, the perfect is

splitting into three independent constructions differentiated by the occurrence or absence of particular adverbs, namely *just*, *yet*, *ever* and *never* (Miller, 2004: 244). Miller interprets the strong association between these adverbs and the present perfect as evidence that the classical interpretation of the meaning of the perfect does not reside in the perfect construction but in the adverbs: "The recent past-time component is signalled by *just*, the experiential component is signalled by *ever* and the resultative component is signalled by *yet*" (Miller, 2000: 335).

In order to study the expression of perfect meaning in spoken HKE, we will analyse all the constructions containing the above-mentioned adverbs (*just, yet, ever* and *never*), since these adverbs, which make the time reference explicit, would unambiguously require the presence of the present perfect according to traditional accounts. We will follow the widespread, traditional analysis of perfect meaning as presented in Comrie (1976; cf. also Dahl, 1985: Ch. 5; 1999: 290-291; Dahl and Hedin, 2000: 385-388; Huddleston and Pullum, 2002: 143-145; Miller, 2000: 327-331; 2004: 230).

1. Use of a verb phrase + *just* to express recent past (also called hot-news perfect, as coined by McCawley, 1971), as in *She has just arrived*.
2. Use of a verb phrase + (*n*)*ever* to express experiential (or indefinite anterior) meaning, as in *Have you ever lived abroad*?
3. Use of a verb phrase + *yet* to express resultative meaning, as in *I haven't read the book yet*.

3. CORPUS ANALYSIS

We selected a sample of data from the *Hong-Kong ICE corpus* (*International Corpus of English*) comprising c. 833,000 words of spoken language (files s1A, private dialogue). We retrieved examples using a concordance program which rendered a total of 2,618 examples of *just*, *yet*, *ever* and *never*. After filtering them manually we obtained 288 examples of the perfect-meaning constructions outlined in Section 2.

In this sample we found 12 different forms to express perfect meaning in combination with *never*, *ever*, *just* and *yet*, as Table 1 illustrates.

Table 1: Forms used in HKE to express perfect meaning in combination with *never*, *ever*, *just* and *yet*

	Form	Example	Frequency
[+ frequent] forms	*Have* + past participle	*Have you ever been to Macau?*	153 (53.1%)
	Preterite	*They just came in.*	72 (25%)
	Present/Base form[3]	*He never book me for a show.*	22 (7.6%)
	Have + Base form	*Have you ever encounter this?*	13 (4.5%)
[- frequent] forms	Present/Participle/Base form	*I just come in Hong Kong by myself.*	6 (2.1%)
	Preterite/Participle/Base form	*Never thought about it.*	6 (2.1%)
	Participle	*I just forgotten it*	5 (1.8%)
	Base form	*It never happen like that in Japan.*	4 (1.4%)
	Past continuous	*I was just watching your map.*	3 (1.1%)
	Be + past participle	*I'm just retired.*	2 (0.7%)
	Be + base form	*I 'm not find a one yet*	1 (0.3%)
	Have + 3rd person form	*Have you ever tries to fall in love with a Chinese?*	1 (0.3%)
TOTAL			288

The results set out in Table 1 show that in combination with *yet, ever, never* and *just*, perfect meaning is expressed by means of the analytic construction *have* + past participle in only 53.1% of all cases, while the synthetic preterite takes up this function in 25%. Additionally, up to 10 different alternative forms are used to express almost one third of the examples (21.8%), in particular the present/base form (7.6%) and the *have* + base form (4.5%). The variation found, then, is very significant. Some of these forms, specially those for which only one case is recorded, could be attributed to performance errors due to a low level of proficiency in English: the latest estimates suggest that only 43% of the population in Hong Kong are proficient in English, and that lower socioeconomic groups do not use English to communicate (Bolton, 2003: 87). Undoubtedly, after this initial description of the expression of perfect meaning, in future research it will be necessary to carry out a micro-sociolinguistic description of usage correlated with sociostylistic parameters as a means of determining what variants correspond to what type of speaker, be it the careful usage of a society's educated population or the spontaneous usage of those without formal education (cf. Schneider, 2007: 18).

The constructions set out in Table 1 are not equally distributed in relation to the time adverbs and their meaning, as Table 2 shows:

[3] Form descriptions containing one or more forward slashes are ambiguous: the verb *book*, for example, can be a simple present form and a base form.

Table 2: Distribution of forms with respect to adverb and meaning

FORMS	Experiential		Recent Past	Resultative	TOTAL
	Never	*Ever*	*Just*	*Yet*	
Have + past participle	59 (50%)	39 (75%)	30 (35.2%)	25 (75.8%)	153
Preterite	33 (28%)	4 (7.7%)	35 (41.2%)	-	72
Present/Base form	16 (13.6%)	1 (1.9%)	4 (4.7%)	1 (3 %)	22
Have + Base form	-	7 (13.5%)	2 (2.4%)	4 (12.1%)	13
Present/Participle/Base form	1 (0.8%)	-	5 (5.9%)	-	6
Preterite/Participle/ Base form	4 (3.4%)	-	1 (1.2%)	1 (3%)	6
Participle	1 (0.8%)	-	4 (4.7%)	-	5
Base form	4 (3.4%)	-	-	-	4
Past continuous	-	-	3 (3.5%)	-	3
Be + past participle	-	-	1 (1.2%)	1 (3%)	2
Be + base form	-	-	-	1 (3%)	1
Have + 3rd person form	-	1 (1.9%)	-	-	1
TOTAL	118 (41%)	52 (18.1)	85 (29.5%)	33 (11.5%)	288

Table 2 shows that experiential meaning in combination with *never* favours the use of *have* + participle, and less frequently, both the synthetic preterite and the present/base form. However, in combination with *ever*, *have* + past participle clearly dominates, followed by *have* + base form. The high proportion of *have* + base forms found in this context, which contrasts with the low number of preterite forms, could be a sign of another change frequently attested in vernacular varieties of English, namely the reduction of verb paradigms, in which the base form also serves as past tense or past participle, as in *Have you ever encounter this?* (cf. Table 1 and Kortmann, 2006: 609).

With respect to the expression of the recent past (instances in combination with *just*), the corpus data show that even if the periphrastic construction is the expected form prescriptively, the dominant form is not the perfect but the preterite form, in agreement with the same tendency observed in American English and spontaneous British English, as noted by Quirk et al (1985: 194) and Miller (2000: 337) among others. Finally, regarding resultative meaning in combination with the adverb *yet*, most examples contain the *have* + past participle form (again followed by the *have* + base form), and not a single preterite was found in this function. This finding contrasts with Quirk et al's (1985: 194) for American English, where synthetic preterite is often preferred to the present perfect in such contexts (e.g. *Did the children come home yet?*). However, the absence of preterites for the expression of resultative meaning in our corpus does corroborate Miller's results (2004: 239-240), who found that a number of constructions other than *have* + past participle are available for the expression of resultative meaning, and that they all involve the past participle (passive), which, according to Haspelmath (1993), was originally resultative.

The semantic type of verb has also been taken into consideration, since it has been found that choice of tense form can also depend on the meaning of the verb (cf. Dahl and Hedin, 2000: 390; Quirk et al, 1985: 193). We classified our tokens following Quirk et al's (1985: 201) 13 categories, but decided to simplify this classification into two basic categories in the analysis (dynamic vs stative) in order to avoid many cells with empty values. The results are shown in Table 3, which only includes [+frequent] forms:

Table 3: Distribution of forms according to semantic type of verb

	Experiential		Recent Past		Resultative	
	Dynamic	Stative	Dynamic	Stative	Dynamic	Stative
Have + past participle	30 (44.1%)	68 (74.7%)	29 (42.7%)	-	10 (76.9%)	15 (88.2%)
Preterite	21 (30.9%)	16 (17.6%)	34 (50%)	1 (50%)	-	-
Present/base form	13 (19.1%)	4 (4.4%)	3 (4.4%)	1 (50%)	-	1 (5.9%)
Have+base form	4 (5.9%)	3 (3.3%)	2 (2.9%)	-	3 (23.1%)	1 (5.9%)
TOTAL	68	91	68	2	13	17

Whereas Table 2 showed that experiential meaning tends to be expressed by *have* + past participle (57% of all cases if we add up *ever* and *never*), Table 3 shows that such association is much stronger when the verb involved is stative (74.7%); with dynamic verbs, however, the association is weaker, the preterite taking up this function in just over 30% of cases. The expression of recent past tends to occur with dynamic verbs, a context dominated by the preterite, immediately followed by the *have* + past participle. Only two stative verbs expressing recent past were found: *have* and *hear*.[4] Finally, in the case of the expression of resultative meaning, the distribution is more balanced, although they tend to appear with stative verbs. In both cases, *have* + past participle is the only frequent form.

A third variable is the polarity of the clause in which the perfect construction appears. Although initially we chose a more detailed classification specifying the type of negative polarity item in the category 'negative', in this study they were conflated into a general category under the cover term 'polarity' which distinguishes between 'positive' and 'negative' polarities. The results are included in Table 4:

[4] Two additional examples have been found among the [- frequent] forms, namely participles, in both cases represented by stative *be* (*been* and *been back*).

Table 4: Distribution of forms according to polarity and meaning

	Experiential		Recent Past	Resultative	
	Positive	Negative	Positive	Positive	Negative
Have + past participle	37 (80.4%)	61 (54%)	29 (41.4%)	5 (100%)	20 (80%)
Preterite	2 (4.3%)	35 (31%)	35 (50%)	-	-
Present/base form	1 (2.2%)	16 (14.2%)	4 (5.7%)	-	1 (4%)
Have+base form	6 (13.1)	1 (0.8%)	2 (2.9%)	-	4 (16%)
TOTAL	46	113	70	5	25

Our findings are in accordance with the crosslinguistic tendency of experiential meaning to occur in non-assertive contexts, negation being one of them (cf. Dahl, 1985: 143; Dahl and Hedin, 2000: 388). Regarding the forms, in negative experiential environments there is more competition than in the positive counterpart, clearly dominated by *have* + participle; in negative contexts *have* + participle, preterite and present/base form coexist. The canonical perfect, therefore, is still strongly associated with positive experiential meaning, while for the expression of negative experiential meaning, which is much more frequent in languages crosslinguistically, there is much more freedom of expression and variation. As to the expression of the recent past, it only occurs in positive contexts. Finally, the resultative meaning tends to occur in negative contexts, and to be expressed by *have* + participle, and much less frequently by the *have* + base form construction.

Table 5: Distribution of forms according to type of clause and meaning

	Experiential	
	Declarative	Question
Have + past participle	62 (53.4%)	36 (83.7%)
Preterite	35 (30.2%)	2 (4.7%)
Present/base form	17 (14.7%)	-
Have+base form	2 (1.7%)	5 (11.6%)
TOTAL	116 (72.9%)	43 (27.1%)

As for clause type, our results show that questions occur only with experiential meaning (27.1% of experiential are interrogative), which once more confirms the tendency of experiential meaning to occur in non-assertive contexts (cf. Dahl, 1985: 143; Dahl and Hedin, 2000: 388). The competition between the present perfect and the preterite for the expression of experiential meaning is stronger in declarative contexts (53.4% as against 30.2%) than in interrogative ones, where the present perfect is used in 83.7% of cases and the preterite in only 4.7%.

4. CONCLUSION

This paper has studied the expression of perfect meaning in combination with *never*, *ever*, *just* and *yet* in spoken HKE. The traditional view, that it is the present perfect in combination with these adverbs that expresses experiential (*never, ever*), recent past (*just*) and resultative meaning (*yet*), is not confirmed for spoken HKE. Our corpus data show that (i) the experiential perfect is in strong competition with the simple past, (ii) the perfect of recent past is just a minority construction, and (iii) the perfect of result is widely used but is only one of a number of resultative constructions. For each of these semantic values there are at least six different forms to express it. Particularly interesting is the occurrence of the present/base form and the *have* + base form, which were tentatively interpreted as manifestations of a vernacular universal: the reduction of verb paradigms.

Significantly, some of the findings for HKE coincide with those for spontaneous British English, such as the dominance of the preterite for the expression of recent past, and differ from other findings that are exclusive of American English, such as the use of the preterite for the expression of resultative meaning. Findings like these seem to make sense if we recall that Hong Kong was a British colony from 1842 to 1997.

In future research we intend to extend this study to further registers of spoken HKE and carry out a sociolinguistic analysis of these data in an attempt to explore further the underlying motivations for this variation.

REFERENCES

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bolton, K. (2003). *Chinese Englishes. A Sociolinguistic History*. Cambridge: Cambridge University Press.

Chambers, J.K. (2004). Dynamic typology and vernacular universals. In B. Kortmann (Ed.), *Dialectology Meets Typology,* (pp. 127-145). Berlin, New York: Mouton.

Comrie, B. (1976). *Aspect*. Cambridge: Cambridge University Press.

Dahl, Ö. (1985). *Tense and Aspect Systems*. Oxford: Blackwell.

Dahl, Ö. and E. Hedin. (2000). Current relevance and event reference. In Ö. Dahl (Ed.), *Tense and Aspect in the Languages of Europe* (Empirical Approaches to Language Typology. Eurotyp 20-6), (pp. 385-401). Berlin and New York: Mouton de Gruyter.

Elsness, J. (1997). The perfect and the Preterite in Contemporary and Earlier English. Berlin: Mouton de Gruyter.

Elsness, J. (2009). The present perfect and the preterite. In G. Rohdenburg and J. Schlütter (Eds.), *One Language, Two Grammars?* (pp. 228-246). Cambridge: Cambridge University Press.

Haspelmath, M. (1993). Passive participles across languages. In B. Fox and P. J. Hopper (Eds.), *Voice. Form and Function,* (pp. 151-177). Amsterdam: John Benjamins.

Huddleston, R. and G. Pullum. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

International Corpus of English: http://ice-corpora.net/ice.

Kirk, J. (2009). Tense and aspect in Irish Standard English. Paper presented at *The Third International Conference on the Linguistics of Contemporary English* (ICLCE3), London, July, 14-17.

Kortmann, B. (2006). Syntactic variation in English. In B. Aarts and A. McMahon (Eds.), *The Handbook of English Linguistics,* (pp. 603-624). Oxford: Blackwell.

Kortmann, B. and E. W. Schneider. (2004). *A Handbook of Varieties of English*. Berlin: Mouton de Gruyter.

Kortmann, B. and B. Szmrecsanyi. (2004). Global synopsis: morphological and syntactic variation in English. In B. Kortmann, K. Burridge, R. Mesthrie, E. W. Schneider and C. Upton (Eds.), *A Handbook of Varieties of English. Vol. II: Morphology and Syntax,* (pp. 1142-1202). Berlin and New York: Mouton de Gruyter.

McCawley, J. D. (1971). Tense and time reference in English. In Ch. Fillmore and T. Langendoen (Eds.), *Studies in Linguistic Semantics* (pp. 96-113). New York: Holt, Rinehart & Winston.

Miller, J. (2000). The perfect in spoken and written English. *Transactions of the Philological Society*, 98(2), 323-352.

Miller, J. (2004). Perfect and resultative constructions in spoken and non-standard English. In O. Fischer, M. Norde and H. Perridon (Eds), *Up and Down the Cline The Nature of Grammaticalization,* (pp. 229-246). Amsterdam: John Benjamins.

Miller, J. (2006). Spoken and written English. In B. Aarts and A. McMahon (Eds.), *The Handbook of English Linguistics,* (pp. 670-691). Oxford: Blackwell.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Schneider, E. W. (2007). *Postcolonial English. Varieties around the World*. Cambridge: Cambridge University Press.

The expression of politeness in research articles: Authorial presence vs. authorial invisibility in the discussion

CARMEN SOLER MONREAL

LUZ GIL SALOM

Universidad Politécnica de Valencia

Abstract

This paper investigates interpersonal relations in written scientific discourse and explores the range of negative politeness strategies where there is implicit and explicit presence of the writer in research article discussions from engineering fields. This study focuses on the writer's ability in both defending their claims and establishing a deferential relationship with the reader in the commenting on results move. The analysis of the corpus shows that both impersonalisation and personal attribution serve the writer's face-redressive purposes. They co-exist in some steps of this move: interpreting, accounting for, evaluating and exemplifying results. However, writers choose to mask their presence in the text when they criticise other works and acknowledge the limitations of the research presented.

Keywords: Politeness, authorial presence, authorial invisibility, politeness strategies, research article

Resumen

En este artículo investigamos las relaciones interpersonales que se dan en el discurso científico escrito. Asimismo identificamos las estrategias de cortesía negativa, donde encontramos muestras de la presencia implícita y explícita del autor en las secciones de discusión de artículos científicos del campo de la ingeniería. El estudio se centra en la capacidad del autor de defender su reivindicación científica a la vez que establece una relación de solidaridad con el lector en el movimiento dedicado a comentar los resultados de la investigación. El análisis del corpus demuestra que tanto la impersonalización como la personalización sirven a los propósitos de reparación de la imagen del escritor. Ambas estrategias están presentes en los sub-movimientos de interpretación, justificación, evaluación y ejemplificación de resultados. No obstante, el escritor opta por enmascarar su presencia en el texto cuando critica otros trabajos y reconoce las limitaciones de la investigación que presenta.

Palabras clave: Cortesía, presencia del autor, invisibilidad del autor, estrategias de cortesía, artículo de investigación

1. INTRODUCTION

Scientific discourse in general and research articles (RAs) in particular display a careful balance of factual information and social interaction between the writer and the reader. A number of studies have analysed how writers construct their texts considering the expectations of their potential audiences (Bazerman, 1988; Hyland, 1996, 2002; Swales, 1990). The maintenance of face has been associated to successful interaction among participants. Face refers to the "public self-image that every member of a society wants to claim for himself" (Brown & Levinson, 1987: 66). To maintain each other's face in interaction is to defend ones's face if threatened and to recognise and respect others' claims. In research articles certain acts, such as making a claim, criticising, speculating or asserting

one's priority (Myers, 1989) threaten face. The writer, in order to obtain credit from the scientific community, tends to mitigate the threat with some sort of face-protective, polite interpersonal strategy aiming at saving both the reader's and the writer's face. This strategy rests on the individual's need to be approved interpersonally and to be unimpeded personally.

In pragmatics, the need to mitigate a claim and develop a successful writer-reader relationship is at the origin of scientific hedging (Hyland, 2005; Markkanen & Schröder, 1989; Myers, 1989). Hyland (1996) and Myers (1989) agree in that hedging devices weaken the claim by presenting it as a personal belief, as an alternative view rather than as universal scientific knowledge. According to Hyland (1996: 446), categorical assertions relegate readers to a passive role, while hedged statements mark claims as provisional and invite the readers to participate in a dialogue. The use of hedges reduces the researcher's explicit role in the process of interpretation and evaluation of the phenomena under study. In this way, the writer diminishes the weight of the imposition on the reader, thus attending to his/her negative face, i.e. the want to be unimpeded by others (Brown & Levinson, 1987). The writer's strategy deliberately anticipates the reader's reactions and shows deference to other opinions. This results in tentative and tactful argumentation which connects hedging to negative politeness because it plays a double face-saving function: to protect the writer from a potential refutation and to show consideration to somebody else in order to obtain avoidance of commitment and reader acceptance of research claims.

Writers may choose to "personally intervene into their writing" or to "linguistically suppress their presence" (Hyland, 2005: 105). Personal attribution and impersonalisation devices become effective ways of presenting claims and engaging readers. The explicit presence of the author reveals the researcher's commitment and attitude towards the work carried out. In many cases, however, the writer feels obliged to disguise his/her personal voice in the discourse due to the face-threatening nature of his/her claims and denials. In this study we investigate interpersonal relations in written scientific discourse. We explore the range of negative politeness strategies where there is implicit and explicit presence of the writer in RAs discussions from engineering fields.

In discussions, more than in any other section of a research article, writers make their claims explicit and seek to persuade the reader of their credibility and the need to accept them as members of the discipline community. Thus, the exposition of facts gets entangled with face-saving strategies intended to impart a certain effect on the reader. Several studies have proposed models to describe the rhetorical structure of this closing section of a RA (Holmes, 1997; Lewin, Fine & Young, 2001; Peacock; 2002; Ruiying & Allison, 2003). These studies

concentrate on descriptions of the macrostructure of discussions, their rhetorical functions and linguistic markers, but do not examine writer-reader interaction. This paper pays attention to the writer's ability in both defending her/his claims and establishing a deferential relationship with the reader in the *commenting on results* move, a typical rhetorical unit in RA discussions.

2. CORPUS AND METHODOLOGY

The corpus comprises 46 closing sections including the word *discussion* in their headings. They belong to RAs published between 2000 and 2003 in 18 prestigious journals in the fields of computing, nanotechnology, robotics and telecommunications (see Appendix). The study was based on both a genre and a pragmatic approach. A move-step analysis of the corpus followed Ruiying & Allison's (2003) model for applied linguistics RAs. The authors found that the information in discussions can be organised in seven moves: *background information*, *reporting results*, *summarising results*, *commenting on results, summarising the study*, *evaluating the study* and *deductions from the research*. However, they stated that the communicative focus of these sections is on commenting on results.

Our analysis showed that *commenting on results* is the most frequent move in the corpus (found in 44 RAs), which coincided with Ruiying and Allison's deductions. We then examined the expressions that served the purposes of negative politeness in the distinguishable rhetorical functions or steps we have identified in this move: i*nterpreting results,* a*ccounting for results, comparing results, evaluating results, indicating limitations of results* and *exemplifying* (Soler-Monreal & Gil-Salom, in press). We considered they might be representative of the scientific writers' ways of defending their claims and interacting with readers. The identification of the politeness-marked linguistic forms required a careful analysis of the specific pragmatic function of every form and was especially attentive to the textual context. We focused on the strategies that contributed to the author's visibility and invisibility in the text.

3. THE EXPRESSION OF AUTHORIAL COMMENT

Writer-oriented and reader-oriented politeness mechanisms allow writers to strengthen credibility while keeping social links so as to cause an effect on the reader and obtain acceptance. Writer-oriented devices redress the writers' faces by reducing their presence in the text. They aim to protect the writer against possible criticism by limiting personal commitment. Reader-oriented devices, on the other hand, aim to redress the reader's face. The writer expresses personal responsibility for the content of the message, which allows the reader to choose a different interpretation.

Nominalisations, passives and impersonal constructions that eliminate writer agency are often used to give the appearance of objectivity and neutrality typical of scientific writing. But a close analysis of these expressions in context reveals that some of them are effective means of writer-reader communication and can be seen as negative politeness devices. They allow the writer's detachment and make the message more acceptable to the reader. Self-reference, on the other hand, brings a concrete presence into the text and also becomes a negative politeness device when the writer uses it as a face-saving device. First person pronouns and possessive determiners clearly signal attribution to the author. When readers' beliefs and attitudes are in conflict with those the author expresses the device avoids the threat and gets the reader actively involved.

Both impersonalisation and personal attribution co-exist in the commenting on results move in discussions, and serve the writer's face-redressive purposes in each step of this move. Examples of these practices are illustrated below.

3.1. Interpreting results

Most writers in our corpus include this step, in which arguments are interwoven with authorial comments directed at the reader. The step contains the researchers' interpretations of their findings and positioning towards them. This implies a high degree of personal exposure. Authorial non-prominence redresses the writers' face. Nominalisation is a common impersonalising strategy used to remove the writer from the text. In scientific writing authors tend to focus on the object rather than the subject, so as to consider findings the property not of the individual researcher but of the whole scientific community. In the examples that follow, the responsibility of the author for the truth of a proposition is diminished by the use of active constructions in which the personal subject is replaced by some "abstract rhetors"

(Hyland 1996: 444), such as *inspection, experiments* or *examination,* which nominalise a personal act:

(1) Inspection of the expression for α (see eq (B1)) shows that $\alpha = 0$ only when $w = win$ or $w = w$. (Robotics)

(2) Given such a higher penalty, **experiments show** that layered multicast is not necessarily superior to stream replication for many network topologies [14]. (Telecommunications)

(3) **Ultrasonic examination did not reveal** gross defects in any of the bonds and the bend tests showed the bonds to be stronger than the interlayer even although this could not be quantified for the conditions of testing employed. (Nanotechnology)

The writer also avoids personal responsibility for the validity of propositional truth by using non-human entities (e.g. *manipulator, method* or *interpolator*) as subjects of the scientific process:

(4) **The manipulator has demonstrated** higher accuracy [27], as has the WFLC in 1-DOF tremor canceling tests [14]. (Robotics)

(5) The phantom harvesting experiment showed that **the new method has the capacity to generate** stronger and longer negative pressure by the actuator, […]. (Robotics)

(6) **The speed-controlled interpolator does quite well** in terms of regulating the feed rate. (Computing)

Deductions from the findings are usually supported by figures and tables. Making explicit reference to visuals in this step gives the writer the role of a spectator rather than of an actor:

(7) Fig. 3 shows the height variation for 20 weeks. (Nanotechnology)

The same purpose is achieved by agentless passive constructions, in which the writer's reputation and commitment are saved by avoiding the agency of the action and focusing on the description of results:

(8) Hence, when the platform is in one of these Cartesian orientations, **all three actuators can be moved** freely without any motion of the platform, […]. (Robotics)

(9) Interestingly, between silica surface layer and crept Si-C-N ceramic **an intermediate zone (Fig.4(a)) was identified** where vast Si2N2O nanocrystallites precipitated homogeneously. In addition, **the Si2N2O nanocrystallites were often detected** at the nano-sized porosity (Fig. 5(c) and (d)). (Nanotechnology)

Another negative politeness strategy when interpreting results is to use impersonal *it* subjects:

(10) As a by-product of this section, **it is proven that**, […], all the translational degrees of freedom of the C joints of the mechanism cannot be actuated simultaneously. (Robotics)

(11) In this way, the number of beacons used may remain approximately constant, as **it was shown** in the second example of the paper. (Robotics)

(12) **It was also observed that** the 3D point depth relative to the cameras is very important to the uncertainty coefficients. (Robotics)

The indefinite pronoun *one* is also a distancing device (Garcés-Conejos & Sánchez-Macarro, 1998). In our corpus *one* refers to any researcher in general (including the writer) and the purpose is to stress the shared interest, aim, or knowledge. In the example *one* refers to any researcher, and the writer is suggesting a possibility of something to happen. This hedging softens his/her commitment to the truth of the proposition:

(13) For the interactive search task, as actual web pages were used, **one would expect** participants to use these automatic attention responses. (Computing)

However, certain strategies indicate that the authors remain behind their arguments in order to guide readers through the discourse. This happens in a variety of expressions conveying attitude. The impersonal pronoun *it* followed by a periphrastic verbal expression plus an attitudinal marker, such as *easy* and *critical* indicates evaluation in terms of social value, establishes interpersonal bonds with readers and guides them in the acceptance of authors' claims:

(14) Computing analytically γ for $w = win$ or $w = wfin$**, it is easy to see that** γ is negative only in the trivial cases where the initial or final (desired) position for the system's point under consideration is inside an obstacle. (Robotics)

(15) Moreover, **it is critical to select** the precursor with highly polymerization or strong coagulation ability for forming the shell particles. (Nanotechnology)

Epistemic expressions (e.g. *possible* and *certain*) mask personal convictions but attempt to guide the reader in the interpretation of the facts:

(16) Thus **it is possible to point out** which is the critical factor for each method. (Robotics)

(17) **It is also not certain that** increased prior consolidation would favour bonding. (Nanotechnology)

But personal attribution is also found where writers make a knowledge claim. Writers choose to make themselves visible through the use of the first personal pronoun *we* to signal their presence as researchers, to emphasise their role in the research process and to stress a contribution to their field of research. The use of *we* associated with verbs of cognition (e.g. *note* and *believe*) or judgement (e.g. *conclude*) shows the writers' distinctive commitment and confidence:

(18) Due to space limitation the results for single fiber are not shown here, however, **we note that** in multifiber networks there is a slight improvement in the performance in terms of T(a), M(a), and W(a) (over single-fiber networks), […]. (Telecommunications)

(19) After detail analysis of these protocols, **we believe that** anomaly detection works better on a routing protocol in which a degree of redundancy exists within its infrastructure. (Telecommunications)

(20) **We conclude that**, for the DCCE method, both in the differential and in the discrete formulations, the critical factor is the disparity […]. (Robotics)

Personal attribution can also be considered as a strategy to mitigate categorical claims. In the following examples the determiner *our* followed by *results, interpretation and belief* in the process of deduction offers readers the possibility of having alternative opinions and invites them to become involved in the interpretation of propositions:

(21) **Our results**, especially those shown in Fig. 5(a), **suggest** that the end mill does not necessarily drill faster with increasing rotation speed, […]. (Robotics)

(22) Once again, **our interpretation of these findings is that** more informed attitudes about the Internet, albeit skeptical attitudes, lead to greater Internet use. (Computing)

(23) **Following from our belief that** the exact definition of consistency is application-specific, the question of whether consistency is "continuous" or how "continuous" it is also depends upon application semantics. (Computing)

The interpretation of results can also be attributed to visuals, supporting sources for readers to agree with the writer's own position:

(24) **From the figures**, **we notice that** the total number of ports in the network and the maximum number of ports at a node among all nodes by using BPHT is less than those using BTMH and much less than those using WBO-RWA. (Telecommunications)

3.2. Comparing results

In this step authors refer to earlier accomplishments in the field and challenge other writers' statements. This gives way to potentially face-threatening acts to the writers' own face and that of members of the scientific community with whom they disagree. It is usual for writers to avoid being responsible of the criticising position through choosing non-human actors that hide personal implication:

(25) Compared to the paper by Schilling et al. [33], the path following algorithm does not suffer from singularities. (Robotics)

(26) **This finding also challenges the idea that** blue links may distract users from other tasks (due to them being easily seen in the periphery; Murch, 1984) and […]. (Computing)

However, when they compare approaches or viewpoints and seek recognition for their own merits, self-reference is also frequent, even if it is disguised by attributing the claim to a third person. The strategy is reader-oriented: authors emphasise their personal contributions, which allows readers to think differently and reduces the threat to their faces:

(27) […] **the main difference between our *lpbcast* algorithm and *pbcast* [Birman et al. 1999]** is that **our approach** melts the two phases of *pbcast* (dissemination of events and exchange of digests) into a single phase. (Computing)

(28) **Sung and Poggio assume** that the number of faces in Test set 2 is 149, however, **Rowley et al. assume** that there are 155 faces in the same test set. **The author also assumes** that the number of faces in Test set 2 is 155. (Computing)

3.3. Accounting for results

In this step authors provide an explanation for the results. Again, eliminating writer agency by means of visuals, passives and impersonalisations redresses the writers' faces:

(29) In fact, Fig. 5(c) and (d) provide sound evidence to support this idea in which only Si2N2O phases were detected. (Nanotechnology)

(30) The improvements **are attributed to** the low accumulated link dispersion and the low crosstalk level from AM-VSB and 256-QAM to OC-48 channel. (Telecommunications)

(31) **It needs to be noted that**, […], the signal may show a large fluctuation strongly depending on sampling and counting from the statistics point of view. (Nanotechnology)

The tendency is also to soften appreciations and avoid presenting information categorically (modals, epistemic expressions). This has the effect of conveying deference, humility and respect for colleagues' views, while protecting the writer from criticism:

(32) **This can be explained** as follows: sometimes, to reduce port count, a longer path that utilizes a wavelength in a band may be chosen even though a shorter path (that cannot be packed into a band) exists. (Telecommunications)

(33) […] **it can be attributed to** the appropriate received optical power at the receiver site. (Telecommunications)

(34) **It is understandable that** smaller grains with higher resistance to dislocation motion **should act as** obstacles to deformation. […]. (Nanotechnology)

Getting personally involved in the explanation of the findings increases writer responsibility:

(35) […], which **we attribute** to the inherent partial wavelength conversion capability. (Telecommunications)

(36) For this reason we have proposed a model of meshing friction. (Robotics)

(37) **We have been assuming that** the numerical weight and order weight of a write are independent of the database state, **which enables** the application to determine the weight solely based on the write itself. (Computing)

This downplays the imposition on the reader. However, the persuasion process is at work by means of evidential expressions:

(38) **It is evident that** grain growth (comparing Fig. 3c with Fig. 4b) may soften the material and enable the intra-granular deformation to become easier, […]. (Nanotechnology)

(39) On the other hand, **we can prove that** with the optimal allocation, increasing the number of layers always leads to a higher degree of fairness [10]. (Telecommunications)

3.4. Evaluating results

This step evaluates the outcome and makes a claim about the generalisability of some or all of the reported results. References to a technique, the work done, or the contributions to research are effective means of hedging personal commitment:

(40) **This technique makes possible** gigabit Ethernet to the classrooms through the existed MMF in the campus. (Telecommunications)

(41) **The present work** does not change current concepts of ceramic–ceramic bonding but it **does enlarge** the range of parameters under which such bonding can be carried out. (Nanotechnology)

(42) […], **this research suggests** that more thought has to be put into the designing of measures for psychological experiments […]. (Computing)

At this stage authors highlight the strengths of their study. They stress the advantages offered by their work and this allows them to claim that their contribution to research is valuable. This justifies self-mention, combined with evaluative language with positive meanings. The strategy serves several purposes: writers emphasise the importance of their work, writers protect themselves from opposing opinions and writers guide the readers in the evaluation of the research while letting them have the impression that they are free to judge the research presented:

(43) Thus, **we believe our metric set is a good tradeoff** between complexity and the semantics space covered. Furthermore, **our model** mainly targets wide-area applications and **our experiences suggest** these three metrics **are a good fit** for these applications. (Computing)

(44) However, we find that the total port number and maximum port number for WBO-RWA increase rapidly with an increase in the number of fibers, further indicating **the effectiveness of our proposed BPHT algorithm**. (Telecommunications)

3.5. Indicating limitations of results

Writers honestly acknowledge gaps and weaknesses in order to present claims and results in the light of accepted knowledge. Expressions of contrast and lexis with negative meaning are characteristic of this step. They convey pessimism about fairly strong claims. Since mentioning the limitations of the research process can be a threatening act to the writer's face, personal attribution is avoided:

(45) In addition, **this method fails to change the orientation** of the platform if the initial and final location of ho*G* are the same. (Robotics)

(46) An extension to model 3D changes is needed. (Computing)

(47) Although **these shortcomings can be addressed** with some additional burden, **more serious problems appear** when it comes to obstacle avoidance. (Robotics)

3.6. Exemplifying

In this step examples are provided to clarify or specify a statement. This strategy maintains the bond with the reader and guides her/him through the argumentation:

(48) […] having a large number of links resembles continuous morphologies **as seen from the examples given**. (Robotics)

When the author explicitly asserts the truth of a proposition, this is left open to the reader's judgement:

(49) As an example of a general holding time distribution, we also consider the Weibull distribution with shape parameters which show the hyper and hypo exponential behaviors, respectively. (Telecommunications)

4. CONCLUSION

This study has analysed negative politeness in the *commenting on results* move in engineering RA discussions. It has described strategies of authorial involvement and authorial detachment aiming at redressing the threat of research claims to the writer's and reader's faces. When they interpret, account for, evaluate and exemplify results, writers use the strategies of vagueness and impersonalisation to reduce their commitment to claims. But they also choose to make their presence explicit, thus reducing the imposition on readers and contributing to a more intimate discourse which leaves room for discussion and differences of opinion. On the other hand, writer effacement seems to be preferred for redressing the face-threatening acts of criticising other works and acknowledging the limitations of the research presented. The analysis demonstrates that in the construction of a text the writer counts on rhetorical and pragmatic options. Controlling the resources for structuring information, interacting with the reader and personally intervening in the text determine successful participation in academic environments.

APPENDIX

Journals in computing:

ACM Transactions on Computer Systems, Computer-Aided Design Computer Vision and Image Understanding, International Journal of Human-Computer Studies.

Journals in robotics:

Robotics and Autonomous Systems, Artificial Intelligence, IEEE Transactions on Robotics and Automation, International Journal of Robotics Research, Robotics and Autonomous Systems, Robotics and Computer Integrated Manufacturing.

Journals in telecommunications:

IEEE Journal on Selected Areas in Communications, Wireless Networks, IEEE Network, IEEE Transactions on Broadcasting.

Journals in nanotechnology:

IEEE Communications Magazine, Acta Materialia, Journal of Magnetism and Magnetic Materials and Nanotechnology.

REFERENCES

Bazerman, C. (1988). *Shaping written knowledge.* Wisconsin: University of Wisconsin Press.

Brown, P. & Levinson, S. C. (1987). *Politeness: Some universals in language usage.* Cambridge: Cambridge University Press.

Garcés-Conejos, P. & Sánchez-Macarro, A. (1998). Scientific discourse as interaction: Scientific articles vs. popularizations. In A. Sánchez-Macarro & R. Carter (Eds.), *Linguistic choice across genres. Variation in spoken and written English* (Vol. 158) (pp. 173-190). Amsterdam/Philadelphia: John Benjamins

Holmes, R. (1997). Genre analysis and the social sciences: an investigation of the structure of research article discussion sections in three disciplines. *English for Specific Purposes, 16*(4), 321-337.

Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied Linguistics*, *17*(4), 433-454.

Hyland, K. (2002). Authority and invisibility: authorial identity in academic writing. *Journal of Pragmatics*, *34,* 1091-1112.

Hyland, K. (2005). Prudence, precision, and politeness: hedges in academic writing. In M.A. Olivares Pardo & F. Suau Jiménez (Eds.). *Las lenguas de especialidad: nuevas perspectivas de investigación* (Quaderns de Filologia. Estudis Linguistics 10) (pp. 99-112). Valencia: Universitat de València.

Lauren, C. & Noerdman, M. (Eds.). (1989). *Special language: From humans thinking to thinking machines*. Clevedon: Multilingual Matters.

Lewin, B.A., Fine, J. & Young, L. (2001). *Expository discourse*. London: Continuum.

Markkanen, R. & Schröder, H. (1989). Hedging as a translation problem in scientific texts. In C. Lauren & M. Noerdman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 171-179). Clevedon: Multilingual Matters.

Myers, G. (1989). The pragmatics of politeness in scientific articles. *Applied Linguistics, 10*(1), 1-35.

Olivares Pardo, M.A. & Suau Jiménez, F. (Eds.). (2005). *Las lenguas de especialidad: nuevas perspectivas de investigación* (Quaderns de Filologia. Estudis Linguistics 10). Valencia: Universitat de València.

Peacock, M. (2002). Communicative moves in the discussion section of research articles, *System*, *30*, 479-497.

Ruiying, Y. & Allison, D. (2003). Research articles in applied linguistics: moving from results to conclusions. *English for Specific Purposes, 22*(4), 365-385.

Sánchez-Macarro, A. & Carter, R. (Eds.). (1998). *Linguistic choice across genres. Variation in spoken and written English* (Vol. 158). Amsterdam/Philadelphia: John Benjamins.

Soler-Monreal, C. & L. Gil-Salom (in press). Moves, steps and linguistic signals in scientific RA discussions.

Swales, J.M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Variación en el uso de conectores causales en alemán según tipos textuales de la lengua hablada

OLIVER STRUNK
Universitat de Barcelona
CLAUDIA BUCHER
Universität Freiburg

Resumen
El uso de conectores causales es en alemán uno de los recursos más usados para crear relaciones causales entre oraciones. Para determinar si en lengua hablada hay diferencias entre conectores, si su uso está en función del tipo textual o del hablante analizaremos datos de hablantes nativos extraídos del corpus Varkom en alemán. Este estudio prospectivo se completará con un estudio comparativo entre hablantes nativos y aprendices de alemán como LE. Gracias al contraste entre nativos y aprendices esperamos poder confirmar que el grupo de los conectores causales es indicador de estadio de aprendizaje.
Palabras clave: Corpus lingüístico, alemán lengua extranjera, conectores, causalidad

Abstract
In order to create causal relations between utterances in German language, causal connectors are one of the most frequently used methods. With the aim of describing the differences in the usage of connectors in spoken language and to find out if their use is related to the text type or the speaker, we are analyzing data from the native speakers of the German part of the corpus Varkom. This preliminary study will be completed with a comparative study between native and non-native speakers, which hopefully will show that the use of causal connectors is an indicator for language proficiency.
Keywords: Corpus linguistics, German as foreign language, connectors, causality

1. INTRODUCCIÓN

La causalidad expresa una relación semántica entre proposiciones y constituye una de las principales categorías de relaciones semánticas. La causalidad puede realizarse por medio de recursos léxicos (como por ejemplo conectores), pero también por medio de fenómenos relacionados con el contenido proposicional, que permiten crear textos a partir de la unión y sucesión de oraciones. La causalidad se expresa por tanto de variadas formas implícitas y explícitas (entre otras por medio de adverbios, conjunciones y preposiciones, véase Frohning 2005: 365, Gohl 2000: 193), abarcando muy diversos fenómenos léxicos y gramaticales, por lo que aglutina y refleja un conocimiento lingüístico complejo que puede utilizarse para la descripción de la variación de un texto; en este sentido posiblemente contribuya también a la determinación del nivel de lengua alcanzado por un aprendiz.

La realización de la causalidad en la lengua se relaciona además estrechamente con los contenidos referidos y las relaciones semánticas subyacentes, hecho que se refleja de modo diferenciado en una tipología textual basada en funciones textuales (tanto de la lengua hablada como de la escrita; véase Linke, Nussbaumer, Portmann 2001: 221ss).

Los conectores, al tratarse de una lista cerrada de elementos léxicos (si bien muchos de ellos polifuncionales)(Peldszus et al. 2008: 74), constituyen un modo de realización de la causalidad que ofrece al investigador la posibilidad de búsqueda semiautomática, con lo que es posible analizar un número elevado de textos en relativamente poco tiempo. Son por esta razón un buen recurso para determinar si la causalidad es o no relevante en cuanto al nivel de LE.

2. OBJETIVOS

Partiendo del material textual del corpus Varkom (Fernández-Villanueva, Strunk: 2009), queremos describir el uso de conectores causales primero en lengua materna para poder contrastar los datos obtenidos en futuros estudios con la producción en alemán como lengua extranjera.

El objetivo principal del estudio es determinar si existe algún tipo de constante en el uso de los conectores causales en hablantes nativos del alemán en el corpus Varkom (y que pueda servirnos luego para establecer comparaciones con las producciones en lengua hablada de aprendices de alemán como LE). Para ello comprobaremos si hay conectores causales más frecuentes que otros a nivel de lengua hablada y en general (siempre a partir de los datos del corpus Varkom); si el tipo textual influye en el número de conectores causales utilizado en un texto; y finalmente, si hay diferencias significativas en el uso de conectores en función de los hablantes individuales, tanto a nivel global, como a nivel de tipos textuales.

La clarificación de estos aspectos constituye la base que permitirá en estudios posteriores, por medio de la comparación con textos producidos por hablantes no nativos, determinar si el uso de conectores causales es de utilidad para describir el nivel lingüístico de un aprendiz.

3. METODOLOGÍA

En primer lugar delimitaremos el término *conector causal* aplicado al alemán y crearemos la lista de entradas a buscar dentro del corpus a partir de la clasificación semántica que propone el *Institut für Deutsche Sprache*. Este listado se reducirá en función de las limitaciones

técnicas que puedan imponer las características específicas de cada conector causal y posteriormente en función de la frecuencia de las entradas individuales.

Para los análisis de frecuencia de los conectores, se utilizarán las herramientas propias del corpus (Exmaralda; Elan), la hoja de cálculo Excel y el paquete estadístico PASW en su versión 17. Partiendo de los textos individuales, éstos se asociaron con información sobre el hablante y el tipo textual correspondiente y se calculó el número de palabras del segmento, el número de conectores individuales, la suma del número de conectores dentro del segmento y su frecuencia relativa.

Respecto a la pregunta de si hay conectores más usados que otros, determinaremos la frecuencia de cada uno de los conectores de la lista en *todos* los textos, independientemente del tipo textual al que pertenezcan o hablante del que procedan[1]. Para determinar si en determinados tipos textuales se utilizan más conectores causales que en otros, contrastaremos la densidad de conectores en cada tipo textual. Al mismo tiempo reflejaremos la distribución individual de conectores en los diferentes tipos textuales para ver si manifiestan particularidades en este sentido. Finalmente, para averiguar si existe variación interpersonal en el uso de conectores causales según informantes, reproduciremos los valores de uso según hablantes individuales.

4. CONCEPTOS BÁSICOS

4.1. Conector

El Institut für Deutsche Sprache de Mannheim (IdS) define *conectores* como "elementos léxicos que permiten crear una relación semántica específica entre frases". En su lista de conectores, basada en Pasch et al. (2003), incluye unas 350 entradas léxicas, entre ellas unidades compuestas por varias palabras[2]. De éstas, unas 17 pueden adscribirse a la clase semántica de la causalidad.

Conceptualmente, la definición de conectores usada por el IdS presenta el problema que el conector es un agrupamiento de diferentes clases de palabra (conjunciones, adverbios…) con una base descriptiva funcional común, pero no sintáctica. Esta pertenencia a diferentes

[1] Los valores obtenidos serán válidos para la muestra de textos, asumiendo una distribución equilibrada en cuanto a género y procedencia geográfica de los hablantes y un valor constante en cuanto a nivel educativo y edad (Fernández-Villanueva, Strunk, 2009) dentro del corpus Varkom.

[2] „Konnektoren sind sprachliche Ausdrücke, die Sätze in eine spezifische semantische Beziehung zueinander setzen können. Das sind im Deutschen etwa 350 Ausdrücke, die traditionell als Konjunktionen und als bestimmte Subklassen von Adverbien und Partikeln beschrieben werden." http://hypermedia.ids-mannheim.de/pls/public/sysgram.ansicht?v_typ=d&v_id=1182

clases de palabra comporta que una palabra concreta pueda tener varios significados y funciones, por lo que durante el análisis concreto es preciso desambiguar las apariciones y determinar si el uso real corresponde a la función de establecer una relación entre oraciones o se trata de un mero homónimo.

Los conectores causales comparten esta característica y se subclasifican generalmente según la clase de palabra a la que pertenecen en conjunciones, subjunciones, adverbios y preposiciones. Para nuestro estudio excluimos de la lista las preposiciones, porque si bien permiten crear relaciones semánticas, no tienen las mismas implicaciones sintácticas propias de conjunciones y adverbios conjuncionales.

4.2. La estrecha relación entre causalidad y consecutividad

Aparte de los conectores clasificados semánticamente como causales, es preciso tener en cuenta otro grupo de conectores que, por sus características semánticas, mantienen una estrecha relación con la causalidad: los conectores consecutivos. Porque si para definir la relación causal puede decirse que *en la segunda oración se formula la razón del contenido de la primera* (invirtiéndose el orden según el conector), para definir la relación consecutiva puede decirse que *en la segunda oración se formula la consecuencia de la primera*. Entre los conectores consecutivos figuran entradas como *also, daher, darum, deshalb, infolgedessen, somit, mitin*, todos ellos con una fuerte connotación causal. Por esta razón optamos por ampliar el grupo de conectores causales definido por el IdS incluyendo el grupo de los conectores consecutivos, ampliándose la lista de 17 conectores causales con 21 conectores consecutivos, llegando hasta un total de 38 conectores.

5. DATOS UTILIZADOS

5.1. Conectores

El IdS lista los siguientes conectores como conectores semánticamente causales:

- *da*
- *schließlich*
- *wegen*
- *immerhin*
- *umso weniger, als*
- *sintemal(en)*
- *aufgrund dessen (...), dass*
- *nämlich*
- *weil*
- *denn*
- *um dessentwillen*
- *doch*
- *umso mehr, als*
- *zumal (da)*
- *wo*
- *infolge*
- *alldieweil*
- *ob*
- *doch*
- *nachdem*

En el grupo de los conectores consecutivos del IdS, aparecen los siguientes:

- *dass*
- *dass*
- *deswegen*
- *weshalb*
- *daher*
- *demnach*
- *von daher*
- *demzufolge*
- *drum*
- *aufgrund dessen*
- *weswegen*
- *deshalb*
- *dementsprechend*
- *damit*
- *ergo*
- *darum*
- *sodass*
- *danach*
- *demgemäß*
- *infolgedessen*
- *mithin*

5.2. Limitaciones intrínsecas a los conectores individuales

Tal como se ha indicado en el apartado dedicado a metodología, hay conectores de estas dos listas que presentan limitaciones en cuanto a la posibilidad de identificación en un corpus. Se trata en primer lugar de conectores que, por razones de tamaño y de tipología, no aparecen en el corpus Varkom. Coinciden básicamente con los conectores menos usados según otras fuentes y estudios de corpus relacionados con conectores[3]. Sin embargo, hay también considerables diferencias en el uso de conectores causales según se trate de lengua escrita o hablada, como se aprecia a partir de Frohning (2005: 52), donde se realiza un estudio de corpus sobre conectores causales basado en un corpus de textos periodísticos del IdS Mannheim.

Pero aparte de eliminar de la lista a los conectores que no figuran como tales en el corpus Varkom, también se optó por eliminar de la lista a los conectores compuestos por más de una palabra, por las dificultades computacionales que presentan (se trata de elementos discontinuos que podían además aparecer flexionados, dificultando así la desambiguación) y por las reducidas frecuencias que tenían.

De este modo se conservaron finalmente los siguientes conectores causales o consecutivos:

causales: *da; wo; zumal; denn; wegen; weil*

[3] Mencionamos aquí el ejemplo de los datos de la base de datos Wortschatz Leipzig, http://wortschatz.uni-leipzig.de. Creando una lista de frecuencia a partir de los datos disponibles en esta base de datos, sin proceder a la desambiguación de las formas homónimas, y siendo la entrada más frecuente cuanto menos elevado es su índice de frecuencia, obtenemos los siguientes valores: clase 5: *damit*; clase 6: *weil, da, doch, denn*; clase 7: *deshalb*; clase 8: *nämlich*, *daher*; clase 9: *aufgrund*; clase 10: *deswegen*, *weshalb*; clase 11: *umso* (*weniger*); clase 13: *mithin*; clase 16: *schliesslich*, *infolgedessen*; clase 20: *alldieweil*

consecutivos: *deswegen; weswegen; deshalb; daher; weshalb*

Estos conectores se han buscado en las transcripciones ortográficas de las entrevistas utilizadas y se han desambiguado, si las características lo requerían.

5.3 Corpus de textos

Se han usado como base inicial los textos de Varkom[4] presentados en Fernández-Villanueva y Strunk 2009. El corpus Varkom es un extracto parcial a partir de una serie de entrevistas realizadas entre 2003 y 2009 a estudiantes alemanes y españoles, que dará lugar a un corpus contrastivo mayor y más complejo, actualmente en fase de elaboración. En el corpus Varkom aquí utilizado se recoge únicamente la producción de una parte de los hablantes alemanes nativos. Se trata de la producción en lengua hablada de 18 hablantes, procedente de la transcripción de una entrevista grabada en vídeo. Las entrevistas han sido segmentadas en función de la tarea asignada a cada parte de la misma (descripción del propio piso, instrucción para hacer el camino desde casa a la universidad, argumentación sobre un tema determinado…). A cada segmento así obtenido (entre 20 y 30 por entrevista) se le ha asignado un tipo textual en función de la tarea lingüística realizada[5]:

Tabla 2: Tipos textuales del corpus Varkom

Núm.	*Tipo textual*	Núm. textos	Núm. palabras
1	*Erzählung* (narración)	82	19.391
2	*Beschreibung* (descripción)	55	11.840
3	*Anleitung* (instrucción)	218	10.581
4	*Erörterung*/*Darstellung* (explicación)	42	16.861
5	*Argumentation* (argumentación)	35	17.034
		432	75.707

[4] Varkom se ha compilado dentro del marco de los proyectos "VARCOM Variació, comunicació multimodal i plurilingüisme: estils discursius i consciència lingüística" (2001-2004), "PRAGMAESTIL Pragmatik, Stil und Identitäten. Untersuchung verbalen und non-verbalen Verhaltens mehrsprachiger Sprecher (2005-2008)" y "COHESTIL"(2009-2012), liderados por Lluís Payrató de la Universitat de Barcelona, financiados por el Ministerio de Ciencia y Tecnología (BFF2001-2004) y el Ministerio de Educación y Ciencia (HUM2005-04936/FILO, FFI2008-01230/FILO). La parte en lenguas alemana, castellana y catalana se elabora bajo la dirección de. Fernández-Villanueva y Oliver Strunk con la colaboración de Yurena Alcalá, Ines Bernauer, Ramona Borsch, Claudia Bucher, Judit del Mestre, Karin Fischer, Kathleen Räthel, Katrin Schmidt, Andrea Schreiner, Ferran Sunyer, Eduard Tapia y Katharina Wörner.

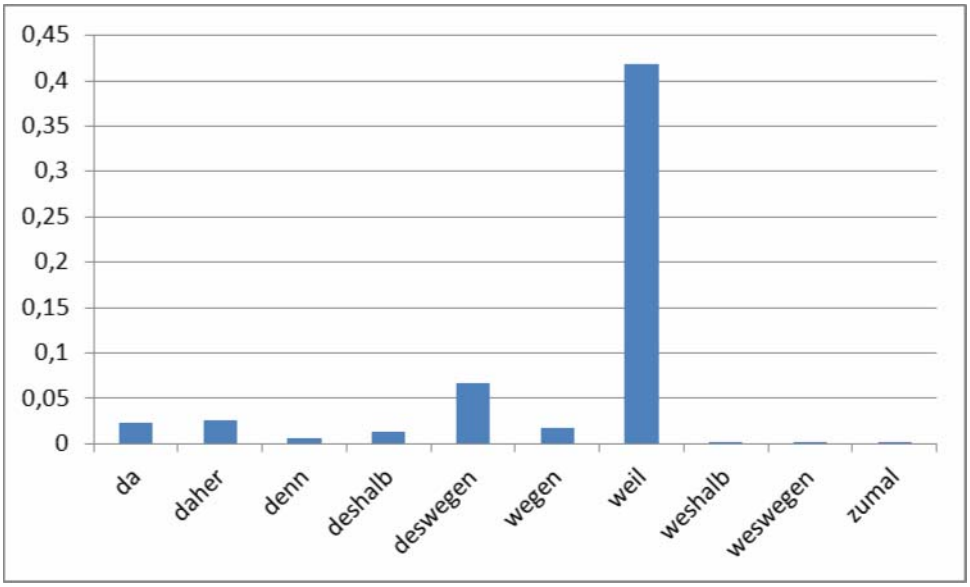
[5] En el corpus Varkom se incluyen las grabaciones en vídeo de las entrevistas y las correspondientes transcripciones ortográficas y GAT de 18 hablantes alemanes nativos. Los hablantes se han seleccionado por su procedencia geográfica (Antigua RFA, antigua RDA, Austria) y sexo (9 mujeres, 9 hombres, equilibrados también según procedencia geográfica).

6. Distribución por frecuencias de conectores causales en Varkom con independencia de la tipología textual

Hemos visto ya que no todos los conectores causales listados por el IdS aparecen también en el corpus Varkom, y que otros no se han contemplado por motivos técnicos. La lista final de conectores causales usados incluye 10 entradas; al parecer, la lengua hablada es menos diversa en cuanto al uso de conectores causales que la lengua escrita. Quedan eliminados conectores de la lengua escrita y comúnmente considerados cultos como *aufgrund dessen, infolge, sintemal, alldieweil,* etc., pero también conectores considerados prototípicos de la lengua hablada, como *wo* (Günthner, 2002: 310). Entre los conectores usados para el análisis, cinco pueden considerarse estrictamente *causales* (*weil, da, wegen, denn, zumal*), mientras que los restantes cinco pertenecen al grupo de los conectores consecutivos, si bien tienen función causal (*deswegen, daher, wegen, deshalb, denn*).

Una vez ejecutada la concordancia de los conectores causales de la lista y desambiguadas las listas KWIC, hemos calculado la frecuencia de conectores causales relativa basándonos en el número total de palabras del corpus (75.707 palabras para todos los segmentos).

Tabla 3: Frecuencia relativa de conectores causales

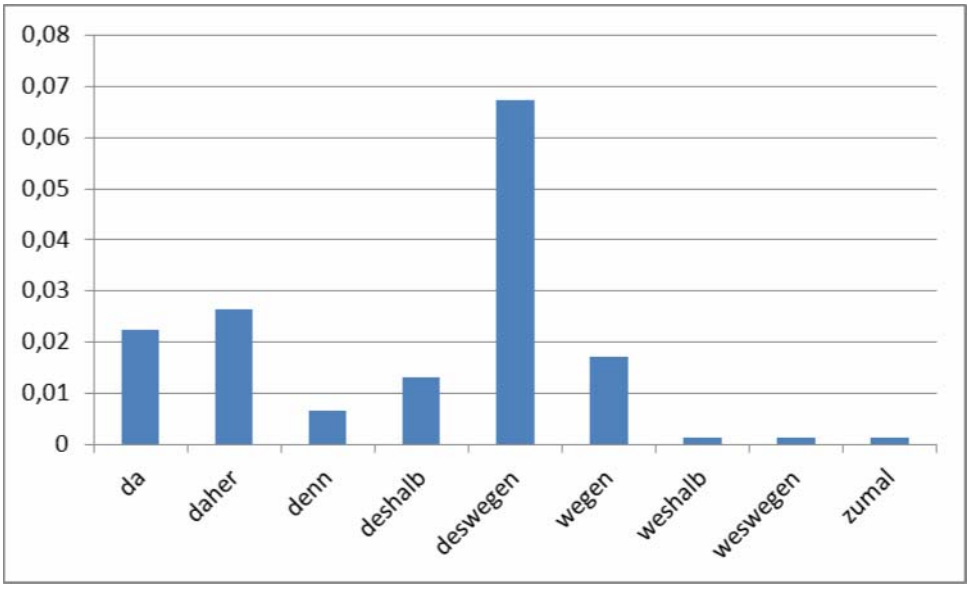


Los 436 usos detectados de conectores causales representan una densidad de un 0,58% o, lo que es lo mismo, un conector por cada 172 palabras. En comparación con otras clases de palabra (conjunciones, adjetivos…) presentan valores muy reducidos, fuertemente marcados además por la presencia de un elemente extremadamente frecuente en comparación con los

demás, *weil*. Este elemento es más frecuente que la suma de todos los demás y constituye probablemente un caso aparte. Su valor relativo de 0,42% implica que se utiliza aproximadamente una vez cada 240 palabras. Partiendo de un promedio de 174 palabras por segmento aparecería algo menos de una vez por segmento. Si eliminamos los segmentos correspondientes al tipo textual 3 (a menudo con un número reducido de palabras) y recalculamos los valores, el promedio de palabras por segmento sería de 300, en cuyo caso *weil* aparecería unas 1,5 veces por segmento. Situándose en este nivel, el uso de *weil* podría utilizarse como elemento de contraste entre nativos y aprendices, siempre y cuando se demuestre que éstos presentan una frecuencia de uso diferente.

En cuanto al resto de los conectores, parecen "insignificantes" a primera vista, pero la suma de los elementos individuales sí puede arrojar alguna información útil. A continuación presentamos en el gráfico las frecuencias relativas sin la presencia de *weil* como elemento dominante:

Tabla 4: Frecuencia relativa de conectores causales sin *weil*



La tabla numérica ordenada por frecuencia del grupo entero es la siguiente:

Tabla 5: Conectores causales en Varkom

Conector	Frecuencia absoluta	Frecuencia relativa en %
weshalb	1	0,0013
weswegen	1	0,0013
zumal	1	0,0013
denn	5	0,0066
deshalb	10	0,0132
wegen	13	0,0172
da	17	0,0225
daher	20	0,0264
deswegen	51	0,0674
weil	317	0,4187

Comparando estos valores con los arrojados por el corpus periodístico anteriormente mencionado, obtenemos una considerable diferencia, aun teniendo en cuenta que en algunos casos la frecuencia absoluta tan reducida de conectores como *weshalb* o *weswegen* los hace inadecuados para presentaciones cuantitativas.

Frohning (2005) analizó textos de los años 1991 y 1994-96, que sumaban un total de 17,14 millones de tokens. El recuento de conectores dio lugar a la siguiente tabla de frecuencias:

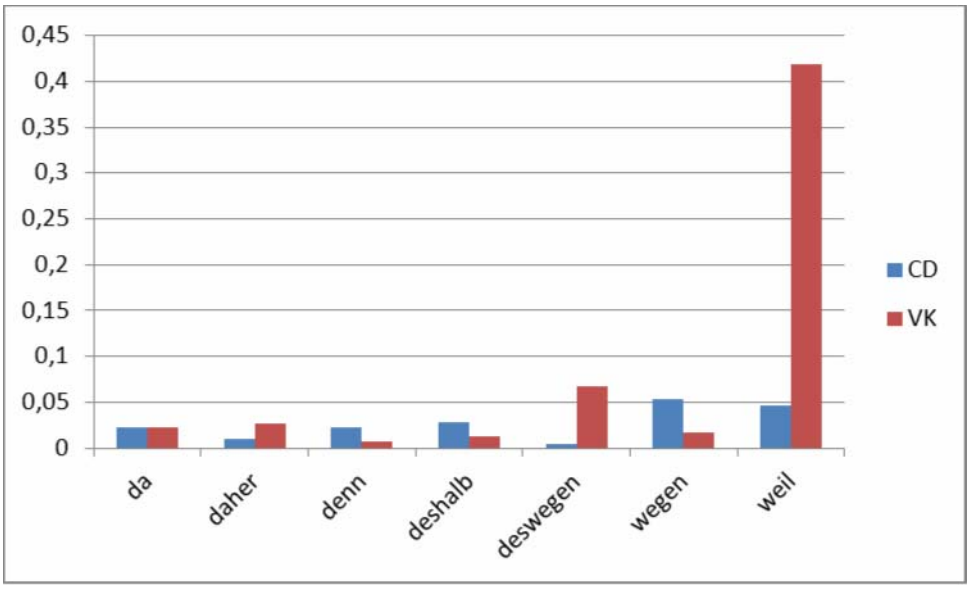
Tabla 6: Conectores en Frohning (2005)

Conector	Frecuencia absoluta	Frecuencia relativa en %
wegen	9179	0,0536
weil	8045	0,0469
deshalb	4779	0,0279
denn	3968	0,0232
da	3939	0,0230
daher	1696	0,0099
aufgrund	1694	0,0099
nämlich	1457	0,0085
deswegen	653	0,0038
darum	151	0,0009
suma	35561	0,2075

La frecuencia relativa de la suma de estos conectores de Frohning se sitúa en 0,21%, lo que equivale a un conector por cada 481 palabras.

Para comparar estos valores de un corpus de textos escrito con nuestro corpus oral nos basamos exclusivamente en el común denominador de conectores, eliminando aquellos conectores que no han sido tenidos en cuenta por uno de los dos estudios. Comparando las frecuencias restantes, se aprecia que el corpus de diarios utiliza un 0,188% de conectores causales, frente a un 0,571 del corpus Varkom. Las diferencias más significativas se dan en unos pocos conectores, como se aprecia en el gráfico:

Tabla 7: Comparativa entre el corpus periodístico y el corpus oral Varkom



Los conectores más destacados, *deswegen*, *wegen* y *weil*, presentan grandes diferencias en un corpus y en otro, pero el más significativo es sin duda *weil*, diez veces más frecuente en Varkom que en el corpus de diarios. Si se comparan los conectores individuales y se toma el uso en el corpus de diarios como 100%, se obtienen las siguientes desviaciones en el corpus Varkom:

Tabla 8: Desviación entre el corpus perioddístico y Varkom

Conector	Diarios (en %)	Varkom (en %)	Desviación (en %)
da	0,022981	0,022456	97,71
daher	0,009895	0,026418	266,99
denn	0,023151	0,006605	28,53
deshalb	0,027882	0,013209	47,38
deswegen	0,00381	0,067367	1768,25
wegen	0,053553	0,017172	32,07
weil	0,046937	0,418731	892,11

A partir de esta información, y a falta de estudios más exhaustivos con tipos textuales más diversificados, puede empezar a dibujarse la idea que los conectores causales más frecuentemente utilizados en el corpus Varkom constituyen una marca de oralidad. Un uso frecuente de *deswegen, weil, daher* y *da* parece ser elemento distintivo de la lengua hablada.

Deswegen y *weil*, más habituales en la lengua hablada de Varkom, implican que sigue una estructura verbal, estructura comúnmente asociada a la lengua hablada; mientas que *wegen*, más frecuente en el corpus de diarios, implica un genitivo y con frecuencia una estructura nominal, que se asocia habitualmente con textos escritos.

Sin embargo, es preciso tener en cuenta que las propiedades semánticas de los conectores causales hacen que estén estrechamente ligados al tipo textual utilizado, por lo que todavía queda por confirmar que las grandes diferencias entre los textos escritos del corpus de diarios y los textos orales del corpus Varkom se mantienen estables en otras comparaciones entre textos orales y escritos.

7. DISTRIBUCIÓN POR FRECUENCIAS DE CONECTORES CAUSALES EN VARKROM EN FUNCIÓN DE LA TIPOLOGÍA TEXTUAL

El segundo objetivo de este trabajo contribuirá en parte a ver si las diferencias detectadas son estables. Para desentrañar la relación entre el uso semántico de un conector causal y el tipo textual en que se aplica, vamos a analizar su distribución en los 5 tipos textuales reflejados en el corpus Varkom (*Erzählung, Beschreibung, Anleitung, Erörterung, Argumentation*).

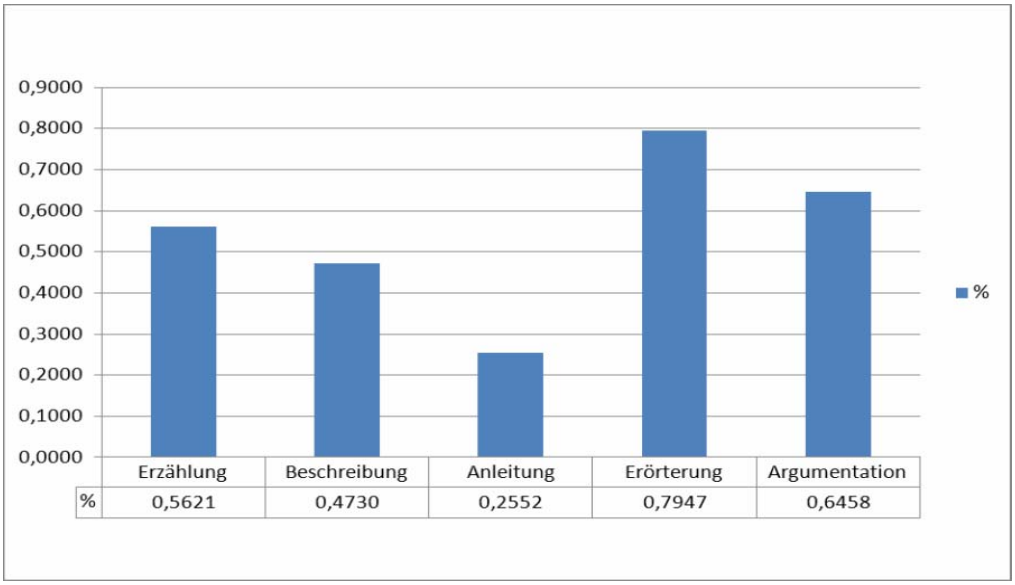
Para ello los datos se han sometido primero a una serie de pruebas de distribución normal. Para el conjunto de los datos de frecuencias relativas según los 432 textos individuales no se ha podido comprobar que correspondan a una distribución normal. Únicamente los textos del tipo textual 5 (Argumentation) obtienen un valor de 0,2 en la prueba de Kolmogorov-Smirnov. No obstante, aplicando la correspondiente prueba paramétrica, la prueba de normalidad Kruskal Wallis arroja una diferencia estadística significativa según tipo textual, con los siguientes rangos:

Tabla 9: Prueba de normalidad Kruskal-Wallis: relación entre frecuencia de conectores y tipos textuales

Rangos	Tipo textual	N	Rango medio
diension1	1	82	236,30
	2	55	249,31
	3	207	157,01
	4	42	298,52
	5	35	305,83
	suma	421	

Ello indica que las diferencias en el uso de conectores dependen del tipo textual utilizado y que los dos tipos textuales que más consistentemente demandan la presencia de conectores son los que presentan una mayor frecuencia relativa:

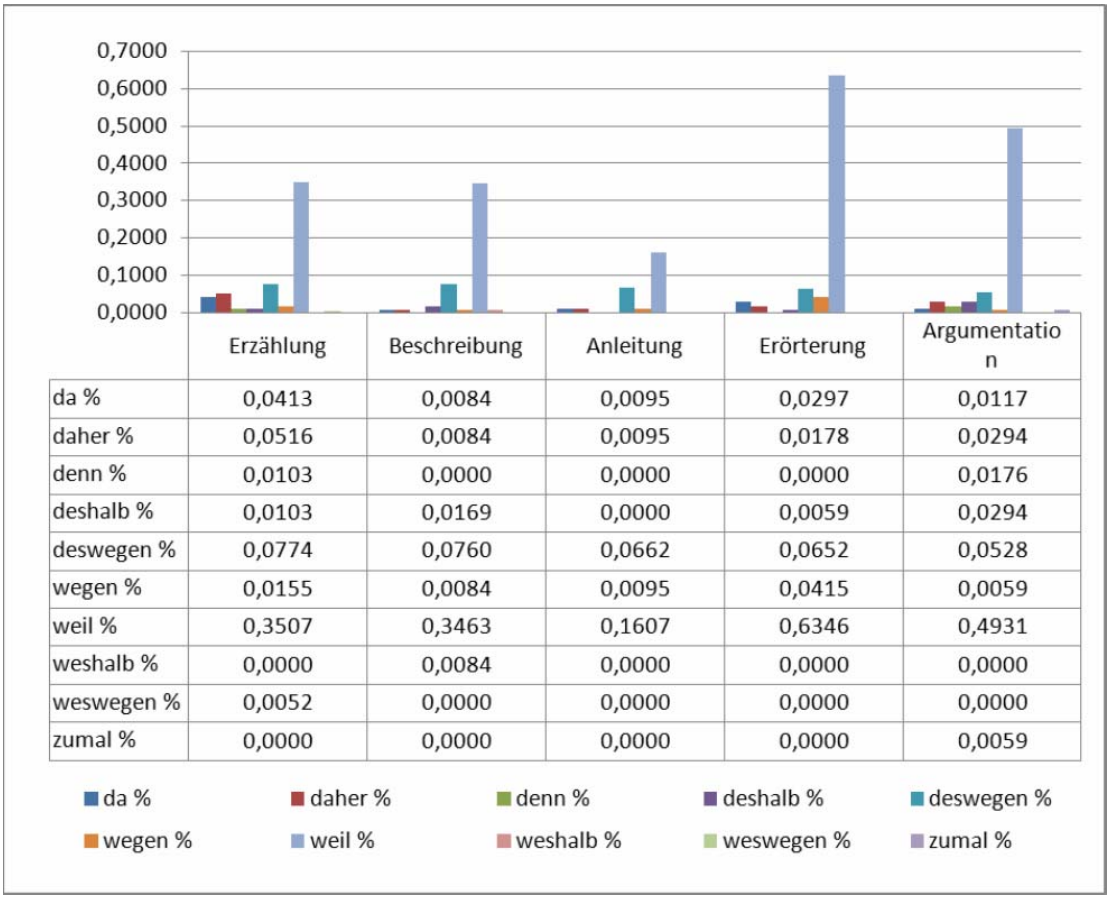
Tabla 10: Porcentaje de conectores causales según tipo textual



La elevada presencia de conectores causales puede considerarse por tanto indicio de que nos encontramos ante un texto argumentativo o explicativo.

En cuanto al uso de conectores individuales, y basándonos únicamente en los valores relativos, las diferencias se acentúan:

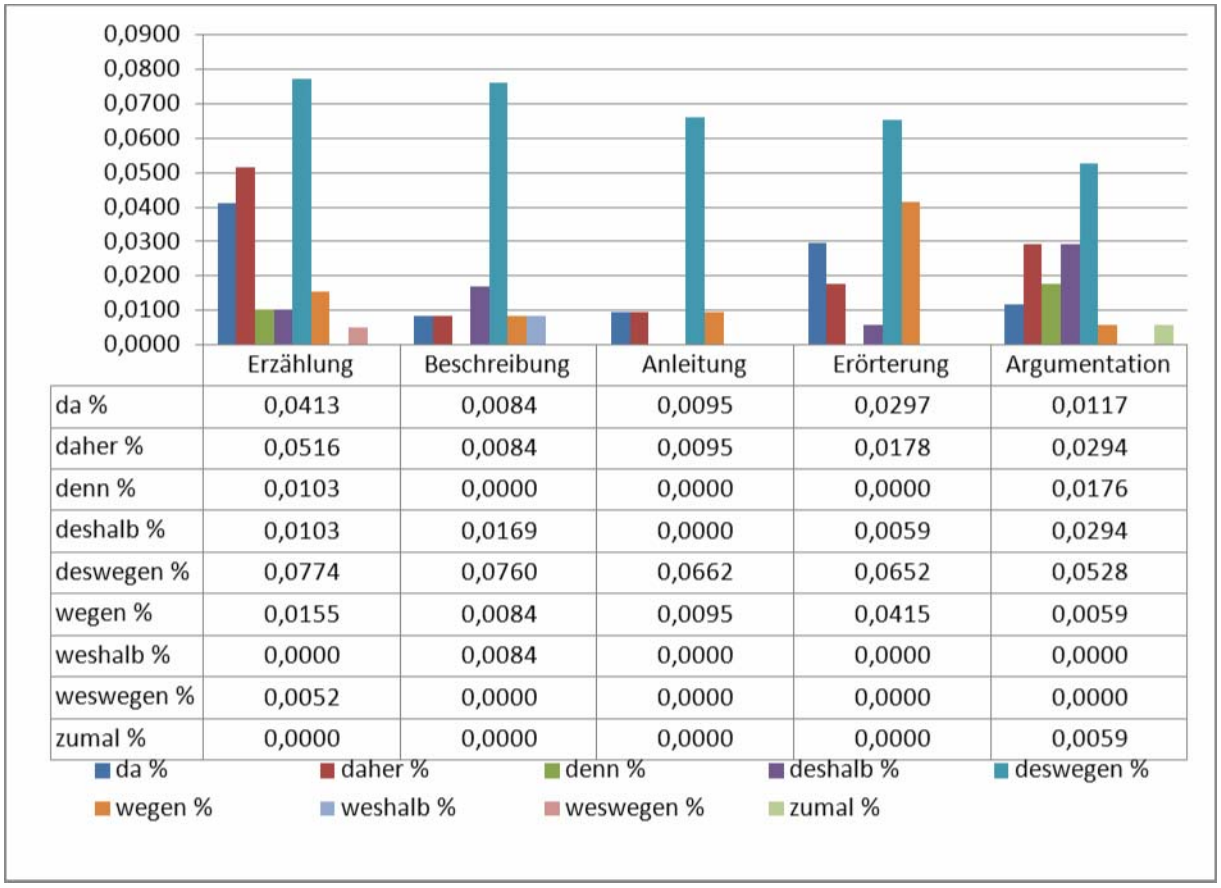
Tabla 11: Distribución de conectores causales individuales según tipos textuales



	Erzählung	Beschreibung	Anleitung	Erörterung	Argumentation
da %	0,0413	0,0084	0,0095	0,0297	0,0117
daher %	0,0516	0,0084	0,0095	0,0178	0,0294
denn %	0,0103	0,0000	0,0000	0,0000	0,0176
deshalb %	0,0103	0,0169	0,0000	0,0059	0,0294
deswegen %	0,0774	0,0760	0,0662	0,0652	0,0528
wegen %	0,0155	0,0084	0,0095	0,0415	0,0059
weil %	0,3507	0,3463	0,1607	0,6346	0,4931
weshalb %	0,0000	0,0084	0,0000	0,0000	0,0000
weswegen %	0,0052	0,0000	0,0000	0,0000	0,0000
zumal %	0,0000	0,0000	0,0000	0,0000	0,0059

Existe un claro predominio del conector causal *weil*, muy alejado porcentualmente de los demás. Su distribución varía además de un tipo textual a otro; excluyendo el caso de las *Anleitungen*, *weil* domina en las *Erörterungen* y las *Argumentationen*, al parecer incluso en detrimento de la diversidad de los demás conectores causales. La frecuencia en el uso de *weil*, igual que la frecuencia global en el uso de conectores causales, parece ser pues indicador de un tipo textual determinado. El resto de los conectores, excluyendo *weil* como elemento distorsionante, da lugar al siguiente gráfico:

Tabla 12: Distribución de conectores causales individuales según tipo textual sin "weil"



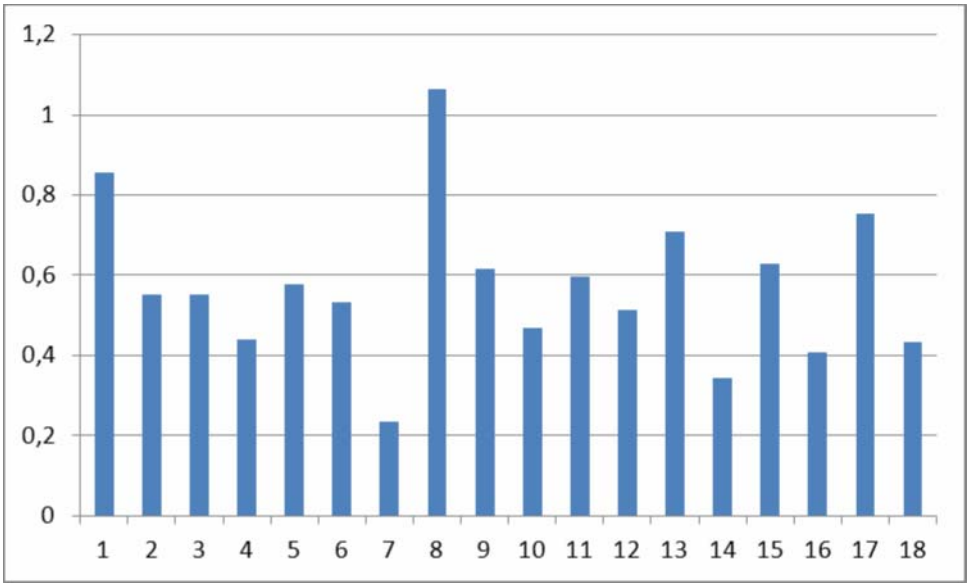
	Erzählung	Beschreibung	Anleitung	Erörterung	Argumentation
da %	0,0413	0,0084	0,0095	0,0297	0,0117
daher %	0,0516	0,0084	0,0095	0,0178	0,0294
denn %	0,0103	0,0000	0,0000	0,0000	0,0176
deshalb %	0,0103	0,0169	0,0000	0,0059	0,0294
deswegen %	0,0774	0,0760	0,0662	0,0652	0,0528
wegen %	0,0155	0,0084	0,0095	0,0415	0,0059
weshalb %	0,0000	0,0084	0,0000	0,0000	0,0000
weswegen %	0,0052	0,0000	0,0000	0,0000	0,0000
zumal %	0,0000	0,0000	0,0000	0,0000	0,0059

Una vez eliminado, emerge *deswegen* como conector causal más frecuente en los textos hablados. A partir de estos gráficos se aprecia cierta dispersión en el uso de conectores causales.

8. Distribución de conectores causales en Varkom en función del hablante

Para estudiar la distribución de conectores causales por hablante hemos sumado primero el número de palabras de cada fragmento de cada hablante, luego el número de conectores y finalmente calculado el porcentaje correspondiente.

Tabla 13: Frecuencia de conectores según hablante



Según la prueba de Kruskali Wallis (Asymp. Sig. = 0,454) no puede establecerse una relación entre hablante y frecuencia en el uso de conectores. Podría intentar establecerse una relación adicional entre hablante y tipo textual (específicamente el tipo 5, la argumentación, que manifestaba una distribución normal en cuanto a porcentaje de conectores) y entre hablante y conector causal concreto (para detectar preferencias individuales). No obstante, utilizando los textos del subcorpus elegido, los datos serían demasiado reducidos, por lo que conviene incorporar datos adicionales para este tipo de análisis.

9. Resumen y conclusiones

Los hablantes nativos de alemán presentan un uso diferenciado de conectores causales. De todos los conectores causales, el más frecuente en lengua hablada es *weil*, aproximadamente diez veces más frecuente que cualquiera de los demás, y mucho más frecuente también que su uso en un corpus de periódicos de referencia. *Weil* puede utilizarse con dos estructuras sintácticas diferentes (formalmente o en lengua escrita con verbo al final o a nivel coloquial con verbo en segunda posición, equivalente en este caso al uso de *denn*, véase Häcker, 1994: 30), lo que aumenta todavía más su usabilidad y requiere un estudio individualizado de su uso en el corpus. La diversidad sintáctica en el uso de *weil* y su frecuencia de uso sugiere que se excluya del perfil de distribución de conectores causales y que, antes de usarlo como referencia o como indicador de nivel lingüístico, se estudie la frecuencia de las diferentes estructuras en nativos y luego se contraste con aprendices. La frecuencia de los demás

conectores en cambio permite crear perfiles de frecuencias de la lengua del hablante, que puede posteriormente dar lugar al contraste de perfiles entre nativos y aprendices.

En cuanto a los tipos textuales, el tipo es un factor que influye en la frecuencia de conectores, siendo esta diferencia especialmente destacada en el tipo de textos argumentativo, seguido del explicativo. En este sentido los conectores causales pueden considerarse diferenciadores de tipos textuales, y será necesario comprobar si esta diferenciación se reproduce en hablantes no nativos.

En cuanto a la variable hablante, ésta no parece influir a nivel global en la frecuencia de conectores, si bien queda por comprobar si existen preferencias individuales por conectores determinados o si existe relación entre hablante y frecuencia de conectores en determinados tipos textuales; la reducida base de datos ha aconsejado no realizar los cálculos correspondientes con los datos usados.

La frecuencia de conectores causales en un texto hablado sigue por tanto determinadas características que dibujan un perfil de uso propio de un hablante nativo: un uso frecuente de *weil* situado en un 0,4 por ciento frente a un uso mucho más reducido de otros conectores, y un uso de conectores más intensivo en unos tipos textuales que en otros. A partir de estos dos datos habrá que comprobar si hablantes aprendices de alemán como LE obtienen valores similares o divergentes. Si el uso de *weil*, por ejemplo, es significativamente mayor o menor en los aprendices, este conector causal sería significativo para la determinación de estadios de aprendizaje, a semejanza del uso de conectores como *und*, cuyo sobreuso se relaciona con un nivel inicial de aprendizaje del alemán como LE (Strunk, 1999).

REFERENCIAS BIBLIOGRÁFICAS

Fernández-Villanueva, Marta; Strunk, Oliver (2009): Das Korpus Varkom - Variation und Kommunikation in der gesprochenen Sprache. En: Deutsch als Fremdsprache 46 (2009) 2, S. 67-73

Frohning, Dagmar (2005) „Kausalmarker zwischen Pragmatik und Kognition. Korpusbasierte Funktionsprofile und Analysen zur Variation im Deutschen. Phil. Diss., Univ. Freiburg."

Frohning, Dagmar (2005). Das universelle weil: korpusbasierte Evidenzen. En: Blühdorn, Hardarik/Breindl, Eva/Waßner, Ulrich Hermann (Hg.): *Text — verstehen. Grammatik*

und darüber hinaus, (365-367). Berlin/New York: de Gruyter. (= Jahrbuch des IDS 2005).

Gohl, Christine (2000). Zwischen Kausalität und Konditionalität: Begründende wenn-Konstruktionen. Konstanz: Tausch. (= Interaction and linguistic structures 24). *Deutsche Sprache* 3/2002. 193-219.

Günthner, Susanne (2002): Zum kausalen und konzessiven Gebrauch des Konnektors wo im gesprochenen Umgangsdeutsch. *Zeitschrift für germanistische Linguistik* 30/3, 310–341.

Häcker, Martina (1994). The death of English for and German denn: linguistic change in progress. *Grazer linguistische Studien* 42, 29-35.

Linke, Angelika; Nussbaumer, Markus; Portmann, Paul R. (2001). *Studienbuch Linguistik*. Tübinegen: Niemeyer.

Pasch, Renate; Brauße, Ursula; Breindl, Eva; Waßner, Ulrich Hermann (2003). *Handbuch der deutschen Konnektoren*. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln), Berlin/New York: de Gruyter

Peldszus, Andreas; Herzog, André; Hofmann, Florian; Stede, Manfred (2008): Zur Annotation von kausalen Verknüpfungen in Texten. En: *Proceedings der Konferenz zur Verarbeitung natürlicher Sprache* (KONVENS), Ergänzungsband. Berlin.

Strunk, Oliver (1999): *Parameter der akademischen Textproduktion*. Tesis doctoral, Universidad de Barcelona.

Linguistic features in a nanotechnology corpus

KEITH STUART

ANA BOTELLA

Universidad Politécnica de Valencia

Abstract

This paper describes the process followed so as to extract and explore lingüístic features in academic discourse. In particular, we present the methodology adopted to analyse a selected set of linguistic features in a corpus of nanotechnology articles using a quantitative approach. Our initial hypothesis is that discourse communities share linguistic conventions although they also show their own particularities. The methodology section presents how data was collected and prepared to carry out the analysis. We finally present empirical results about language use in our Nano-Corpus by comparing the data with former studies. Significant results are examined to draw qualitative conclusions about variations encountered within the same genre. In this sense, corpus linguistcs techniques combined with computer software open up numerous possibilities to explore the linguistic features that make up the structure of nanotechnology texts.

Keywords: corpus linguistics, academic discourse, nanotechnology corpus, lexico-grammatical variation.

Resumen

En este artículo se describe el proceso que se ha seguido para extraer y analizar rasgos lingüísticos utilizados en el discurso académico. En concreto, presentamos las metodología que se adoptó para analizar determinados rasgos lingüísticos en un corpus de artículos de nanotecnología desde una perspectiva cuantitativa. Nuestra hypótesis de partida es que las comunidades discursivas comparten convenciones lingüísticas, aunque también poseen sus rasgos particulares. En la sección de metodología se explica cómo se recopiló y se preparó la información para llevar a cabo el análisis. Finalmente, presentamos algunos datos empíricos sobre el uso del lenguaje en el Nano-Corpus mediante la comparación con anteriores estudios. Se revisan resultados relevantes para establecer conclusiones cualitativas sobre las variaciones detectadas en un mismo género discursivo. En este sentido, las técnicas de la lingüística de corpus junto con el uso de programas informáticos abren numerosas posibilidades en el estudio de la variación léxico-gramatical que conforma la estructuraa de los artículos de nanotecnología.

Palabras clave: lingüística de corpus, discurso académico, corpus de nanotecnología,, variación léxico-gramatical

1. INTRODUCTION

In this article we describe the process followed to extract lexical and grammatical information from a corpus of academic articles. In particular, we present the analysis that has been carried out to gather linguistic information from a corpus of nanotechnology articles. The study involves corpus based techniques combined with quantitative analyses using computer programmes. Our main research purpose in this study is to report on the lexico-grammatical features of nanotechnology articles. There has as of yet been no linguistic

description of articles from this discipline. We compare the linguistic features of these articles with those found in Biber's study of Academic Prose (1995).

The structure, grammar, and vocabulary of written texts vary depending on why we are writing, who we are writing for, and what we are writing about. We refer to these predictable patterns in written language as genres. In this study, we analyse an academic genre, the article. Genres transcend the limits of disciplines and usually share features of different disciplines but are also susceptible to disciplinary variations. Genres are not static; on the contrary, they are dynamic and respond to novel social demands and communicative purposes (in this study, the communicative purpose of our articles is to transmit knowledge about the relatively new discipline of nanotechnology. The relationship between genre (social activity) and text is illustrated below:

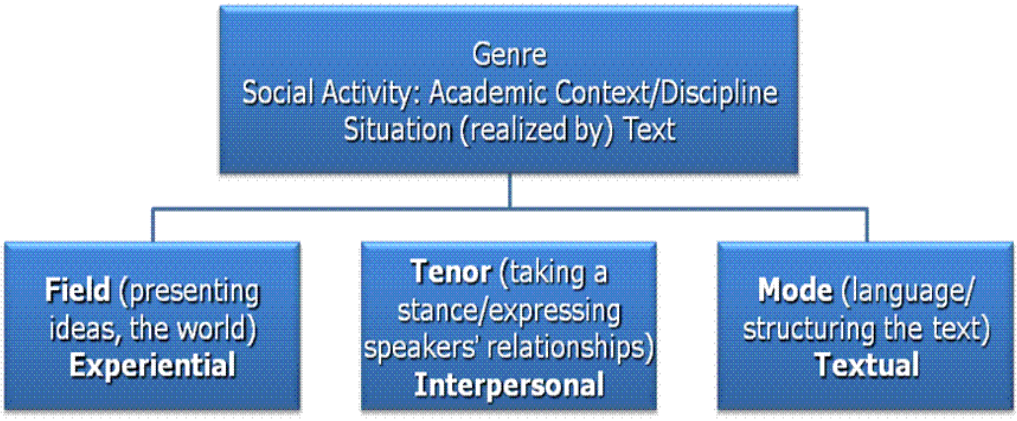


Figure 1: Relationship between social activity and text

A writer always simultaneously writes about something (field) – that is, the ideas and world of nanotechnology (ideational metafunction); enacts an interpersonal relationship with the reader - that is, social relationships are established and maintained (interpersonal metafunction); and creates a textual context for the information to be presented as a message - that is, a text is being structured (textual metafunction). These choices are in turn realized by lexico-grammatical choices. Typical experiential choices are noun phrases, nominal groups (participants); verbs (material, behavioural, mental, verbal, relational processes); prepositional phrases, adverbial adjuncts, and other resources for expressing circumstance (time, place, manner). Typical interpersonal choices are mood (statements, questions, demands); modality (modal verbs, and adverbs) and other resources for evaluative and attitudinal meaning. Typical textual choices are cohesive devices (conjunctions, connectors,

etc.); clause-combining strategies and thematic organization. In this study, we concentrate on the lexico-grammatical choices that realize the metafunctions that configure the most important academic genre, the article.

Our research analyses 370 nanotechnology articles (from now on referred to as Nano-Corpus) using various software tools to be able to quickly and efficiently gather information about the structure of these texts, lexical clusters, collocations, and grammatical features. The total number of running words in the texts is 1,185,407. Having described the methodology used for extracting and processing linguistic data using quantitative procedures, we then suggest how this statistical data may be compared and contrasted with previous linguistic research, especially Biber's study of Academic Prose (1995). Then, from a qualitative perspective we will explore some relevant salient features in our Nano-Corpus.

2. METHODOLOGY

The first stage in our study involved corpus compilation and design. At this stage, questions referring to corpus purpose, size, representativeness, target audience, etc., were addressed (Torruella & Llisterri, 1999: 45; Hunston, 2002: 26; Biber et al., 1998: 246). Moreover, decisions were taken with regards to corpus design (tagging and annotation). As for the Nano-Corpus, it comprises whole texts, taken form leading journals with a high impact factor (all articles have been written between 2006 and 2009). Table 1 below shows journal names as well as their impact factor and the number of texts collected from each of them. A portion of our corpus includes articles which have been written by researchers and lecturers from our Institution (Universidad Politécnica de Valencia).

Table 1: Journal Names/ Distribution of texts in Nano-Corpus

Journal Title	ISSN	JCR Data			
		Total Cites	Impact Factor	5-Year Impact Factor	Number of Articles Nano-Corpus
IEEE Transactions on Nanotechnology	1536-125x	1098	2.110	2.814	50
Journal of Micromechanics & Microengineering	0960-1317	4672	1.930	2.495	20
Journal of Nanoparticle Research	1388-0764	1277	2.338	2.758	25
Nano Letters	1530-6984	26246	9.627	11.048	60
Nano Today	1748-0132	114	5.929	6.357	20
Nanoscale Research Letters	1931-7573	62	2.158	2.158	20
Nanotechnology	0957-4484	10839	3.310	3.511	50
Nature Nanotechnology	1748-3387	688	14.917	14.958	20
Scripta Materalia	1359-6462	11290	2.481	2.739	30
Small	1613-6810	2696	6.408	6.419	25
UPV articles					50
				TOTAL	**370**

Once these crucial aspects were determined, the next step in the process consisted in saving the texts as plain text documents so as to have them in the format required to carry out subsequent tasks. Tags were then assigned manually to different elements of the document (see Table 2 below). Then, the corpus was transferred to a Microsoft Access database with a Visual Basic Application, which will allow us to retrieve relevant information from the articles easily at macroscopic level. The tags allow us to extract the different section headings and, therefore, we are able to examine the global structure of nanotechnology articles.

Table 2: Tags used for main structural features of the articles in Nano-Corpus

DOCUMENT ELEMENTS	TAG
Journal Name/	<n> , </n>
Author Name	<a> ,
Article Title	<tit> , </tit>
Article Sections	<s> , </s>
Author Institution	,
Publication Date	<pub> , </pub>

Then, the Nano-Corpus was uploaded to Wmatrix (Rayson, 2003), a web based computer tool which allowed us have our texts automatically part-of-speech (POS) tagged using a Tag Wizard. Next, a POS frequency list was produced using the same software. Wmatrix uses the CLAWS tagger (Garside & Smith, 1997). It is at this point that we had our data ready to be analysed.

In order to carry out the analysis, Wmatrix was used in parallel with Wordsmith Tools. The reason for adopting this method is that for certain features a POS list was generated at a first stage, to then process the list obtained with the different instances of the grammatical category. To do this, we used the utility Concord in Wordsmith, which offers the user the possibility of batch processing, thus, generating concordance lines for a list of realisations of a selected feature.

To achieve our goal, we followed the model implemented by Biber et al. (1995, 1998), who designs a typology of linguistic features (a total of 65) to extract and analyse lexico-grammatical variations across various genres (15 genres). He adopts a Multi-Dimensional Approach (MDA) to linguistic variation to describe textual relations among spoken and written genres. Each dimension represents a set of linguistic features that co-occur in texts.

Biber's model has served as the basis for research in this paper. To develop our comparative study, we calculated the mean frequency of a number of selected linguistic features. It was carried out by normalizing the frequency counts of linguistic features in the

Nano-Corpus to a text length of 1,000 words. According to this procedure, the mean frequency of amplifiers in the Nano-Corpus was calculated as follows:

$$2{,}309 / 1{,}185{,}407 \times 1{,}000 = 1.44$$

There are a total of 2,309 amplifiers in our corpus. We divided this by the total running words in the corpus (1,185,407) and then multiplied by 1,000.

Corpus-based together with corpus-driven techniques were applied to achieve our research purposes.

3. RESULTS

In previous stages in the study, we discovered how words used in scientific terminology dependent on a specialized field of knowledge, generally display low frequency statistics in the normal discourse of general English. These specialized terms help to define the communities that use them in the same way as these communities define their terms. The information compiled in the different stages of the research has made use of the notions of word frequency, keywords and lexico-grammatical relations, that is to say, the lexico-grammatical phenomena of collocation, semantic prosody and colligation. Similarly, basing ourselves on statistical relevance, we have evaluated frequency counts for 24 linguistic features in our Nano Corpus. The results obtained have revealed substantial information relating linguistic co-occurrences within the same genre.

Besides the intratextual study realized, certain intertextual aspects have been considered that have allowed us to detect variations which are produced within the same genre: academic prose. The table below shows total frequencies and mean frequencies for the Nano-Corpus as well as mean frequencies for these features in Academic Prose.

Table 3: Linguistic features in the Corpus of Nanotechnology articles and in Biber's study of Academic Prose

Linguistic Feature	Total Frequency (Nano-Corpus)	Mean Frequency (Nano-Corpus)	Mean Frequency (Academic Prose)
Amplifiers	2,309	1.94	1.4
Analytic negation	2,365	1.99	4.3
Causative clauses	552	0.46	0.3
Concessive clauses	521	0.43	0.5
Conditional clauses	556	0.46	2.1
Downtoners	2,010	1.69	2.5
Emphatics	3,344	2.82	3.6
Existential 'there'	736	0,62	1.8
First person pronouns	4,540	3,8	5.7
Hedges	354	0.3	0.2
Indefinte Pronouns	484	0.4	0.2
Necessity modals	919	0.77	2.2
Nominalisation	56,147	47.365	35.8
Possibility modals	6,449	5.4	5.6
Predictive modals	1,586	1,33	3.7
Prepositions	158,861	134.01	139.5
Private verbs	3,078	2.59	12.5
Pronoun IT	3,115	2.6	5.9
Public verbs	707	0.59	5.7
Second person pronouns	23	0.01	0.2
Seem / appear	552	0.46	1
Suasive verbs	307	0.25	4.0
Synthetic negation	1,021	0.86	1.3
Third person pronoun	3,745	3.15	11.5
Type/token ratio	-	37.88standardised TTR	50.6
Word length	-	5.26	4.8

The first question we would like to address in our review of the results obtained is that of the use of downtoners. Following Biber's remarks, we have confirmed in our corpus that downtoners are used in academic articles to establish comparisons between items as well as their use to express interpersonal meanings. Chafe (1985) states that they indicate probability and reliability. Moreover, they are a way of expressing uncertainty towards a proposition.

We find 'slightly' as the most frequent collocate of 'different'. To provide evidence of this particular usage, concordance lines with the pattern 'downtoner+different' in our corpus are presented below.

- This task can be implemented using the CMOL DSP fabric with an almost similar pixel structure with a just **slightly different** data flow
- The elementary cell is **slightly different** from those currently used in literature
- The three search methods provide **slightly different** results
- The three search methods show **slightly different** results in the high impact technology fields
- It should be noted that **slightly different** mask
- The minima and maxima at the interface experience **slightly different** field strengths

Likewise, hedging in academic articles is a pervasive feature, which can be understood as indicators of politeness or deference towards the addressee. This communicative purpose is also achieved by means of perception verbs such as 'seem' and 'appear'.

Next in our analysis we would like to review the use of first person pronouns. The results showed that, 'we' is the most frequent pronoun, with a wide range of referents that might go from groups of people to only the writer. This leads us to confirm Biber's Dimension 1, which deals with Involved versus Informational Production (Biber, 1995: 122). From this perspective, conversation is characterised as highly interactive and not edited, whereas academic prose is highly edited and not interactive. It should be also noted that mean frequencies for third person pronouns in both corpora are substantially different, being higher in Biber's corpus (3.15 vs. 11.5). Again, this reveals the low degree of interactive language in specialised discourse. Other outstanding salient features in the study for this dimension are private and public verbs, being both of them much lower in the Nano-Corpus than in academic prose. Specialized genre reveals itself as non-affective and relatively free from interpersonal content.

Other relevant aspects relating to this multidimensional approach to linguistic features is the use of verb tenses (Dimension 2: Narrative vs. Non-narrative). Present indicative tense appears to be the most widespread in scientific texts. This is due to the writer's intention to present expository and procedural information. Researchers use academic articles to describe current findings or current state of affairs.

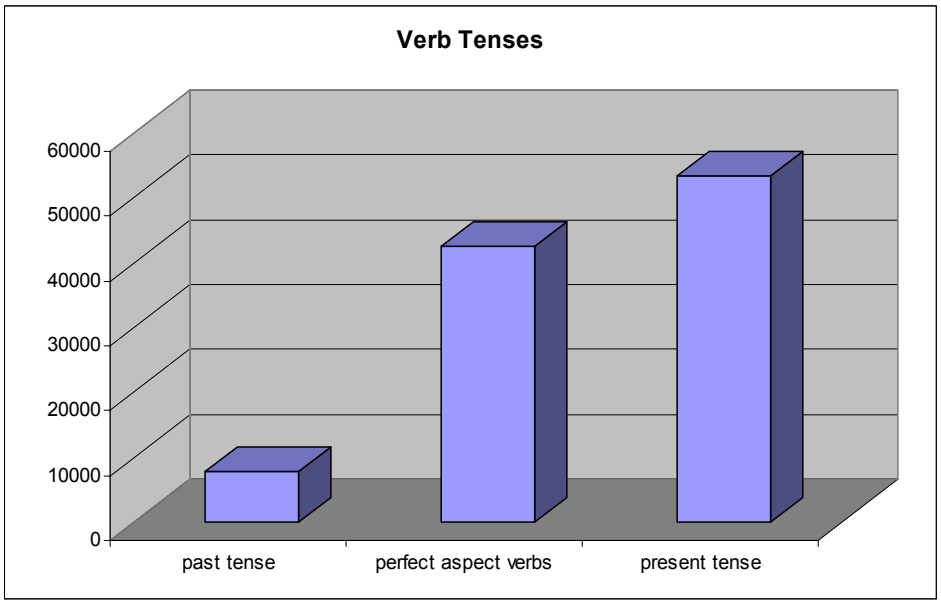


Figure 2: Verb Tenses in the Nano-Corpus

Dimension 4 is presented in terms of persuasive use of discourse. Linguistic features in this group include predictive modals, necessity modals, possibility modals, conditional clauses, suasive verbs, infinitives and split auxiliaries. They refer to the speaker's/writer's persuasion, assessment of likelihood or advisability. After having examined these features in both corpora, it was found that there is no significant variation in the use of possibility modals. This responds to the writer's intention to assess what he thinks it is likely to occur when considering different perspectives to the issue being researched. Conditional clauses and predictive verbs are less frequent in nanotechnology texts, as it is not the author's purpose to predict what will or will not happen but to report on his findings.

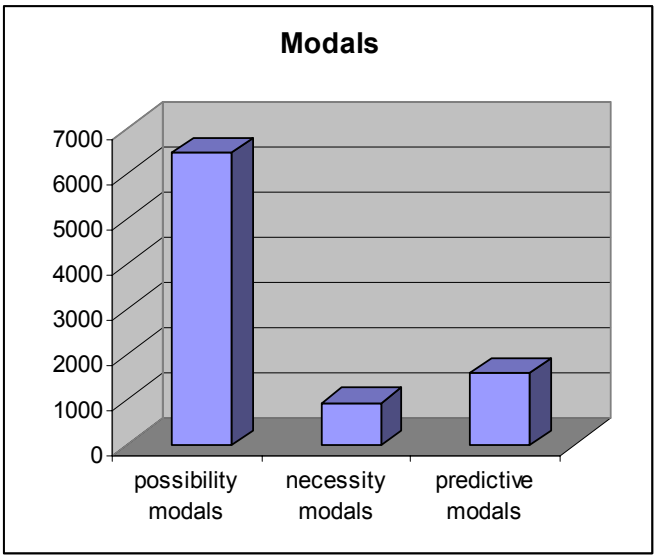


Figure 3: Modal verbs in the Nano-Corpus

Suasive verbs are generally used to mark an intention either to persuade or to convince the reader. In our case, it might be stated that the sub-genre of science and technology shows a lower persuasive communicative function than the super-genre of academic discourse.

As for Dimension 5 (Abstract versus Non-abstract), academic articles are highly abstract. This issue is worth mentioning with regards to the Nano-Corpus. Mean frequencies of nominalisations are found to be higher in our corpus (47.36) than in that of academic prose (35.8). Academic discourse is characterised by "the production of highly informational discourse" (Biber, 1995: 233). Nominalisations are a means of packing information in a dense manner (as in the Hallidayan sense of lexical density of written text). Nominalisations were extracted from our corpus for all words ending in –tion, -ment, -ness and –ity and their

plurals. The examples below illustrate the use of abstract, conceptual lexis in a specialised corpus which helps to give the impression of more objectiveness and precision.

- However this can be of importance in the **development** of new functional biomimetic
- reversible mass **movement** and controlled molecular transport on the nanometer level is a
- can also be switched by simple **adjustment** of the pH
- the use of suitable building blocks allows the **development** of functional
- the other is the **displacement** protocol which also involves the use of a binding site
- in a **displacement** reaction the binding site coordinates to the guest
- Nanostructured materials are attracting an increasing interest and even **excitement**
- we have investigated how an **improvement** in the ZnO nanocrystals
- Possible **assignment** for the different bands is also given
- after a surface silylation **treatment** was used as support for the preparation
- the sample was kept under protective atmosphere after the reduction **treatment** for used

The reason for the higher score of this linguistic feature can be attributed to changes and variations in language use within the same genre (Biber's corpus is about 40 years older than ours). Our study has been developed for an emerging technology, which shows how genres adapt to specific, exact and novel ways of conveying knowledge. Examples of nominalisations in our corpus reveal the specificity and informational density of the language of science and technology.

Table 4: Examples of nominalisations in the Nano-Corpus

- complementarity	- dimensionality	- smoothness
- functionality	- hydrophobicity	- translocation
- applicability	- polarity	- exploration
- affinity	- darkness	- photodimerization
- mesoporosity	- effectiveness	- irradiation
- intensity	- lightness	- titration
- crystallinity	- usefulness	- protonation
- responsiveness	- thickness	- transformation
- robustness	- fastness	- calcination
- randomness	- roughness	- deconvolution
- stiffness	- hardness	- functionalization
- correctness	- sharpness	- calibration

Likewise, long nominal groups and careful selection of vocabulary realize the highly ideational nature of this kind of discourse where the packaging of ideas is foremost.

Table 5: Examples of noun groups in the Nano-Corpus

– molecular entities	– anion-induced polarity modulations
– functionalized nanosized cavities	– polarity-induced colour modulation
– advanced organic molecular cavities	– carbon nanotubes layer thickness
– feed conductivities	– back-gate oxide thickness
– hydrogen-carbon possibilities	– attractive robustness properties
– room temperature sensitivities	– light-induced switching transformation
– enhanced chemical compatibility	– aqueous dendrime-containing solution
– thermal conductivity detector	– citrate calibration curves
– size-and polarity-controlled gate-like scaffolding	– current density deposition time

The approach and examples presented in this section have provided us with empirical data about the use of a selection of linguistic features from a corpus of specialised texts. This type of analysis could be further extended to other linguistic co-ocurrences within the Nano-Corpus.

4. CONCLUSION

Nowadays, the availability of academic articles in electronic format and the existence of readily available computer software offer the linguist numerous possibilities to analyse linguistic variation. Such a study requires processing large amounts of texts in order to obtain representative, significant empirical evidence about lexical and grammatical co-occurrence. Corpus-based together with corpus-driven techniques have provided the methodological basis for our linguistic research.

Texts are seen in our analysis as multidimensional constructs which meet the communicative requirements of a discourse community. They are dynamic and in constant change as they are influenced by disciplinary knowledge. As new research areas emerge, linguistic patterns adapt themselves and find new ways of conveying emerging knowledge. Our main research purpose in this study has been to report on the lexico-grammatical features of nanotechnology articles. There has as of yet been no linguistic description of articles from this emerging discipline which will need further work including information about the social context of the production of nanotechnology articles as well as greater linguistic detail about the schematic structure of nanotechnology texts (in what order and how are the ideas organized in the text?); intertextual analysis (linguistic features of anything drawn from other texts, how information is attributed to sources?, what kind of shared knowledge is expected

of readers?); and, finally, further linguistic analysis of lexico-grammatical features for realizing the metafunctions of language: experiential, interpersonal and textual meanings which configure the discourse structure of nanotechnology texts.

REFERENCES

Biber, D.(1995). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics*. Cambridge: Cambridge University Press.

Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. In D. Olson, N. Torrance and A. Hildyard (Eds.), *Literature, Language and Learning: The Nature and Consequences of Reading and Writing*, (pp. 105-123). Cambridge: Cambridge University Press.

Garside, R., and Smith, N. (1997) A hybrid grammatical tagger: CLAWS4. In Garside, R., Leech, G., and McEnery, A. (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora,* (pp. 102-121). Longman, London,.

Hunston, S.and Francis, G. (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. John Benjamins: Amsterdam.

Rayson, P. (2003). Matrix: *A statistical method and software tool for linguistic analysis through corpus comparison*. *Ph.D. thesis*, Lancaster University.

Scott, M. (2004). *WordSmith Tools version 4*. Oxford: Oxford University Press.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Torruella, J. and Llisterri, J. (1999). Diseño de corpus textuales y orales. Filología e informática. *Nuevas tecnologías en los estudios filológicos*. (pp. 45-77). In J. Blecua, G. Clavería, C. Sánchez and J. Torruella (Eds.), Barcelona: Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona - Editorial Milenio.

The concept of 'circumcollocate' and its significance for lexicography: A discussion with particular reference to the Japanese language

TADAHARU TANOMURA

Osaka University

Abstract

The concept of 'circumcollocate' will be proposed and defined as a habitually occurring discontinuous word sequences which surrounds, i.e. both precedes and follows, a given expression. Several kinds of instances of circumcollocates in the Japanese language obtained though corpus analysis will be presented, and relevant issues will be discussed. It will be shown that the concept of circumcollocate will be of lexicographic significance at least as far as Japanese is concerned. The corpus of Japanese used here will be a huge Web corpus constructed by the author in 2008, consisting of about 45 billion words, which amounts to 150 Giga bytes in file size.

Keywords: collocation, circumcollocate, lexicography, Web corpus, the Japanese language

Resumen

El concepto de "circumcollocate" será propuesto y definido como secuencias discontinuas habituales de palabras que flanquean a una expresión dada, tanto precediéndola como siguiéndola. Se presentarán distintos tipos de ejemplos de "circumcollocates" en japonés obtenidos de nuestro corpus, y se tratará también de algunos temas relacionados. Se mostrará que el concepto de "circumcollocate" es significativo desde el punto de vista lexicográfico, al menos en cuanto a la lengua japonesa se refiere. El corpus que hemos usado, creado para Internet por nosotros en 2008, consiste en 45 mil millones de palabras y es un archivo de 150 GB.

Palabras clave: colocación, "circumcollocate", lexicografía, Corpus para Internet, lengua japonesa

1. THE CONCEPT OF 'CIRCUMCOLLATE'

The concept of collocation has typically been viewed as a binary relation between a pair of co-occurring words, such as a relation between a particular verb and the nouns it habitually takes as its object, or a relation between a particular adverb and the adjectives it is habitually used to modify.

In this presentation, we will argue that there is a need to identify a complex type of collocate, to be termed 'circumcollocate', which will be of significance from a lexicographic point of view. A circumcollocate is defined as a habitually occurring discontinuous word sequences which surrounds, i.e. both precedes and follows, a given expression. If we call, for convenience, a collocate in the preceding context of a given expression a 'precollocate', and a collocate in the following context of a given expression a 'postcollocate', then a circumcollocate may be defined as a habitual combination of a precollocate and a postcollocate, as we may represent schematically as follows:

circumcollocate = precollocate … postcollocate.

As probably may be surmised, the term ‘circumcollocate’ was coined after ‘circumfix’, which denotes a kind of affix which is placed around the stem, as may be exemplified by the German past participle affix *ge-…-t* as in *ge-lieb-t* (‘loved’).

Before we turn to the observation of concrete cases of circumcollocates in the Japanese language, a few preliminary remarks will be in order. (A discussion of issues related to the computational treatment of Japanese texts may be found in Tanomura (2009).)

2. PRELIMINARY REMARKS

2.1. Japanese grammar

First we will briefly sketch the grammar of Japanese, so that the main discussion later on may be understood better by those who are not familiar with the language.

Japanese is a consistent SOV, head final language, as may be exemplified by the following simple sentence.

(1) *watasi-wa ongaku-o aisuru.*
I-TOPIC music-ACC love
'I love music.'

Another important characteristic of Japanese grammar is its agglutinative morphology. Grammatical elements are appended to nouns and verbs as shown in (2).

(2) a. *watasi-ga* ('I-NOM')
watasi-o ('I-ACC')
watasi-ni ('I-OBL', ‘to me’)
watasi-ni-mo ('I-OBL-also’, ‘also to me’)
watasi-ni-sae ('I-OBL-even’, ‘even to me’)

b. *tabe-ru* ('eat-PRES', ’eat’)
tabe-ta ('eat-PAST', ’ate’)
tabe-nai ('eat-NEG', ’do not eat’)
tabe-rare-ru ('eat-PASS-PRES', ’be eaten’)
tabe-rare-ta ('eat-PASS-PAST', ’was(were) eaten’)
tabe-rare-nai ('eat-PASS-NEG', ’be not eaten’)
tabe-sase-ru ('eat-CAUS-PRES', ’make sb eat’)
tabe-sase-rare-ta ('eat-CAUS-PASS-PAST', ’was(were) made to eat’)

2.2. The corpus to be used

A large amount of linguistic data is required for collocational analysis. In this study, a huge Web corpus constructed by the author in 2008 will be used. It is a collection of some ten million Web pages, and consists of about 75 billion characters. This amounts to about 45 billion words, or 150 Giga bytes in file size.

The Web corpus was constructed by the following procedure of five steps.

(3) i) Make a large list of (sets of) words

ii) Search with Yahoo! using those (sets of) words as keywords

iii) Acquire the first hundred URLs in each search result

iv) Acquire the documents referred to by the URLs

v) Eliminate HTML tags, etc. from the documents

The nature of the acquired documents may differ depending on the keywords given to the search engine in the step (3ii). Basically the keywords were prepared by means of a mechanical segmentation of various types of Japanese texts so that the keywords would be unbiased as a whole.

Although a number of other minor problems which had to be coped with were encountered in the actual processing, they will not be mentioned here.

3. Examples of circumcollocates in Japanese

Now we will see the lexicographic significance and usefulness of the concept of circumcollocate by observing several kinds of examples.

3.1. Circumcollocates of verbs

Let us take an intransitive verb *tukiru* as the first example. It denotes 'run out, be exhausted', and may be morphologically decomposed into *tuki-ru*, where *tuki-* is the stem and *–ru* is a non-past suffix.

Through analysis of the Web corpus, we may obtain, as nouns which frequently appear as the subject of this verb, such nouns as *tikara* (‘power’), *bansaku* (‘every mean’), *aisoo* (‘patience’), *kyoomi* (‘interest’), *gimon* (‘doubt’), *nayami* (‘worry’). But importantly these nouns fall under two groups depending on the overall pattern of expression in which they co-occur with the verb *tukiru.*

(4) a. {*tikara/bansaku/aisoo-ga*} *tuki-ta*
{power/every mean/patience-NOM} be exhausted-PAST

'{power/every mean/patience} was exhausted'

b. {*kyoomi-ga/gimon-wa/nayami-wa*} *tuki-nai*
{interest-NOM/doubt-TOPIC/worry-TOPIC} be exhausted-NEG
'{interests/doubts/worries} will never be exhausted'

Note that in the first group of nouns shown in (4a), they co-occur with the affirmative, past tense form of the verb *tukiru*. On the other hand, in the second group of nouns shown in (4b), they co-occur with the negative, non-past tense form of the verb. The nouns in question might sometimes appear in a different configuration as well, but there is an overwhelming tendency for them to be used in ways indicated above.

So it does not suffice to simply say that the verb *tukiru* frequently follows these nouns, or for that matter that it sometimes precedes the suffixes of past tense or negation. We need to pay attention to the combination of the noun as a precollocate and the verbal suffix as a postcollocate. Our proposal is that we identify discontinuous word (or morpheme) sequences such as *tikara…ta* ('power…PAST') and *kyoomi-ga…nai* ('interest-NOM…NEG') as circumcollocates of the verb *tukiru*. Such information concerning circumcollocates of a given expression in a dictionary of the Japanese language, or a collocational dictionary in particular, will be no less useful than information concerning precollocates and postcollocates.

The second example we will look at is a transitive verb *yurusu*. It denotes 'permit, allow, forgive', and is analyzable as *yurus-u*, *-u* being an allomorph of the non-past suffix *–ru* we saw in the last example.

The nouns which frequently appear as the subject of this verb include *zikan* ('time'), *tairyoku* ('bodily powers, stamina'), *yosan* ('budget'), *zizyoo* ('circumstances, conditions'), *zyookyoo* ('state of affairs, situation'), *puraido* ('pride, self-respect'), *ryoosin* ('conscience'). But, as in the last example, the idiomaticity of their use depends on the elements which follow the verb.

(5) a. {*zikan/tairyoku*}*-no* *yurus-u* *kagiri*
{time/stamina}-NOM permit-NON-PAST limit
'as far as {time/one's stamina} permits'

b. {*zikan/yosan/zizyoo/zyookyoo*}*-ga* *yurus-eba*
{time/budget/circumstances/situation}-NOM permit-COND
'if {time/budget/circumstances/situation} permits'

c. {*puraido/ryoosin*}*-ga* *yurus-anai*
{pride/conscience}-NOM permit-NEG
'one's {pride/conscience} does not permit (him/her to do something)'

The three patterns in which *yurusu* occurs as shown in (5) denote 'as far as SUBJECT permits', 'if SUBJECT permits' and 'SUBJECT does not permit …' respectively, and the typical nouns which appear as the subject of *tukiru* depend on the pattern.

We may also observe a correlation between the object noun and the predicate form of *yurusu*. The nouns which are relevant here includes *yodan* ('prediction'), *dakyooo* ('compromise'), *tuizui* ('overtake, catch up'), *husei* ('injustice'), *dokusoo* ('runaway lead'), *gyakuten* ('reversal (of the situation in game or battle)').

(6) a. {*yodan/dakyoo/tuizui*}*-o* *yurus-anai*
{prediction/compromise/overtake}-ACC permit-NEG
'does not allow {prediction/compromise/overtake}'

b. {*husei/dokusoo/gyakuten*}*-o* *yurus-ita*
{injustice/runaway lead/reversal}-ACC permit-PAST
'permitted (=failed to prevent) {injustice/runaway lead/reversal}'

These two patterns denote '… does not allow OBJECT' and '… failed to prevent (the occurrence of) OBJECT' respectively.

Thus it will be helpful to provide, in a dictionary entry for the verb *yurusu*, information about circumcollocates of the verb such as *zikan-ga…kagiri* ('time-NOM…limit'), *yosan-ga…eba* ('budget-NOM…COND'), *yodan-o…anai* ('prediction-ACC…NEG')

3.2. Circumcollocates of adjectives

Next let us see another example of circumcollocate using *ooki*, an adjective of an archaic style, denoting '(be) many'.

The adjective *ooki* is typically used in the phrase pattern *N1 ooki N2*, where *N1 ooki* is a relative clause modifying *N2*, as exemplified by the following.

(7) *koi ooki onna*
love be many woman
'woman with many love affairs, (lit.) woman with whom loves are many'

What is worth noting about this phrase pattern is that the habitual combination of *N1* and *N2* is rather restricted. The expressions which frequently occur in the corpus include the following.

(8) a. *koi ooki* {*onna/otoko/otome*}
love be many {woman/man/maiden}
'{woman/man/maiden} with many love affairs'

b. {*nayami/yume*} *ooki* {*tosigoro/zinsei/hibi*}
{worries/dreams} be many {age range (of a person)/life/days}

'{age range/life/days} with many {worries/dreams}'

The phrase pattern *N1 ooki N2* will not be used idiomatically when saying, e.g., 'man with a lot of money' or 'season with a lot of rain', in spite of the grammaticality of the expressions. Hence it may be of use to identify *koi...onna* ('love...woman'), *nayami...tosigoro* ('worries...age range'), etc. as circumcollocate of the adjective *ooki*, and to include such information in the entry for *ooki* in a collocational dictionary.

3.3. Circumcollocates of adverbs

Circumcollocates of adverbs will be exemplified with the adverb *massugu*, denoting 'straight'.

The patterns in which *massugu* habitually occurs include the following:

(9) a. {*kono miti/toori/syootengai/singoo*}*-o* *massugu* {*iku/susumu*}
{this road/street/shopping mall/traffic lights}-ACC straight {go/proceed}
'{go/proceed} straight along this road, etc.'

b. (…*no*) {*me/kao*}*-o* *massug*u {*miru/mitumeru*}
(…GEN) {eyes/face}-ACC straight {look/stare}
'{look/stare} at (sb's) {eyes/face} straight on'

c. {*sesuzi/asi/te*}*-o* *massugu*(*-ni*) {*nobasu/suru*}
{back/legs/arms}-ACC straight {stretch/make}
'stretch one's {back/legs/arms}'

Thus, discontinuous word sequences such as *kono miti-o…iku* ('this road-ACC…go'), (…*no*) *me-o…miru* ((…GEN) eyes-ACC…look'), *sesuzi-o…nobasu* ('back-ACC…stretch') can be counted as circumcollocates of the adverb *massugu*.

4. Conclusion

There would be no reason to restrict the concept of collocation to a relation between two words. Rather, collocation should be regarded as a relation between multiple words in general, especially if our goal is to make a collocational dictionary. What counts from a practical viewpoint of the dictionary user is the overall patterns in which a given expression habitually occur, rather than simple relations between the expression and a co-occurring single word. Thus, the inclusion of information concerning circumcollocates of dictionary entries will be an important maneuver to be taken in order to enhance the usefulness and value of the dictionary.

McIntosh *et al.* (2009), a very well compiled dictionary of English collocations, lists the pattern *in*…{*to/with*} as what we would call a circumcollocate of the noun *contrast*, but the number of such circumcollocates described therein is relatively few. One reason for this will be that this dictionary gives obvious priority to simple relations between a pair of words. But I suppose, in addition, that the extent to which the notion of circumcollocate has lexicographic significance may depend on the word order of the language, and hence differ from language to language. Trivial circumcollocates such as *lions...meat*, *horses...grass* and *cats...fish* for the verb *eat* will be found in a great quantity in English, but they will be of relatively less use in a dictionary. The actual significance of the concept of circumcollocate for English, Spanish and other languages will be an open question which needs to be examined and evaluated individually for those languages.

REFERENCES

McIntosh, C. Francis, B., & Poole, R. (eds.) (2009) *Oxford Collocations Dictionary for Students of English.* Second edition. Oxford: Oxford University Press.

Tanomura, T. (2009). Retrieving collocational information from Japanese corpora: An attempt towards the creation of a dictionary of collocations (in Japanese). *Handai Nihongo Kenkyuu*, 21, 21-41, Osaka: Osaka University.

Diátesis léxica de gehen y kommen en un corpus de lengua oral en alemán

EDUARD TAPIA YEPES

Universitat de Barcelona

Resumen

Este artículo presenta los datos, la orientación metodológica y los resultados de una clasificación de diátesis léxicas en los verbos dinámicos alemanes gehen y kommen, extraídos del corpus de lengua oral VARKOM. El análisis topológico de las relaciones sintáctico-semánticas en estos predicados ha permitido diferenciar perspectivas funcionales oracionales prototípicas agentivas de gehen y kommen (PROT: acciones/actividades con agente explícito en topic), de perspectivas no prototípicas o menos/no agentivas (NPROT1: acciones/actividades con sujeto no agentivo en focus y agente bloqueado implícito; NPROT2: estados/procesos sin agente).

Palabras clave: diátesis léxica, perspectiva funcional oracional, agente, sujeto gramatical, clase de predicado

Abstract

This paper presents study data, methodology and results of a lexical diathesis classification on the German dynamic verbs gehen and kommen, extracted from the oral language corpus VARKOM. The topological analysis of the syntactic-semantic predicate relations has allowed distinction between prototypical agentive use (PROT) of gehen and kommen (actions/activities with explicit agent as topic) and their non prototypical less agentive one (NPROT1: actions/activities with non agentive subject as focus, and blocked yet implicit agent; NPROT2: states/events with no agent).

Keywords: lexical diathesis, functional sentence perspective, agent, grammatical subject, predicate typology

1. INTRODUCCIÓN[1]

En el marco de un estudio que pretende explorar la relación entre el comportamiento de la alternancia en la perspectiva[2] funcional oracional y el contexto comunicativo, se detalla aquí una clasificación de diátesis léxicas realizadas con los verbos dinámicos alemanes *gehen* y *kommen*. El cambio diatético en un predicado implica una reordenación en su estructura sintáctico-semántica y, a menudo, un cambio en su tipología (estado, proceso, acción/actividad). Es por ello que los cambios diatéticos determinan la perspectiva funcional oracional.

[1] El corpus está elaborado por el grupo LADA (Lingüística Aplicada i Didàctica de l'Alemany) de la Universitat de Barcelona, en el marco del proyecto de investigación COHESTIL (FFI2008-01230/FILO), y financiado por el Ministerio de Ciencia e Innovación. Agradezco los comentarios y las sugerencias de Yurena Alcalá y de la Dra. Marta Fernández-Villanueva

[2] Una perspectiva que no tiene que ver aquí con el concepto pragmático y "específico" de perspectiva (Sandig, 1996: 38).

La realización de una perspectiva más o menos agentiva de un predicado depende de la intención de cada hablante y/o de la situación comunicativa en la que éste se encuentre, ya que el sistema no prevé una realización diatética concreta como forma básica o no marcada (cf. Brinker, 1971: 15; Eroms, 2000: 383). Uno de los factores del contexto comunicativo que pueden presentar alguna correlación con la alternancia en la perspectiva funcional oracional es el desarrollo temático textual: argumentación, descripción, exposición, instrucción o narración (cf. Heinemann y Viehweger, 1991; Fernández-Villanueva, 2002). Una narración se entiende como una estructuración cronológica de sucesos; una descripción, como una representación de las partes, características o circunstancias de seres animados, objetos o conceptos abstractos. Así pues, resulta plausible asociar la primera a perspectivas y predicados agentivos (acciones o actividades) y la segunda a predicados no agentivos (estados o procesos).

Este planteamiento se enmarca en la línea de trabajos anteriores (Taylor y Tversky, 1992 y 1996; Stutterheim, 1997; Tappe, 2001). El estudio proyectado se entiende como novedad dentro de este ámbito, en la medida en que supone una observación de perspectivas funcionales en cinco desarrollos temáticos textuales en lengua oral. El análisis de categorías, frecuencias y distribución de diátesis léxicas presentado aquí, supone el punto de partida para dicho estudio.

2. DATOS

Los datos analizados proceden del corpus de lengua oral VARKOM (Fernández-Villanueva y Strunk, 2009). Para este estudio se han escogido siete entrevistas de registro semiformal a mujeres de entre 25 y 30 años, realizadas en alemán como lengua materna. Cada entrevista consta de cinco tareas de elicitación experiencial (dialogada) y experimental (monologada). Las tareas se corresponden, a su vez, con los cinco desarrollos temáticos textuales: narración, descripción, instrucción, argumentación y exposición. A continuación se especifica la duración y el número de *tokens* de las entrevistas completas del corpus (v. tabla 1) y de las partes de elicitación monologada, de donde provienen los predicados analizados (v. tabla 2).

Tabla 1: Duración y *tokens* de las siete entrevistas completas: elicitación experiencial y experimental

	Duración (s.)	***Tokens***
Media	1849,1429	3232,4286
Mediana	1888,0000	3015,0000
Varianza	84008,810	686263,952
Desv. típ.	289,84273	828,41050
Mínimo	1422,00	1874,00
Máximo	2348,00	4167,00
Rango	926,00	2293,00
Amplitud intercuartil	325,00	1290,00
Asimetría	,399	-,385
Curtosis	1,011	-,490

Tabla 2: Duración y *tokens* de las partes monologadas de las entrevistas

	Duración (s.)	***Tokens* totales**
Media	1214,5714	1970,2857
Mediana	1167,0000	1843,0000
Varianza	66064,952	491962,905
Desv. típ.	257,03103	701,40067
Mínimo	898,00	1209,00
Máximo	1677,00	3319,00
Rango	779,00	2110,00
Amplitud intercuartil	376,00	817,00
Asimetría	,899	1,263
Curtosis	,841	1,818

En la parte monologada, los *tokens* verbales realizados presentan una distribución normal, como demuestra la prueba Kolmogorov-Smirnov (v. tabla 3).

Tabla 3: Test de normalidad para la distribución de *tokens* verbales en las tareas de elicitación monologada

Z de Kolmogorov-Smirnov	,578
Sig. asintót. (bilateral)	,892

3. METODOLOGÍA DE ANÁLISIS

3.1. Fase cuantitativa

Para asegurar una aparición suficientemente representativa de estas realizaciones, la fase cuantitativa ha tenido como objetivo determinar presencia y frecuencia de los lexemas susceptibles de presentar alternancia diatética, es decir, los *types* con mayor número de *tokens*. Esto ha requerido una lematización del corpus, en forma de transliteraciones de las producciones elicitadas mediante tareas, con *TreeTagger* y *Lexic Tools* (Schmid, 1994; Strunk, 2008). Gracias a este paso se han seleccionado los cuatro verbos más frecuentes (*gehen*, *kommen*, *stehen* y *sein*). Su agrupación parte del rasgo semántico dinámico/estático y de la clase de predicado que expresan a priori en su significado central: *gehen* y *kommen* ('ir' y 'venir'), verbos dinámicos o de movimiento (156 y 121 *tokens* respectivamente) y *stehen* y *sein* ('estar de pie' y 'hallarse en'), verbos estáticos o de situación (31 y 112 *tokens* respectivamente). La clasificación descrita aquí se centra únicamente en los verbos dinámicos.

3.2. Fase cualitativa: análisis de las relaciones sintáctico-semánticas

Para clasificar perspectivas funcionales oracionales en los predicados de *gehen* y *kommen* ya localizados, se diseñó un análisis sintáctico-semántico basado en la combinación de las siguientes variables: función sintáctica, topología oracional y rol semántico. Esto permite distinguir entre perspectivas funcionales más o menos agentivas. Una coincidencia entre el rol agente y el sujeto gramatical o complemento nominativo (NomE) en predicados dinámicos apuntaría claramente a acciones y actividades (usos prototípicos). Sin embargo, la ocurrencia de cualquier otro rol semántico con la misma función sintáctica convertiría un predicado (a priori) dinámico en un proceso o un estado (usos no prototípicos).

La tabla 4 es un ejemplo concreto del análisis sintáctico-semántico llevado a cabo para cada complemento realizado, previsto por la valencia del verbo o no.

Tabla 4: Detalle de análisis sintáctico-semántico para el complemento E1

clase de predicado	**E1**							
		categoría léxica	**supracat. léxica**	**realización morfosint.**	**función sintáctica**	**topología**	**rol sem.**	**subrol sem.**
acción	*wir*	Persona	ser vivo	grupo pronominal	compl. nominativo	*Vorfeld*[3]	agente	

[3] En oraciones enunciativas alemanas se considera *Vorfeld* toda información a la izquierda del verbo conjugado.

Esta disposición de los datos hace posible el cruce entre las variables mencionadas y facilita la observación de las diferentes realizaciones diatéticas de los predicados *gehen* y de los predicados *kommen*. Las conclusiones de esta observación conducen a una clasificación de perspectivas funcionales que se especifica en el apartado 4.

4. RESULTADO DE LA CLASIFICACIÓN

El análisis diseñado en el apartado 3 se ha efectuado para 277 *tokens* realizados de los verbos dinámicos *gehen* y *kommen*.

La clasificación parte de la observación de la (no) aparición del rol semántico agente como sujeto gramatical en la posición *topic*[4]. La variación en dicho esquema implica una alternancia diatética o reestructuración topológica de roles semánticos que determina perspectivas o visiones más o menos agentivas de *gehen* y *kommen*.

Del análisis de ocurrencias resultan dos grupos de perspectivas funcionales realizadas para estos dos predicados dinámicos: perspectivas prototípicas (PROT) y perspectivas no prototípicas (NPROT).

4.1. Verbos dinámicos en uso prototípico (PROT)

Según las frecuencias observadas, se puede hablar de un primer grupo consistente y bien representado de predicados: 119 *tokens* de *gehen* (76,28%) y 82 *tokens* de *kommen* (67,76%).

Todos estos predicados presentan una coincidencia entre el rol semántico agente y la función sintáctica de sujeto gramatical o complemento nominativo (NomE). Desde el punto de vista topológico, NomE aparece tendencialmente en la posición no marcada de *Vorfeld*. Sólo en un 28,85% de los casos PROT (58 *tokens* de 201) el sujeto aparece en *Mittelfeld*, es decir, después del verbo conjugado en oración simple enunciativa. Esto sucede únicamente en textos descriptivos e instructivos.

Los predicados prototípicos son semánticamente dinámicos. La fuerza motriz inherente al agente en la posición no marcada de *Vorfeld* favorece que estos predicados pertenezcan a la clase de predicado acción/actividad.

Estas características definen los usos PROT de *gehen* y *kommen* y permiten, además, establecer la diferencia entre perspectivas funcionales prototípicas y no prototípicas (v. 4.2.).

4.2. Verbos dinámicos en uso no prototípico (NPROT)

Existe otro grupo de predicados que aparece menos representado: 39 *tokens* de *gehen* (23,71%) y 37 *tokens* de *kommen* (32,23%). Por este motivo, los predicados correspondientes se han etiquetado como no prototípicos (NPROT). Dentro de este grupo se pueden diferenciar a su vez dos subgrupos NPROT: los predicados NPROT1 difieren solamente de un rasgo PROT (v. 4.1.) y los predicados NPROT2, de todos ellos. Aún así, el grupo NPROT presenta un rasgo transversal en todas sus realizaciones: la disociación de agente y NomE. A continuación se detallan las características de los dos subgrupos NPROT.

4.2.1. NPROT1

El subgrupo NPROT1 consta de 9 *tokens* en el caso de *gehen* y de 8 *tokens* en el caso de *kommen*, es decir, un 22,36% de los predicados NPROT y un 6,13% sobre el total de realizaciones (PROT+NPROT). Cabe destacar aquí que el contexto de aparición de NPROT1 es restringido: únicamente aparece en desarrollos temáticos descriptivos e instructivos.

Los predicados NPROT1 también expresan acciones/actividades dinámicas, como los predicados PROT. Sin embargo, el sujeto gramatical de los predicados en voz activa de NPROT1 deja de ser agente. Los significados de *gehen* y *kommen* que encajan en el esquema PROT1 aparecen detallados en la tabla 5. En ella se incluye también el rol semántico *no* agentivo realizado como NomE en cada caso.

Tabla 5: Frecuencias de los significados de NPROT1 según el rol semántico como NomE

Verbo	**Significado**	**NomE**		**Reparto (en *tokens* y % en NPROT1)**
		Características léxicas	**Rol semántico**	
gehen	'ir hacia', en el sentido de 'conducir hacia'	objeto concreto	instrumento	5 *tokens*, 29,41%
	'desde algún lugar se llega a otro'	concepto abstracto	tema[5]	4 *tokens*, 23,52%
kommen	'aparecer', 'venir después de', en el sentido de 'ocupar un lugar en una enumeración ordenada'	local/ concepto abstracto/ lugar geográfico	punto del recorrido	7 *tokens*, 41,17%
	'ir hacia', en el sentido de 'conducir hacia'	objeto concreto	instrumento	1 *token*, 5,88%

[4] En alemán, el sujeto gramatical se presenta tendencialmente como *topic* o información no marcada (de partida) y aparece en el *Vorfeld* (VF) o a la izquierda del verbo (cf. Weinrich, 2007: 64).

[5] Entendido aquí como rol semántico no afectado. Presenta similitudes con el caso "objective" de Fillmore (1968: 25) o el "patient" de Comrie (1993: 910).

No se habla aquí únicamente de la desaparición del agente como NomE, sino de su completo bloqueo (v. 1). Esto tiene como consecuencia que dicha entidad semántica pase a pertenecer al terreno de la información implícita o de segundo plano.

(1a) *da <geht> dann auch ne Treppe hoch* (DtMsKS10)
('y ahí también va/conduce una escalera hacia arriba')

(1b) * *da <geht> dann auch ne Treppe hoch* [von mir/dir/…]
(*'y ahí también va/conduce una escalera hacia arriba *por mí/ti/...*')

Sin el agente realizado, el dinamismo se intuye exclusivamente a través de otros roles semánticos que forman parte del campo semántico del dinamismo (v. 1a: *da* como locativo *topic* y *auch ne Treppe* como instrumento/medio de movimiento *focus*).

Como se observa en el ejemplo anterior (v. 1a), la no realización del agente provoca una reordenación[6] sintáctica/topológica de los demás roles semánticos. El movimiento sigue existiendo, aunque ahora se percibe a través de entidades no agentivas. Se habla en este caso de cambio en la perspectiva funcional de *gehen* y *kommen* respecto a sus usos prototípicos (cf. Eroms, 2000: 383).

De este modo, el rol semántico con función de sujeto gramatical en NPROT1, que en la mayoría de casos (70,58%, 12 de los 17 *tokens*) aparece marcado en la posición *focus* o *Mittelfeld*, es no animado o abstracto y no es fuente de movimiento en sí (v. tabla 5). Aún así, los predicados NPROT1 no pueden considerarse de situación o estáticos. El dinamismo sigue presente en ellos y no se interrumpe: los roles-sujeto son puntos, marcas o partes, tangibles o no, del camino recorrido/por recorrer, que se van haciendo visibles al caminante (agente) a medida que éste va avanzando (v. 2 y 3).

(2) *und danach <kommt> dann n Friseurladen und da musst du dann rechts* (DtMsKS13)
(‚y entonces *viene* una peluquería y ahí es donde tienes que girar a la derecha')

(3) *Da <geht> eine kleine Treppe runter oder hoch* (DtMsJdM11)
(‚Ahí una pequeña escalera *va/conduce* hacia abajo o hacia arriba')

A través de estos puntos de referencia se intuye la acción/actividad de *ir/venir*, llevada a cabo por un agente implícito. No es sino gracias a la existencia de este agente, que la

escalera (3) puede *ir* o bien *hacia* arriba, o bien *hacia abajo*. La dirección que toma la escalera viene siempre determinada por la entidad agentiva. Es la prueba de la existencia de un agente. Sin esta entidad asociada a un predicado claramente dinámico, la peluquería (2) no llegaría a aparecer nunca.

Queda demostrado que en el caso de NPROT1 se trata de las mismas acciones/actividades dinámicas de PROT, presentadas a través de una diátesis o conversión léxica menos agentiva. En efecto, si bien la perspectiva funcional oracional resultante de *gehen* y *kommen* en NPROT1 convierte a estos predicados en equivalentes pasivos[7] de sus variantes activas PROT, hay que apuntar que, desde el punto de vista formal, esto se expresa sin recurrir al mecanismo de la pasiva gramatical (en el caso de verbos intransitivos como *gehen* y *kommen*, mediante la pasiva impersonal: *es wird gegangen/gekommen*), sino con la misma forma verbal activa simple.

PROT y NPROT1 constituyen de este modo diátesis léxicas diferentes que describen la misma realidad desde perspectivas o puntos de vista diferentes. La característica de NPROT1 de compartir, aunque sintácticamente reestructurados, roles semánticos con PROT, constituye la diferencia más destacada entre NPROT1 y NPROT2.

4.2.2. NPROT2

El subgrupo NPROT2 está representado por 28 *tokens* en el caso de *gehen* y por 31 *tokens* en el de *kommen*, es decir, un 77,63% de los predicados NPROT y un 21,29% del total de realizaciones (PROT+NPROT). Si se observan los desarrollos temáticos textuales en los que aparecen usos NPROT2, se puede afirmar que los contextos son complementarios respecto a los de NPROT1, ya que NPROT2 aparece principalmente (83,05%) en textos argumentativos y expositivos (26 y 23 *tokens* respectivamente).

NPROT2 no coincide con ninguna de las características de PROT y sólo coincide con NPROT1 en la desaparición/bloqueo del agente. Los predicados NPROT2 pierden todo su dinamismo y se apartan así del campo semántico del movimiento. En ellos falta toda referencia a las entidades semánticas asociables al dinamismo (cf. Porzig, 1971: 125s.; Jackendoff, 1991: 25s.; Lyons, 1996: 264). No se habla tanto aquí de reordenación o reestructuración sintáctica/topológica de roles semánticos (v. 4.2.1.), sino de su reasignación. En la tabla 6 se detalla la lista de los nuevos roles semánticos que aparecen como NomE en

---

[6] Por ello, los roles semánticos implicados son también compartidos entre PROT y NPROT1: agente como caminante (implícito o no), locativo (destino), punto del recorrido o instrumento/medio de movimiento.

[7] No se entiende aquí "pasivo" desde el punto de vista gramatical o de paradigma verbal, sino desde el punto de vista de semántica interna del predicado.

los usos NPROT2, así como también los nuevos significados *no* dinámicos de los predicados correspondientes.

Tabla 6: Frecuencias de los significados de NPROT2 según el rol semántico como NomE

Verbo	**Significado**	**NomE**		**Reparto (en *tokens* y % en NPROT2)**
		Características léxicas	**Rol semántico**	
gehen	‚ser realizable o posible de alguna manera'	concepto abstracto	tema	11 *tokens*, 39,28%
	‚transcurrir o desarrollarse de alguna manera'	concepto abstracto	tema	6 *tokens*, 21,42%
	‚alguien toma algo como punto de partida de sus investigaciones, pensamientos o comportamiento'	ser animado	experientivo	4 *tokens*, 14,28%
	‚alguien se encuentra de alguna manera en una situación determinada'	ser animado	experientivo	2 *tokens*, 7,14%
	‚tratar a alguien de determinada manera'	ser animado	agente	2 *tokens*, 7,14%
	‚tratarse de alguien o de algo'	concepto abstracto	tema	1 *token*, 3,57%
	‚desaparecer'	concepto abstracto	privativo	1 *token*, 3,57%
	‚ser correcto'	concepto abstracto	tema	1 *token*, 3,57%
kommen	‚algo depende de algo'	concepto abstracto	tema/ causativo	10 *tokens*, 32,25%
	‚crear, engendrar algo'	ser animado	objeto afectado	4 *tokens*, 12,9 %
	‚conformarse, figurarse'	concepto abstracto	objeto efectuado	2 *tokens*, 6,45%
	‚alguien sale de algo o se mete en algo'	ser animado	experientivo	2 *tokens*, 6,45%
	‚llegar a saber; averiguar'	ser animado	experientivo	2 *tokens*, 6,45%
	'obtener (a cambio de algo)'	ser animado	benefactivo	2 *tokens*, 6,45%
	‚algo se crea mediante de algo'	concepto abstracto	objeto efectuado	1 *token*, 3,22%
	‚un proceso se desarrolla de alguna manera'	concepto abstracto	tema	1 *token*, 3,22%
	‚algo se extiende desde alguna parte'	concepto abstracto	objeto efectuado	1 *token*, 3,22%
	'nacer'	ser animado	objeto efectuado	1 *token*, 3,22%
	‚superar dificultades'	ser animado	benefactivo	1 *token*, 3,22%

El cambio absoluto de campo semántico implica, además, un cambio en la tipología de predicado, ya que la gran mayoría (89,83%) de predicados NPROT2 son procesos o estados.

Los predicados NPROT2 son producto de una abstracción o desplazamiento semántico respecto al uso considerado aquí como prototípico (PROT). La consiguiente reasignación de roles aparta a NPROT2 de una estructura semántica tendencialmente asociable a PROT. Si hay un rol semántico claramente asociable a NomE para NPROT2, éste sólo puede observarse en *gehen* (tema), en un 67,85% (19 *tokens*). En cambio, los usos NPROT2 de *kommen* presentan roles semánticos como causativo, objeto afectado, objeto efectuado, experientivo y benefactivo en la función de sujeto gramatical (v. tabla 6).

Las características de NPROT2 detalladas en este apartado no impiden afirmar que NPROT2 y NPROT1 conforman un grupo consistente (NPROT) en relación a PROT: todos los usos no prototípicos de *gehen* y *kommen* analizados experimentan una desaparición del rol semántico agente. Aunque por abstracción se sitúen en un campo semántico totalmente desplazado respecto a PROT, los predicados NPROT2 también pueden considerarse como usos no agentivos en voz activa de *gehen* y *kommen*.

5. OBSERVACIONES FINALES

Según las realizaciones observadas, se pueden diferenciar dos perspectivas funcionales oracionales: prototípicas y no prototípicas. La diferencia más relevante entre usos o perspectivas funcionales prototípicas y no prototípicas se observa a nivel de semántica oracional y tiene que ver con la (no) realización del rol agentivo como sujeto gramatical (NomE): todos los usos PROT presentan un agente en NomE; los usos NPROT *no* lo presentan. Esta característica convierte a los predicados NPROT en perspectivas funcionales, situadas o no en el mismo campo semántico de PROT, poco o en absoluto activas.

Compartir o no campo semántico con PROT ([±abstracción]) es lo que separa usos *dentro* de NPROT. El subgrupo NPROT1 lo comparte, es decir, sigue presentando la clase de predicado acción/actividad. Esto provoca que NPROT1 comparta roles semánticos con PROT. La característica principal del fenómeno diatético entre NPROT1 y PROT es la reestructuración sintáctica/topológica de los roles semánticos compartidos.

La mayoría de los usos NPROT, el subgrupo NPROT2, experimenta una abstracción semántica respecto a PROT, por lo que en el plano semántico hay que hablar de reasignación de roles en NPROT2. Fruto de esta reasignación, desaparecen los roles del campo semántico del dinamismo y son roles semánticos como tema o experientivo los que desempeñan en NPROT2 la función NomE. Esto explica la predominancia de predicados proceso o estado en el subgrupo NPROT2.

Todos los usos NPROT se han expresado mediante la voz activa y pueden considerarse como perspectivas funcionales pasivas o *no* directamente agentivas de los lexemas a priori agentivos *gehen* y *kommen*.

Aunque la distribución de los predicados NPROT según el desarrollo temático textual sea complementaria (NPROT1 en descripciones e instrucciones; NPROT en argumentaciones y exposiciones), dato que subrayaría la idea de una eventual relación entre desarrollo temático textual y perspectiva funcional oracional, el número reducido de predicados analizados (277) no permite comprobar una correlación fiable entre las variables mencionadas.

REFERENCIAS BIBLIOGRÁFICAS

Brinker, K. (1971). *Das Passiv im heutigen Deutsch. Form und Funktion.* München/Düsseldorf: Max Hueber Verlag/Pädagogischer Verlag Schwann (=Linguistische und didaktische Beiträge für den deutschen Sprachunterricht; I/2).

Brinker, K. (1985). *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden*. Berlin: Erich Schmidt (=Grundlagen der Germanistik; 29).

Comrie, B. (1993). *Argument Structure*. En J. Jacobs, A. von Stechow, W. Sternefeld y T. Vennemann (Eds.), *Syntax. Ein internationales Handbuch zeitgenossischer Forschung* (pp. 905-914). Berlin/New York: Walter de Gruyter (=Handbücher zur Sprach- und Kommunikationswissenschaft; 9.1).

Eroms, H. W. (2000). *Syntax der deutschen Sprache.* Berlin/New York: Walter de Gruyter.

Fabricius-Hansen, C. (1991). *Verbklassifikation*. En A. Stechow y D. Wunderlich (Eds.), *Semantik. Ein internationales Handbuch zeitgenossischer Forschung* (pp. 692-709). Berlin/New York: Walter de Gruyter (=Handbücher zur Sprach- und Kommunikationswissenschaft; 6).

Fernández-Villanueva, M. (2002). *Textthematische Entfaltung: Textmodelle und Textexemplare*. Barcelona: PPU.

Fernández-Villanueva, M. y Strunk, O. (2009). Das Korpus Varkom - Variation und Kommunikation in der gesprochenen Sprache. *Deutsch als Fremdsprache, 46,* 67-73.

Figge, U. L. (1995). Valenzen und Diathesen. En C. Schmitt y W. Schweickard (Eds.), *Die romanischen Sprachen im Vergleich.* Akten der gleichnamigen Sektion des Potsdamer Romanistentages (27. - 30.9.1993) (pp. 86-110). Bonn: Romanistischer Verlag.

Fillmore, C. J. (1968). The Case for Case. En E. Bach y R. J. Harms (Eds.), *Universals in Linguistic Theory* (pp. 1-88) New York: Holt, Rinehart, and Winston.

Goergen, P. (1994). *Das lexikalische Feld der deutschen inchoativen Verben*. München: Iudicium Verlag.

Griesbach, H. y Uhlig, G. (1998). *Die starken Verben im Sprachgebrauch. Syntax – Valenz – Kollokationen.* (5ª ed.). Leipzig/Berlin/München/Wien/Zürich/New York: Langescheidt Verlag Enzyklopädie.

Gruber, J. S. (1976). *Lexical structures in syntax and semantics.* Amsterdam, New York, Oxford: North-Holland Publishing Company.

Heinemann, W. y Viehweger, D. (1991). *Textlinguistik. Eine Einführung.* Tübingen: Max Niemeyer Verlag (=Reihe Germanistische Linguistik; 115).

Jackendoff, R. (1991). *Semantic Structures*. (2ª ed.). Massachusetts: The MIT Press.

Lakoff, G. y Johnson, M. (1980). *Metaphors we Live by*. Chicago: Chicago University Press.

Lehrer, A. (1985). Markedness and antonymy. *Journal of Linguistics, 21*, 397-429.

Lyons, J. (1996). *Semantics* (Vol. 1). (11ª ed.). Cambridge: Cambridge University Press.

Manzotti, E., Putsch, L. y Schwarze, C. (1975). Sorten von Prädikaten und - Wohlgeformtheitsbedingung für eine Semantiksprache. *Zeitschrift für germanistische Linguistik, 3,* 15-39.

Mourelatos, A. P. D. (1981). Events, Processes and States. En P. J. Tedeschi (Ed.), *Tense and Aspect* (pp. 191-211). New York/London/Toronto/Sydney/San Francisco: Academic Press. (= Syntax and Semantics; 14).

Pittner, K. (1995). Valenz und Relevanz - eine informationsstrukturelle Erklärung für "obligatorische" Adverbiale. En R. J. Pittner y K. Pittner (Eds.), *Beiträge zu Sprache und Sprachen. Vorträge der 4. Münchner Linguistik-Tage der Gesellschaft für Sprache und Sprachen. (GESUS) e.V*. (pp. 95-106). München: Lincom Europa.

Polenz, P. von (1988). *Deutsche Satzsemantik. Grundbegriffe des Zwischen-den-Zeilen-Lesens.* (2ª ed.). Berlin/NewYork: Walter de Gruyter (=Sammlung Göschen; 2226).

Porzig, W. (1971). *Das Wunder der Sprache*. (5ª ed.). München: Francke Verlag (=Uni-Taschenbücher; 32).

Sánchez, P. (2009). Estudio lexicológico semántico sobre los verbos *geben-bekommen* y la relación antonímica de inversión. *Revista de Filología Alemana, 17*, 143-158.

Sandig, B. (1996). Sprachliche Perspektivierung und perspektivierende Stile. *LiLi – Zeitschrift für Literaturwissenschaft und Linguistik, 26,* 36-63.

Schmid, H. (1994). TreeTagger - a language independent part-of-speech tagger. Disponible en http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTree Tagger. html

Schreiber, H., Sommerfeldt, K. E. y Starke, G. (1990). *Deutsche Wortfelder für den Sprachunterricht. Verbgruppen*. (2ª ed.). Leipzig: Verlag Enzyklopädie.

Schröder, G. (1995). *Lexikon deutscher Präfixverben*. (4ª ed.). Berlin/München/Leipzig: Langenscheidt Verlag Enzyklopädie.

Schumacher, H., Kubczak, J., Schmidt, R. y de Ruiter, V. (2004). *VALBU – Valenzwörterbuch deutscher Verben*. Tübingen: Gunter Narr Verlag (=Forschungen des Instituts für deutsche Sprache; 31).

Strunk, O. (2008). Lexic Tools: Una herramienta integrada para el etiquetado gramatical y el trabajo con concordancias. En C. M. Bretones Callejas y J. R. Ibáñez Ibáñez (Eds.), *De la lingüística aplicada a la lingüística de la mente: Hitos, prácticas y tendencias.* XXVI Congreso Internacional de la Asociación Española de Lingüística Aplicada (AESLA) (pp. 236-237). Almería: Universidad de Almería.

Stutterheim, C. von (1997). *Einige Prinzipien des Textaufbaus. Empirische Untersuchungen zur Produktion mündlicher Texte.* Tübingen: Max Niemeyer Verlag.

Tappe, H. (2000). Perspektivenwahl in Beschreibungen dynamischer und statischer Wegeskizzen. En C. Habel y C. von Stutterheim (Eds.), *Räumliche Konzepte und sprachliche Strukturen,* (pp. 69-95). Tübingen: Max Niemeyer Verlag (=Linguistische Arbeiten; 417).

Taylor, H. A. y Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. *Journal of Memory and Language, 31,* 261-292.

Taylor, H. A. y Tversky, B. (1996). Perspective in spatial descriptions. *Journal of Memory and Language, 35,* 371-391.

Weinrich, H. (2007). *Textgrammatik der deutschen Sprache.* (4ª ed.). Hildesheim, Zürich, New York: Georg Olms Verlag.

Wunderlich, D. (1993). Diathesen. En J. Jacobs, A. Stechow, W. Sternefeld y T. Vennemann (Eds.), *Syntax. Ein internationales Handbuch zeitgenossischer Forschung* (pp. 730-747). Berlin/New York: Walter de Gruyter (=Handbücher zur Sprach- und Kommunikationswissenschaft; 9.1).

Zifonun, G., Hoffmann, L. y Strecker, B. (1997). *Grammatik der deutschen Sprache.* Berlin, New York: Walter de Gruyter.

THAI-NEST: A framework for Thai named entity tagging specification and tools

THANARUK THEERAMUNKONG, MONTHIKA BORIBOON, CHOOCHART HARUECHAIYASAK,
NICHNAN KITTIPHATTANABAWON, KRIT KOSAWAT, CHUTAMANEE ONSUWAN,
ISSARIYAPOL SIRIWAT, THAWATCHAI SUWANAPONG, AND NATTAPONG TONGTEP
Thammasat University
National Electronics and Computer Technology Center

Abstract
A THAI-NEST framework is presented for a construction of Thai news corpus with named entity (NE) tagging process. Three main components of the framework are corpus tagging specification, tagging process, and tagging tools. To be in line with the Text Encoding Initiative (TEI) standardization, a tagging specification is developed by taking into account some characteristics of Thai NEs, including proper nouns, expressions of date, time, and quantity, and other extended named entities. The developed specification includes a tag set and its tagging schema. A set of tagging tools is designed and implemented with an effective GUI. The tool set supports two tagging levels of NE type and NE structure. Results and statistics of our ongoing corpus construction are reported.
Keywords: Named Entity, Thai Language, News Corpus, Language Resource

Resumen
Se presenta el marco THAI-NEST para construir un corpus de noticias tailandesas mediante un proceso de etiquetaje de entidades nombradas (NE). Los tres componentes principales del marco son la especificación del etiquetaje del corpus, el proceso de etiquetaje y las herramientas de etiquetaje. Para seguir la línea de la estandarización de la Text Encoding Initiative (TEI), se desarrolla una especificación de etiquetaje teniendo en cuenta algunas características de entidades nombradas tailandesas, incluyendo nombres propios, expresiones de fecha, hora y cantidad, así como otras entidades nombradas. La especificación desarrollada incluye un conjunto de etiquetas y su esquema de etiquetaje. Se diseña un conjunto de herramientas de etiquetaje que se implementa con un GUI efectivo. El conjunto de herramientas admite dos niveles de tipo y estructura de entidades nombradas. Se informará de los resultados y estadísticas de nuestro corpus en construcción.
Palabras clave: Entidad Nombrada, Lengua Tailandesa, Corpus de Noticias, Recurso Lingüístico

1. INTRODUCTION[1]

Named Entity Recognition (NER) is considered one of the fundamental tasks in NLP. In Thai NLP community, however, the topic of NE related tasks is not as widely discussed as in other languages. The early Thai NER had some limitations. Firstly, it was specific to a very few

[1] This research is supported by the National Electronics and Computer Technology Center under the project code NT-B-22-KE-38-52-01, and partially supported by National Research Council of Thailand (NRCT) via Thammasat University as well as Thailand Research Fund under the project number BRG5080013.

areas such as agricultural (Kawtrakul *et al*., 2001) and political news (Chanlekha & Kawtrakul, 2004). Secondly, only a small class of NE types (person name, organization, and place) was mentioned. Lastly, evaluation tasks were usually performed on small-size corpora. With this in mind, the main goal of this project is to design a new framework for constructing a large-scale NE corpus with a larger set of NE types from various domains.

The proposed framework (THAI-NEST) includes the following components: (1) a specification for Thai NE tag set, (2) a tagging process, and (3) tagging tools. Our tag set specification was adapted from the TEI guidelines (Barnard & Ide, 1997) to suit the Thai language characteristics. The tag set followed some specifications proposed by TEI guidelines with additional modifications. Our NE tag set includes person name, organization name, place name, date and time expression, quantitative expression. In addition, we also considered an annotation of other named entities previously proposed by the Sekine's Extended Named Entity Hierarchy (Sekine, 2007).

The tagging process consists of three main steps: (1) news collection, (2) news article metadata and structure tagging, and (3) NE tagging and verification. Also due to a very short project time frame of one year, our tagging process was designed such that many steps can be carried out in a pipeline and parallel processing manner. For example, the news article structure tagging and the NE tagging can be performed as a pipeline process. Moreover, the tagging of different NE types can be done in parallel. In this paper, we will discuss and share our experience in designing the framework to allow the maximum resource allocation. With 10,000 news articles being annotated, our corpus will be the largest Thai NE corpus to date.

The remainder of paper is organized as follows. The related work in NE related tasks, i.e., corpus construction and NER, is given in the next section with special focus on Thai language. In Section 3, we present the proposed tagging framework which includes tag set design, tagging process and tools. A flowchart diagram with some tagging examples will be given to illustrate the proposed framework. Section 4 provides a summary of current corpus statistics. A summary with discussion is presented in the last section.

2. RELATED WORK

Named entities recognition (NER) is one of the most extensively studied topics in NLP. There have been many organized conferences and workshops to discuss related issues including corpus designs and algorithms for recognizing and extracting NEs. Early conferences include MUC (Message Understanding Conference) and CoNLL (Conference on

Computational Natural Language Learning). MUC, initiated and funded by the Defense Advanced Research Projects Agency (DARPA), is perhaps the first widely recognized conference which focused on designing and creating a large-scale corpus specifically for NE related tasks (Chinchor, 1998). There are three-main subtasks considered under MUC: (1) entity names (person, organization, and location), (2) temporal expressions (date and time) and (3) number expressions (monetary expression and percentage).

CoNLL, on the other hand, focuses on language-independent NER task (Sang *et al*., 2003). In addition to basic NE types, CoNLL corpus also includes miscellaneous (MISC) names belonging to different domains such as adjectives (e.g., Italian) and events (e.g., World Cup, Olympics). Another effort in NE tasks is the ACE (Automatic Content Extraction) conference organized by the National Institute of Standards and Technology (NIST) (Linguistic Data Consortium, 2008). Under ACE, entities are categorized into seven types: person (PER), organization (ORG), geo-political entity (GPE), location (LOC), facility (FAC), vehicle (VEH), and weapon (WEA). Each tag is allowed to have subtypes such as *PER.Individual* to denote individual person and *PER.Group* to denote a group of people. ACE aims to develop some automatic content extraction techniques from different text sources such as newswire, broadcast conversation, and weblogs.

Recently the NE related tasks are increasingly recognized and have been discussed among Asian NLP community. Kumano *et al*. (2003) constructed a cross-lingual Japanese-English broadcast news corpus of 1,100 article pairs with NE tags. Their goal was to acquire NE translation knowledge by utilizing NE extraction techniques. Takenobu *et al*., (2006) reported a collaboration among many countries in Asia to create a common standard for Asian language resources. Their project focuses on constructing a lexicon set with an upper-layer ontology. The NE related tasks have yet to be discussed. In IJCNLP 2008 (The Third International Joint Conference on Natural Language Processing), there were two workshops related to NER tasks: (1) Named Entity Recognition for South and South East Asian Languages (NERSSEAL) and (2) Asian Language Resources (ALR). There are many reported NE corpora and tools in many different Asian languages including Indian, Telugu, Bengali, and Tamil (Ekbal & Bandyopadhyay, 2008; Saha *et al*., 2008; Sangal *et al*., 2008).

As for NE related tasks in Thai, Tongtep and Theeramunkong (2008) have presented a pattern-based approach for named entity extraction from Thai news documents. This named entity extraction is later applied as preprocessing for relation extraction from Thai news documents in Tongtep and Theeramunkong (2009). During the most recent SNLP 2009 (The Eighth International Symposium on Natural Language Processing), there were some reports

on Thai NE related tasks. Lertcheva and Aroonmanakun (2009) applied the CRF (Conditional Random Fields) algorithm to construct an NER model for Thai language. The corpus is based on the BEST 2009 word segmentation corpus (Kosawat *et al*., 2009) and contains only 90,000 words. Only three named entities (person, organization and place) were considered. Suwanapong and Theeramunkong (2009) proposed a method based on the SVD (Singular Value Decomposition) algorithm to identify aliases in Thai sports news articles. Inyaem *et al*. (2009) presented a method based on domain-specific NE to extract terrorism events from Thai news articles. Sutheebanjerd and Premchaiswadi (2009) proposed a different method to extract NEs from Thai news articles. Their work only covered person names and their models were trained using only approximately 1,000 news articles. Tirasaroj and Aroonmanakun (2009) presented a study on linguistic structure of Thai product names from economic news articles.

In sum, previous work on Thai NE related tasks was very limited to a few NE types and also domain-specific to certain topics. There has not been any open large-scale NE corpus with multiple NE types for Thai language. Perhaps the most related works to ours are the development of "Simple Named Entity Guidelines for Thai" and "Time Annotation Guidelines for Less Commonly Taught Languages: Thai" carried out by a research group at the Linguistic Data Consortium (LDC) (Linguistic Data Consortium, 2006a; Linguistic Data Consortium, 2006b). The simple named entity guidelines are based on the MUC-7 NE Guidelines which cover three basic NE types of person, organization and location. The time annotation guidelines are based on the TIMEX2 standard which provides a tag set for tagging temporal expressions.

Compared with the guidelines above, our proposed NE tag set is designed based on the adaptation of TEI guidelines which provide rich details for the designed tag sets (Barnard & Ide, 1997). Some of the tags are modified and added to make them suitable for the Thai language characteristics. For the application of detection and recognition tasks, tag sets are modified. For example, unlike TEI guidelines which classify <*placeName*> into various geopolitical subtypes, <*district*>, <*settlement*>, <*region*>, <*country*>, and <*bloc*>, we propose <*prefix*>, <*infix*>, and <*suffix*> to be used as a feature set for training the NER model. Such design provides additional syntactic features to compensate for the lack of capitalization in Thai.

Our guidelines cover seven NE types along with the framework of process and tools to assist the tagging process. We use Thai news articles collected from the Web as the main resource for the corpus. Once completed, we plan to release the corpus to the Thai NLP community.

3. THE PROPOSED TAGGING FRAMEWORK

In this section, we describe the tagging specification, process, and tools. The details on the tag set design with a tagging example are given to illustrate our proposed framework.

3.1. Tag Set Design

Many single and multiple word NEs in Thai derive from existing lexicon. Thus, some components of NE may be confused with ordinary words. For example, in a sentence หัวหน้าขอบอ้างอภิสิทธิ์ (อภิสิทธิ์, a single word, has two meanings: the name of a person (NE) or a privilege). Regular dictionaries often do not include NEs because they are an open set, i.e., new NEs could be coined everyday. So, it is not reasonable to recognize NEs by mapping with common dictionary, except for some frequently used ones. A more practical approach is by using Machine Learning (ML) technique in which annotated corpora are employed to train a model. Features such as capitalization, word boundaries are used as contextual clues in Latin-based writing systems. However, these features do not exist in Thai as shown in the above example where written text appears as a long string of uniform connected symbols. Contextual understanding and background knowledge are required to identify Thai NEs.

NE-Type Level	NE-Structure Level	Description	Example
<persName>		■ Proper noun or proper noun phrase referring to people	<persName> <roleName>ดร.</roleName> <forename>อนารักษ์</forename> (<addname>ปิง</addname>) <surname>ธีระมั่นคง</surname> </persName>
	<roleName> <forename> <middlename> <surname> <addName>	• Titles, honorific titles, ranks, kinship terms, etc. used before given name • First name • Middle name (if any) • Family or last name • Additional or alias name	
<orgName>		■ Name of organization, institution, nation, etc.	<orgName> <component> <prefix>สาขาวิชา</prefix> ภาษาศาสตร์ </component> <component> <prefix>คณะ</prefix>ศิลปศาสตร์ </component> <component> <prefix>มหาวิทยาลัย</prefix> ธรรมศาสตร์ <infix>ศูนย์</infix>รังสิต </component> </orgName>
	<prefix> <infix> <suffix> <addName> <component>	• Set of words indicating types used before a given organization name • Set of words indicating sub-types used inside a given organization name • Set of words indicating types used after a given organization name • Additional or alias name, metonyms of a given organization name • Members of single contiguous organization sequence	
<placeName>		■ Name of place, location, geographical bodies, country, etc.	<placeName> <prefix>พระที่นั่ง</prefix>จักรี <suffix>มหาปราสาท</suffix> </placeName>
	<prefix> <infix> <suffix> <addName> <component>	• Set of words indicating types used before a given place name • Set of words indicating sub-types used inside a given place name • Set of words indicating types used after a given place name • Additional or alias name, metonyms of a given place name • Members of single contiguous place name sequence	
<date>		■ Date in any format	<date> <offset>ก่อน</offset> วันที่ 1 มีนาคม 2009 </date>
	<offset>	• Set of words (before or after date expression) including repositions, subordinating conjunctions	
<time>		■ Time of day in any format	<time>19.30 น. <offset>เป็นต้นไป</offset> </time>
	<offset>	• Set of words (before or after time expression) including repositions, subordinating conjunctions	
<measure>		■ Word or phrase referring to some quantity of an object or commodity	<measure> <offset>ประมาณ</offset> <quantity>30</quantity> <unit>คน</unit> </measure>
	<offset> <quantity> <unit>	• Set of words (before or after quantitative expression) including adverbial and adjectival phrases • Digit and characters signifying measurement, amount, value, and ordinal number • Noun classifiers	
<name>		■ Extended named entities excluding those that refer to the person name, organization name, and place name (e.g., food's name, disease, product's name, etc.)	<name>รถโตโยต้า</name>

Figure 1: NE-Types, structures, and examples

Like other languages, Thai NEs are also presented with clue words that are specific to some types of NE; for example, the following terms นาย (Mr.) นาง or (Mrs.) can

be used to indicate the beginning of a Person Name, while จำกัด (Ltd.) typically used to indicate the end of an Organization Name. Those clue words are tagged separately as NE internal structure. Their positions will later be learned by ML algorithms to better predict the NE types and boundaries. Unlike NEs which are open set and hard to recognize, clue words are expected to some extent to be a closed set. Since, at this stage, we develop our corpus manually, word segmentation is beyond the scope of this paper.

In our proposed framework, Thai NE tagging comprises two levels 1) NE-Type Level, and 2) NE-Structure Level. The types and structures of NEs are fully listed in Figure 1 along with tagging examples.

3.2. Tagging Process

A flowchart (shown in Figure 2) summarizes main steps involved in the tagging process.

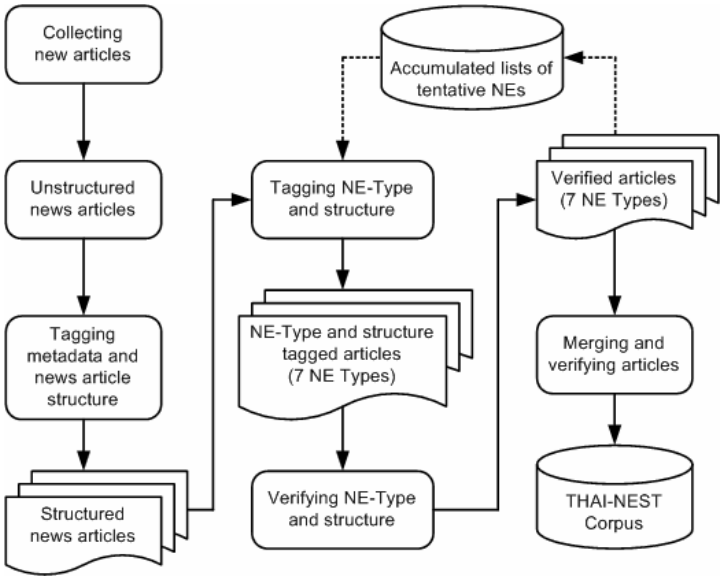


Figure 2: The tagging process

3.3. Collecting news articles

News articles during January to December 2009 have been collected from twenty-one Thai online newspaper publishers including seven major categories: crimes (CR), politics (PO), foreign affairs (FO), sports (SP), education (ED), entertainment (EN), and economic issues (EC). Over 300,000 news articles have been collected and imported to the next step using our developed Thai news structure tagging tool.

3.4. Tagging metadata and news article structure

From the collected news articles, we selected and balanced a set of 10,000 news articles in terms of news categories, publishers and time periods. Thai news structure tagging tool automatically converted the selected news articles into XML format with UTF-8 encoding. Then, metadata and news article structure were systematically assigned to the texts.

As shown in Figure 3, assigned metadata contains news title, author, publisher name, publisher URL, published date, news category, and news source URL. The news article structure includes headline, lead, and body of the news. Lastly, file names were given according to published date, news category, publisher, and file status (original, pending, submitted, and verified).

```
<?xml version="1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>news title or headline</title>
        <author>journalist or publisher</author>
      </titleStmt>
      <publicationStmt>
        <publisher>news publisher</publisher>
        <pubPlace>publisher URL</pubPlace>
        <date>published date (yyyy-mm-dd hh:mm:ss)</date>
      </publicatonStmt>
      <notesStmt>
        <note>news category</note>
      </notesStmt>
      <sourceDesc>
        <bibl>news source URL</bibl>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <news>
    <headline>news headline</headline>
    <lead>news lead</lead>
    <body>
      <p>news details</p>
      …
    </body>
  </news>
</TEI>
```

Figure 3: Thai news article metadata and structure

3.5. Tagging NE-Type and structure

Using our developed Thai NE Annotation Tool as shown in Figure 4, NE of seven types in news articles were tagged in parallel by linguists according to our NE annotation guidelines.

Therefore, one original new article produces seven separate files according to their tagged NEs.

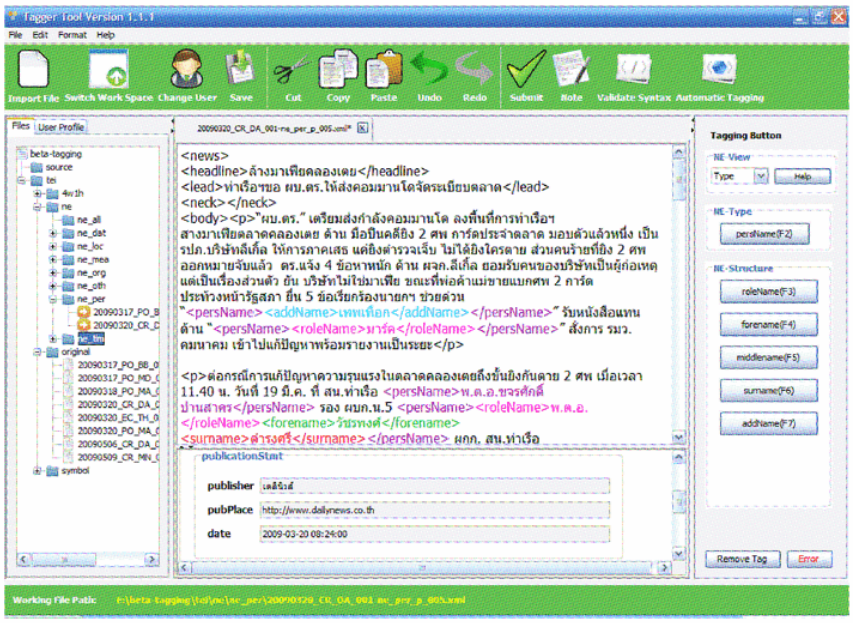


Figure 4: Snapshot of Thai NE Annotation Tool

To save time and reduce annotators' effort, the NE annotation tool was designed to incorporate several useful functions such as syntax validation, note/memo taking, log and status control, and customized GUI. Misspellings and other irregularities may be marked and kept separately. In addition, an automatic tagging function, which can be used to speed up the process, was implemented by using accumulated lists of tentative NEs either created manually by annotators or obtained from the upcoming verifying process.

3.6. Verifying NE-Type and structure

To ensure the validity and consistency of our tagging, NE of seven types was separately listed, double-checked, and corrected (if necessary). In this process, another tool – NE Tagging Verification Tool was introduced. As mentioned above, lists of each NE type can be exported to the automatic tagging function.

3.7. Merging and verifying articles

A designed function available in NE Tagging Verification Tool merges a set of separate files (from the same original file) into one file with completed NE tagging. There are three merging result status: finished, incomplete (more files are needed), and failed (addition or

deletion of lines and/or characters). Only files with finished status will proceed to our THAI-NEST corpus.

4. CORPUS STATISTICS

A summary of NE tagging statistics can be generated from the NE Tagging Verification Tool. The reports include (1) the number of files, (2) the total and distinct number of tagged NEs, and (3) file size, with respect to year, month, publisher, and news category. Moreover, the user notes, logs, and errors can be viewed separately. Figure 5 illustrates a snapshot of the system showing the total and distinct number of tagged NEs.

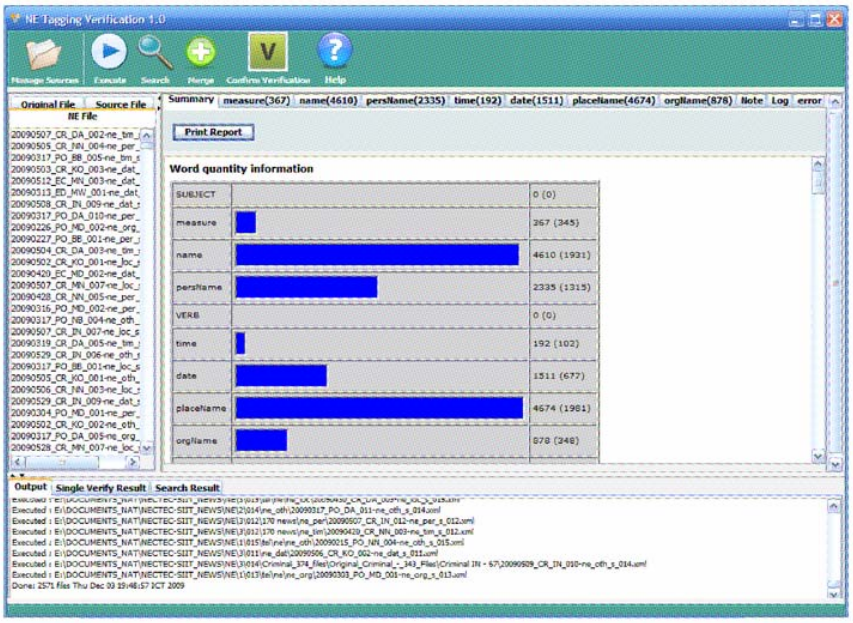


Figure 5: Snapshot of NE Statistics reported by NE Tagging Verification Tool

The current status of NE tagging process can be summarized in Table 1. To date, the numbers in the table are calculated from 7,520 tagged files.

Table 1: The number of NEs in the NE-Type level

NE-Type	#tokens	#unique tokens	#docs	#tokens/doc
<*persName*>	4,110	2,071	584	7.04
<*orgName*>	18,374	5,339	2,110	8.71
<*placeName*>	6,411	2,319	981	6.54
<*date*>	3,197	1,314	1,224	2.61
<*time*>	413	165	1,221	0.34
<*measure*>	3,068	2,348	339	9.05
<*name*>	9,482	3,842	1,061	8.94
Total	**45,055**	**17,398**	**7,520**	**43.22**

From Table 1, we observed that the largest average number of tokens per document (last column) is <*name*> – extended named entities, the smallest number is <*time*>. On average, the number of NEs found in each news article is 43.22. It is worth mentioning that the discrepancy in the number of tagged documents to date can be explained by the parallel process we have adopted (one annotator for each NE-type tag set). Due to the difference in annotators' experience and knowledge and task difficulty, the numbers of tagged documents varied among the NE types.

5. SUMMARY AND DISCUSSION

We proposed a new framework called THAI-NEST (THAI-Named Entities Specification and Tools) to support the NE corpus construction from Thai news articles. Since the project time frame is set within one year, the framework was carefully designed to allow efficient resource allocation and usage (i.e., time and humans). In this framework, three issues were considered: (1) a specification for Thai NE tag set, (2) a tagging process, and (3) tagging tools. Our tag set specification was adapted from the TEI guidelines to suit the Thai language characteristics. The tagging process includes the verification step which allows the tagged corpora of different NE types to be merged and checked for any error. To allow the maximum efficiency, the tagging tools were designed according to the proposed tagging process.

It is worth noting that one of the common problems found during the tagging process is difficulties in recognizing various jargons and metonym from different domains. Therefore, annotators should be assigned based on their familiarity with the domain. Another problem is ambiguous cases between organization and place names. Determining these two NE types requires deeper interpretation of contextual information and background knowledge. As a

final step of this project, we hope to review each individual tagging process and its shortcomings and to provide some suggestions for a creation of other Thai corpora.

As for future works, we plan to train and evaluate models using available machine learning techniques such as the CRFs. We will provide an initial technical report on training the NER model by using the corpus. To promote the NER task for Thai language, we plan to make the corpus publicly available. Another plan is to extend the use of this corpus by including another tag set of 4W1H (*Who*, *What*, *When*, *Where*, and *How many*). The 4W1H tag set can be partially transformed from the current NE tag set with further modification. For example, person names from the NE tag set can be mapped into *Who* of the 4W1H tag set. In addition, we will include a tag set of *subject* and *verb* to provide the *action* part of sentences. The newly designed 4W1H plus *subject* and *verb* tag set would be very useful for implementing a QA system from news articles.

REFERENCES

Barnard, D.T. & Ide, N.M. (1997). The text encoding initiative: flexible and extensible document encoding. *Journal of the American Society for Information Science*. 48(7), 622-628.

Chanlekha, H. & Kawtrakul, A. (2004). Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information, *Proceedings of the IJCNLP 2004*. (pp. 49–55).

Chinchor, N.A. (1998). Overview of Proceedings of the Seventh Message Understanding Conference (MUC-7)/MET-2, *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. (pp. 5).

Ekbal, A. & Bandyopadhyay, S. (2008). Development of Bengali Named Entity Tagged Corpus and its Use in NER Systems, *Proceedings of the IJCNLP 2008 workshop on Asian Language Resources*. (pp. 1-8).

Inyaem, U., Meesad, P., & Haruechaiyasak, C. (2009). Named Entity Techniques for Terrorism Event Extraction and Classification, *Proceedings of the 8th International Symposium on Natural Language Processing*. (pp. 175-179).

Kawtrakul, A., Collier, N., Takeuchi, K., Ono, K., Suktarachan, M., Chanlekha, H., Waiyamai, K. (2001). Collaboration on Named Entity Discovery in Thai Agricultural

Texts, *Proceedings of the 8th International Workshop on Academic Information Networks and Systems*. (pp. 77-82).

Kosawat, K., Boriboon, M., Chootrakool, P., Chotimongkol, A., Klaithin, S., Kongyoung, S., Kriengket, K., Phaholphinyo, S., Purodakananda, S., Thanakulwarapas, T., & Wutiwiwatchai, C. (2009). BEST 2009: Thai Word Segmentation Software Contest, *Proceedings the 8th International Symposium on Natural Language Processing*. (pp. 83-88).

Kumano, T., Kashioka, H., Tanaka, H., & Fukusima, T. (2003). Construction and analysis of Japanese-English broadcast news corpus with named entity tags, *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition*. (pp. 17-24).

Lertcheva, N. & Aroonmanakun, W. (2009). A Linguistic Study of Product Names in Thai Economic News, *Proceedings the 8th International Symposium on Natural Language Processing*. (pp. 26-29).

Linguistic Data Consortium (LDC). (2008). ACE (Automatic Content Extraction) *English Annotation Guidelines for Entities Version 6.6 2008.06.13*.

Linguistic Data Consortium (LDC). (2006). Simple Named Entity Guidelines Version 6.4 Thai. Retrieved from *http://projects.ldc.upenn.edu/LCTL/Specifications/SimpleNamedEntityGuidelinesV6.4-Thai.pdf*

Linguistic Data Consortium (LDC). (2006). Time Annotation Guidelines for Less Commonly Taught Languages: Thai (Based upon the TIMEX2 Standard) Version 1.0. Retrieved from *http://www.ldc.upenn.edu/Projects/LCTL/Specifications/TimeAnnotation GuidelinesV1. 0-Thai.pdf.*

Saha, S.K., Sarkar, S. & Mitra, P. (2008). Gazetteer Preparation for Named Entity Recognition in Indian Languages, *Proceedings of the IJCNLP 2008 workshop on Asian Language Resources*. (pp. 9-16).

Sangal, R., Misra Sharma, D., & Kumar Singh, A., Eds, *Proceedings of the IJCNLP 2008 workshop on Named Entity Recognition for South and South East Asian Languages*.

Sekine, S. Sekine's Extended Named Entity Hierarchy. Retrieved from *http://nlp.cs.nyu.edu/ene/.*

Sutheebanjerd, P. & Premchaiswadi, W. (2009). Thai Personal Named Entity Extraction without Using Word Segmentation or POS Tagging, *Proceedings of the 8th International Symposium on Natural Language Processing*. (pp. 221-226).

Suwanapong, T. & Theeramunkong, T. (2009). Aliases Discovered in Thai Sports News Articles, *Proceedings of the 8th International Symposium on Natural Language Processing*. (pp. 63-66).

Takenobu, T., Calzolari, N., Huang, C. R., Prevot, L., Sornlertlamvanich, V., Monachini, M., YingJu, X., Kiyoaki, S., Charoenporn, T., Soria, C., & Hao, Y. (2006). Infrastructure for standardization of Asian language resources, *Proceedings of the COLING/ACL on Main conference poster sessions*. (pp. 827-834).

Tirasaroj, N. & Aroonmanakun, W. (2009). Thai Named Entity Recognition Based on Conditional Random Fields, *Proceedings of the 8th International Symposium on Natural Language Processing*. (pp. 216-220).

Tjong, K. S., Erik, F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: language-independent named entity recognition, *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. (pp. 142-147).

Tongtep, N. & Theeramunkong, T. (2008). Pattern-based Named Entity Extraction for Thai News Documents, *Proceedings of the 3rd International Conference on Knowledge, Information and Creativity Support Systems (KICSS'08)*. (pp. 82-89).

Tongtep, N. & Theeramunkong, T. (2009). A Feature-based Approach for Relation Extraction from Thai News Documents, *The Pacific Asia Workshop on Intelligence and Security Informatics 2009 (PAISI-09)*. (pp. 155-160).

The semantic function of affixation in a corpus of Old English derived nouns

ROBERTO TORRE ALONSO

Universidad de La Rioja

Abstract

The aim of this paper is to examine the function of Old English affixation in terms of meaning modification. For this purpose, a corpus has been analysed of 4,084 derived nouns retrieved from the lexical database of Old English Nerthus (www.nerthusproject.com). Extensive data analysis shows that there are differences in the semantic function of affixes, which can be related to the change undergone by the type of derivational morphology of Old English, which shifted from variable bases (stems) to invariable bases (words). The conclusion is reached that affixes add new meaning when attached to derived nouns whereas some examples have been found of lack of semantic modification when the input to the morphological process is an underived form.

KeyWords: Morphology, word-formation, affixation, semantics

Resumen

*Durante el periodo de inglés antiguo se produjo un cambio tipológico, por el cual las nuevas creaciones léxicas dejaron de estar basadas en formas variables (stems) y comenzaron a basarse en formas invariables (palabras). Así, asumiendo que este cambio tipológico se ha completado de manera homogénea, podemos analizar la formación de palabras en relación con lexemas existentes en la lengua. Desde una perspectiva semántica, este trabajo pretende definir las implicaciones que la adición de afijos tiene en la creación léxica. El estudio de un corpus de 4.084 predicados obtenidos de la base de datos Nerthus_(*www.nerthusproject.com*) muestra divergencias semánticas en la incorporación de afiijos a palabras que han sido previamente derivadas. Los afijos aportan significado cuando la base es afijada, sin embargo, con bases cero derivadas se observan casos donde la adición de afijos no implica modificación semántica.*

Palabras Clave: Morfología, formación de palabras, afijación, semántica

1. INTRODUCTION[1]

This paper aims at describing, in a tentative way, the initial steps undertaken in the semantic study of Old English word formation, in the corpus provided by the database *Nerthus* (Martín Arista *et al.* 2009). For so doing, a total of 4,084 derived predicates have been reviewed in a double analysis. In the first place, a morphological deconstruction of the predicates has been carried out, following the methodological principles stated in *Fundamentos empíricos y metodológicos de una base de datos léxica de la morfología derivativa del inglés antiguo* (Torre Alonso *et al.* 2008), which allows to identify the individual constituents taking part in each word formation. Secondly, a contrastive analysis has been made, by which the meanings

[1] This research has been carried out with the funding of the *Ministerio de Educación y Ciencia* through the grant FFI2008-04448/FILO.

of the base constituents and those of the resulting derived predicates are compared, thus allowing for a more concrete view of the function of affixation.

The semantic studies of the Old English lexicon so far have been based upon text translations. This fact is responsible for the major shortcomings of this type of analysis. Translations do not constitute systematic and stable analytical processes, and not unconnected to it, translations may be imbued with lexicographical bias. The existence of a corpus that allows for the identification of the different constituents of complex words, and the ordering of related words around morphologically related families constitutes a gigantic step forward in the functional and semantic studies of Old English affixation.

Nonetheless, until the morphological analysis *Nerthus* aims at is fully developed, I am forced to trust and operate with the existing dictionary definitions, some of which are theoretically old-fashioned while others are not complete yet. As for the semantic value of affixes, several works have been devoted to establish the meanings and functions of these bound morphemes. Difficulties arise first in what scholars decide to be grammatical or lexical elements, on the establishment of a close set of affixes. Not less important is the overlapping of functions a single affix may carry out, thus leading to an extension of their semantic definitions. Last but not least, spelling variations lead to the mixing and overlapping of affix forms, creating an obscure and not-easy-to-explore field of analysis.

While these questions will be treated in turn, I shall not close this section without devoting a word to recursivity and recategorisation.. According to Martin Arista (2009) these two phenomena constitute the basis of a morphological study on word formation. This paper aims at examining the consequences of affixation upon already derived bases, and try and establish differences between the relationship among the different derivational processes taken into account, that is, affixation (both prefixation and suffixation), compounding and zero-derivation.

2. EXPLORING A DATABASE CORPUS OF DERIVED NOUNS

The database *Nerthus* (www.nerthusproject.com) upon which this research is based, includes morphological and semantic information compiled from different sources, mainly from *A Concise Anglo-Saxon Dictionary* (Clark Hall, 1996), but also from *An Anglo-Saxon dictionary* (Bosworth and *Toller*, 1973) and *The student's dictionary of Anglo-Saxon* (Sweet, 1976). Without taking more dictionaries into account, some differences in the meanings offered for some predicates arise, as (1) demonstrates:

(1)

mierring: (CH) ‘hindering, squandering, waste’; (Sweet) ‘leading astray, squandering, waste’; (BT) ‘(I) hindering, leading astray, (II) squandering, waste’

mearcung: (CH) ‘(-) marking, branding; mark, characteristic; (+/-) description, arrangement; constellation; title, chapter’; (Sweet) ‘(ge-) marking, mark, characteristic; marking out, description; constellation’; (BT) ‘(I) a marking, mark; (II) a marking out, description, arrangement, disposition’

Differences lay not only on the meanings provided, but on the morphological form the predicate has when it is assigned a precise meaning, and on the distribution and organisation of the meanings assigned to the dictionary entry. Thus, we can observe that Bosworth and Toller (1973)separate meanings according to different senses of words, while Clark Hall (1996) prefers a separation according to the morphological structure of the word based on the presence or absence of the prefix *ge-*. Sweet also takes this factor into account, thus allowing of another element of divergence between these proposals. For Sweet (1976), the meanings ‘marking, mark, characteristic’ apply when the predicate *mearcung* is prefixed, while Clark Hall (1996) includes these meanings, when the predicate lacks *ge-*.

Erasing these differences is what a corpus-based study aims at. But for so doing, a definite, precise and well-established corpus of analysis must be defined. In this case, I deal with a total of 4,084 affixed Old English nouns obtained from the database *Nerthus*. I shall not begin without acknowledging the manner in which the inventory of affixes involved in the creation of derived nouns has been established. Affixed predicates are formed by the attachment of a bound morpheme to a free predicate, both in its prefield or in the postfield, in the case of Old English. As for compounding, it includes those complex predicates formed by means of the combination of two free lexemes. Nevertheless, the distinction between bound and free morphemes is not clear in functional terms. Mairal Usón and Cortés Rodríguez (2000-2001) have analyzed derivational morphemes as predicates, thus doing away with the distinction between free forms (lexemes) and bound forms (derivational morphemes) because both are listed as predicates in the lexicon. In the same line, Martín Arista (2008, 2009) has demonstrated that the same word-functions can be performed by free and bound morphemes, that is, there is no functional difference between the insertion of a free or a bound form into the word slots. Although the borderline between derivation and compounding is not always clear, the distinction between both processes is maintained in this analysis in order to perform

the gradual study of processes and focus on the restrictions that may be imposed on the different combinatory elements. This distinction however, poses the problem of 'affixoids' (Kastovsky, 1992), borderline cases between derivation and compounding. They are elements that exist as independent lexemes in the lexicon of the language and which are going through a process of grammaticalization, whereby a lexical item becomes a bound form (Bauer, 2007). The inventory of affixoids includes the prefixoids shown in (2a) and the suffixoids in (2b):

(2)

a. *æfter-* 'after', *be-* 'by, near', *fær-* 'calamity, sudden danger, peril, sudden attack', *for-* 'before, from', *fore-* 'before', *forð-* 'forth, forwards', *ful-* 'full', *in-* 'in', *of-* 'over, above', *ofer-* 'over', *on-* 'on', *to:-* 'to', *ðurh-* 'through', *under-* 'under', *up-* 'up', *ut-* 'out, without', *wan-* 'lack of', *wið-* 'with, near, against', *wiðer* 'against' and *ymbe-* 'around, about'.
b. *-bora* 'bearer', *-do:m* 'doom, condition', *-ha:d* 'person, condition, state', *-la:c* 'play, sacrifice', *-mæ:l* 'mark, measure', *-ræ:den* 'terms, condition' and *-wist* 'being, existence'.

For the purposes of this paper, the question of the separation between affixation and compounding regarding the affixoids has been solved by analysing the predicates in which these elements appear. When the number of lexicalized predicates is relevant, the affixoid has been treated as a pure affix. In the postfield, this treatment does not cause further problem, for in Modern English, those affixoids have been fully grammaticalized, as in *fre:onscipe* 'friendship' or *wi:sdo:m* 'wisdom'. In the prefield, however, the question is more complex. By assuming total grammaticalization, I am considering as inseparable some forms which can, nowadays, be detached from the base predicate, as in *incuman* 'to come in, to go into' or *forðsendan* 'to send forth'. The reason for this decision is that this analysis is more oriented towards form. at this stage of the analysis.

The full inventory of the affixes identified for this research is as follows. Brackets represent spelling variants. The prefixes are stated in (3a), while suffixes are grouped in (3b):[2]

[2] The forms in brackets represent alternative spellings.

(3)

a. *a:- (æ:-), æ:-, æfter-, and- (an-, on-, ond-), ante-, arce-, be- (bi-, bi:-, big-), ed- (æd-, et-, æt-, ead-, eð-), el- (æl-, ell-), fær-, for- (fore-), forð- , ful-, ge-, in-, med- (met-), mis-, of- (æf-, ef-), ofer-, on- (an-), or-, sa:m-, sam-, sin-, sub-, to:- ðurh-, un- (on-), under-, up-, ut-, wan-, wið-, wið er-,* and *ymb- (ymbe-).*

b. *-að (-oð 4), -noð, -uð, -eð), -bora , -do:m , -el (-ol, -ul, -ele, -la, -elle, -le, -l, -il), -els, -en (-n, -in), -en, -end, -ere (-era), -estre (-ystre, -istre), -et (-ett), -ha:d, -icge (-ecge, -ige), -incel, -ing (-ung), -la:c, -ling (-lung), -mæ:l, -ness (-nes, -nis, -nyss, -nys), -ræ:den, -scipe (-scype), -t (-ð, -ðo, -ðu)* and *-wist.*

Although the current research within the *Nerthus* group is analysing the endings *-a, -e, -o, -u*, as inflective derivatives (thus González Torres, 2009), I shall take here a more conservative perspective and treat them as purely inflective endings, leaving them out of the inventory of suffixes identified in this work.

Regarding the meaning of affixes, we must lay on their function as word formers to be consistent. The aim of affixation is to generate new meanings through new lexical forms. Following Martin Arista (2009) recursivity and recategorization are the two major properties of morphology in a functional structural approach to word-formation. Old English, a stage of the language where lexical creation was mainly based upon the use of already existing word, is a suitable language to prove this statement. (4) presents some cases of recursivity in which affixes are attached to bases previously derived on the same side of the structure[3]:

(4)

Prefixation of prefixed predicates: *underto:dal* 'secondary division' (*to:da:l* 'partition'); *ungeðwæ:re 2* 'disturbance' (*(ge)ðwæ:re* 'united', adjective).
Suffixation of suffixed bases: *cri:stendo:m mæ:nsumung, feohle:asnes*

But for the purpose of this research, recursivity has been understood in the widest sense of derivation, thus allowing for the study of affixation applied to previously derived words created by processes other than affixation as in (5).

[3] The bases of the derived words and their translations are offered in brackets.

(5)

Prefixation of compound predicates: *midyrfenuma* 'coheir' (*yrfenuma* 'heir, succesor'); *oferealdormann* 'chief officer' (*ealdorman* 'alderman, ruler, prince')

Suffixation of compound predicates: *fæstræ:dnes* 'constancy, fortitude' (*fæstræ:d* 'firm, constant'); *wordsnoterung* 'sophism'(*wordsnoter* 'eloquent, wise in words')

Prefixation of zero-derived predicates: *bi:cwide* 'byword, proverb' (*cwide* 'speech, saying, word'); *onspræ:c* (*spræ:c*);

Suffixation of zero-derived predicates: *gifung* 'consent' (*gif 2* 'gift, grace'); *wi:fla:c* 'cohabitation, fornication' (*wi:f* 'woman, female, lady');

The second main feature of morphology in an approach as the one here presented is recategorization, a function mainly fulfilled by suffixes as (6) shows:

(6)

a:birging 'taste'(*a:birgan* 'to taste, eat') (verb); *forwerodnes* 'old age' (*forwerod* 'worn out, very old') (adjective); *ha:rung* 'hoariness, old age'(*ha:r* 1 'hoar, hoary, grey, old') (adjective);

However, this research has also proved the existence of recategorization in the formation of prefixed predicates. I have identified a total of 23 predicates with this structure. Consider the cases in (7):

(7)

oferfyrr 'excessive distance'(*feorr* 1 'far, remote') (adjective); *oferwriten* 'superscription' (*(ge)wri:tan* 'to write') (verb); *unlanda:gende* 'not owning land' (*landa:gende* 'owning land') (adjective)

As stated above, in order to provide a functional explanation of process feeding or recursive word-formation in Old English complex nouns it is necessary to consider what the function of derivational morphology is, namely to generate new meanings by means of new forms.. As a result of a methodological decision made as a general guideline of the *Nerthus* project, the definitions of meanings will be dealt with once the derivational morphology of the language has been fully described and explained. Although insightful observations of

semantic primes have been carried out by de la Cruz Cabanillas (2007) and Guarddon Anelo (fc. a, b), for the time being, our research is based on the existing dictionaries, some of which are theoretically old-fashioned and others are not complete yet. In spite of this conditioning, the question must be answered, however preliminary, of the semantic compatibility of affixes raised by following Lieber (2004). Lieber states that new affixes can be attached to previouly derived words if they contribute additional meaning and remarks that "lack of content limits recursivity" (Lieber, 2004: 169) but she makes the provision that "repeating the same features is possible as long as the result is useful and interpretable" (Lieber, 2004: 166). In general, throughout the analysis of the derived nouns of Old English, I have found a meaning difference between the base and the derivative that motivates the activation of the derivational process. The only remarkable exception arises when zero-derivation feeds affixation. Excluding cases in which there is partial synonymy between base and derivative, in such a way that the derivative specializes in one of the meanings of the base, as in the afffixal *oferbru:* 'eye-brow' vs. the zero-derived *bru:* 'brow, eye-brow, eye-lid, eye-lash', there remain the instances of zero-derivation feeding prefixation given in (8a) and zero-derivation feeding suffixation given in (8b).

(8)

a. *onforwyrd* 'destruction' vs. *forwyrd* 'destruction, ruin, fall, death'
to:gehlytto 'fellowship, union' vs. *gehlytto* 'fellowship, lot'
æfgrynde 'abyss' vs. *grynde* 'abyss'
midhelp 'help, assistance' vs. *help* 'help, succour, aid'
anhoga 'care, anxiety' vs. *hoga 2* 'fear, care; attempt, struggle'
æfterle:an 'reward, recompense, restitution, retribution'
vs. *le:an 1* 'reward, loan, compensation, remuneration, retribution'
edle:an 'reward, retribution, recompense, requital'
vs. *le:an 1* 'reward, loan, compensation, remuneration, retribution'
ansto:r 'incense' vs. *sto:r 1* 'incense'
bi:swæc 'tripping up, treachery'
vs. *swic* 'illusion; deceit, treachery'

b. *a:gennes* 'property' vs. *a:gen* property, own country'
æ:bylgð 'anger' vs. *æ:bylg* 'anger'
æ:bylgnes 'anger, offence' vs. *æ:bylg* 'anger'

(ge)anbidung ‘waiting for, expectation; delay’
vs. *anbid* ‘waiting, expectation,’
by:sting ‘beestings’ vs. *be:ost* ‘beestings’
bebodræ:den ‘command, authority’ vs. *bebod* ‘command’
blinnes ‘cessation, intermission’ vs. *blinn* ‘cessation’
by:ing 2 ‘dwelling’ vs. *bu: 1* ‘dwelling’
gebu:nes ‘dwelling’ vs. *bu: 1* ‘dwelling’
(ge)ce:lnes ‘coolness, cool air, breeze’
vs. *ciele* ‘coolness, cold, chill, frost’
e:htnes ‘persecution’ vs. *e:ht 1* ‘pursuit’
e:htung ‘persecution’ vs. *e:ht 1* ‘pursuit’
fle:amdo:m ‘flight’ vs. *fle:am* ‘flight’
forebodung ‘prophecy’ vs. *forebod* ‘prophecy, preaching’
frihtrung ‘soothsaying, divination’ vs. *friht* ‘divination’
la:rdo:m ‘teaching, instruction’
vs. *la:r* ‘art of teaching, preaching, doctrine’
le:odscipe ‘nation, people; country, region’
vs. *le:od 2* ‘people, nation’
sæ:dnað ‘sowing’ vs. *sæ:d* ‘sowing’
trahtað ‘commentary’ vs. *traht* ‘text, passage; exposition, treatise, commentary’
wæ:dlung ‘poverty, want; begging’ vs. *wæ:dl* ‘poverty; barrenness’

These data have to be interpreted with caution for two reasons. In the first place, the degree of synonymy of the pairs given above is open to question. In the second place, and related to the first reason, shortcomings of lexicographical work cannot be completely ruled out. Despite the little evidence that has been found and the disclaimers just given, it seems that zero-derivation can feed affixation without significant meaning change. This happens more often in suffixation than in prefixation (which is a consequence of the higher figures yielded by suffixation) and with a variety of affixes.

3. CONCLUSION

Despite the difficulties in carrying out semantic studies in diachrony, corpus-based approaches as the one here presented allow for the regularization and standardization of the analytical procedures in such a way that the inconsistencies of prior proposals can be observed and a new view on semantic research can be set on the solid grounds of systematicity. This paper just remarks inconsitencies in previous approaches, while opening the doors to new and more deep semantic analyisis to be done.

This paper puts forwards the possibility for affixes, both prefixes and suffixes, to combine with all kinds of derived predicates, and remarks the ability of some Old English prefixes to create new lexical items by means of recategorization.

As regards meaning, this paper remarks the existence of affixations in which no meaning addition is observed. However, the debate whether the existence of meaningless addition of affixes responds to the shortcomings of previous studies or responds to a regular and identifiable process is yet an open question. The fact that this phenomenon occurs only with zero derived predicates seems to indicate some degree of regularity. Very tentatively, the reason why recursivity without meaning change is allowed may have to do with the decay of zero-derivation as a productive process in the derivational morphology of Old English. More work, however, is needed in this area.

REFERENCES

Bauer, L. (2007). *The Linguistics Student's Handbook.* Edimburgh: Edimburgh University Press.

Bosworth, J., and T. N. Toller. (1973 (1898)). *An Anglo-Saxon Dictionary*. Oxford: Oxford University Press.

Clark Hall, J. R. (1996 (1896)). *A Concise Anglo-Saxon Dictionary*. Toronto: University of Toronto Press.

De la Cruz Cabanillas, I. (2007). Semantic Primes in Old English: A Preliminary study of descriptors. *SELIM* 14: 37-58

González Torres, E. (2009). Affixal nouns in Old English: morphological description, multiple bases and recursivity. PhD dissertation, University of La Rioja.

Guarddon Anelo, C. Un análisis de las propiedades combinatorias de los primitivos semánticos a través de las adposiciones complejas del inglés antiguo. *Revista de la Sociedad Española de Lingüística.* Forthcoming a.

Guarddon Anelo, C. The natural semantic metalanguageof Old English compound adpositions. *English Studies*. Forthcoming b.

Kastovsky, D. (1992). Semantics and vocabulary. En R. Hogg (Ed.) *The Cambridge History of the English Language I: The Beginnings to 1066,* (pp. 290-408). Cambridge: Cambridge University Press.

Lieber, R. (2004). *Morphology and Lexical Semantics*. Cambridge: Cambridge University Press.

Mairal Usón, R., and F. Cortés Rodríguez. (2000-2001). Semantic Packaging and Syntactic Projections in Word Formation Processes: the Case of Agent Nominalizations. *RESLA* 14, 271-294.

Martín Arista, J. (2008). Unification and Separation in a Functional Theory of Morphology. En R. Van Valin (Ed.) *Investigations of the Syntax-Semantics-Pragmatics Interface,* (pp. 119-145). Amsterdam: John Benjamins.

Martín Arista, J. (2009). A Typology of Morphological Constructions. En C. Butler y J. Martín Arista (Eds.) *Deconstructing Constructions,* (pp. 85-115). Amsterdam: John Benjamins.

Martín Arista, J. *et al*. (2009). Nerthus: An Online Lexical Database of Old English. http://www.nerthusproject.com

Torre Alonso, R *et al*. (2008). Fundamentos empíricos y metodológicos de una base de datos léxica de la morfología derivativa del inglés antiguo. *Revista de lingüística y lenguas aplicadas* 3, 129-144.

Sweet, H. (1976 (1896)). *The Student´s Dictionary of Anglo-Saxon*. Cambridge: Cambridge University Press.

Método general de lematización con una gramática mínima y un diccionario óptimo. Aplicación a un corpus dialectal escrito

HIROTO UEDA
Universidad de Tokio
MARIA-PILAR PEREA
Universitat de Barcelona

Resumen
A falta de lematizadores y etiquetadores para lenguas minoritarias proponemos elaborar nuestro propio procesador programado por Microsoft Excel VBA. El proceso consiste en asignar provisionalmente la categoría de partes de oración por un diccionario preparado en procesos anteriores y desambiguar los homógrafos por una lista de reglas gramaticales, también almacenadas anteriormente, para proceder finalmente a la lematización utilizando el mismo diccionario ahora dotado de información ortográfica y gramatical (categoría).
Este método es flexible y aplicable a distintas lenguas europeas y consigue ofrecer un resultado cada vez mejor a medida que en cada operación se nutren tanto el diccionario como la gramática. Nuestra idea es crear un aparato de procesamiento común, que se activa con parámetros léxicos y gramaticales específicos de cada lengua, objeto de investigación.
Palabras clave: lematización, Microsoft Excel VBA, catalán, lengua minoritaria, diccionario, gramática
Abstract
Owing to the lack of lemmatizers and taggers for minority languages, we propose to develop our own processor programmed by Microsoft Excel VBA. The process consists in assigning temporarily the category of parts of sentences through a dictionary prepared in previous processes and disambiguating homographs using a list of grammatical rules, also stored previously, in order to finally lemmatize the text using the same dictionary now provided with spelling and grammar (category) information.
This method is flexible and applicable to different European languages and offers a better result as both the dictionary and the grammar are fed on each transaction. Our idea is to create a common processing apparatus, which is activated with language specific lexical and grammatical parameters.
Keywords: lemmatization, Microsoft Excel VBA, Catalan, minority languages, dictionary, grammar

1. INTRODUCCIÓN

Uno de los procesos fundamentales para tratar los textos lingüísticos es la lematización, que consiste en asignar una forma representativa a distintas formas concretas variables: formas conjugadas del verbo, cambios según el género y número de adjetivos y sustantivos, etc. Es necesario, a la hora de realizar unos estudios estadísticos, redactar un vocabulario o diccionario, intentar una búsqueda de información por palabras clave y analizar la combinación de elementos, entre otros objetivos de investigación.[1]

[1] Para la información general de anotaciones del texto lingüístico, véanse por ejemplo Meyer (2002) y McErney (2003). Contamos con una monografía sobre el tema de lematización del español en la obra de Gómez Díaz (2005).

Aparte de los procesadores existentes para lenguas mayoritarias, como el inglés, el español, el japonés, etc., los investigadores de lenguas minoritarias se ven obligados a efectuar la lematización manualmente. Para salvar esta dificultad, vamos a proponer un método sencillo, que podría ser aplicable a distintas lenguas europeas, junto con algunas consideraciones teóricas y prácticas. Nuestro objeto de lematización es el corpus preparado en el proyecto de edición en CD-ROM del *Bolletí del Diccionari de la Llengua Catalana* (2002-2004).[2]

2. DESCRIPCIÓN DEL CORPUS

En dialectología catalana es bien conocida la figura de Antoni M. Alcover (Manacor 1881 - Palma 1931). Gestor y redactor del *Diccionari català-valencià- balear*, emprendió prolongadas encuestas dialectales, recopiló narraciones populares y, entre otras muchas actividades, redactó casi íntegramente dos publicaciones que consiguieron una muy buena acogida en su época: la revista titulada *Bolletí del Diccionari de la Llengua Catalana* (BDLC) y el semanario *La Aurora*, editado en su villa natal, Manacor.

El *Bolletí del Diccionari de la Llengua Catalana* se considera la primera revista de carácter filológico publicada en Cataluña y en el Estado español. Alcover la creó para que se convirtiera en la tribuna de divulgación y de propaganda de su *Obra del Diccionari i de la Gramàtica*, nombre con que se conocía el proyecto de elaboración del diccionario citado anteriormente.

Bajo la dirección de Alcover, se publicaron catorce volúmenes del BDLC (entre el 1901 y el 1926), y, a pesar de que contó con colaboraciones más o menos esporádicas de distintos eruditos, la mayor parte de la redacción fue obra exclusiva del dialectólogo. La revista fue concebida inicialmente como una publicación mensual de 16 páginas, pero a menudo la puntualidad se incumplía, y el lector recibía números mucho más extensos aunque aparecieran dos o tres meses después. En general, la unión de los números de dos años ha dado lugar a un tomo (Tabla 1), excepto en el tomo V, que corresponde a la publicación de la primera excursión de Alcover a países europeos, y el XI, que agrupa exclusivamente los números publicados en 1920.

A grandes rasgos, el *BDLC*, en su primera época, contiene las informaciones siguientes:

---

[2] El proyecto fue financiado por la Conselleria d'Educació i Cultura del Govern de les Illes Balears. Véase Perea (2003, 2004).

1) de manera mayoritaria, textos, artículos y estudios del propio Alcover, que tratan, por un lado, de temas filológicos, dialectales, onomásticos, históricos, lingüísticos, y reúnen, por otro lado, numerosas reseñas bibliográficas, necrologías, los dietarios y los manifiestos;

2) textos y artículos de Francesc de B. Moll, que colabora en el *Bolletí* desde el volumen XIII, y se convierte en su editor en 1933, iniciando la segunda y última etapa de la revista, con la publicación del volumen XV;

3) artículos más o menos extensos de escritores, lingüistas y eruditos de la época;

4) artículos periodísticos publicados en la prensa de la época, que ilustran acontecimientos concretos, como la realización del Primer Congrés Internacional de la llengua catalana, en 1906, o las diversas manifestaciones del movimiento de recuperación de la lengua catalana que se llevaron a cabo a inicios del siglo XX;

5) las observaciones dialectales —las llamadas «Notes dialectals»— que enviaban algunos corresponsales y colaboradores de la *Obra del Diccionari*.

El contenido del BDLC se ha publicado en CD-ROM en dos ediciones (2003) y (2004). Y ha sido de ésta última de donde se ha seleccionado el corpus que se pretende estudiar.

Para nuestro estudio, que se centra en el corpus escrito alcoveriano, únicamente se toma en consideración el contenido del apartado (1). Las otras informaciones se omiten, puesto que no corresponden al mismo autor.

Se indica a continuación la extensión de cada tomo, los números que contiene, el año de publicación y el número de palabras, una vez que se han eliminado las informaciones no alcoverianas.

Tabla 1: Descripción formal del *Bolletí del Diccionari de la Llengua Catalana* (BDLC)

Tomo/núm.	Año	pág.	palabras	Vol.	Año	pág.	palabras
BDLC I (17)	1902-1903	587	228.518	BDLC VIII (8)	1914-1915	268+114	157.298
BDLC II (13)	1904-1905	408	138.358	BDLC IX (13)	1916-1917	384	145.240
BDLC III (9)	1906-1907	414	165.719	BDLC X (13)	1918-1919	524	217.883
BDLC IV (11)	1908-1909	405	176.599	BDLC XI (4)	1920	336	125.158
BDLC V (dietario)	1908	378	167.600	BDLC XII (6)	1921-1922	368	153.293
BDLC VI (21)	1910-1911	392	171.785	BDLC XIII (7)	1923-1924	376	168.496
BDLC VII (13)	1912-1913	436	183.627	BDLC XIV (5)	1925-1926	352	121.805

3. PROCESOS DE LEMATIZACIÓN

3.1. Texto

Los textos no presentan siempre un estado ideal para el procesamiento automático, el cual cuenta con reglas gramaticales y ortografía establecida. Ante la realidad complicada de la lengua, lo normal sería tratar tales textos "problemáticos" de forma manual para salvar las dificultades técnicas que puedan presentar algunos tratamientos automáticos. La manipulación manual es necesaria en el caso de algunos datos especiales y no lo es para procesamientos regulares. Cabe distinguir entre el procesamiento de tipo específico y el de tipo general. En esta sección intentaremos presentar nuestro método de lematización de un corpus dialectal a través de algunos experimentos reales y de los resultados obtenidos a través de procesamientos manuales y automáticos.

El Texto 1 muestra un fragmento del texto a que nos hemos referido en la introducción. Cada párrafo está situado en la celda de Excel junto con el número de identificación correspondiente en la celda izquierda. Nuestro propósito es lematizar todas las formas verbales asociándolas con la forma representativa.

22 (...) Bo es de veure que quedam amichs. Mostra a l'alemanya unes castanyetes noves ab uns grans flochs y borlins. —Això no es de Catalunya, li dich. —No, diu ell, es d'Andalusia. Si un estranger se'n vol dur res característich d'Espanya, compra unes castanyetes; si de Catalunya, una barretina y unes espardenyes. —Prou que hi ha molts d'espanyols que donen peu a formar tals judicis d'Espanya.

Texto. 1: Texto llano (un fragmento)

A partir del texto llano digitalizado podemos elaborar automáticamente una lista de frecuencia de cada voz con relación al total del texto: *de* (10.191), *y* (8.849), *a* (5.251), *la* (4.332), *l* (3.410), *el* (3.241), *d* (3.139), *un* (2.867), *hi* (2.839), *que* (2.800), *les* (2.415), *una* (2.294), *ha* (2.277), ... hasta multitud de voces de una sola ocurrencia.

Como primer intento, pensamos asignar un lema a cada voz, y aplicar la lista de correspondencia Forma - Lema al texto. Esta lista de correspondencia sería útil para evitar el trabajo repetitivo de asignación de las mismas formas de alta frecuencia. Una lista así preparada serviría también para trabajos posteriores con otros textos del mismo corpus, añadiendo las nuevas voces que aparecen en el nuevo texto. Este método sería la única solución, al carecer de un programa de etiquetado específico, que se han elaborado en mayor

envergadura para las grandes lenguas europeas.[3] Se trata de un método directo con un máximo de diccionario (lista de correspondencia) y sin gramática (categorización y reglas).

En esta ocasión, en cambio, proponemos un método ecléctico buscando un punto óptimo de colaboración entre la gramática y el diccionario. Por no poseer una gramática aplicable al procesamiento de textos, de momento contaremos con un mínimo de información gramatical: la categoría de partes de oración.

3.2. Lista de Lema-Formas

Una de las características comunes de las lenguas indoeuropeas es su morfología verbal basada en la conjugación, flexión de las formas terminales. Es decir, cada forma verbal está constituida básicamente por una parte anterior invariable (Raíz) y otra posterior variable (Terminación). Es una regla morfológica sencilla, aparte de los casos de cambios de radicales, por ejemplo, *hago, haces, hice, hecho, etc.* del verbo español *hacer*; o *faig, fas, fa, fem feu, fan* del catalán *fer*; y de supletismo, *voy, vas, va* ... del verbo español *ir* y *vaig, vas, va, anem, aneu, van* del verbo catalán *anar*. Nuestra idea es preparar provisionalmente una bolsa de infinitivos más frecuentes y, a partir de esta bolsa, asignar automáticamente un lema a las nuevas formas que aparecen en el texto eligiendo la forma más parecida posible. Por ejemplo, la voz *abeuren* debe corresponder al lema *abeurar* 'abrevar' por tener una parte común *abeur*.

Con tal de que esté preparada una buena lista de correspondencia Raíz - Lema, en general la mayoría de las veces una forma se identifica con su lema correcto. Ahora bien, nuestro trabajo manual ya no es asignar uno por uno el lema correspondiente a la nueva forma que aparece en el nuevo texto, sino simplemente registrar la nueva forma de la raíz en la lista de raíces:

[3] Veánse los sitios de Brill's Tagger: http://www.tech.plym.ac.uk/soc/staff/guidbugm/pysoftware.htm; CLAWS: http://ucrel.lancs.ac.uk/claws/
TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/; de los cuales Brill's Tagger y CLAWS son etiquetadores específicos de inglés, mientras que Tree Tagger permite uso de varias lenguas europeas.

LEMA	Cat.	Otras formas
a	P	
això	M	
alemany	S+A	
amb	P	
amich	S	
anar	V	vaig, vas, va, an, vag
Andalusia	E	
aprendre	V	aprend
aqueix	M	aquexos, aqueixa, aqueixes

Figura 1: Lista de Lema-Formas

La Figura 2 muestra una parte del resultado del análisis del Texto:

&	LEMA
a	_P_a
ab	@
això	_X_això
alemanya	_A_alemany
amichs	_S_amich
andalusia	_E_Andalusia
barretina	_S_barretina
bo	_A_bo
borlins	_A_bo

Figura 2: Lista del resultado

donde observamos que *amichs* se ha identificado correctamente con el lema *amich*,[4] aunque la forma plural no figure en la lista. Las palabras que no han sido analizadas por falta de los datos del Diccionario aparecen con el signo de arroba (@), lo cual quiere decir que se debería asignar la categoría y lema en el Diccionario. El caso de "borlins", asignado equivocadamente con "A_bo", se debe a que el programa ha buscado el adjetivo "bo" como último recurso. Con la inclusión de *borlin* en el Diccionario se resuelve este problema.[5]

Nuestro programa también puede ofrecer un texto de resultado. En este caso nos limitaremos a presentar la asignación automática de la categoría gramatical.[6]

[4] La grafia <amich> corresponde a <amic> del catalán estándar.
[5] El caso anterior de *ab* @ es distinto. No se identifica con a (P_a), puesto que la preposición exige una identificación total, mientras que las palabras variables (verbos, adjetivos, sustantivos) permiten una identificación parcial. La identificación de *borlins* con *bo* podría evitarse asignando algunas condiciones especiales de la terminación de adjetivos. De momento, sin embargo, seguimos trabajando sin especificaciones de condiciones específicas para conservar la característica generalizadora del programa, aplicable a distintas lenguas.
[6] Los signos utilizados son: A: Adjetivo, C: Conjunción, D: Adverbio, I: Interjección, M: Demostrativo, N: Numeral, O: Posesivo, P: Preposición, R: Relativo, S: Sustantivo, T: Artículo, V: Verbo, X: Pronombre. Naturalmente estos signos se pueden cambiar según la conveniencia del usuario.

22 (...) Bo_A es_T_V_X de_P veure_V que_C quedam_V amichs_S. Mostra_S_V a_P l_T_X'alemanya_A unes_T castanyetes_S noves_A ab @uns_T grans_A flochs_S y_C borlins_A. —Això_X no_D es_T_V_X de_P Catalunya_E, li_X dich_V. —No_D, diu_V ell_X, es_T_V_X d_P'Andalusia_E. [...]

Texto. 2: Texto asignado de Categorías

3.3. Casos ambiguos y desambiguación

Al observar el Texto 2, nos damos cuenta de que algunas Formas poseen más de una asignación, por ejemplo, "es_T_V_X", "l_T_X", "alemanya_S_A", etc. Se trata de casos ambiguos respecto a la categoría, puesto que se ha hecho a partir de la lista de formas, sin tener en cuenta su sintaxis. Precisamente para poder ofrecer estos casos ambiguos, hemos introducido el procesamiento de categorización. El mérito de asignación de categoría gramatical es su capacidad de distinguir entre varios homógrafos: *trabajo* como sustantivo y *trabajo* como primera persona singular del verbo *trabajar* en español; y *sebre* 'saber' (sustantivo) y *sebre* 'saber' (verbo) en catalán dialectal. La mayoría de las veces se distinguen por la categoría, por ejemplo S(ustantivo) y V(erbo): trabajo_S, trabajo_V. A partir de esta asignación por "V", podemos proceder a la lematización verbal, excluyendo los casos de "S" (sustantivo).

Para distinguir entre T (artículo) y X (pronombre), contamos con la información sintáctica siguiente:[7]

1) Delante de un sustantivo (S), _A_X (adjetivo / pronombre) debe convertirse en _A (adjetivo), por ejemplo: tals_A_X judicis_S

2) Delante de un sustantivo (S), _T_X (artículo / pronombre) debe convertirse en _T (artículo), por ejemplo: sebre'l_T_X castellà_S

3) Delante de un verbo (V), TX (artículo / pronombre) debe convertirse en _X (pronombre), por ejemplo: per que el_X vejen_V

Para estas Reglas gramaticales, se elaboran las siguientes fórmulas, que se basan en las Expresiones Regulares:[8]

[7] Sin recurrir a las reglas gramaticales, se podría solucionar el problema de ambigüedad por medidas estadísticas, que consisten en buscar la mayor probabilidad posible de secuencias de tres elementos (trigramas) extraídos de textos anotados. Véase Voutilainen (2003).

[8] Estas fórmulas están basadas principalmente en la versión de Expresiones Regulares de Microsoft VBScript, con algunas modificaciones simplificadoras.

1) (&)_A_X(@&_S)=>1_A2

2) (&)_T_X(@&_S)=>1_T2

3) (&)_T_X(@&_V)=>1_X2

donde "&" representa una secuencia de letras utilizadas en las palabras, "@" es una secuencia de letras no utilizadas en las palabras, $1 corresponde a la secuencia de letras entre la primera paréntesis (&) y $2, a la de la segunda paréntesis (@&_S). El signo de "=>" significa que la fórmula izquierda se convierte en la fórmula derecha.

Estas asignaciones se almacenan en la Lista de Reglas, que se utiliza cada vez que se obtiene un texto ambiguo. El resultado es:

22 (...) Bo_A es_V de_P veure_V que_C quedam_V amichs_S. Mostra_V a_P l_T' alemanya_A unes_T castanyetes_S noves_A ab_P uns_T grans_A flochs_N y_C borlins_S. —Això_M no_D es_V de_P Catalunya_E, li_X dich_V. —No_D, diu_V ell_X, es_V d_P' Andalusia_E. [...]

Texto. 3: Texto desambiguado

3.4. Lematización

Una vez desambiguado el texto por varias fórmulas gramaticales, se procede finalmente a su lematización. La lematización consiste en asignar el lema correspondiente único. Con la información de la forma de la voz y su categoría gramatical, se identifica con su lema correspondiente, con alta precisión.

22 (...) Bo_A_bo es_V_ésser de_P_de veure_V_veure que_C_que quedam_V_quedar amichs_S_amich. Mostra_V_mostrar a_P_a l_T_el' alemanya_A_alemany unes_T_un castanyetes_S_castanya noves_A_nova ab_P uns_T_un grans_A_gran flochs_N y_C_i borlins_S. —Això_M_això no_D_no es_V_ésser de_P_de Catalunya_E_Catalunya, li_X_li dich_V_dir. —No_D_no, diu_V_dir ell_X_ell, es_V_ésser d_P_de' Andalusia_E_Andalusia. [...]

Texto. 4: Texto lematizado

El proceso general de lematización de un texto consiste en: (1) Primera categorización gramatical automática; (2) Corrección manual de errores del Diccionario; (3) Segunda categorización gramatical automática; (4) Corrección manual de errores de la Gramática; (5) Lematización automática.

En el método propuesto se combina el procesamiento automático y la asignación manual. En cuanto al procesamiento automático, hemos elaborado un programa general que podría ser aplicable a distintos idiomas flexivos de terminación, de tipo indoeuropeo. La diferencia consiste en la formación del Diccionario y en sus propias reglas de la Gramática. La tarea del investigador se divide en dos partes: preparación de un Diccionario óptimo y preparación de una Gramática mínima. Cuánto menor es la cantidad de elementos en el Diccionario, la búsqueda resulta más rápida. En la recopilación de la Gramática, se tiene que tener en cuenta no solamente la formulación adecuada, sino también su orden de aplicación. Aquí también se desea la cantidad mínima de las reglas, para lo cual se debería estar versado tanto en la sintaxis de la lengua como en el manejo de Expresiones Regulares. Lo ideal sería que colaboraran un lingüista y un especialista de ciencias informáticas.

Existen varios proyectos de lematizadores y etiquetadores en el mundo académico de la lingüística de corpus. Algunos son aplicables a solo una lengua y otros, aplicables a varios idiomas, alcanzan apenas 90 o 95% de precisión. Los programas se realizan de manera independiente de nuestro ámbito de trabajo. El método que hemos propuesto difiere de los anteriores en los puntos siguientes: (1) es aplicable a múltiples lenguas europeas; (2) se asciende el grado de precisión a medida que se nutren tanto el Diccionario como la Gramática;[9] (3) los procesos están programados para Microsoft Excel, nuestro ámbito de trabajo de siempre, de modo que hay una buena continuación del lugar de trabajo y de los textos tratados de cantidad relativamente reducida.[10]

4. EXPERIMENTO Y RESULTADO

En cuanto a la labor de identificación, en teoría se supone que la lematización por categorías resulta más económica que la de por formas. El proceso mismo de la lematización serviría como una prueba de esta hipótesis. Ahora nos preguntamos qué grado de diferencia existe entre los dos métodos. Para realizar el experimento de comprobación utilizamos dos textos del volumen 5 de BDLC: uno de texto llano donde aparecen voces diferentes, y otro de texto lematizado donde aparecen lemas unificados de las voces. Si nos fijamos en las formas

[9] El estado del resultado es flexible dependiendo de la cantidad de información disponible y de la buena constitución de las reglas. Consideramos que la flexibilidad de uso que permiten los programas es importante a la hora de aplicar a distintas lenguas y a distintos objetivos de investigación. Nuestro método no exige unas etiquetas previamente establecidas ni reglas gramaticales incorporadas, sino que es adaptable a las condiciones del usuario. Para las directrices diferentes del programa, véase Mueller (2009).

[10] Cf. Tabla 1. Para la utilización de programas codificados en macro de Excel, véase Ueda (2005), donde hemos explicado las funciones de nuestro sistema SIAL (Sistema Integral para Análisis Lingüísticos).

verbales, en el corpus aparecen 19.703 voces en total que se distribuyen de la manera siguiente:

Tabla 2: Voces y lemas

	Voces	Formas	Lemas
Cantidad total	19.703	2.625	822
Valor máximo	2.277	2.277	3.512

En la Tabla 2 observamos que hay mucha más cantidad en voces unificadas que en lemas. Si se trabaja manualmente con formas (2.277), el coste de la labor en el tratamiento es tres veces mayor grande que el de lemas (822). Es impensable trabajar con las voces concretas que aparecen en el Texto (19.703).

La cantidad máxima en las voces corresponde a la forma *ha*, que posee el valor de 2.277, mientras que en la de los lemas es del verbo *haver*, 3.512. Las formas y los lemas subsiguientes son los que se presentan en las Figuras siguientes:

V,C,L:Voz	B-05	Acum.
ha	2277	2277
es	1908	4185
son	698	4883
he	636	5519
fa	591	6110
té	383	6493
som	337	6830
veure	307	7137
està	246	7383
fan	217	7600

Figura 3: Voces

V,C,L:LEMA	B-05 (2)	Acum
haver	3512	3512
ésser	3483	6995
fer	1260	8255
tenir	730	8985
anar	657	9642
veure	625	10267
estar	475	10742
dir	410	11152
trobar	398	11550
poder	289	11839

Figura 4: Lemas

Para observar la curva de las cantidades acumuladas de los elementos en orden decreciente, hemos elaborado la Figura 5:

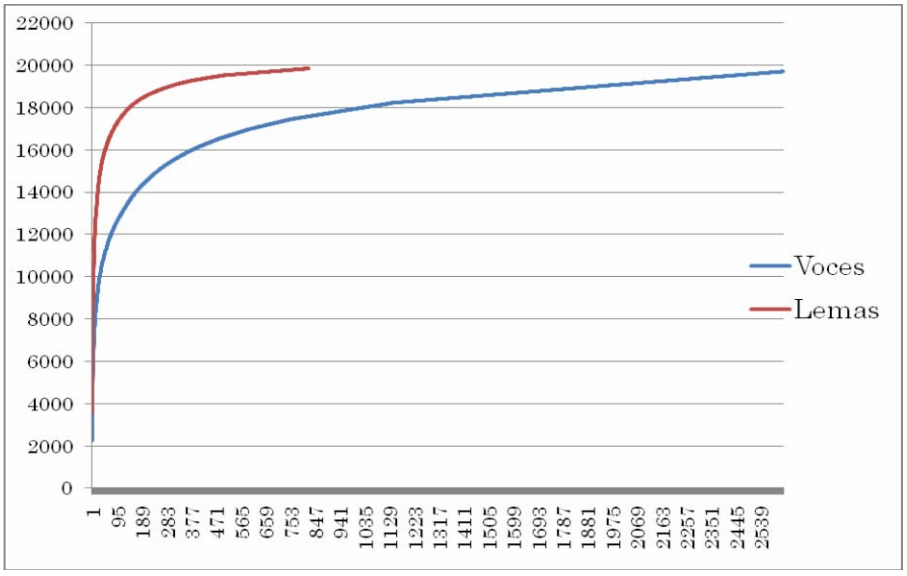


Figura 5: Voces y Lemas

En esta Figura se observa que la curva ascendiente de lemas es considerablemente más destacable que la de las voces. En la práctica de lematización, se supone que en los primeros textos tratados aparecen las voces frecuentes. Una vez registrados las voces y sus lemas en el Diccionario, el coste de los trabajos siguientes se reduce notablemente. Lo mismo puede decirse de la alimentación de la lista de Reglas gramaticales, puesto que por ser gramatical se espera que la lista sea aplicable a otros contextos en general.

5. CONCLUSIÓN

En nuestro primer intento de lematización de un texto dialectal catalán, hemos comprobado la eficacia del método con un Diccionario y una Gramática, lo mismo que se hace en el ámbito de enseñanza-aprendizaje de lenguas extranjeras en la escuela tradicional. Los nuevos métodos comunicativos, por otra parte, son más prácticos que teóricos: se enseña y se aprende a través de distintas actividades lingüísticas. Esta tendencia también se presenta en el tratamiento de datos lingüísticos, que consiste en reunir los textos anotados y almacenar informaciones estadísticas para aplicarlas a los nuevos textos.

Los dos métodos tienen sus méritos y desventajas. En el método teórico tradicional se busca la exhaustividad de tratamiento, donde se exige la preparación de un buen Diccionario y de una Gramática infalible. En el método práctico de las últimas tendencias, se busca la mayor aplicabilidad posible con un coste relativamente bajo. Se consideraría satisfactorio que el resultado llegase al nivel de más del 95% de respuestas acertadas. Diríamos que el primero

sería trabajo de lingüista, mientras que el segundo, trabajo de especialista de ingeniería informática.[11]

En la lingüística se persiguen la precisión, la concisión y la exhaustividad. En la descripción de una lengua, para llegar al nivel deseado, se tiene en cuenta el estado tanto del Diccionario como de la Gramática[12] y no se consideraría satisfactoria una precisión de un 95%. La cuestión de tratamiento lingüístico del texto no obstante no es preparar un diccionario gigantesco ni una gramática escrupulosa sin necesidad. Nuestra propuesta es buscar un punto equilibrado donde colaboren un diccionario óptimo y una gramática mínima.

REFERENCIAS BIBLIOGRÁFICAS

Chrupała, G. (2006). "Simple data-driven context-sensitive lemmatization". *Proceedings of SEPLN*. http://www.sepln.org/revistaSEPLN/revista/37/16.pdf (2010/ 2/8)

Gómez Díaz, R. (2005). *La lematización en español: una aplicación para la recuperación de información.* Gijón: Ediciones Trea.

Guirao, J. M. & Moreno-Sandoval, A. (2004) "A "toolbox" for tagging the Spanish C-ORAL-ROM corpus". *IV International Conference on Language Resources and Evaluation (LREC2004) Proceedings.* http://lablita.dit.unifi.it/coralrom/papers/ toolbox-final.pdf (2010/2/8).

Loftsson, H. (2008). "Tagging Icelandic text: A linguistic rule-based approach". *Nordic Journal of Linguistics*, 31.p.1-29. http://www.ru.is/faculty/hrafn/Papers/ IceTagger_final.pdf (2010/2/8).

McEnery, T. (2003). "Corpus linguistics". En R. Mitkov (ed.), *The Oxford Handbook of Computational linguistics, (*pp. 448-463). Oxford: Oxford University Press

[11] Megyesi (2002) se refiere al mérito del etiquetador estadístico derivado de datos suecos, mientras que según el experimento que ha realizado Loftsson (2008) con los textos islándicos, el etiquetador basado en reglas gramaticales presenta mayor eficacia que otros programas de carácter estadístico. Samuelsson y Voutilainen (1997) informan que el programa basado en reglas presenta menos errores que el basado en estadística. Chrupała (2006) propone la combinación de los dos métodos, lo cual está realizado en el proyecto de Guirao y Moreno-Sandoval (2004).

[12] En esta ocasión no hemos tratado la cuestión morfológica, limitándonos a presentar unos ejemplos de reglas sintácticas. Somos conscientes de que es necesario incluir en la Gramática la información morfológica: terminaciones verbales y adjetivales, singular y plural de los sustantivos, sufijos adverbiales, etc. Para la información morfográfica que se toma en consideración, véanse Plissen et al. (2004) y O'Donovan y Troussov (2003).

Mueller, M. (2009). “NUPOS: A part of speech tag set for written English from Chaucer to the present”, http://panini.northwestern.edu/mmueller/nupos.pdf.

Meyer, C. F. (2002). *English corpus linguistics. An introduction.* Cambridge: Cambridge University Press.

Megyesi, B. (2002). “Shallow Parsing with PoS Taggers and Linguistic Features”, *Journal of Machine Learning Research*, 2, 639-668. http://jmlr.csail.mit.edu/papers/volume2/megyesi02a/megyesi02a.pdf (2010/2/10).

O’Donovan, B. & Troussov, A. (2003). “Morphosyntactic annotation and lemmatization based on the finite-state dictionary of wordformation elements”. *Proceeding of the International Conference Speech and Computer*, Moscow, Russia. http://www.iol.ie/~bodonovan/pubs/SPECOM-03.pdf.

Perea, M.-P. (ed.) (2003). *Bolletí del Diccionari de la Llengua Catalana*. Palma: Conselleria d’Educació i Cultura. Govern de les Illes Balears. CD-ROM edition.

Perea, M.-P. (ed.) (2004). *Bolletí del Diccionari de la Llengua Catalana* nova edició ampliada amb índexs). Palma: Conselleria d’Educació i Cultura. Govern de les Illes Balears. CD-ROM edition.

Plisson, J., Lavrac, N. & Mladenic, D. (2004). “A Rule based Approach to Word Lemmatization”. *Proceedings of the 7th International Multi-Conference Information Society*, 1(1): 83-86. http://eprints.pascal-network.org/archive/00000715/01/Pillson-Lematization.pdf (2010/2/8).

Samuelsson, C. & Voutilainen, A. (1997). “Comparing a Linguistic and a Stochastic Tagger”. *8th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference,* (pp. 246-253). Madrid: UNED. http://www.aclweb.org/anthology/P/P97/P97-1032.pdf (2010/2/10).

Siemens, R. G. (1996). “Lemmatization and parsing with TACT preprocessing programs”. *Computing in the Humanities Working Papers*. http://www.chass.toronto.edu/epc/chwp/siemens2/index.html (2010/2/10).

Ueda, H. (2005). “Methods of ‘hand-made’ corpus linguistics - A bilingual data base and the programming of analyzers". *Usage-Based Linguistic Informatics 1, Linguistic Informatics -State of the Art and the Future*, (pp. 145-166). John Benjamins Publishing Company.

Voutilainen, A. (2003). “Part-of-speech tagging”. In R. Mitkov (ed.) *The Oxford Handbook of Computational linguistics*, (pp. 219-232). Oxford University Press.

Doublets and nominalization in Early Modern scientific English

VERA VÁZQUEZ LÓPEZ

Universidade de Santiago de Compostela

Abstract

The main aim of this corpus-based study is to analyze the possible structural differences between doublets formed by action nouns that have the same verbal root but a different suffix, such as declaring *and* declaration*, both deriving from the verb* declare*. The nominalizations analyzed will be those ending in the Romance suffixes* -al, -age, -ance, -(a)tion/-sion, -ment, *and* -ure *and the native suffix* -ing*. The scope of the analysis will be Early Modern scientific English, since there is a preference for the use of nominalizations in this genre and it is in this period when English becomes one of the main languages of science. The working assumption is that the new words needed to describe scientific concepts will be obtained not only by borrowing, but also by resorting to methods of word formation such as affixation. This being the case, it is conceivable that doublets will flourish during the period under analysis.*

Keywords: Doublets, nominalization, suffixation

Resumen

Este estudio de corpus tiene por objeto analizar las posibles diferencias estructurales que puedan existir en los dobletes formados por nombres de acción que comparten una misma raíz verbal, pero contienen un sufijo diferente, como declaring *y* declaration*, ambos derivados del verbo* declare*. Los sufijos analizados son* -al, -age, -ance, -(a)tion/-sion, -ment *y* -ure*, de origen romance, y* -ing, de origen nativo*. El estudio se centra en el inglés científico del período Moderno Temprano debido a que el uso de nominalizaciones en este género es muy frecuente y es precisamente en este período cuando el inglés se convierte en una de las lenguas principales de transmisión del conocimiento científico. La hipótesis de trabajo es que los nuevos términos necesarios para referirse a conceptos científicos no sólo se obtendrían mediante préstamos, sino también por métodos de formación de palabras como la afijación. En este contexto, es probable que los dobletes prosperaran en este período.*

Palabras clave: Dobletes, nominalización, sufijación

1. INTRODUCTION[1]

Complete synonymy in language is said not to occur because one of the basic characteristics of language is the preference for economy; that is, two ways of referring to the same thing are usually avoided. However, when evidence from real language in use is considered, things are not always so clear-cut. In the domain of nominalizations,[2] it is common, for example, to find

[1] For generous financial support, I am grateful to the following institutions: Spanish Ministry for Education (grant AP2007-04509), Spanish Ministry for Science and Innovation and European Regional Development Fund (grant HUM2007-60706), Autonomous Government of Galicia (grants 2008-047 and INCITE-08PXIB204016PR). Many thanks to Teresa Fanego and María José López-Couso for valuable comments and feedback on an earlier version of this paper.

[2] The term 'nominalization' is used here in a broad sense, to subsume both nominalizations proper, that is gerundial (*Liking her is impossible*), infinitival (*To like her is impossible*), and *that*-clauses (*That I liker her is a fact*) which occur in clause structure in slots usually filled by ordinary nouns, and the kind of related formation

doublets[3] deriving from the same verbal root, as is the case of *declaring* and *declaration*, both from the verb *declare.* As can be seen in instances (1a) and (1b) below, there is no apparent difference in meaning, both members of the doublet referring to the action of the base verb.

> (1a) E1 1548 Vicary *Anatomie*, 13.23. Nowe the thirde way to knowe what thing Chirurgerie is, It is also to be knowen by his beeing or *declaring*[4] of his owne properties, […]
>
> (1b) E1 1551 Record *Geometrie* 2FR.323. your eye may iudg without muche *declaracion*, so that I shall not neede to make more exposition therof, […]

Since complete synonymy is thought not to be permissible, the aim of this paper is analyzing the possible structural differences between doublets that have led to the coexistence of both kinds of nominalizations. The focus of the analysis is scientific English of the Early Modern English (EModE) period, both because there is a preference here for the use of nominalizations in scientific texts (Biber, 1988: 107, 128), and because EModE was the period when English started to develop in its role as one of the principal languages of science (Atkinson, 1999). Thus, the demand for new words to describe concepts at this time would lead not only to the borrowing of new terms but also to the creation of new ones, using the resources of the language such as affixation. It is not surprising that this being the state of things, the use of doublets flourished in this period.

2. THE CORPUS AND ITS ANALYSIS

The corpus chosen for this study is *The Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME; cf. Kroch, Santorini and Delfs, 2004). It contains roughly 1.8 million words subdivided into three directories, Helsinki, Penn 1 and Penn 2. The Helsinki directory consists of about 573,000 words. Both Penn 1 and Penn 2 are supplements to the Helsinki

---

that Chomsky (1970:188) labelled *derived nominals*, such as *destruction* in *The destruction of the city was a disaster*.

[3] The term *doublet* is usually applied to grouping two elements of the same category. However, in this study, it will be used in a wider sense, since, as will be seen in section 3.2, some of the *-ing* formations are verbal in nature.

[4] An anonymous reviewer points out that the choice of *declaring* in this example might be related to the fact that it is cooordinated to another *–ing* form for which there is no possible Romance alternative. It is possible, of course, that a factor such as this can account for the choice of a gerund insteand of a Romance suffix in some cases, this is an aspect whose investigation will be left for future research.

Corpus (see Kytö, 1996), and whenever possible they consist of samples taken from the same authors and editions as the material used in the Helsinki Corpus. Penn 1 and Penn 2 contain roughly 615,000 and 606,000 words respectively. However, for this study, only the Helsinki and the Penn 1 directories have been analyzed, as the number of words was judged to be adequate for our purposes.

The texts in the PPCEME belong to three different subperiods: E1 encompasses the period 1500 to 1570; E2 covers 1570 to 1640, and E3 from 1640 to 1710. All three subperiods contain samples of eighteen genres, namely Bible, Biography/ Autobiography, Biography (other), Diary (private), Drama/Comedy, Educational Treatises, Fiction, Handbooks (other), History, Law, Letters (non-private), Letters (private), Philosophy, Proceedings/Trials, Science/Medicine, Science (other), Sermons and Travelogue. However, for the present study, the only genres analyzed are Science/Medicine and Science (other). General information about the texts used is presented in Table 1.

Table 1: Number of words analyzed by subcorpus, text and time period

PERIOD	AUTHOR	HELSINKI	PENN 1	TOTAL
E1 (1500-1570)	Vicary	6,280	6,779	27,033
	Record	6,768	7,206	
E2 (1570-1640)	Blundevile	6,573	7,190	28,747
	Clowes	7,330	7,654	
E3 (1640-1710)	Hooke	6,381	7,275	23,419
	Boyle	5,423	4,340	
	TOTAL	38,755	40,444	79,199

As can be seen in Table 1, a total of six texts were analyzed, two per subperiod. These works belong either to physical or medical sciences. The texts from E1 (1500 to 1570) are *The anatomie of the bodie of man* by Thomas Vicary, a medical text dating from 1548, and *The path-way to knowledge, containing the first principles of geometrie* by Robert Record, a work on Geometry of 1551. Texts from E2 are William Clowes's 1602 medical treatise, *Treatise for the artificiall cure of struma*, and two works on Geometry and Cosmography by Thomas Blundevile, *A briefe description of the tables of the three speciall right lines belonging to a circle, called signes*, *lines tangent, and lines secant* (1593) and *A plaine Treatise of the first principles of Cosmographie, and specially of the Spheare, representing the shape of the whole world* (1594). The E3 period is represented by Robert Hooke's *Micrographia* (1665) and Robert Boyle's *Electricity & Magnetism* (1675-1676).

Searches were made using the WordSmith programme. The Concordance, one of its specific tools, enables the creation of a list of items sought in their contexts. Thus, not only

can information on frequency of occurrence be obtained, but also the collocates of an item. In order to identify doublets, searches were made using suffixes. This process is carried out by introducing as search-word the intended suffix preceded by a wildcard (e.g. **ing*). It must be noted that in EModE no fixed standard yet existed, and variation in the spelling of words was common. Therefore, all the possible spelling variants for a suffix (according to the *Oxford English Dictionary* (OED)) were searched for. Thus, for instance, the suffix *-ing* may also be spelt as *-yng(e), -enge* or *-inge*. Results of the search were then edited manually; in the case of searching for the suffix *-ing*, for example, forms like *standing*, a participle, and nouns such as *thing* and *king*, had to be removed.

3. RESULTS

The main aim of this study is to analyze doublets formed by action nouns that have the same base but a different suffix. The suffixes analyzed were the Romance *-al, -age, -ance, -(a)tion/-sion, -ment,* and *-ure* and the native suffix *-ing.*

After consulting the OED, it is clear that there is a tendency for Romance nominalizations to be used as both action and result nominals,[5] whereas in the case of *-ing* nominalizations the stronger tendency is for them to be used only as action nouns. However, this shade of meaning cannot account for the existence of the doublets considered in the current study because, as noted above, only action nouns were taken into account. Therefore, aspects such as their chronology, frequency, and their dependents were assessed to try to account for the coexistence of both nominalizations.

3.1. Data obtained

21 doublets were found in the corpus; in all cases, both for the formations in *-ing* and the formations with a Romance suffix, the bases were Romance. When looking at the Romance suffixes used, only three of the suffixes searched for were found in the material, namely, *-(a)tion, -ment* and *-ance.* Of these three, *-(a)tion* is by far the most frequent since it was used 19 times (95% of the cases), while *-ment* and *-ance* are used only once (2.5% each). The complete list of doublets found follows here.

[5] Action nouns are those which do not refer to concrete entities but to actions, having "an associated event structure" (Grimshaw, 1990: 49), e.g. *The constant assignment of unsolvable problems is to be avoided* (Grimshaw, 1990: 50), while result nominals are those naming the "output of a process or an element associated with the process" (Grimshaw, 1990: 49), such as for instance *The assignment is to be avoided* (Grimshaw, 1990: 50).

Table 2: Doublets listed by alphabetical order; frequency between brackets.

-ING	ROMANCE
Accomplishing (2)	Accomplishment (1)
Acting (1)	Action (7)
Adding (2)	Addition (1)
Agitating (1)	Agitation (3)
Applying (2)	Application (3)
Attracting (3)	Attraction (10)
Concluding (1)	Conclusion (4)
Considering (1)	Consideration (3)
Declaring (1)	Declaration (1)
Describing (1)	Description (1)
Directing (1)	Direction (2)
Distilling (1)	Distillation (1)
Emitting (2)	Emission (2)
Evaporating (1)	Evaporation (1)
Examining (1)	Examination (1)
Explicating (1)	Explication (1)
Judging (1)	Judgment (1)
Mentioning (1)	Mention (3)
Moving (13)	Motion (3)
Performing (2)	Performance (2)
Publishing (1)	Publication (1)

3.2. Breakdown of the data: nominal versus verbal nominalizations

Whereas all Romance nominalizations have a nominal nature, not all *-ing* nominalizations can be considered as nouns, since as is well known, some of the *-ing* formations or 'gerunds', started to acquire verbal features since Middle English times (Fanego, 1996: 120). Thus, *-ing* nominalizations will be classified as nominal or verbal[6] attending to their internal syntax: if they are accompanied by determiners, possessives, adjectives or prepositional phrases (henceforth PPs), they will be considered nominal (cf. 2a). If followed by bare noun phrases (henceforth NPs) or showing adverbial modification,[7] they will be considered verbal (cf. 2b). If there are no elements indicating the nature of the *-ing* nominalization or the indicators are ambiguous, these nominalizations will be considered as structurally ambiguous (cf. 2c), being excluded from the analysis.

[6] An anonymous reviewer inquires whether there are any hybrid *–ing* forms with both nominal and verbal features. They do not occur in the material examined for this paper, but examples can indeed be found in this period.

[7] Since locative and temporal adverbials can also appear with nouns in Present Day English (Jack, 1988: 56-58), they cannot be considered as indicators of the verbal nature of the *-ing* nominalizations. Thus, *-ing* formations having no other syntactic indicators of their nature will be considered as ambiguous in this study.

(2a) E1 1548 Vicary *Anatomie* 59.191. One is, that one cote is not sufficient nor able to withstande the violent *mouing* and steering of the spirite of lyfe [...]

(2b) E3 Boyle 6M.51. Haking a very small fragment of a Loadstone, I found, agreeable to my conjecture, that by *applying* sometimes one Pole, sometimes the other, [...]

(2c) E2 1602 Clowes *Treatise* 17.124. [...] and so to open the powers of the skinne by *euaporating*, breathing and scattering abroad, and make thinne the grosse matter and Phlegme.

As can be seen in Table 3 below, the decrease in the number of nominal *-ing* formations, from a normalized frequency[8] of 2.58 in E1 to 0.85 in E3, is offset by the increase of verbal ones, rising from 0.36 in E1 to 7.25 in E3, as a consequence of the verbalization process of the gerund mentioned above.

Table 3: Nature of *-ing* nominalizations per period, normalized frequency per 10,000 words between brackets

	E1	E2	E3
NOUNS	7 (2.58)	4 (1.39)	2 (0.85)
VERBS	1 (0.36)	2 (0.69)	17 (7.25)
AMBIGUOUS	3 (1.10)	1 (0.34)	3 (1.28)

3.3. Chronology

Table 4 shows the first appearance of the earlier member of the doublet. Thus, when the earlier formation appearing is the *-ing* nominalization, it is reflected in the second column (*-ing* nominals). If the earlier formation is the Romance nominalization, it appears in the third column. Finally, if both members appear during the same period, they are grouped in the fourth column under the label *both*.

Table 4: Date of appearance and origin of the first element of the doublet.

	-ing nominals	Romance	Both	TOTAL
E1	1 (12.5%)	5 (62.5%)	2 (25%)	8 (38.1%)
E2	1 (12.5%)	4 (50%)	3 (37.5%)	8 (38.1%)
E3	-	-	5 (100%)	5 (23.8%)
TOTAL	2	9	10	21

[8] The normalized frequency has been calculated as follows: dividing the number of occurrences of a particular feature by the total amount of words in the text, and multiplying the result by 10,000 (Biber, 1988: 14, fn 3; 75-76).

These results indicate that most of the earlier elements of the doublet are already introduced in E1 and E2 (38.1% in each period). In these periods, it is the Romance nominalization that appears first (62.5% in E1 and 50% in E2). In all the instances found in E3, both members of the doublet appear at the same time.

This seems to indicate that, generally, Romance nominalizations are borrowed as full words into English. Later, when they are assimilated into the language, their bases are used to create new words using native suffixes. It is at this point that syntactic aspects should be analyzed, so as to understand why these new hybrid[9] formations, apparently synonyms of already existing words, are needed in the language.

3.4. Constituents of each nominalization phrase

In order to give an account of possible differences in the grammar of the two elements of the doublets, they were then classified according to the type of constituents with which they usually collocate, that is, according to the structure of the phrase as a whole. Thus, determiners (*the*), possessives, possessive phrases, adjectives and adverbs that may appear before the nominal will be considered pre-head dependents. Post-head dependents will be all those PPs, NPs and adverbs acting as complements or adjuncts to the head.

In the corpus, phrases having a nominalization as their head show a variety of constituents, such as TYPE 1 structure, having the nominal head as the only constituent (cf. [3a]); TYPE 2 structure, having just pre-head dependents, as in [3b]; TYPE 3, showing post-head dependents only (cf. [3c]); and TYPE 4, having both pre- and post-head constituents (cf. [3d]):

(3a) E2 1675-6 Boyle *Electricity* 24E.81. […] Amber it self is wont to do before it be committed to *Distillation*.

(3b) E2 1602 Clowes *Treatise* 19.133. […] there is but few mens labors at the first made so perfect, but that in processe of time & further *consideration*, they may be bettered, corrected and amended.

(3c) E3 1675-6 Boyle *Electricity* 24E.81. […] so that they who believe the virtue of *attracting* light Bodies to flow from the substantial form of Amber, […]

(3d) E3 1675-6 Boyle *Electricity* 29E.100. […] or whether the Effect were not due rather to the *Emission* and Retraction of Effluvia, which being of a viscous nature may consist of Particles […]

[9] A hybrid is here understood as a nominalization whose base and suffix have a different origin. Take, for instance, *describing*, its base being Romance and its suffix Germanic.

Tables 5, 6 and 7 show the distribution of the dependents collocating with the different kinds of nominalization along the diachronic dimension.

First of all, it must be noted that in Tables 5 and 6 there is no instance of TYPE 1 *-ing* nominalizations. Their ambiguous nature (cf. section 3.2 above) makes it impossible to classify them either as verbal or nominal and that is why they were excluded from the tables. However, the data show that there was a decrease in this type of pattern, from a normalized frequency of 0.73 in E1 to 0.42 in E3.

As shown in Table 5, verbal *-ing* formations are mostly used in structures of TYPE 3, increasing from 0.36 in E1 to 6.40 in E3. This pattern is strongly reminiscent of that of a clause in which the verb is followed by its object as in, for instance, *That matter attracts light bodies.*

Both *-ing* formations of a nominal nature and Romance nominalizations follow a similar development (see Tables 6 and 7). Thus, TYPE 4 *-ing* nominalizations are the most frequent, although there is a reduction from 1.84 in E1 to 0.42 in E3. Such a decrease is related to the general reduction in the total amount of *-ing* formations having a nominal nature. In the case of Romance nominalizations, TYPE 4 pattern is also quite frequent, although they show a stronger preference for TYPE 2. The nominal nature of these two kinds of nominalization is indicated by the fact that they are preferably preceded by determiners, possessives or adjectives and followed by PPs, and they are not allowed to be followed by bare NPs, as verbal *-ing* formations do.

Table 5: Type of dependents collocating with verbal *-ing* nominalizations (absolute figures and normalized frequencies per 10,000 words)

	E1	E2	E3	TOTAL
TYPE 1 (No Dependents)	-	-	-	-
TYPE 2 (Pre-head Dep.)	-	-	1 (0.42)	1 (0.12)
TYPE 3 (Post-head Dep.)	1 (0.36)	2 (0.69)	15 (6.40)	18 (2.27)
TYPE 4 (Pre- & Post-H.)	-	-	1 (0.42)	1 (0.12)
TOTAL	1 (0.36)	2 (0.69)	17 (7.25)	20 (2.52)

Table 6: Type of dependents collocating with nominal -ing nominalizations (absolute figures and normalized frequencies per 10,000 words)

	E1	E2	E3	TOTAL
TYPE 1 (No Dependents)	-	-	-	-
TYPE 2 (Pre-head Dep.)	2 (0.73)		-	2 (0.25)
TYPE 3 (Post-head Dep.)	-	-	1 (0.42)	1 (0.12)
TYPE 4 (Pre- & Post-H.)	5 (1.84)	4 (1.39)	1 (0.42)	10 (1.26)
TOTAL	7 (2.58)	4 (1.39)	2 (0.85)	13 (1.64)

Table 7: Type of dependents collocating with Romance nominalizations (absolute figures and normalized frequencies per 10,000 words)

	E1	E2	E3	TOTAL
TYPE 1 (No Dependents)	2 (0.73)	4 (1.39)	3 (1.28)	9 (1.13)
TYPE 2 (Pre-head Dep.)	4 (1.39)	6 (2.08)	13 (5.55)	23 (2.90)
TYPE 3 (Post-head Dep.)	2 (0.73)	1 (0.34)	1 (0.42)	4 (0.50)
TYPE 4 (Pre- & Post-H.)	1 (0.36)	5 (1.73)	10 (4.27)	17 (2.14)
TOTAL	9 (3.32)	16 (5.56)	27 (11.52)	52 (6.56)

3.4.1. Post-head dependents in doublets

The kind of post-head dependents of *-ing* and Romance nominalizations may be relevant to explain the existence of doublets. The different post-head dependents found in the corpus can be classified into: a) *of*-PP identified semantically with the subject of the nominalization (*of*-PP Subject) [cf. 4a]; b) *of*-PP identified semantically with the object of the nominalization (*of*-PP Object) [cf. 4b]; c) NP identified with the object of the nominalization [cf. 4c]; and d) other dependents (OTHER) [cf. 4d], including adverbial phrases and other PPs apart from the ones mentioned in (a) and (b).

(4a) E2 1602 Clowes *Treatise*. 8.72. […] requireth a fauourable acceptation, which is as well to be esteemed, as *the performance of them* […]

(4b) E2 1602 Clowes *Treatise*. 49.299. […] notwithstanding all our turmoiling, much care, industry and diligence, with *the application of most excellent medicines* […]

(4c) E2 1597 Blundevile *Cosmographie* 49V.50. […] I thinke it good therefore to shew you the order of the said tables by *describing the same* as followeth.

(4d) E3 1665 Hooke *Life* 115.94. for upon *examination with my Microscope*, I have found that the pith of an Elder, or almost any other Tree, [...]

As can be seen in Table 8 below, verbal *-ing* nominalizations almost always collocate with NPs, never below 93.8% of the number of occurrences. However, Tables 9 and 10 show that the post-head dependents of nominal *-ing* and Romance nominalizations are of a completely different kind: as seen in section 3.4, nouns are not allowed to be post-modified by a bare NP. Thus, *of*-PPs are the preferred dependents of both types of nominalizations. Having a look at the total amount of occurrences, nominal *-ing* appears mostly with *of*-PPs (Subject), 54.5% of the examples, while Romance nominals prefer *of*-PPs (Object), 50% of the occurrences. However, when having a look at the diachronic dimension, it shows a tendency for *-ing* nominals to move from *of*-PPs (Subject) in E1 (80%) to *of*-PPs (Object) in E3 (100%), whereas Romance nominalizations move from no instances of *of*-PPs (Subjects) in E1 to 45.4% in E3.

In the light of the results, it seems that a specialization in the kind of dependents collocating with each type of nominalization is taking place. This syntactic difference may be a reason for the coexistence of the different types of nominalizations analyzed.

Table 8: Post-head dependents of verbal *-ing* nominalizations

	E1	E2	E3	TOTAL
Of-PP (Subject)	-	-	-	-
Of-PP (Object)	-	-	-	-
NP	1 (100 %)	2 (100%)	15 (93.8%)	18 (94.7%)
Other	-	-	1 (6.2%)	1 (5.3%)
TOTAL	1	2	16	19

Table 9: Post-head dependents of nominal *-ing* nominalizations

	E1	E2	E3	TOTAL
Of-PP (Subject)	4 (80%)	2 (50%)	-	6 (54.5%)
Of-PP (Object)	1 (20%)	2 (50%)	2 (100%)	5 (45.4%)
NP	-	-	-	-
Other	-	-	-	-
TOTAL	5	4	2	11

Table 10: Post-head dependents of Romance nominalizations

	E1	E2	E3	TOTAL
Of-PP(Subject)	-	1 (16.7%)	5 (45.4%)	6 (30%)
Of-PP (Object)	2 (66.7%)	4 (66.6%)	4 (36.4%)	10 (50%)
NP	-	-	-	-
Other	1 (33.3%)	1 (16.7%)	2 (18.2%)	4 (20%)
TOTAL	3	6	11	20

4. CONCLUSION

This paper has shown that, although close in meaning, the two members of the doublets analyzed have some identifiable differences, and that these justify the coexistence of both. It is clear that once Romance nominalizations were introduced into English, they were assimilated in the receptor language and their bases used to create new nominalizations by the use of the native suffix -ing. The factor that justifies the existence of the doublet is the type of dependents used with each kind of nominalization. Nominal -ing and Romance nominalizations are nouns, and as such they usually have pre-head dependents or both pre- and post-head dependents. Furthermore, their post-head dependents are usually of-PPs, as in the case of nouns, -ing formations acquiring a preference for of-PPs (Object) and Romance nominalizations for of-PPs (Subject). The case of -ing nominalizations, however, is different. The pattern they usually follow is TYPE 3, and the kind of post-head dependent most frequently used is the NP. This possibility of combining with NPs with no requirement of a preposition in between is due to the acquisition of verbal characteristics of the gerund. Thus, it seems that the gerund is specialized in filling a gap in syntax that cannot be filled by nominalizations having a nominal nature.

REFERENCES

Atkinson, D. (1999). *Scientific Discourse in Sociohistorical Context. The Philosophical Transactions of the Royal Society of London, 1675-1975*. Mahwah, New Jersey: Lawrence Erlbaum.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Chomsky, N. (1970). Remarks on nominalization. In: A. J. Roderick & P. S. Rosenbaum (Eds.), *Readings in English Transformational Grammar,* (pp. 184-221). Waltham, Mass., Toronto & London: Ginn and Company.

Fanego, T. (1996). The gerund in Early Modern English: Evidence from the Helsinki Corpus, Folia Linguistica Historica 17, 97-152.

Grimshaw, J. (1990). *Argument Structure*. Cambridge, MA: The MIT Press.

Jack, G. B. (1988). The origins of the English gerund, NOWELE 12, 15-75.

Kroch, A., Santorini, B. & Delfs, L. (2004). *Penn-Helsinki Parsed Corpus of Early Modern* English. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/

Kytö, M. (1996). *Manual to the Diachronic part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts.* (3rd ed.). Helsinki: Department of English, University of Helsinki.

OED = Simpson, J. A. (Ed). (1989). *Oxford English Dictionary* (2nd ed.) on CD-ROM Version 3.1.1. Oxford: Oxford University Press.

El verbo débil como base de la derivación léxica en inglés antiguo

RAQUEL VEA ESCARZA

Universidad de La Rioja

Resumen

El objetivo de esta comunicación es ofrecer un análisis preliminar del papel desempeñado por los verbos débiles en la morfología derivativa del inglés antiguo. Partiendo de trabajos de lingüistas que han destacado la contribución central de los verbos fuertes a la formación de palabras en este estadio de la lengua inglesa, como Bammesberger (1965), Hinderling (1967) o Kastovsky (1992), se lleva a cabo un análisis de un corpus de verbos débiles, proporcionado por la base de datos léxica de inglés antiguo Nerthus (www.nerthusproject.com), para llegar a la conclusión de que los de los verbos débiles se puede derivar un paradigma derivativo relevante desde un punto de vista cuantitativo y también cualitativo.

Palabras clave: *morfología, formación de palabras, inglés antiguo, base de datos léxica*

Abstract

The purpose of this paper is to offer a preliminary analysis of the role played by weak verbs in the derivational morphology of Old English. By drawing on linguists who have emphasized the core contribution of strong verbs in word formation at this stage of the English language, such as Bammesberger (1965), Hinderling (1967), or Kastovsky (1992), the corpus of Old English verbs provided by the lexical database of old English Nerthus (www.nerthusproject.com) has been analyzed, to conclude that weak verbs repreent a productive source for the derivation of paradigms, both from a qualitative and a quantitative point of view.

Keywords: morphology, word formation, Old English, lexical database of Old English

1. INTRODUCCIÓN[1]

Kastovsky (1986, 1989, 1990, 1992, 2005, 2006) es el autor que más atención ha dedicado a la formación de palabras en inglés antiguo, centrándose en la evolución de la morfología de base variable a la de base invariable. Martín Arista (en prensa-a, en prensa-b, en prensa-c) también se ha ocupado de la cuestión, desde la perspectiva de la teoría morfológica de corte estrutural-funcional. Otros trabajos a tener en cuenta que han tenido como objeto de estudio la formación de palabras en inglés antiguo y su aplicación en la creación de una base de datos léxica son Caballero González *et al.* (2004-2005), Torre Alonso *et al*. (2008), González Torres (2009), Pesquera Fernández (2009) y Torre Alonso (2009).

Como han puesto de manifiesto varios lingüistas destacados (Bammesberger 1965; Hinderling 1967; Kastovsky 1992) los verbos fuertes son una fuente principal para la formación de palabras en inglés antiguo. Por ejemplo, a partir del verbo fuerte *(ge)drifan*

[1] Investigación financiada con cargo el proyecto FFI08-04448/FILO.

‘drive’ se obtienen los derivados que aparecen, por clase léxica, en el ejemplo (1), basado en Martín Arista (en prensa-c):

(1) *(Ge)dri:fan* (strong I) pret. sing. *dra:f*, pret. plur. *drifon, dreofon*, past part. *drifen*‘to drive, force, hunt, follow up, pursue; drive away, expel; practise, carry on; rush against, impel, drive forwards or backwards; undergo’.
Nombres femeninos: *dra:f* ‘action of driving’, *(ge)drif* ‘fever’, *fordrifnes* ‘opposition’, *onwega:drifennes* ‘a driving away’, *to:dræ:fednes* ‘dispersion’, underdrifennes ‘subjection’, *u:tdræ:f* ‘decree of expulsion’.
Nombres masculinos: *dræ:fend* ‘hunter’, *u:tdræ:fere* ‘driver out’.
Nombres neutros: *gedri:f* ‘a drive’.
Verbos fuertes de la clase I: *a:dri:fan* ‘to drive’, *bedri:fan* ‘to beat’, *efta:dri:fan* ‘to reject’, *eftfordri:fan* ‘to drive away’, *fordri:fan* ‘to sweep away’, *frama:dri:fan* ‘to remove’, *frama:dry:fan* ‘to drive away’, *indri:fan* ‘to ejaculate’, *oferdri:fan* ‘to overcome’, *onwega:dri:fan* ‘to drive away’, *to:dri:fan* ‘to scatter’, *ðurhdri:fan* ‘to drive through’, *u:ta:dri:fan* ‘to drive out’, *u:tdri:fan* ‘to expel’, *wiðdri:fan* ‘to repel.’

Sin embargo también hay evidencia de que es posible organizar paradigmas derivativos en torno a los verbos débiles. Para discutir adecuadamente la cuestión, he analizado 103 paradigmas derivativos con base derivativa en un verbo débil. Los datos se han obtenido de la base de datos léxica del inglés antiguo *Nerthus*, que comprende algo más de 30.000 registros basados en la información proporcionada por tres fuentes principales: *A Concise Anglo-Saxon Dictionary* (1996) de Clark Hall, *An Anglo-Saxon Dictionary* (1973) de Bosworth y Toller, y *The Student's Dictionary of Anglo-Saxon* (1976) de Sweet.

2. ANÁLISIS

Del análisis de los 103 paradigmas derivativos organizados a partir de verbos débiles, surgen 646 predicados derivados, que se agrupan por categorías como se ve en (2):

(2) Resultados cuantitativos por clase léxica
- Nombres : 289
- Adjetivos : 86
- Adverbios : 15

- Verbos débiles : 254
- Interjecciones : 2

De esta clasificación se concluye que la categoría léxica que más destaca en la formación de palabras a partir de verbos débiles es el nombre, con un total de 289 predicados, seguido muy de cerca por la categoría de verbo débil, que alcanza los 254 del total de predicados analizados.

En cuanto al proceso de formación de palabras que produce dichos predicados, los datos obtenidos pueden verse en (3):

(3) Resultados cuantitativos por proceso derivativo

- Básicos : 92
- Afijados : 312 (total)
    - Prefijados : 173
    - Sufijados : 139
- Compuestos : 174
- Derivación cero : 65
- Conversión : 1
- Sin determinar : 2

Una breve nota es necesaria respecto a la definición de los procesos. El uso de los términos afijación y composición es el comunmente aceptado, es decir: la afijación incluye la prefijación y la sufijación de formas trabadas mientras que la composición implica la combinación de formas libres. La conversión es la extensión categorial sin modificación formal, lo que en inglés antiguo significa que sólo hay conversión hacia clases morfológicamente invariables. La derivación cero, para terminar esta revisión terminológica, es la derivación sin morfemas flexivos o la derivación por medio de morfemas derivativos.

Los resultados del análisis, presentados por categoría léxica y proceso derivativo son los que se dan en (4):

(4) Resultados cuantitativos por categoría léxica y proceso derivativo

- Predicados básicos : 92 (total)
    - verbos débiles : 91
    - nombres : 1
- Predicados prefijados : 173 (total)

    - verbos débiles : 128
    - nombres : 17
    - adjetivos : 26
    - adverbios : 2
  - Predicados sufijados : 139 (total)
    - verbos débiles : 5
    - nombres : 87
    - adjetivos : 34
    - adverbios : 13
- Predicados compuestos : 174 (total)
    - nombres : 146
    - adjetivos : 18
    - verbos débiles : 9
    - interjecciones : 1
  - Derivados cero : 69 (total)
    - nombres : 38
    - verbos débiles : 23
    - adjetivos : 8
  - Conversión : 1 (total)
    - interjección : 1
  - 2 nombres por determinar en cuanto a proceso derivativo

Como se aprecia en esta clasificación son los nombres afijados los que más predicados reúnen, un dato que coincide con los resultados que Martín Arista (en prensa-b) obtiene tras llevar a cabo un análisis similar a éste pero con los verbos fuertes como punto de partida. Sin embargo, si consideramos la prefijación y la sufijación de forma aislada, entonces la composición, y concretamente los nombres compuestos, adquieren una mayor relevancia.

En lo que resta de esta comunicación analizo en detalle cuatro paradigmas derivativos organizados en torno a otros tantos verbos débiles. El criterio para seleccionarlos ha sido que se tratase de los paradigmas más amplios tanto en lo que respecta al número de derivados como a la variedad de clases léxicas derivadas. He seguido las convenciones de la base de datos léxica del inglés antiguo *Nerthus* en lo que respecta a utilizar dos puntos para indicar las vocales largas como en el uso de predicados numerados para diferenciar predicados

homófonos y/o sinónimos entre los que se establecen contrastes morfológicos como el género, la clase derivativa, la categoría léxica, etc.

El análisis del paradigma derivativo de HERIAN 1 puede verse en (5):

(5) **HERIAN 1** 'to extol, praise, commend; to help' (14 predicados)

herian 1 'to extol, praise, commend; to help'; *a:herian* 'to praise'; *efenherian* 'to praise together'; *(ge)herigendli:ce* 'praiseworthily'; *(ge)herigendlic* 'laudable, commendable; praising; excellent'; *herigend* 'flatterer'; *herigendsang* 'song of praise'; *herlic 1* 'noble'; *herung* 'praise'; *lofherung* 'praising'; *samodherian* 'to praise together'; *samodherigendlic* 'engaged in worshipping together'; *samodhering* 'praising'; *unherigendlic* 'not praiseworthy'.

Categorías: 4 adjetivos: *(ge)herigendlic, herlic 1, samodherigendlic, unherigendlic*; 1 adverbio: *(ge)herigendli:ce;* 5 nombres: *herigend, herigendsang, herung, lofherung, samodhering*; 4 verbos débiles: *a:herian, efenherian, herian 1, samodherian.*

Procesos derivativos: 1 básico: *herian 1*; 6 compuestos: *samodherigendlic, herigendsang, lofherung, samodhering, efenherian, samodherian*; 7 afijados: 5 sufijados: *(ge)herigendlic*, *herlic 1, (ge)herigendli:ce, herigend, herung*, y 2 prefijados: *a:herian*, *unherigendlic.*

Categoría y estatuto derivativo: 1 verbo débil básico; 2 verbos débiles compuestos; 3 nombres compuestos; 1 adjetivo compuesto; 1 verbos débil prefijado; 1 adjetivo prefijado; 1 adverbio sufijado; 2 adjetivos sufijados; 2 nombres sufijados.

El paradigma derivativo de HERIAN 1 consta de 14 predicados, de los que 5 son nombres, 4 adjetivos, 4 verbos débiles y 1 adverbio. En cuanto al estatuto derivativo, cabe decir que hay un sólo predicado básico, frente a los 6 compuestos y 7 afijados, 5 de ellos sufijados y 2 prefijados. Después de esta descripción, se concluye que el nombre es la categoría que más predicados concentra en el paradigma de HERIAN 1, seguida de cerca por el adjetivo y el verbo débil, siendo, por tanto, el adverbio la minoritaria. En cuanto al estatuto derivativo de los predicados, teniendo en cuanta que sólo hay un básico, es la derivación por medio de prefijos y sufijos el fenómeno lingüístico más destacado, y en similar medida también lo es la composición. El paradigma derivativo de HERIAN 1 se representa como se muestra en la figura 1.

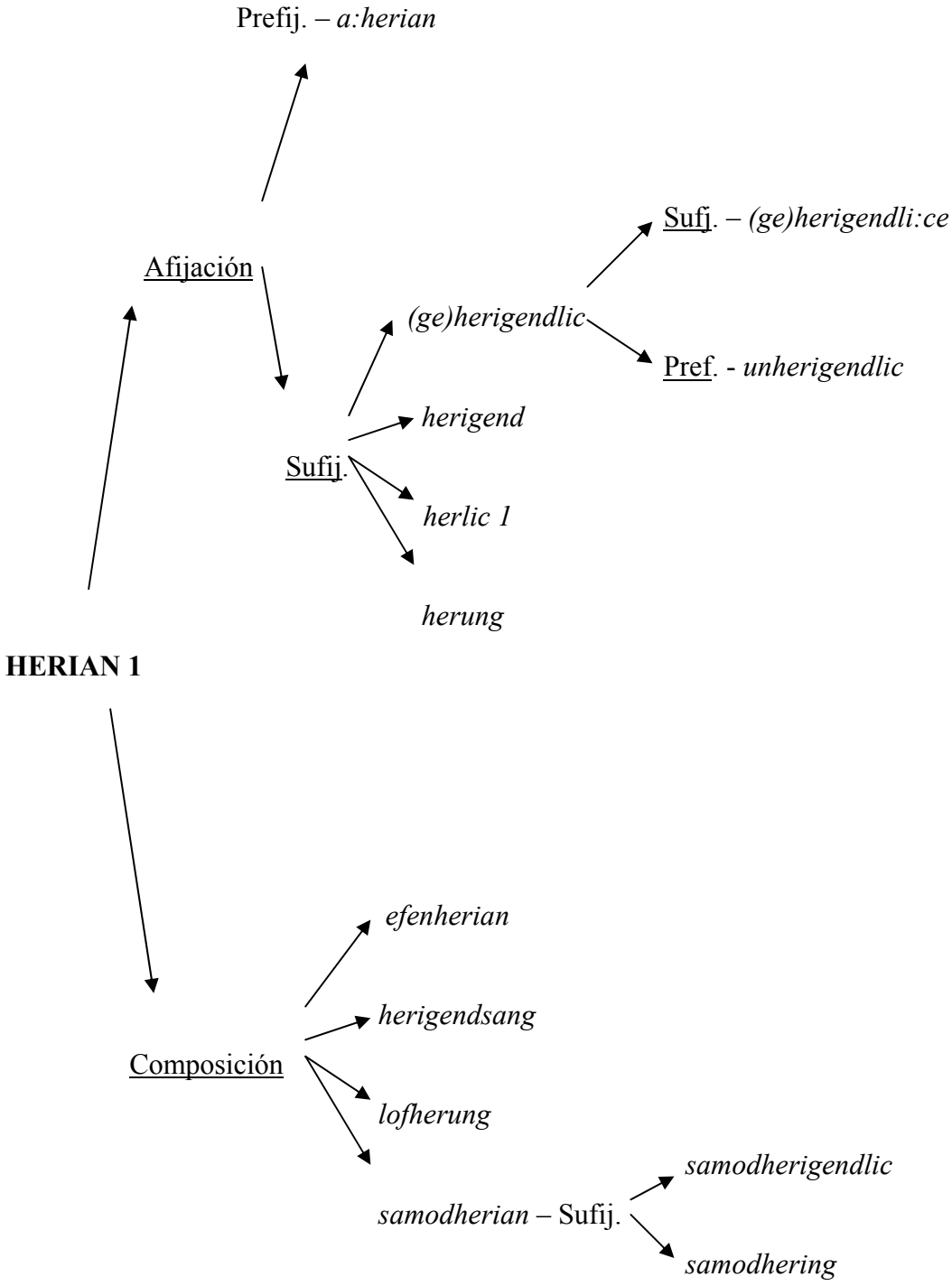


Figura 1: Representación del paradigma derivativo de HERIAN 1

A partir de la base de derivación HERIAN 1 se obtienen directamente, bien por medio de afijación o composición, los siguientes predicados: por un lado, la prefijación ha originado *a:herian*, mientras que por medio de la sufijación se han creado *(ge)herigendlic, herigend, herlic 1* y *herung*. A su vez, el predicado sufijado *(ge)herigendlic* ha dado por prefijación *unherigendlic* y por sufijación *(ge)herigendli:c*. Por otro lado, la composición ha formado predicados directamente derivados de la base sin ningún paso intermedio, incluidos *efenherian, herigendsang, lofherung* y *samodherian*; de éste último, esta vez por sufijación, se han obtenido *samodherigendlic* y *samodhering*.

El paradigma derivativo de (GE)HI:ERAN se da en (6):

(6) **(GE)HI:ERAN** 'to hear, listen (to); obey, follow; accede to, grant; be subject to, belong to, serve' (25 predicados)

(ge)hi:eran 'to hear, listen (to); obey, follow; accede to, grant; be subject to, belong to, serve'; *(ge)he:oran* 'to hear, listen (to); obey, follow; accede to, grant; be subject to, belong to, serve'; *(ge)hi:ernes* '(+) hearing, report; (+/-) obedience, subjection, allegiance, jurisdiction, district'; *(ge)hi:ersum* 'obedient, docile'; *(ge)hi:ersumian* 'to obey, serve'; *(ge)hi:ersumnes* 'obedience, submission; service; humility'; *gehi:eran* 'to judge'; *gehi:erend* 'hearer'; *gehi:erendlic* 'audible'; *gehi:ersumian* 'to reduce, subject, conquer'; *hi:ering 1* 'hearing, hearsay'; *hi:eringman* 'subject'; *hi:ersumlic* 'willing'; *mishy:ran* 'to hear amiss, not to listen to, disobey'; *ni:edhi:ernes* 'slavery'; *oferhi:eran* 'to overhear, hear; disobey, disregard, neglect'; *oferhi:ernes* 'neglect, disobedience; fine for transgression of law or legal orders'; *ofhi:eran* 'to hear, overhear'; *onhy:rsumian* 'to be busied with'; *ungehi:ersum* 'disobedient, rebellious'; *ungehy:red* 'unheard of, untold'; *ungehy:rnes* 'harness of hearing, deafness'; *unhi:ersum* 'disobedient'; *unhi:ersumli:ce* 'disobediently'; *unhi:ersumnes* 'disobedience'

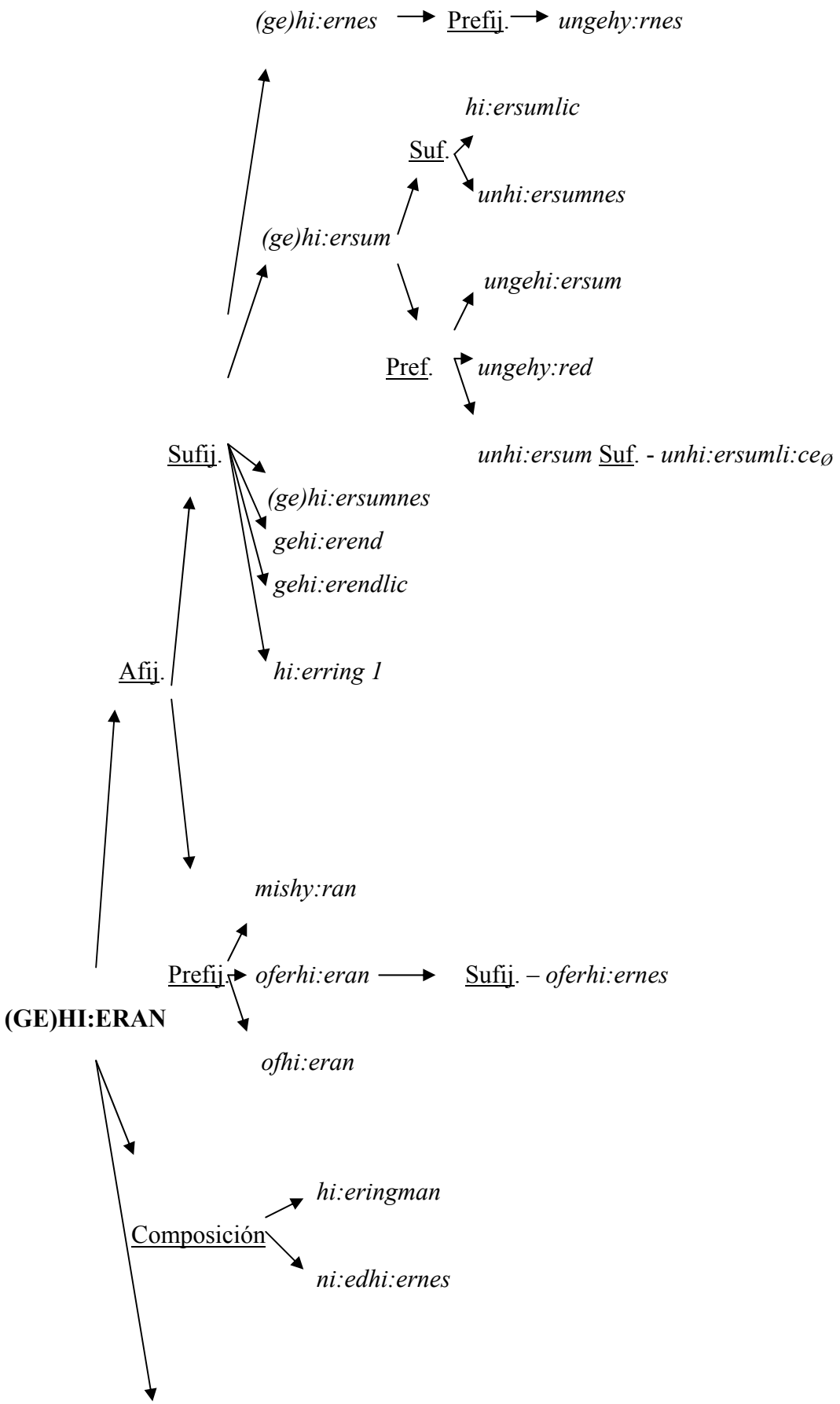
Categorías: 6 adjetivos: *(ge)hi:ersum, gehi:erendlic, hi:ersumlic, ungehi:ersum, ungehy:red, unhi:ersum;* 1 adverb: *unhi:ersumli:ce;* 9 nombres: *(ge)hi:ernes, (ge)hi:ersumnes, gehi:erend, hi:ering 1, hi:eringman, ni:edhi:ernes, oferhi:ernes, ungehy:rnes, unhi:ersumnes;* 9 verbos débiles: *(ge)hi:eran, (ge)hi:ersumian, gehi:eran, gehi:ersumian, mishy:ran, oferhi:eran, ofhi:eran, onhy:rsumian*

Procesos derivativos: 1 básico: *(ge)hi:eran*; 2 compuestos: *hi:eringman, ni:edhi:ernes;* 18 afijados: 8 prefijados: *mishy:ran, oferhi:eran, ofhi:eran, onhy:rsumian, ungehi:ersum, ungehy:red, ungehy:rnes, unhi:ersum*, y 10

sufijados: *(ge)hi:ernes, (ge)hi:ersum, (ge)hi:ersumnes, gehi:erend, gehi:erendlic, hi:ering 1, hi:ersumlic, oferhi:ernes, unhi:ersumli:ce, unhi:ersumnes*; 3 derivados cero: *(ge)he:oran, gehi:eran, gehi:ersumian*, y *(ge)hi:ersumian*

Categorías y estatuto derivativo: 1 verbo débil básico; 1 verbo débil derivado cero; 2 nombres compuestos; 3 adjetivos prefijados; 4 verbos débiles prefijados; 1 nombre prefijado; 6 nombres sufijados; 3 adjetivos sufijados; 1 adverbio sufijado; 3 verbos débiles derivados cero.

El paradigma derivativo de (GE)HI:ERAN, como puede verse en (6), consta de 25 predicados, de ellos son los nombres y los verbos débiles las categorías más numerosas, 9 de cada uno de ellos; mientras que vuelve a ser el adverbio la categoría menos relevante. En el caso de esta familia, el proceso de formación de palabras que más interviene es con diferencia respecto al resto la afijación, con más de la mitad de predicados que se han originado bien por prefijación o por sufijación. El paradigma de (GE)HI:ERAN se representa como muestra la figura 2:


(ge)hi:ernes
Prefij.
ungehy:rnes
hi:ersumlic
Suf.
unhi:ersumnes
(ge)hi:ersum
ungehi:ersum
Pref.
ungehy:red
Sufij.
unhi:ersum Suf. - unhi:ersumli:ceØ
(ge)hi:ersumnes
gehi:erend
gehi:erendlic
Afij.
hi:erring 1
mishy:ran
Prefij.
oferhi:eran
Sufij. – oferhi:ernes
(GE)HI:ERAN
ofhi:eran
hi:eringman
Composición
ni:edhi:ernes

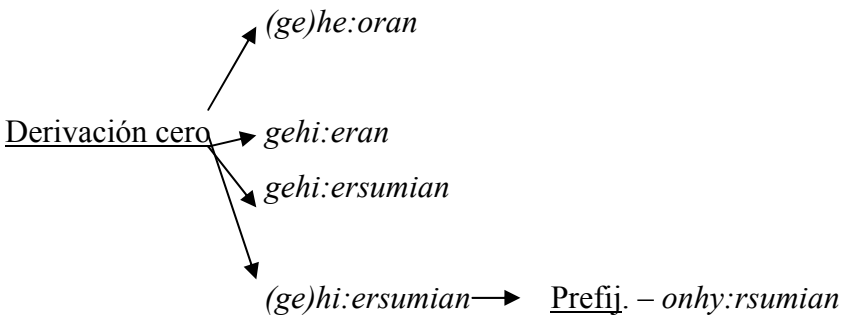


Figura 2: Representación del paradigma derivativo de (GE)HI:ERAN

De la base de derivación (GE)HI:ERAN se derivan varios procesos de formación de palabras para originar un total de 25 predicados. De este total, por medio de afijación se derivan directamente los siguientes: sufijados son *(ge)hi:ernes, (ge)hi:ersum, (ge)hi:ersumnes, gehi:erend, gehi:erendlic* y *hi:erring 1*; de *(ge)hi:ernes* deriva por prefijación *ungehy:rnes*; además, a partir del predicado sufijado *(ge)hi:ersum* se obtienen por sufijación *hi:ersumlic* y *unhi:ersumnes* y por prefijación *ungehy:red* y *unhi:ersum*; de éste último, a su vez, se obtiene por sufijación *unhi:ersumli:ce*$_{\emptyset}$, un predicado que no está atestiguado en la lengua o, al menos, en las fuentes consultadas, y se utiliza como paso intermedio entre dos predicados (fuente y meta) que existen. Por otro lado, a través de una prefijación de la base de derivación se han originado *mishy:ran, ofhi:eran* y *oferhi:eran*, éste último es fuente del predicado sufijado *oferhi:ernes*. A través de la composición, (GE)HI:ERAN ha producido *hi:eringman* y *ni:edhi:ernes*. Por derivación cero se han obtenido *(ge)he:oran, gehi:eran, gehi:ersumian*, y *(ge)hi:ersumian*, que mediante prefijación origina *onhy:rsumian*.

El paradigma derivativo de (GE)∂ENCAN sigue en (7):

> (7) **(GE)∂ENCAN 1** 'to think, imagine, think of, meditate, reason, consider; remember, recollect; intend, purpose, attempt, devise; learn; wish, desire, long for' (16 predicados)
> *(ge)∂encan* 'to think, imagine, think of, meditate, reason, consider; remember, recollect; intend, purpose, attempt, devise; learn; wish, desire, long for'*; æ:rbe∂oht* 'premeditated'; *(ge)∂o:ht* 'process of thinking, thought; mind; a thought, idea, purpose; decree; compassion, viscera'; *(ge)∂ync∂o* 'dignity, rank. office; meeting, assembly, court of justice; private agreement (to defeat justice)';

(ge)∂yncan 'to appear, seem'; *ge∂o:htung* 'counsel'; *inge∂o:ht* 'conscience'; *mo:dge∂o:ht* 'thought, understanding, mind'; *mys∂yncan* 'to be mistaken'; *of∂yncan* 'to give offence, insult, vex, displease, weary, grieve'; *ofer∂o:ht* 'thought over, considered'; *synne∂o:ht* 'sinful thought'; *unbe∂o:ht* 'unexpected'; *unbe∂o:hte* 'unthinkingly'; *unfor∂o:ht* 'unexpected'; *woruldge∂o:ht* 'worldly thought'.

Categorías: 4 adjetivos: *æ:rbe∂oht, ofer∂o:ht, unbe∂o:ht, unfor∂o:ht*; 1 adverbio: *unbe∂o:hte*; 7 nombres: *(ge)∂o:ht, (ge)∂ync∂o, ge∂o:htung, inge∂o:ht, mo:dge∂o:ht, synne∂o:ht, woruldge∂o:ht*; 4 verbos débiles: *(ge)∂encan, (ge)∂yncan, mys∂yncan, of∂yncan.*

Procesos derivativos: 1 básico: *(ge)∂encan*; 4 compuestos: æ:rbe∂oht, mo:dge∂o:ht, synne∂o:ht, woruldge∂o:ht; 8 afijados: 7 prefijados: *inge∂o:ht, mys∂yncan, of∂yncan, ofer∂o:ht, unbe∂o:ht, unbe∂o:hte, unfor∂o:ht*, y 1 sufijado: *ge∂o:htung*; 3 derivados cero: *(ge)∂o:ht, (ge)∂ync∂o, (ge)∂yncan*.

Categoría y estatuto derivativo: 1 adjetivo compuesto; 3 nombres compuestos; 1 nombre prefijado; 2 verbos débiles prefijados; 3 adjetivos prefijados; 1 adverbio prefijado; 1 nombre sufijado; 2 nombres derivados cero; 1 verbo débil derivado cero.

El paradigma derivativo de (GE)∂ENCAN cuenta con un total de 16 predicados, siendo el nombre la categoría más destacada, frente a un sólo predicado correspondiente a la categoría adverbio. La afijación vuelve a mostrarse como el proceso que mayor número de predicados ha originado, especialmente la prefijación. La representación del paradigma aparece en la figura 3:

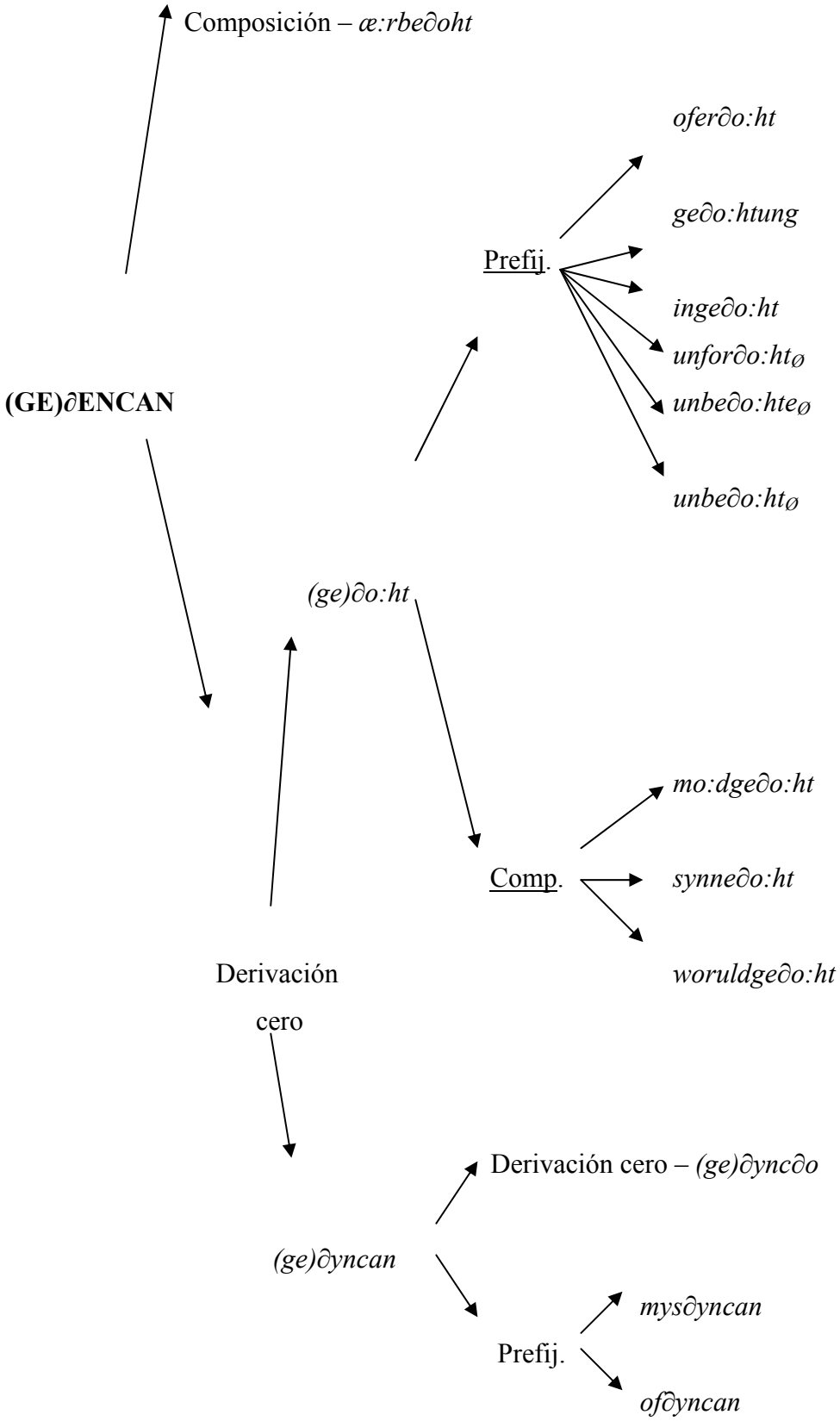


Figura 3: Representación del paradigma de (GE)ðENCAN

A partir de la base de derivación (GE)∂ENCAN se origina dos primeros procesos de formación de palabras, por un lado la composición, y, por otro, la derivación cero. Creado por composición directamente de la base de derivación, encontramos el predicado *æ:rbe∂oht*. Mediante derivación cero se han obtenido *(ge)∂o:ht* y *(ge)∂yncan*; el primer predicado ha servido de fuente de derivación por medio de prefijación a los siguientes predicados: *ofer∂o:ht*, *ge∂o:htung*, *inge∂o:ht*, *unfor∂o:ht*$_{Ø}$, *unbe∂o:hte*$_{Ø}$, y *unbe∂o:ht*$_{Ø}$ (los tres últimos son hipotéticos o reconstruidos). Asimismo, por medio de la composición de *(ge)∂o:ht* también se han derivado los predicados *mo:dge∂o:ht, synne∂o:ht*, y *woruldge∂o:ht*. El segundo predicado obtenido por derivación cero directamente de la base de derivación es *(ge)∂yncan*, que, a su vez, ha producido *(ge)∂ync∂o*, mediante un proceso de derivación cero en segundo nivel; además, por medio de prefijación también se derivan los predicados *mys∂yncan* y *of∂yncan*.

El paradigma derivativo de SLECCAN se muestra en (8):

(8) **SLECCAN** 'to weaken, disable'

a:slacian 'to become slack, decline, diminish; grow tired; make slack, loosen, relax, dissolve'; *a:slæccan* 'to slacken, loosen'; *slacful* 'lazy'; *sleac* 'slack, remiss, lax, sluggish, indolent, languid; slow, gentle, easy'; *sleacian* 'to delay, retard, slacken, relax efforts'; *sleacli:ce* 'slothfully'; *sleaclic* 'slow, languid, idle'; *sleacmo:dnes* 'slackness, laziness'; *sleacnes* 'slowness; remissness, laziness'; *sleacornes* 'laziness'; *sleccan* 'to weaken, disable'; *unsleac* 'not remiss, active, diligent'; *unsleacli:ce* 'energetically'

Categorías: 4 adjetivos: *slacful; sleaclic; unsleac*; 2 adverbios: *sleacli:ce*; *unsleacli:ce*; 3 nombres: *sleacmo:dnes*; *sleacnes*; *sleacornes;* 4 verbos débiles: *a:slacian*; *a:slæccan*; *sleacian*; *sleccan*

Status:1 básico: *sleccan*; 10 affixed: 3 prefijados: *a:slacian, a:slæccan, unsleac*, y 7 sufijados: *slacful, sleacli:ce, sleaclic, sleacmo:dnes, sleacnes, sleacornes, unsleacli:ce*; 2 derivados cero: *sleac, sleacian.*

Categorías y status: 1 verbo débil básico; 2 verbos débiles prefijados; 1 adjetivo prefijado; 2 adjetivos sufijados; 2 adverbios sufijados; 3 nombres sufijados; 1 adjetivo derivado cero; 1 verbo débil derivado cero.

El paradigma de SLECCAN reúne un total de 13 predicados que se reparten de manera más equitativa entre las distintas categorías, al contrario que en las anteriores familias; aún

así son el adjetivo y el verbo débil las más destacadas. Una vez más, es la afijación el proceso más repetido entre los predicados del paradigma, si bien en este caso el número de predicados sufijados supera al de los prefijados. En la figura 4 se ofrece la representación del paradigma derivativo de SLECCAN.

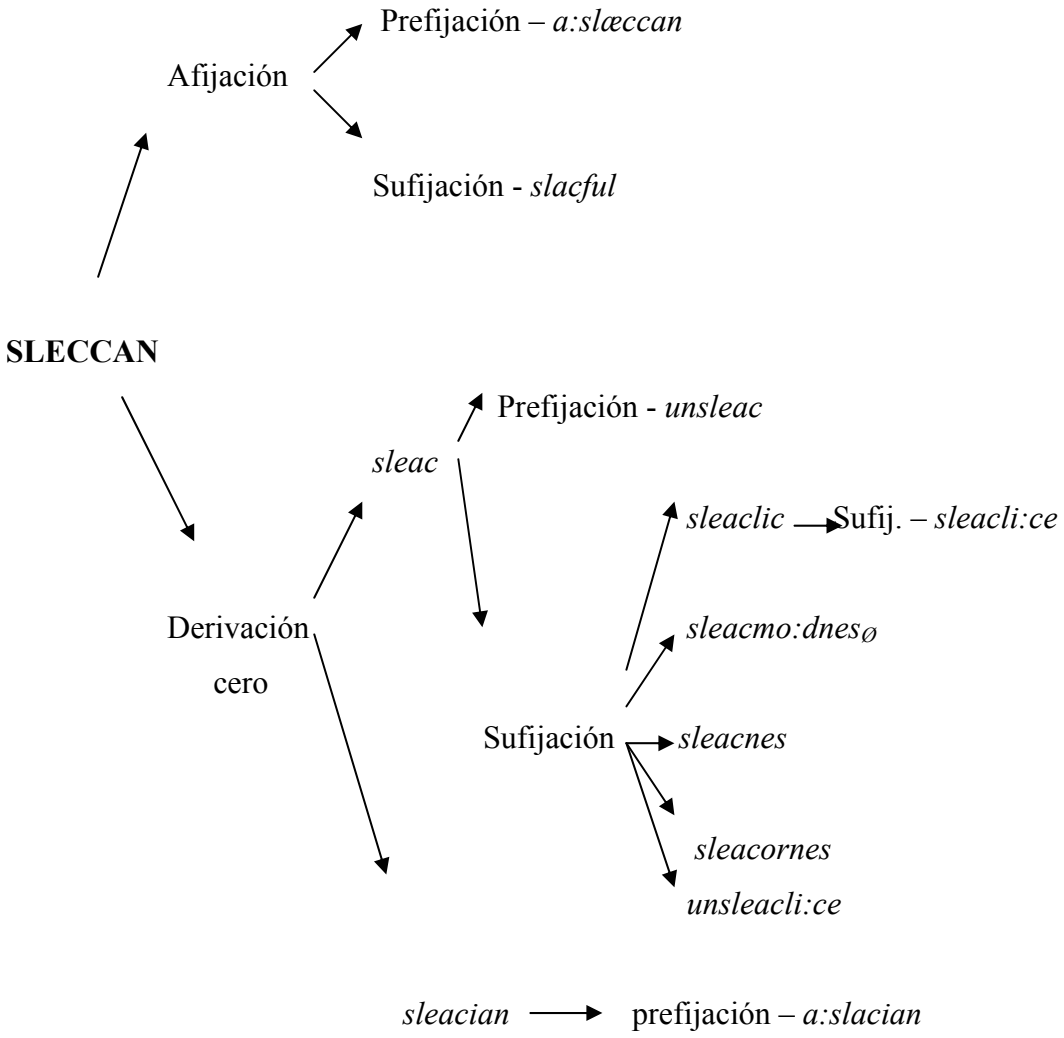


Figura 4: Representación de la familia de SLECCAN

Directamente de la base de derivación SLECCAN se obtienen predicados por medio de afijación y derivación cero. Por prefijación se obtiene *a:slæccan*, y por sufijación *slacful*. Por medio de lo que se denomina derivación cero se originan dos predicados, *sleac* y s*leacian*. *Sleac* produce a través de la prefijación el predicado *unsleac*, mientras que por sufijación *sleaclic*, el hipotético *sleacmo:dnes*$_{Ø}$, *sleacnes, sleacornes* y *unsleacli:ce*. De *sleaclic*, además, se obtiene de nuevo por sufijación el predicado *sleacli:ce*. Por último, *sleacian* forma *a:slacian* tras ser prefijado.

3. CONCLUSIÓN

A modo de conclusión cabe decir que los verbos débiles son una fuente relevante para la formación de palabras en inglés antiguo. En el estadio actual del proyecto *Nerthus* se han podido identificar 103 paradigmas derivativos de verbos débiles, que dan un total de 646 predicados a partir de los siguientes procesos de formación de palabras: afijación (prefijación y sufijación), composición, derivación cero y un caso de conversión, incluyendo aquéllos que son básicos, y, como tales, no se ven afectados por ninguno de estos procesos.

La afijación, en términos generales, es decir incluyendo la prefijación y la sufijación, resulta ser el proceso que más palabras ha originado a raiz de verbos débiles, aunque, considerados individualmente, la composición sobrepasa en un predicado a la prefijación. Además, teniendo en cuenta conjuntamente la categoría y el estatuto derivativo, son los nombre compuestos los que mayor número de predicados constituyen, seguidos de cerca por los verbos débiles formados por prefijación.

Frente a los nombres y a los verbos débiles, que son las categorías léxicas más destacadas cuantitativamente, los adjetivos y, sobre todo, los adverbios resultan ser las categorías que menos frecuentemente derivan de verbos débiles. La conversión, en concreto, no se considera un proceso destacado puesto que sólo se ha obtenido un predicado de este tipo en el análisis llevado a cabo.

REFERENCIAS BIBLIOGRÁFICAS

Bammesberger, A. (1965). *Deverbative jan-Verba des Altenglischen, vergleichend mit den übrigen altgermanischen Sprachen dargestellt*. München: Ludwig-Maximilians Universität.

Caballero González, L. *et al*. (2004-2005). Predicados verbales primitivos y derivados en inglés antiguo. Implicaciones para la elaboración de una base de datos léxica. *RESLA* 17-18 : 35-49.

González Torres, E. (2009). *Affixal Nouns in Old English: Morphological Description, Multiple Bases and Recursivity.* Tesis doctoral, Universidad de La Rioja.

Hinderling, R. (1967). *Studien zu den starken Verbalabstrakten des Germanischen.* Berlin: Walter de Gruyter.

Kastovsky, D. (1992). Semantics and vocabulary. In R. Hogg (ed.) *The Cambridge History of the English Language I: The Beginnings to 1066*, (pp. 290-408). Cambridge: Cambridge University Press.

Martín Arista, J. *et al.* (2009). *Nerthus: An Online Lexical Database of Old English*. http://www.nerthusproject.com

Martín Arista, J. OE strong verbs derived from strong verbs. *SKASE Journal of Theoretical Linguistics*. En prensa-a

Martín Arista, J. Morphological Relatedness and Zero Alternation in Old English. In Butler C. and P. Guerrero Medina (eds.) *Morphosyntactic Alternations in English*. London: Equinox. En prensa-b.

Martín Arista, J. Building a lexical database of Old English: issues and landmarks. In Considine, J. ed. *Current projects in historical lexicography*. Newcastle: Cambridge Scholars Publishing. En prensa-c.

Pesquera Fernández, L. (2009). *Transparent and Opaque Word-Formation in the Derivational Paradigms of Old English Strong Verbs.* Tesis doctoral, Universidad de La Rioja.

Torre Alonso, R. *et al.* (2008). Fundamentos empíricos y metodológicos de una base de datos léxica de la morfología derivativa del inglés antiguo. *Revista de lingüística y lenguas aplicadas* 3: 129-144.

Torre Alonso, R. (2009). *Morphological process feeding in the formation of old English nouns: Zero-derivation, affixation and compounding*. Tesis doctoral, Universidad de La Rioja.

Using multilingual parallel corpora for contrastive studies and translation studies: A case study of the verbs of sitting, standing, and lying

ÅKE VIBERG
Uppsala University

Abstract
From the combined perspectives of contrastive studies and of translation studies, the paper presents a case study of the Swedish postural verbs referring to sitting standing and lying and their correspondents in English based on data from the English Swedish Parallel Corpus (ESPC). A comparison of English original and translated texts shows clear signs of "translationese" or crosslinguistic influence due to the much more frequent use of the postural verbs in Swedish. This study is complemented with a semantically more fine-grained analysis of one of the postural verbs STAND based on the ESPC and The Multilingual Parallel corpus which contains data also of the translations of Swedish stå 'stand' into German, French and Finnish.
Keywords: multilingual corpora, translation, contrastive linguistics

Resumen
El trabajo presenta, a partir de una combinación de las perspectivas aportadas por la lingüística contrastiva y los estudios de traducción, un estudio sobre los verbos suecos de posición corporal (con el significado de 'estar parado', 'estar sentado' y 'estar acostado') y sus equivalentes ingleses, basado en datos extraídos del Corpus Paralelo Inglés-Sueco (ESPC). La comparación entre textos ingleses originales y traducidos muestra señales claras de "traslacionismo", o influencia interlingüística debida a la mayor frecuencia de uso de los verbos de postura corporal en la lengua sueca. El estudio se completa con un análisis semántico detallado del verbo stand 'estar parado' basado en el corpus ESPC y el Corpus Paralelo Multilingüe, que posee datos acerca de la traducción del verbo equivalente sueco stå al alemán, el francés y el finlandés.
Palabras clave: corpus multilingüe, traducción, linguüística contrastiva

1. USING MULTILINGUAL PARALLEL CORPORA FOR CONTRASTIVE STUDIES AND TRANSLATION STUDIES

Within contrastive studies, corpus-based analysis has led to the revitalization of a field that was dormant for many years. The use of parallel (or translation) corpora and comparable corpora allows fine-grained corpus-based contrastive comparisons (Altenberg & Granger 2002, Johansson 2007, Gómez González et al 2008). Parallel corpora can also be used to study translational phenomena such as translation universals, "translationese" and the treatment of "unique elements" in translations (Gellerstam 1986, Laviosa 2004, Mauranen & Kujamäki 2004, Halverson 2003, forthc.).

Using the Swedish verbs of sitting, standing and lying as an example, the present paper will present data from two translation corpora. One is the Multilingual Parallel Corpus

(MPC), which is being compiled by the author and consists of extracts from Swedish novels and their published translations into English, German, French and Finnish. The total number of words in the Swedish originals is around 250 000 words in the version used for this study. The other corpus is the English Swedish Parallel Corpus (ESPC) compiled by Altenberg and Aijmer (2000). The ESPC corpus consists of original printed texts in Swedish and English together with their translations. The texts are divided into two broad genres: Fiction and Non-fiction with several subcategories. The number of words in various subcorpora of the ESPC are shown in Table 1.

Table 1: The composition and size of of the English Swedish Parallel Corpus (ESPC)

		English originals	**Swedish translations**	**Swedish originals**	**English translations**	**Total**
Fiction	Number of words	340.745	346.649	308.160	333.375	1.328.929
Non-Fiction	Number of words	364.648	344.131	353.303	413.500	1.475.582
Total	Number of words	705.393	690.780	661.463	746.875	2.804.511

2. THE MAJOR USES OF THE POSTURAL VERBS IN SWEDISH

In their use as postural verbs indicating the position of the human body, the Swedish verbs *sitta, stå* and *ligga* are very similar to their English cognates *sit*, *stand* and *lie*. The Swedish verbs indicate a state or activity: *Peter sitter* 'Peter sits/is sitting'. As shown in Table 2, all postural verbs have specific causative forms, e.g. *Peter satte babyn i stolen* 'Peter put (seated) the baby in the chair'. The causative postural verbs can also be used in a reflexive form (with the reflexive pronoun *sig* 'oneself') to indicate an inchoative meaning (assume a bodily posture): *Peter satte sig* 'Peter sat down'.

Table 2: The Swedish Postural verbs

Location	***Posture***		
State	***State/Activity***	***Postural change (Inchoative)***	***Causative***
HORIZONTAL DIMENSION DOMINANT	ligga 'lie'	lägga sig 'lie down'	lägga 'lay'
1.VERTICAL DIMENSION DOMINANT	stå 'stand'	ställa sig 'stand up' resa sig (upp) 'rise (up)'	ställa 'stand'
2. FUNCTIONAL UPPER SIDE			
ATTACHED	sitta 'sit'	sätta sig 'sit down' sätta sig upp 'sit up' slå sig ner 'sit down'; 'settle'	sätta 'seat'

An important use of the causative forms, which will not be treated here, is as verbs of putting with physical objects. English *put* is usually translated with one of these three verbs (Author 1998). The Swedish postural verbs *sitta*, *stå* and *ligga* are often used as locational verbs to indicate the position of physical objects. The major contrasts in that use are briefly indicated in the leftmost column in Table 2. *Stå* is used when the vertical dimension is most salient as in the normal position of a vase: *Vasen står på bordet* 'The vase is (standing) on the table', whereas *ligga* is used when the horizontal dimension is dominant as in *Vasen ligger på bordet* 'The vase is (lying) on the table'. The verb *stå* is also used about objects which have a functional upper side and appear in their canonical position as in *Tallriken står på bordet* 'The plate is (standing) on the table' (Cf. *Tallriken ligger upp och ner på bordet* 'The plate is (lying) upside down on the table'). The verb *sitta* is used when something is attached to the landmark as in *Räkningarna sitter i pärmen* 'The bills are in the file' (If the bills are field away.) Cf. *Räkningarna ligger i pärmen* 'The bills are (lying loose) in the file'. A special case of the locational use, is the use of the postural verbs in presentational (existential) constructions: *Det står en vas på bordet* 'There is (lit. it stands) a vase on the table', *Det sitter ett frimärke på brevet* 'There is (lit. it sits) a stamp on the letter'. Figurative uses are also prominent. These functions have parallels in many other languages as evidenced by the work presented in Newman (ed. 2002). In his introduction, Newman states that the extent to which posture verbs can be extended to non-human referents in posture-based locational expressions varies. The correspondents of the Swedish postural verbs in Romance languages in which postural verbs have a limited function have already been studied in rather great detail (see

Kortteinen 2008 for Swedish-French and Svensson 2005 for Swedish-Italian. For a study of the Swedish posture verbs see Author 1985 and Jacobson 1996 and for a study of English posture verbs in English corpora, see Newman & Rice 2004).

3. The use of postural verbs in original and translated texts. A case of "translationese"

What makes a contrastive comparison between Swedish and English particularly interesting is the striking differences in usage patterns in spite of the fact that all three canonical posture verbs have direct equivalents which are transparent cognates. Table 3 shows the occurrence of the English and Swedish canonical postural verbs in the English Swedish Parallel Corpus (ESPC). A pattern that applies to both languages is the difference between genres. As can be observed in Table 3, the postural verbs are much more frequent in Fiction than in Non-fiction in both English and Swedish, in spite of the fact that the total number of words is somewhat higher for Non-fiction. If we compare the use of postural verbs in the original texts, we find that the postural verbs are much more frequent in Swedish than in English in spite of the fact that the basic meaning is very similar. In total, the three English verbs have the frequency 932 versus 2357 for the Swedish verbs. (The same difference is found if the individual verb pairs are compared.) Since the number of words in the subcorpora are not exactly equal, the number of occurrences per 100 000 have been calculated and are indicated below the total number of postural verbs in the subcorpora. As can be observed, the differences are clear even when the sizes of the subcorpora are taken into consideration.

Table 3: Postural verbs in English and Swedish in the ESPC corpus

	Original texts			Translated texts		
	Fiction	NonFict.	Total	Total	Fiction	NonFict.
English						
sit	302	50	352	504	419	85
stand	231	90	321	515	442	73
lie*	155	104	259	369	294	75
Total			932	1388		
Frequency per 100 000 words			132	186		
Swedish						
sitta**	489	125	614	317	277	40
stå	666	273	939	724	487	237
ligga	493	309	804	665	389	276
Total			2357	1706		
Frequency per 100 000 words			358	247		

(The figures for *stand* and *stå* are exact, based on a line-by-line reading of all examples. The other figures are based on automatic extractions of the various inflected forms of the verbs. The figures for the English lemma *lie* include occurrences of *lie*, *lies*, *lying* in the sense 'tell a lie'. The figures for Swedish *sitta* do not include the imperative form *sitt*, which is homonymous with the frequent reflexive possessive pronoun *sitt* but include the past form *satt* which is homonymous with the supine *satt* of *sätta* 'put', which is less frequent than the past of *sitta*.)

The translated texts can be studied from different perspectives. Translations can be compared to originals in the same language to see if there is any systematic influence on the target language from the source language, a phenomenon referred to as "translationese" by Gellerstam (1986). As can be observed, the English postural verbs are used with a much greater frequency in translations than in original texts. Together the three verbs *sit*, *stand* and *lie* occur 932 times in the English original texts versus 1388 times in translated texts. The English postural verbs are clearly overrepresented in translated texts. There is a clear difference between originals and translations also at the level of individual verbs. Comparing Swedish original texts with Swedish translated texts, we find the opposite pattern. Swedish postural verbs are clearly underrepresented in translations, both if compared as a group (2357 in originals versus 1706 in translations) and at the level of individual verbs. There is a clear influence from the source language on the translated texts in both languages. This is a clear case of translationese in Gellerstam's (1986) sense and is a parallel to what is called transfer, interference or crosslinguistic influence in bilingualism and second language acquisition research (Jarvis & Pavlenko 2008).

4. A CLOSER LOOK AT ENGLISH *STAND* AND SWEDISH *STÅ*

It is possible to get a better grasp of the reasons why English translated texts differ so much from original texts by looking more closely at the different meanings (or uses) of the postural verbs. The verbs referring to standing will be used as an example. The various uses of English *stand* and Swedish *stå* are shown in Table 4 and Table 5, respectively. The categories are rather broad and clear-cut in order to make the statistical differences more clear. More fine-grained differences will be discussed later but such distinctions in many cases are difficult to quantify since many of these uses form a continuum. The term postural is used when the subject is human (or in a few cases an animal) and refers to a concrete situation where a certain body posture is clearly involved. In many cases, the body posture may be back-grounded and a location or an accompanying activity be fore-grounded (*stand talking in the yard*) but such distinctions form a continuum and will be discussed later. A distinction is made between a Postural state or activity (no change of posture) as in (1) and postural change which refer to the case where a new posture is assumed as in (2) and (3). (In the examples, original texts will be quoted in the first line followed by a reference in capitals to the individual text from which the example is taken.)

Postural state

(1) Sometimes I ***stand*** in the bathtub, elbows resting on the sill, SG
Ibland ***står*** jag i badkaret med armbågarna på fönsterbrädan,

Postural change

(2) — ***Stand up*** till we see you. RDO
— ***Res dig upp*** så vi ser dig. (lit. 'raise yourself up')

(3) Kevin ***stood up*** again and scouted for a watchman. RDO
Kevin ***ställde sig upp*** igen och spanade efter vakten.

The term Location will be used when the subject refers to a concrete inanimate object to indicate its location in physical space as in (4).

Location (of a concrete, inanimate physical object)

(4) The father took a golf-club that ***was standing*** in the corner. RD
Hennes pappa tog en golfklubba som ***stod*** i ett hörn.

Since in Swedish, a very frequent use of *stå* refers to a written message, this use is singled out as a special category named Writing. This use is clearly based on the idea that letters are standing on a line but often the reference concerns the message rather than the letters as such as in (5).

Writing

(5) Men det måste ***stå*** på skylten hur långt det är till fornminnet. SC
[lit. 'it must stand on the sign']
But it ought to ***say*** on the sign how far it is to the ancient monument.

In addition to the broad categories mentioned so far, there are many interesting figurative uses and set phrases of various types but the parallel corpus is not large enough to make quantitative comparisons meaningful of the many different uses of these types which are simply referred to as Other (see 6 for an example).

Other (Figurative, set phrases etc)

(6) He was prepared to ***stand by*** that report, FF
Han var beredd att ***hålla fast*** vid rapporten,

Table 4 shows a comparison of English original and translated texts. As can be observed, the use of *stand* to indicate a postural state is clearly over-represented. (The proportion F^{transl}/F^{orig} is 2.08.) In most cases, the Swedish originals contain the verb *stå*. The use of *stand* to refer to the location of a concrete inanimate object is also more frequent in translations than in originals (proportion: 1.58) but the total number of occurrences of this use is lower. As already mentioned, Swedish *stå* is not used to indicate a change of position. Writing will be commented on in section 5.

Table 4: The major meanings of the English verb stand and their major Swedish translations:

	English original texts		*English translated texts*	
Meaning	F^{orig}	*stå* as correspondent of *stand* in Swedish translation	F^{transl}	*stå* as correspondent of *stand* in Swedish original
Postural state/activity	162	143	337	298
Postural change	41	0	27	
Location	43	26	68	48
Writing	0	0	0	
Other	56	21	55	18
Total	302	190	487	364

Table 5 shows the use of *stå* in Swedish original texts. Compared to the use of *stand* in English original texts, Swedish *stå* is used to refer to a postural state almost three times as much as *stand* in English originals. Swedish *stå* is not used quite as much to refer to location but this use is more than four times as ferequent as the use of *stand* in this function in English originals. In this function, *be* is more frequent than *stand* as a translation of *stå*, even though the use of *stand* as a translation appears to be over-represented.

Table 5: The major meanings of the Swedish verb stå and their major English translations

	Swedish original texts			
Meaning	F	Major translations		
		stand	be	say
Postural state/activity	443	298	37	
Postural change	0			
Location	184	48	85	
Writing	89			24
Other	223	18		
Total	939			

A comparison of the figures for the category Other indicates that figurative uses and conventionalized set phrases are more prominent in Swedish than in English. In this case, *stand* is used as a translation in relatively few cases.

To sum up: In spite of the fact that Swedish *stå* and English *stand* have the same basic meaning when referring to the posture of a human being, posture is indicated much more in Swedish originals than in English originals even when this feature is back-grounded and other characteristics of the situation, in particular the location, are in the foreground. This is a contrast in usage pattern rather than in meaning. Crosslinguistic influence is strongest in the English translations in this basic use. There is a certain amount of crosslinguistic influence even in the extended use as a locational verb referring to the location of a concrete inanimate object, but in this case the crosslinguistic influence is weaker. The relatively frequent use of

Swedish *stå* referring to Writing is language-specific and has no counterpart in the use of English *stand*. There is no indication of direct transfer of the Swedish meaning in this case since the contrast is so obvious. (Such influence might be expected with non-advanced second language learners.)

5. THE TRANSLATIONAL EQUIVALENTS OF *STÅ* ACROSS FOUR LANGUAGES

The Multilingual Parallel Corpus (MPC, see Section 1) makes it possible to look at translations of Swedish originals across four different languages. The major translations of *stå* in this corpus are displayed in Table 6. French has a different set of major translations from the other languages which will be commented on later.

Table 6: The major translations of *stå* into English, German, French and Finnsih in the MPC Corpus

Swedish		English		German		French		Finnish	
Meaning	stå	stand	be	stehen	sein	debout*	être	seisoa	olla 'be'
Postural	241	156	4	196	5	19	18	165	15
Locational	99	16	43	76	4	2	9	13	63
Resistence	10	3		1		0		0	
		say		stehen		écrire 'write'		lukea 'read'	
Writing	38	11		30		15		13	
Various	56								
Total	444								

*) All expressions where *debout* forms a part: *être debout, rester debout, se tenir debout*,

As indicated above, the category Postural covers all cases where a verb refers to the body posture of a human (or animal) whether the posture is profiled as in (7) or back-grounded as in (8), where the location of the human is profiled.

(7) a Men han tyckte nästan det var lika kallt när han ***stod upp***. KE
b but he found it almost as cold when he ***stood up***.
c Er fand aber, daß es ***im Stehen*** fast genauso kalt war.
d De toute façon, il avait presque aussi froid quand il ***était debout***.
e Mutta oikeastaan oli yhtä kylmää kun ***seisoi***.

(8) a Men det var i alla fall hennes skuld att han ***stod*** i brunnen. KE
b But all the same, it was her fault he ***was*** down a well.

c Es war aber auf alle Halle ihre Schuld, daß er in dem Brunnen ***stand***.

d Mais en tout cas c'était de la faute à Gudrun s'il ***se trouvait*** dans ce puits.

e Gudrunin vika joka tapauksessa, että hän nyt ***oli*** täällä kaivossa.

As can be observed in Table 6, the verb meaning 'stand' is dominant as a translation in all the MPC languages except French, when the subject is human, in spite of the fact that the posture is clearly profiled as in (7) only in a limited number of cases.

When the subject refers to the location of a concrete inanimate object as in (9) and (10), the copula in its locative function is the most frequent translation in both English and Finnish in spite of the fact that there is a well-established correspondent of *stå* in its prototypical use as postural verb. German, on the other hand, predominantly uses *stehen* 'stand' as a translation also when *stå* refers to the location of an object. French differs completely from the other languages. As noted above, posture is seldom indicated even when the subject is human. To indicate location, French can use *être* 'be' as a translation as in (9) but this is just one of several options in French. Since these have been described in a detailed and insightful way in Kortteinen (2008), they will be briefly treated here. Among the most frequent options are *se trouver* 'find oneself' and *se dresser*. A frequent alternative identified by Kortteinen is the use of a resultative expression consisting of *être* 'be' + a participial form of an action verb such as *être posé* 'be placed' in (10) or *être garé* 'be parked'. Apart from French, which represents a separate type of language, the languages displayed in Table 6 form a continuum from Swedish and German with predominant use of a postural verb to indicate location to English and Finnish where the copula with a locative complement predominates.

(9) a Hon visste inte var hennes sang ***stod*** eller hur gammal hon var. MF

b She didn't know where her bed ***was*** or how old she was,

c Sie wußte nicht, wo ihr Bett ***stand*** oder wie alt sie war.

d Quel âge avait-elle ? Où ***était*** son lit? Elle l'ignorait. ('be')

e Hän ei tiennyt, missä hänen sänkynsä ***oli*** tai miten vanha hän oli. ('be')

(10) a Tallriken ***stod*** på en grå vaxduk med röd bård. IB

b The plate ***is*** on grey oilcloth with a red border,

c Der Teller ***stand*** auf einem grauen Wachstuch mit roter Umrandung.

d L'assiette ***est posée*** sur une toile cirée grise à bordure rouge. (‘be placed’)

e Lautanen ***oli*** harmaalla vahakangasliinalla, jota reunusti punainen raita.

In clauses where *stå* indicates a location the identification of the place is the most important information and the meaning of the postural verb is partly redundant. The translation is several cases involves restructuring where the postural verb has a zero correspondent. In most cases, the restructuring consists in the replacement of a relative clause containing *stå* + Place with a single prepositional phrase as in all languages except Finnish in (11).

(11) a Han betraktade böckerna som ***stod*** där i hyllan. HM
[lit. He looked at the books that stood there on the shelf]

b He looked at the books on the shelf.

c Er betrachtete die Bücher im Regal.

d Il contempla les livres sur l'étagère.

e Hän katseli hyllyssä olevia kirjoja. [He looked at in shelf being books]

In Swedish, the postural verbs are sometimes used with an adjective as complement in a way that is similar to the descriptive use of a copula. In example (12), the French and Finnish translations consist of copula + Adjective. However, the Swedish postural verbs in such uses retain some of their original meaning and have not developed into pure copulas.

(12) a Bussdörren ***stod öppen*** i värmen (’stood open’)

b The bus doors ***had been left open*** because of the heat,

c Die Bustür ***stand offen*** wegen der Hitze, (’stood open’)

d La portière de l'autocar ***était ouverte*** à cause de la chaleur, (’was open’)

e Koska oli lämmin, bussin ovi ***oli auki***, (’was open’)

As already noted, one special, but actually rather frequent use is when *stå* ‘stand’ is used with reference to various kinds of writing, which obviously is motivated by the fact that the letters are conceived of as standing on a line. As can be observed in (13) other languages

can conceptualize this in several different ways. Whereas ‘say’ often is used in English, the most frequent translation in Finnish uses ‘read’ (*lukea*) with a generic subject (expressed as zero. In anaphoric reference, an explicit pronoun is required in the third person in Finnish.). In French, the most frequent translation involves some form of the verb *écrire* ‘write’. German even in this case can use *stehen* in a similar way as *stå*.

(13) a Barbro Lund med son ***stod*** på listan. KE
b Barbro Lund and son, ***it said*** on the list.
c Barbro Lund mit Sohn, das ***stand*** auf der Liste.
d Barbro Lund et son fils - ***c'était écrit*** sur le carnet.
e Barbro Lund ja poika, niin listassa ***luki***.

The use of *stå* to refer to writing can be further extended and refer to the propositional content rather than the actual letters or words as in (14).

(14) Hon fick ett papper där det ***stod*** att kungen av Karelen, han som gick på tunet i stövlar och väst, hade skjutits av ryssarna redan i början av kriget. AP
[lit. ’a paper where it stood that the King of Karelia…]
She was given a paper which ***said*** that the King of Karelia, the one who had walked about in boots and waistcoat, had been shot by the Russians right at the beginning of the war.

To sum up: In the use as a postural verb, the Swedish verb *stå* has a simple verb as a close correspondent in all the MPC language except French. The extent to which this equivalent is used varies across languages. In the postural use, it is the most frequent translation (except in French). The close equivalents may be overrepresented in translations but this cannot be tested as long as only translations in one direction are studied. In the locational use, there is a clear contrast between the target languages. German uses *stehen* as a translation in most cases, whereas English and Finnish predominantly uses ‘be’ as a translation and this must reflect a genuine contrast between these languages and Swedish. French has a system that belongs to a different type than the other languages represented in the MPC corpus. The use of STAND to refer to Writing appears to be a very language-specific feature even if it is shared with German which is closely related to Swedish and

appears to be very similar to Swedish in the use of *stehen* (except, perhaps, with respect to figurative meanings and phraseological units based on STAND).

6. TOWARDS A MORE FINE-GRAINED ANALYSIS

It is possible to break down the major uses into a number of more fine-grained categories. Space only allows one examples of this type. Postural state (and activity) as represented in the tables above covers a construction known as pseudo-coordination in Swedish grammar (see Darnell 2008 for a recent in-depth study), which represents an interesting case of emergent grammaticalization. In a pseudo-coordination, the first verb is de-accented and the construction also behaves idiosyncratically with respect to certain word order regularities. When the first verb is a postural verb coordinated with another Verb ($V^{Postural}$ + *och* 'and' + V), the construction indicates progressive aspect. One frequent type of English correspondent consists of only the second verb in progressive form as in (15).

(15) Hon ***stod*** antagligen ***och vinkade*** åt honom. KE [lit. 'stood an waved at him']
She ***was*** presumably ***waving*** to him.

Another relatively frequent English correspondent is the combination of Postural verb and a verb in the progressive form ($V^{Postural}$ + V-*ing*), a construction which Newman & Rice (2004) refer to as s*imultaneous conjunction*.

(16) He would drop the spoon in the sink and ***stand sipping*** from his mug while the cat wove between his feet. AT
Sedan skulle han släppa ner skeden i diskhon och ***stå och läppja*** på kaffet medan katten slingrade sig ut och in mellan fötterna på honom. [lit. 'would… stand and sip']

(16) is taken from an English original text. In Swedish translations, this construction is probably over-represented but the fine-grained coding of the examples has not yet been completed, so no statistics will be presented. In pseudo-coordination, the postural verb is completely back-grounded and this is clearest when nothing intervenes between the postural verb and the rest of the construction as in (16) whereas the postural verb is not back-grounded to the same degree in (17).

(17) Nere i ångloket ***stod*** lokföraren Lukas Kelly, före detta sjöman, ***och spanade*** ut genom sidofönstret. ARP [lit. 'stood /… / and looked']
Down in the engine cab, Lukas Kelly, engine-driver and ex-seaman, ***was looking*** out of the side window.

An interesting, but not very frequent translation of the Swedish pseudo-coordination is shown in (18), in which the pseudo-coordination refers a characteristic and repeated activity and is translated with the English past habitual marker *used to* V.

(18) Det hela hade slutat med att fadern hade kastat sin kaffekopp i tulpanrabatten och stängt in sig i gavelbyggnaden där han **stod och målade** sina tavlor med samma, ständigt upprepade motiv: HM2
It ended up with his father throwing his coffee cup into the tulip bed and locking himself in the shed where he **used to paint** his pictures with the same motif, repeated over and over again:

The most direct equivalents to the Swedish pseudo-coordination in the MPC languages are present in example (19).

(19) a Till sist tog jag mod till mig och frågade pappa som ***stod och tvättade*** vår nya Volvo PV. MN [lit. 'stood and washed']
b In the end I picked up courage and went over to Dad, who ***was*** busy ***washing*** our new Volvo PV:
c /---/ Papa, der gerade unseren neuen Volvo PV ***wusch***: [Zero + 'washed']
d /---/ papa qui ***était en train de*** laver notre nouvelle Volvo PV:
e /---/ isältä, joka ***oli pesemässä*** uutta Volvo PV: [lit. 'was in washing']

In (19), English uses the second verb in the progressive form as translation, whereas German uses a Zero translation of the postural verb in combination with the second verb carrying the ordinary tense marking (past in this case). This is a frequent alternative in German, but also a literal translation of postural verb + 'and' + V occurs. It remains to be seen whether the last alternative is overrepresented in the translations. French uses the progressive marker *être en train de* as a translation in (19) but this is not a frequent

correspondent. Usually, the postural verb corresponds to Zero and the second verb is used alone. In examples in past tense, the second verb often appears in an imperfective past form (imparfait), which can be regarded as a partial correspondent of the progressive meaning in the Swedish original. Finnish rather frequently uses a progressive construction consisting of *olla* 'be' (as in 19) or sometimes a postural verb in combination with the second verb in the 3rd infinitive form marked with inessive case (*pese-mä-ssa* V-3Inf-Inessive).

Many figurative uses are based on image schemas of the type studied in cognitive linguistics. For English, Gibbs et al (1994) found that the dominant image schemas were RESISTANCE, CENTER-PERIPHERY, BALANCE, VERTICALITY and LINKAGE. These image schemas motivate many of the figurative uses of Swedish *stå* as well (see Jacobsson 1996). Translational equivalence is usually low in such uses in spite of the fact that the same image schemas often are involved in English and Swedish and this is due to the frequent use of phraseological units to express the figurative meanings.

7. CONCLUSION

From a contrastive perspective, the present study shows how Swedish and German differ from the other languages in particular with respect to the use of the postural verbs to describe the location of a concrete inanimate object in physical space. The use of STAND to refer to Writing turned out to be more or less unique characteristic of Swedish and German with respect to the other languages. The use of postural verbs in pseudo-coordination to express progressive meaning in Swedish turned out to lack a direct structural equivalent in all the languages in the study including German, the most closely related language.

As a contribution to translation studies, the most interesting result was the amount of crosslinguistic influence found in English translations from Swedish. The rather great amount of such influence found in particular with the use of English *stand* as a translation of Swedish *stå* is probably due to the fact that the basic meaning is very similar and that the contrast between the two languages in many respects can be characterized as a difference in usage patterns. The influence was also strongest when *stå* was used in the basic postural meaning. This result can be compared to what has been found in earlier studies as regards the translation strategies used to express the meaning of elements which are more or less unique in the source language. One such element is the use of Swedish *gå* 'go' to express modal meanings as in (20).

(20) Somliga människor ***går*** helt enkelt inte att beskriva. MG
[lit. 'some people goes simply not to describe']
Some people simply ***can***'t be described.

To render this meaning, English translators often take recourse to can in combination with the main verb in the passive. The other MPC languages take recourse to partly different translation strategies which share the characteristic of referring to a vague or generic agent such as the agentless passive or various types of impersonal constructions (Author 1999, forthc.) Another type of translation strategy is the relatively frequent use of Zero translations of certain of the language-specific meanings of one of the most frequent Swedish verbs *få* 'get;may' (Author 2002). Different translation strategies are used depending on the type of relationships between the elements involved in the source and target languages.

REFERENCES

Altenberg, B. & Aijmer, K., (2000). The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies, in: C. Mair & M. Hundt (eds.) *Corpus Linguistics and Linguistic Theory*, pp. 15-33, Rodopi, Amsterdam – Atlanta/GA.

Altenberg, B. & S. Granger (eds.) (2002). *Lexis in contrast*. Amsterdam: Benjamins.

Author, A. (1985). Lexikal andraspråksinlärning. *SUM-rapport 2*. Centre for Research on Bilingualism. Stockholm University.

Author, A. (1998). Contrasts in polysemy and differentiation: Running and putting in English and Swedish. In Johansson, S. & Oksefjell, S. (eds.), *Corpora and cross-linguistic research*. Amsterdam: Rodopi, 343-376.

Author, A. (1999). The polysemous cognates Swedish *gå* and English *go*. Universal and language-specific characteristics. *Languages in Contrast*, 2, 89-115.

Author, A. (2002). Polysemy and disambiguation cues across languages. The case of Swedish *få* and English *get*. In B. Altenberg & S. Granger (eds.), *Lexis in contrast*, 119-150. Amsterdam: Benjamins.

Author, A. forthcoming. Basic verbs in typological and contrastive perspective.

Darnell, Ulrika Kvist. (2008). *Pseudosamordningar i svenska : särskilt sådana med verben sitta, ligga och stå.* PhD thesis. Department of Linguistics, Stockholm University.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindqvist (eds.) *Translation studies in Scandinavia*. [Lund Studies in English 75] Lund: CWK Gleerup, 88-95.

Gibbs, R., Beitel, D., Harrington, M., and Sanders. P. (1994). Taking a stand on the meaning of Stand: Bodily experience as motivation for polysemy. *Journal of Semantics 11*, 231-251.

Gómez González, M., J. L. Mackenzie & E. González Álvarez (eds.) (2008). *Current trends in contrastive linguistics*. Amsterdam: Benjamins.

Halverson, S. (2003). The cognitive basis of translation universals. Target 15:2, 197-241.

Halverson, S. forthcoming. Cognitive translation studies. Developments in theory and method. In Gregory Shreve and Erik Angelone (eds), *Translation and cognition.* Philadelphia: John Benjamins [25 pages]

Jacobsson, Ulrika. (1996). Familjelika betydelser hos *stå, sitta* och *ligga* – en analys ur den kognitiva semantikens perspektiv. *Nordlund 21*. Lund University: Department of Scandinavian Languages.

Jarvis, S. & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. London & New York: Routledge.

Johansson, S. (2007). *Seeing through multilingual corpora*. Amsterdam: Benjamins.

Kortteinen, P. (2008).*Les verbes de position suédois* stå, sitta, ligga *et leurs équivalents français*. *Étude contrastive*. [Acta Univeristatis Gothoburgensis LX.] Göteborg: Göteborgs universitet.

Laviosa, S. (2002). *Corpus-based translation studies*. Amsterdam - New York: Rodopi.

Mauranen, A. & Kujamäki, P. (eds.) (2004). *Translation universals*. Amsterdam: Benjamins.

Newman, J. (ed.) (2002). *The linguistics of sitting, standing, and lying*. Amsterdam: Benjamins.

Newman, J. & Rice, S. (2004). Patterns of usage for English SIT, STAND, AND LIE: A cognitively inspired exploration in corpus linguistics. *Cognitive Linguistics 15-3*, 351-396.

Svensson, K. (2005). *Uno studio contrastivo svedese-italiano sui verbi svedesi* stå, sitta *e* ligga: Thèse pour le doctorat. Göteborg: Göteborgs universitet.

Subordinación sustantiva en redacciones de estudiantes de licenciatura en educación secundaria

IRMA GUADALUPE VILLASANA MERCADO

Centro de Actualización del Magisterio (Zacatecas)

Resumen

Para que los docentes tengan un dominio adecuado del español, la enseñanza de la lengua materna tiene que darse a través de una planificación lingüística que parta de los saberes que ya se poseen. Estudiar el dominio lingüístico de estudiantes normalistas en México resulta indispensable para planificar cursos que coadyuven al mejoramiento de la relación que se inaugura entre educadores y alumnos a través de la palabra. Por ello, la presente disertación analiza la subordinación sustantiva en 64 redacciones de estudiantes de licenciatura en educación secundaria (LES) del Centro de Actualización del Magisterio en Zacatecas; muestra el dominio que los futuros docentes poseen de esta estructura, corrobora que el grado escolar influye en el mismo y describe el tipo de estructura con que se actualizan los diversos tipos de subordinación sustantiva.

Palabras claves: gramática intensional, competencia comunicativa, complejidad sintáctica, relaciones subordinadas sustantivas, tipos de subordinación sustantiva

Abstract

In order that the teachers have a suitable command of Spanish, the education of mother tongue has to be given across a linguistic planning that departs on the knowledge that is already possessed. Studying the students' linguistic domain in Mexico turns out indispensable to plan courses that contribute to the improvement of the relation that is inaugurated among educators and pupils across the word. For it, the present paper analyzes noun subordination in drafts of 64 undergraduate students in Secondary Education of the Centro de Actualización del Magisterio en Zacatecas; it shows the control of this structure that the future teacher has, it corroborates that the school degree influences it and describes the type of structure with which the diverse types of substantive subordination are updated

Keywords: intensional grammar, communicative competition, syntactic complexity, substantive relations, types of substantive subordination

> Cuando una persona aprende a leer y escribir *convenientemente* aprende muchas cosas más sobre su idioma y aprende también a ver el mundo de una manera diferente, una manera más rica y global. Así, adquiere las estructuras lingüísticas necesarias no sólo para expresar y comprender una gama muy amplia de sentimientos y pensamientos sino incluso para concebir adecuadamente tales sentimientos y pensamientos. (López Chávez y Arjona Iglesias 2001: 12)

1. INTRODUCCIÓN

En la escuela, el portador de la tarea de enseñar lengua es el docente. Aunque la asignatura a impartir no sea Español, el medio por el que el maestro interactúa con el alumno es la palabra. Para que se inaugure dicha vinculación intersubjetiva, el profesor debe poseer

esquemas lingüísticos flexibles que le permitan expandir coherentemente su espacio en el mundo y el de sus alumnos.

Para que los docentes tengan un dominio adecuado del español, la enseñanza de la lengua materna tiene que darse a través de una planificación lingüística que parta de los saberes que ya se poseen. Estudiar el dominio lingüístico de estudiantes normalistas en México resulta indispensable para planificar cursos que coadyuven al mejoramiento de la relación que se inaugura entre educadores y alumnos a través de la palabra. Por ello, la presente disertación analiza la subordinación sustantiva en 64 redacciones de estudiantes de Licenciatura en Educación Secundaria (LES) del Centro de Actualización del Magisterio en Zacatecas; se pretende saber si los futuros docentes poseen un dominio adecuado de dicha estructura y si el grado escolar influye en el mismo.

2. El docente y la enseñanza del Español

La enseñanza de la lengua en niveles básicos y superiores debería propiciar el incremento de la competencia comunicativa, es decir, las habilidades que permiten actuar en varias situaciones. Esto conlleva una competencia léxica, "el conocimiento que cada hablante tiene del lexicón de su lengua", una competencia gramatical, "el conjunto de reglas sintácticas que permitirán (al sujeto) producir –o comprender- un número infinito de oraciones por medio de una cantidad limitada de reglas", y las modalidades de ejecución de un contexto específico (López Chávez, 2001: 24).

El futuro docente tendría que desarrollar conscientemente estas competencias como lo marca el perfil de egreso de las escuelas normales: al egresar el alumno "expresa sus ideas con claridad, sencillez y corrección en forma escrita y oral; en especial, ha desarrollado las capacidades de describir, narrar, explicar y argumentar, adaptándose al desarrollo y características culturales de sus alumnos" (2004: 10).

Para Arjona Iglesias el escolar debería adquirir y desarrollar la gramática intensional o de la partícula, aquella que corresponde a los elementos vacíos de significado, como los llama Lyons (1983), y que implica el dominio de las piezas que permiten la combinatoria de los vocablos. Scholes y Willis definen estos componentes como "entidades que no hacen referencia a nada fuera del sistema lingüístico mismo. Sus significados se encuentran dentro de la propia gramática [preposiciones y conjunciones se agrupan aquí]" (1998: 297).

Si el normalista no desarrolla su competencia comunicativa y, por ende, domina la gramática intensional, tendrá mayores dificultades para inaugurar un puente armónico en el

aula o forjar ciudadanos. Cuando finaliza la clase, ¿qué permanece en la memoria del alumno? Las palabras del profesor, que pueden ser un jeroglífico impenetrable o una ventana a otros mundos.

3. La subordinación sustantiva

Las oraciones subordinadas sustantivas desempeñan respecto de la oración en que se insertan funciones típicamente reservadas a los grupos nominales. Se trata de los oficios sintácticos de sujeto, predicado nominal, objeto directo, régimen de verbo prepositivo y objeto indirecto (Delbecque y Lamiroy, 1999). De acuerdo a los trabajos de Lope Blanch, la sustantivación abarca 14.9% del total de núcleos verbales empleados en el habla de Hispanoamérica. Dentro de la misma, las oraciones objetivas gozan de mayor vitalidad con 10.3%; después se ubican las sujetivas con 3.3%. Los otros tipos presentan una ocurrencia menor a 1%: las predicativas comprenden solamente 0.6%; las de régimen de verbo prepositivo, 0.5%; las objetivas indirectas 0.2%. (1993: 50)

Las sujetivas cumplen el rol de sujeto de la oración a la cual se subordinan; las predicativas, de predicado nominal "al relacionarse con el sujeto de la oración a la cual se subordina a través de los verbos copulativos *ser* o *estar*" (Luna Traill, 2007: 1006). Como las subordinadas no se hallan siempre en el mismo grado de necesidad respecto al elemento principal, las sujetivas y predicativas son indispensables, ya que "no se pueden separar sin que el enunciado pierda su sentido" (Gili Gaya, 2002: 985). La frecuencia y la adecuada actualización de estas oraciones es una marca de complejidad sintáctica, porque refleja la capacidad del hablante para expandir elementos esenciales dentro de la oración.

Dentro de las sustantivas, las objetivas directas que cumplen el papel de complemento directo por su frecuencia integran los casos más representativos de la subordinación sustantiva. Según Martínez Lara, la producción de las mismas es más fácil "debido a que para ello sólo requieren de un verbo transitivo" (2006: 82), cuya ocurrencia en español es alta. A partir de lo anterior, la autora propone que las redacciones que contengan más subordinación sustantiva objetiva directa "serán redacciones con bajo nivel de complejidad sintáctica" (2006: 82).

Las oraciones de régimen de verbo prepositivo funcionan como término de una estructura cuyo núcleo es un verbo prepositivo. Esta clase de verbos "se construye obligatoriamente con una preposición, de tal manera que si ésta se suprime, tanto la estructura de la oración como la significación del verbo resultan afectadas" (Luna Traill, 2007: 1513).

por sus características estructurales, el manejo adecuado y diversificado de esta estructura sustantiva es un indicador del dominio de la gramática intensional. Para la presente investigación, dentro de la subordinación sustantiva, los informantes que produzcan mayor cantidad de objetivas directas y no muestren una distribución y variación de las clases de subordinación sustantiva reflejarán menor grado de aprehensión de los elementos intensionales.

4. La subordinación sustantiva en redacciones de estudiantes de LES

En las composiciones de estudiantes de LES, hay 476 oraciones sustantivas: 268 objetivas directas (igual a 56.30%), 95 sujetivas (19.96%), 67 predicativas (14.08%), 45 de verbo de régimen prepositivo (9.45%) y sólo un testimonio de objetiva indirecta (0.21%). Esta distribución es análoga a la presentada por Lope Blanch con un coeficiente de correlación de 0.98. La oración subordinada sustantiva objetiva directa cubre 56.30%; entre los otros cuatro tipos, el porcentaje restante, 43.70%. Estas cifras corroboran la predominancia de las objetivas directas sobre las otras y reflejan la necesidad de que en la escuela se refuerce el uso de las otras oraciones sustantivas.

Tabla 14: Oraciones subordinadas sustantivas por escolaridad

SEMESTRE	Primero		Tercero		Quinto		Séptimo		Total por tipo	
Tipo de subordinación sustantiva	Frec.	%	Frec.	%	Frec.	%	Frec.	%	Frec.	%
Objetiva directa	68	55.74	76	59.38	65	59.63	59	50.43	268	56.30
Objetiva indirecta	0	0	0	0	1	0.92	0	0	1	0.21
Predicativa	15	12.30	14	10.94	17	15.60	21	17.95	67	14.08
Sujetiva	33	27.05	26	20.31	19	17.43	17	14.53	95	19.96
De verbo de régimen prepositivo	6	4.92	12	9.38	7	6.42	20	17.09	45	9.45
Total	122	100	128	100	109	100	117	100	476	100

Por escolaridad en tercero se trasponen más oraciones sustantivas, 128. Primero y séptimo se ubican en el segundo lugar con 122 y 117 casos, respectivamente. En las muestras de quinto semestre sólo hay 109 testimonios, 19 menos que en tercero. En primero 10 estructuras tienen algún tipo de anomalía; siete, en tercero y quinto; cuatro, en séptimo. Por tanto, se observa que entre primero y tercero aumenta el uso de sustantivas, en el siguiente

año disminuye y en el último grado vuelve a ascender sin llegar a equiparar o superar la producción de los primeros semestres.

En los cuatro grados las objetivas prevalecen sobre las otras clases. En tercero aparece 76 veces (equivalente a 59.38%); en primero, 68 (55.74%); en quinto, 65 (59.63%); en séptimo, 59 (50.43%). En los ciclos intermedios el porcentaje de este tipo es mayor que en el inicial y el final. En el último año disminuye la producción de objetivas en 9.2% con relación al grado anterior.

Después de las objetivas se ubican las sujetivas en tres grados: primero con 33 ocurrencias (27.05%), tercero con 26 (20.31%) y quinto con 19 (17.43%). En séptimo se colocan en cuarto con 17 casos (14.53%). En el manejo de estructuras complejas que desempeñen el rol de sujeto se observa una disminución en la cantidad producida conforme aumenta el grado escolar. En la elaboración de predicativas se invierte el comportamiento: en séptimo hay 21 oraciones (17.95%); en quinto, 17 (15.60%); mientras en primero, 15 (12.30%), y, en quinto, 14 (10.94%).

Los estudiantes de séptimo generan más sustantivas de verbo de régimen prepositivo que los de otros grados, 20 casos (17.09%). En tercero esta construcción aparece 12 veces (9.38%); en quinto, siete (6.42%); en primero, seis (4.92%). Con relación a las otras clases la aparición de ésta es inferior. Sólo hay un testimonio de oración objetiva indirecta en quinto semestre. Dicha frecuencia no resulta relevante frente al total del grado, representa 0.92%.

Por frecuencia y distribución los aprendices de séptimo reflejan una mayor habilidad para insertar subordinadas sustantivas adecuadamente. Aunque elaboran un número inferior de éstas, usan proporcionalmente más predicativas, de verbo de régimen prepositivo y sujetivas; el porcentaje de las objetivas directas es inferior al promedio por 5.87 puntos porcentuales. Los de primero ocupan el siguiente lugar, ya que la aparición de las objetivas directas tampoco rebasa la media, las sujetivas tienen una frecuencia de 27.05%, 7.54% sobre el promedio; sin embargo, el uso de predicativas y de verbo de régimen prepositivo requiere trabajarse.

En tercero y quinto el porcentaje de objetivas directas es mayor a la media; el de verbo de régimen, inferior. En tercero las sujetivas están por encima del promedio y las predicativas por debajo; en quinto se muestra el patrón inverso. Estos resultados indican que los alumnos de los grados intermedios poseen una competencia más baja que los de los otros grados para elaborar oraciones subordinadas sustantivas.

5. Estructura de las relaciones subordinadas sustantivas

A continuación se trabajan las estructuras oracionales de cada una de las clases de subordinación sustantiva, así como los verbos regentes de las mismas.

Tabla 15: Estructura oracional de las relaciones subordinadas objetivas directas

Estructura	Total	
	Frec.	%
Introducida por *que*	143	53.36
De relativo sustantivada	27	10.07
Interrogativa indirecta	24	8.96
De infinitivo	72	26.87
De estilo directo	2	0.75
Total	268	100

Al revisar la forma en que se incrustan las oraciones subordinadas objetivas directas se encuentra que en la muestra aparecen cinco actualizaciones distintas: introducida por *que* en 143 casos (53.36%), de infinitivo en 72 (26.87%), de relativo sustantivada en 27 (10.07%), interrogativa indirecta en 24 (8.96%) y de estilo directo sólo dos veces (0.75%). Los verbos distintos que rigen las oraciones objetivas directas son 65. El vocablo más frecuente es querer con 23 apariciones, además hay ejemplos del mismo en los cuatro grados. Luego se ubican hacer con 22, creer con 21, decir con 20 y saber con 16. El primero se presenta sólo en tres grados; los segundos en todos. Después de estos verbos la frecuencia de aparición es menor a 11. 30 verbos aparecen una sola vez.

Tabla 16: Estructura oracional de las relaciones subordinadas predicativas

Estructura	Total	
	Frec.	%
Introducida por *que*	16	23.88
De relativo sustantivada	21	31.34
De infinitivo	30	44.78
Total	67	100

Las predicativas se actualizan a través de tres estructuras distintas. En orden descendente éstas son de infinitivo (30 casos, 44.78%), de relativo sustantivada (21, 31.34%) e introducidas por *que* (16, 23.44%). Sólo aparecen tres verbos distintos como regentes de estas oraciones subordinadas. El verbo atributivo ser aparece 65 veces (equivalentes a 97.01%). Los verbos intransitivos parecer y significar son utilizados en una sola ocasión (1.49%).

Tabla 17: Estructura oracional de las relaciones subordinadas sujetivas

Estructura	Total	
	Frec.	%
Introducida por *que*	19	20
De relativo sustantivada	14	14.74
Interrogativa indirecta	5	5.26
De infinitivo	57	60
Total	95	100

Por su parte las sujetivas se actualizan de cuatro formas distintas: de infinitivo con 57 testimonios (60%), introducida por *que* con 19 (20%), de relativo sustantivada con 14 (14.74%) e interrogativa indirecta con 5 (5.26%). El tipo de verbo regente más frecuente en la subordinación sujetiva es el atributivo, hay 46 ocurrencias, igual a 48.42%. Los testimonios de intransitivos abarcan 36.84% y los de transitivos 12.63%. Sólo se dan dos apariciones de verbo de régimen prepositivo como regente de una relación sujetiva.

En el total de la muestra hay 18 verbos distintos rigiendo las oraciones subordinadas sujetivas. El verbo atributivo ser posee la más elevada frecuencia, 44 casos, igual a 46.32%. Gustar ocupa el segundo sitio con 15, equivalentes a 15.76%. Entre estos dos verbos abarcan solos 62.08%. Se dan nueve testimonios con hacer, cuatro con dar, tres con implicar y tocar, dos con apasionarse, estar, parecer, proporcionar y resultar, uno con asemejar, basarse en, dificultarse, encantar, recaer, redundar en y representar. Estos datos hacen pensar en la necesidad de ejercitar el uso de verbos diversos, sobre todo intransitivos, que puedan fungir como regentes de oraciones subjetivas.

Tabla 18: Estructura oracional de las relaciones subordinadas de verbo de régimen prepositivo

Estructura	Total	
	Frec.	%
Introducida por *que*	9	20
De relativo sustantivada	1	2.22
Interrogativa indirecta	4	8.89
De infinitivo	31	68.89
Total	45	100

Las relaciones de verbo de régimen prepositivo se actualizan de cuatro maneras: de infinitivo 31 veces (68.89%), introducidas por *que* nueve (20%), interrogativas indirectas cuatro (8.89%) y de relativo sustantivadas una (2.22%). En las redacciones de estudiantes de LES hay 21 verbos de régimen prepositivo fungiendo como regentes de oraciones subordinadas. Ayudar a posee la mayor frecuencia, nueve casos, igual a 20%, y la más amplia difusión, aparece en tres semestres. Sigue enfocarse en con siete (15.56%), con una difusión de uno. Consistir en ocupa el tercer sitio con cuatro apariciones (8.89%). En cuatro lugar se sitúa pensar en con tres testimonios (6.67%). Estos verbos cubren 51.12%. El resto de los verbos aparecen dos o una vez en un solo semestre. Sólo conformarse con, decidirse por y hablar de se presentan en dos grados. Del total de verbos de régimen prepositivo en séptimo aparecen 10, en tercero ocho, en primero cinco y en quinto cuatro.

6. Conclusión

El presente análisis arroja que el comportamiento de la subordinación sustantiva en estudiantes de LES se asemeja al patrón propuesto por Lope Blanch: las objetivas directas superan en ocurrencia a las otras sustantivas. Por otro lado, se ha comprobado que a mayor grado escolar disminuye el número de relaciones sustantivas, pero la distribución por clases es mejor, ya que disminuye el porcentaje de objetivas directas y aumenta la presencia de sujetivas, predicativas y de verbo de régimen prepositivo. No obstante, la forma en que se estructuran las mismas refleja la carencia de recursos intensionales y léxicos.

La enseñanza planificada de la lengua materna, aquella que parte de las estructuras base que poseen los alumnos como anclas para aprehender las metas, propiciará una distribución

homogénea en el uso de las diversas estructuras sintácticas, así como una mente letrada capaz de comprender la multiplicidad de factores que inciden en su lengua y su entorno.

El docente tiene la misión de forjar hombres y el manejo adecuado de la palabra es fundamental para que pueda hacer comprender a sus alumnos que el mundo es un complejo entramado de relaciones paradigmáticas y sintagmáticas, coordinadas y subordinadas. La toma de conciencia, por parte de todos los integrantes de la sociedad, de la lengua como una herramienta efectiva para la supervivencia es el primer paso para que la preocupación por el proceso de enseñanza aprendizaje tenga buen fin.

REFERENCIAS BIBLIOGRÁFICAS

Alarcos Llorach, E. (1951). *Gramática estructural*. Madrid.

Alcina Franch, J. & Blecua, J. M. (1980). *Gramática española*. Ariel. España.

Arjona Iglesias, M. (2001). *Usos verbales en México y su enseñanza*. Edĕre. México.

Bosque, I. & Demonte, V. (1999). *Gramática descriptiva de la lengua española*. Espasa-Calpe. Madrid.

Demonte, V. (1977). *La subordinación sustantivada*. Cátedra. España.

Gaya, G. (2002). *Curso superior de sintaxis española*. Vox. España.

Halliday, M. A. K. (1982). *El lenguaje como semiótica social*. FCE. México.

Lope Blanch, J. M. (1979). *El concepto de oración en la Lingüística español.* UNAM. México.

Lope Blanch, J. M. (1983). *Análisis gramatical del discurso*. UNAM. México.

Lope Blanch, J. M. (1993). *Nuevos estudios de Lingüística hispánica*. UNAM. México.

López Chávez, J. (1980). *Análisis morfosintáctico*. Colegio de bachilleres/ Centro de Actualización y Formación de Profesores. México.

López Chávez, J. & Arjona Iglesias, M. (2001). *Sobre la enseñanza del español como lengua materna*. Edĕre. México.

López Morales, H. (1984). *La enseñanza de la lengua materna. Lingüística para maestros de español*. Playor. España.

Lyons, J. (1977). *Introducción a la lingüística teórica*. Teide. España.

Luna Traill, E. (1980). *Sintaxis de los verboides en el habla culta de la Ciudad de México*. UNAM. México.

Luna Traill, E. (2007). *Diccionario básico de lingüística*. UNAM. México.

Martínez Lara, M. G. (2006). *Complejidad sintáctica en narraciones de estudiantes de contaduría pública de la Universidad Autónoma de Aguascalientes.* Tesis inédita en Maestría en Enseñanza de la Lengua Materna. UAZ. Zacatecas.

Pérez Molina, M. (1997). *Nivel de desarrollo sintáctico en alumnos egresados de sexto*. Tesis inédita en Licenciatura en Lengua y Literatura Hispánicas. UNAM. México.

Plan de estudios 1997. Licenciatura en Educación Primaria (2006). SEP. México.

Plan de estudios 1999. Licenciatura en Educación Secundaria (2004). SEP. México.

Real Academia Española (1962). *Gramática de la lengua española*. Espasa-Calpe. Madrid.

Roca-Pons, J. (1975). *Introducción a la gramática*. SEP. México.

Rodríguez Guerra, F. (1998). *Las oraciones subordinadas sustantivas en el habla culta en la Ciudad de México.* Tesis inédita en Lingüística Hispana. UNAM. México.

Sholes, R. J. & Willis, B. J. (1998). "Los lingüistas, la cultura escrita y la intensionalidad del hombre occidental de Marshall McLuhan". En Olson, D. & Torrance. N. (Eds). *Cultura escrita y oralidad,* (pp. 285-311). Gedisa. España.