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 I
A-K

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

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*

Medición de fluidez oral en lengua materna y extranjera: percepción y medidas cuantitativas a partir de audio y texto en un corpus de alemán

YURENA ALCALÁ HERNÁNDEZ
Universitat de Barcelona

Resumen
En este estudio se investigan diferentes medidas de fluidez oral en producciones narrativas monologadas en alemán como lengua materna y como lengua extranjera, calculables a partir de archivos de audio, transcripciones ortográficas y transcripciones con pausas medidas. La primera cuestión de este estudio pretende explorar la relación entre la percepción de fluidez por parte del oyente y una serie de variantes de ratio de habla. El objetivo es analizar la fiabilidad de estas variantes estableciendo niveles generales (alto, medio y bajo) y grados (clasificación dentro de estos niveles) de fluidez oral. La segunda cuestión explora formas de combinar ratio de habla con otras medidas de fluidez oral que permitan obtener mediciones más fiables de este parámetro de la competencia comunicativa oral.
Palabras clave: evaluación de fluidez oral, producción oral en lengua extranjera, medidas de fluidez oral, percepción de fluidez oral, ratio de habla, producción oral en lengua materna

Abstract
This study investigates different ways of measuring oral fluency using audio files, orthographic transcriptions and transcriptions with measured pauses from a VARCOM corpus sample (elicited narratives in German L1 and FL). The first research question of this study examines the relationship between different variants of speech rate and listeners' perceptions of fluency. This question aims at analyzing the accuracy of different variants at establishing levels (high, intermediate and low) and sublevels of oral fluency. The second question explores the possibility of combining speech rate with oral fluency measures in order to obtain a more accurate assessment of fluency.
Keywords: oral fluency assessment, foreign language speech production, oral fluency measures, oral fluency perception, speech rate, first language speech production

1. INTRODUCCIÓN[1]

Este estudio gira en torno a la determinación cuantitativa del grado de fluidez en producciones orales narrativas monologadas en alemán como lengua materna y como lengua extranjera.

El término de fluidez oral puede usarse en un sentido amplio, esto es, como sinónimo de competencia comunicativa oral o en sentido estricto, haciendo referencia a un aspecto de esta competencia (Lennon, 1990, 2000). En este último sentido, Lennon define la fluidez como *"an impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently"* (Lennon, 1990: 391).

[1] Este estudio se ha realizado gracias al programa FPU del Ministerio de Ciencia e Innovación. Se enmarca dentro del proyecto de investigación COHESTIL (FFI2008-01230/FILO), financiado por el organismo mencionado.

La fluidez como aspecto de la competencia comunicativa oral aparece como parámetro tanto en el contexto de evaluación de competencia en lengua extranjera como en el de investigación sobre adquisición de segundas lenguas. Dentro de estos contextos se ha intentado establecer qué características de la producción influyen en la percepción de fluidez.

En la investigación en adquisición de segundas lenguas, la fluidez se ha investigado a cuatro niveles: temporal, interactivo, fonológico y léxico (Kormos y Dénes, 2004). Este estudio sigue la línea de las investigaciones de enfoque temporal y fonológico que han intentado averiguar qué aspectos de la producción contribuyen a la percepción de fluidez oral y qué medidas cuantitativas son más adecuadas para evaluar productivamente este parámetro.

En esta línea se han identificado una serie de medidas cuantitativas que correlacionan con evaluaciones perceptivas de fluidez. Entre estas medidas destacan ratio de habla, longitud media de run – esto es, promedio de sílabas producidas entre pausas medidas – (p.e. Freed, 1995; Kormos y Dénes, 2004; Lennon, 1990; Towell, Hawkins y Barzegui, 1996), ratio fonación-tiempo (p.e. Kormos y Dénes, 2004; Lennon, 1990; Towell y otros, 1996) y duración media de pausa (p.e. Kormos y Dénes, 2004; Towell y otros, 1996). En Kormos y Dénes (2004) se destaca asimismo la medida de sílabas acentuadas por minuto, una medida de carácter fonológico aún poco investigada como indicador de fluidez oral.

El ratio de habla es una medida consistente en calcular palabras o sílabas por minuto, esto es, palabras o sílabas producidas entre tiempo de producción total (incluido tiempo en pausas) en segundos, multiplicado por 60. Se trata de una de las medidas más consolidadas como discriminadora de grados de fluidez oral percibida (p.e. Derwing, Rossiter, Munro y Thomson, 2004; Freed, 1995; Kormos y Dénes, 2004; Lennon, 1990; Rossiter, 2009; Towell y otros, 1996). Respecto a las otras tres medidas temporales mencionadas anteriormente tiene además la ventaja de que para calcularla no es necesario identificar ni medir pausas. Sin embargo, el cálculo de esta medida no parece estar estandarizado. Existen distintas variantes en cuanto a la unidad de habla escogida, cada una con un coste temporal propio debido a los diferentes procesos requeridos para su cálculo (véase tabla 1).

Tabla 1: Variantes comunes de ratio de habla

Variantes comunes de ratio de habla	**Unidad**		**Procesos requeridos**	
	Tipo	***Cantidad***	***Cómputo silábico***	***Identificación de discontinuidades***
Palabras por minuto	Palabras	Todas las realizadas	-	-
Palabras por minuto, excluyendo ***discontinuidades***		Sin fenómenos considerados no fluidos	-	+
Sílabas por minuto	Sílabas	Todas las realizadas	+	-
Sílabas por minuto, excluyendo ***discontinuidades***		Sin fenómenos considerados no fluidos	+	+

Los procesos discontinuos o discontinuidades de construcción del discurso (Alcalá, 2009; Gülich y Kotschi, 1996; Hossbach, 1997; Kotschi, 2001) suelen considerarse fenómenos no fluidos, puesto que interrumpen el desarrollo discursivo progresivo y lineal. Algunas medidas basadas en indicadores no verbales (pausas silenciosas) de procesos discontinuos sí discriminan grados de fluidez percibida (duración media de pausa, ratio fonación-tiempo), aunque con menor precisión que las medidas relacionadas con producción de unidades de habla (ratio de habla, longitud media de *run*). Sin embargo, las medidas de frecuencia o número de procesos discontinuos con elementos verbales (por ejemplo pausas rellenas, repeticiones no enfáticas, auto-correcciones, comentarios) investigadas en estudios cuantitativos no suelen dar resultados significativos a la hora de diferenciar producciones con distinto grado de fluidez temporal, esto es, no correlacionan con las evaluaciones de fluidez percibida ni con otras medidas temporales (p.e. Freed, 1995; Freed, Segalowitz y Dewey, 2004; Kormos y Dénes 2004; Lennon, 1990).

En los estudios en los que se utiliza ratio de habla no siempre se menciona la existencia de variantes de esta medida y, por lo tanto, no se argumenta la elección de una variante determinada más allá de referencias a estudios anteriores que, si bien usan la misma variante, tampoco tematizan la posibilidad de un cálculo diferente (p.e. Freed, 1995; Freed y otros, 2004; Kormos y Dénes, 2004; Segalowitz y Freed, 2004; Towell y otros, 1996). Para la opción sílabas o palabras no hemos encontrado estudios dónde se mencionen y usen las dos opciones. En cuanto a la opción con o sin discontinuidades, el uso de sus dos variantes – ratio de habla con discontinuidades verbales (*unpruned speech rate, speech rate A*) y sin ellas (*pruned speech rate, speech rate B*) – es más común (p.e. Derwing, Rossiter, Munro y Thomson, 2004; Gilabert, 2005; Lennon, 1990; Mehnert, 1998; Rossiter, 2009, Yuan y Ellis, 2003), especialmente en la investigación sobre complejidad de tarea (*task complexity*). En estos casos sí se argumenta el uso de ambas variantes, pero no suele contrastarse su grado de precisión diagnóstica. Por ejemplo, en Mehnert (1998) se argumenta que el objetivo de la

variante sin discontinuidades verbales es capturar la influencia de este tipo de fenómenos (Mehnert, 1998: 85). En Gilabert (2005) se justifica el uso de la variante sin discontinuidades a efectos de comparación, citando a Mehnert (1998) y a otros estudios dónde sólo se ha usado ésta última variante, y no las dos (Gilabert, 2005: 210). En ambos estudios se cita a Lennon (1990), un estudio sobre medidas cuantitativas de fluidez oral ampliamente citado que parece haber sido el primero en proponer la variante sin discontinuidades de ratio de habla.

Teniendo en cuenta que cada una de las variantes de ratio de habla implica un coste temporal diferente (o de recursos humanos y económicos) sería útil saber hasta qué punto y en qué casos (lengua materna o extranjera, niveles de competencia o de fluidez) las variantes presentadas son igual de fiables y por lo tanto intercambiables.

2. CUESTIONES DE INVESTIGACIÓN

Puesto que ratio de habla es una de las medidas más consolidadas y menos costosas de calcular, la primera cuestión de este estudio se centra en esta medida como instrumento único de diagnóstico de fluidez oral. Se pretende explorar la precisión de una serie de variantes de ratio de habla a la hora de establecer niveles generales (alto, medio y bajo) de fluidez y grados (clasificación dentro de estos niveles) en producciones monologadas de alemán como lengua materna y como lengua extranjera.

Dentro de los distintos aspectos que parecen influir en la percepción de una producción como fluida, la medida de ratio de habla está relacionada sobre todo con el aspecto de la velocidad. La segunda cuestión de este estudio explora formas de combinar ratio de habla con otro tipo de medidas de fluidez oral que, al reflejar mejor otros componentes como el ritmo o las interrupciones del flujo discursivo, permitan obtener mayor precisión en el establecimiento de niveles y grados de fluidez.

Cuestión 1: Variantes de ratio de habla y fluidez perceptiva

¿Cómo correlacionan grado de fluidez percibida y distintas variantes de ratio habla en diferentes niveles y grados de fluidez en producciones monologadas de alemán como lengua materna y como lengua extranjera?

Cuestión 2: Combinación ponderada de medidas de fluidez oral y fluidez perceptiva

¿Qué ponderación o ponderaciones pueden aplicarse a una combinación de medidas de fluidez oral para que la clasificación de producciones resultante sea lo más similar posible a la obtenida mediante evaluaciones fluidez percibida? ¿Supone la clasificación resultante de

esta combinación una mejora significativa respecto a la obtenida únicamente mediante ratio de habla?

3. DATOS

En este estudio se analizan 60 fragmentos de narraciones en alemán como lengua materna y como lengua extranjera producidas por 52 hablantes (ver tablas 2 y 3). Se trata de narraciones monologadas elicitadas a partir de un estímulo visual, la historia en viñetas Frog, where are you? (Mayer, 1969).

Tabla 2: Información de hablantes: sexo, origen y número de producciones

Grupo	**Sexo**	**Origen**	**Producciones**
ALM**: ALEMÁN LENGUA MATERNA (24 hablantes)**	Hombres (12) Mujeres (12)	Alemania Occidental (6), Alemania Oriental (6), Austria (6), Hispano-alemán con escolarización en comunidades de habla catalana (6)	24: 1 producción por hablante
ALE**: ALEMÁN LENGUA EXTRANJERA (28 hablantes)**	Hombres (10) Mujeres (18)	Bilingües catalán-castellano con escolarización en comunidades de habla catalana (26), monolingües de castellano (2)	**36**: 20 hablantes de producción única; 8 hablantes con dos producciones (previa y posterior a estancia en país de habla alemana)

Tabla 3: Duración de las producciones originales y de los fragmentos analizados

Duración (minutos)	**Producciones**		**Fragmentos**	
Grupo (N)	***ALM (24)***	***ALE (36)***	***ALM (24)***	***ALE (36)***
Suma	40,57	99,82	33,04	54,22
Media	1,69	2,77	1,38	1,51
Mediana	1,50	2,36	1,50	1,53
Desviación estándar	0,70	1,27	0,25	0,14
Min	0,70	1,12	0,70	1,12
Max	3,73	5,80	1,64	1,67
Rango	3,03	4,68	0,95	0,56

Los datos pertenecen al subcorpus alemán-catalán-castellano (Fernández-Villanueva y Strunk, 2009) desarrollado por el grupo LADA (Lingüística Aplicada y Didáctica del Alemán) de la Universidad de Barcelona dentro del proyecto VARCOM (Payrató, Álamo, Fitó y Juanhuix, 2004). Se trata de un corpus audiovisual de entrevistas semi-estructuradas diseñadas para elicitar experiencial y experimentalmente producciones orales semi-espontáneas de distinto desarrollo temático. Los entrevistados son hablantes multilingües con estudios universitarios y edades comprendidas entre los 19 y los 32 años.

Se eligen para el análisis narraciones en formato audio. Este tipo de tarea y formato se han investigado con gran frecuencia en estudios relacionados con fluidez, lo que aumenta la posibilidad de comparación de los resultados que se obtengan Se han desechado las narraciones experienciales (dialogadas) para minimizar la posible influencia del entrevistador en la percepción de fluidez oral. Se han seleccionado, por tanto, sólo las narraciones experimentales (monologadas). Estas producciones se han fragmentado para incrementar la cantidad de producciones evaluadas y la de evaluadores utilizados dentro de un coste temporal razonable en términos de eficacia y eficiencia – práctica común en estudios perceptivos de fluidez oral (p.e. Derwing y otros, 2004; Derwing, Thomson y Munro, 2006; Derwing y otros, 2009; Freed, 1995; Rossiter, 2009).

4. METODOLOGÍA

4.1. Evaluación perceptiva de fluidez oral

Para determinar el grado de fluidez percibida de los fragmentos se recurre a evaluadores expertos (docentes de alemán como lengua extranjera), nativos y no nativos, y a nativos no expertos. En estudios previos de esta área en inglés como lengua extranjera no se han encontrado diferencias significativas en las valoraciones de estos tres grupos (Derwing y otros, 2004, Rossiter, 2009). La evaluación consiste en un juicio holístico mediante escala numérica. Para las producciones en lengua materna se utiliza una escala de 5 niveles. Para las narraciones en lengua extranjera se usan 9 niveles a fin de poder reflejar el mayor rango de variación esperado en estas producciones.

Para orientar a los evaluadores, el concepto de fluidez se define previamente como una impresión general del oyente de poder seguir el discurso sin esfuerzo, impresión a la que contribuyen un discurso de velocidad óptima, ritmo adecuado y con interrupciones (discontinuidades) naturales producidas de forma ocasional. Esta definición es una adaptación de las descripciones de 5 niveles de fluidez elicitadas en Brown, Iwashita y McNamara (2005). El concepto de fluidez también se diferencia explícitamente de otros parámetros de competencia lingüística como la corrección y complejidad gramatical, corrección y densidad léxica, variedad dialectal, correcta pronunciación o calidad y cantidad de contenido.

Una vez obtenidos los resultados de esta evaluación, esto es, el grado de fluidez, se clasifican los fragmentos en 3 niveles (alto, medio y bajo) por grupo (lengua materna y lengua extranjera).

4.2. Variantes de ratio de habla

Las variantes de este estudio resultan del tipo de unidad (palabras o sílabas), la cantidad de éstas (todas las realizadas o excluyendo discontinuidades) y del método utilizado para calcularlas (automático, semiautomático o manual), que a su vez depende en parte del formato base (audio o texto) y en parte de la unidad escogida. De estas posibilidades surgen las siguientes variantes:

Tabla 4: Ratio de habla: Variantes calculadas

Variante	Unidad	Cantidad de unidad	Formato base	Método de cálculo	Software / Procedimiento
PM12A	**PALABRAS**	TODAS	TEXTO	AUTOMÁTICO	Microsoft Word ® (función: contar palabras)
PM22A		SIN DISCONTINUIDADES	TEXTO	SEMI-AUTOMÁTICO	Identificación manual de las discontinuidades Microsoft Word ® (función: contar palabras)
SM11A	**SÍLABAS**	TODAS	AUDIO	AUTOMÁTICO	Praat[2] script[3]
SM12A			TEXTO	AUTOMÁTICO	Contador online: www.leichtlesbar.ch
SM12M			TEXTO	MANUAL	---
SM22A		SIN DISCONTINUIDADES	TEXTO	SEMI-AUTOMÁTICO	Identificación manual de las discontinuidades Contador silábico online: www.leichtlesbar.ch
SM22M			TEXTO	MANUAL	---

Estas variantes se correlacionan seguidamente con las evaluaciones perceptivas de fluidez. Asimismo se calcula para cada una de las variantes el índice de error determinando niveles de fluidez y grados dentro de estos niveles.

4.3. Combinación de ratio de habla y otras medidas de fluidez oral

En este estudio se parte de un concepto de fluidez como impresión del oyente que se basa en gran medida en la velocidad, el ritmo y la frecuencia y tipo de interrupciones del flujo discursivo presentes en la producción del hablante. En algún estudio (p.e. Mehnert, 1998) se ha argumentado que las variantes sin discontinuidades de ratio de habla reflejan la influencia de este tipo de interrupciones del flujo discursivo. También se ha argumentado (p.e. Yuan y Ellis, 2003) que la medida de ratio de habla tiene en cuenta en todas sus variantes las interrupciones no verbales (pausas silenciosas), al ser el divisor el tiempo de producción total

[2] Boersma, P. y Weenink, D. (2010). Praat: doing phonetics by computer (Versión 5.1.25) [Programa informático] Descargado el 20 de enero de 2010 en http://www.praat.org/.
[3] De Jong y Wempe (2009).

y no el de fonación. Sin embargo, la ratio de habla es una medida relacionada sobre todo con el parámetro de la velocidad. A fin de intentar establecer un perfil de fluidez basado varias medidas ponderadas y combinadas, escogemos las medidas de sílabas acentuadas por minuto y de duración media de pausa. Estas medidas se relacionan respectivamente con los parámetros de ritmo y de interrupciones, complementando así a ratio de habla.

Las variantes resultantes del formato base y del método de cálculo para estas dos medidas y el procedimiento seguido para su cálculo se resumen en las siguientes tablas:

Tabla 5: Duración media de pausa: Variantes calculadas

Variante	**Formato base**	**Método de cálculo**	**Software /Procedimiento**
DMP1A	AUDIO	AUTOMÁTICO	Praat (función: to textgrid (silences))
DMP2M	TEXTO	SEMI-AUTOMÁTICO	Praat (función: to textgrid (silences)) Revisión, medición y atribución de pausas al hablante correspondiente manuales

Tabla 6: Sílabas acentuadas por minuto: Variantes calculadas

Variante	**Formato base**	**Método de cálculo**	**Software / Procedimiento**
SAM1A	AUDIO	SEMI-AUTOMÁTICO	Praat script[4] para identificación de sílabas Localización manual de sílabas con mayor intensidad
SAM1M	TEXTO	MANUAL	Identificación y transcripción de sílabas acentuadas por parte de hablante nativo

A partir de la magnitud de correlación y del coeficiente de determinación de estas medidas y de ratio de habla respecto a las evaluaciones perceptivas de fluidez escogemos una serie de ponderaciones posibles y se aplican a la mitad de las producciones para determinar la más adecuada. La ponderación escogida se aplica a la otra mitad de las producciones para examinar su fiabilidad. A continuación se comprueba si la evaluación de fluidez oral obtenida a partir de esta combinación ponderada de medidas mejora significativamente respecto a la obtenida con ratio de habla como medida única.

En la presente comunicación discutimos los resultados obtenidos, valorando las distintas medidas y sus variantes según precisión y coste temporal.

[4] De Jong y Wempe (2009).

REFERENCIAS BIBLIOGRÁFICAS

Alcalá, Y. (2009). Construcción discontinua del discurso en alemán como lengua materna: Aspectos de la relación entre procedimientos de verbalización y de referencia. *Revista de Filología Alemana*, *17*. (pp. 159-181). Disponible en www.ucm.es /BUCM/revistas/fll/11330406/.../RFAL0909110159A.PDF

Brown, A., Iwashita, N. y McNamara, T. (2005). An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks. (*TOEFL Monograph Series #MS29*). Princeton, NJ: Educational Testing Service.

De Jong, N.H. y Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, *41*(2). (pp. 385-390).

Derwing, T. M., Rossiter, M. J., Munro, M. J. y Thomson, R. I. (2004). L2 fluency: Judgments on different tasks. *Language Learning, 54*. (pp. 655-679).

Derwing, T. M., Thomson, R. I. y Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System, 34*. (pp. 183-193).

Fernández-Villanueva, M. y Strunk, O. (2009). Das Korpus Varkom – Variation und Kommunikation in der gesprochenen Sprache. *Deutsch als Fremdsprache, 46*(2). (pp. 67-73).

Freed, B.F. (1995). What makes us think that students who study abroad become fluent? En B.F. Freed, (Ed.). *Second Language Acquisition in a Study Abroad Context* (pp.123-148). Amsterdam: Benjamins.

Freed, B.F., Segalowitz, N. y Dewey, D.P. (2004). Context of learning and second language fluency in French. *Studies in Second Language Acquisition, 26*. (pp. 275-301).

Gilabert, R. (2005). Task Complexity and L2 Narrative Oral Production. Tesis doctoral (Universidad de Barcelona). Disponible en: http://www.tdx.cat/TDX-1220105-085713.

Gülich, E. y Kotschi, T. (1996). Textherstellungsverfahren in mündlicher Kommunikation. Ein Beitrag am Beispiel des Französischen. En W. Motsch (Ed.), *Ebenen der Textstruktur: sprachliche und kommunikative Prinzipien* (pp.37-80). Tübingen: Niemeyer.

Hossbach, S. (1997). *Zur Redewiederaufnahme im Diskurs*. Münster: Lit.

Kormos, J. y Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System 32*. (pp. 145-164).

Kotschi, T. (2001). Formulierungspraxis als Mittel der Gesprächsaufrechterhaltung. En K. Brinker, G. Antos, W. Heinemann y S.F. Sager (Eds.). *Text- und Gesprächslinguistik :*

ein internationales Handbuch zeitgenössischer Forschung (Vol. 2). (pp.1340-1348). Berlin, New York: de Gruyter.

Lennon, P. (1990). Investigating fluency in EFL: a quantitative approach. *Language Learning 40*(3). (pp. 387-417).

Mayer, M. (1969). *Frog, where are you?* New York: Dial Press.

Mehnert, U. (1998). The effects of different lengths of timer for planning on second language performance. *Studies in Second Language Acquisition*, *20*. (pp. 83-108).

Payrató, Ll., Àlamo, M., Fitó, J. y Juanhuix, M. (2004). El projecte VARCOM. Variació, comunicació multimodal i multilingüisme: estils discursius i consciència lingüística en la producció de textos orals. En Ll. Payrató, N. Alturo y M. Payà (Eds.). *Les fronteres del llenguatge. Lingüística i comunicació no verbal.* (pp. 111-121). Col·lecció Lingüística Catalana, 7. Barcelona: Promociones y Publicaciones Universitarias (PPU).

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review, 65*(3). (pp. 395-412).

Segalowitz, N. y Freed, B. (2004). Context, contact and cognition in oral fluency acquisition: learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition* 26. (pp. 173-199).

Towell, R., Hawkins, R. y Barzegui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics 17*(1). (pp. 84-119).

Yuan, F. y Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*(1). (pp. 1-27).

Sense-division and grammatical information in the *Dictionary of Adjectival Complementation in Old English (DACOE)* and in the *Dictionary of Old English (DOE)*

ALEJANDRO ALCARAZ SINTES

Universidad de Jaén

Abstract

Old English adjectives permit or require different complementational patterns, semantic and syntactic. This paper describes and shows how these patterns are used to create senses and subsenses in the Dictionary of Adjectival Complementation in Old English (DACOE) that I am presently compiling, containing some 600 adjectives. This is compared to the approach taken by the editors of the Dictionary of Old English, where argumental and syntactic information is not presented in a systematic manner. Each entry in the DACOE is divided into twelve sections, namely, lemma, definition, synonyms, antonyms, semantically-related adjectives, translation equivalents, collocations, syntactic function, argumental structure, syntactic structure, examples and translation of examples. This lexicon accounts for all the patterns found in attested examples from the Dictionary of Old English Corpus. For this purpose, I am using Tshwanelex, a dictionary compilation software that proves convenient because of the clarity of the layout of the different entry fields, the possibility to view the result while working on the entries and the existence of hyperlinks or cross-references between adjectives.

Keywords: adjective, argument, complement, complementation, dictionary, lexicography, Old English

Resumen

Los adjetivos del inglés antiguo permiten o exigen diferentes estructuras de complementación, semánticas y sintácticas. Esta comunicación describe y muestra cómo se usan estas estructuras para crear las entradas y subentradas en un diccionario de la complementación del adjetivo en inglés antiguo (DACOE) , que incluye unos 600 adjetivos y que se halla en fase de compilación. Esta presentación se compara con la adoptada por los editores del Dictionary of Old English, en el que no se presentan las características argumentales y sintácticas de una forma sistemática. Cada entrada se divide en doce secciones, a saber, lema, definición, sinónimos, antónimos, adjetivos semánticamente relacionados, equivalentes de traducción, colocaciones, función sintáctica, estructura argumental, estructura sintáctica, ejemplo y traducción de ejemplos. Este lexicón da cuenta de todas las estructuras observadas en ejemplos atestiguados del Dictionary of Old English Corpus. Con este propósito, estoy utilizando Tshwanelex, un compilador lexicográfico que destaca por la presentación clara de las mencionadas secciones, por la posibilidad de ver el resultado mientras se trabaja en las entradas, así como por la existencia de enlaces o referencias cruzadas entre adjetivos.

Palabras clave: adjetivo, argumento, complementación, complemento, diccionario, inglés antiguo, lexicografía

1. INTRODUCTION

The purpose of this paper is to analyze and compare how syntactic and semantic information, specifically such as regards the syntactic and semantic complementation of adjectives, is presented in two lexicographic words: the *Dictionary of Adjectival Complementation in Old English* (*DACOE*) that I am currently compiling, and the major standard dictionary of Old

English (OE), that is, the monumental *Dictionary of Old English* (*DOE*; diPaolo Healey *et al.* 2007), which so far covers *A* to *G*.[1]

The objective of the *DACOE* is different from that of the *DOE*. We do not intend it to be used so much as an aid to translation, but as a linguistic description of lexical items (adjectives) from which comparisons can be established with other word-classes (e.g., verbs or nouns) in OE synchronic studies, or as a starting point for diachronic studies on the subsequent evolution of complementation structures in Middle English or Modern English. In other words, we are interested in registering not all the semantic meanings and nuances that were created, for example, as a result an adjective collocating with specific words or being used in a particular register or genre, but rather all corpus-attested argumental and syntactic patterns. Thus, the *DACOE*, a single-manned project, is not in competition with a collective lexicographic work of outstanding scholarship as the *DOE*.[2] If aims are not the same—we are not writing a translation dictionary, unlike the *DOE* editors—, it follows that the strategies employed in some of the entry fields must also be dissimilar. These similarities and differences are shown in this paper.

2. BRIEF DESCRIPTION OF THE *DACOE*

The *DACOE* is a major development from my research on the complementation of OE adjectives (Alcaraz-Sintes 2006a: 121). Before we started compiling the dictionary, we announced imminent inception of the work at the HEL-LEX 2005 conference in Helsinki (see Alcaraz-Sintes 2006c). Now, five years later and three-quarters of the way through,[3] we present one particular aspect of our work in this forum.

The *DACOE* comprises 600 adjectives selected on the basis of three principles: frequency of occurrence in the *DOEC*, previous knowledge of specific adjectives being complemented (derived from academic literature), and complexity or interest of complementational structures (see Alcaraz-Sintes 2006a: 25-36).

[1] Hall (1960 [1894]) has been left out of our comparative study since no definitions, but only translation equivalents, are given for the headwords, nor is any grammatical information (other than the part of speech) included in the entries. The undoubted utility of this dictionary lies elsewhere. Bosworth & Toller (Toller 1998, 1921) has also been excluded from this comparative analysis for space reasons: a three-way comparison would exceed the permitted limits for this paper.

[2] Indeed, we would like to express our deepest gratitude to Professor diPaolo Healey, editor of the DOE and of the Dictionary of Old English Corpus (DOEC; diPaolo Healey 2009), for allowing us to use the dictionary facilities and materials at Robarts Library for 6 months in 2006, and to the whole editorial team for the warm response to and valuable comments on my work.

[3] Albeit with a few modifications regarding argument labels and entry layout.

We have retrieved[4] all the occurrences in the *DOEC* of each of these adjectives in order to build the *DACOE* corpus. The examples selected from the dictionary naturally belong to this corpus. As regards the dictionary-building process, we are using a software application specifically designed to compile dictionaries, *TshwaneLex*.[5]

The different types of structures liable to complement an adjective are *Noun Phrases* (inflected in the dative or genitive, occasionally the accusative or the instrumental), *Prepositional Phrases*, introduced by an assortment of different prepositions, and *Clauses*, which can be finite (introduced by such complementisers as *þæt*, *hu*, *hwæt*, etc.) and non-finite or infinitival (where the infinitive is a simple infinitive ending in *-an* or an inflected infinitive, that is, *tō -enne*). The *DACOE* accounts for both personal constructions (those in which the subject is realised by a NP) and impersonal constructions (those in which the subject is an anticipatory pronoun (*hit*, *þæt*, *þis*) or a clause (finite or non-finite), or *zero*).

The semantic roles or arguments are a pertinent selection from those suggested in Cook (1998) and Comesaña-Rincón (2001: 40–46); see Alcaraz-Sintes (2006b: 14-18) too. We have loosely adapted their definitions and roles to the description of adjectival predicates: A (AGENT) is the participant who consciously or deliberately takes part in an action; (E) EXPERIENCER is the participant experiencing the quality (emotion, sensation or cognitive process) denoted by the adjective (whether as a result of external agency or not); CAUSE (C) is the entity that accounts for what the E feels; SC (SCOPE) is the entity expressing the extent to which the meaning of the adjective is valid; LOC (LOCATIVE) is the entity where the state of affairs occurs; and TH (THEME) is the participant unconsciously or passively involved in a state of affairs or an action.

2.1. Entry fields in the dacoe

This section presents the different entry fields used in the *DACOE*. The explanations are preceded by the symbols and abbreviations used.

- **Headword**. The spelling is that given in the *DOE*, for A to G, and in Hall (1960 [1894]), for *H* onwards.
- **Participial adjective**. If the adjective is a past participle used as an adjective, the infinitive of the verb is given.

[4] We have used the application *Search & Replace*, developed by FUNDUC Software. <http://www.funduc.com/search_replace.htm> (08/03/2010).

[5] *TshwaneLex* is fully customizable and allows for as many different fields and sub-fields in each entry as necessary. Results can be exported to different formats, including .doc, .rtf , .xml, and .htm, or added to an ODBC database. The sample entries shown in this paper were all originally "generated" with *TshwaneLex*. (*TshwaneLex. Dictionary Compilation Software*. Developed by David Joffe and Gilles-Maurice de Schryver. Tshwane (Pretoria): TshwaneDJe HLT. <http://tshwanedje.com> (11/03/10)).

- **Definition**. The definition is phrased in such a way that the arguments allowed or required are made obvious, though not their actual linguistic realization.
- **Synonyms**. Synonymous adjectives are linked to other adjectives in the *DACOE*. Simbol: =.
- **Antonyms**. Antonymous adjectives are linked to other adjectives in the *DACOE*. Symbol: ≠.
- **Collocates**. Collocates may be coordinate adjectives, adjectives appearing very near the headword, nouns or phrases recurrently used with the headword (subjects or modified nouns). Symbol: ≈.
- **Syntactic function of the adjective**. We indicate whether the adjective is used as Complement of the Subject (C_S) or of the Object (C_O) in a Predicative (Pred.) construction, in a Supplementive verbless clause (Supp.), or as a Modifier (Mod.) of a Noun in attribution (Att.) or postposition (Post.).
- **Argumental structure**. The arguments involved are listed here. Since the argumental structure belongs to the sense of an adjective, it is used for sense division.
- **Clause structure**. The different clause elements are listed here. Actual syntagmatic order is irrelevant. Optionality is indicated by means of round brackets. Square brackets indicate a relevant element outside the complementation structure, normally an Adverbial. Example: S-V-Adj-C-[A] or S-V-Adj-(C).
- **Verb**. When the adjective is C_S or in a supplementive clause, the verbs found in the corpus of examples are listed here, with an indication of the role or nature: copula (Cop), intransitive verb (Int.), others. For instance: V= Cop: *beon/wesan, weorþan*; Int.: *arisan, cuman, feran, gan, hweorfan, wunian*; others: *beran, gecyrran, lædan*.
- **Information of clausal subject**. We indicate the grammatical nature of the subject (S): whether it is a noun phrase (NP) or a clause (Cl), whether its referent is personal (pers) or abstract (abs)…, and we give its semantic role. For example: S = NP; pers, abs; E.
- **Information on the complement**. We indicate whether the complement is optional by round brackets, (C), and we give its semantic role. For example: (C); A. The complements given under the same section are considered to be interchangeable insofar as they materialize the same argument. The grammatical structure of the complement is indicated through standard abbreviations.
- **Other clause elements**. We give information on the clausal elements we consider relevant to understand the semantic structure of the adjective. For example, a subordinate clause (causal, temporal or conditional) may be an argument of the adjectival predicate, e.g., C, (even if it does not materialize as a complement syntactically). The specific conjunctions found are also indicated. For example: [A] = Finite clause (Cl Fin), conditional, *gif*, C.
- **Further collocates**. Collocates specifically found within the complementation pattern are given here, together with a translation. Frequent LOC adverbials are also indicated. For example: ≈ *gedrefed* + *for* / *fram* / *mid* / *of* / *on* + *ansyne* / *dome* / *ege* / *geswencednesse* / *hatheortnesse* / *hathige* / *hreohness* / *irre* / *ormodnesse* / *sare* / *smeagunge* / *sorge* / *spræce* / *stefne* / *wacmodnesse* / *yrmþ* ('troubled + for (God's) face / judgement / fear / affliction / wrath / anger / rough [weather] / ire / despair / sorrow / meditation / sorrow / speech / voice / weakness / poverty'); *blīþe* + *on andwlite* (blithe in countenance).
- **Examples from the *DOEC***. The example field contains three sections: 1) Short *DOEC* reference; 2) example from the corpus; 3) translation into Present Day

English. The adjective and its complements are highlighted in bold. Comments are given within square brackets.

2.2. Comparison of DACOE *and* DOE

This third part of the paper deals with the differences and similarities between the two dictionaries with respect to the following features: headwords, type of definition and translation equivalents, synonyms and antonyms, sense division explicitness of information on arguments and types of complement, and selection, ordering and treatment of the examples from the *DOEC*.

3. HEADWORDS

The first obvious difference between the two works is that the *DACOE* contains adjectives only, while the *DOE* is a general dictionary. Secondly, not all adjectives are included. We have left out those for which no complementation patterns are attested in the *DOEC*. However, when an predicative adjective, for example, is attested with a complement, then the description of its pattern without complements is also given.

A third difference is that we consider some participles as adjectives, and include them as headwords, if certain semantic, morphological and syntactic conditions are met. For examples, past participles denoting a quality or a state are considered adjectives ((GDPref 1 (C) 4.29) *hit is ny afyled mid þy duste eorðlicra dæda* 'it is now defiled with the dust of earthly deeds'). So are those which show comparative inflection or have a negative prefix ((LibSc 16.7) *se þe soðlice unbesceawud ys to specenne he ongytt yfele* 'he who truly is careless to speak / in speaking, he knows evil things'). The adjectival nature of the participle is also manifest when it is modified by an intensifying adverb or is coordinated to an adjective ((Or3 7.65.25) *for eow Romanum sindon þa ærran gewin swa welgelicad & swa lustsumlice on leoðcwidum to gehieranne* 'why to you Romans preterite victories are so likable and pleasant to listen to in poems').[6]

Therefore, to give an example, the *DACOE* contains a past participle *ābysgod*, while the *DOE*, s.v. *ābysgian*, indicates that it is used "mainly in passive constructions", though in some of the illustrations offered they are clearly adjectives, in our view, e.g., (Bo 41.143.7) *Se wæs swiðe abisgod mid ðære ilcan spræce* 'he was very troubled with that same speech'.

[6] More on this in Sweet 1891-1898 I: 96, Jespersen 1909-49 IV: 230, Visser 1963-73: §§985 and ff., 1145, 1789, p. 1931, Nickel 1966, 1967, Mitchell 1981: §§101, 975 and ff., Granger 1983, Comesaña-Rincón 1986: 264-72, Denison 1993: 372 and ff.

Likewise, we include a present participle *forhtigend* on account of examples such as (HomS 12 18) *And heo beoð þonne ealle swyðe forhtiende hwider heo þonne hwyrfan sculon* 'And then they will all be very afraid whither they must go then', while the *DOE*, s.v *forhtian* 1.b, gives as definition or translation equivalent "trembling with fear, fearful."

4. Definitions and translation equivalents

The strategy used to write a definition and give the senses of a headword depends to a very great extent on the purpose of the dictionary. In the *DOE* online website,[7] the editors explain that they consider themselves "primarily to be writing a translation dictionary". Their goals are not to provide a "deeply analytical" definition, in the manner of the *OED*, or a very precise but "misleading" contextual gloss. Rather, they aim "to provide a good translation, to outline the range of meaning and application of a word, to indicate typical collocations and uses, and to present a clear overview". To achieve this purpose, the entries are "partly analytical, partly contextual, and partly translation-oriented".

The purpose of the *DACOE* is quite different. We want to provide information about the complementation patterns of adjectives that we have found in the *DOEC*. Therefore, as Herbst et al. (2004: xxxviii) said of their dictionary, we "do not claim to provide an extensive semantic analysis of the word in question". Besides, our strategy for defining is not just to indicate the meaning(s) of an adjective through translation equivalents, but to phrase the definition in such a way that the complementation properties of an adjective are reflected most clearly, to paraphrase Herbst et al. (2004: xxxvii).

However, the most usual strategy for defining adjectives in the *DOE* is through translation equivalents, even though other methods are certainly used. For example, for the adjective *blīþe*, the *DOE* gives two major senses, indicated by means of translation equivalents:

1. joyful, happy;
2. mild, gentle.

The *DACOE*, on the other hand, presents three different senses through definitions, followed by translation equivalents:

[7] "The Content and Format of the Entries" <http://tapor.library.utoronto.ca/doe/dict/help/sample.html#sense> (09.03.2010).

1. feeling happiness on account of something; content, happy, joyful, merry
2. having or showing kindness to somebody; beneficent, gentle, kind, mild, well-disposed
3. having qualities pleasing to somebody; agreeable, delightful, pleasant

It may seem that, for this particulary adjective, the *DACOE* contains an extra sense not covered by the *DOE*. Of course, this is not the case: our third sense is duly recorded in the *DOE*. The reason is simply the approach taken to provide the senses and to subdivide them, as will be seen further down.

However, for other adjectives, the *DOE* does give definitions, followed by translation equivalents. For example, s.v. *drēorig*, we find:

1. suffering anguish, grief, horror or misery
1.a agitated, anguished, distraught; grief-stricken, sorrowful, doleful
1.b wretched, miserable; *drēorig heap* (of the sinful on Doomsday)
2. causing anguish, grief, horror or misery; grievous, horrible, grim

This, as we have seen with *blīþe* above, is the procedure followed in the *DACOE*, but we do not descend to such specific meanings as, for example, that which *drēorig* acquires in combination with *hēap* (sense 1.b. in the *DOE*): 'of the sinful on Doomsday'. But, although the basic sense division is the same in both dictionaries, given the purpose of the *DACOE*, our phrasing is more consistent with that used to define *blīþe*, unlike the *DOE*.

1. feeling sorrow on account of something; anguished, distressed, miserable, sorrowful
2. causing sorrow or grief to somebody; grievous, horrible, sorrowful

Sometimes, the wording of a definition in the *DOE* is somewhat misleading, for it seems to point at a particular complementation structure, but it does not. For example, under sense 1 of *blīþe* (joyful, happy), we find that subsense 1.c says "cheerful, willing (do to or suffer something)". We expect the examples to show the adjective used predicatively (as C_S) and complemented by an infinitive clause. Yet, in the four examples provided, the adjective modifies a noun and is not complemented. The referent in three cases is non-personal: *andwlitan* ('countenance') and *mōd* ('spirit'). Besides, the meaning of "suffering" indicated in the parenthetical explanation derives in two cases from the verbs used in the context (*forberan* 'bear', *þolian* 'suffer'). Still, the translation equivalents and the explanation, the subsense indeed, in the *DOE* is not wrong, but it shows that, since the purpose of the *DOE* is different from that of the *DACOE*, the strategies employed are also different.

Finally, it should be noted that no translation equivalent can really convey exactly the same meaning as the headword, nor can the synonyms and antonyms offered in the *DACOE*. This is the reason why both dictionaries offer more than one equivalent and why the *DACOE* (but not the *DOE*) arranges translation equivalents (and synonyms and antonyms) alphabetically, so that the user is not misled into thinking that those listed first are more exact. In the *DOE*, however, the definitions of some words (or, as is often the case, the translation equivalents), are "arranged to reflect the semantic development" of the word, or else the most common, or general, or literal senses are given first.[8]

5. SYNONYMS AND ANTONYMS

One important feature of the *DACOE*, not found in the *DOE*, is the inclusion of three fields in the entry, underneath each of an adjective's senses: a field for synonyms, one for antonyms and, when pertinent, a field for semantically related words. All of them are also headwords in the *DACOE* and are hyperlinked to their entries.

The inclusion of these fields is—we believe—very helpful, in that it allows users to compare the complementational patterns of different but related adjectives, and to check whether these patterns are recurrently used with adjective of a similar meaning or semantic class.[9]

The selection of the adjectives to be included in these fields was made exclusively by studying the meanings, the examples and the complementation patterns of the adjectives while we were compiling the dictionary.[10] However, just as it is often hard to find a good translation equivalent, it proves just as difficult to find a real synonym. For example, in the case of *blīþe*, the following synonyms and antonyms are listed in the *DACOE* for each of the three meaning provided.[11]

> 1. feeling happiness on account of something
> = *fægen*, *fægnigend*, *glæd*[1], *lustfulliend*, *þancfull*[2]
> ≠ *earm*[1], *gedrefed*[1], *geunrotsod*, *sarig*, *unrot*
> 2. having or showing kindness to somebody
> = *ardæde*, *arfull*[4], *eaþe*, *eaþmōd*[3], *fremsum*, *glæd*[2], *mild, mildheort*, *rumheort*[1]
> ≠ *dreorig*, *egesfull*, *reþe*, *wraþ*

[8] "The Content and Format of the Entries" <http://tapor.library.utoronto.ca/doe/dict/help/sample.html#sense> (09.03.2010).
[9] If publication of the *DACOE* is electronic (htm format), the synonyms and antonyms would be linked to their corresponding entries.
[10] We have occasionally consulted Roberts and Kay (1995) to confirm our observations.
[11] The superscript numbers following the adjectives indicate the sense number in their corresponding entry.

3. having qualities pleasing to someone
eaþe[2a], *gecweme*[1], *glæd*, *lustsumlic*, *þancfull*[1]
≠ *sarig*, *sariglic*, *sarlic*, *unrot*

As would be expected, the *DACOE* definitions of these adjectives resemble that of the headword in question, from which they are linked. For example, the *DACOE* definitions of the adjectives synonymous to *blīþe* in sense 1 are:

1. *fægen* and *fægnigend*: feeling joy on accound of something
2. *glæd*[1]: feeling happiness on account of something
3. *lustfulliend*: feeling pleasure or joy on account of something
4. *þancfull*[2]: feeling pleasure or contentment on account of something

6. SENSE DIVISIONS, ARGUMENTS AND COMPLEMENTS

The *DOE* editors assume that "the senses of the words under consideration are essentially the same" and they "let the evidence force [them] into the creation of subsenses". When "more elaborate definitions" or "finer subdivisions" than those in Toller (1998, 1921) are made, the reason is that they have a "broader body of evidence at [their] disposal", but it "may merely be a matter of approach".[12] Thus, the two major senses of *blīþe* that we saw in the previous section are further subdivided, but the rationale behind the division is not clear:[13]

1. joyful, happy
    1.a. joyful, exultant; jubilant, merry
    1.b. happy, content; free from care, at ease
        1.b.i happy (about something *gen*., *for* / *to* and *dat*.)
    1.c. cheerful, willing (to do or suffer something)
    1.d. of objects, places, periods of time: pleasant, agreeable, delightful
2. mild, gentle.
    2.a. of one's demeanour, appearance: cheerful; gentle, serene
    […]
    2.b. generally: gentle, mild; kind, beneficent, gracious
        2.b.i. kind, gracious; well-disposed, beneficent (to someone dat.)

It may be seen that the same grammatical feature—the form of the complement—is used to create sub-senses at different levels: 1.b.i (for inflected NP and PP) and 1.c (for infinitive clause, as may be presumed from the phrasing). Likewise, subsense 2.b.i seems to exist to cover those examples where the adjective is complemented by a dative NP with personal reference, but this is not really a new semantic division, rather a grammatical one. In

[12] "The Content and Format of the Entries" <http://tapor.library.utoronto.ca/doe/dict/help/sample.html#sense> (09.03.2010).
[13] Our claim must not be taken to mean that the sense division in the *DOE* is wrong or misleading or incomplete, of course.

fact, the translation equivalents for 2.b.i are the same as those for 2.b, except for "mild". On the other hand, the type of referent of the noun predicated of or modified by the adjective is explicity stated in 1.d—"objects, places, periods of time"—, or in 2.b—"demeanour, appearance." However, no mention is made of the type of referent in the other divisions.

With some headwords, it is the type of referent (a person or a thing) of the noun predicated of or modified that constitutes the criterion for subdividing a sense in the DOE. Thus, for cearig, the sense division is the following:

1. full of care
    1.a. of persons: sorrowful, anxious, troubled: *carig ymb* 'anxious about'
    1.b. of things: grievous, troublesome
2. taking care, attentive (in attitude or action)

Our sense division, on the contrary, is based on the complementation patterns that are explicitly stated in special fields, where we also inform on the type of reference of the noun. The "formulae" used to present the information is standardized througout the *DACOE*. Thus, our entry reads:

1. feeling troubled on account of something
    E = S = NP; pers
    TH = C = $PP^{æt,\ ymb}$; abs
2. showing vigilance or watchfulness towards something
    A = S = NP; pers
    TH = C = $PP^{tō}$; abs
3. causing pain or grievance to somebody
    A = NP; phys
    E = |

The correspondence is the following: senses 1.a, 2 and 1.b in the *DOE* correspond to senses 1, 2, 3 in the *DACOE*.

Also, other types of grammatical information are always explicitly and formally mentioned for all adjectives in the *DACOE*, which is not the case in the *DOE*. For example, within each sense, between the 'antonyms' and the 'complementation structure' fields, a major division is established on the basis of the syntactic function of the adjective, whether it is a C_S or a C_O in a clause or a modifier of the noun in a NP (in which case we indicate whether the adjective is used attributively or in postposition).[14] Thus, we may have "Pred C_S, C_O, Supp" or "Mod Attr Post". Likewise, when the adjective is used predicatively, we include a special field to inform about which verbs are used in the clause, rounding off the information provided in the collocation field. For example, s.v. *afyrht*, "seized with fear of

[14] Often the adjective belongs to a subjectless supplementive clause (see Quirk *et al.* 1985: 1124-1127), which is sometimes difficult to distinguish from a postpositive adjective modifier.

something or on account of something", there are two different syntactic structures, one in which the argument C is not expressed syntactically, and one in which it is. For each of them the verbs are given:

Pred. $C_{S,}$ Supp.;˜ E-C; S-V-Adj-(C)-[A]
V = Cop: *beon*/*wesan*, *weorþan*, *beon*/*wesan geworden*; Intr. *standan*, *sittan*, *arisan*, *feallan*, *feran*, *fleon*, *tofesian*.
Pred. $C_{S,}$ Supp.;˜ E-C; S-V-Adj-C-[A]
V = Cop: *beon*/*wesan*, *weorþan*; Intr. *standan*, *feallan*, *fleon*.

Finally, in order to show how lexicographic choices may be affected by the type of dictionary being compiled or its overall purpose, we now turn to the way information on syntactic complementation is given in the two dictionaries. We will use the adjectives *behygdig* and *glæd* for this purpose. The adjective *behygdig* is attested in only 6 occurrences in the *DOEC*, while *glæd* has 100 occurrences. The treatment given to them by the *DOE* clearly reflects this as regards the manner in which grammatical information is conveyed.

In the case of behygdig, the DOE gives two senses, defined through translation equivalents, followed directly by 4 and 2 examples, respectively.

1. heedful (of), attentive (to)
2. concerned.

Practically no grammatical information is provided at all. The only inference we can make from the wording of sense 1, or rather, from the fact that the prepositions are given in brackets, is that a complement is optionally allowed. But it is through reading the examples that users of the *DOE* can really retrieve grammatical information: from one example we learn that the adjective can be used as an attributive modifier (*mid behygdine mode* 'with attentive spirit'), from two examples we learn that it can be used with the copulas *wesan* 'be' and *weorþan* 'become' as C_S and that the TH argument is realized by means of either a PP headed by *be* or a genitive NP (*heo wearð bihygdid be þissum* 'she was attentive to these'; *wes þu behydig & gemyndig Marian þinga* 'be thou heedful and solicitous of Mary's things'). Also, we learn that an Infinitive clause may also be used as complement, though the adjective is substantized (*se behydegæsta þa* [the rules] *to healdenne* 'the most heedful to hold [the rules]'). In the two examples for sense 2, the adjective is predicative, in one of which it is a C_O complemented by a PP headed by *be* (*wæs seo abbudisse* […] *bighygdig & sorgende* 'the abbess was concerned and troubled'; *þu me næfre behydigne and sorhfulne be þise wise ne læte* 'do not ever leave me troubled and sorrowful about this way').

On the other hand, the *DOE* treats adjective *glæd* with a far greater level of detail. Thus, sense 2 is divided in the following way:

2. of animate beings: glad, cheerful, joyous
  2.a. glad, cheerful, joyous in disposition
    2.a.i *glæd on mode* 'glad / cheerful of heart'
    2.a.iii of the body personified: glad, cheerful, joyous
    2.a.iv of a plant personified: glad, cheerful, joyous
    2.a.v of the Chruch personified: glad cheerful joyous
    2.a.vi used as a plural substantive: the glad / cheerful / joyous
    2.a.vii specifically: glad, cheerful, joyous on account of a particular circumstance, event, etc.
      2.a.vii.a. with cause of joy indicated contextually (by subordinate clause or independent sentence)
      2.a.vii.b. with cause of joy indicated by a preposition: *glæd fore* / *in* / *on* 'glad, joyous because of / about (something); in Northumbrian gospels
      2.a.vii.c with cause of joy indicated by dependent genitive
  2.b of a person: glossing *ludibundus* 'playful, light-hearted'
  2.c. of a person: glossing *alacer* 'enthusiastic, eager in action'
  2.d. of people / God / Christ: well-disposed, agreeable, pleasant; ? kind, gracious
    2.d.1. well-disposed towards (someone, *wið* and *dat*. / *acc*.)
    2.d.2. rendering *placatus* 'kindly-disposed' (freq. as a result of being appeased)

Three things may be observed in this sense division. Firstly, the criteria used for sense division are unequal and, therefore, incompatible with our aims in the *DACOE*: the presence of a locative PP for sense 2.a.1, the specific type of referent of the noun predicated of or modified for senses 2.a.iii to v, the ability of the adjective to be used a plural collective noun for sense 2.a.vi, and the presence of a C argument realized as an adverbial or a complement for senses 2.a.vii.a to c.

Secondly, the underlying arguments for senses 2.a and 2.d are not the same: E and C for the former, and A and E for the latter. This alone will shape the phrasing of our definitions and the major sense division in a manner different from that of the *DOE*, as we shall see below.

Thirdly, in sense 2.a.vii, we find the explanation "on account of a particular circumstance, event, etc.". In our view, this is also the case for sense 2.a, since the argument C is also present, albeit not realized, in the examples.

Unlike the *DOE*, the *DACOE* treats both adjectives in a similar manner, notwithstanding the disparity in the number of occurrences in the *DOEC*. It should be noticed

that the formal realization of the complements is indicated in the same place, consecutively, and not in different senses, as was the case of *glæd* in the *DOE*.

behygdig

1. 'showing awareness of something, paying attention to something'
   attentive, heedful
   Pred. C_S; A-TH; S-V-Adj-C
   = *behealden*, *gemyndig*[1aiii]
   S = NP; pers; A
   C = NP *gen*, PP *be*; abs; TH
   (HomS 21 (BlHom6) 36) *Martha, Martha, wes* ***þu behydig*** *7 gemyndig Marian* ***þinga*** 'Martha, Martha, you were attentive and solicitous about Mary's things' ● (LS 7 (Euphr) 47) *And* ***heo*** *wearð* ***bihydig be þissum*** *[þæt ece lif]* 'And she became attentive to this [eternal life]'
   Mod. Att; A-TH; Adj-N-(C)
   N = con; A
   C = │˜ Inf. *tō -enne*; abs; TH
   (Bede 4 3.264.30) *Ða wunade he ðær sum fæc tide wundriende 7 wafiende, 7 mid* ***behygdige mode*** *þohte 7 smeade, hwæt þa þing beon sceolde* 'Then he remained there some time wondering and amazed, and with solicitous spirit thought and pondered what the things must be' ● (Bede 5 18.466.25) *[Acca] wæs in reogolum cyriclicre gesetnesse* ***se behydegæsta þa to healdenne*** '[Acca] was the most heedful in the rules of ecclesiastical institution to observe them'
2. 'feeling troubled on account of somebody or something'
   concerned, troubled, worried
   Pred. C_S; E-C; S-V-Adj-(C)
   = *abysgod*[2], *carfull*[1], *sorgiende*, *sorhful*[1]
   S = NP; pers; E
   C = │; PP *be*; abs; C
   Coll.: behyghig + sorgiende / sorhful 'troubled + concerned / distressed'
   (Bede 4 8.282.28) *Þa wæs* ***seo abbudisse*** *[...] bighygdig 7 sorgende* 'Then the abbess and the mother of the congregation was concerned and sorrowful' ● LS 10.1 ((Guth) 20.76) *Forþon ic þe bidde and halsige, þæt þu* ***me*** *næfre* ***behydigne*** *and sorhfulne* ***be þisse wisan*** *ne læte æfter þinre forðfore* 'Therefore I pray and entreat you that you never let me concerned and distressed about this way after your death'

glæd

1. 'feeling happiness on account of something'
   cheerful, glad, happy
   Pred. C_S, Supp.; E-C; S-V-Adj-(C)-[A^1]-[A^1]
   = *blīþe*[1,3], *fægen*, *fægnigend*, *gecweme*, *lustfulliend*, *lustsumlic*, *rumheort*[2]
   ≠ *earm*[1], *gedrefed*[1]
   ≈ *unrot*[1]
   S = NP; pers; abs; E
   C = NP; *gen*, PP *be*; abs; C
   [A^1] = PP *on*; abs; LOC
   [A^2] = Fin. Cl. *forþon* / *forþam* (*þe*), *mid þy þe*; abs; C
   C = │, NP *dat*, *gen*; PP *for* / *on*; Fin. Cl. *þæt*; abs; C

Coll.: *glæd and blissigende* 'glad and exulting'; *gefeade* + *glæd* 'joyful + glad' (ÆCHom I, 31, 456.25) *Æfre* ***he*** *bið anes modes. and* ***glæd*** *þurhwunað* 'He was always of one mind and remained happy' ● (ÆLet 4 (SigeweardZ), 1051) *Ða geseah Iohannes* ***sumne cniht*** *[...] swiðe* ***glæd*** ***on mode*** 'Then John saw a young man […] very cheerful of spirit' ● (LS 29 (Nicholas) 146) ***Biscopes*** *wæron* ***glæde*** ***forþon þe hi mosten begeton swylcne gefera*** 'Bishops were glad because they are allowed to receive such a companion' ● (HomS 24.1 (Scragg) 265) ***Mid þy þe heo þa hine gesawon swa bismerlice geteodne****, þa wæron* ***hi þæs*** *swiðe* ***glæde*** 'When they then saw him so ignominiously determined, then they were very glad for that' ● (HyGl 2 (Stevenson) 99.1) ***þu*** *crist [...]* ***glæd halgum benum*** 'you Christ […] happy for holy prayers' ● (El 955) ***Sefa*** *wæs þe* ***glædra þæs þe heo gehyrde*** 'Spirit was the gladder on account of that which he heard' ● (JnGl (Li) 3.29) ***friond*** *[...]* ***ðæs brydgumes*** *[...] bið* ***glæd fore stefne ðæs brydgumes*** 'let the bridegroom's friend […] be glad for the bridegroom's voice' ● (LS 9 (Giles) 80) ***[he]*** *asteh þa into þæt scip and for [...] ut on þære sæ* ***glæde*** *and blissigende* ***on his drihtenes mildheortnesse*** '[he] went aboard the ship and sailed […] out into the sea, glad and joyful on his Lord's mercy' ● (HyGl 2 (Milfull) C18.2) ***glæd þæt he yrne weg*** 'glad that he ran'

2. 'having or showing kindness to somebody
favourable, kind, propitious
Pred. C_S; A-E; S-V-Adj-(C)-[A]
= *arfæst*[2], *bliþe*[2], *eaþmod*[3], *mild*[1], *mildheort*[1], *rummod*[3]
≠ *bewependlic*[3], *dreorig*[2], *earm*[2], *egesfull*[1], *geswincful*, *þrealic*[1], *reþe*, *sorhful*[2], *sorhlic*[1]
S = NP; pers, abs; A
C = ⎪, NP *dat*, PP *wiþ*; pers; E
[A] = PP *on*; abs; SC
(HyGl 2 (Stevenson) 129.2) *beon dagas bliþe 7* ***glæde nihta*** 'days will be blithe and nights beneficent' ● (LS 9 (Giles) 306) **He wæs […] wacol on gebedu and glæd on gastlice dæde** 'He was vigilant in prayers and kind in spiritual deeds' ● (Ch 779 (Rob 48) 12) *þa munecas libban heora lif æfter regole þæs halgan Benedictes [...] þæt we* ***þone hælend*** *habban* ***us glædne*** 'the monks live their life after holy Benedict's rule […] that we may have the Lord favourable to us and he may direct us' ● (Gen (Ker) 43.14) *Min drihten* ***hine*** *do* ***glædne wið eow*** 'My Lord, make him favourable to you'
Mod. Att; A-E; Adj-N-⎪
N = pers, abs; A
C = ⎪; pers; E
(Beo 1180) *Ic* ***minne*** *can* ***glædne Hroþulf*** 'I know gracious Hrothulf' ● (BenR 5.22) *He is us þeah to gefarenne mid rumheortum mode and mid godum and* ***glædum geþance*** 'He is, however, to go with large-hearted spirit and with good and kind thought'

That the *DOE* is not primarily interested in presenting the grammar of the adjectives, despite subsenses 2.a.vii.a to c for *glæd*, may also be seen in the treatment given to an example of great interest from the point of view of syntax and complementation. The third

sense given by the *DOE* to *glæd* reads: "3. of abstractions, physical entities, etc.". Within subsense 3.c, "of night, the sea: calm, tranquil", we find the first of the following examples, that contains an adverbial realized by a dative NP (*onsiene*). However, in the syntactically more interesting example that follows, there is a SC (SCOPE) argument realized through an infinitive clause complement, but in the *DOE* it is simply added parenthetically and in an abbreviated manner, as if the semantic similarity with the preceding example made it unworthy of full quotation. It follows that the grammar of the construction is not commented on. Besides, there seems to be some kind of incompatibility between the translation equivalents given ('calm, tranquil') and the meaning of the infinitive verb.

> (Met 5.7) *hrioh bið þonne seo* ***þe*** *ær* ***gladu onsiene*** *wæs* (it will be troubled then that had been beautiful of countenance before)
> (Bo 6.14.12) *þonne heo þonne swa gemenged wyrð mid ðan yþum, þonne wyrð heo swiðe hraðe ungladu, þeah* ***heo*** *ær* ***gladu*** *wære* ***on to locienne*** (when [the sea] then becomes so disturbed by/with the waves, then it will quickly become unfair, though she had been beautiful to look at before)

7. COLLOCATIONS

We believe that collocations provide valuable information, not only to identify the best translation equivalent or gloss for a particular context, but also to allow comparison with synonymous adjectives. In this respect, the *DACOE* is also different from the *DOE*: collocations are provided for subject nouns, modified nouns and coordinate adjectives. Thus, to give just one example, in the case of the adjective *behygdig*, the *DOE* does not provide usual collocates, while collocates in the *DACOE* have their own entry field: *behygdig & gemyndig* / *sorgende* / *sorhful*. See also entry for *glæd* above.

8. CORPUS EXAMPLES

There are a few differences between the *DOE* and the *DACOE* regarding examples or citations from the *DOEC*.

First of all, the *DOE* contains many citations with the same argumental and syntactic structure as regards adjectival complementation, while the *DACOE* normally has only one per pattern. For example, the *DOE*, s.v. *gearo* sense 2 'ready, prepared, willing (to do something specific'), provides 5 examples in which the complement is a clause introduced by *þæt*, while the *DACOE* gives only one.

Secondly, the *DOE* gives first "a citation which validates the definition, then a second citation which strengthens the support for the definition, then a few citations which show some slight variation in sense while covering a representative range of genres and spellings".[15] This is clearly not our aim. In the *DACOE*, examples are arranged in the same order as is used to formally present the different complementational patterns and inform on reference types, or different copulas. This may be seen in the examples from the *DACOE* above.

Thirdly, we often abbreviate examples and use square brackets to delete or insert words from the immediate context that we consider necessary for the user to understand the example or the complementation pattern illustrated by the example. For this reason, the *DOE* examples tend to be longer than ours.

Finally, all our examples are accompanied by a translation into Modern English. This is not the policy of the *DOE*.

9. CONCLUSION

This paper is a description of the *Dictionary on Adjectival Complementation in Old English* whose compilation and publication, either electronic or printed, we hope to announce soon. We have explained the reason why we are engaged in this project, where the dictionary derives from, and what the entries and their contents will be like.

We believe that this new dictionary fills an important lacuna in English historical linguistics. In order to demonstrate its value as a work complementing, not competing with, the major lexicographic project currently in progress, the *DOE*, we have attempted to show the similarities and differences between our own lexicographic project, the *DACOE*, and the *DOE*, naturally only as regards adjectival complementation. We have seen how the strategy employed to provide or explain the 'grammar of the adjectives' is closely intertwined with the type of definition provided and with the number of senses and subsenses a headword may be given. We have also explained the reason for the differences between both works, namely, their nature —general vs. specific— and their purpose —providing good translations to convey the smallest semantic nuances vs. providing information on the argumental and complementational patterns of adjectives—.

[15] "The Content and Format of the Entries" <http://tapor.library.utoronto.ca/doe/dict/help/sample.html#sense> (09.03.2010).

REFERENCES

Alcaraz Sintes, Alejandro. (2006a). *La complementación del adjetivo en inglés antiguo*. Jaén: Servicio de Publicaciones de la Universidad de Jaén.

Alcaraz Sintes, Alejandro. (2006b). Old English Ditransitive Adjectives. *Journal of the Spanish Association for Mediaeval English Language and Literature*, *13*, 9-49.

Alcaraz Sintes, Alejandro. (2006c). Proposal for a Dictionary of Syntactic and Semantic Complementation of Old English Adjectives. In R.W. McConchie, Olga Timofeeva, Heli Tissari and Tanja Säily. (Eds.). *Selected Proceedings of the 2005 Symposium on New Approaches in English Historical Lexis (HEL-LEX)*. Somerville (MA): Cascadilla Proceedings Project (pp. 34-40).

Comesaña Rincón, Joaquín. (1986). *La complementación adjetiva en inglés contemporáneo*. PhD thesis. Universidad de Sevilla.

Comesaña Rincón, Joaquín. (2001). Decoding and Encoding Grammatical Information in Adjectival Entries: Processes and Cases. *Atlantis – Revista de la Asociación Española de Estudios Anglo-Norteamericanos 23*(2), (pp. 31-48).

Cook, Walter A. S. J. (1998). *Case Grammar Applied*. Publications in Linguistics 127. Dallas (Texas): The Summer Institute of Linguistics and The University of Texas at Arlington.

Denison, David. (1993). *English Historical Syntax*. London & New York: Longman.

diPaolo Healey, Antonette *et al*. (Eds.) (2007). *Dictionary of Old English: A to G* online. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto. <http://tapor.library.utoronto.ca/doe/dict/index.html> (08/03/10).

diPaolo Healey, Antonette (Ed.) with John Price Wilkin and Xin Xiang. (2009). *Dictionary of Old English Web Corpus*. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto. <http://tapor.library.utoronto.ca/doe/dict/index.html> (09/03/10).

Granger, Sylvia. (1983). *The 'be' + Past Participle Construction in Spoken English – with Special Emphasis on the Passives*. Amsterdam: North Holland.

Hall, John Richard Clark, (1960 [1894]) *A Concise Anglo-Saxon Dictionary*. With a Supplement by Herbert D. Meritt. Medieval Academy Reprints for Teaching, 14. Fourth edition. Toronto, Buffalo and London: University of Toronto Press in association with the Medieval Academy of America.

Herbst, Thomas, David Heath, Ian F. Roe & Dieter Götz. (2004). *A Valency Dictionary of English. A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Topics in English Linguistics 40. Berlin & New York: Mouton de Gruyter.

Jespersen, Otto. (1909-1949). *A Modern English Grammar on Historical Principles*. Copenhagen & New York: Swan Sonnenschein & Macmillan.

Mitchell, Bruce. (1985). *Old English Syntax*. Oxford: Clarendon Press.

Nickle, Gerhard. (1966). *Die Expanded Form in Altenglischen. Vorkommen, Funktion und Herkunft der Umschreibung 'beon/wesan' + partizip präsens*. Neumünster: Wachholz.

Nickle, Gerhard. (1967). An example of syntactic blend in Old English. *Indogermanische Forschungen 72*, (pp. 261-274).

OED: Simpson, Joseph & Weiner, Edmund (Eds.) (2002) [1989]. *The Oxford English Dictionary Second Edition on Compact Disc*. Version 3.00. Oxford: Oxford University Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Startvik. (1985). *A Comprehensive Grammar of the English Language*. London and New York: Longman.

Roberts, Jane & Christian Kay, with Lynne Grundy. (1995). *A Thesaurus of Old English*. London: King's College London, Centre for Late Antique and Medieval Studies.

Toller, Thomas Northcote. (1898). *An Anglo-Saxon Dictionary: Based on the Manuscript Collections of the Late Joseph Bosworth*. Oxford: Clarendon Press.

Toller, Thomas Northcote. (1921). *An Anglo-Saxon Dictionary: Based on the Manuscript Collections of the Late Joseph Bosworth*. Supplement. Also Alistair Campbell (1972) *Enlarged Addenda and Corrigenda*. Oxford: Oxford University Press.

Visser, Frederic Th. (1963-1973). *An Historical Syntax of the English Language*. Leiden: E.J. Brill.

Evidential 'may' in medical scientific abstracts

FRANCISCO ALONSO ALMEIDA

Universidad de Las Palmas de Gran Canaria

Abstract
This paper deals with the role of evidentiality in a corpus of medical abstracts. The main objective is to see whether the modal verb 'may' encodes evidential meaning. To achieve this aim, I will focus on a selection of texts from the Corpus of Specialized Papers in English. My ultimate goal is to evaluate the existence of grammatical evidentiality in English. Modals represent a first step in this study. Admittedly there are contested positions as for the grammatical value of evidentiality, since for many scholars this phenomenon in English only occurs lexically. Notwithstanding, in my view, modals seem to be a good starting point for debate since these forms can only be used when some previous knowledge is presupposed on the part of the writer/speaker.
Keywords: evidentiality, epistemic modality, modal verbs, cognition, stance

Resumen
En este trabajo, se estudia la categoría evidencial en un corpus de resúmenes médicos. El principal objetivo consiste en comprobar si el verbo modal 'may' manifiesta un significado evidencial. Para esto, me centraré en un selección de textos tomados del Corpus of Specialized Papers in English. Uno de los objetivos finales es evaluar la existencia de evidencialidad gramatical en inglés. Muchos estudios han intentado demostrar que la evidencialidad en esta lengua es únicamente léxica; sin embargo, los modales parecen encontrarse en un estadio intermedio entre lo que es gramatical y lo que es léxico. En mi *opinión, los verbos modales pueden manifestar origen de la información sin que necesariamente presenten modalidad epistémica puesto que son dos categorías independientes. De igual manera, un mismo verbo modal puede ofrecer las dos lecturas categoriales.*
Palabras clave: evidencialidad, modalidad epistémica, verbos modales, cognición, 'stance'

1. INTRODUCTION[1]

This paper seeks to explore the use of may as an evidential marker in medical scientific abstracts. In doing so, I will address to key issues in the study of evidentiality in English. The first one is related to the lack of studies in the field for the English language, especially with respect to the marking of evidentiality. There is a linguistic tenet supported by Lazard (2001), among others, which denies the existence of the grammatical evidentiality type to mark source of knowledge in English. Dendale and Tasmowski (2001) are on the other side and claim that grammatical evidentials are possible in this language. They plead for more research in the field in which these markings are identified and categorised accordingly.

The second issue is connected to the concept of epistemic modality and authorial stance in relation to evidentiality. Evidentiality in its broadest sense allows for the interpretation of

[1] Dr. Francisco Alonso is part of the Research Project "Evidencialidad en un corpus multidisciplinar de artículos científico-técnicos en lengua inglesa," grant FFI2009-10801 (FEDER, Spanish Ministry for Science and Innovation). This grant is hereby gratefully acknowledged. Many thanks also are due to two anonymous CILC reviewers for their generous suggestions to improve the paper.

the author's intention and degree of certainty in their use of expressions indicating their sources of knowledge. Previous research in the area suggests that the use of evidential marking rarely is free from the imprint of the author, and therefore other epistemic meanings may be conveyed in evidential markers. However, this line of thought is not a global position, and for some evidentiality and epistemic modality are clear-cut categories, to the extent that a modal can be both evidential and epistemic. This study focuses on the modal verb *may* to show possible epistemic and evidential readings.

The initial corpus of analysis contains 30 abstracts in the field of medicine published in 2008. In this study, I follow Chafe (1986), and Dendale and Tasmowski (2001), among others, for the concept of evidentiality. As for modality, I mainly follow Hoye (1997), Palmer (1996), and Biber et al. (1999). The description of the relationship between evidentiality and epistemic modality is based on Carretero (2004), as well as Cornillie (2009). The paper is organised, as follows: Section 2 offers the description of corpus and the data under analysis. Next section covers a very short description of the main theoretical tenets followed in the description of modality in scientific abstracts. Section 4 presents the analyses and discussion of samples. Last section gives the conclusions drawn from this study.

2. CORPUS AND DATA DESCRIPTION

The data for analysis have been extracted from the *Corpus of Specialized Papers in English*, currently under compilation and tagging in the Instituto Universitario para el Desarrollo y la Innovación en las Comunicaciones (IDeTIC) at the University of Las Palmas de Gran Canaria. The corpus is monogeneric, as it only includes abstracts preceding scientific articles; these have been randomly selected from several databases of scientific journals. Journals include *International Immunology*, *Harm Reduction Journal*, *Academic Emergency Medicine*, *Fetal and Maternal Medicine Review*, and *The New England Journal of Medicine*, The *Cambridge Law Journal*, *Legal Theory*, *Journal of Law and Society*, *International Journal of Computer Vision*, *IEEE Transactions on Software Engineering*, among others. The corpus is divided according to registers into medicine, law, and computing, and later according to chronology, as the corpus covers a time span of ten years: 1998-2008. For the present paper, the 2008 entries of the medical field have been selected.

The corpus texts were published in machine-readable soft format, and hence they are retrieved online, although some of these can be also found in hard format. All the abstracts are written by native speakers of English, and this allows for a unified account of the

findings. The number of abstracts in this study amounts to thirty for the present paper, but in my research prospects I plan to include the complete number of abstracts in the *Corpus of Specialized Papers in English* for representativeness and accuracy of results.

I also envisage dividing analyses according to time in a diachronic dimension of a decade (1998-2008) to examine the degree of variation in the use of modals for this genre. The quick changes in scientific thought and technological advances may have a strong effect on scientific methods and procedures, and this in turn has an effect on language use. More confidence in results may follow from these advances in technology, and this implies bigger authorial commitment among other aspects.

3. EVIDENTIALITY AND EPISTEMIC MODALITY

Evidentiality and epistemic modality are often seen as related concepts. Traditionally, the study of these two concepts overlaps, as in Chafe (1986), although there is certainly a distinction between them. Evidentiality is manifested by means of evidentials, which are used "to qualify the reliability of information communicated in four primary ways. They specify the source of evidence on which statements are based, their degree of precision, their probability, and expectations concerning their probability" (Mithun, 1986: 89). The ongoing debate as to whether evidentiality is one case of epistemic modality, as in Palmer (1986) has led to two different types of evidentiality: (a) broad evidentiality, and (b) narrow evidentiality. The former refers to evidentials as showing the source of knowledge, and the inferred degree of certainty as for the propositions expressed. The latter means that evidentials are only a manifestation of the source of knowledge.

Dendale and Tasmowski (2001) divide the existing approaches to the relationship between evidentiality and epistemic modality into three types: disjunction, inclusion and intersection. A disjunctive relation matches with the concept of narrow evidentiality, and so evidentials imply the evidence of the speaker's utterance (De Haan, 1999: 85). The inclusive type is supported by Palmer (1986) and evidentiality is seen as a subdomain of propositional modality. Finally, the last relation, i.e. intersection, implies an overlap between inferential evidentiality and epistemic necessity (van der Auwera and Plungian, 1998: 86).

As for modality itself, modals affect the meaning of the complete proposition in which they are embedded. According to Hoye (1997), following the modal logic tradition in Von Wright (1951: 1-2), modals can show deontic or epistemic meanings. Epistemic modals are "concerned with matters of knowledge or believe on which basis speakers express their

judgements about state of affairs, events or actions" (Hoye, 1997: 42). In the case of deontic modals, they refer to the "necessity of acts in terms of which the speaker gives permission or lays and obligation for the performance of actions at some time in the future" (Hoye 1997: 43). This twofold distinction of modality coincides with Biber et al.'s (1999: 485) concepts of intrinsic and extrinsic modality, as shown below:

> Each modal can have two different types of meaning, which can be labeled intrinsic and extrinsic (also referred to as 'deontic' and 'epistemic' meanings). Intrinsic modality refers to actions and events that humans (or other agents) directly control: meanings relating to permission, obligation, and volition (or intention). Extrinsic modality refers to the logical status of events or states, usually relating to assessments of likelihood: possibility, necessity, or prediction.

Carretero (2004) proposes an intersective approach to the study of evidentiality and epistemic modality. She sees this intersection in terms of a continuum from evidential to epistemic expressions, and categorises them "depending on the commitment to the truth of the utterance which they encode or implicate" (2004: 27-28). Cornillie (2009) presents a rather different view, and considers epistemic modality and evidentiality as two distinct categories. These are not necessary mutually exclusive, and so a particular modal verb, such as must, may present an evidential reading as well as an epistemic reading. Nuyts (2004) argues that modals cannot show more than one qualification per clause[2]. Confusion in this respect arises from the frequent association of the mode of knowing with the degree of the speaker's commitment as for the proposition manifested. Cornillie (2009) concludes that modes of knowing do not imply any degree of certainty, commitment or likelihood of a future event to be true. In short, an evidential verb is only considered as such in that it shows some kind of mode of knowing without any relation as for how far the author commits himself to the truth of his own proposition. Modes of knowing can be direct or indirect, depending on whether the speaker has obtained the information visually, through his own inferences or from others' inference processes. In this paper, I follow Cornillie's view, and so, whereas evidentiality "refers to the reasoning processes that lead to a proposition", epistemic modality "evaluates the likelihood that the proposition is true". He does reject the inclusive and overlapping combinations to describe the relationship of epistemic modality and evidentiality.

[2] As recorded in Cornillie (2009: 54).

4. THE ANALYSIS OF *MAY*

The corpus has been analysed using CasualConc software (by Yasu Imao) for data retrieval, although manual analyses have been also performed at times. I have firstly produced a list of occurrences in the corpus to learn the presence of modal verbs in texts. After that, I interrogated the corpus for the item may and subforms using wildcards. This revealed that may is by large the commonest verb in the abstracts, amounting to 46.15% of the cases in which a modal is used, as seen in Figure 1, below. The meaning of may as marking possibility and probability makes it ideal to introduce new knowledge at this stage of the paper in which this genre is embedded without fully committing to it. Some examples of may in corpus are the following:

(1) Lowering low-density lipoprotein cholesterol with statin therapy results in substantial reductions in cardiovascular events, and larger reductions in cholesterol may produce larger benefits (JM).

(2) It may coexist with asbestos related to pleural plaques but has a distinctly different pathology (MT).

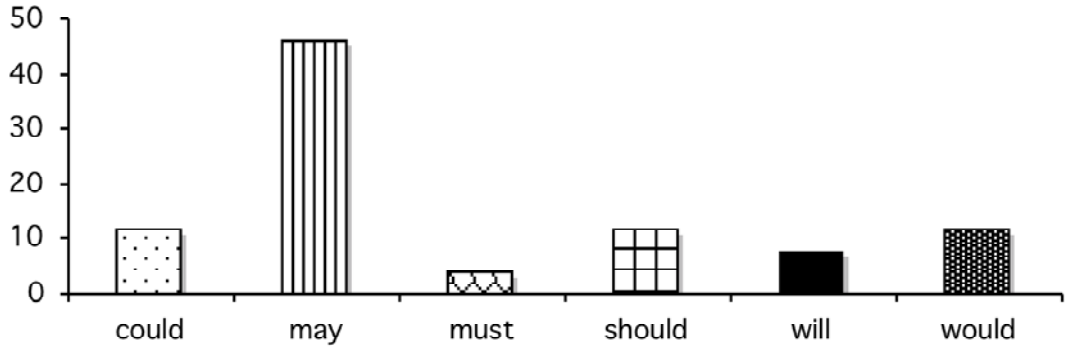


Figure 1: Modals in corpus (percentage).

These two examples will serve as the basis for discussion as to whether modal verbs should be treated as evidentials or not. This is a long debated aspect, not yet solved. Palmer (1986), for instance, includes evidentials as a subcategory of epistemological stance[3], but this is also a matter of controversy, as pointed out in Mushin (2001: 51-83). Aikhenvald (2004)

[3] A definition of epistemological stance is given in Mushin (2001: 58): "construal of information with respect to a speaker's assessment of their epistemological status". This concept relates then to the way in which a speaker shows how he has obtained his information, even if it does not correspond with his actual source.

has plainly manifested that "saying that English has 'evidentiality' (cf. Fox, 2001) is misleading: this implies a confusion between what is grammaticalized and what is lexical in a language. Lexical expressions may, of course, provide historical sources for evidential systems" (2004: 10). With this, Aikhenvald does not clarify whether modals may show epistemic and evidential meaning at the same time. Actually, the author goes over the issue to conclude that modals represent a case in the borderline. Be it as it may, the author considers modals as a case of lexical evidentiality:

> Modal verbs present a separate problem. In many languages they are a closed subclass. One may wonder whether their evidential extensions, if any, should be treated on a par with lexical expression of evidentiality, or as evidentiality strategies... That a modal verb can express inference does not mean that it is an evidential... Evidentiality is not among its primary meanings. But is this an evidentiality strategy?
>
> The answer to this question depends on the status of modal verbs in the language -whether they are indeed a closed class, and whether they form special grammatical constructions in which they acquire additional meanings related to information source. Generally speaking, they are on the borderline between lexical evidentiality and evidentiality strategy (Aikhenvald, 2004: 150).

Going back to examples (1) and (2), *may* implies epistemic modality showing probability, and, apparently, no intention of evidential meaning. By "apparently", I mean that, in the particular case of (1), *may* appears to be used as a consequence of the first proposition, if we consider only the linguistic environment. Therefore, the modal form is here an indicator of the stance of the author towards the text, rather than source of knowledge. However, the benefits of cholesterol reduction are common knowledge, and with this the author may also seek to support his claim for the case of this type of cholesterol. There are some other cases in the corpus where epistemic meaning alone is beyond doubt, as in (3) and (4), below:

(3) Screening ultrasound *may* depict small, node-negative breast cancers not seen on mammography (CS).
(4) Genotyping these variants *may* help to achieve the benefits of statin-induced therapy more safely and effectively (VS).

The use of the modals in these samples seems to refer back to previous knowledge, although mode of knowledge is not clearly indicated at this point of the complete academic article. In both cases, tentative probability is manifested, and the authors do not blatantly commit themselves to the truth of the statement. In the specific case of (4), the authors express a future uncontrolled event by means of this verbal form. They cannot guarantee that

the action will eventually take place. In other words, *may* is used in this context with a sense of predictability, but it also manifests the authors' expectation towards a desired event.

The following example shows both an evidential and an epistemic reading of the clause in which *may* is embedded:

(5) Lead, mercury, and arsenic have been detected in a substantial proportion of Indian-manufactured traditional Ayurvedic medicines. Metals *may* be present due to the practice of rasa shastra (combining herbs with metals, minerals, and gems) (LM).

In this example, the use of may clearly manifests the authors' deductive inferential reasoning from which they obtain their evidence leading to this proposition. As put forward by Cornillie (2009: 58), "inferences have generally been associated with strong speaker commitment, but... such an association does not always hold". The reliability of this proposition depends on how accepted or common this idea is within the medical literature rather than with the degree of commitment or likelihood. As an epistemic marker, may is canonically used to refer to a weaker epistemic commitment (cf. Palmer 1990). It seems that inference should be expressed by means of must rather than may, since the former appears to be more reliable than the latter (Cornillie 2009: 58), as in She must be there. The light is on vs. She may be there. The light is on. Why is may selected here then? This choice may arguably show authorial tentativeness regarding the truth of their proposition, even if they are highly confident about it. However, this selection does not undermine the value of may as an evidential marker, since the authors' lack of commitment does not determine the status of the evidence.

5. Conclusion

In this paper, I have focused on the use of may in medical research abstracts to evaluate possible epistemic and evidential readings. The modal under study is most often used to indicate an epistemic qualification of the clause in which it is embedded, but it is also likely to express evidential meaning. The evaluation of context is relevant in the disambiguation of what types of meanings this modal may present. The linguistic and environmental contexts allow for the interpretation of may as showing epistemic or evidential meanings, or both. I have thus tried to keep both the evidential and epistemic modality as two distinct categories by avoiding an evaluation of the mode of knowing in terms of authorial commitment. This

study represents work in progress, and more research will be done in a near future with a larger corpus of texts.

REFERENCES

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford: Oxford University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan (1999). *Longman Grammar of Spoken and Written English*. New York: Longman.

Carretero, M. (2004). The Role of Evidentiality and Epistemic Modality in Three English Spoken Texts from Legal Proceedings. In J. I. Marín-Arrese (Ed.). *Perspectives on Evidentiality and Modality* (pp. 25-62). Madrid: Editorial Complutense.

Cornillie, B. (2009). Evidentiality and Epistemic Modality. On the Close Relationship between Two Different Categories. *Functions of Language 16:1*. (pp. 44-62).

Chafe, W. L. (1986). Evidentiality in English Conversation and Academic Writing. In W. L. Chafe & J. Nichols (Eds.), *Evidentiality: The Linguistic Coding of Epistemology* (pp. 261-272). Norwood, NJ: Ablex Publishing Co.

De Haan, F. (1999). Evidentiality and Epistemic Modality: Setting Boundaries. *Southwest Journal of Linguistics, 18*: 83–101.

Dendale, P., Tasmowski, L. (2001). Introduction: Evidentiality and Related Notions. *Journal of Pragmatics, 33*: 339–348.

Fox, B. (2001). Evidentiality: Authority, Responsibility, and Entitlement in English Conversation. *Journal of Linguistic Anthropology, 11*(2): 167-192

Hoye, L. (1997). *Adverbs and Modality in English*. Essex: Longman.

Lazard, G. (2001). On the Grammaticalization of Evidentiality. In Dendale, P., Tasmowski, L. (Eds.), Evidentiality (special issue). *Journal of Pragmatics, 33*: 359–367.

Mithun, M. (1986). Evidential diachrony in Northern Iroquoian. In W. Chafe and J. Nichols (eds.), *Evidentiality*: *The Linguistic Coding of Epistemology* (pp. 89-112). Norwood, NJ: Ablex.

Mushin, I. (2001). *Evidentiality and Epistemological Stance: Narrative Retelling*. Amsterdam: Benjamin.

Nuyts, J. (2004). Over de (beperkte) combineerbaarheid van deontische, epistemische en evidentiële uitdrukkingen in het Nederlands. *Antwerp Papers in Linguistics* 108.

Palmer, F. R. (1986). *Mood and Modality*. Cambridge: Cambridge University Press.

Palmer, F.R. (1990) [1979]. Modality and the English Modals. London and New York: Longman.

Van der Auwera, J., Plungian, V.A. (1998). Modality's Semantic Map. *Linguistic Typology 2*, 79–124.

Von Wright, G.H. (1951). An Essay in Modal Logic. Amsterdam: North Holland.

Exploring the Use of Wordsmith Tools for Sociolinguistics Purposes: a Case Study of Cultural Loaded Language Uses in White and Black Rappers' Corpora

PEDRO ÁLVAREZ MOSQUERA

University of Salamanca

Abstract

The study of processes of language crossing between white and black rappers (Álvarez-Mosquera, 2009), it allowed us to achieve two independent corpora corresponding to these two ethnic groups. Although different linguistic features were the subject of study, we also analyzed the presence of some cultural components in this music genre deeply rooted in the Black Oral Tradition. In this study, we are presenting some remarkable results that were obtained by processing this data through Wordsmith Tools in order to track cultural similarities and differences in their language uses in rap compositions. The use of personal pronouns like I, you or we, the presence of violent terminology or the references to one of the most symbolic space for rappers, that is, the hood draw some ethnic lines in this specific linguistic context.

Keywords: Rap music, Wordsmith Tools, ethnicity, authenticity, Black Oral Tradition, violence, community, hood

Resumen

El estudio de procesos de language crossing llevados a cabo por raperos blancos en AAVE (African American Vernacular English) (Álvarez-Mosquera, 2009), nos permitió crear dos corpora lingüísticos correspondientes a ambos grupos étnicos. Aunque en nuestro estudio diferentes rasgos lingüísticos fueron objeto de análisis, prestamos especial atención a la presencia de elementos cuturales presentes en este género musical tan enraizado en la cultura afroamericana. En este trabajo de investigación, presentamos algunos de los resultados más importantes obtenidos al procesar nuestros datos a través de Wordsmith Tools con el objetivo de localizar similitudes y diferencias cuturales en los usos lingüísticos recogidos en sus composiciones musicales. El uso de pronombres personales como I, you o we, la presencia de determinada terminología violenta o las referencias a uno de los espacios más simbólicos para los raperos, es decir, the hood, dibujan delimitaciones étnicas en este contexto lingüístico tan específico.

Palabras clave: Música rap, Wordsmith Tools, etnia, autenticidad, Balck Oral Tradition, violencia, comunidad, hood

1. INTRODUCTION[1]

Music has been a source of study from many different perspectives in the linguistics field (Alim, 2006; Newman, 2005; Best & Keller, 1999; Rampton, 1995; among others) and corpus-based approaches have not been an exception in recent years (e.g. Beal, 2009). While carrying out a sociolinguistics research project on processes of language crossing between white and black rappers (Álvarez-Mosquera, 2009) we created two independent corpora corresponding to the two groups subject of analysis: European Americans and African

[1] This article is funded by the Junta de Castilla and Leon project SA012A09.

Americans. The origins of rap music are strongly rooted in the African American culture. The existence of oral and cultural ethnic patterns in this music genre, the use of features of AAVE (African American Vernacular English) or the strong concept of authenticity, keep rap music linked to the ethnic group in which it was born. However, the success of rap music has gone beyond the ethnic borders, setting the linguistic component at the center of the controversy on many occasions. This fact is not a minor issue due to the strong correlation between language, identity and power. Since "ethnic distinctions are rarely neutral" as Giddens asserts (qtd. in Milroy & Gordon, 2003: 108), the study of individuals located at the borderline between two language varieties, like white rappers, let us explore the complexities of such linguistic interactions in order to identify linguistic patterns or even language barriers. The aim of this study is to shed some light on these particular contexts where identity and language constructions are at issue between both communities.

2. MATERIALS AND METHOD

Before starting to analyze the results of the Wordsmith Tools approach, we need to clarify some previous aspects concerning the nature of our corpora. To begin with, we have to point out that the size of the corpora is limited due to sociolinguistic reasons. In order to maintain different social variables such as origin (all rappers come from New York area[2]), age (in order to avoid generational differences) and gender (all rappers analyzed in this study were masculine) which could alter our results. Concerning the choice of New York, we would like to say that city where rap music was born is large enough to find six rappers with consolidated artistic careers and, at the same time, there is less overlapping among local language varieties associates with European Americans and AAVE (Wolfram & Christian, 1989: 18-19). Also, we considered necessary that the albums analyzed here were published around the same year to restrict the possible variation resulting from trendy language uses. By maintaining this approach we were able to keep the ethnic component as isolated as possible. Once we have satisfied the sociolinguistics needs of our study, we also faced the fact that most rap songs present choruses that could significantly vary our final results. In other words, the presence of a particular word (e.g. jail) a couple of times in a chorus that can be repeated over ten times in only one song might appear very high in the wordlist provided by Wordsmith Tools despite the fact that it could be the only song that contains such word. In

[2] This parameter might be affected by rapper's mobility. For instance, Chris Palko (Cage) was born in the city of Würzburg (Germany) although his American parents moved to New York when he was four years old. Similarly, the rapper Erik Schrody (Everlast) was born in Long Island but he also lived in Los Angeles.

order to avoid this type of situation, we have only taken the choruses into account once. We have also removed any technical word referred to the format of the lyrics such as title, album, chorus, etc. so that only the words uttered by the rappers are actually analyzed here. Finally, by listening to every song and taking into account the information provided by the CDs and the website web Ohhla.com[3] (Original hip hop lyrics archive), we also removed the parts of the songs sung by other artists whose ethnic, age or origins did not correspond to our research parameters.

For this study, we have selected an album (around 12 songs) by every rapper, namely *Beastie Boys* and *Public Enemy* for the 80's, *Everlast* and 2*Pac* in the 90's and finally, *Cage* and *50 Cent* for the beginning of the 21st century. These songs were edited following the provided guidelines. Due to the nature of this data, our corpora were importantly reduced, so that the white corpus contains 14,865 words and we have 17,221 words in the black one. Nevertheless, it is our intention to increase the amount of tokens for further research.

3. CORPUS ANALYSIS

Wordsmith Tools represents an authentic opportunity to explore the resulting corpora from a sociolinguistic approach by looking for patterns in rappers' language uses. The potential and possibilities of analysis with Wordsmith Tools are certainly numerous, however we point out to those that are relevant from the sociolinguistic point of view. Through the Wordlist option we obtain the list of words organized in terms of frequency, that is, the number of times a particular word has been used in every corpora. For both lists, we have added the same Stopword list containing fifty grammatical words (e.g. the, a, to, of, etc.). The fact that their nature and high index of frequency would have placed them on the first positions of our lists, it would make our analysis more difficult and, at the same time, these types of words are also non significant for our study.

[3] We have used this website to obtain the lyrics that were not included in the original CDs. This website has been used by other authors, such as Morgan (2002), in their studies.

White rappers:

Table 1: WST (European American Rappers)

N	Word	Freq.	%	Texts	%	Lemmas	Set
1	I	489	3.3135	3	100		
2	MY	264	1.7889	3	100		
3	YOU	196	1.3281	3	100		
4	I'M	187	1.2671	3	100		
5	ME	151	1.0232	3	100		
6	LIKE	119	0.8063	3	100		
7	GOT	117	0.7928	3	100		
8	HER	101	0.6844	3	100		
9	ALL	84	0.5692	3	100		
10	GET	82	0.5556	3	100		

Black rappers:

Table 2: WST (African American Rappers)

N	Word	Freq.	%	Texts	%	Lemmas	Set
1	I	550	3.2031	3	100		
2	YOU	429	2.4984	3	100		
3	MY	286	1.6656	3	100		
4	ME	252	1.4676	3	100		
5	I'M	236	1.3744	3	100		
6	GOT	152	0.8852	3	100		
7	GET	148	0.8619	3	100		
8	YOUR	138	0.8037	3	100		
9	WE	121	0.7047	3	100		
10	THEY	117	0.6814	3	100		

Due to the characteristics of this presentation, we are only focusing on some remarkable aspects of our study. First of all, the personal pronoun *I* is the most frequent word in both groups (489 times for white rappers (WR) and 550 times for black rappers (BR)). This fact confirms that rap remains as personal message type of genre connecting to the concept of the African *griot* in the *Black Oral tradition*. As Smiterman asserts, a rapper represents "a post-modern African griot, the verbally gifted storyteller and cultural historian in traditional African society" (Smitherman, 2000: 269). Actually, if we consider the lemmas and other references to the first-person singular pronoun such as *my* (WR: 264; BR: 286), *me* (WR: 151; BR: 252), *I'm* (WR: 187; BR: 236), among others, we observe that they are also placed in the high part of our lists, thus reinforcing our statement. However, we observe a slight quantitative difference in favor of Black rappers that draws a thin ethnic line between both groups. In any case, the role of the rapper could be considered as prominent in both

groups. Nevertheless, this seems to be one of the few similarities since ethnic differences become more prominent in the following sections.

Despite the fact that the pronoun *you* appears in the third position for white rappers and in the second one for the African American group, the latter group obtains a much higher frequency. This information leads us to think that the message uttered by the Black rappers may appeal to the receiver substantially more than the *White message* (WR 196; BR: 429), which would again highlight the role of the African *griot*. Another personal pronoun, *we*, together with the references to the *hood* keep drawing ethnic lines at rapping, that is, the sense of community in this case. In Black communities, "[t]he rapper must literally be the homeboy next door…except now a neighbor who's up on stage, rich and famous, via his *entitlement* to speak to, of and for his community" (Costello & Foster, 1990: 115). In our study, *we* is used twice as much by the black rappers than the white group (WR: 61; BR: 121). From this result, we can infer that the concept of community and identifying the message to their people (Rose, 1994: 9-10) is also much stronger in the black group. In relation to this concept of community in rap music, one of the places for *authentification* is the *hood* or *ghetto*. The rapper JT the Bigga Figga express this idea in a interview: "[s]o this is the voice of the ghetto. The rap comes from the voice of the ghetto […] Straight from the streets" (Alim 2006: 1, see also Bennett, 1999: 3). By using *Wordsmith Tools* to locate terms in their contexts, we looked for explicit references to the hood[4]. The results show that white rappers have mentioned the *hood* twice in different ways:

1. "like I'm from the hood" (Cage).

Example that emphasizes the importance of this space in the rap world.

2. "not everyone can relate to the hood" (Everlast).

Everlast would acknowledge that there are limits, although he insists he spends time in this location.

However, in the black rappers corpus we find seven clear references to the *hood*, reinforcing the relevance of this location (see table 3).

[4] *Hood* and *ghetto* were subjects of analysis in this part of the study, but the rappers did not use the latter in their lyrics. We need to specify that we are dealing with explicit references because many rap songs are implicitly contextualized in this symbolic space.

Table 3: The use of the word *hood* by African American rappers.

N	Concordance
1	No problem, it's all good I ain't fresh out the **hood**, I'm still in the hood Black rims
2	it's all good I ain't fresh out the hood, I'm still in the **hood** Black rims, black hemi,
3	That's why we never ever ever see you in the **hood** with it Man e'rybody know
4	I have ya outlined in chalk (I-I Get It) In the **hood** if ya ask about me They'll tell
5	my prayers for me I come creepin through the **hood** wearin teflon Hit the corners
6	cowards tried to murder me From **hood** to the 'burbs, everyone of you niggas
7	cheap tricks from gettin on her... life in tha **hood**... is all good for nobody remember

It is particularly interesting that the relationship between *hood* and *authenticity* is established in the seven cases, not only to portrait that they belong to the hood, but also to transmit messages to their addressees such as, "we never ever ever see you in the hood with it" (50 Cent) where it matters, in the *hood*. Despite the low rate of appearances, it can therefore be established that the nature of the examples shows a greater attachment to this symbolic space in the African American group.

Continuing with our analysis, we wanted to know what was true about the general belief of the violent nature of rap music. Since its very beginning, rap has been attacked from many social sectors that alert us about the dangerous content of its lyrics and it has constituted a source of pejorative images of African American communities as a threat to American society. Its increasing presence in the media "has also fulfilled national fantasies about the violence and danger that purportedly consume the poorest and most economically fragile communities of color" (Rose, 1994: 11). As for our corpora, the word *gun* (20 repetitions), ranked in number 66 in the *white* list, attracts our attention since its frequency is higher than common verbs like *want* (position 68) or *been* (position 71). A closer approach to the topic of violence allows us to confirm this tendency, that is to say, the case of *gun* was not an exception in this ethnic group and the amount of violent terms appears significant in both corpora. In our study, we compare the use of explicit violent terminology to determine whether white and black rappers present similar or different rates. In order to enrich our approach to the study of violence, we took four terms into account (*gun*, *kill*, *shot* and *fight*) and their *lemmas* (e.g. *kill* includes all its variations in our corpora: *kills*, *killed*, *killing*, *killers*). (See table 4).

Table 4: WST (violent words).

White Rappers	Frequency	Position	**Black Rappers**	Frequency	Position
GUN	27 times	43	GUN	17 times	97
KILL	14 times	110	KILL	14 times	127
SHOT	17 times	89	SHOT	15 times	119
FIGHT	7 times	229	FIGHT	4 times	485
POLICE	1 time	2.431	POLICE	7 times	312

The results show that white rappers present a higher use of this type of explicitly violent vocabulary, being only inferior in the category of *police* and presenting the same number of instances for the term *kill*. The word *police*, a non-violent word in principle, has been included due to the negative connotations in the ghetto context, since policemen can be considered as an oppressive force. In our study white rappers use this word only once while black rappers present seven instances. The study of this term in its context confirms again the existence of an ethnic barrier resulting from the *black experience*[5], since white rappers cannot appropriate the local knowledge coming from lived experiences (Morgan, 2001: 194). In the light of these results, we can question the general belief that African American rappers incite more to violence, since, at least in terms of frequency, their white counterparts present higher numbers in this case.

Another language feature that appears to be relevant for the analysis of rap corpora is the use of imperative *don't*[6]. As we have pointed out earlier, rappers, as the new *griots*, appeal significantly to their audiences, a fact that might indicate the existence of a high rate of imperative uses because their stories would be orientated to alert others and transfer relevant knowledge. However, due to the type of analysis we are using here, that is, the study of a word in its context, we will only be able to analyze the negative cases of imperative (e.g. don't do this), lacking the possibility of analyzing non-negative orders (e.g. look at me; be yourself, etc.) due to the numerous variations. According to our results, 31.7% of the uses of *don't* are imperative in the European American group, for 27.3% in the African American one. This data states that almost one out of every three uses of *don't* has an imperative function in both groups. Therefore, we can affirm that white and black rappers present similar levels of imperative uses in their music.

[5] In general terms, the expression *Black Experience* refers to the collection of historical, social and economic conditions that this ethnic group have suffered throughout its history in the American continent.
[6] A future step will consist of including affirmative imperative forms.

Finally, the last feature that is the subject of analysis is a preliminary approach to the study of *repetition*[7]. *Wordsmith Tools* informs us of the ratio of any corpus, data that becomes very relevant in the particular case of rap due to the fact that repetition patterns are part of the language possibilities present in AAVE. We obtain this data by dividing the total number of types by the total number of items. Therefore, higher percentages would represent a greater amount of different words, that is to say, less repetition. For a better understanding of the use of this pattern, we study every rapper separately and we also provide the totals for both ethnic groups. Here the following results were obtained:

Table 5: WST (Ratio)

White Rappers	Type/Item Ratio (%)	Black Rappers	Type/Item Ratio (%)
Beastie Boys	25	Public Enemy	26
Everlast	31	2Pac	23
Cage	33	50Cent	23
TOTAL	**22**		**17**

The partial and total results (with the exception of Beastie Boys) show a greater amount of different words in the European American groups. In other words, African American rappers tend to repeat more. This fact, far form demonstrating a lack of linguistic skills, can reflect the maintenance of one of the most characteristic features of AAVE speakers, that is, repetition patterns as a means of communication and reinforcement of their identity. As Snead states "repetition is an important and telling element in culture, a means by which a sense of continuity, security, and identification are maintained" (qtd. in Rose, 1994: 68-69) and these ideas also apply to rap music: "[African American] rappers call upon the use of repetition at will and use it to perform a variety of functions, such as: to tell cautionary tales, to drive important points/themes home, to elicit laughter, and to display their lyrical skillz" (Alim, 2006: 84). According to our results, only the groups in the 80's present similar ratios, although in the following decades both ethnic groups experiment opposing tendencies. Two possible reasons might explain this phenomenon. On the one hand, the commercial and social implications derived from the fact that rap music was starting to be treated as a product for the mass audience might have conditioned the linguistic production of rappers since Public Enemy and Beastie Boys have a lower rate of other AAVE features (Alvarez Mosquera, 2009) in comparison with their own ethnic colleagues. On the other hand, the fact that the

[7] We would like to emphasize again that we have only included every chorus once. Therefore, a complete study of the repetition patterns will require taking choruses into account in future studies.

repetition ratio increases in the last two decades (in the case of Black rappers) might also be explained by the tendency to increase of AAVE features in their rap songs as rappers from other ethnic groups proliferate and this genre gets totally consolidated. However, a further analysis including the choruses is necessary in order to reinforce or reject these preliminary results.

4. CONCLUSION

Although further research has to be done in order to explore more possibilities of the use of *Wordsmith Tools* for sociolinguistic purposes, our preliminary study has shown that this tool has been successful at reinforcing and clarifying some cultural aspects related to rap music. Similarities were found in the use of the personal pronoun *I* (and other references to the first person) that is related with the nature of this music genre connected to the *Black Oral Tradition* and the role of the African *griot* as well as the use of the imperative *don't*. Despite this fact, quantitative differences were observed in favor of the African American group in most of the cases. Nowadays, it is easy to observe how hip hop characteristics are reproduced by young people all over the planet, embracing fashion, language, dancing, etc. However, although some aspects might be easy to *appropriate*, some others are highly disputed. The appealing nature of rap, that is, the fact that rap songs are intended to transmit a message to the community, constitutes the first difference between both ethnic groups since the pronoun *you* is doubled in the African American group. Actually, the sense of community is another source of divergence between white and black rappers. The use of the pronoun *we* and the references to the *hood* set white rappers further from this site of *authentification*, mainly because of the consequences of what is called the *black experience*. Defined by Rose as "the experience of domination and the hidden transcripts produced in relation to these experiences of domination are culturally coded and culturally specific" (Rose, 1994: 123), it also influences the last part of our study, in particular the use of violent terms in rap songs and the use of repetition. *Gun*, *kill*, *shot*, *fight* and *police* are widely used by both ethnic groups, although white rappers demonstrate more instances in total. This fact contradicts the general belief that African American rappers promote violence in particular due to their ethnic origins. However, the term *police* draws another ethnic line between both groups as a figment of the mentioned *black experience*. In the case of repetitive patterns, the results provided by the ratio point out another source of divergence between both groups when doing rap, although contextual reasons may also explain their opposing tendencies. There is no doubt

that further research will shed more light on the language uses of white and black rappers and *Wordsmith Tools* will allow us to explore more similarities and differences in this music genre, not only restricted to the sociolinguistic field.

REFERENCES

Alim, S. (2006). *Roc the Mic* Right: *The Language of Hip Hop Culture*. London: Routledge.

Álvarez Mosquera, P. (2009). El uso de AAVE por raperos blancos: ¿un caso real de language crossing? MA dissertation. University of Salamanca. Unpublished.

Beal, J. (2009). 'You're Not from New York City, You're from Rotherham': Dialect and Identity in British Indie Music. *Journal of English Linguistics*, 37 (3). (pp. 223-240).

Bennett, A. (1999). Rappin' on the Tyne: white hip hop culture in Northeast England - an ethnographic study. *The Sociological Review*, 47(1). (pp. 1-24).

Best, S. & D. Keller (1999). Rap, black rage and racial difference. *Enculturation*. Spring 2.2.

Costello, M. & D. Foster (1990). *Signifying Rappers*: *Rap and Race in the Urban Present*. New York: Ecco.

Milroy, L. & M. Gordon (2003). *Sociolinguistics*: *Method and Interpretation*. Oxford: Blackwell.

Morgan, M. (2001). 'Nuthin' but a G thang:' grammar and language ideology in hip hop identity. In Sonja L. Lanehart (ed.). *Varieties of English around the World: Sociocultural and Historical Context of African American English*. Amsterdam: Benjamins. Pp.185-207.

Newman, M. (2005). Rap as literacy: a genre analysis of hip-hop ciphers. *Text,* 25 (3). (pp. 399-436).

Rampton, B. (1995). *Crossing: Language and Ethnicity Among Adolescents*. New York: Longman.

Rose, T. (1994). *Black Noise*: *Rap Music and Black Culture in Contemporary America*. New England: Wesleyan U P.

Smitherman, G. (2000). *Talkin That Talk: Language, Culture and Education in African America.* New York: Routledge.

Wolfram, W. & D. Christian (1989). *Dialects and Education: Issues and Answers*. Englewood Cliffs: Prentice Hall.

Reconsideraciones sobre el diseño inicial de un corpus digital de lengua de signos española

PATRICIA ÁLVAREZ SÁNCHEZ

Universidade de Vigo

Resumen

Durante los últimos quince años el grupo de investigación sobre lenguas signadas y lingüística de la Universidad de Vigo ha trabajado en el desarrollo de un corpus digital de lengua de signos española (LSE).

Tras una primera fase de registro de datos y el etiquetado parcial de éstos mediante glosas, hemos ampliado nuestros objetivos y nos hemos replanteado varios aspectos en la creación de este corpus, principalmente en lo que respecta a la naturaleza de las muestras de lengua y al sistema de anotación de las mismas. Centraremos nuestra comunicación en estas reconsideraciones que han surgido fundamentalmente gracias a los avances teóricos en el campo de la lingüística de las lenguas de signos y a las posibilidades que brindan las nuevas tecnologías para el desarrollo de los corpus.

Además, repasaremos las características específicas que definen los corpus de lenguas signadas, sus posibles explotaciones en los campos de la lexicografía, fonología y psicolingüística entre otros, y las cuestiones técnicas que derivan de la recogida de datos en formato visual.

Palabras clave: lengua de signos, lengua de signos española (LSE), corpus, anotación, lingüística, nuevas tecnologías, signoblog, vlog, teclado

Abstract

During the last fifteen years, the research group on sign languages and linguistics at the University of Vigo has worked in the development of a digital corpus of Spanish Sign Language (LSE).

After a first stage of data recording and its partial annotation using glosses, we have widened our aims and reconsider several aspects in the development of our corpus, mainly with regard to the nature of the language samples and the annotation system. We will focus this paper in the reconsiderations that have come up thanks to the many theoretical advances in the field of sign language phonology and the possibilities offered by the new technologies for the development of a corpus.

Moreover, we will revise the specific features that define sign language corpora, their potential exploitation within the fields of lexicography, phonology, and psycholinguistics among others, and every technical question deriving from the gathering of digital data in a visual format.

Keywords: sign languages, Spanish Sign Language (LSE), corpus, annotation, linguistics, new technologies, signoblog, vlog, keyboard

1. INTRODUCCIÓN

El grupo de investigación sobre lenguas signadas y lingüística de la Universidad de Vigo lleva más de quince años trabajando en el desarrollo de un corpus digital de lengua de signos española (LSE) a partir de las muestras de lengua recogidas entre personas signantes de toda España.

Los objetivos que impulsaron al primer grupo de investigadores de lenguas de signos de nuestra universidad a emprender esta ardua tarea fueron principalmente comenzar la descripción de la LSE, investigar cuáles eran las unidades mínimas lingüísticas de esta lengua

y conocer a fondo su gramática y, por último, desarrollar una serie de herramientas útiles para la recogida, transcripción y etiquetado de muestras de lengua en formato visual.

Ana Fernández Soneira realiza la primera transcripción parcial de nuestro corpus con el fin de utilizar los datos etiquetados como base para la tesis que publica en 2004 sobre los cuantificadores en lengua de signos española (Fernández Soneira, 2004). En 2008 ve la luz un manual didáctico de LSE titulado "*Defiéndete en LSE*" (Báez Montero, Cabeza Pereiro, Fernández Soneira, y Eijo Santos, 2008) que cuenta con numerosas fotografías de signos en LSE glosados en español.

Nuestro corpus es un conjunto de datos lingüísticos reales almacenados en este caso en un medio electrónico y reunidos con el fin último de estudiar la lengua de signos española. Los datos que se han recogido son muestras de lengua visuales en formato de vídeo que están siendo transcritas y etiquetadas para facilitar búsquedas complejas dentro del corpus. Estas muestras se han guardado en cintas de vídeo VHS, grabadores externos de DVD, ordenadores personales y discos duros externos. El formato de vídeo resulta imprescindible en nuestra tarea si tenemos en cuenta que las lenguas de signos no cuentan hoy en día con un sistema fiel de transcripción suficientemente extendido que permita codificar este discurso de manera efectiva.

Las muestras de lengua recogidas hasta el momento provienen de más de ochenta informantes usuarios de la LSE. Algunos de los datos recogidos acerca de estos signantes se corresponden con los metadatos propuestos por la iniciativa ISLE (IMDI – Isle Meta Data Initiative) desarrollada en el Instituto Max Planck. Entre otros detalles significativos hemos anotado los datos personales del informante, su origen y desarrollo social, la educación recibida y sus habilidades lingüísticas. Además, consideramos que es fundamental tener en cuenta parámetros como el grado de sordera, las relaciones familiares o el tipo de centro educativo en el que los informantes han sido escolarizados para poder refinar los criterios de búsqueda en el corpus y, en un futuro, poder analizar la relación entre estas variables y sus muestras de LSE correspondientes.

Hemos creado una base de datos para recoger toda esta información relevante acerca de los informantes y poder estudiar los índices de variación lingüística presentes también en las lenguas de signos. Esperamos que estos metadatos puedan cruzarse en un futuro con las propias muestras de LSE con el fin de permitir búsquedas complejas en el corpus. De esta manera, el usuario del corpus podrá localizar los signos correspondientes a un concepto teniendo en cuenta además otros criterios como la edad, procedencia o educación recibida por el signante.

Hasta ahora, los tipos de muestras de lengua recogidos se habían limitado a tres clases de textos signados (Báez Montero y Cabeza Pereiro, 1999):

a) Un monólogo guiado en el que se pedía al informante que describiese elementos de su entorno (su casa, su familia, etc.), una anécdota o historia corta, etc. A pesar de que el discurso recogido siguiendo este método resultó ser poco natural, obtuvimos muestras de carácter descriptivo y narrativo con una importante carga gramatical y semántica.
b) Una entrevista semidirigida en la que se pedía a una persona signante que conversase con el informante acerca de temas diversos que variaban en función de la edad y aficiones del entrevistado. El objetivo de este tipo de tareas para la obtención de muestras era conseguir signos naturales y espontáneos que no requiriesen de una organización o reflexión previa por parte del signante.
c) El discurso público recogido durante la interpretación en LSE de conferencias, mesas redondas o cursos académicos y el discurso público de personas sordas.

Por supuesto, se ha intentado cuidar el aspecto ético de este trabajo mediante la revisión y aceptación por parte de los informantes de unas cláusulas de cesión de imagen con fines académicos y de investigación que también contemplaban la posible publicación *online* de los datos.

Las primeras grabaciones para nuestro corpus fueron recogidas en cintas de vídeo VHS. La calidad de estas muestras es ciertamente mejorable y eso nos ha impulsado a seguir adelante utilizando nuevos tecnologías de grabación, desarrollando nuevos sistemas de etiquetado y ampliando los tipos de muestras de habla a nuevos contenidos, registros y variedades lingüísticas.

La falta de tradición en la creación de corpus de lenguas de signos ha sido la causa que nos ha obligado a iniciar, pausar y retomar el proceso de desarrollo de nuestro proyecto ante obstáculos no previstos, numerosos avances teóricos y mejoras tecnológicas (Báez Montero y Cabeza Pereiro, 1995).

En esta comunicación nos centraremos, por una parte, en las reconsideraciones de aspectos fundamentales del diseño, creación y explotación de nuestro corpus que han surgido en los últimos años y por otra, en las soluciones prácticas que hemos encontrado gracias a la revisión de otros corpus internacionales y a la integración de otros departamentos y facultades como colaboradores en nuestro proyecto.

2. NUEVOS OBJETIVOS

Teniendo en cuenta la diversidad de muestras de habla que recoge nuestro corpus podemos agrupar los objetivos que perseguimos en las siguientes cinco líneas generales:

2.1. Gramática, fonología y análisis del discurso

Del mismo modo en que las lenguas orales se sirven de sus corpus para conocer con más detalle los distintos niveles de la lengua y las variedades lingüísticas, el corpus de lengua de signos española tendrá como principal objetivo la descripción detallada, todavía pendiente, de esta lengua en todos los niveles lingüísticos.

Dentro de este punto nos interesa resaltar las recientes investigaciones sobre fonología de las lenguas de signos. Esta disciplina surge a finales de los años 80 con los primeros artículos dedicados al tema publicados por Robert Johnson y Scott Liddell (Johnson y Liddell, 1989). De la misma manera en que los fonemas del español se pueden dividir en rasgos fonéticos que conforman su articulación, Johnson & Liddell conciben los signos como elementos también compuestos por rasgos articulatorios.

Gracias a las nuevas muestras de lengua que suponen las entradas de los signoblogs de personas sordas, de los que hablaremos en el siguiente apartado, nos proponemos estudiar también el discurso signado de los informantes como actos de comunicación en diferentes contextos de uso.

2.2. Lexicografía

Otro de los objetivos lingüísticos que persigue nuestro corpus es el de desarrollar, a partir de los datos obtenidos, una herramienta lexicográfica que pueda ser utilizada de manera sencilla e intuitiva tanto por personas oyentes como por usuarios sordos. La interfaz de búsqueda permitirá encontrar signos eligiendo parámetros dentro de su propia configuración, buscar signos a partir de su interpretación en español y viceversa.

2.3. Psicolingüística y didáctica de lenguas

El análisis en profundidad de la gramática de la lengua de signos española nos lleva a su vez a la enseñanza de esta gramática.

El uso de los datos de este corpus para la creación de herramientas didácticas atiende a la creciente demanda de métodos de enseñanza de la LSE. En nuestra universidad comprobamos que el interés acerca de esta lengua no decrece desde el primer año que la impartimos formalmente como complemento de formación universitaria. Al contrario, cada

año son más las personas que se acercan a esta disciplina desde todos los ámbitos de la sociedad y por los más diversos motivos.

Para todos ellos, nuestro grupo de investigación editó en 2008 *Defiéndete en LSE (Lengua de signos española)* (Báez Montero, 2008), un método comunicativo de autoaprendizaje que ha sido distribuido a centros de enseñanza de todos los niveles educativos con el fin de dar a conocer esta lengua y a la comunidad de signantes que la utiliza.

Siguiendo la misma línea, consideramos que nuestro corpus permitirá a las personas interesadas conocer o mejorar su LSE gracias a la multitud de muestras de lengua y ejemplos de uso que podrán encontrar en él. La interfaz de búsqueda será capaz de recuperar signos, interpretaciones en lengua escrita e incluso narraciones completas partiendo de parámetros de configuración manual o desde el español.

2.4. Interpretación

Hemos contado con la colaboración de intérpretes de toda España para la recogida de muestras de lengua. Consideramos que sus grabaciones son un recurso valiosos si tenemos en cuenta que otra de las posibles explotaciones de este corpus será analizar el discurso interpretado con el objetivo de conocer mejor el proceso de adquisición de la lengua de signos por hablantes no nativos, qué recursos de expresión utilizan los intérpretes y las características del proceso de interpretación simultáneo de textos científicos. Por supuesto, se ha tenido en cuenta su condición de hablantes no nativos a la hora de etiquetar el corpus

2.5. Formación académica de las personas sordas

Alcanzados los objetivos previos, el corpus supondrá sin duda un paso adelante en la educación de las personas sordas, que ya podrán contar con personal docente y administrativo instruido en el uso de la LSE, material descriptivo y didáctico de su propia lengua y una amplia selección de textos académicos y científicos adaptados e interpretados en LSE.

3. NUEVOS DATOS

En relación a los contenidos, hemos querido incluir en nuestro corpus cinco nuevos tipos de muestras de lengua en formato de vídeo:

3.1. Conferencias científicas interpretadas

Hasta el momento, contamos con cinco horas de conferencias sobre lingüística, grabadas, interpretadas en LSE y con subtítulos adaptados para personas sordas.

3.2. Interpretaciones de actas de congresos

Hemos editado en lengua de signos las actas del III Congreso de Lingüística Hispánica celebrado en nuestra ciudad en 2007. Se trata de resúmenes de textos científicos presentados en este congreso que, gracias a esta iniciativa, ahora están al alcance de cualquier persona sorda interesada en la lingüística.

3.3. "Píldoras" de gestión académica

Además de los vídeos mencionados anteriormente, también hemos colaborado en otros proyectos con los Servicios Multimedia de la Universidad de Vigo[1]. En la siguiente imagen podemos ver uno de los vídeos explicativos que se han grabado con el fin de facilitar el acceso a los contenidos de gestión académica de la Universidad de Vigo para todas aquellas personas sordas que lo necesiten. Entre otras gestiones, nuestra intérprete explica cómo realizar la matrícula del curso por internet o el sistema de reconocimiento de créditos académicos.

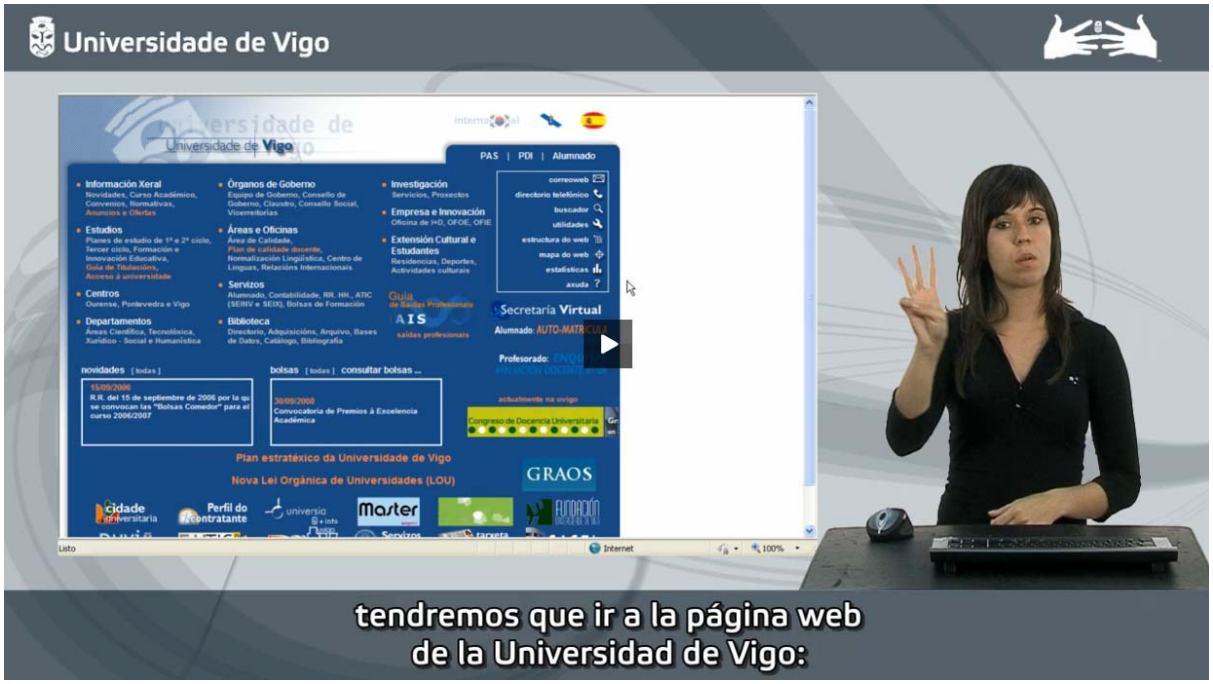


Figura 1: Imagen de uno de los vídeos de gestión académica editados en LSE y realizados en colaboración con UVigo TV.

[1] Página web de UVigo TV: http://tv.uvigo.es/ .

3.4. Programas de televisión signados

Televisión de Galicia emite cada día un espacio informativo en lengua de signos y también programas dirigidos a la comunidad sorda que nosotros grabamos, verificamos y posteriormente seleccionamos para incluirlos en nuestro corpus.

3.5. Signoblogs

Los vídeo blogs (o *vlogs*) cumplen en 2010 su primera década de vida y ya forman parte de la extensa red de contribuciones personales de los usuarios en internet. Se trata de la evolución natural de los blogs con la diferencia de que las entradas de estas webs no son únicamente textuales, sino que también incorporan audio y vídeo a la composición (Hamill, 2004). Nuestro objetivo será incluir el mayor número de aportaciones posibles a nuestro corpus, ya que son muestras de habla naturales, de signantes sordos cuya lengua materna es la LSE y otros que la han aprendido como si se tratase de una lengua extranjera. Existen diferentes tipos de signoblogs, según su temática: aficiones (diarios de viajes, recetarios, vlogs deportivos, cinéfilos, etc.), cultura e identidad sorda (espectáculos y obras realizadas por personas sordas, variedades de la lengua de signos en España, etc.), nuevas tecnologías (últimos *gadgets* o utilidades, software útil, etc.) y actualidad (noticias nacionales e internacionales acerca de la comunidad sorda, etc.).

4. NUEVOS SISTEMAS DE RECOGIDA DE DATOS

Nuestro grupo de investigación comenzó el proceso de recogida de datos para el corpus de lengua de signos española en el año 1995. Por aquel entonces, las grabaciones se realizaban en formato VHS con cámaras que proporcionaban escasa calidad audiovisual. En esta nueva etapa hemos contado con el apoyo del Servicio Multimedia y de TV de la Universidad de Vigo. Con ellos hemos grabado una serie de vídeos cuya duración varía entre los 2 y los 80 minutos, en los que alguno de nuestros colaboradores signantes interpreta textos científicos (clases magistrales, comunicaciones, conferencias, etc.). La calidad de la grabación, tanto en términos audiovisuales como de diseño, es muy superior en este caso.

La figura del signante se ha capturado en formato estándar DVPAL sobre croma blanco. A su lado, se ha incorporado una presentación PowerPoint capturada a resolución 800x600 pixels, con códec CamStudioCodec. El resultado son vídeos en formato FLV/H264 a 1.5Mbps.

Esta nueva serie de vídeos producidos en colaboración con UvigoTV, junto a los antiguos capturados con cámara de vídeo, ya están siendo almacenados en discos duros

internos y externos y en grabadoras de DVD a la espera de que se inicie el proceso de anotación.

5. NUEVAS HERRAMIENTAS PARA LA ANOTACIÓN DEL CORPUS

William C. Stokoe ya demostró en su obra *Sign Language Structure* (1960) que los signos del léxico de la lengua de signos norteamericana (ASL) no son integrales sino que están compuestos por otras unidades más pequeñas y sin significado que, reorganizadas a su vez, producen un léxico más amplio (Sandler, 2003).

El desarrollo de estas investigaciones ha dado lugar a una nueva línea de estudio sobre la fonología de las lenguas de signos, es decir, sobre el estudio de la organización y estructuración de las unidades mínimas contrastivas de la lengua de signos.

Debido al carácter visual-gestual de estas lenguas podemos clasificar las características de los signos en cinco parámetros diferentes, los cuales son responsables de las diferencias en el significado de los signos.

Ceil Lucas (2000) distingue estos cinco parámetros que constituyen la articulación de los signos para la lengua de signos americana (ASL): configuración de la mano, movimiento, localización, orientación y rasgos no manuales. La inclusión de este último parámetro como una de las categorías básicas de la fonología de la lengua de signos ha sido discutida por numerosos autores.

Para la etiquetación de nuestro corpus hemos tenido en cuenta estas cinco categorías: los signos se realizan en un lugar al que llamamos *lugar de articulación*; los signos adoptan, en el lugar de articulación o en contacto con él, una *configuración manual* (palma de la mano extendida, palma cerrada en puño, dedos extendidos, flexionados o cerrados, que la articulación se indique en los cinco dedos o en dedos específicos, etc.); los signos, al ejecutarse, adoptan una determinada *orientación* de la mano (hacia arriba, hacia abajo, hacia el frente, hacia uno mismo, etc.), los signos realizan normalmente un *movimiento* a partir de su ubicación inicial y por último, estos signos se acompañan de *rasgos no manuales* que pueden indicar aspectos como la intensidad o la intención del signante (movimiento de cejas, posición de los labios, etc.).

Es decir, para realizar un signo, la mano activa o dominante (la derecha en los diestros y la izquierda en el caso de las personas zurdas) se dirige a un lugar, adopta en él o en contacto con él una determinada configuración y orientación y ejecuta un movimiento a partir de esa ubicación inicial. Además de estos parámetros, nuestro corpus también recogerá los

rasgos no manuales que acompañan al signo matizando su significado, como la expresión facial, apertura labial, posición de la lengua, movimiento de las cejas, etc.

En el siguiente gráfico se muestra una de las herramientas que estamos desarrollando en colaboración con el grupo de Reconocimiento de Imagen de la Facultad de Ingeniería Industrial de la Universidad de Vigo. Se trata de un teclado con piezas coloreadas y símbolos impresos que se corresponden con las distintas etiquetas fonológicas con las que queremos anotar nuestro corpus. Es decir, este teclado contará con cinco tipos de teclas coloreadas por categorías: articulación, orientación, movimiento, localización, rasgos no manuales, etc. En cada una de ellas se puede ver un símbolo correspondiente al sistema de escritura de lenguas signadas *Signwriting*[2]. El resultado será compatible con nuestro programa de anotación de corpus, que recibirá estas etiquetas en sistema estándar de codificación UNICODE.

Consideramos que la mejor etiquetación posible será la realizada por un signante nativo de la lengua del corpus. Por ello, la creación de este teclado facilitará el trabajo de personas sordas en este proyecto en concreto y en futuras colaboraciones con nuestro grupo de investigación.

Figura 2: Imagen del teclado en construcción para facilitar la anotación del corpus de LSE.

Ésta no es la primera vez que se adapta un teclado para personas con necesidades específicas. Bajo el mismo criterio de distinción por colores y funciones, el teclado para disléxicos se comercializa en todo el mundo desde hace más de 10 años.

6. CONCLUSIÓN

El proceso de creación de un corpus de lengua de signos no es tarea fácil. Nuestro grupo de investigación ha encontrado muchos obstáculos en los últimos años, la mayoría de los cuales han sido salvados gracias a las nuevas tecnologías. Se han producido importantes mejoras en las posibilidades de captura y edición de vídeos. Actualmente nuestras grabaciones tienen

[2] Signwriting - http://www.signwriting.org/ []

calidad de alta definición y son archivadas en dispositivos de almacenamiento externo preparados para recibir, en una etapa posterior, las anotaciones oportunas.

Estos adelantos técnicos nos han permitido ampliar nuestro corpus a nuevos tipos de muestras de lengua. Entre otros, destacamos la suma de ediciones de textos orales científicos en lengua de signos española y las entradas de signoblogs realizados por signantes nativos de LSE y aprendices de esta lengua de toda España y publicados en línea.

La incorporación de estos nuevos datos, a su vez, nos ha impulsado a perseguir nuevos objetivos, como la descripción gramatical, fonológica y del discurso de la LSE. Además, pretendemos estudiar en un futuro el proceso de adquisición de la LSE por hablantes nativos y no nativos y caracterizar el discurso signado por intérpretes en contextos informativos y científicos.

Tanto la edición y adaptación del discurso científico para personas sordas como el propio uso del corpus en sus múltiples posibilidades de búsqueda y recuperación de datos supondrá un importantísimo avance en la educación de las personas sordas. Mantendremos este objetivo siempre en mente cuando iniciemos el proceso de etiquetación de nuestro corpus.

Teniendo en cuenta las futuras necesidades de los usuarios potenciales de nuestro corpus y gracias a los avances teóricos en materia de fonología de las lenguas de signos, hemos desarrollado un sistema de transcripción y etiquetación que permitirá realizar búsquedas detalladas por transcripción en español del signo, por rasgos fonológicos del signo, etc. Uno de los proyectos que actualmente llevamos a cabo pretende crear un teclado con distinción de colores por funciones que ayude a nuestros colaboradores sordos a ser más eficaces en el tratamiento de los datos del corpus. Es decir, esta herramienta contará con teclas de distintos colores en función del parámetro de la lengua de signos con el que se relacionen: teclas para etiquetar la orientación de la mano, su localización, su movimiento y su configuración, así como los rasgos no manuales más destacables en el discurso analizado.

La línea de investigación que seguiremos a partir de ahora nos llevará a elegir un programa de transcripción válido para captar en la medida de lo posible las complejidades lingüísticas que presentan los datos de nuestro corpus, compatible con el formato multimedia que hemos utilizado en las grabaciones y que posibilite la anotación colaborativa y a distancia de los datos.

REFERENCIAS BIBLIOGRÁFICAS

Báez Montero, I. C. y Cabeza Pereiro, M. C. (1995). Diseño de un corpus de lengua de señas española. Comunicación presentada en: *Actas del XXV Simposium de la Sociedad Española de Lingüística*. Zaragoza.

Báez Montero, I. C. y Cabeza Pereiro, M. C. (1999). Elaboración del corpus de lengua de signos española de la Universidad de Vigo. Comunicación presentada en *Taller de Lingüística y Psicolingüística de las lenguas de signos.* A Coruña.

Báez Montero, I. C.; Cabeza Pereiro, M. C.; Fernández Soneira, A.; Eijo Santos, F. (2008). *Defiéndete en LSE (Lengua de signos española)*. Madrid: Anaya.

Fernández Soneira, A. (2004). *La cuantificación en la lengua de signos española.* Tesis doctoral. Vigo: Universidade de Vigo.

Hamill, A. (2004). *Empowerment in the Deaf community: Analyzing the posts of internet weblogs. A thesis.* Bowling Green State University.

Johnson, R. y Liddell S. (1989). American Sign Language: The phonological base. *Sign Language Studies*, 64. (pp. 195-278).

Lucas, C. y Valli, C. (2000). *Linguistics of American Sign Language: an introduction.* Washington: Gallaudet University Press.

Sandler, W. (2003). Sign language phonology. En Frawley, W. (Ed.) *The Oxford International Encyclopedia of Linguistics.* Newark: University of Delaware.

Stokoe, William C. (1960). Sign language structure: An outline of the visual communication systems of the American deaf. En: *Studies in linguistics, Occasional papers*, 8. Silver Spring: Md: Linstok Press.

Synonymous prepositional phrases in a corpus-based cognitive analyses of radial categories: the case of *in the island* vs. *on the island*

ANNA BĄCZKOWSKA

Kazimierz Wielki University

Abstract

Cognitive linguistics and corpus linguistics provide us with highly efficient theoretical and empirical tools applicable to natural language analysis. An investigation wherein both these branches converge in order to uncover semantic-conceptual meanings of two prepositional phrases– in an/the island and on an/the island – is the purpose of the present paper. In accordance with cognitive linguistics, English prepositions are believed to be highly polysemous lexical items with their senses arranged in a network of meanings. There are several theories built around the network structure of categorization, such as family resemblance, the prototype theory and the radial sets model. It is the last model that will be invoked in our analysis. Some didactic applications of cognitive linguistics and language corpora will be proposed towards the end of the paper.

Keywords: prepositions, cognitive linguistics, corpus-based language teaching, radial sets

Resumen

La lingüística cognitiva y la lingüística de corpus nos proporcionan unas herramientas teóricas y empíricas muy útiles para el análisis del lenguaje natural. La presente disertación tiene como objetivo la investigación, en la que convergen estas dos ramas de la ciencia para descubrir los significados semánticos-conceptuales de dos frases preposicionales: in an/the island y on an/the island. De acuerdo con la lingüística cognitiva las preposiciones del idioma inglés se consideran como unos elementos léxicos altamente polisémicos cuyos significados están estructurados en forma de una red. Existen varias teorías en torno a la estructura de la red de categorización, como semejanzas de familia, la teoría del prototipos y el modelo de los categorías radiales. Esta estructura es la última que hemos planteado en nuestro análisis. Además, en la parte final del estudio se han enseñado algunas aplicaciones didácticas de la lingüística cognitiva y del corpus de idioma.

Palabras clave: preposiciones, lingüística cognitiva, enseñanza de idiomas basada en el corpus, redes radiales

1. INTRODUCTION

The analysis of the conceptual meaning of two English prepositions mentioned in the title of this paper relies on the apparatus offered by cognitive linguistic on the one hand and real data retrieved from a corpus of English on the other. The purpose of our investigation is concerned with polysemy and radial sets model of language categorization. Given corpus attestation of relevant data and allowing for prior research into the conceptual meaning of English prepositions it will be shown how two lexical items that are treated as synonymous may undergo a transformation along radial chains and thus display increasing polysemy.

2. COGNITIVE LINGUISTICS ON PREPOSITIONS

Despite the attention the preposition has received thus far, the definition of what the preposition is (in contrast to the particle or the adverb) has often been questioned, and the taxonomy of its types has not been limited to a single proposal. A classic definition has been offered by Quirk, Greenbaum, Leech and Svartvik (1985: 188), who treat the preposition as an uninflected closed class item whose role is specified as 1. linking two elements within a sentence, and 2. dictating the type of relation held between them. Traditionally, and in particular in generative grammar, the preposition was treated as a semantically vacuous word which gained meaning only when attached to another word with which it created a phrase. The role of the preposition in a sentence and a phrase was thus downgraded and pushed to the margin of semantic speculation.

Cognitive linguistics represents a radically different standpoint. The preposition is an example of a relation (contrary to things, i.e. nouns, and processes, i.e. verbs) that is atemporal. As it co-occurs with nominals to make a prepositional phrase, wherein the preposition functions as the head of the phrase, it enters into an atemporal relation with the noun, or, to be more precise, with the object conceptualized by an addressee triggered by the noun. The conceptualization of the noun is not imperious to changes. It depends on the preposition with which it collocates. In accordance with cognitive linguistics, different fragments of the object are envisaged whenever the preposition is substituted by another one. To give an example, in the phrase *on the table* the visualization of the table focuses on its surface, while, in the phrase *under the table* the lower part is highlighted (i.e. profiled). The meaning thus construed allows for a dynamic interpretation of the semantic-conceptual meaning of the noun on the one hand, and acknowledges valence relations wherein the preposition plays a decisive role (of profile determinant) on the other.

The entity profiled in a scene is the figure, or the trajector, while the background is the landmark. The phrase *sugar in a jar* permits profiling the interior part of the jar. The jar, being a larger and less mobile element than sugar constitutes the background facet of thus portrayed picture (i.e. it is its landmark), whereas sugar is assigned the role of the trajector. In a sentence, it is the first noun of a sentence that is the trajector, as it naturally catches the addressee's attention to a greater extent than the second noun. In *John washes his car in the garage*, the trajectory is *John washes his car*, while the landmark comprises *his car in the garage* (Hawkins, 1984: 321; discussed in more detail in Bączkowska, 2010).

Also connected with attention is another issue stressed by cognitive linguists. Standing in sharp contrast to what is known as objective reality, it is claimed (Langacker, 1987) that an object observed and its mental representation are two distinct issues, as the former is rarely, if at all, mapped in an objective way. In principle, our conceptualizations are only *our* images of the world, and thus they are highly subjective. To give an example, our sense of the passing of time changes drastically depending on the circumstances in which we experience it. Research on subjective time passing assessment is extensive and it shows that the human being misjudges the lapse of time as either longer or shorter depending on a number of factors, physical, environmental and physiological. They include the size of the place in which is the observer (e.g. the size of the room), the size of the observer, his age, mood, level of hormones, body temperature, medicine or drugs taken, etc. These data support theses voiced by cognitive linguists about the subjective nature of mental representations.

Langacker (1987) also discusses the problem of what he calls 'virtual boundaries' to illustrate the divergence of what one conceives and perceives while speaking about the meaning of a *ditch*. The conceptualization of a ditch requires that a cognizer create some fictive pattern (a term coined by Talmy (2000)) that reflects the object observed, wherein a dent in the ground receives an axiologically negative charge, being assessed as a place wherein something is missing (a piece of ground) due to the expectation of the level of the ground extending horizontally along a straight line. Any departures from the expected shape are conceived of as unnatural. An observer recognizes the missing part and 'repairs' it by filling it with the missing material through mental operations (i.e. virtually). The process of filling in is encouraged by the fact that the missing part is sanctioned as a hole on the grounds of its ontological dependence on its host, therefore making the relation between them of parasitic nature. The process finishes the moment a cognizer reaches the imaginary line (virtual boundary) that closes, so to speak, the hole, and the line that links, like a bridge over a river, two points wherein the dent begins. As the closing of the area occupied by the hole is virtual, the dependence of the hole on its host seems to be epistemic. The argument of virtual boundaries is consonant with Gestalt rules of perception (this particular one to the rule of closure) which cognitive linguistics often calls for.

From what we have discussed thus far it transpires that meaning is not only conceptualization (as it resists in the way one envisages the object observed), as Langacker (1987) maintains, but also that it varies relative to the attention an object is given. In such configurations, the type of relation held between the two (conceptually autonomous) nouns is profiled by the preposition (conceptually incomplete, i.e. dependent). Taken together,

cognitive linguistics promotes the thesis that meaning equals conceptualization and stresses psychological factors that contribute to meaning, such as attention and the cognizer's state of mind.

3. CATEGORIZATION IN COGNITIVE LINGUISTICS

Prepositions are good examples of polysemy. Their multiplicity of meanings evoked by distinct contexts has been proved by the seminal work of Lakoff (1987). His research uncovers a network of senses tied together through links based on metaphors, metonymies, (image schema) transformation, etc. (cf. Lewandowska-Tomaszczyk, 2007). Such a network creates a radial structure, wherein each node may initiate another radial category becoming its prototype. Following Lakoff (1987), the radial categories approach, sometimes called an extended version or the second generation of the Prototype Theory, was later pursued by Janda (1990), Goldberg (1992) and others.

In the first generation of the Prototype Theory (cf. Rosch), members of a category are graded, with entities with the greatest number of core features in the centre and their attenuation towards the periphery of a network (thus a *robin* is a prototype and an *ostrich* a peripheral example of the category BIRD in American English). The resultant asymmetries emerging between (proto)typical and borderline cases demonstrate membership gradience. What distinguishes radial categorization from prototypical categorization is the fact that in the later there is only one central prototype. On the other hand, core features cannot be recognized between entities in the family resemblance approach (cf. Geeraets, 1995; Cuyckens, 1990) , although some features are shared by some (but not all) members of a family (family 1 contains features: a, b, c; 2: c, d, e; 3: e, f, g). Both the family resemblance model and the radial sets model may be mapped onto the Prototype model. In the same vein, all these models can be captured by a schematic network[1] which, as radial sets, shows how meanings are connected in a network by means of relational links and, additionally, explicates the relations by relying on the processes of generalization (schematicity) and extension to account for different levels of abstractedness (e.g. thing → mammal → rodent → squirrel).

In our analysis we shall adhere to the radial categories approach mapped onto a schematic network. What will be presented in the empirical part cannot prove that the two

[1] The schematic networks model was formulated by Langacker (1987) and later developed by Rudzka-Ostyn (1989), Taylor (1992) and others (discussed in Lewandowska-Tomaszczyk, 2007; Tuggy, 2007).

prepositions in question make a radial network, as there are only four phrases examined. Thus, we need to bear in mind that other investigations support the radial categories claim and the modest analysis presented in this short paper contributes positively to what has already been stated by providing further evidence.

4. ANALYSIS

There is no denying that while some prepositional phrases are built around distinct prepositions, their semantic meaning is largely preserved, leaving sense variations to subtle, often minute changes. The following synonymous pairs whose analysis was based on corpus data are a case in point: *on a bus* vs. *in a bus* or *nibble on* vs. *nibble at* or *sip on* vs. *sip at* vs. *sip* (for details see Bączkowska, 2010). The *bus* patterns together with *in* or *on*, which are the two prepositions we shall examine in the remainder of our investigation. Let us then briefly recall the conclusions concerning the difference in meaning between *on a bus* and *in a bus.* Whenever the bus is used as a means of transport, whose main function is to relocate the objects inside it, the vehicle is envisaged as a supporting platform-like entity for these objects. As support is the notion ascribed traditionally to the meaning of *on*, such a construal would be elaborated by *on a/the bus*. On the other hand, when the platform-like element is de-highlighted and the boundaries of the vehicle are put to the fore, then a container-like entity is in focus. As a container schema is typically associated with the meaning of *in*, in this case it would be coded by *in a/the bus*. Thus, one would preferably use two distinct prepositions in the contexts below:

(1) I heard some gossip about John in a bus.

(2) I was on a bus on my way to school when we were hit by a truck. (BNC corpus data)

The first sentence recognizes the trajectory hovering in the air, so to speak, within the boundaries of the container-bus. The highlighted part of the bus in the second sentence is the platform along with its traditional function of carrying objects, and thus the traditional function of the bus used for transportation.

Despite these subtle differences, the choice of the preposition in a number of contexts remains unstable, as can be seen in the following example:

(3) I left my umbrella on?/in? a bus.

In what follows, we shall be preoccupied with phrases that display synonymy marshalled in a radial fashion with the proviso that the initial synonymy may evolve following a chain of nodes tied together by transformation links which weakens shared features, or even leads to polysemous antonymy.

The synonymous pair we shall tackle is one already discussed by Fillmore (1975: 17) and Dirven (1993), namely *on the/an island* vs. *in the/an island*. In line with the specifications of *on* and *in* mentioned above, flat surfaces and spatial extension are conceptualizations likely to be invoked by *on* whilst inclusion is likely to be called up by *in*. Corpus data seem to support this claim and they provide us with further details.

The first group of examples, wherein *island* collocates with *in*, is used of a collection of entities (people, inhabitants, etc.). Thus, in the concordances below[2] (appendix, Table 1.) we can find (people of) Spain oppressing the people of Cuba, a poet staying physically in an island, inhabitants of an island being encouraged to pay regular subscription, two parties co-existing in the island, etc.

The second group of examples stresses the container schema by making reference to its interior which is either unavailable or hardly penetrable. The contexts in concordances 5 and 6 illustrate the case. The preposition *in* is thus used to speak about archaeological finds that are unearthed, and to denote the unspecified location of a volcano.

There are thus two contexts used for the container schema. One stresses its visually accessible ingredients of a dispersed and collective nature, and the other emphasizes the inaccessibility of its interior. The two cases are presented graphically below.

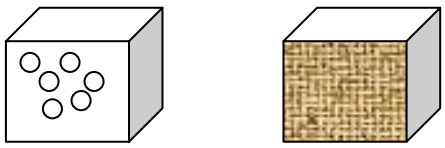


Figure 1: Two subschemas of *in an/the island*.

We shall now move on to the other preposition in question. *On the island* triggers associations with space with little or no obstacles (adhering thus to the classical and cognitive definition of *on* denoting a flat surface and open space), wherein objects are clearly visible

[2] All concordances have been written by the author of this paper. They are based on contexts originally presented in BNC, yet they have been simplified to cater for the needs of language learners.

and easily noticeable (cf. appendix, Table 2.). Alternatively, objects are sufficiently big for an observer to be noticed and singled out from an environmental context, even if it is not completely bare and perfectly flat. Thus relatively big structures, such as buildings, typically pattern with *on+island*. Open space is seen in the example below where a trajectory (a person) is not expected to be on an island any more but rather on a peninsula, which urges a step-back perspective to widen the scope of the scene observed so that a juxtaposition of two entities can be exposed. A flat surface, in turn, is profiled in the scene where campers are camped on an island; the island serving as a supporting plain. Similarly, a football pitch invokes associations with a flat piece of ground.

When two participants of a scene are profiled, and other vertical elements of the island are ignored, again *on* is employed, as in the case of two people falling in love while being together *on* an island. Horizontality is profiled in another example wherein an island is almost completely deprived of trees and on a flat-like platform several extremely tall statues stand erect (profiles). It is also seen in the example which describes the habits of Galapagos doves. The birds tend to perform destruction display, i.e. a series of actions that lure attention away from its eggs lying in a ground nest when an intruder approaches (e.g. simulated injury, such as struggling away from the nest with an apparent broken wing, calling loudly, etc.). This anti-predator behaviour triggers a horizontal trajectory along which a ground enemy (avian predators are rarely approached away in this way) approaches a nest, creating an imaginary path schema with the terminus profiled.

In the extracts analysed, the preposition *on* calls for contexts with either conspicuously tall objects standing against a flat-like background (vertical profiling), and then the distance between the conceptualizer and the scene conceptualized being extended, or, alternatively, it evokes the path schema with a flat-like plain supporting the scene conceptualized.

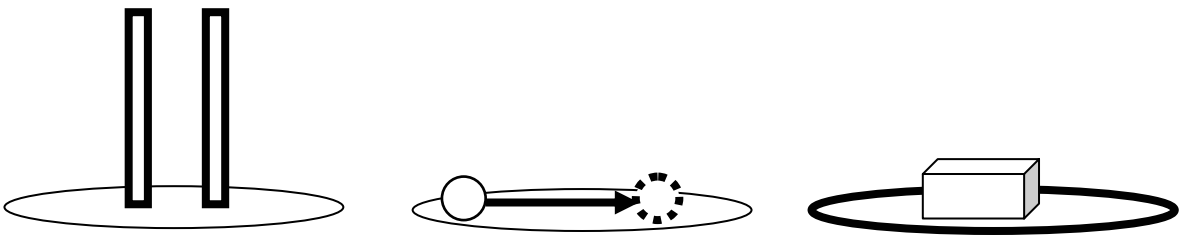


Figure 2: Subschemas of *on an/the island*.

Taken together, the phrases with *in* express the container schema seen either as a closed inaccessible entity or as a host for a collective trajectory. As regards *on*, it is illustrative of the concept of a flat surface on which horizontal trajectory extends, with connotations of support, or, alternatively, it encodes scenes wherein tall vertical entities are profiled and the flat supporting surface remains in its shadow. The analysis has thus demonstrated that within the network structure of each phrase the contexts under investigation are synonymous to the extent that they roughly correspond to the same location in space, identifying the coordinates of the trajectory in a geographically converging place. Put differently, they are synonymous at their general, schematic level. When the properties of these places (*island*, *bus*) are magnified by the lens of a range of actual scenes that present reality filtered through an observer's subjective mind (captured in a language corpus), distinct facets of these concepts (their mental representations) are profiled marking a hiatus at their conceptual structure noticeable with each unfolding subschematic level. As was noticed in the theoretical part of this paper, unlike in the prototype theory of meaning wherein all members of a category share some necessary features, and contrary to the family resemblance approach that underscores only partial overlap of features, in a radial network initial features are shared while subschematic cases may vary to the extent that they may not resemble the central category. In our analysis, the schema of a container triggers associations with its interior, pushing cases with the exterior profiled to the periphery of the radial network. Likewise, the schema of a flat surface seems to have little in common with verticality. These conclusions are valid for contexts encoded by the same lexical unit. They indicate that several conceptual meanings that demonstrate opposite notions may be expressed by one phrase (*in the island*$_1$ vs. *in the island*$_2$; etc., and *on the island*$_1$, vs. *on the island*$_2$, etc.).

Let us now turn to the main thesis of this paper which is concerned with pairs of synonymous phrases that contain *different* prepositions (*in the island* vs. *on the island*). Synonymy contained in these phrases is both intuitively sensed and empirically verifiable. There are languages where both phrases have only one translation equivalent. For example in Polish both phrases are rendered into only one phrase (*na wyspie*, on+island). Let us mention in passing that there are also phrases when the situation is reverse: there are two Polish variants of one English phrase. For example the phrase *in the rain* can have two readings – *na deszczu* (on+rain) or *w deszczu* (in+rain), while *in the sun* may be translated as either *na słońcu* (on+sun) or *w słońcu* (in+sun). Subtle shifts of meaning in Polish versions can be easily observed (Bączkowska 2008). The cases of 4 to 5 translation equivalents (or 5 to 4) are not rare and they seem to support their synonymy.

Returning to the phrases at issue, both of them can make reference to the geographic place and the same objective time, leaving the choice of the preposition to the conceptualiser's decision of which element of the construed scene should be profiled. Thus optics initiated at sub-schematic conceptual level entails choices at the lexical level. The choice, however, seems to be rather fluid. It is not unlikely that a horizontal perspective (typical of *on*) has also an interior perspective (typical of *in*), as in

(4) I saw a snake on the island crawling towards the nest of a Killdear. (horizontal external)

(5)? There are many snakes in the island crawling at night towards nests of Killdears. (horizontal internal)

Along with subjective perspective, the choice of the prepositions might be dependent on the complexity of the trajectory (multiplex, *snake*) as well as the processes involved in the portrayed scene (verb, *be*, *see*).

Similarly, a horizontal external perspective is also possible. Juxtaposition of two entities (e.g. associated with two different places) requires a step-back perspective (external) along with traversing (mentally) an imaginary path linking the initial point with path terminal (path schema).

(6) The tourists were marooned on the island by the storm. (horizontal external)

In the same fashion, an interior perspective may be combined with verticality, as in:

(7) She lived on the island on the 120 floor of a tower block with a beautiful panorama on the whole island. (vertical internal)

The above discussion can be summarized graphically by Figure 3.

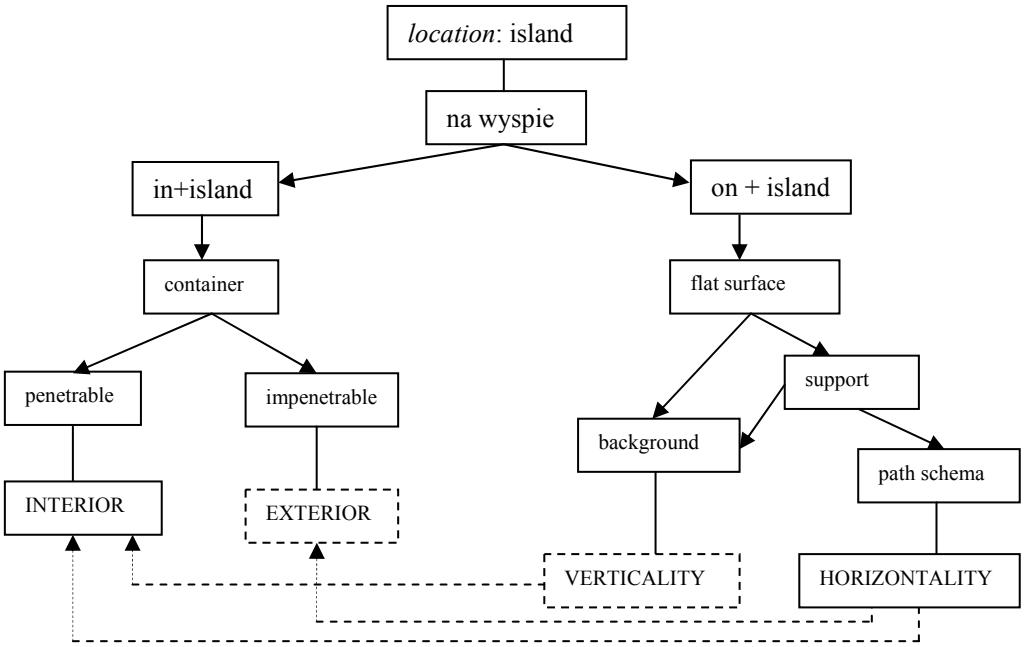


Figure 3: Network of meanings encoded by *in an/the island* and *on an/the island*.

The applicability of the above discussion to practical language teaching may take the form of a set of concordances with words blanked out to be filled in by students (appendix, Table 3).

Grouping the above examples into several categories identified earlier in our discussion could be a follow up activity.

5. CONCLUSION

Bearing in mind the extensive research into the conceptual meaning of *in* and *on* presented by cognitive linguistics, it is legitimate to adhere to the claim that polysemy of meaning occurs within the preposition itself and that it is marshalled in a radial network. Now, let us recall that the preposition, although conceptually dependent, functions as a profile determinant for the subsequent nominal, therefore features of a preposition necessitate shifts in the conceptual meaning of the nominal (prepositional valence). As a consequence, the bifurcation of meaning of the whole prepositional phrase is likely to take place. Thus, while *in an/the island* vs. *on an/the island* permit some degree of semantic overlap at high schematic levels, i.e. there are contexts wherein these prepositions could be used interchangeably, if one wants to achieve greater granularity of the portrayed scene and achieve a higher precision of expression, only one of the possible prepositions tends to be involved. The selected

preposition may display a high degree of prototypicality effect although it does not have to as class membership features are not always preserved in non-central members.

It can be concluded from the above discussion that being a part of a radial set of meanings, the prepositions *in* and *on* may in some contexts display antonymic inclinations at the level of their subschematic structures (vertical divergence), nevertheless some meanings of the whole phrases (*in an/the island* and *on an/the island*) may traverse across the senses distinguished within the network ascribed to a particular preposition and then they sustain phrasal synonymy at different levels of schematicity (horizontal convergence).

APPENDIX

Table 1: Concordances with *in an/the island*.

allusions that depend on the poet's being physically	in the island.	The same is true of another poem by the same
in order to put an end to the oppression by Spain	in the Island	of Cuba. The head of the Labour Party asked
is open and encourages temporal residents to stay	in the island	if they pay regular subscription fee and then
le for both loyalists and conservatives to co-exist	in the island	as a whole. The government wants to secure t
e legend about the volcano that is located somewhere	in the island	and that erupts whenever the God of the islan
nd archeologists have found many valuable vessels	in the island.	Some of them were made by the first inhabit
ere he might be but Peter is believed to be no longer	on the island.	Rather he is expected to be on the peninsula
so at first he hesitated but he finally decided to camp	on the island.	He built a fire over which he cooked some l
f such problems football matches should be played	on the island	where the club can fully satisfy the requirem
hat in different circumstances, if there were left along	on the island	they would fall in love for sure as they seem
Killdear was noticed performing a distraction display	on the island	even though there had been no predators for
had been flipped by a huge wave and then washed up	on an island	near Porto. Police was surprised to find him s
w it he held his breadth in amazement - the school sat	on an island	slightly off the coast, you could see it from t
at the city from his office, floor 120, the highest point	on the island,	with amazement, gazing at the panorama of t
aricatures and some of these statues are 30 meters tall.	On an island	almost denuded of wood, transportation of m

Activity 1. Please fill in the blanks with either *in* or *on*.

Table 2: Concordances with on an/the island.

nd allusions are depend on the poet's being physically		the island. The same is true of another poem by the same
way in order to put an end to the oppression by Spain		the island of Cuba. The head of the Labour Party asked
had been flipped by a huge wave and then washed up		an island near Porto. Police was surprised to find him s
w it he held his breadth in amazement - the school sat		an island slightly off the coast, you could see it from t
at the city from his office, floor 120, the highest point		the island with amazement, gazing at the panorama of t
aricatures and some of these statues are 30 meters tall.		an island almost denuded of wood, transportation of m
ship is open and encourages temporal residents to stay		the island if they pay regular subscription fee and then
os pidgin was noticed performing a distraction display		the island even though there had been no predetors for
hat in different circumstances, if there were left along		the island they would fall in love for sure as they seem
as able for both loyalists and conservatives to co-exist		the island as a whole. The government wants to secure t

REFERENCES

Bączkowska, A. (2008). Contrasting Polish and English prepositions: the case of *w/na deszczu* vs. *in the rain* and *w/na słońcu* vs. *in the sun,* In: B. Lewandowska-Tomaszczyk (ed.), *Corpus Linguistics, Computer Tools, and Applications – State of the Art* (pp. 329-347).

Bączkowska, A. (2010). *Space, Time and Language: A Cognitive Analysis of English Prepositions*. Bydgoszcz: Kazimierz Wielki University Press.

Cuyckens, H. (1991). *Linguistic Semantics of Spatial Prepositions in Dutch: A Cognitive-Linguistic Exercise*. PhD dissertation. Antwerp: University of Antwerp.

Dirven, R. (1993). Dividing up Physical and Mental Space into Conceptual Categories by means of English Prepositions. In C. Zielinski-Wibbelt (ed.), *The Semantics of Prepositions: From Mental Processing to Natural Language Processing* (pp. 73-97). Berlin: Mouton de Gruyter.

Fillmore, C. (1975). *Santa Cruz Lectures on Deixis 1971*. Indiana: Indiana University Press.

Geeraerts, D. (1995). Representational formats in cognitive semantics. *Folia Linguistica, 29,* 21-41.

Goldberg, A. (1992). The inherent semantics of argument structure: the case of the English ditransitive construction. *Cognitive Linguistics*, 3*,* 37-74.

Hawkins, B. (1984). *The Semantics of English Spatial Prepositions*. San Diego: University of California. Unpublished PhD. dissertation.

Janda, L. (1990). The radial network of a grammatical category – its genesis and dynamic structure. *Cognitive Linguistics,* 1*,* 269-288.

Lakoff, G. (1990). *Women, Fire, and Dangerous Things*. Chicago: Chicago University Press.

Lewandowska-Tomaszczyk, B. (2007). Polysemy, Prototypes, and Radial Categories. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics* (pp. 139-169). Oxford: OUP.

Langacker, R. (1987). *Foundations of Cognitive Linguistics*. Stanford: Stanford University Press.

Quirk, R., S. Greenbaum, G. Leech, & J. Svartvik. (1985). *A Comprehensive Grammar of the English Language*. Harlow: Pearson-Longman.

Rosch, E. (1973). Natural categories. *Cognitive Psychology,* 4*,* 328-350.

Rudzka-Ostyn, B. (1989). Prototypes, schemat, and cross-category correspondences: the case of *as*. *Linguistics, 27,* 613-662.

Talmy, L. (2000). *Toward a Cognitive Semantics* (Vols. 1-2). Cambridge: CUP.

Taylor, J. R. (1992). Old problems: adjectives in cognitive grammar. *Cognitive Linguistics, 3,*1-36.

Tuggy, (2007). Schematicity. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics* (pp. 82-116). Oxford: OUP.

Nominalizations in astronomical texts in the eighteenth century

IRIA BELLO VIRUEGA

Universidade da Coruña

Abstract

The main aim of this paper is to study the evolution of nominalizations in scientific register in the 18th century. Nominalizations are well-known markers of scientific discourse. They are complex linguistic devices, whose syntactic ambiguity may result in semantic ambiguity for readers. Studying the evolution of this process across time may help understand its present-day situation and facilitate disambiguation. Halliday asserts that the evolution of the language of science in the last 400 or 500 years has developed new ways of nominalizing (2004:105). In this paper I will attempt to demonstrate the veracity of Halliday's claim. For that purpose, I will consider not only morphology and syntax but also issues of thematic and textual structures as well as sociolinguistic and cognitive concerns.

Key words: nominalization, scientific register, Historical Linguistics, Science History

Resumen

El objetivo de este trabajo es el estudio de la evolución de las nominalizaciones en el registro científico en el siglo XVIII. Las nominalizaciones son complejos marcadores del discurso científico. Uno de los problemas que entrañan para los lectores es la ambigüedad sintáctica que entraña su construcción. El estudio de la evolución de este proceso puede facilitar la comprensión de su situación actual y su desambiguación. Halliday afirma que la evolución del lenguaje de la ciencia en los últimos 400 o 500 años ha desarrollado nuevas formas de nominalización (2004:105). En este trabajo intentaré demostrar la veracidad de la afirmación de Halliday. Para ello, tomaré en consideración cuestiones de morfología, sintaxis, estructuras temáticas e informacionales, sociolingüística y cognición.

Palabras clave: nominalización,registro científico,lingüística histórica, historia de la ciencia

1. INTRODUCTION

The aim of this paper is to study the evolution in the use of the process of nominalization in the 18th century in astronomy texts. To this end, this paper is subdivided in five sections. Section 1 is concerned with sociohistorical issues about the period studied. In this section I will give a brief account of the situation and development of astronomy science in the18th century. In section 2, I will attempt to define the concept of nominalization. Section 3 is concerned with the description of the corpus of data used in this study, the *Coruña Corpus of English Scientific Writing* (*CC*, henceforth). Corpus exploitations will be shown and analyzed in section 4. Finally, section 5 will contain some of the conclusions reached after combining the empirical results obtained from the corpus exploitation, the internal considerations about the nature of nominalizations and the external evidence of the period studied.

2. SCIENCE AND ASTRONOMY IN THE 18TH CENTURY

The 17th century meant the establishment of some decisive changes that affected the development of science history in England. The humanist movement at the end of the 15th century meant a development of experimentation as a method of obtaining knowledge. As Crespo (2004) points out, previous scholastic models of knowledge were based on the establishment of a series of reasoned, purely theoretical deductions derived from a set of previously established principles. Humanists were rather concerned with searching solutions to specific problems. Their interest was not on immutable divine truths but on specific issues that could affect man.

As far as the implications of scientific revolutions on language are concerned, according to Barber (1993), the influence of science was to be seen not only in the expansion of vocabulary but also in the adoption of a plain style written in prose as the most usual way of conveying not only scientific but also any kind of written knowledge: "The rise of scientific writing in English helped to establish a simple referential kind of prose as the central kind in Modern English" (1993: 214). This may be a good starting point for the study of grammatical metaphor and nominalizations. Since nominalizations are a complex way of encoding processes into nouns, it might be supposed that the earlier the text is, the least chances we have to find them. Nominalizations are linguistic structures that contribute to increase the complexity of texts. Thus, as the development of science advances and new complex scientific theories are enunciated, it is reasonable to expect an increase in their use.

Astronomy, the scientific study of celestial objects, is one of the oldest and most popular sciences. In spite of the interest of ancient civilizations in the discipline, modern astronomy began with the invention of the telescope at the beginning of the 17th century. In 1453 Nicholas Copernicus' reassertion of the heliocentric theory meant an impulse for other authors to formulate new theories. Thus, Johannes Kepler (1571-1630) was able to formulate a series of laws of planetary motion that described the elliptical orbits of planets and Galileo Galilei (1564-1642) used the telescope to discover the phases of Venus and the four moons orbiting Jupiter. Sir Isaac Newton (1642-1727) was the leading astronomer in the 17th century. Although he is considered to be rather a physicist and a mathematician, his formulation of the laws of motion and gravitation meant the birth of modern astronomy and astrophysics. His *Philosophiae Naturalis Principia Mathematica* was considered central in astronomy until Einstein's theory of relativity.

The Age of Reason cast a new light on astronomy and it meant the establishment of an independent astronomical discipline. As Rothenberg points out "the older form of science persisted until 1700 in the form of almanacs"(1985: 118). Most scholars were concerned with the study, application and development of the newtonian theory. The universe began to be conceived as a clockwork-like mechanism and astronomers devoted themselves to the calculation and prediction of planetary orbits. However, the eighteenth century was also important because it meant the final separation between religion and astronomy. Up to this century, astronomy was closely related to religion and theology. The immutability of the universe was seen as a proof of the existence of God. Any discrepancy with the doctrine established by the Church was seen as an offence to the institution itself and created bitter social debates. That was the case with Copernicus and Galileo, whose theories were considered heretic. It was not until the Enlightenment that we can find a sharp distinction between religion and astronomy. For scientists of that age, reason and the application of scientific laws could explain how Nature works.

The development of astronomy in North America is to be considered separately. The corpus material for this paper includes texts written in English by English-speaking authors. The Declaration of Independence of the American colonies in 1776 was not only political but also social and linguistic and, consequently, the effects of the independence rapidly affected the development of science in America. Green describes the involvement of American astronomers as observers rather than expanders of science: "American astronomers contributed no great discoveries either empirical or theoretical, but they kept abreast of the latest developments, made and published useful observations, and propounded theories of their own to account for what they observed." (1954: 339). Several authors (Greene, 1954; Yeomans, 1977; Rothenberg, 1985) cite John Winthrop as the responsible for the introduction of the discipline in the seventeenth century. During the colonial period, colleges in the area of Boston monopolized the astronomical activity of the country. American scholars depended on European Journals for publication until 1771, when the American Philosophical Society started to publish its own *Transactions*.

3. APPROACHES TO NOMINALIZATION

In order to study the device of nominalization, it is helpful to consider the treatment that they have received from different linguistic schools, namely the functionalist and the generative ones. Within the functionalist sphere, Banks has described nominalization as a "form of grammatical metaphor whereby a process, which could be encoded as a verb, is encoded non-congruently as a noun." (2005a: 78). Other authors have produced similar definitions for this process (Guillén, 1998; Halliday, 1985, 2004; Ravelli, 1988; Ventola, 1996). The concept of "grammatical metaphor" recalls on the basics of the functionalist approach. Guillén (1998: 368) defines it as "the transference of the linguistic representation of the semantic components of a situation between different lexicogrammatical categories". This notion of "transference of linguistic representation" is linked to the existence of some prototypical realizations of the semantic components in terms of lexical categories. Grammatical metaphors create tension between semantics and lexis, disrupt this prototypical configuration, and assign new realizations to semantic components. Prototypically, the unmarked function of nouns is to express an entity (or Thing) and that of verbs is to express a process. In nominalizations, however, the function of expressing a process is realized by a nominal group (Banks, 2005b). The merging of function and form results in a structure "in which verbal processes are coded in nominal structures" (Ventola, 1996: 153). In nominalizations the semantic component of verbs, that of expressing a process, is encoded into a nominal group. In other words, nominalizations retain the semantic component of the "process" but they present it in the form of an "entity".

Functionalists have mostly tried to describe the functions and advantages involved in the use of nominalizations (Halliday, 1985, 2004; Ventola, 1996; Guillén, 1998; Ravelli, Banks, 2005a, 2005b). A brief account of them would include: lexical cohesion (repetition and summarization); economy, conciseness and packing of information; backgrounding of information (related to theme, rheme and information structures); advancement of discourse and reification. The packing of information and the dynamism nominalizations add to the thematic structure of a text are perhaps the most salient features of this process. Nominalizations "made it possible on the one hand to construct hierarchies of technical terms, and on the other hand to develop an argument step by step, using complex passages 'packaged' in nominal forms as Themes" (Halliday, in Guillén, 1998: 371). Indeed, thematization appears usually tied to the development of information systems in the text

(Halliday, 2005: 55-110). As a logical consequence, the systematic backgrounding of information via nominalized processes allows some degree of systematicity in the balance of backgrounded and foregrounded informatio. However, despite nominalized processes have become standard markers of scientific discourse, the first examples of scientific English tended to code information congruently and respect word classes. The tendency from the 17th century onwards has been to include as much information as possible into the foregrounded nominal group so as to facilitate the introduction of new information.

Concerning the historical evolution of the process, Halliday asserts that the language of science in English has developed into more complex ways of nominalizing processes (2005: 155). The role of verbs is being progressively reduced because their function is no longer that of expressing processes or actions, but that of establishing a set of relations between the processes that are being expressed via nominal groups. As a consequence of this, the current situation of the scientific language in English, Halliday (2004) explains, is the result of a process that started 400 or 500 years ago, when the first scientific texts in English began to be published. The tendency is to nominalize as much as possible.

The generative approach to nominalization tries to describe the transformational rule that best comprises the process of nominalization. Nominalizations are defined as transformations of a verbal phrase into a nominal form[1] (Grefenstette & Teufel, 1995: 98). In the nominalization process, the verb is ejected from its syntactic role into a nominal position. This nominal form can be fulfilled by either a gerundive or a nominal form of the verb (a deverbal noun). As a result of the transformation of the verb into a nominal form, a highly-unpredictable, semantically-emptied support verb is introduced to link the nominalizations with the rest of the sentence. This vision basically coincides with the one presented by Halliday. Grefenstette & Teufel do provide a functional explanation for the process of nominalization. They claim that "nominalizations are used for a variety of stylistic reasons: to avoid repetitions of a verb, to avoid awkward intransitive uses of transitive verbs, in technical descriptions where passive is commonly used, etc." (1995: 98)

One of the main advantages of the generative approach is that it provides two different explanations for the process of nominalization. The main concern in this approach is to determine whether nominalizations derive from verbs - and thus, to some extent, depend on

[1] This definition of "nominalization" establishes a difference between the generative and the functionalist approaches: whereas functionalists consider that nouns that have no cognate verb form should be considered nominalizations as long as they encode a process, the generative approach does not include those verbs under the definition of nominalization because they are not results of any transformation.

them, too-. Thus, the transformational hypothesis focuses on the dependency between the nominal and the original Verb phrase, whereas the lexicalist hypothesis states the independence of the nominalized form. For followers of the lexicalist hypothesis, such as Chomsky (1970), derived nominals are considered nouns in deep structure, not deep-structure transformations. He found that there were two main types of nominalizations and called them gerundive –recognisable by their –ing suffix- and derived nominals. The transformationalist approach tries to demonstrate that nominalizations are the result of a movement that turns a verb into a noun. Lees (1960) was the first one to develop a system according to this approach. The tendency in the following years was to progressively ascribe to the lexicalist approach because it could better explain the semantic differences between the two kinds of nominals. (Grimshaw, 1990; Newmeyer, 1971; Siloni, 1997; Zucchi, 1993).

One of the points that most clearly differentiates both approaches is the issue of whether nominalizations depend on verbs. Whereas generativists question the dependence of nominalizations from verbs, functionalists do not pay much attention to that issue and tend to assume that, even though they are independent, the verbal realization is superior. Indeed, the concept of grammatical metaphor itself is built around the presumption that the verbal component is previous to the nominal one and also, in some way, that the verb is going to rule over the noun.

There is some evidence that shows that nominalizations are totally independent processes. In this sense, I will argue that it is necessary to differentiate between verbal and nominal encodings of processes and that nominalizations are a completely independent way of expressing a process. It has been argued (Banks, 2005b; Halliday, 1985) that grammatical metaphor could be compared to semantic metaphor because both metaphors are based on the shift of either the function or meaning to create certain effects in the text. This view, however, does not acknowledge that, unlike grammatical metaphors, the level of de-automatization that semantic metaphors create is not easily interpreted and it usually requires some kind of preparation. Grammatical metaphors, on the other hand, are easily interpreted and sometimes they are hardly recognizable in texts. Another distinction would relate to the low productivity of semantic metaphors: whereas nominalizations are a marker of adult speech, not every adult speaker is able to generate new ones. Further evidence of the independence of nominalizations may be found in the existence of dead nominalizations and in the tendency to produce automatized expressions. Instances of grammaticalized verbs followed by a nominalization may include: “to perform his Revolution” (Hodgson, 1749)

instead of "to revolve" or "make frequent mention" (Fuller, 1732) instead of "frequently mention".

The device of nominalization can be seen as part of a process. Now there may be more refined and more primitive nominalizations as far as the independence from the Verb group is concerned. There may be instances of nominalizations at different stages of the process VG/Process > NG/Process > NG/Entity, always considering that this is a continuum process rather than a clear-cut set of steps. The fact that, synchronically, some nominalizations are nearer the VG/Process stage and others approach to the NG/Entity do not interfere with the diachronic evolution of the process. Nominalizations should be considered as independent linguistic devices: the fact that a NG indicates the presence of a process does not show any deviance from regular use. This use is rather motivated by other series of parameters -sociolinguistics, cognition, thematic structure-. These considerations should not obstruct the study of nominalization as a changing process over time. Indeed, this linguistic device is key in the study of language change in scientific registers.

4. CORPUS MATERIAL

The corpus material for this study has been taken from the *CETA*, one of the subcorpus contained in the *Coruña Corpus*. The *CC* has been designed to contribute to the diachronic study of English at several linguistic levels. The *CC* includes texts from all scientific disciplines except Medicine. For this study I have selected only the astronomical discipline (*CETA* subcorpus). The corpus contains two texts per decade and discipline. Samples contain around 10,000 words excluding figures, tables, formulae and graph. Issues of representativeness and balance (Moskowich-Spiegel and Crespo, 2007: 349; Lareo and Esteve, 2008: 70; Montoya, 2008: 140) have been taken into account, both from linguistic and extralinguistic perspectives –use of first editions only and balance in text types, gender and origin of authors. There are 21 texts in this study, containing 199326 words. The list of the sample texts and their authors can be found in Annex 1.

5. ANALYSIS OF THE DATA

The initial approach to corpus exploitations consists of the data includes a quantitative analysis of the nominalizations contained in the corpus material. According to Vazquez (2006: 410), "quantification of nominalizations is a matter of counting the number of instances of nominalizations per 1000 words. The measure represents the proportions of total

instances of nominalizations over the total number of words per text" (2006: 410). The total number of nominalizations found reaches 2449, which represents the 12.23 ‰ of the total. Only those nominalizations that are the result of a process of suffixation have been taken into account in this study.

Concerning the deverbal suffixes selected to carry out this analysis, there seems to be no unanimity in delimiting the number of possible deverbal suffixes. According to Biber (1988), the four most common noun derivational suffixes are *-tion, -ness* and *-ity*. From this list, *-ity* should be left out of the present analysis because it is a deadjectival suffix forming nouns. For Merlo and Ferrer (2006) the nine most common nominal endings indicating deverbal derivation are: *-ant, -ee, -er/-or, -age, -al, -ion, sion, -ing* and *-ment*. From this list of deverbal suffixes, *-ant, -ee, -er/-or* and *-al* indicate participants, which places them out of the scope of nominalized processes. As a result, the deverbal suffixes that have been taken into account for this analysis are: -age (*passage*), -ing (*bursting*, *rising*), -ment (*commencement*, *measurement*), -tion (*destruction*), -sion (*regression*).

The lack of agreement in the definition of nominalizations renders the task of selecting nominalizations difficult. Some restrictions apply to the scope of this analysis. Firstly, deverbal nouns that have been lost their verbal meaning are not under the scope of this study. However, ambiguous cases which could be read with both as processes or events have been counted as nominalizations. Words such as *meaning* fall into the first group of verbs because although their form is identical to prototypical nominalizations, they can only be read as events. Similarly, *beginning* and other similar words have been included because one can still read these words as nominalized processes.

The total amount of occurrences is shown in table 1, grouped in periods of twenty years -each period contains four texts-. The mean percentages indicate that there are no big oscillations in the rate of nominalizations in the 18th century and the mean rate remains steady. These data clearly contradict Halliday's hypothesis that scientific language shows a "steady drift towards the nominalizing region" (Halliday, 2004: 174). Acknowledging that Halliday's claims were referred to the evolution of the process of nominalization in the last 400-500 years, it may be argued that the data in this study does not cover the time span intended by him. It is thus expected that the amount of nominalizations increases in texts from the 19th century onwards.

Table 1: Occurrences of nominalizations

Period of time	Occurrences	Mean percentage
1700-1720	466	1,18%
1721-1740	501	1,25%
1741-1760	503	1,26%
1761-1780	466	1,17%
1781-1800	515	1,29%

Having a closer look at the evolution in the use of the different suffixes, it is noticeable that there are great differences in the use of the different suffixes, as table 2 shows:

Table 2: Distribution of different suffixes

Suffix	Occurrences	Percentage
-age	11	0,43%
-ing	442	17,37%
-ment	79	3,10%
-tion	1841	72,36%
-sion	171	6,72%

The most productive suffix is *-tion*, with almost three thirds of the total. The low percentage of occurrences with *-ing* can be surprising. If we consider that *-ing* is a suffix that can be found not only in verbs but also in nouns, and that nominalizations are in-between nouns and verbs, it would be highly expectable to find a high amount of *-ing* nominalizations. The inaccuracy of this hypothesis is linked to the existence of non-finite forms, which are not nouns but can function as them, as in the example:

> (1) "For *it cannot be, that the Rules for directing Ships into Ports through the vaſt Ocean, ſhould certainly and infallibly have their intended Effect (...).*"
> (1715, Whiston; emphasis added)

This non-finite verbs cannot be considered nominalizations and thus they have been excluded from this analysis. Only words with *-ing* suffixes functioning as true nouns have been accepted, as in:

> (2) "Aſcenſion is the **riſing** of any Star, or any part of the Equinoctial above the Horizon." (1702, Curson; emphasis added)

Table 3 shows the evolution of the two most productive nominalizing suffixes. The data contained in it shows that whereas the number of nominalizations with the suffix *-tion* progressively augmented, nominalizations with *-ing* tended to decrease.

Table 3. Use of *-tion* and *-ing*.

Period of time	Occurrences with suffix *-tion*	Mean Percentage	Occurrences with suffix *-ing*	Mean Percentage
1700-1720	300	0,76%	132	0,33%
1721-1740	295	0,73%	139	0,35%
1741-1760	473	1,18%	83	0,20%
1761-1780	389	1,03%	44	0,11%
1781-1800	384	0,96%	74	0,10%

If we consider that the amount of nominalizations during the period studied remained steady -around 1,25%-, the fact that the evolution in the use of the suffixes *-tion* and *-ing* are complementary should be read as a motivated change. This motivation could be that there is a minimum percentage of nominalizations needed to guarantee a level of cohesion in a text. The implications may be that nominalizations are markers of scientific discourse and they play indeed an important role in the configuration of thematic and information structures in texts. As Halliday claims, nominalizations are a "resource for the construction and transmission of knowledge" (2004: 170).

6. CONCLUSION

Nominalization is a complex linguistic device in which a verb is turned into a noun by the process of conversion of affixation. But nominalizations are not only that: by being turned into nouns, they acquire some of the features of nouns and they develop some functions that characterize them. In scientific discourse they play an important role in the inner organization of texts and are essential for the development of discourse, which is one of the most basic aims of scientific texts. The data exposed in this paper shows that the hypothetical tendency to find more nominalizations as time advances is not grounded. There is the underlying supposition, however, that in the 19th century we may find a progressive augmentation of nominalizations. The breakdown of occurrences per suffixes shows the opposed evolutions of the two most productive suffixes –the tendency of -tion is to augment and that of -ing, to decrease-. This fact confirms that nominalizations are essential discourse organizers.

APPENDIX

Annex 1: Table of texts and authors

Autor	Title	Date	Words
Curson, Henry	*The Theory of Sciences illustrated, or the grounds and principles of the seven arts: grammar, logick, rhetorick, musick, arithmetick, geometry, astronomy.*	1702	9846
Morden, Robert	*An introduction to astronomy, geography, navigation, and other mathematical sciences [...]*	1702	10006
Whiston, William	*Astronomical lectures.*	1715	9757
Harris, John	*Astronomical dialogues between a gentleman and a lady.*	1719	9884
Gordon, George	*An introduction to geography, astronomy, and dialling.*	1726	10003
Watts, Isaac	*The knowledge of the heavens and the earth made easy: or, the first principles of astronomy and geography explain'd by the use of globes and maps.*	1726	10049
Fuller, Samuel	*Practical astronomy, in the description and use of both globes, orrery and telescopes. ... with ten curious copper-plates.*	1732	10035
Charlton, Jasper	*The Ladies Astronomy and Chronology in four parts.*	1735	10082
Long, Roger	*Astronomy, in five books.*	1742	10045
Hodgson, James	*The theory of Jupiter's satellites.*	1749	9930
Hill, John	*Urania: or, a compleat view of the heavens.*	1754	10029
Ferguson, James	*Astronomy explained upon Sir Isaac Newton's.*	1756	10042
Stewart, Matthew	*Tracts, physical and mathematical: containing, an explication of several important points in physical astronomy and a new method for ascertaining the sun's distance from the earth.*	1761	9881
Costard, George	*The history of astronomy, with its application to geography, history, and chronology; occasionally exemplified by the globes.*	1767	9959
Wilson, Alexander	*Philosophical transactions - Observations on the solar spots.*	1774	4136
Adams, George	*A Treatise describing the construction and explaining the use of celestial and terrestrial globes.*	1777	9899
Lacy, John	*The universal system: or mechanical cause of all the appearances and movements of the visible heavens.*	1779	5845
Nicholson, William	*An introduction to natural philosophy.*	1782	9932
Bonnycastle, John	*An introduction to astronomy in a series of letters from a preceptor to his pupil.*	1786	9909
Vince, Samuel	*A treatise on practical astronomy.*	1790	9993
Bryan, Margaret	*A compendious system of astronomy.*	1797	10064

REFERENCES

Banks, D. (2005a). Emerging scientific discourse in the late seventeenth century. *Functions of Language,* 12(1): 65-86.

Banks, D. (2005b). On the historical origins of nominalized process in scientific text. *English for Specific Purposes,* 24(3): 347-357.

Barber, C. (1993). *The English language: A historical introduction*. Cambridge: Cambridge University Press.

Chomsky, N. (1979 [orig. English: 1970]). *Sintáctica y semántica en la gramática generativa (C.* Peregrín Otero Trans.). Siglo XXI.

Crespo García, B. (2004). General survey of the growth of scientific culture. A historical approach. In Woodward, Elizabeth et al. (Ed.), *About culture* (pp. 57-65). A Coruña: Universidade da Coruña.

Greene, J. C. (1954). Some aspects of American Astronomy 1750-1815. *Isis*, 45(4) 339-358.

Grefenstette, G., & Teufel, S. (1995). Corpus-based method for automatic identification of support verbs for nominalizations. *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics,* Dublin, Ireland.

Grimshaw, J. (1990). *Argument structure*. Cambridge: MIT Press.

Guillén Calve, I. (1998). The textual interplay of grammatical metaphor on the nominalizations occurring in written medical English. *Journal of Pragmatics, 30* 363-385.

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Continuum.

Halliday, M. A. K. (2004). In Webster J. (Ed.), *The language of science*. London: Continuum.

Halliday, M. A. K. (2004). On the language of physical science. In J. Webster (Ed.), *The language of physical science* (162-178). London: Continuum.

Lareo, I., & Esteve Ramos, María José. (2008). 18th century scientific writing. A study of *make* complex predicates in the *Coruña Corpus*. *Icame Journal, 32:* 69-96.

Lees, R. (1960). *The grammar of English nominalizations*. The Hague: Mouton.

Merlo, P., & Esteve Ferrer, E. (2006). The notion of argument in prepositional phrase attachment. *Computational Linguistics*, 32(3): 341-377.

Montoya Reyes, A. (2008). The presence of cognitive verbs in mathematical texts (1800-1900) of the Coruña Corpus. *Icame Journal, 32:* 139-155.

Moskowich-Spiegel, I., & Crespo, B. (2007). Presenting the *Coruña Corpus*: A collection of samples for the historical study of English scientific writing. In J. L. Pérez-Guerra, D.

Bueno Alonso, González-Álvarez & E. Rama Martínez (Eds.), *Of varying language and opposing creed: New insights into Late Modern English* (341-357). Berlin/New York: Peter Lang.

Newmeyer, F. J. (1971). The source of derived nominals in English. *Language, 47*(4) 786-796.

Ravelli, L. J. (1988). Grammatical metaphor: An initial analysis. In E. H. Steiner, & R. Veltman (Eds.), *Pragmatics, discourse and text. Some systemically-inspired approaches* (133-147). London: Pinter.

Rothenberg, M. (1981). Organization and control: Professionals and amateurs in American Astronomy, 1899-1918. *Social Studies of Science*, 11(3): 305-325.

Siloni, T. (1997). *Noun phrases and nominalizations. The syntax of DPs*. Dordrecht/ Boston/ London: Kluwer Academic Publishers.

Vazquez, I. (2006). A corpus-based approach to the distribution of nominalization in academic discourse. In Hornero, A.M., M.J. Luzón and S. Murillo (Eds.). *Corpus Linguistics: Applications for the study of English* (399-417). Bern: Peter Lang.

Ventola, E. (1996). Packing and unpacking of information in academic texts. In E. Ventola, & A. Mauranen (Eds.), *Academic writing, intercultural and textual issues* (153-194). Amsterdam: John Benjamins.

Yeomans, D. K. (1977). The origin of North American Astronomy - seventeenth century. *Isis*, 68(3): 414-425.

Zucchi, A. (1993). *The language of propositions and events*. Dordrecht/ Boston/ London: Kluwer Academic Publishers.

La constitución de corpus para el estudio de la historia lingüística del Uruguay

VIRGINIA BERTOLOTTI
Universidad de la República

Resumen
Describo aquí las principales fuentes documentales para el estudio de la historia lingüística del Uruguay. Doy cuenta de los criterios subyacentes a la selección de los documentos y de las formas de presentación de estos en el Corpus para el estudio del español en el Uruguay y en el Corpus para el estudio del portugués en el Uruguay. Señalo los avances en la constitución de fuentes para otras lenguas, que han tenido contacto con el español a lo largo de la historia del Uruguay.
Palabras clave: Fuentes documentales, corpus, historia lingüística del Uruguay

Abstract
In this article I intend to characterize the main documentary sources for the study of linguistic history in Uruguay. I explain the underlying criteria in order to select the documents and describe the ways of presentation for the case of the Corpus for the study of Spanish in Uruguay and the Corpus for the study of Portuguese in Uruguay. I highlight the advances in the constitution of sources for the analysis of Indian and African languages, which have had contact with Spanish throughout the whole history of Uruguay. Finally, I analyze briefly the limitations of the described corpus for the investigation.
Keywords: documentary sources, corpora, linguistic history in Uruguay

1. ACERCA DE LA NECESIDAD DE CORPUS HISTÓRICOS

La utilización de corpus para el estudio de lenguas iberorrománicas se remonta a la fundación de la propia lingüística románica (Enrique-Arias, 2009: 11). Sin embargo, la conciencia de la existencia de corpus específicos, distintos de los apoyos documentales clásicos es más reciente y ha sido señalada, por ejemplo, en Company (2001a: 117; 2001b: 208). Esta autora sostiene que los corpus construidos con documentos de archivo, específicamente seleccionados con criterios que emanan de la Lingüística Histórica, constituyen la infraestructura necesaria para comenzar a completar el conocimiento (fónico, morfológico, léxico y sintáctico) del español –americano-, sustituyendo así los textos literarios y los documentos publicados por historiadores como fuentes para la historia de la lengua.

Este giro hacia corpus conformados por algunos tipos de documentos predeterminados se relaciona con la consideración de la lengua oral como el ideal a describir.

En el caso de América, y más específicamente de Uruguay, la necesidad de constituir corpus documentales es mayor, debido a que las manifestaciones lingüísticas ya recogidas y procesadas provenientes de los primeros tiempos de la colonización son escasas y están

compuestas, primordialmente, por manifestaciones literarias o por documentos recogidos por historiadores.

2. El corpus para el estudio del español en el Uruguay

Dos décadas atrás, la Asociación de Lingüística y Filología de América Latina decidió auspiciar la creación de un proyecto coordinado entre diferentes países de América para el estudio de la *Historia del español de América*, coordinado por la tempranamente fallecida lingüista argentina Beatriz Fontanella de Weinberg. El responsable de la puesta en marcha del capítulo Uruguay[1] fue Adolfo Elizaincín, y actualmente, esta responsabilidad recae en Magdalena Coll y Virginia Bertolotti.

No describiré en esta ocasión los nuevos conocimientos sobre el español en el Uruguay que esta línea de trabajo[2] ha generado, sino que me centraré en las características del corpus constituido y en el camino recorrido, con la expectativa de que estas reflexiones puedan ser de utilidad para otros colegas.

3. Constitución del corpus: descripción cuantitativa

Actualmente, el *Corpus para la historia del español en el Uruguay* (siglos XVIII y XIX) se compone de 593 documentos incorporados en diversas etapas de desarrollo del proyecto, que describiré en el apartado siguiente.

El documento más temprano es de 1726 y el que cierra el corpus es de 1904. Esta datación reciente, en comparación con otros corpus europeos y americanos, tiene una explicación histórica: Montevideo[3] fue fundada tardíamente (1724 a 1726) en el contexto americano.

El corpus está constituido por unos mil quinientos folios. El número de folios por documento es muy disímil, ya que, como veremos, incluye desde cartas familiares hasta expedientes judiciales. El número total de palabras es aproximadamente 150.000.

No todos los documentos incluidos en el corpus han sido transcriptos e informatizados. Hay 395 documentos transcriptos de los cuales 380 están en versión digital.

[1] Este corpus tienen un carácter nacional, sin embargo, algunos de los resultados obtenidos confirman la existencia histórica de una zona dialectal del español que pone en evidencia que los recortes nacionales no son pertinentes.

[2] Esta se ha llevado adelante en el Instituto de Lingüística de la Facultad de Humanidades y Ciencias de la Educación, Universidad de la República, Montevideo, con la financiación de la Comisión Sectorial de Investigación Científica de dicha Universidad, en varios bienios.

[3] Capital del Uruguay; actualmente, la mitad de la población del país (1,5 millones de habitantes) se concentra allí.

4. CONSTITUCIÓN DEL CORPUS: DESCRIPCIÓN CUALITATIVA

Para la selección de los documentos se tuvieron en cuenta varios parámetros. Este cuidado en la conformación compensa, parcialmente, el hecho de que se trate de un corpus que no está lematizado ni etiquetado.

Los autores de los documentos. En la medida de lo posible, se ha identificado el origen geográfico de los autores, sin embargo, en el caso de "desconocidos" que el archivo rescata en forma casual, esto no ha sido posible. Cabe aclarar, asimismo, que no todos los autores de los documentos son nacidos en América. Sobre todo en las primeras década fue imposible contar con plumas criollas, ya que el número de personas que sabían escribir entre la población nacida en América era exiguo.

El nivel socio-cultural de estos autores, todos ellos pertenecientes a la minoría alfabetizada sobre todo el siglo XVIII, es heterogéneo. Hemos realizado una tipificación de los autores de los documentos de acuerdo con dos categorías: *culto* y *semiculto*, siguiendo la terminología y la caracterización de Oesterreicher (1996: 324-325)[4], que ha sido probada en algunos textos del corpus.

El lugar de origen de los documentos. La mayor parte de los documentos que constituyen el corpus fueron escritos en la zona del actual Uruguay, sin embargo, también se incluyen documentos de escritores indudablemente identificados como nacidos y criados en la zona, aunque fechados en otras partes.

El tipo de documento es un criterio de selección fundamental para la constitución del corpus. Como mencioné más arriba, la búsqueda de la oralidad en la escritura llevó a privilegiar la selección de cartas, preferentemente familiares, y los pasajes de textos jurídicos que pudieran reflejar la oralidad, típicamente las declaraciones de testigos.[5]

Siguiendo a Barbosa (http://www.mundoalfal.org/indexe.htm, cf. Capítulo IV) hemos tipificado el corpus de acuerdo con el origen social del documento, considerando su pertenencia a la *Administración pública*, a la *Administración privada* o al ámbito de lo *Personal* (Polakof, inédito). Esta tipologización, si bien constituye un principio de

[4] Allí se caracteriza como semiculto a los autores que "escriben o dictan un texto sin conocer suficientemente ni la variedad lingüística exigida por el género respectivo ni las reglas discursivas válidas para la estructuración del texto, y que, muchas veces, no saben aprovechar las posibilidades de la comunicación escrita". Este concepto se puede complementar en cuanto a la materialidad de la escritura con el de maõs inábeis de Marquillas (1998: 763).

[5] La ampliación de los objetivos del trabajo y la incorporación desde hace unos años de los enfoques desde las Tradiciones Discursivas (cf. por ejemplo, Kabatek, 2008), nos han llevado a una mayor permeabilidad en la selección de los documentos, incorporando textos, que quizás en una etapa anterior de la conformación del corpus hubiéramos descartado.

organización para trabajar con los datos, nos muestra que, de todas formas, textos de muy diversa índole se adscriben a una misma categoría.

Con el fin de afinar la caracterización del material al que se enfrentarán los investigadores, hemos tipificado la relación interpersonal entre autor y lector. Bertolotti (en proceso) propone considerar tres indicadores: el grado de *coloquialidad*, el *contenido* y la *cortesía* para determinarla y lo ha contrastado en un corpus de 100 cartas con resultados positivos. Esto no ha sido todavía aplicado al resto del corpus.

Los repositorios de los cuales se seleccionaron textos para su incorporación al corpus fueron diversos: el *Archivo General de la Nación* (Secciones Particulares, Judiciales y el Ex Archivo Administrativo); el *Archivo del Cabildo de Montevideo*, *Archivo de la Curia* (Montevideo), el archivo del *Museo Histórico Nacional*, el *Archivo Saravia* del Centro de Estudios Históricos del Comando General del Ejército; el *Archivo Artigas*[6] y la *Revista Histórica*[7], el *Archivo General de la Nación* (Buenos Aires) y el *Archivo General de Indias* (Sevilla).

5. La presentación actual

Una proporción interesante de los documentos que constituyen el corpus (más de 80%) salen del archivo por primera vez. Se los presenta transcriptos de acuerdo con las *Normas de transcripción de los documentos*[8], ya utilizadas para otras publicaciones y ahora actualizadas. Los documentos están editados cuasipaleográficamente, titulados, ubicados temporalmente y con señalamiento del repositorio del cual provienen, como se puede ver en el ejemplo debajo.

[6] Compilación y publicación de documentos históricos relacionados directa o indirectamente con la vida pública o privada de José Gervasio Artigas, héroe patrio.
[7] Publicación del Archivo General de la Nación en la que, además de diversos estudios de carácter histórico, se incluyen transcripciones o facsimilares de documentos.
[8] Las normas de transcripción utilizadas para los documentos son una adaptación de las "Normas para transcrição de documentos manuscritos" tomadas de Mattos e Silva, R. V. (org.). (2001) de Mackenzie, D. (1986).

117. Carta de Josefa Alfonsín de Lamas a su hijo Andrés Lamas

Siglo: XIX Año: 1829
Ubicación: Archivo General de la Nación, Ex – Archivo y Museo Histórico Nacional, Caja 117, Carpeta 10. Doc. s/i, 1 f.

Josefa Alfonsín señala a su hijo su condición de tal.
[*fol. 1r]
1. Monte bideo, nobienbre 2 d^{e} 829

Miestima do andrecito eresibido tua preciable
del 27 ýensu contes tacion digo que nosolo es
denececida tener asunto para ser este un mo
5. tibo para tener relacion, amas que enti
es un deber, pero aunque no fuese noporeso
debias degar detenerla, porque larelasion
no esta solo sifrada en interes ý por lo
tanto jamas debes deolbidar que soi tu
10. ma, ý que nada mas justo, es que unigo
tenga precente ala que fue sutodo, en
fin andrecito mucho tediria sobreelpar
ticular, pero lodego al tienpo, me alegro
[*fol. 1v]
1. detubuena salu, enel ýnter queda
tuma ýserbidora
Josefa Alfonsin
tantas cosas deLuisito que noseolbida deti
5. ýlomis mo Maria ý Fransisca

Existe, además, una base de datos en la que se resume toda la información ya mencionada.

Los documentos y sus facsimilares así como la base de datos serán hechos públicos próximamente en www.historiadelaslenguasenuruguay.edu.uy y se realizará una versión impresa de la trascripción de 200 de ellos (Bertolotti, Coll, Polakof, en proceso) con el título *Documentos para el estudio de la Historia del español en el Uruguay I*. Esperamos que esto dé lugar a nuevos estudios sobre el español en el Uruguay.

La historia del Uruguay, sin embargo, no es una historia monolingüe.

6. El corpus para el estudio del portugués en el Uruguay

Como ya se ha señalado (Elizaincín, 2003: 605; Bertolotti, Coll, Caviglia y Fernández, 2005:11 y Coll, 2008: 25) solo es posible la comprensión de la realidad lingüística de parte Uruguay indagando sobre la presencia histórica de la lengua portuguesa en este territorio,

que, es sabido, se remonta a los comienzos mismos de la colonización portuguesa en América.

Por esta razón, y a partir del 2004, constituimos un corpus de documentos escritos en el actual territorio del Uruguay[9], tanto con el objetivo tanto de recuperar la presencia histórica de la lengua portuguesa como con el de poder estudiar su contacto con el español.

7. CONSTITUCIÓN DE CORPUS

La guía de nuestra búsqueda, fue, en este caso, la lengua: el criterio que privó fue evidenciar la presencia de la lengua portuguesa en documentos escritos en territorio uruguayo en el siglo XIX o principios del XX. La mayor parte de los autores no son identificables. Si bien el lugar del territorio donde fueron escritos los documentos no fue determinante para la selección, comenzamos la búsqueda en los archivos de la zona fronteriza de Uruguay con Brasil.

El corpus reúne documentos de variada índole: cartas (formales e informales), testamentos, causas civiles, inventarios, recibos, textos argumentativos, proclamas, edictos, oraciones religiosas, anuncios de prensa, cartas de particulares en periódicos, recetas de cocina.

Dado que el objetivo de la constitución del corpus era la documentar la presencia de la lengua portuguesa en el territorio uruguayo y el tipo de contacto con el español, cobra valor una tipología que exprese formas diversas de contacto entre las lenguas. La establecimos en Bertolotti et alii (2005a: 18-24) y la ajustamos en Caviglia, Bertolotti, Coll, 2007). Esta tipología aprehende diversas manifestaciones del contacto entre el español y el portugués y permite deducir si el escribiente es hablante nativo de español, de portugués o de una variedad producto del contacto entre ambas lenguas, como muestro en Bertolotti (2008).

8. LA PRESENTACIÓN ACTUAL

Ochenta y ocho de los documentos seleccionados fueron publicados en una versión impresa (Bertolotti et alii (2005a). Un número mayor de textos (121) lo fueron en CD (Bertolotti, Coll, Caviglia y Fernández, 2005b), incluyendo los facsimilares de 96 de los documentos.

Tal como en el *Corpus para el estudio del español en el Uruguay*, los documentos están transcriptos de acuerdo con las *Normas de transcripción de los documentos* ya

[9] Este proyecto investigación se llevó adelante en el Instituto de Lingüística de la Facultad de Humanidades y Ciencias de la Educación, Universidad de la República, Montevideo, con la financiación de la Comisión Sectorial de Investigación Científica de dicha Universidad.

mencionadas y están encabezados por los datos temporales del documento (el siglo y la fecha del mismo) y su ubicación, como puede verse debajo.

88. Recetas de cocina.
Siglo: XX Fecha: 1911
Gentileza de la familia Notejane.
Trans: VB. Rev: MC, VB. Riv,1

[fol. 70r]
Mostachón grande
6 libras de harina 2 lb de asucar
12 ovos 1 onza de carbonato
[5] enpastase todo junto y esti
rase con el rollo asta quedar
de el groso de un cobre
se cortase redondo latas
untadas con mantega mo
[10] llanse con agua y se pasa
apronella forme quete.

Cuando la complejidad del contenido lo amerita, se encuentra una síntesis del contenido de la pieza.

El material que recoge el corpus se relevó en diversos museos, archivos y bibliotecas. En Montevideo se revisaron los Legajos 1 (1838-1854), 2 (1855-1856), 15 (1873-1874) y 94 (1892) del departamento de Tacuarembó y las Cajas 3 (1826), 4 (1828) y 5 (1829) de Cerro Largo del *Archivo General de la Nación*, Judiciales. También allí, en Particulares se consultó el Archivo Aparicio Saravia, y en Escribanía de Gobierno y Hacienda, los Legajos 128, 129 y 140 (1822-1825). Se consultó también el *Archivo Aparicio Saravia* del Centro de Estudios Históricos del Comando General del Ejército (21 cajas, 1894-1910), el *Archivo Artigas* (Tomos 2, 4, 5, 8, 9, 18, 28, 31-33) y la *Revista Histórica* (Tomo 35), ya referidos.

Dado que los objetivos de esta línea de trabajo son más amplios, se busca también documentar cómo y en qué ámbitos se dio la presencia del portugués en el Uruguay, se consultaron diarios decimonónicos de la región de frontera como por ejemplo *El deber cívico* (de la ciudad de Melo conservados en la Biblioteca Nacional, el Diario *La Verdad* (1897-1900) y algunos números de *La France* de 1908, conservados en la *Biblioteca Artigas* en el departamento de Rivera. En el *Museo da Folha Popular*, en Sant´Ana do Livramento (Brasil), se revisaron *O canabarro* (1894-96), *O cidadão* (1886) y *El Debate* de principios de siglo XX. Fueron consultados también algunos ejemplares de los diarios; *O Maragato* (1908)

y *La Voz de Rivera* (febrero 1886), en manos de particulares. Asimismo se estudiaron algunos libros copiadores de la Policía de 1906 y 1908.

Tanto el corpus como la bibliografía referida estarán disponibles en www.historiadelaslenguasenuruguay.edu.uy.

9. Cómo conformar corpus para el contacto lingüístico: otras lenguas europeas, lenguas indígenas y lenguas africanas

La presencia de otras lenguas europeas todavía no ha sido estudiada en perspectiva histórica, por lo cual, no ha habido impulsos para construir sus corpus.

En cuanto a las lenguas indígenas, esta región, como otras de América, ya estaba poblada cuando llegan los europeos. Sin embargo, no se ha considerado el peso que puede haber tenido el componente indígena en la conformación de nuestro español, quizás porque no hay grupos étnicos originarios que se mantengan como tales y porque existe una fuerte ideología que sostiene que la sociedad uruguaya tiene bases europeas.

Si los estudios sobre lenguas indígenas habladas en este territorio son escasos –a excepción, hecha, por cierto del guaraní, la posibilidad de reconstruirlas a través de corpus documentales es nula, ya que se trató de sociedades ágrafas. Su estudio y el de su contacto con el español queda relegado a fuentes secundarias tal como pueden ser registros de religiosos u otros intelectuales de la época o comentarios de viajeros. Ese trabajo está aun sin realizar.

En el caso de las lenguas africanas, la situación es recientemente más promisoria. Coll (2009; 2010) muestra algunos caminos para el estudio de las lenguas africanas en el Uruguay, a través de causas judiciales, y de fuentes secundarias como relatos de viajeros y crónicas de épocas. Ha conformado un corpus, esencialmente distinto de los dos anteriormente presentados, ya que se trata de la compilación de fuentes literarias para el estudio del contacto de lenguas africanas con el español en el Río de la Plata compuesto por 56 textos, de cuyas características lingüísticas ha avanzado algunos análisis (Coll, 2010).

10. Conclusión

La comparación de las situaciones descriptas nos permite afirmar algo bastante obvio: cuanta más disponibilidad de datos tenemos más exigencias podemos tener con los criterios de constitución del corpus.

Los análisis lingüísticos realizados a partir de los corpus referidos arriba, nos muestran, asimismo, que por sofisticados que sean los criterios de conformación de un corpus documental, hay fenómenos lingüísticos fuertemente ligados a la oralidad que no aparecen allí. Por ejemplo, el *che* rioplatense casi no presenta registros en corpus documentales creados con el fin de acercarse a la oralidad y, sin embargo, se documenta profusamente en la literatura y en las letras de tango, como muestra Bertolotti (inédito), lo cual confirma la relevando de lo señalado en la corriente de las Tradiciones Discursivas.

En síntesis, el trabajo de conformación y análisis de estos corpus nos confirman que un corpus histórico no es lo que queremos sino lo que podemos tener, lo cual requiere entonces, definir estrategias de análisis que nos permitan acercarnos a lo que no tenemos.

REFERENCIAS BIBLIOGRÁFICAS

Bertolotti, V. (2008) El caso del falso Aparicio Saravia: análisis de dos cartas escritas en la frontera uruguayo brasileña en En Espiga, J.; Elizaincín, E. (org.). *Español y Portugués: um (velho) Novo Mundo de fronteiras e contatos*. Pelotas: Educat, (pp. 299-318).

Bertolotti, V. (inédito). Notas sobre el *che*.

Bertolotti, V., Coll, M., Caviglia, S. y Fernández, M. (2005a). *Documentos para la historia del portugués en el Uruguay*. Montevideo: Facultad de Humanidades y Ciencias de la Educación.

Bertolotti, V., Coll, M., Caviglia, S. y Fernández, M. (2005b). *Documentos para la historia del portugués en el Uruguay: Transcripciones y Facsimilares.* Montevideo: Facultad de Humanidades y Ciencias de la Educación. Edición en CD.

Bertolotti, V. Coll; M. Polakof, A.C. (en proceso). *Documentos para el estudio de la Historia del español en el Uruguay I*. Instituto de Lingüística, Universidad de la República.

Caviglia, S.; Bertolotti, V.; Coll, M. (2008). La frontera Uruguay–Brasil: análisis lingüístico de un corpus del siglo XIX. *Spanish in Context 5: 1,* (pp. 20-39).

Coll, M. (2008). Estudios sobre la historia del portugués en el Uruguay: estado de la cuestión. En Espiga, J.; Elizaincín, E. (org.). *Español y Portugués: um (velho) Novo Mundo de fronteiras e contatos*. Pelotas: Educat, (pp. 23-64).

Coll, M. (2009) Fuentes para el estudio del contacto entre el español y las lenguas africanas en Montevideo en los siglos XVIII y XIX. Ponencia presentada en *Rosae – I*

Congresso Internacional de Lingüística Histórica. Universidad Federal da Bahia, Universidade Estadual de Feira de Santana y Universidade do Estado do Bahia.

Coll, M. (2010). *El habla de los esclavos africanos y sus descendientes en Montevideo en los siglos XVIII y XIX: representación y realidad*. Con prólogo de José María Obaldía. Montevideo: Ediciones de la Banda Oriental.

Company, C. (2001a). Aspectos metodológicos prácticos para una filología lingüística del español colonial de México. En Curiel, F., Clark, B. (coords.). *Filología mexicana*, México: UNAM, 111-140.

Company, C. (2001b). Para una historia del español americano. La edición crítica de documentos coloniales de interés lingüístico. *Studia in honorem Germán Orduna*, Alcalá de Henares: Universidad de Alcalá, (pp. 207-224).

Elizaincín, A. (2003). *Testimonios sobre la peculiaridad lingüística fronteriza uruguayo-brasileña*. En Moreno Fernández, F., Gimeno Menéndez, F., Samper, J.A., Gutiérrez Araus, M.L., Vaquero, M., Hernández, C. (eds.) *Lengua, variación y contexto*. *Estudios dedicados a Humberto López Morales* tomo II, Madrid: Arco/Libros, (pp. 605-610).

Enrique-Arias, A. (2009). *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Frankfurt am Main-Madrid: Vervuert-Iberoamericana.

Kabatek, Johanes. (2008). *Sintaxis histórica del español y cambio lingüístico: Nuevas perspectivas desde las Tradiciones Discursivas*. Frankfurt am Main-Madrid: Vervuert-Iberoamericana.

Mackenzie, D. (1986). *Manual de Transcripción para el Diccionario del Español Antiguo*. 4ta. ed. de Victoria A. Burus, trad. de Aurora Santa Olalla. Madison: The Hispanic Seminary of Medieval Studies.

Marquillas, R. (1998). Maõs inábeis nos arquivos da Inquisiçaõ. Fontes para o estudo fonológico do português do século XVIII. En Kremer, D. (ed). *Homenaxe a Ramon Lorenzo*. Vigo: Galaxia, (pp. 761-767).

Mattos e Silva, R. V. (2001). (org.). *Para a história do português brasileiro: primeiros estudos*. São Paulo: Humanitas/FADESP, (pp. 553-555).

Oesterreicher, W. (1996). Lo hablado en lo escrito. Reflexiones metodológicas y aproximación a una tipología. En Kotschi, T., Oesterreicher, W., Zimmermann, K. (eds.). *El español hablado y la cultura oral en España e Hispanoamérica*. Frankfurt am Main-Madrid: Vervuert-Iberoamericana, (pp. 317-340).

Polakof, A. inédito. Algunas cuestiones metodológicas para la sistematización de los datos del corpus de *Historia del Español del Uruguay*. Informe para la asignatura Taller Metodológico II de la Licenciatura en Lingüística de la Facultad de Humanidades y Ciencias de la Educación, Universidad de la República.

A corpus-based approach to the origin and development of the intensifier *deadly* in English

ZELTIA BLANCO SUÁREZ

Universidade de Santiago de Compostela

Abstract

The necessity which speakers feel for constant innovation in language so as to achieve their communicative requirements makes intensification one of the most creative areas of language and, hence, one of the most unsteady. Intensifiers, as one of the strategies to mark intensification in language, have been a very fruitful topic of academic discussion, already from the beginning of the twentieth century. The advent of corpus linguistics, however, has propitiated new approaches to the subject, with innumerable variationist and grammaticalisation studies on a wide variety of intensifiers.

Following a corpus-based approach, this paper tackles the origin and development of the intensifier deadly in the history of the English language, from its beginnings until the early twentieth century. Two diachronic corpora of English, in addition to two major historical dictionaries (the Oxford English Dictionary and the Middle English Dictionary), have been selected as sources of evidence.

Keywords: deadly, intensifier, grammaticalisation, subjectification, corpus-based approach

Resumen

La necesidad de innovación lingüística por parte de los hablantes para satisfacer así sus necesidades comunicativas convierte a la intensificación en una de las áreas más creativas de la lengua y, por consiguiente, en una de las más inestables. Los intensificadores, como uno de los posibles mecanismos de intensificación, han constituido un tema muy fructífero en discusiones académicas, ya desde comienzos del siglo xx. Las contribuciones recientes en lingüística de corpus, sin embargo, han aportado una nueva perspectiva a estos estudios mediante innumerables análisis sobre variación y gramaticalización en diferentes intensificadores.

Siguiendo un enfoque de corpus, este artículo aborda el origen y la evolución del intensificador deadly en la historia de la lengua inglesa, desde sus comienzos hasta principios del siglo xx. Además del Oxford English Dictionary y del Middle English Dictionary, dos corpus diacrónicos del inglés han sido seleccionados como fuentes de información para el presente estudio.

Palabras clave: deadly, intensificador, gramaticalización, subjetivización, enfoque de corpus

1. INTRODUCTION[1]

The apparent human "taste for hyperbolic expression" (Peters, 1994: 271) makes intensifiers one of the most prolific areas for the expression of exaggeration. As prototypical markers of intensification, they are constantly undergoing renewal, and hence, as Bolinger puts it, "afford a picture of fevered invention and competition that would be hard to come by elsewhere" (1972, 18). On this account, intensifiers have always been pivotal in scholarly discussions, both from the point of view of synchrony and of diachrony.

[1] For generous financial support, I am grateful to the following institutions: Spanish Ministry for Education (FPU grant, reference number: AP2008-03626), Spanish Ministry for Science and Innovation and European Regional Development Fund (grant HUM2007-60706), Autonomous Government of Galicia (grants 2008-047 and INCITE-08PXIB204016PR).

In this piece of research, intensification is taken in the sense of Bolinger as the 'linguistic expression of exaggeration and depreciation' (1972: 20), and the label intensifier, along the lines of Bolinger (1972) and Quirk et al. (1985), is to be understood as 'any device that scales a quality, whether up or down or somewhere between the two' (Bolinger, 1972: 17). Furthermore, the term is interpreted here in its broadest sense, so that intensifiers can be seen as modifying not only adjectives and adverbs, but also nouns, as will be shown below.

This paper aims at tracing the historical evolution of the intensifier deadly in the course of time, in order to present its major collocations and how these might have evolved, as well as the grammaticalisation (cf. Hopper and Traugott, 2003) and subjectification (cf. Traugott, 1995; Traugott and Dasher, 2002, among others) processes which this intensifier has undergone. Data for this paper have been drawn from two diachronic corpora of English (cf. section 2). In addition, two historical dictionaries, viz. the *Oxford English Dictionary* (*OED*) and the *Middle English Dictionary* (*MED*), have been consulted as further sources of evidence. Section 3 discusses the results provided by the data which have been consulted for this study. Finally, section 4 presents some general concluding remarks on the overall data in relation to the phenomena of grammaticalisation and subjectification.

2. SOURCES OF EVIDENCE

The online edition of the *OED* will be taken as a point of departure for the analysis of the history of the intensifier deadly in this study, with additional instances taken from the *MED*. Later this evidence will be compared with the data retrieved from two corpora, namely the *Helsinki Corpus of English Texts* (*HC*) and the extended version of the *Corpus of Late Modern English Texts* (*CLMETEV*). Therefore, the present paper will cover the time span going from the Old English period until the early twentieth century. The *HC* will supply data on Old (OE), Middle (ME) and Early Modern English (eModE), whereas the *CLMETEV* will provide information concerning Late Modern English (lModE) and the first decades of the twentieth century, until 1920.

The *HC*, compiled by Matti Rissanen et al. at the Department of English of the University of Helsinki, is a multi-genre corpus which includes a selection of texts from approximately 750 A.D. to 1710. These comprise religious treatises, medical and legal texts, historical and biblical accounts, travelogues, sermons, letters (both private and official), diaries and fictional works. It consists of a total number of 1,572,800 words, distributed across different periods and subperiods, as shown in Table 1.

Table 1: Periodisation of the HC and number of words per period.

OE	OE1 (-850)	OE2 (850-950)	OE3 (950-1050)	OE4 (1050-1150)	TOTAL 413,300 words
ME	ME1 (1150-1250)	ME2 (1250-1350)	ME3 (1350-1420)	ME4 (1420-1500)	TOTAL 608,600 words
eModE	eModE1 (1500-1570)	eModE2 (1570-1640)	eModE3 (1640-1710)		TOTAL 551,000 words

The extended version of the CLMETEV, as its name suggests, covers the Late Modern English period, including texts from 1710 until 1920. It was compiled at the University of Leuven by Hendrik De Smet. As for the genres represented in this corpus, it contains fictional works, scientific writings and personal letters. The total number of words amounts to 14,970,622, and it is subdivided into three different periods: 1710-1780, 1780-1850, and 1850-1920.

3. Analysis and discussion of the data

3.1. Evidence from the historical dictionaries

As can be gathered from the *OED*, deadly, both an adjective and an adverb, originally applied to entities or individuals which were able to cause death, thus being a synonym for 'fatal(ly), mortal(ly)', as in (1) and (2) below:

> (1) The snakes bite deadly, fatall are their teeth. (1627, *OED*, s.v. deadly adv. 1).
>
> (2) □er is no wounde so cruelle; for with out remedye it is dedlych. (c.1430, *OED*, s.v. deadly a. 4a).

This quality was not limited to the physical domain, but could also denote spiritual death, as the following example from the *MED* suggests:

> (3) Þir are þe seuen dedly synnes: Pryde and Envy, Ire, Slouth, Couetyse, Glotony and Lechery. (c. 1440, *MED*, s.v. dedli a. 3).

Deadly could also be used with the meaning 'to the death', as in (4):

(4) Junoes long fostred deadlye reuengement. (1583, *OED*, s.v. deadly a.6).

In a more figurative sense, it could also apply to qualities which were related to death, such as paleness, silence or darkness (cf. example (5) below):

(5) His coloure gan to chaunge in-to a dedely hewe. (c.1460, *MED*, s.v. deadly a. 7.a).

According to the *OED*, as an intensifier meaning 'extreme(ly), excessive(ly)', deadly is not attested until the fourteenth century in its use as an adverb (cf. (6)), while its first record as an intensifying adjective goes back to 1660 (cf. (7)):

(6) I □at es sa dedli dill. (a. 1300, *OED*, s.v. deadly adv. 4).

(7) A deadly drinker he is, and grown exceedingly fat. (1660, *OED*, s.v. deadly a. 8.a).

Examples (1) to (7) above, therefore, reveal the progressive subjectification and grammaticalisation of the intensifier deadly. In (1) to (3), deadly is a manner adverb and a descriptive adjective, respectively, with the meaning 'fatal(ly), mortal(ly)'. Examples (4) and (5), however, do not imply a descriptive meaning but rather favour an evaluative or affective meaning, which provides the bridging context to the final use of deadly as an intensifier, and, hence, its grammaticalisation, as represented in (6) and (7) above. Therefore, the development of deadly fits in with the grammaticalisation clines which have been observed in the literature on the topic, namely that of descriptive adjective > affective adjective > intensifier (cf. Adamson (2000, 55)), and the so-called 'modal-to-intensifier shift' (Partington, 1993) or qualitative adverbs > boosters (Peters, 1994).

4. EVIDENCE FROM THE CORPORA[2]

4.1. The Helsinki Corpus of English Texts

The evidence retrieved from the *HC*, with a total of 67 tokens, points to a dramatic increase in the frequency of occurrence of deadly in the Middle English period, with a rise from 2.17 NF in OE to 7.72 in ME, followed by a drastic reduction to 1.99 NF in Early Modern English (5.73 NF), as shown in Table 2:

Table 2: Distribution of deadly across the different periods in the HC (absolute figures and normalised frequencies (NF) per 100,000 words).

	OE1 (-850)	OE2 (850-950)	OE3 (950-1050)	OE4 (1050-1150)	TOTAL
Deadly (and variant forms)[3]	0 tokens	1 token NF 1.08	8 tokens NF 3.17	0 tokens	9 tokens NF 2.17
	ME1 (1150-1250)	ME2 (1250-1350)	ME3 (1350-1420)	ME4 (1420-1500)	TOTAL
Deadly (and variant forms)	8 tokens NF 7.07	5 tokens NF 5.12	14 tokens NF 7.59	20 tokens NF 9.35	47 tokens NF 7.72
	eModE1 (1500-1570)	eModE2 (1570-1640)	eModE3 (1640-1710)	TOTAL	
Deadly (and variant forms)	7 tokens NF 3.68	3 tokens NF 1.58	1 token NF 0.58	11 tokens NF 1.99	

[2] In order to process the data from the corpora and obtain the concordances, I have made use of WordSmith and, more specifically, of its Concord tool.

[3] The variant forms which have been attested in the *HC* include, in decreasing order of frequency: *dedly*, *deedly*, *deadly*, *deadliche*, *dedliche*, *deadlican*, *deadlich*, *deadlic*, *deadlicum*, *deadlices*, *deedli*, *dea□lich*, *deadly*, *deadlicne*, and *deadliest*.

The data from the corpus also suggest that *deadly* sin,[4] occurring 31 times, is the most frequent collocation, amounting to 46.27% of the examples recorded. This was one of the original descriptive meanings of *deadly* ('entailing spiritual death'). Tables 3 and 4 supply information on the collocations of *deadly* with a descriptive meaning as an adj. and as an adv., respectively, while Table 5 provides data on the subjective meanings of the adjective *deadly* (no subjective meanings as an adverb have been found).

Table 3: Collocations of deadly (adj.) with a descriptive meaning in the HC and their percentage of occurrence.

	Tokens	Percentage
Sin	31	46.27%
Life	7	10.45%
Wound(ed)	3	4.48%
Lichama ('body, corpse')	3	4.48%
Flesh	2	2.98%
Man	2	2.98%
World	1	1.49 %
Creature	1	1.49 %
Land	1	1.49 %
Case	1	1.49 %
Weapon	1	1.49 %
Fight	1	1.49 %
List ('cunning, skill')	1	1.49%

Table 4: Collocations of deadly (adv.) with a descriptive meaning in the HC and their rate of occurrence.

	Tokens	Percentage
Wounded	1	1.49%
Idoruen ('hurt')	1	1.49%
Slain	1	1.49%

Table 5: Collocations of deadly (adj.) with a subjective meaning in the HC and their rate of occurrence.

	Tokens	Percentage
Enemy	3	4.48%
Grief	2	2.98%
Feud	1	1.49%
Ifoan ('enemy')	1	1.49%
Foe	1	1.49%
Leor ('countenance')	1	1.49%

[4] In what follows, items are provided in their Present-day English form, except in those cases in which the word no longer exists.

Table 6 presents the percentages of descriptive and subjective meanings per period in the HC and the number of tokens which have been found of each of the collocations:

Table 6: Distribution of descriptive and subjective meanings of deadly (adj. and adv.) in the HC in the different periods.

Period	Descriptive meanings	Subjective meanings	Total number of tokens	Percentage with descriptive meanings	Percentage with subjective meanings
OE	Life (2), lichoma (3), flesh (1), world (1), man (1), list (1)		9	100%	0%
ME	Sin (27), flesh (1), life (5), man (1), wound (3), slain (1), case (1), land (1), creature (1), idoruen (1), fight (1)	Ifoan (1), enemy (1), foe (1), leor (1)	47	91.48%	8.51%
eModE	Sin (4), weapon (1), wounded (1)	Grief (2), enemy (2), feud (1)	11	54.54%	45.45%

A close examination of the total number of items found in the corpus reveals a clear preponderance of descriptive meanings (58 tokens; 86.57%) versus subjective meanings of deadly (9 tokens; 13.43%). In particular, Table 6 shows that the percentage of descriptive meanings in the HC has witnessed a substantial reduction from OE to eModE (from 100% to 54.54%), and, conversely, the rate of affective meanings in the corpus has increased considerably in the same interval (up to 45.45%).

Regarding the distribution of deadly across the different periods, OE only witnesses combinations with descriptive meanings, viz. lichama ('body, corpse'), flesh, man, world and life. Two illustrative examples are provided under (8) and (9):

(8) & on □issum deadlican fl©sce he hine selfne ©teowde. (OE2. Cura Pastorialis R 52.405.31).
and on this deadly flesh he him self revealed

(9) ond he n©fre ma wona□ ne ne weaxe□ on his endebyrdnesse, ac □enden □a tunglu her lyhta□ and he never more wins not not grows on his position but while the star here shines on □ysse deadlican worolde. (OE3. Martyrology R454). on this deadly world

However, as shown in Table 2, the number of instances of deadly increases significantly in ME, on account of the frequent combination of deadly and sin, as exemplified in (10):

> (10) For sothe ['truly, indeed'], synne is in two maneres; outher it is venial or deedly synne. (ME3. The Parson's Tale P298.C1).

It is also at this point in time when a growth in subjectivity occurs, hence collocations such as deadly foe, leor, ifoan, and enemy. These nouns do not have death as an intrinsic quality, but have some feature which is reminiscent of death. Thus, we have an enemy 'to death' (foe and ifoan are included here), or a countenance which, due to paleness, suggests death. This evinces a more affective value, rather than the original descriptive meaning. Examples (11) and (12) show these affective connotations:

> (11) & swi□e bihalden of ham alle. for lonk he is. & leane. & his leor dea□lich.
> and intently behold of him all for skinny he is and lean and his countenance deadly (ME1. Sawles Warde P169).

> (12) Hwa durste slepen hwil his deadliche fa heolde an itohe sweord up on his heaued?
> Who dared sleep while his deadly foe holds a drawn sword up on his head (ME1. Ancrene Wisse P166).

The occurrence of deadly in combination with these nouns might point to an ongoing process of subjectification. However, as shown in Table 6, in ME there is still a very high percentage of descriptive meanings on account of the very frequent combination of deadly with sin. Furthermore, it is important to bear in mind that the collocations of deadly with descriptive meanings have either negative or neutral connotations, while the ones with subjective meanings have all negative connotations.

The eModE period witnesses a drastic reduction of the number of tokens of deadly from the ME figures (from 7.72 NF to 1.99 NF). There are still collocations with sin, but the number is considerably reduced, only 4 examples. There exist other collocates which also indicate the subjectification of deadly, viz. deadly grief, deadly enemy, and deadly feud, for they do not entail death per se, but rather suggest some feature which is connected with it. In (13) the grief is so intense that it is almost fatal, and in (14) somebody wants the enemy to be dead:

> (13) A deadly grief vnto me […] is that I perceiue my good sonne your husband […] in great displeasure and daunger of great harme therby. (eModE1. Thomas More. Letters to daughter P509).

(14) and among the rest here hee [God] forbiddeth Vsurie, as one of her deadliest enemies: (eModE2. Two sermons on "Of Usurie" PB4R).

Of these examples, only deadly grief was not attested in previous stages. In addition, two other combinations, namely deadly wounded and deadly weapon have been found in the corpus. In these cases, however, deadly retains its descriptive meaning 'mortal(ly)'.

The data from the HC, therefore, accord partially with the potential diachronic evolution depicted in the OED. It has been shown that deadly has progressively acquired more subjective meanings in the course of time. Thus, in OE only descriptive meanings are recorded, but this situation changes in ME, which registers the first signs of subjectification of deadly. Furthermore, ME is the period in which the collocation deadly sin is at its height, which explains the high number of descriptive meanings which can still be found at the time. The situation changes in eModE, since deadly shows more signs of subjectification, with approximately 45% of subjective meanings. Nonetheless, there are still no examples of deadly as an intensifier in the corpus. Furthermore, it is remarkable that the overwhelming majority of the uses of deadly in this corpus correspond to its use as an adjective. In fact, there are only 3 instances in which deadly is an adverb modifying a past participle, to wit those in which it modifies wounded, idoruen, and slain.

5. The Corpus of Late Modern English Texts Extended Version (CLMETEV)

The *CLMETEV*, with 221 tokens of deadly (NF per 100,000 words: 1.48), offers a much wider range of collocations than the *HC*. Thus, we also find deadly sin, which was the most frequent combination in the *HC*, but its proportion is drastically reduced (only 12 tokens; 5.43%). Since the maximum of occurrences of any of the collocations with deadly in this corpus is 13, the variation is enormous. Tables 7 and 8 show the collocations of deadly as an adjective and as an adverb with a rate of occurrence over 1%.[5]

[5] A sample of the remaining collocations follows: *acid*, *animosity*, *arrow*, *atmosphere*, *atom*, *blow*, *colour*, *darkness*, *draught*, *disease*, *doom*, *execration*, *fang*, *Gnome*, *grip*, *hatred*, *instrument*, *intent*, *languor*, *malaria*, *malignity*, *net*, *Nitrous Gas*, *pallid*, *reek*, *serpent*, *sigh*, *skeleton*, *snake*, *stain*, *still*, *stroke*, *war*, and *weapon*.

Table 7: Collocations of deadly (adj.) in the CLMETEV and their rate of occurrence.

	Tokens	Tokens of deadly	Percentage
Sin	12	221	5.43%
Poison	7	221	3.17%
Peril	4	221	1.81%
Paleness	3	221	1.36%
Fear	3	221	1.36%
Nature	3	221	1.36%
Blight	3	221	1.36%
Enemy	3	221	1.36%
Struggle	3	221	1.36%

Table 8: Collocations of deadly (adv.) in the CLMETEV and their rate of occurrence.

	Tokens	Tokens of deadly	Percentage
Pale	13	221	5.88%
White	6	221	2.71%
Sick	5	221	2.26%
Cold	3	221	1.36%

Such a broad spectrum of collocations of deadly might indicate that at this point in time it can virtually combine with any word, thus showing features of grammaticalised elements. Examples (15) and (16) are in this sense far from the typical combinations with wound or sin, and point in this direction:

> (15) The deadly atoms especially lurk in all kinds of clothes and furs. (1844. Eothen, ch.3).

(16) The fiend Ennui awhile consents to pine, There growls, and curses, like a deadly Gnome, (1812. Rejected Addresses, Cui Bono? I).

However, a close examination of the collocates of deadly shows that the vast majority of the tokens have either negative or neutral connotations, and that only 6 of them, representing 2.71% of the total, have positive meanings, albeit the overall meaning of the sentence is negative, as in (17) and (18):

(17) There was a sad, a deadly charm still about the journey. (1920. The Happy Foreigner, ch. XV).

(18) Then Carver Doone, with his deadly smile, […] pointed full at Lorna's heart. (1869. Lorna Doone, ch. XLVIII).

As far as the distribution of descriptive, affective and intensifying uses of deadly, we find significant differences with respect to the *HC*. Table 9 shows the percentages of each reading in the corpus for deadly (adj. and adv.) individually; Table 10, by contrast, presents the overall rate for each of these readings:

Table 9: Individual percentages of descriptive, subjective, and intensifying uses of deadly (adj.) and deadly (adv.) in the *CLMETEV*.

	Percentage of descriptive meanings	Percentage of subjective meanings	Percentage of intensifying meanings
deadly adj.	37.56%	43.89%	0.90%
deadly adv.	2.71%	14.03%	0.90%

Table 10: Overall percentages of descriptive, subjective, and intensifying uses of deadly (adj. and adv.) in the *CLMETEV*.

Percentage of descriptive meanings	Percentage of subjective meanings	Percentage of intensifying meanings
40.27%	57.92%	1.80%

A significant difference with respect to the data from the *HC* is the rise in the proportion of subjective readings attested in this corpus. The subjective meanings in the ME section of the *HC* suggests a rate of 8.51%. In the eModE data from the *HC* this percentage increases up to about 45%, while in lModE, represented by the *CLMETEV*, affective meanings amount to almost 58% of the instances recorded.

Regarding the descriptive meanings in the corpus, we find examples like deadly malaria, sin, disease, acid, fang, or herb, among others. Concerning the affective uses, we can

mention deadly pale(ness), cold, reek, malignity, hatred, or quiet, to mention but a few. The most conspicuous difference with respect to the *HC*, however, is the actual presence of intensifying uses in the corpus, which, although very low (1.80%), is crucial, since it shows that deadly is undergoing a grammaticalisation process. Thus, we find deadly little, deadly deal, deadly fierce, and deadly interest. Two of these instances are given under (19) and (20):

> (19) He held up the deadly little dagger called the misericorde. (1870. The Caged Lion, ch. II).
>
> (20) Master Hardcastle's! Lock-a-daisy, my masters, you're come a deadly deal wrong! (1773. She Stoops to Conquer, Act I).

Moreover, as was already the case in the *HC*, it seems that deadly tends to occur more commonly as an adjective modifier of nouns than as an adverb modifying adjectives or other adverbs. In the case of the *CLMETEV*, this tendency is even more conspicuous than in the HC, since only 39 of the 221 occurrences of deadly in this corpus (17.65% of the total) are instances of the adverb deadly in which it modifies the adjectives pale, white, sick, cold, still, and pallid.[6]

The data from the *CLMETEV*, then, seem to confirm the gradual subjectification of deadly in the course of time, with a significant rise in the number of affective or subjective meanings. In addition, the *CLMETEV* also records instances of intensifying uses, which although still very rare, point to the progressive grammaticalisation of deadly over time, as suggested in the *OED*.

6. CONCLUSION

The analysis of the HC and the *CLMETEV* and of the historical dictionaries has revealed the progressive subjectification and grammaticalisation of deadly over time. Already from the OE period, it could denote both physical and spiritual death. The results suggest, however, that the descriptive meaning of spiritual death clearly prevails over the physical value in ME, in the light of the high number of tokens of the combination deadly sin registered in the HC. This might be due to the paramount importance of religion for the society of the time, epitomised by the seven deadly sins, and to the amount of religious texts contained in this corpus. This is manifested not only in the common collocation deadly sin, but also in the use of other combinations like deadly lichama ('body') or flesh. The momentousness of combats

[6] Cf. Table 8 and footnote 5. Note also that there are 10 tokens of *deadly* (adv.) which have not been included in Table 8, since their rate of occurrence is lower than 1%.

and battles at this time also has a linguistic reflection in other descriptive combinations of deadly with wound(ed), slain, and weapon. Furthermore, it has been suggested that the potential collocates of deadly have expanded along with its progressive subjectification, a process which started in the ME period, continued in eModE and seems to be at its height in lModE. This expansion is perfectly exemplified by the *CLMETEV*, which offers a wide variety of collocations with subjective values. In addition, it seems that in the main deadly collocates with nouns with either negative or neutral meanings, while combinations with positive nouns are infrequent in the corpus. Finally, instances of grammaticalised intensifying uses of *deadly* are very rare in the corpora, in spite of the evidence shown by the OED. Therefore, the diachronic development of *deadly* (adj. and adv.) tallies with the grammaticalisation clines which have hitherto been observed in the literature on the topic. Thus, as can be gathered from the data consulted, in the course of time *deadly* (adj.) has progressively developed more subjective meanings, eventually acquiring intensifying uses; in turn, *deadly* (adv.) has also gained in subjectivity, and from a manner adverb has developed into a degree adverb.

All in all, though limited, the sources of information which have been consulted show in the main the progressive subjectification of *deadly*. As for its grammaticalisation, a more comprehensive survey of the actual uses of deadly and its distribution is needed in order to corroborate the data provided by the *OED*.

REFERENCES

Adamson, S. (2000). A lovely little example: Word order options and category shift in the premodifying string. In O. Fischer, A. Rosenbach and D. Stein (Eds.), *Pathways of Change. Grammaticalization in English* (pp.39-66). Amsterdam: John Benjamins.

Bolinger, D. L. (1972). *Degree Words*. The Hague and Paris: Mouton.

Hopper, P. J. and Traugott, E. C. (2003). *Grammaticalization*. (2nd ed.). Cambridge: Cambridge University Press.

MED = Kurath, H. et al. (Eds.) (1952-2001). *Middle English Dictionary*. Ann Arbor, MI: University of Michigan Press. Online version. Available at: http://quod.lib.umich.edu/m/med/.

OED = Oxford English Dictionary. (1989). (2nd ed.). Oxford: Oxford University Press. Online version with revisions. Available at: http://www.oed.com.

Partington, A. (1993). Corpus evidence of language change. The case of the intensifier. In M. Baker, G. Francis and E. Tognini-Bonelli (Eds.), *Text and Technology. In Honour of John Sinclair* (pp. 177-192). Amsterdam: John Benjamins.

Peters, H. (1994). Degree adverbs in Early Modern English. In D. Kastovsky (Ed.), *Studies in Early Modern English* (pp. 269-288). Berlin: Mouton de Gruyter.

Quirk, R., et al. (1985). *A Comprehensive Grammar of the English Language. London: Longman.*

Traugott, E. C. (1995). Subjectification in grammaticalisation. In D. Stein and S. Wright (Eds.), *Subjectivity and Subjectivisation: Linguistic Perspectives* (pp. 31-54). Cambridge: Cambridge University Press.

Traugott, E. C. and Dasher, R. B. (2002). *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

SOURCES

CLMETEV = Corpus *of Late Modern English Texts Extended* Version. (2006). Leuven: Department of Linguistics.

HC = The Helsinki Corpus of English Texts. (1991). Helsinki: Department of English.

Annotation of linguistic phenomena and topics in query logs

JEAN-LEON BOURAOUI, BENOIT GAILLARD, EMILIE GUIMIER DE NEEF and MALEK BOUALEM

Orange Labs

Abstract

We present a thorough analysis of query logs of various search engines. We first propose a methodology to annotate these logs and explain how it is applied to the corpus. Finally, we report three main observations issued from this study: the distributions of categories of queries significantly vary according to the search engines they come from; Named Entities are very frequent; the queries are often ambiguous and include spelling errors.

Keywords: corpus analysis, query logs, corpus annotation

Resumen

Presentamos un análisis minucioso de las solicitudes en varios motores de búsqueda. Primero, proponemos una metodología para anotar las entradas de datos que se efectúan y explicar cómo se relacionan con el conjunto. Para terminar, establecemos tres observaciones principales a partir de nuestro estudio: la repartición de las categorías de solicitudes cambia de manera significativa en función de los motores de búsqueda ; los nombres propios son muy frecuentes; muchas veces las solicitudes resultan ambiguas y mal ortografiadas.

Palabras clave: análisis de corpus, registros de consultas, cuerpo anotación

1. INTRODUCTION

Analyses of query logs on research engines are quite rare, due to the fact that only search engines editors can access to this kind of corpus. However, categorizing the queries and thematic subjects that interest most of the users is useful for web providers to adjust their offer. From a more theoretical perspective, the categorizing queries also presents some interest since a good classification can yield significant performance gains for search engines. For example, Query Expansion (QE) or *Cross Language Information Retrieval* (CLIR) techniques can benefit greatly from accurate linguistic descriptions of query corpora. This was shown especially by Jensen (2006, p. 185).

In the first part of the paper, the different resources used to carry out the study are described. In a second section an annotation methodology is proposed. Finally, the most interesting phenomena that have been highlighted are presented.

2. PRESENTATION OF THE ANALYZED CORPORA

Five different corpora were used for this research. Four of them are composed of user's queries on various search applications.

The first corpus comes from the query log of a search engine prototype targeting video contents in the domain of news. The corpus contains 3 380 queries with some rare repetitions, submitted to the engine from July 2008 to January 2009. The whole corpus has been annotated.

The second corpus includes query logs of a search engine specialized in User Generated Content (UGC): videos that users generated and uploaded themselves (similar to *YouTube*). The corpus includes 10 078 query patterns[1], ranked by frequency of occurrence (the most frequent query has been entered 695 times). 71.56% of the total number of queries were annotated.

The third corpus is related to the previous one, and corresponds to the index of the search engine (extracted single word patterns, ranked by frequency of occurrence). There are 123 335 words in this corpus; we annotated 39 392 of them, that is to say 27.07%. This resource was used to compare the classification that emerges from the queries to the one that can be observed in the index.

The fourth corpus contains query patterns submitted by users to the news section of 3 separate search engines. These queries bear on textual contents (and not video contents as for the second corpus).

The corpus contains 186 300 queries, ranked by frequency of occurrence (the most frequent query has been performed 29 395 times). The annotation was carried out on the 1000 most frequent query patterns, which corresponds to 45.36% of the total number of actual queries.

The fifth corpus is made of query patterns submitted from mobile phones, on a general search engine. It contains 4 655 433 queries; the most frequent was repeated 630 546 times, and the less frequent 26 times. The annotation was carried out on the 1000 most frequent query patterns, which corresponds to 85.07% of the total number of actual queries.

[1] A query pattern corresponds to the same query entered several times by users. For example, in the second corpus, the query "noel" has been entered 332 times.

3. ANNOTATION METHODOLOGY

3.1. State of the art for query logs annotation

In 1999, some analysis of query logs have been carried out and described in (Silverstein and al., 1999). In 2000, a study by (Jansen and al., 2000) showed that most queries are made of 2 to 4 words; this study was carried out in the framework of Information Retrieval in multimedia data. A similar, more recent study is made by (Chau and al., 2005). In her Phd thesis (Léon, 2008), S. Léon introduces the notion "complex lexical units"[2], that gathers locutions, compound, Named Entities, and describes some methods of extraction and translation of such units. As we will show, the queries often include such units.

More recently (2006), the American provider AOL put on a web sites some query logs corresponding to 3 month (a total number of 20 millions queries, submitted by 650 000 different users). Initially, the goal of AOL was to provide data for the researchers. Each query from these logs is made of several fields: an anonymous ID for each single user, the query itself, the day and hour of the query, and, when applicable, the link clicked by the user among the answers to his queries. Consequently, these logs include all the necessary information for a detailed analysis of the queries submitted on a "common use" search engine. Several websites[3] propose various services devoted to the study of these logs: search engines, ranking of the most frequent sites or words used in queries, etc.

These lists show that the most frequent queries relate to other search engine (firstly Google) or sites with a large audience such as Myspace, Ebay, etc.

The methodology for annotating a given corpus, that is to say set of rules and categories used, is called an "annotation scheme". Such a scheme has to be validated to be considered as robust. In order to do it, it is necessary to compare the annotation of a same corpus by several annotators. It can be done by the use of the "kappa measure", described for example in ((Krippendorff (1980), cited in (Carletta 1996)); basically, it is based on the number of the inter-annotator differences. A more recent study of the methodologies of Named Entities annotation is described in (Fort and al., 2009)

[2] In french: « unités lexicales complexes ».
[3] For example, http://data.aolsearchlogs.com/ or http://www.seosleuth.com/site/ .

3.2. Description of our annotation methodology

We initially have identified some classification topics, with corresponding categories. The goal was to take into account the main relevant linguistic phenomena and topics to represent the queries: morphological and syntactic features, semantic relations, etc. For example, we defined the classification topics for the domain concerned by the query, starting with categories such as: TV, politics …

From there, two annotators carried out a manual analysis and annotation of the two first corpora. Each time they found a query that highlighting an interesting phenomenon that they had not yet identified, they created a new category or topic in which to classify the query. The resulting annotations were frequently compared to each other. The goal was to obtain a unified annotation scheme. Thus, it was possible to assess the importance of topics and categories during the analysis process.

This methodology enabled to dynamically suppress, merge, or detail several categories and classification topics.

The resulting annotation scheme can be represented by trees (that are not necessarily deep). A query belongs to several trees. Thus, annotators and users of the analysis can choose which perspective they want to adopt to study the corpus or the results.

4. PRESENTATION OF THE CLASSIFICATION TOPICS AND CATEGORIES

In the process of annotating the logs presented in section 1, according to the methodology described in section 2, topics and categories were dynamically defined by the annotators. The final topics and categories already are in themselves a significant result as they highlight the salient features of the corpora. They constitute an annotation scheme that is very relevant to the corpus because it was defined by a bottom-up approach, based on the data.

There are 12 different first-level classification topics in the scheme; each one is subdivided into several categories. A category can become a classification topic; for example, the category “music” is also a classification topic that has “rock” as category.

We present in Table 1 an excerpt from the resulting set of topics and categories. Each column corresponds to a given category or topic, and contains some instances.

Table 1: Presentation of the most representative topics and categories of the proposed classification

Lexical categories	**Grammatical categories**	**Categories of error**	**Domain**	**Linguistic phenomena**	**Ambiguities**
Named Entities	Name	Missing or added accentuation	Culture	SMS style[4]	2 different Named Entities
Expression[5]	Commun noun	Gender/number agreement between several words	Sport	Abbreviation	Named Entity or Commun noun
Single Word	Last name	Unrecognized character	Motor engines	Diminutive[6]	Polysemy
Date	Proper name	Deletion of one or several characters	Services	Implicit	Grammatical Ambiguity
Quote[7]	Noun phrase	Insertion of one or several characters	Policy	Play on words[8]	Correction alternatives
	First name	Transposition of one or several characters	General		Different senses according to the language
	Nickname	Inversion of one or several characters	Economy		Different chunking available
	Verb	Segmentation	News		
	Adjective	More than one error in a word	Enterprises		
	Sentence	Phonetic spelling	Health		
	Acronym	Repetition of one or several characters	International		
	Key words		Geography		
	Miscalleanous		Miscalleanous		

Some of the categories are used as topics which are themselves divided into categories. This feature of the annotation scheme can be represented by trees. For example, one of the categories of the topic "*domain*" is "*culture*", which is itself subdivided into categories such as "*music*", which itself is divided into categories such as "*artist*" or "*title*".

[4] Any query written using the same abbreviations than SMS. For example, the english word "before", when written "be4" will be labelled as "SMS style".
[5] We use this term in its linguistic sense: an expression is a set of words that is used as a single unit. For example, "grippe aviaire" ("asian influenza" in English)
[6] Usually used for any short nickname (for example, "Manu" for the French name "Emmanuel").
[7] Any query that correponds to a famous quote (example: "I have a dream").
[8] Any query that produce a humorous play on word.

An instance of such trees is presented in Figure 1 below.

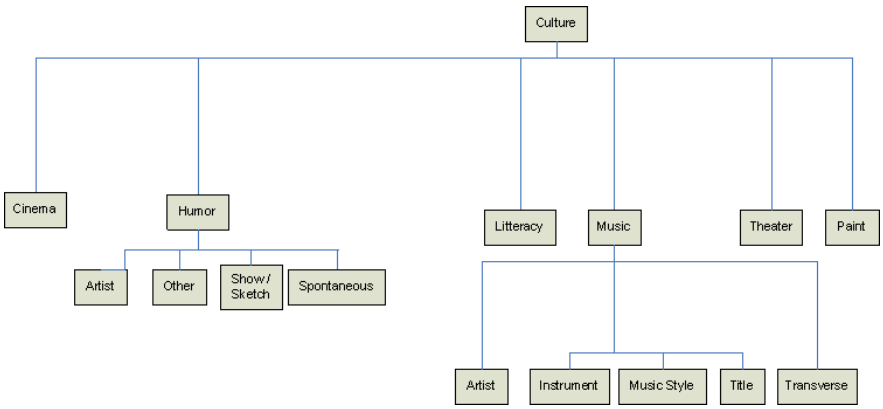


Figure 1: Tree representation of the "culture" classification topic

5. MAIN OBSERVATIONS

Although some observations are common to all corpora, there are also some significant discrepancies between the results of the annotation of the various logs. In consequence, we present in a first section the observations that are common to all corpora, and in a second one those which are specific.

5.1. Observations common to all corpora

A very large number of queries involve Named Entities. The majority of them concern people, places, and TV broadcasts. Figure 2 below displays the distribution, in the first corpus, of Named Entities categories (according to the classification described in section 3.1.); in this corpus, the Named Entities represent 41.06% of the total number of queries (the same trend is observed in the other corpus).

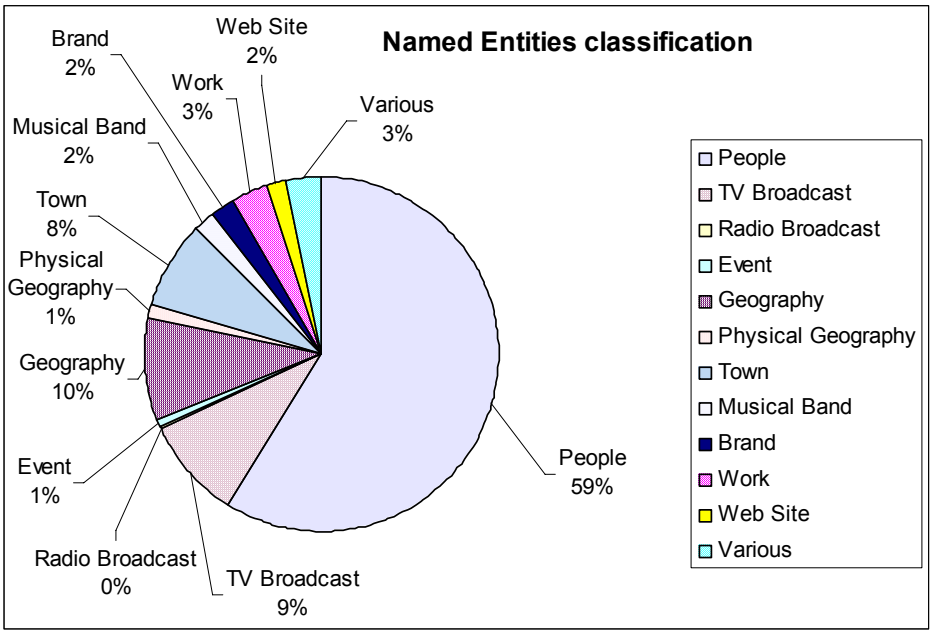


Figure 2 : Distribution of Named Entities categories in the first corpus, according to the proposed classification

In addition to Named Entities, a large number of queries contain phrases, words or compounds.

We noticed frequent spelling mistakes in the queries. Besides overwhelming accentuation and capitalization approximations, most of the mistakes are distributed among omissions, insertion and phonetic errors. These mistakes can cause difficulties while automatically processing queries for application such as CLIR or QE. For example the application can fail to recognize a Named Entity. The typology of errors and its distribution in corpus 1 is displayed in Figure 3 (the "0%" numbers correspond to only a couple of occurrences of the related categories)

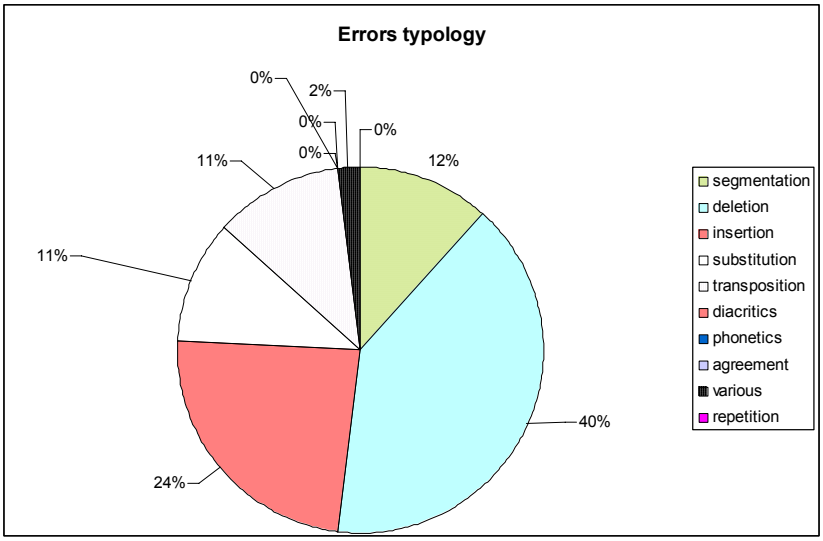


Figure 3 : Distribution of errors in the first corpus

An interesting observation resulting from this annotation work is the significant number of ambiguous queries. An ambiguous query is a query for which various alternative processing can be applied, and for which the choice between these various alternatives is not straightforward. A significant number of them arises from the fact that a word or (group of words) can refer to a Named Entity or to an usual word (for example, the word "cruise" in "Tom Cruise"). Another important cause of ambiguities is the misspelling. This is illustrated by the Figure 4 below.

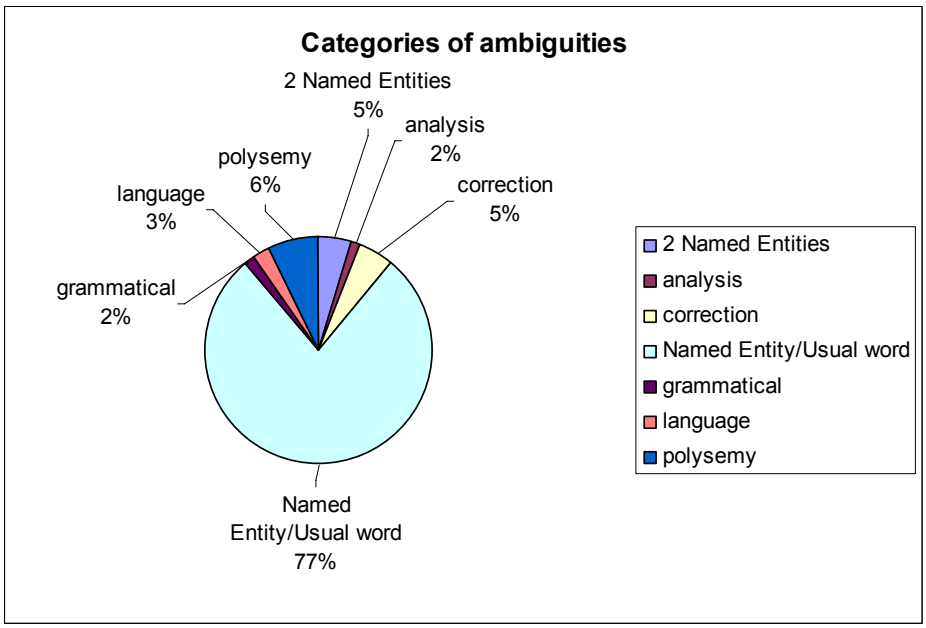


Figure 4: Distribution of ambiguous queries

5.2. Corpus specific observations

The distribution of query categories clearly varies across the analyzed logs. Here, we present two examples of enlevé "the" corpus specific observations.

Since the first corpus results from "test" queries, submitted by researchers on a private prototype, it contains very little number of "adult" queries, whereas the other logs display a significant number of such queries.

We observed, on the mobile portal corpus, that a significant number of queries aim at directly finding the *url* of a known web service (e.g. "Google"); these queries seem to be used by the user as a shortcut or a bookmark to the service. This kind of queries is considerably less frequent in the other logs.

6. CONCLUSION

We presented a method of annotation of web-based query logs. This study involved several resources, following a methodology that uses the data to define the annotation scheme. We showed the benefits that this analysis and annotation scheme can bring to Information Retrieval. It was applied to a significant amount of the available resources. We analyzed and commented the various observed query categories.

From this analysis we can outline several needed processing for a better handling of queries: lemmatizing and interpreting them, using of orthographic correction and identifying their various components according to their typology. Our annotation scheme must also be validated by the use of the kappa coefficient, that we described in section 3.1. above. Finally, query logs analysis is sufficiently rich and adaptable to be used in a more systematic way. Some projects are in progress to automatically label the queries of the first corpus, with the *Tilt* platform (described in (Heinecke *and al.*, 2008)). Thus, it will be possible to compare the labelling done by Tilt with the same work by human annotators.

REFERENCES

Carletta J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, Vol.2, Issue 2. (pp. 249-254).

Chau M., Fang X., Liu Sheng O. R. (2005). Analysis of the query logs of a Web site search engine. *Journal of the American Society for Information Science and Technology*, Volume 56 Issue 13. (pp. 1363-1376).

Fort K., Ehrmann M., Nazarenko A. (2009). Vers une méthodologie d'annotation des entités nommées en corpus ?. *TALN'09*: Senlis, France

Heineke J., Smits G., Chardenon C., Guimier De Neef E., Maillebuau E., Boualem M. (2008). TiLT: plate-forme pour le traitement automatique des langues naturelles. *Traitement Automatique des Langues*, Volume 49 Numéro 2. (pp. 17-41).

Jansen J., Goodrum, A. and Spink, A., (2000). Searching for multimedia: analysis of audio, video and image Web queries. *World Wide Web Journal 3(4)*.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA: Sage.

Léon S. (2008). Acquisition automatique de traductions d'unités lexicales complexes à partir du web. PhD Thesis, 8 november 2008. Aix Marseille University.

Silverstein C., Henzinger M., Marais H., Moricz M. (1999). Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33 (1). (pp. 6-12).

Dormir el sueño de los justos. Fraseología y valores pragmáticos a partir de corpus textuales en alemán y español

PATRICIA BUJÁN OTERO
Universidade de Vigo
CARMEN MELLADO BLANCO
Universidade de Santiago de Compostela

Resumen
El objetivo de este trabajo es poner de relieve las ventajas que conlleva la utilización de herramientas informáticas, como son los corpus textuales, para la determinación del significado connotativo y pragmático de los fraseologismos. Para ello se elige la unidad esp. dormir el sueño de los justos / al. den Schlaf der Gerechten schlafen, con un paralelismo formal total, procedente de la base de datos del campo conceptual MUERTE de nuestro proyecto de investigación[1]*. Frente a una primera intuición que hace pensar que ambos fraseologismos son equivalentes plenos con un único significado, dado su total paralelismo formal, el estudio empírico textual nos lleva a la conclusión de que se trata de un FR polisémico en ambas lenguas, pero con sememas no coincidentes ni en su significado ni en su frecuencia de uso.*
Palabras clave: fraseología, fraseografía, corpus textuales, pragmática

Abstract
The main aim of this paper is to highlight the advantages of using computer tools, such as textual corpora, in order to establish the connotativ and pragmatic meaning of phraseologisms. The phraseologisms sp. dormir el sueño de los justos / ge. den Schlaf der Gerechten schlafen are analyzed, both included under the conceptual field DEATH of the database created within the research project FRASESPAL. Due to their formal parallelism, our first intuition is that they are total equivalents. However, and thanks to the empirical study, we come to the conclusion that they are polysemic phraseologisms in both languages, but with no matching sememes neither in its meaning, nor in its frequency of use.
Key words: phraseology, phraseography, textual corpora, pragmatics

1. BREVE PRESENTACIÓN DEL PROYECTO DE INVESTIGACIÓN LA ESTRUCTURA IDIOMÁTICA DEL ESPAÑOL. UN ESTUDIO COGNITIVO A PARTIR DE UN CORPUS ONOMASIOLÓGICO

El citado proyecto está siendo realizado por el grupo de investigación FRASESPAL y se ha marcado los siguientes objetivos (*cfr.* www.usc.es/frasespal/):

1. Creación de un tesauro onomasiológico del alemán y del español en torno a varios campos conceptuales de especial actividad fraseológica. Se incluye información relativa a la

[1] El proyecto, que es interuniversitario y en el que colaboran ocho universidades, está financiado por el Ministerio de Ciencia e Innovación (Plan Nacional I+D) desde 2007 (HUM2007-62198/FILO) y lleva por título *La estructura idiomática del español*. *Un estudio cognitivo a partir de un corpus onomasiológico.* El presente trabajo se enmarca dentro de este proyecto.

fuente, marca estilística, significado, subcampo conceptual, modelo metafórico, ejemplos, lematización, comentarios pragmáticos y, por último, equivalencias interlingüísticas.

2. Análisis de los modelos metafóricos que subyacen los distintos campos y subcampos semánticos del alemán y del español, con el fin de determinar los paralelismos y las divergencias entre ambas lenguas teniendo como base teórica los principios de la semántica cognitiva. De esta manera se podrá constatar qué es lo común y lo específico de cada una de ambas lenguas en la manera de concebir y verbalizar la realidad extralingüística por medio de estructuras idiomáticas.

3. Creación de un diccionario onomasiológico *on line* con los resultados de 1 y 2.

En cuanto al primer punto, los campos conceptuales elegidos son los siguientes: SALUD/ENFERMEDAD, ETAPAS DE LA VIDA/MUERTE, HABLAR/CALLAR. En este trabajo nos centraremos en los valores pragmáticos de los fraseologismos (FR) del campo MUERTE en alemán y español, campo que en cada una de las lenguas estudiadas alberga en torno a 300 unidades y que nos parece especialmente interesante para un estudio pragmático y contrastivo a partir de corpus textuales.

Quisiéramos resaltar que la principal particularidad de nuestro tesauro radica en que la ordenación macroestructural de los FR se ha hecho siguiendo un principio onomasiológico[2], que va de la idea a la palabra, lo cual tiene numerosas ventajas frente a los diccionarios idiomáticos alfabéticos (*vid.* Mellado Blanco y Buján Otero, 2007). La principal ventaja en el ámbito didáctico es la posibilidad de que los FR se aprendan de manera asociativa, facilitando su memorización y propiciando su uso activo. También para traductores y cualquier persona que utilice el diccionario como herramienta de codificación de textos, los tesauros onomasiológicos brindan la posibilidad de elegir entre varios FR que designan el mismo referente, pero con matices y connotaciones distintas.

Y es en este punto donde consideramos que nuestra aportación es especialmente enriquecedora, puesto que los FR se describen no solo en su significado denotativo, sino también connotativo y pragmático, a partir del análisis minucioso de un número significativo de testimonios textuales y contextos de uso. Gracias a la lingüística de corpus, en la actualidad se puede perfilar el significado y uso de las unidades lingüísticas basándose en testimonios reales, que en no pocas ocasiones pueden llegar a contradecir nuestra primera intuición como hablantes nativos e incluso como lingüistas. No obstante, la introspección

[2] La taxonomía utilizada para clasificar conceptualmente los FR fue creada con una metodología inductiva, una vez estudiada la semántica de los FR recopilados.

sigue siendo desde nuestro punto de vista un método válido para discriminar, sistematizar y priorizar información procedente de bases de datos y corpus textuales.

El tesauro alemán-español elaborado en el marco del proyecto FRASESPAL no ha sido concebido puramente como una herramienta de uso lexicográfico, sino como la materialización del análisis detallado (léxico, semántico, pragmático y cognitivo) de unidades fraseológicas en alemán y español para desarrollar posteriormente un estudio cognitivo de la motivación (fundamentalmente de base metafórica) de las unidades y comparar mecanismos en ambas lenguas/culturas. No obstante, la descripción detallada de estas unidades, el establecimiento de equivalencias a nivel de sistema entre ambas lenguas y la posibilidad de consulta por sinónimos (gracias a la ordenación onomasiológica) enriquecen los posibles usos del tesauro, de forma que resulta útil tanto para el aprendizaje de los fraseologismos en ambas lenguas como para la traducción.

2. METODOLOGÍA

Para llegar a esta descripción detallada de las unidades hemos recurrido al uso de corpus textuales monolingües. Particularmente, mediante la búsqueda de ejemplos representativos del significado de cada unidad para introducir la información correspondiente en el campo EJEMPLO, hemos podido, por un lado, determinar si el significado inicial concebido o que ofrecen los diccionarios consultados se corresponde con el uso real en la lengua y, por otro, delimitar más el significado registrado y enriquecerlo con información pragmática. Esta información constituye la base necesaria para establecer posteriormente relaciones de equivalencia entre ambas lenguas.

En la siguiente sección analizamos la semántica y pragmática de dos de los FR que forman parte del tesauro elaborado por FRASESPAL. Como ya señalamos en la introducción, este análisis se basa en el estudio de ejemplos tomados de textos auténticos, entendiendo por tales aquellos textos en los que el uso del FR es más o menos espontáneo, no condicionado por querer mostrar un ejemplo de uso. Para la localización de ejemplos se han utilizado, para el español, el Corpus de Referencia del Español Actual (CREA[3]) de la Real Academia de la Lengua Española y, para el alemán, la versión de consulta en web de COSMAS II, proyecto elaborado en el Institut für deutsche Sprache[4] de Mannheim. Así mismo, dado el bajo número de ocurrencias que ofrece CREA (en parte, debido a que las búsquedas se han restringido al

[3] Dirección URL de consulta: http://corpus.rae.es/creanet.html
[4] Dirección URL de consulta: https://cosmas2.ids-mannheim.de/cosmas2-web/

ámbito geográfico de España), se ha utilizado como soporte una herramienta de explotación de la web como corpus, *WebCorp Linguist's Search Machine*[5].

3. VALORES PRAGMÁTICOS EN LA FRASEOLOGÍA DEL CAMPO CONCEPTUAL *MUERTE*

Los trabajos que hemos consultado sobre la fraseología del campo conceptual MUERTE abordan el tema generalmente desde una perspectiva cognitiva, en el marco de la teoría de la metáfora de Lakoff y Johnson (1984); así, por ejemplo, Piirainen (2002) para el alemán, Marín-Arrese (1996) para el inglés y el español, o Bierich (1998) para el checo, ruso, croata y serbio. Para el alemán y el español, Piñel (2003) realiza un estudio intercultural analizando las imágenes que subyacen a los FR de ambas lenguas. En estas obras, los valores pragmáticos, fundamentales para establecer equivalencias interlingüísticas fiables, no se mencionan. El aspecto eufemístico de la fraseología de MUERTE se aborda en Luchtenberg (1985: 95-102) y Hessky (2001), lo que supone un valioso acercamiento pragmático al tema. Un enfoque especialmente original es el de Baranov y Dobrovol'skij (2009: 453-459), quienes relacionan la forma interna (imagen) de los FR que designan 'muerte' con la posibilidad de desarrollar polisemia en su significado y con posibles restricciones y preferencias de uso.

El siguiente análisis del FR esp. *(dormir) el sueño de los justos* y al. *den Schlaf der Gerechten (schlafen)* pretende ofrecer un enfoque puramente pragmático y demostrar la utilidad de los corpus textuales para determinar el significado exacto de los FR y los grados de (in-)equivalencia interlingüística.

3.1. Análisis empírico del FR (dormir) el sueño de los justos: significado(s) real(es), colocativos y valores pragmáticos

El estudio que a continuación presentamos tiene por objeto elaborar una definición del FR que se ajuste a su uso real contextualizado[6]. En un principio, introdujimos la unidad *dormir el*

[5] Consulta desde el sitio web http://www.webcorp.org.uk/. También se ha aplicado un criterio de restricción de la consulta a solo las páginas generadas en España (en este caso, sitios web con dominio .es).

[6] Para poder contrastar su significado en base a usos reales con el registrado en diccionarios, hemos consultado varias obras lexicográficas de referencia del español. A este respecto, ni María Moliner, ni DRAE ni Seco lo registran. No obstante, sí está recogido en Cantera (2007, 145) bajo "sueño" como la locución *dormir el sueño eterno/el sueño de los justos* con el significado siguiente: "a. Se dice de un difunto. b. También se dice para significar que ha caído en el olvido". Estos significados se desdicen parcialmente de los resultados obtenidos en el estudio empírico, como también lo hacen con respecto a la explicación recogida en Buitrago (2003: 236) en la entrada *dormir el sueño de los justos* y que reproducimos a continuación: "Estar una persona inactiva o permanecer un objeto sin ser utilizado durante mucho tiempo. *Carlos es un tipo realmente válido para la empresa. No sé por qué lo tienen en ese despacho sin pegar ni golpe, durmiendo el sueño de los justos.*| *Resulta que te empeñaste en comprar el vídeo y lleva quince días en la caja, durmiendo el sueño de los justos.* La frase hace referencia al estado (de inactividad, reposo, vagancia, relax, paz, tranquilidad, ataraxia..., ponga lo que

sueño de los justos en nuestro tesauro con el significado 'morir' dentro del campo conceptual MUERTE. El estudio de las 25 ocurrencias ofrecidas por el CREA para *el sueño de los justos*, usando como única restricción "España" como ámbito geográfico, ofrece interesantes resultados en cuanto a la especialización semántica, y nos ha permitido corregir nuestra primera intuición y reubicar correctamente el FR.

En primer lugar, cabe señalar que, además de usarse con el verbo *dormir*, también encontramos otros colocativos, como *conciliar* o *guardar*. Semánticamente, destaca el hecho de que, en 21 de las ocurrencias, el FR *(dormir) el sueño de los justos* implica 'olvido', 'desatención', y hace referencia a sujeto inanimado [-anim], que suele actualizarse en el habla mediante lexemas que indican proyectos, ideas, propuestas o investigaciones. En cuanto a la tipología textual, este FR suele estar vinculado al lenguaje periodístico[7] o al ensayo. (*Vid.* ejemplos 1a y 1b en apéndice.)

El significado de *(dormir) el sueño de los justos* se podría describir, en el caso de un sujeto inanimado, (1a) y (1b), como 'haber caído en el olvido'. Desde un punto de vista pragmático, con este FR el hablante expresa crítica hacia el estado denotado. El acto discursivo es expresivo. Se trata de un eufemismo usado irónicamente.

En relación a sujeto animado [+anim] o [+hum], el FR adquiere el significado 'descansar plácidamente' (*vid.* ejemplo 1c en apéndice). Ninguna de las ocurrencias ofrecidas por CREA reproduce el significado 'estar muerto', y solo 3 de las 96 ofrecidas por WebCorp denotan este significado (vid. ejemplos 1d, 1e y 1f en apéndice).

Como principal consecuencia que desprendemos de este análisis para nuestro corpus es que la unidad no debería ser incluida en el tesauro, ya que no responde a uno de los principales criterios: uso actual y frecuente con el significado relacionado con la muerte[8]. Podría incluirse, sin embargo, en el corpus SALUD/ENFERMEDAD, ya que hace referencia a dormir plácidamente, no relacionado necesariamente con tener buena conciencia (caso diferente es la unidad *den Schlaf der Gerechten schlafen* que veremos más adelante, en el análisis contrastivo).

El análisis semántico de *(dormir) el sueño de los justos*, de la mano de los corpus textuales consultados, arroja los siguientes resultados:

---

gusten) de quienes, habiendo sido acogidos en el cielo por sus buenas acciones en la tierra, permanecen allí eternamente tranquilos (o tranquilos eternamente)".

[7] 11 textos de prensa escrita, 2 textos de prensa oral, 5 textos de ensayo, 3 textos literarios (2 de narrativa y 1 de teatro).

[8] Aunque aquí no lo detallamos pormenorizadamente, estos mismos resultados son los que se derivan de la búsqueda en WebCorp. La búsqueda se ha restringido a los sitios web con dominio .es y excluyendo aquellas páginas que contengan *comparsa* y *house*. Se han obtenido en total 96 resultados.

—Este FR se basa en una imagen que aparece ya en la Antigüedad clásica (*cfr.* Piñel, 2003: 233) y que se basa en la metáfora MORIR ES DESCANSAR[9], también presente en los FR esp. *descansar en paz, dormir el sueño eterno*, al. *in den ewigen Frieden eingehen, zur ewigen Ruhe eingehen*, y en los monolexemas esp. *descansar, reposar* (la palabra *cementerio* significa 'dormitorio', *cfr.* Crespo Fernández, 2008); al. *einschlafen, entschafen, einschlummern*. Este componente cultural arcaico se ve completado con la visión religiosa de la muerte (presente en el sintagma adnominal *de los justos*), según la creencia del premio final de los buenos[10]. Aunque el significado 'estar muerto' es muy marginal y convendría ahondar en su estudio en base a un mayor número de usos reales, podríamos estar ante un FR con doble vertiente: eufemística (*vid.* ejemplo 1f en apéndice), pues denota un objeto tabuizado culturalmente que se evita nombrar directamente, intentando así "dulcificar" y desdramatizar el acto de la muerte, y disfemística-humorística, como muestran los ejemplos 1d y 1e.

—El FR, con el sujeto [+hum], puede significar también 'dormir plácidamente'. En este caso estamos ante un FR de valor eufemístico con el que el hablante hace una valoración humorística de la acción descrita por el verbo.

—El FR desarrolla con el tiempo un tercer significado ('haber caído en el olvido', muy próximo al valor del FR *estar muerto de risa* con sujeto [-hum]), que se puede caracterizar de irónico o humorístico, cayendo los dos primeros sememas prácticamente en desuso[11]. Se trata de un claro caso de polisemia regular (*cfr.* Baranov / Dobrovol'skij, 2009: 453-459) que surge cuando el sujeto, que tiene la marca [+hum], puede ser usado con la marca [-anim], desarrollando así un nuevo significado (lo mismo sucede con el FR *dormir el sueño eterno*[12]). Se aprecia una clara intencionalidad crítica por parte del hablante, por la

[9] Ya en la *Ilíada*, Homero llama a la muerte el "hermano gemelo del sueño" (*vid.* Bierich, 1998: 21). Crespo Fernández (2008) demuestra en su interesante estudio empírico que la metáfora DESCANSAR ES MORIR es la más frecuente en los epitafios de las tumbas. Se trataría de un eufemismo (no humorístico) en el que la muerte se presenta como una liberación de esta vida, como un paso a un estado mejor (*cfr.* Piirainen, 2002: 226).

[10] Röhrich (2004: 1346-1347) cita como origen del FR los pasajes bíblicos *Libro de los Proverbios* 24,15 y *Moisés* 26,6. Sin embargo, no se trata de un bibleísmo literal procedente de una cita bíblica concreta, sino de un bibleísmo situacional que procede de la sintetización de uno o varios pasajes de los que se extrae una idea (para la tipología de los *bibleísmos, cfr.* Mellado Blanco, 2008).

[11] Hessky (2001: 171) comenta, a propósito de FR religiosos como *das letzte Vaterunser beten* ('rezar el último Padrenuestro': 'morir'), que la desaparición de ciertos tabús en la sociedad provoca que estos FR desaparezcan o que desaparezcan sus rasgos eufemísticos. En el caso de *(dormir) el sueño de los justos* podría haber tenido lugar un proceso semántico de este tipo.

[12] El caso de polisemia de *(dormir) el sueño eterno* es semejante al de *dormir el sueño eterno*. Para sujeto inanimado [-anim], Seco recoge la siguiente acepción: "dormir el sueño eterno [alguien]. v (lit) Estar enterrado [...] dormir el sueño eterno [algo]. v (lit) Permanecer en inactividad definitiva. || PReverte *Maestro* 273: En las paredes había viejas panoplias en las que dormían el sueño eterno herrumbrosos aceros condenados al silencio". Aquí solo podemos corroborar el segundo significado, destacando además una marca pragmática, y es que siempre implica una intencionalidad del hablante crítica, como podemos apreciar en el siguiente ejemplo: "La

desatención a la que se ve sometido el objeto denotado por el sujeto (documentos, archivos, proyectos, etc.). Al contrario que en el caso anterior, el valor pragmático eufemístico no es inherente al propio FR, sino que es dependiente de la situación de uso, y lo denotado no es un objeto tabuizado. Hessky, en su estudio sobre el eufemismo (2001: 170), habla en estos casos de "individuelle bewertende Stellungnahme" ('posición individual valoradora' del hablante). En caso del significado del FR como 'estar muerto' (1d) estaríamos, por el contrario, ante un caso de una "sozial verbindliche Wertung" ('valoración vinculante socialmente') de un objeto denotado tabuizado, con la que se pretende, en palabras de Hessky (2001:170), "hacer inofensivo" lo denotado.

3.2. Análisis empírico del FR den Schlaf der Gerechten (schlafen): significado(s) real(es), colocativos y valores pragmáticos

La definición que ofrece Röhrich (2004: 1346) de este FR es: "tief und ruhig schlafen, ohne sich stören zu lassen" ('dormir profunda y tranquilamente sin ser molestado'), definición[13] que aparece en algunos de los usos registrados y cuyo análisis se presenta a continuación. Así mismo, Schemann (1993) ofrece los siguientes ejemplos tipificantes de sus valores semánticos: "**den Schlaf des Gerechten schlafen** *oft iron.* 1. Die Rosemarie schläft einmal wieder den Schlaf des Gerechten. So einen tiefen Schlaf wie die möchte ich auch mal haben, wenigstens für vier Wochen. 2. Während wir hier Wache schieben, damit keiner die Werkzeuge von dem Bauplatz klaut, schläft der Willi den Schlaf des Gerechten und kümmert sich um nichts.- Den wirst du auch mit den wichtigsten Dingen nicht um seine Nachtruhe bringen; die ist ihm geradezu heilig". Como comentamos anteriormente, el análisis de la semántica de *Schlaf der Gerechten* se hizo a partir de los resultados obtenidos en COSMAS II, que arrojó un total de 102 testimonios, de los que han sido válidos para el estudio 73 (los 29 restantes han sido eliminados por tratarse de repeticiones o referencias al título de una obra literaria). Predomina claramente el colocativo *schlafen*, aunque se registra algún caso con *träumen* y *geniessen*. De este estudio se desprenden los siguientes datos:

---

urbanización que, junto con el campo de golf, fue llamada a ser "la salvadora" de la economía sanluqueña con la pretendida creación de miles de puestos de trabajo, duerme el sueño eterno hasta que otro príncipe bese a la bella urbanización para rescatarla de la apatía y desgana en la que sus creadores la tienen sumida. (GOOGLE / http://www.sanlucardigital.es/index.php?option=com_content&task=view&id=2013&Itemid=50)".

[13] Este significado también lo constata *Duden* (1992: 621): "den Schlaf des Gerechten schlafen (ugs.): *tief und fest schlafen:* Während das ganze Dorf versuchte, den Brand zu löschen, schlief der Feuerwehrhauptmann den Schlaf des Gerechten. ... und da lag der Penner ... und schlief den »Schlaf des Gerechten« (Plievier, Stalingrad 73). ◊ Die Wendung bezieht sich darauf, daß der Gerechte keine Gewissensqualen kennt und deshalb ruhig und fest schläft".

—Se verifica el significado documentado en los diccionarios mencionados de ‘descansar, dormir plácidamente’ (35 casos) referido a sujeto con rasgo semántico [+hum], y que podría equivaler al esp. *dormir como un bendito*. (*Vid.* ejemplo 2a en apéndice.)

—Desde un punto de vista pragmático, constatamos en 24 casos el significado ‘no actuar frente a una situación que requiere una actuación’, con sujeto [+hum], referido siempre a la persona o colectivo (casi siempre un organismo, a menudo de carácter institucional) sobre quien recae la responsabilidad de solucionar una situación. Presenta en este caso una clara intencionalidad irónica y crítica por parte del hablante, y se acerca al valor 2 señalado por Schemann (*cfr. supra*). Se trata, por lo tanto, de un FR irónico, cuyo valor eufemístico está vinculado a una determinada situación de uso y no es intrínseco al propio significado del FR[14]. (*Vid.* ejemplo 2b en apéndice.)

—En 13 casos se registra un uso relacionado con el anteriormente señalado, “inactividad”, pero específicamente empleado en textos periodísticos de carácter deportivo para especificar que, en un partido, la defensa de uno de los equipos ha permanecido pasiva. (*Vid.* ejemplo 2c en apéndice.)

—Aunque muy periféricamente (y, de hecho, fuera de los resultados obtenidos en COSMAS II), también se localizan casos de uso del FR con el significado ‘haber caído en el olvido’ y sujeto [-anim], lo que requiere un estudio más amplio en base a más casos de uso. (*Vid.* ejemplo 2d en apéndice.)

4. Conclusión

El estudio contrastivo de ambos FR pone de manifiesto lo siguiente:

—La frecuencia de uso de los distintos sememas existentes en el FR ocupa un papel relevante para establecer la equivalencia interlingual de los FR. En nuestro caso, la configuración de los sememas es dependiente de los rasgos semánticos del sujeto gramatical en cada caso, [+hum] o [-anim].

[14] En este caso registramos variaciones con los colocativos, *schrecken, rütteln* y *stören*, en las que se utiliza para llamar a la acción a aquel (en ambos casos, organismos) que duerme “den Schlaf der Gerechten” y despertar las conciencias: “Ihr dürft nie vergessen!” Das Musiktheater “Die Endlösung” hat einige Leute unsanft aus dem Schlaf der Gerechten gerüttelt. Der Ring Deutscher Soldatenverbände (RDS) ist offenbar vergeßlich: Millionen Deutsche haben den Völkermord der Nazis zumindest geduldet oder nicht verhindert. Wer dies als ein “ungeheuerliches Pauschalurteil” hinstellt, will Geschehenes verharmlosen. Auch der RDS müßte wissen, daß der Leidensweg der Juden nicht erst in Auschwitz begann” (COSMAS / RHZ96/AUG.17785 Rhein-Zeitung, 31.08.1996; Vergeßliche Veteranen) || “die erstrangige Aufgabe zu, die freie Welt über die Lage der Juden im besetzten Europa zu informieren. Berühmt sind Riegners Telegramme. Über diplomatische Kuriere weitergereicht, suchten sie den Schlaf der Gerechten in London und Washington zu stören” (COSMAS / (E98/JAN.01248 Zürcher Tagesanzeiger, 20.01.1998, S. 11, Ressort: Schweiz; “Am besten würden alle einmal schweigen”).

—Lo divergente en ambos FR es la frecuencia con la que se usan los distintos sememas del FR en alemán y español.

—En español no se observa el significado 'estar inactivo' con sujeto [+hum], por lo que ambos FR son falsos amigos parciales. En los significados coincidentes se observa una fuerte disimetría en la frecuencia de uso.

—El significado primero del FR español ('haber caído en el olvido') no coincide con el semema más usual del FR alemán, que es 'dormir plácidamente'.

De este examen contextual basado en bases de datos concluimos que la amplia coincidencia formal de determinados bibleísmos en varias lenguas no es garantía de equivalencia fraseológica en todos sus niveles. Si como dice Corpas Pastor (2003: 281-282), la equivalencia plena presupone "el mismo significado denotativo y connotativo, una misma base metafórica, una misma distribución y **frecuencia** de uso, las mismas implicaturas convencionales, la misma carga pragmática y similares connotaciones (restricciones diastráticas, diafásicas y diatópicas)" (la negrita es nuestra), entonces no podemos compartir con esta autora (Corpas Pastor, 2003: 282), la idea de que los europeismos, y por inclusión también los bibleísmos, presenten tal tipo de equivalencia[15].

Con este estudio contrastivo de la mano de corpus textuales hay que resaltar que, al contrario de lo que se viene afirmando casi como tópico, (1) los bibleísmos no pueden ser tratados con los mismos parámetros que los europeísmos procedentes de sagas, cuentos, obras y mitos de la Antigüedad Clásica o de la literatura universal; (2) los bibleísmos no siempre son ejemplos prototípicos de equivalencia plena entre las lenguas; (3) las divergencias interlinguales en los bibleísmos pueden estar condicionadas por la distinta evolución semántica de los FR a la hora de desarrollar nuevos sememas, así como por diferencias de uso contextual e implicaturas.

APÉNDICES

Relación de ejemplos referidos en el texto:

(1a) Hacienda ha vuelto a la carga y ha retomado con ganas una investigación fiscal que parecía **dormir el sueño de los justos**: las operaciones de "lavado de cupón (CREA / El Mundo, 09/01/1996: Impuestos. Hacienda investiga el fraude con el "lavado del cupón"...)

(1b) pero como quiera que estas recomendaciones emanadas de reuniones de científicos no son vinculantes para los gobiernos, en muchos casos, como lo fue el de España, las propuestas durmieron

[15] Para una panorámica del concepto de *equivalencia* en fraseología, *cfr.* Buján Otero (2004) y Mellado Blanco (2010).

el sueño de los justos en archivos ministeriales. (CREA / E. Sánchez-Monge, Del "experto" seleccionador a la Ingeniería Genética en mejora de plantas [Historia de la Genética], 1987).

(1c) Se avecina un cálido verano futbolístico a horas intempestivas, y a buen seguro que a través de las ventanas abiertas se difundirá a altas horas de la noche la voz enfervorizada de los comentaristas cantando los goles, para desesperación de quienes sólo tratan de **conciliar el** reparador **sueño de los justos** (CREA / La Vanguardia, 17/06/1994: Mundial 94: no hay salida)

(1d) El empeño digital de la ciudad se ve hasta en las vidrieras de San Pedro el Viejo, panteón real. Allí **duermen el sueño de los justos** (y de todos, vaya) los Reyes Alfonso el Batallador y Ramiro el Monje. Fueron hermanos. (WebCorp / http://perso.wanadoo.es/e/gabrielpe/Huesca/huesca.htm)

(1e) Sin embargo, la intención de la productora de lanzar para el 2009 una película sobre el superhéroe dejaría abierta la puerta a una hipotética resurrección. Tampoco sería la primera vez que Marvel **rescata del sueño de los justos** a uno de sus sobrehumanos protagonistas. El mismísimo Superman resucitó en 1993 (WebCorp / http://www.laverdad.es/murcia/prensa/20070309/gente/muere-capitan-america_20070309.html)

(1f) Para Mi Padre que **descansa / El sueño de los justos** infinito / Que pasea en los bordes de la lontananza / Que vuela sobre los agrestes Abismos. // Tu padre tú que fuiste / Más que luz un sol, más que un hombre un mito / Tú ,que sólo tú conseguiste / Enseñarme la vida como un manuscrito. (WebCorp / http://www.foropoemas.es/index.php?topic=25099.0;wap2)

(2a) *Er läßt sich auf einer Bank nieder, stellt seinen Armband-Wecker und schläft tatsächlich selig und ungestört, bis der Piepston der Uhr ihn hochschreckt. Und dann ist er doch ziemlich überrascht. [...]Wo sind die anderen, die auch ihre Lieben abholen? Der Bildschirm über seinem Kopf sagt ihm, wo die anderen sind: Auf Terminal A. Dorthin nämlich sind die Issos-Urlauber zu später Stunde umgeleitet worden, während unser Franke* ***den Schlaf der Gerechten schlief****. (COSMAS / NUN98/JUL.00389 Nürnberger Nachrichten, 04.07.1998, S. 33; "Wardn" auf die - Ein Franke in Terminal D)*

(2b) *Der Vorsitzende des Bundes Deutscher Kriminalbeamter (BDK), Klaus Jansen, bezeichnete die Sicherheitsmaßnahmen als unzureichend. »Wir haben nichts unter Kontrolle, in Deutschland* ***schlafen wir den Schlaf der Gerechten****.« In der Debatte über weitere Sicherheitsmaßnahmen bekräftigten Politiker von Union und SPD das Ziel, im Herbst eine neue Anti-Terror-Datei einzurichten. Dadurch soll die Polizei auch Zugriff haben auf Geheimdienstinformationen. Bayerns Innenminister Günther Beckstein (CSU) geht der vorliegende Gesetzentwurf aber nicht weit genug. (COSMAS / NUN06/AUG.01499 Nürnberger Nachrichten, 14.08.2006; Mehr Sicherheit - mit Augenmaß - Minister Schäuble warnt vor »überbordendem Aktionismus«)*

(2c) *Mitten in die Kostheimer Euphorie hinein gelang Oberliederbach der Ausgleich. Nach einem Einwurf von der linken Seite* ***schlief*** *die Kostheimer Abwehr* ***den Schlaf der Gerechten****, und die*

Gäste erzielten per Drehschuss das 1:1 (63.). (COSMAS / RHZ00/OKT.01045 Rhein-Zeitung, 02.10.2000; Jascheks cooles Tor)

(2d) *"Ihr habt doch einen totalen Vogel: 99 % der Firmen handeln nicht mit Daten, da liegen die Daten auf der Festplatte und* ***schlafen den Schlaf der Gerechten****. Und es besteht keinerlei Grund die dort aufzuwecken und die intimsten und persönlichsten Daten der Bundesbürger per Post durch die Lande zu schicken".* (GOOGLE / frank.geekheim.de/?p=435)

REFERENCIAS BIBLIOGRÁFICAS

Obras lexicográficas de consulta

Buitrago Jiménez, Alberto (2002). *Diccionario de dichos y frases hechas*. Madrid: Espasa-Calpe.

Cantera, Jesús (2007). *Diccionario de fraseología española: locuciones, idiotismos, modismos y frases hechas usuales en español*. Madrid: Abada.

Drosdowski, Günther / Scholze-Stubenrecht, Werner (1992). *Duden Redewendungen und sprichwörtliche Redensarten: Wörterbuch der deutschen Idiomatik*. Mannheim: Dudenverlag.

Moliner, María (2007 [1900-1981]). *Diccionario de uso del español*. 3.ª ed. Madrid: Gredos.

Röhrich, Lutz (1994). *Lexikon der sprichwörtlichen Redensarten*. Freiburg: Herder.

Schemann, Hans (1992). *Synonymwörterbuch der deutschen Redensarten*. Stuttgart: Klett.

Schemann, Hans (1993). *Deutsche Idiomatik: Die deutschen Redewendungen im Kontext*. Stuttgart: Klett.

Seco, Manuel (dir.) (2004). *Diccionario fraseológico documentado del español actual: locuciones y modismos españoles*. Madrid: Aguilar.

Referencias bibliográficas

Baranov, Anatolij y Dobrovol'skij, Dmitrij (2009). *Aspectos teóricos da fraseoloxía*. Santiago de Compostela: Xunta de Galicia.

Bierich, Alexander (1998). The semantic field "death" in czech, russian, croatian and serbian phraseology. En P. Durčo (Ed.), *EUROPHRAS '97: Phraseology and Paremiology* (pp. 17-23). Bratislava: Akadémia PZ.

Buján Otero, Patricia (2004). Algunhas consideracións sobre a equivalencia fraseolóxica. En J. Varela Zapata, J. M. Oro y J. Anderson (Eds.), *Lengua y sociedad: lingüística aplicada*

en la era global y multicultural (pp. 457-570). Santiago de Compostela: Servizo de Publicacións da Universidade de Santiago de Compostela.

Burger, Harald (2010). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlín: Erich Schmidt.

Corpas Pastor, Gloria (2003). Acerca de la (in)traducibilidad de la fraseología. En Corpas Pastor, Gloria (Ed.), *Diez años de investigación en fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos* (pp. 275-310). Madrid: Iberoamericana.

Crespo Fernández, Eliecer (2008). La conceptualización metafórica del eufemismo en epitafios, *Estudios Filológicos* 43. (pp. 83-100).

Ettinger, Stefan (2004). 'Zeig Pelz die kalte Schulter': Phraseographie und Sprachwirklichkeit. En R. Brdar-Szabó y E. Knipf-Komlósi (Eds.), *Lexikalische Semantik, Phraseologie und Lexikographie*. *Abgründe und Brücken. Festgabe für Regina Hessky* (pp. 315-329). Frankfurt am Main: Peter Lang.

Hessky, Regina (2001). Das euphemistische Idiom - eine Problemskizze. En A. Häcki Buhofer *et al. (Eds.), Phraseologiae Amor. Aspekte europäischer Phraseologie* (pp. 163-175). Baltmannsweiler: Schneider.

Luchtenberg, Sigrid (1985). *Euphemismen im heutigen Deutsch*. Frankfurt a. M. / Berlín: Peter Lang.

Marín-Arrese, Juana I. (1996). To die, to sleep: a contrastive study of metaphors for death and dying in English and Spanish, *Language Science* 18. (pp. 37-52).

Mellado Blanco, Carmen (2007). La Biblia como fuente de idiomaticidad. Pretextos y contextos en alemán y español. En J. de D. Luque Durán y A. Pamies Bertrán (Eds.), *Interculturalidad y lenguaje. El significado como corolario cultural* (pp. 99-108). Granada: Editorial Método.

Mellado Blanco, Carmen (2009). Utilidad y limitaciones de los corpora informáticos en la elaboración de un tesauro fraseológico (alemán-español). En P. Cantos Gómez y A. Sánchez Pérez (Eds.), *A Survey on Corpus-based Research. Panorama de investigaciones basadas en corpus* (pp. 138-151). Murcia: AELINCO.

Mellado Blanco, Carmen (ed.) (2009). *Theorie und Praxis der idiomatischen Wörterbücher* (Lexicographica Series Maior, 135). Berlín: de Gruyter.

Mellado Blanco, Carmen (2010). Die phraseologische Äquivalenz auf der System- und Textebene. En J. Korhonen *et al.* (Eds.), *Phraseologie global - areal - regional* (pp. 277-284). Tübingen: Narr.

Mellado Blanco, Carmen y Buján Otero, Patricia (2007). Die festen Vergleiche im Deutschen, Spanischen und Galicischen. Eine onomasiologische Untersuchung. En E. Kržišnik y W. Eismann (Eds.), *Phraseologie in der Sprachwissenschaft und anderen Disziplinen* (pp. 501-515). Ljubljani: Editorial Univerze v Ljubljani.

Piirainen, Elisabeth (2002). *Er zahlt keine Steuern mehr*. Phraseologismen für 'sterben' in den deutschen Umgangssprachen. En E. Piirainen e I. T. Piirainen (Eds.), *Phraseologie in Raum und Zeit* (pp. 213-238). Baltmannsweiler: Schneider.

Piñel, Rosa M.ª (2003). Der Tod und das Sterben in der deutschen und spanischen Phraseologie: ein interkultureller Vergleich. En H. Burger *et al.* (Eds.), *Flut von Texten – Vielfalt der Kulturen* (pp. 229-238). Baltmannsweiler: Schneider.

Possibilities and limits of corpora in lexicography - an exemplary study of female nouns in corpora and their representation in dictionaries

EMILIE BURI

ANNINA FISCHER

STEFANIE MEIER

German Institute of the University of Basle

Abstract

Beyond controversy, corpus analysis as a method in lexicography provides a range of advantages (Landau, 2001; Tognini-Bonelli, 2001; Hunston, 2002; Sinclair, 2004). However, there are some aspects that have to be seen critically: when analyzing corpora, frequency is generally considered as a vital criterion for entry making. As research has shown, problems arise from the unrepresentativity of a text corpus and consequently from the fact that frequency does not reflect recent changes of a language. Furthermore, applying the principles of critical discourse analysis, there is a discrepancy between language, society, as well as text corpora and the aim of reducing social injustice. This applies, for instance, to the little number of entries of female job and person designations. Until now, this problem has been treated unsystematically or with neglect. We reveal the unsystematic structures of existing dictionaires and reflect on possible solutions using our experience of writing a Swiss German Dictionary. We discuss the question of contemporary corpus-based gender representation on the basis of different dictionaries: Are we forced to unaltered representation of the corpus data? Or do we have a social responsibility?

Keywords: Corpus analysis, lexicography,gGender linguistics,cCritical discourse analysis

Resumen

Sin duda, el análisis de los corpus ofrece muchas ventajas (Landau 2001; Tognini-Bonelli 2001; Hunston 2002; Sinclair 2004). Sin embargo, existen algunos aspectos que requieren estar reflejados críticamente: al analizar los corpus, la frecuencia de las palabras cobra importancia, ya que determina la incorporación de los vocablos a un diccionario. Como ha mostrado la investigación, problemas surgen por la representatividad inexistente y con ello por el hecho que la frequencia no refleja cambios recientes de un lenguaje. Además, implementando los principios del análisis crítico del discurso, hay una discrepancia entre lenguaje, sociedad así como corpus de textos y el objeto de rdeucir injusticia social. Esto ocurre, por ejemplo, al incorporar nombres femeninos de profesiones o nombres comunes femeninos. Hasta ahora, este problema no ha estado estudiado en absoluto o poco sistemáticamente. Develarémos estructuras no sistemáticas de diccionarios y reflejarémos possibles soluciones, aprovechando nuestra experiencia

En nuestra comunicación, discutirémos la temática de la representación actual de los géneros en los diccionarios, trabajando con un corpus de textos. Nuestra base serán diferentes diccionarios. Las preguntas serán: ¿Estamos obligados y obligadas a una representación inalterable de los datos en los corpus? ¿O tenemos una responsabilidad social?

Palabras claves: análisis de los corpus, lexicografía, lingüística del genero, análisis crítico del discurso

1. GENERAL PROBLEMS OF FREQUENCY IN CORPUS ANALYSIS

When analyzing corpora as a method for lexicography, frequency is often regarded as a very important criterion for entry making in dictionaries (cf. for example frequency indications in dictionaries such as the *Collins COBUILD English Dictionary* or the *Macmillan English*

Dictionary for Advanced Learners (MEDAL)). Frequency is for example used in order to decide on the arrangement of the different meanings of a polysemic word or the decision whether a word should be included or not. Generally, this empirically based approach is considered to be the most scientific procedure for writing a dictionary.

Nevertheless, frequency enquiries also entail weak points. Some problems of a frequency-based corpus analysis have already been discussed in detail in research, especially the question of representativity (cf. for example Atkins / Rundel 2008: 63-68). However, an issue that has not been handled to date is the question of how to deal with cases where frequency mirrors linguistic phenomena that should be dismissed from a critical sociological perspective.

In the following, we illustrate what we mean by such linguistic phenomena: It is self-evident that a text corpus does not reflect politically correct language. Unfortunately, one can find a wide range of rassist, sexist or antisemitic expressions in a text corpus. However, since these words are part of our language too, we are obliged to describe them in a dictionary, however, with a semantic description that explains the respective connotation and implication. But what happens with words that are not found or hardly found in a text corpus? Words, that in order to speak and write a fair language should be used and integrated in a dictionary? In this case we refer to the use of female nouns for profession and person designations and to their lexicographical representation and disposal in German and Swiss German Dictionaries.[1]

Working at a new dialect dictionary (*Das Neue Baseldeutsch Wörterbuch*), it became evident that very often the female version of a profession or person designation is much less – if at all – represented in our text corpus than the male version. As Elminger (2009: 63) states that the visibility of women in language is most evident in person designations (mostly designations of profession or function) that describe socially valorised roles, it is extremely important that this finding has to be critically reviewed on a linguistic as well as sociological level. Thus, the question arose: Should we, following Sinclair (2004), "trust the text" and omit including female versions since it was difficult to find them in the corpus? Or do we have a responsibility to deal with the problem normatively?

The following table shows the current state of how this topic is treated in recent dictionaries:

[1] It has to be noted that the issue of masculine and female profession and person designation is different between German and English: In German, we are almost always able to use a female or masculine variant of a profession and person designation. As a general rule, this happens by using a female suffix at the end of the masculine form in order to build the female version.

Table 1

	Duden 2009	*Langenscheidt Grundwortschatz Französisch 2000*	*DWDS: Das Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts*	*Wörterbuch Spanisch – Deutsch, Deutsch – Spanisch (Serges Medien) 2001*	*Rudolf Suters Baseldeutsch-Wörterbuch 2006*
Italian (m) *Italiener*	x	-	x	x	x
Italian (f) *Italienerin*	x	-	0	0	x
Pedestrian (m) Passant	x	-	x	x	-
Pedestrian (f) Passantin	x	-	0	x	-
Street cleaner (m) Strassenfeger	x	-	0	x	x
Street cleaner (f) Strassenfegerin	x	-	x	x	0
Soccer player (m) Fussballer / Fussballspieler	x	-	x	x	x
Soccer player (f) Fussballerin / Fussballspielerin	x	-	0	0	0
House husband, home keeper (m) Hausmann	x	o	x[2]	x	x
House wife, home keeper (f) Hausfrau	x	x	x	x	0
Flight attendant (m) Steward	x	0	x	x	-
Flight attendant (f) Stewardess	x	x	0	x	-
Captain (m) Kapitän	x	x	x	x	-
Captain (f) Kapitänin	x	0	0	0	-
Weaver (m) Weber	x	-	x	x	x
Weaver (f) Weberin	x	-	x	x	x
Student (m) Student	x	x	x	x	x
Student (f) Studentin	x	0	x	x	0
Professor (m) Professor	x	x	x	x	X
Professor (f) Professorin	x	0	x	x	0

[2] Hausmann is labelled as "veraltet", i.e. "out-dated", describing a "janitor".

Shop assisant (m) Verkäufer	x	x	x	x	x
Shop assisant (f) Verkäuferin	x	0	x	x	x
Farmer (m) Bauer	x	x	x	x	x
Farmer (f) Bäuerin	x	x	x	0	x
Voter (m) Wähler	x	x	x	x	x
Voter (f) Wählerin	x	0	0	0	0

Thus, except of the new Duden, no dictionary displayed follows a consequent strategy regarding the incorporation of male and female nouns for professions and person designations. Therefore, aiming to write a dictionary conscientiously, we have to ask ourselves the questions of how to position and explain female nouns for professions and person designations.

2. THE LIMITS OF A CORPUS AS A LEXICOGRAPHICAL METHOD

In order to answer this question, it is inevitable to step further back to get a fuller picture: What is the aim of a dictionary? What is the role of a corpus as a method for data acquisition? And what role has the lexicographer accordingly?

A dictionary is first and foremost an object of utility. It is designed for the use of a, in many cases broader, audience. It is desirable that a dictionary is based on scientific research, which – as earlier mentioned – asks for the use of a text corpus. Older methods such as the lexicographer's competence and the knowledge of a handful of selected informants cannot should be neglected whenever possible. However, in order to write a dictionary ready to publish, the lexicographer takes a vital part by selecting the information obtained by the corpus. Rules have to be made for criterions for entry or exclusion of words or variants. Those criterions cannot be found in an objective corpus but have to be actively made by the lexicographer. A dictionary, therefore, can never be purely descriptive, even though it is the current research aim to get as far away from normativity as possible. Consequently, the most important principle of a scientific researcher is to be aware of one's own influence. Atkins and Rundell (2008: 428, 430) are very explicit, when acknowledging that "impartiality is a good aspiration, but it's important to recognize that a dictionary will inevitably reflect the values of the culture from which it springs" and that "it should be clear that 'neutrality' isn't always possible, so it is important first to be aware of the belief system in which we are

operating (and its possible impact on the 'stance' of some definitions), and secondly to react to changes in the real world as they occur".

As corpuses reflect to a certain degree the society whose texts are incorporated, it is obvious that also societal injustice is evident in the corpus itself. Corpus analysis can therefore be a new methodology for critical discourse analysis (Mautner 2009: 32). It is therefore extremely important that researchers making use of a corpus for another branch of linguistics, such as lexicography, are aware of the ideological weight such a corpus carries. According to critical discourse analysis providing ideological critique belongs to the duties of academic activity. They should use their expertise in order to make the wider public critically aware of the power of language (Fairclough 1995: 18, 221).

After having elaborated on the various factors influencing methods in lexicography the initially asked question of how to position and explain female nouns for professions and person designations can be answered as follows: Even a dictionary aspiring descriptivity can never fully escape the normative decisions of a lexicographer. A dictionary is not the presentation of a linguistic, scientific research but an object of utility. Users are being instructed of the contexts specific words can be used; they are taught what words belong to the linguistic variety described by the dictionary. It is therefore important that the lexicographer makes conscious decisions and uses his expertise in order to make the users of the dictionary aware of language. This can only be achieved if such matters are thought through and a systematic solution is being presented. Our suggestion is to follow the new Duden's practice of incorporating all profession and person designations in their male and female form respectively. However, the decisions made by the lexicographer – whichever it shall be - should be made transparent to the user and be discussed.

REFERENCES

Atkins, B. T. S. and Rundell, M. (2008). *The Oxford guide to practical lexicography*. New York.

Buri, E. (2008). *Das* Schweizer Textkorpus *als empirische Basis für eine kulturlinguistische Untersuchung. Bedeutungswandel und Kultur im Schweizer Hochdeutschen des 20. Jahrhunderts*. Basel: unpublished manuscript.

Chassard, J-N. & Poloni, B. (Ed.) (2000). *Langenscheid Grundwortschatz Französisch.* Berlin etc: Langenscheidt.

Elminger, D. (2009). Sprachliche Gleichbehandlung von Frau und Mann: Eine korpusgestützte Untersuchung über den Sprachwandel in der Schweiz. In *Linguistik* Online 23, 3/2009, (pp. 61-73).

Fairclough, N. (1995). *Critical Discourse Analysis: the critical study of language*. London, New York: Longman.

Fischer, A. (2006). *Geschlechterstereotype in der Anzeigenwerbung – eine Untersuchung zur Signifikanz von Geschlechterstereotypisierungen in der Schweizer Anzeigenwerbung von heute.* Basel: unpublished manuscript.

Heid, U. (2008). Corpus Linguistics and Lexicography. In Anke Lüdeling, Merja Kytö and Tony McEnery (ed.), *Corpus Linguistics. An international Handbook*. Mouton de Gruyter, Berlin, (pp. 131-153).

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge etc.: Cambridge University Press (=The Cambridge Applied Linguistics Series).

Lakoff, R. T. (1975). *Language and woman's place*. New York etc.: Harper and Row.

Landau, S. I. (2001). *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Mautner, G. (2009). Corpora and Critical Discourse Analysis. In: Baker, P. (ed.), *Contemporary Corpus Linguistics*. London, New York: Continuum

Meier, S. (2009). *Zur Empirie der Umfrage. Die Online-Umfrage als neues Datenerhebungsverfahren der Dialektlexikographie*. Basel: unpublished manuscript.

Rothe, U. (2004). Das einsprachige Wörterbuch als Produkt von'Kultur': Lexikographische Definitionen und Artikelbaupläne im Licht semantischer Theorien. In: Herbst, T., Lorenz, G., Mittmann, B. and Schnell, M. (eds.), *Lexikografie, ihre Basis- und Nachbarwissenschaften*. Tübingen: Niemeyer. (pp.71-87).

Sinclair, J. M. (2004). *Trust the Text. Language, Corpus and Discourse*. London: Routledge.

Suter, R. (2006). *Baseldeutsch-Wörterbuch*. Basel: Christoph Merian Verlag.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam, Philadelphia: John Benjamins Publishing (=Studies in Corpus Linguistics 6).

Van Dijk, T. (2004). Discourse, knowledge and ideology: Reformulating old questions and proposing some new solutions. In Pütz, M., Neff-van Aertselaer, J., van Dijk, T. (eds.), *Communicating Ideologies: Multidisciplinary Perspectives on Language, Discourse, and Social Practice*. Frankfurt am Main: Peter Lang. (pp. 5-38).

La interlengua en el léxico disponible de un grupo alumnos de portugués en México

ERÉNDIRA D. CAMARENA ORTIZ
Universidad Nacional Autónoma de México

Resumen
Esta ponencia es un estudio de caso del léxico de un grupo avanzado de portugués de alumnos universitarios como parte de una investigación de léxico disponible en lenguas extranjeras con 16 centros de interés que se está realizando en el Centro de Lenguas Extranjeras de la Universidad Nacional Autónoma de México. El estudio está basado en el concepto de interlengua desarrollado por Selinker y su objetivo ha sido la identificación y análisis de errores de interferencia, hibridización, invención, analogía y transferencia en léxico basados el criterio de correcto e incorrecto a partir de las gramática y diccionarios de portugués con la intención de implementar mejoras en el proceso de enseñanza – aprendizaje de esta lengua próxima.
Palabras clave: léxico disponible, interlengua, lenguas extranjeras, lenguas próximas

Summary
This presentation is a case study on the vocabulary of an advanced group of university students of Portuguese language. It is part of a research on available lexicon of foreign languages with 16 different areas that is being held in the Foreign Language Center at the National Autonomous University of Mexico. This study is based on the interlanguage concept by Selinker and its main goal was to identify, analyze mistakes and categorize them by interference, hybridization, invention, analogy and transference, based on the correct and incorrect criteria established by general Portuguese grammars and dictionaries with the intention to improve close language teaching and learning.
Keywords: available lexicon, interlanguage, foreign languages, close languages

1. INTRODUCCIÓN

Esta ponencia versa sobre los resultados de un estudio adjunto que forma parte de un proyecto mayor de investigación denominado "Léxico Disponible en Lenguas Extranjeras" realizado en el Centro de Enseñanza de Lenguas Extranjeras de la Universidad Nacional Autónoma de México.

Se trata del estudio de caso del léxico disponible de 16 centros de interés de un grupo avanzado del Departamento de Portugués, que fue analizado con el objetivo de identificar algunas dificultades relacionadas con esta lengua próxima.

2. OBJETIVOS

La tarea fundamental de este trabajo es la detección, análisis y medición de errores basándonos en los criterios de correcto e incorrecto a partir de los diccionarios y de las gramáticas generales de portugués de Brasil y Portugal para relacionarlos con la interlengua.

Su propósito es analizar y controlar ortografía, léxico y otros elementos para implementar mejoras en la enseñanza enfrentando problemas concretos, respondiendo a las preguntas de: ¿cuáles son los errores de interlengua más comunes en el portugués escrito?, ¿cuál es el índice de incidencia? y ¿qué tipo de errores son?

3. PLANTEAMIENTOS TEÓRICOS

Se abordó el modelo de Análisis Contrastivo de lenguas, presentado por Fries (1945) y Lado (1957), quienes postulan que las diferencias entre la lengua materna (L1) y la lengua extranjera (LE) que se va a aprender entrañan dificultades particulares; que la primera lengua interfiere en la lengua extranjera y que la distancia que media entre ambas lenguas y la interferencia consecuente son las responsables de los errores producidos en el aprendizaje de una LE.

También el Análisis de Errores de Corder (1967) quien reformula la hipótesis del Análisis Contrastivo y propone estudiar empírica y sistemáticamente la producción del estudiante de LE, mediante la elaboración de un inventario de errores más frecuentes, en el que se valorará la importancia y gravedad de los mismos, con el objetivo de señalar las áreas de dificultad de aprendizaje. Este investigador propone además, clasificar los errores en dos grandes categorías: errores de interferencia o interlingüísticos y errores intralingüísticos, que se originan dentro de la estructura interna de la lengua.

Principalmente se han tomado en consideración los postulados del Análisis de Interlengua formulados por Selinker (1972) quien la describe como sistema lingüístico especial con el que un aprendiente de lenguas extranjeras intenta expresar significados en la lengua que está estudiando. Es visto también como un sistema lingüístico separado de la lengua nativa del aprendiente y de la lengua meta, pero interrelacionado con ambas a través de identificaciones interlinguales desde la percepción del estudiante.

En este caso hacemos una descripción de los errores contenidos en las encuestas desde el punto de vista de la ortografía -acentuación y grafías- (Perera, 2000: 208-209) y de léxico –transferencia- (Weinreich, 1968), -hibridización, invención- (Payrató, 1984: 46) y de estrategias de comunicación, transferencia de entrenamiento, sobregeneralización e hipercorrección (Selinker, 1972) que responden a las características particulares de los datos analizados.

4. CONTEXTO Y RECOGIDA DE DATOS

La población analizada es hispanohablante y pertenece al Centro de Enseñanza de Lenguas Extranjeras de la Universidad Nacional Autónoma de México, donde los cursos de lengua portuguesa se dividen en seis semestres de 100 horas cada uno: cinco horas por semana, durante cinco meses. Los grupos se organizan en seis niveles y cada grupo tiene entre 6 y 30 alumnos. Las encuestas se realizaron este grupo durante el período 2009-1 con estudiantes voluntarios del quinto semestre que en ese momento contaba con 14 alumnos, seis hombres y ocho mujeres de diferentes carreras, de los cuales se pudieron analizar 13 alumnos.

4.1. Criterios de análisis

De acuerdo a varios criterios de los estudios de interlengua antes mencionados, en esta investigación se analizaron los siguientes puntos:

- Transferencia o interferencia de L1: se entiende como la influencia de la lengua madre, en este caso se trata palabras que se escribieron directamente en castellano.
- Transferencia o interferencia de otra LE: cuando se encontraron influencias de alguna otra lengua extranjera, generalmente otra lengua que conocían previamente como inglés, o que están estudiando simultáneamente.
- Sobregeneralización o hipergeneralizaciones: se percibe como un mecanismo intralingual de generalizaciones de reglas, regularizando el sistema sin tener en cuenta las restricciones. Por ejemplo la regularización de verbos irregulares, o la extensión de un paradigma.
- Hibridización: cuando se forman las palabras con elementos de ambas lenguas.
- Analogía es cuando se refleja en el uso de un término o construcción por otro próximo o formal, funcional o semánticamente; pero que no es adecuado en el contexto o en el registro utilizado.
- Hipercorrección: se entiende como el fenómeno contrario a la generalización, o sea hacer de una excepción una regla. O bien, extrapolar una dificultad proveniente del contraste con la lengua materna a casos en que no existe ese contraste.
- Invención: es cuando los alumnos crean sus propios términos.
- Estrategia de comunicación: se refiere a cuando los alumnos han reproducido o adaptado modelos de su lengua nativa la lengua extranjera.

- Transferencia de entrenamiento: también es común que los alumnos reproduzcan los modelos que escucharon de los profesores o que leyeron de los libros de textos. En este rubro se incluyeron algunos fallos provocados por la fonética particular del portugués.
- Omisión de la regla: se da cuando el alumno omite alguna regla; ya sea de acentuación, tildes, de formación de plurales, de colocación de ç, etc.

4.2. Estudio

Los 16 centros de interés son los siguientes: partes de cuerpo, ropa, alimentos, cosas de la escuela, ciudad, campo, transportes, animales, distracciones, profesiones, colores, medios de comunicación, sensaciones, adjetivos, vocabulario especializado de su carrera y verbos.

Centro de interés 1, partes del cuerpo: de 145 palabras, 112 son propias del portugués, aunque también se advierten tres palabras que no tienen relación directa con el cuerpo humano como: *pretas, brancas* y *pequena*. En el caso de partes del cuerpo que son más de una, casi siempre aparecen en plural como en el caso de *dedos.* Del total de palabras encontramos que 13 son interferencia directa de la lengua materna; mientras que en el campo de la hibridización aparece solamente "*cabelho*". En cuanto a omisión de la regla tenemos el caso de "*unas*" y como hipercorrección tenemos "*torço"*.

Centro de interés 2, ropa: tenemos un total de 75 palabras de las cuales 44 pertenecen al léxico del portugués y cinco son palabras del castellano. Existe interferencia de otra lengua extranjera como son las palabras inglesas *short, blazer* y *jeans*, pero que son de uso común en el léxico de México. Tenemos también sobregeneralización en *"blousa"* con el diptongo *ou* común en portugués. La hibridización es relevante por la mezcla de palabras, pues aparece "*traje de banho*" en vez de *fato de banho* y "*ropa esportiva"*. También aparece "*soutier*" hibridización entre *soutiã* y *brassier* (sostén). En el campo de hipercorrección se usa *banheiro* (baño), por *fato de banho* (traje de baño o bañador). La invención desde luego responde a lo que el alumno supone debe ser portugués y así inventa *"pantalão"* por *calça,* "*jaqueta*" por casaco y "*peticó*" que no sabemos qué es, pero que suponemos puede ser *paletó* (chaqueta) y *"raia"* que quizás sea *saia.* En el caso de transferencia de entrenamiento se encontraron fallos por cuestiones de fonética propia del portugués de Brasil como "*cauça comprida*" por *calça comprida*; *"cacherol"* por *cachecol* y "*saja", "minisaja"* por *saia* y *minisaia*, "*frauda*" tal vez por *fralda* o en confusión por *saia* (falda) y finalmente una palabra que causó problemas por fue "*cinzo*" por su parecido con la palabra *cinza* (gris) en vez de *cinto*. Hay problemas con las letras *v* y *b*, y el fallo más común que son las omisiones de acentos.

Centro de interés 3, alimentos: hay en total 130 palabras, de las cuales 89 son palabras portuguesas y cuatro directamente tomadas del castellano. Entre los problemas más sobresalientes se observan seis palabras tomadas de otras lenguas extranjeras como es el caso del italiano "*prochiuto" prosciutto* y la analogía entre dos palabras del portugués muy similares como son *queixo* (mentón) en vez de *queijo* (queso). En el campo de la sobregeneralización está el diptongo *ou* colocado en palabras como *soupa*. En hibridización hay palabras como: "*cacão, cervelha, feijos"* por *cacau*, *cerveja* y *feijão*. En el campo de la hipercorrección tenemos casos como *"cachaza"* y *"doze"* por *cachaça* e *doce*. En el campo de invención tenemos palabras que no fue posible identificar como: "*fechonda", "salsão", "feixas" y "freio"*. Finalmente hay muchos casos de omisión de acentos y de confusión entre las letras *c* y *s* como en "*senoura*" por *cenoura.*

Centro de interés 4, escuela: tenemos un total de 135 palabras de las cuales 100 son del portugués y 14 del castellano. Hay un pequeño grupo de palabras con interferencia por analogía de otras palabras del portugués como son *salão* por *sala de aula*. En el área de invención hay una palabra no identificable como es "*aposionamento*"; mientras que en hibridización están términos como "*calificação"* y "*exitação*" esta última sin relación aparente con el campo de la escuela. En transferencia de entrenamiento tenemos fallos por entrenamiento fonético como "*folias*" y "*folho*".

Centro de interés 5, ciudad: hay un total de 161 palabras, de las cuales 123 son de portugués y únicamente 9 son del español. La ciudad es un campo donde aparecen además ciertos elementos que se viven diariamente en una gran ciudad como la ciudad de México como son: *tráfego, poluição, policía, cão, marginação, corrupução, favelas, malucos,* etc. La interferencia de otra lengua extranjera proviene básicamente del inglés y la hibridización se manifiesta en palabras que siempre han representado un problema para los estudiantes como hibridizaciones de *árvores* en "*arbores"*.

Centro de interés 6, campo: se contabilizaron 138 palabras de las cuales 92 son de portugués, 9 en castellano y el resto se reparte entre hipercorrección como *"cãos" por cães,* y *"plantras"* por *plantas*; sobregeneralización como *"animaes"* por *animais, "galhina"* por *gallina,* etc; hibridización como *"cabalhos"* por *cavalos, "campesino"* por *camponês,* etc. Son relevantes las secuencias donde aparecen asociadas cosas de la naturaleza como: *grama, árvores, terra*, *monte* etc.; animales: *burro cavalo, porco, pássaros, galinhas, vacas,* etc. pero también cosas relacionadas con las actividades del campo: *agricultura, fazendas*, etc., y elementos más subjetivos como *tranquilidade, gentileza*, etc. que contrastan con el léxico

mencionado en ciudad; y desde luego sobresale la analogía en el uso de la palabra *frango* por *pintinho*.

Centro de interés 7, transportes: este centro muestra un total 111 palabras, 79 en portugués, por sólo 5 en castellano. En el rubro de interferencia de otra lengua extranjera, se encuentran palabras del inglés, pero que son de uso común en México como: *bus* y *kayak.* El resto de palabras muestran una tendencia a problemas de acentuación que probablemente se deba a que mucho vocabulario es igual entre ambas lenguas salvo por los acentos, así encontramos palabras como *"helicoptero", "taxi", "ambulancia",* etc. y en hipercorrección aparece en varias ocasiones "*trêm*". Sin embargo, es necesario apuntar que en las encuestas en castellano, los alumnos generalmente omitieron los acentos del castellano también.

Centro de interés 8, animales: este centro tiene un total de 147 palabras, de las cuales 104 son del portugués y 15 del castellano. El resto de palabras pertenecen en su mayoría al rubro de hibridización como son *"ratoncinho, balena, galhinha",* aparece especialmente repetido "*cabalho".* Son interesantes los casos de fallos por la fonética propia del portugués como *"lião, xacaré, lagartixa";* no hay sobregeneralización y sólo se observa un caso de hipercorrección. En cambio hay cuatro problemas con la omisión de acentos y finalmente es interesante la invención donde aparecen animales como *"boto"* y "*salôn*" que no se sabe qué son y "*alebrije*", que es un animal impreciso e imaginario de la cultura popular mexicana.

Centro de interés 9, distracciones: tenemos 108 palabras en total, de las cuales 78 son del portugués y 11 de interferencia L1, o sea del castellano. Aparecen tres casos de inglés, pero que son de uso común en México, como es *scrabble, bingo* y *tennis.* En este campo es interesante el rango de cosas que fueron consideradas distracciones, aparecen muchos verbos son como: *beber, dormir, falar, nadar, correr,* etc., sitios como *cinema, teatro, museus,* etc., aparatos *internet, televisão, videojogos,* etc., y desde luego deportes como *futebol, basquete,* etc. Además las distracciones son muy distintas entre hombres que prefieren actividades como: *beber, futebol, nadar*, etc., de las mujeres que mencionan diversiones como: *dançar, teatro, cinema, passear,* etc.

Centro de interés 10, profesiones: se observaron 152 palabras, de las cuales 102 son del portugués y 9 del castellano. El mayor problema en este campo se observa en la colocación de acentos, que faltaron en 19 palabras. Es de llamar la atención también que los rangos van desde 4 profesiones hasta 22, lo cual representa una gran diferencia; sin embargo, el promedio general es de 12 palabras por alumno. La generalidad son profesiones como *engenheiro, profesor, biólogo,* etc., que corresponden en su mayoría a las carreras de la

misma universidad y un número menor de palabras son oficios como *carpinteiro, leiteiro,* etc. en los que se observan bastantes problemas de omisión con el diptongo *ei*.

Centro de interés 11, colores: tenemos un total de 123 palabras, de las cuales 91 son del portugués y 10 de interferencia de la L1, es decir del castellano. Uno de los colores que tuvieron problemas por analogía fue *marrom* que aparece como *"café"* que en portugués es sólo la bebida y no un color como en México. Como sobregeneralización se observan "*amarelho*", "*amarehlo*" y "*amarello*" por *amarelo*; y "*açul*" por *azul*; como analogía aparece varias veces "*roxo*" en vez de *rouxo* (morado), muy probablemente por su semejanza con rojo y "*nego*" por negro, que es un término coloquial para llamar a una persona de raza negra en Brasil. En transferencia de entrenamiento están las palabras influenciadas por la pronunciación como "*bermelio*"; en invención está la palabra "*anaranja*" y como parte de su creatividad "*arcoíris*".

Centro de interés 12, medios de comunicación: hay 105 palabras, de las cuales 70 pertenecen al léxico del portugués y 9 al del castellano. Aparecen palabras de interferencia de otra lengua extranjera como "*journal*" del francés por *jornal* y el principal problema es el gran número de omisiones de acentos y tildes, lo cual no es extraño, puesto que los omitieron también en la encuesta de castellano. Tenemos varios casos de sobregeneralización como "*televição*" por *televisão* y en los medios de comunicación que incluyeron tenemos además términos como *corporal*, *gesticulação*, *telepática*, *ocular* y otros relacionados indirectamente con la comunicación como *cultura, educação,* etc.

Centro de interés 13, sensaciones: en este rubro hay 134 palabras, de las cuales 99 son portuguesas y sólo ocho del castellano. En hipercorrección tenemos "*trizte*, y *tristeça*" en varias ocasiones; de hibridización con el italiano tenemos "*apassionado*". Realmente el fallo que representa un problema el de la omisión de acentos.

Centro de interés 14, adjetivos: tenemos 160 palabras, 123 del portugués y 15 del español. El mayor problema es la omisión de acentos, con 15 casos; a continuación llama la atención la invención de palabras como "*mâo", abelho"* y "*divertente*" y en hibridización "*fermoso*", "*exquisito*" y "*horribel*".

El centro de interés 15, vocabulario especializado: éste campo es poco representativo pues se pidió a los alumnos escribir vocabulario en portugués relacionado con su estudios. Dado que las carreras son muy disímiles, así es el léxico que anotaron. Sobresale el hecho de que hay alumnos que leen textos escolares en portugués tuvieron un vocabulario extenso, correcto y variado. Mientras que hay otros que no tienen relación alguna con la lengua portuguesa en sus estudios, por lo que dejaron este centro de interés en blanco.

Centro de interés 16, verbos: es el campo con mayor número de palabras y menor número de errores. Tenemos 279 palabras, de las cuales sólo siete están en castellano, cuatro casos de hibridización *"escreber", "começar"* y "*cocinhar*"; un hibrido de francés "*lir*" y finalmente en dos ocasiones aparece la analogía "*trouxer"* y en una ocasión se presenta "*fedor*" que no es verbo, sino sustantivo.

5. Consideraciones Finales

A través de este estudio se hizo evidente que la enseñanza de portugués comporta unas dificultades específicas en función de la lengua materna de los alumnos, la estrecha relación entre castellano y portugués producen fenómenos de interlengua en grados a veces inimaginables gracias a la intercomprensión que posibilita la cercanía entre ambas lenguas.

Transferencia. Hay una tendencia natural a transferir desde la lengua materna la forma y el significado de un vocablo. Cuando una forma en castellano encuentra un correspondiente con la misma forma y significado en portugués (y la probabilidad de que esto ocurra es muy alta), tenemos una transferencia positiva. Sin embargo, también en el proceso de aprendizaje concurren aspectos contrastivos divergentes en ambas lenguas, entonces se producen errores de transferencia o de interferencia. Como consecuencia de ello, muchos errores tienden a fosilizarse.

La ortografía es el problema más grave en este estudio. Se hace evidente la necesidad de prestar especial atención a las pequeñas, pero significativas diferencias que los alumnos pasan por alto. Los errores ortográficos son muchos y repetidos de manera sistemática, lo que nos habla de un problema importante de fosilización; aunque también en los últimos niveles tienen menor incidencia lo que nos habla de un proceso de sistematización de la lengua. También observamos que es un problema que se presenta desde el castellano. Por último existen pocos problemas en cuanto al significado de las palabras.

La hibridización de palabras es otra dificultad destacada que se presenta en todo tipo de palabras. Los alumnos crean palabras que creen portuguesas a partir de elementos que les parece son del portugués o de ambas lenguas y muchas son invenciones.

Al final de esta parte del proyecto se hace una serie de consideraciones didácticas donde se sugiere entre otras cosas un estudio consciente de la lengua, un replanteamiento didáctico que refuerce la corrección, el uso de estrategias orientadas a la reflexión, etc.; así como tomar en consideración las competencias personales.

Benítez, P. (1993). Anglicismos en la disponibilidad léxica de Madrid. *X Congreso Internacional de ALFAL*.

Benedetti, A. (2001). Interferencias semánticas del portugués en el aprendizaje del español. *Forma.* Asencio, J., Sánchez J. y Larrañaga A. (Eds.). *Interferencias, Cruces y Errores* No. 2. Madrid: Sociedad General Española de Librería, S.A. (pp 9-25).

Camarena, E. (2006). *Análisis de Fenómenos de Error, Interlengua e Interferencia en el Portugués Escrito de Estudiantes Castellanohablantes Universitarios.* Trabajo de investigación no publicado para el 2º año de doctorado en Didáctica de la Lengua bajo la dirección del Dr. Joan Perera i Parramón. Universidad de Barcelona, España.

Carcedo, A. (1998). Sobre las pruebas de disponibilidad léxica para estudiantes de español/LE. *RILCE*: *Español como lengua extranjera: investigación y docencia*, 14.2: 205-224.

Caetano, A. y Filho, J. (2005). *Gramática Portuguesa*. Madrid: Espasa Editorial.

Corder, S. (1967). *The Significance of Learners Errors*. IRAL 5 4: 161-170.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Fernández, S. (1997). *Interlengua y Análisis de Errores en el Aprendizaje del Español como Lengua Extranjera.* Madrid: Edelsa, Grupo Didascalia.

Fries, C. (1945). *Teaching and Learning English as a Second Language*. Ann Arbor: University of Michigan Press.

García Megía, A., (2004). *La disponibilidad léxica en la ciudad de Almería entre los nueve y los doce años.* Almería: Servicio de Publicaciones de la Universidad.

Gargallo, I. (1993). *Análisis Contrastivo, Análisis de Errores e Interlengua en el Marco de las Lingüística Contrastiva*. Madrid: Síntesis.

Grosjean, F. et Py, B. (1991). La restructuration d'une première langue: l'intégration de variantes de contacte dans la compétence de migrantes bilingues. *La Linguistique.* Vol. 27. Fasc. 2.

Herrera, L., Camarena, Martínez y Ortiz. (2003). *Informe Sobre la Encuesta de Opinión*. *Semestre 2003-2*. Centro de Enseñanza de Lenguas Extranjera. Dirección General de Evaluación Educativa. México: UNAM.

Kasper G. and Schmidt, R. (1996). Developmental issues in interlanguage pragmatics. *Studies in Second Language Acquisition*, Num.18, 149-169.

Kasper G. and Blum-Kulka, S. (Eds.). (1993). *Interlanguage Pragmatics*, New York: Oxford University Press.

Klein, W. (1986). *Second Language Acquisition.* Cambridge Textbooks in Linguistics: Cambridge: Cambridge University Press.

Lado, R. (1957). *Linguistics Across Cultures.* Ann Arbor: University of Michigan Press.

Larsen-Freeman, D. and Long, M. (1991). *An Introduction to Second Language Acquisition*. Essex: Longman.

Moreno, F. y Maia González (2006). *Diccionario Esencial, Español-Portugués/Português-Espanhol*. Madrid: Arco Libro, S. L.

Nikula. T. (1997). Interlanguage view on hedging. *Hedging and Discourse Approaches to the Analysis of a Pragmatic Phenomenon*. Academic Texts.

Nunan, D. (1999). *Second Language Teaching and Learning*. Boston: Newbury House. Teacher Development.

Nussbaum, L. (1992). Manifestacions del contacte de llengües en la interlocució. *Treballs de Sociolingüística Catalana*. Núm. 10.

Payrató, L. (1984). *La Interferència Lingüística: Comentaris i Exemples Català-Castellà*. Pub. De l'Abadia de Monserrat. Barcelona: Curial.

Perera, J. (2000). *Les Llengües a l'Educació Secundária.* Barcelona: ICE/ Horsori.

Seco, M., Andrés, O. y Ramos, G. (2000). *El Diccionario Abreviado del Español Actual.* Barcelona: Santillana.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, IRAL 10, 3. Reprinted in Richards 1974: 209-231.

Selinker L. (1992). *Rediscovering Interlanguage.* London: Longman.

Silva-Valdivia B. (1994). Cambios de código, alternancias e interferencias lingüísticas: unha perspectiva didáctica sociocomunicativa. *Didáctica da Língua en Situacións de Contacto Lingüístico.* Universidade de Santiago de Compostela.

Tarone, E. (2000). Interlanguage. *Concise Encyclopedia of Educational Linguistics.* Bar-Ilan: Pergamon. (Pp. 507-510).

Vila i Moreno, F. (1998. Bueno,vale ja de criticar, no? Marques transcòdiques lèxiques i variació funcional en català. *Oralmente. Estudis de Variació Funcional.* PAM.

Weinreich, U. (1968). *Languages in Contact.* The Hague: Mouton.

Wigglesworth, G. (2005). Current approaches to researching second language learner processes. *Annual Review of Applied Linguistics*. 25: 98-11. Cambridge University Press.

Zimmerman, K. (2001). Interculturalidad y contacto de lenguas: condiciones de la influencia mutua de las lenguas amerindias con el español. *Lo Propio y lo Ajeno en las Lenguas Austronésicas y Amerindias*. Madrid: Vervuert-Iberoamericana.

Corpus design for the translation of R&D reports

MIGUEL ÁNGEL CANDEL MORA
ALICIA RICART VAYÁ
Universidad Politécnica de Valencia

Abstract
This paper describes the role of the technical translator, the most important criteria for the selection of parallel texts, and the elaboration of a reference corpus and the use of computer tools to help translators in their task of identifying technical terms, collocations and standard text segments used in R&D reports. Electronic corpora allow the translator to apply computer tools and to process information following some parameters; this will permit translators to verify the use of certain terms or phrases with respect to other units in the text; in this way, translators can create their own databases, based on real texts issued by international institutions and bodies and adapt them to computer-assisted translation tools such as translation memories and terminology management applications.
Keywords: translation-oriented terminology, corpus design and tools

Resumen
Este artículo describe el papel del traductor técnico, los criterios más importantes para la selección de textos paralelos, y la elaboración de un corpus de referencia y el uso de herramientas informáticas para ayudar a los traductores en su tarea de identificación de términos técnicos, colocaciones y segmentos de texto habituales en informes de I+D. El corpus electrónico permite al traductor el uso de herramientas informáticas para procesar la información según determinados parámetros, lo que permitirá a los traductores verificar el uso de determinados términos o frases con respecto a otras unidades en el texto, de esta manera, los traductores pueden elaborar sus propias bases de datos, a partir de textos reales procedentes de entidades y organismos internacionales y adaptarlas a las herramientas de traducción asistida por ordenador, como memorias de traducción y aplicaciones de gestión terminológica.
Palabras clave: gestión terminológica, traducción, diseño de corpus y herramientas

1. INTRODUCTION

The society of the 21st century based on knowledge, information and communication technologies, the Internet, etc, demands a rapid dissemination of information and the immediate transfer of research results and technological advances worldwide. This new scenario has affected the field of translation in different ways: on one hand, the development of new computer resources as support tools for translation and the compilation of terminology databases, and on the other hand, a new approach to the role of translators. This fact becomes particularly evident in the field of R&D reports issued by international bodies and organizations, since their results have to be translated and distributed as fast as possible.

The advances in computer tools have favored the appearance of new approaches in corpus linguistics with the computerized elaboration and storage of parallel corpora and with

the development of computer programs specially designed to align original and translated texts, search for concordances, collocations or key words in context.

The efficiency of a translation-oriented terminological management system is affected by the inconsistencies in the technical terminology used in the elaboration of the original documents; therefore, the most appropriate solution would be to establish a quality control system of terminology from the beginning of the development of R&D management reports, specially when the documents of technology transfer have been conceived to be translated to different languages. It is not then a question of promoting Spalink's concept of "writing for translation" (2000) but of ensuring accuracy and coherence in the terminology employed in the original texts.

The communication of research results, in the form of annual reports, is an essential part of the activities of research centres, since there must be some kind of information transfer to other entities, to the sponsoring institutions or as a record of the activities developed. The importance of these activities of dissemination and communication of information become patent when a revision through the basic documents of the R&D managing bodies reveals clearly that the dissemination of results is one of their priority objectives.

2. TRANSLATOR'S PROFILE

The profile of the technical translator depends on variables such as the specialized field of activity, the working conditions, access to information resources, the IT tools used and the individual skills and strategies. However, among the constants, time is one of the most important factors, since it usually limits the possibilities to extrapolate efforts beyond translation.

Thus, considering time constraints as one of the key points influencing professional translation, the organization of the texts as a translation memory or as a textual database greatly compensates the efforts to elaborate corpora and terminology databases.

Other factors that affect the translator's work and productivity are the skills to create, adapt and recycle language resources as well as those functions known as *language mediation* (Hatim and Mason, 1995:298), that suggest new approaches and strategies in the translation process.

According to Hatim and Masson (1995:32), the translator's role depends on the field of activity. These authors distinguish between freelance translators and translators who work for

a company or international institution. To this we can add the role of the translator working for a translation company.

From the translator's point of view, the elaboration of a terminology database is a hard and cumbersome task. However, in the long run it compensates these efforts since coherence in technical terminology is considered as one of the most relevant quality parameters of a translation both in the theoretical and in the professional practice approaches: "the determination, documentation, and effective application of terminology play a pivotal role in document production" (Wright, 2001:490). From the perspective of the profession, the incorporation of a systematic database to the new computer-based translation tools facilitates decision making and guarantees lexical coherence.

During the process of translation of technical texts, the translator generally deals with specialized language not only because of the terms used, which at first sight could be considered as the most complex aspect of the translation process, but also as a consequence of the register and idiomatic expressions used by the author of the original text, which is what actually provides the target text with the same standards and quality as the original text (Mayoral, 1994:73).

The main difficulty for professional translators in their work is the lack of appropriate reference material available in terms of validity, degree of specialization, linguistic information, reliability of the source, or lack of economic resources or availability of standard reference material. The simplest solution is to use original technical texts both in the source and target languages dealing with the same topic and belonging to the same text type. Textual documentation thus becomes a basic tool because one of the weak points of reference material is that it usually does not provide the use of language in context, but only as an isolated representation of the concepts.

Most of the common technical terms can be found in dictionaries and encyclopaedias; however, for the translator it is of great relevance to verify the use of the terms in renowned journals or documents, together with the textual characteristics of the document to translate. In this sense, it becomes evident the lack of updated specialized and reliable reference material. The contributions of the study of languages for special purposes, the identification of complex technical or semi-technical terminology, and the importance of textual analysis suggest the need for the translator to count on parallel texts, i.e., original texts with textual and thematic features similar to those found in the text to be translated but in the target language.

From the perspective of Translation Studies and Contrastive Rhetoric the use of reference material is essential to have access to the strategies and internal mechanisms of the translation process. If a bilingual corpus can also be included in the translation memory for further processing with a computer-assisted translation system, the possibilities for professional translators greatly increase as they will count on easily accessible relevant reference material. To ensure the reliability of this new documental tool it is be necessary to follow the criteria established in Corpus Linguistics Studies during the process of text collection in terms of currency, quality and reliability of the resources.

3. DELIMITATION AND ANALYSIS OF THE R&D FIELD

Due to the wide range of topics dealt with in R&D management reports, it is almost impossible for a translator to be an expert in all the disciplines; therefore, for the translation of such reports it is not feasible to count with the professional services of experts on all the topics. Therefore, documentation work offers many advantages to translators: through collection and analysis of the documents on the topic to be translated, the translator becomes familiar with the topic and the textual features of the document. In addition, a revision of the source documents gives the translator a first approach to the most common technical and general vocabulary used in that particular type of text. The key point of the process is for the translator to gain experience on the specific topic field which allows him to apprehend the elements for a better understanding of the text.

For professional translators, textual analysis is addressed from two different points of view: on the one hand, as the process to decode the original text (Gommlich, 1993:175), and on the other, as the description of the different types of text for a particular use in translation. In both cases, text typology has a specific purpose: “Its main aim is to describe recurring textual structures in a manner that is appropriate to translation and that allows for simple storage and processing in a database to which the translator should have access while preparing a translation” (Gommlich, 1993:176). In addition the analysis allows the translator to determine the common aspects of a type of text and the differences with other types of text.

Regarding text analysis applied to translation, Gommlich (1993:180) establishes three general steps: the first step consists of sorting out the original text into the class of text type it belongs to, in terms of Gommlich *fundamental interactional* aim (Gommlich, 1993:180); the second step is the identification of the potential target reader and of the target reader's needs; and the last step consists of the verification that the intention of the original text matches the

target text, which will be essential for the subsequent decision making during the translation process.

Most of the approaches found in the literature suggest starting the terminological work from the delimitation of a specific field. However, the majority of methods coincide in emphasizing the role played in the purpose or application of the terminological work as a basic criterion when initiating a task of such characteristics (Cabré, 1993:3). Therefore, although the delimitation of the structure of the field is essential during the stage of selections of terms, not less important is the consideration of certain variables mentioned in the recommendations of the POINTER project (1988), such as political and geographical scope of the field (national, regional, international); audience (public, private, academic, research); application (translation, technical writing, library science and documentation, education, marketing, production, etc); specific field (engineering, medicine, computer science, sociology, etc); or register (standardization, descriptive, normative, lexicographic).

R&D covers a broad spectrum of disciplines as it is directly related to the fields of economy, business, industry, science, technology, international policy and many other disciplines that at some stage during the process of dissemination of results are included in the R&D reports (Cornella, 2000:146). For this reason, this work focuses on the R&D management reports addressed to experts and the public.

A significant aspect of R&D management reports is the need to be translated because most belong to European or international projects. For this reason we have considered of interest to carry out an in-depth analysis of the process of communication of research results, with the purpose of identifying the consequences of considering the possibility of integrating the process of parallel translation to the process of creation or, as suggested by other authors (Spalink, 2000:158; Esselink, 2000:3), to make the writer aware of the fact that the text is to be translated and that the message has to be to simplified.

In addition to the aim of communication among professionals in a specific field and the use of a specific terminology in that field, the texts belonging to a discipline have as their final goal to accurately and clearly inform the target reader, who theoretically belongs to that field of speciality and possesses the same level of knowledge as the authors of the text; as a consequence, the target reader should theoretically have no problems in understanding the message. However, in R&D management reports, the documents do not have a single target reader's profile, therefore the target reader does not necessarily need to be an expert in the topic; in addition, because different disciplines –technical and scientific, legal and

administrative, economic, etc. - converge, the problem of specific terminology tends to decrease or at least to be balanced between the knowledge of expert and non-expert users.

4. CORPUS DESIGN

Parallel texts are half way between text typology and Corpus Linguistics applied to Terminology. On one hand, they meet all the criteria that contribute to describe the features of the original text to be transferred to the target text during the translation process. On the other hand, they define the scope of the search for suitable texts and provide the translator with a valuable reference database for further management of terminological data. Therefore, the higher the specificity of the source text, the more necessary it will be for the translator to count with a reference corpus to verify language usage by the experts in the topic.

For an efficient exploitation of these parallel texts, the use of electronic texts helps to perform a detailed analysis of textual features and typology. Although not specifically designed for translation, Zampolli (1991:186) suggests the elaboration of textual databanks, defined as a collection of homogeneous texts used as a textual source for a specific task or for a particular language subset.

The development of a bilingual corpus becomes a useful tool for the translator as it not only provides updated reference material but it also includes information not usually found in conventional reference materials, like dictionaries and encyclopaedias.

The first stage in corpus development consists in identifying the topic or disciplines of the text to be translated in order to determine the working field. At this stage, the translator has the first contact with the linguistic and extra-linguistic features of the text to translate, such as author's style, target audience, terminology or text format, among others.

The second stage of the documentation work consists of the dynamization of the resources for the collection and further analysis of the reference material. For this end, in addition to the international or national standards published on the topic, the most common searching method consists of locating the material in libraries or in the Internet using specific and advanced search strategies and criteria. Cabré (1999:19) establishes four types of data sources useful for translators: grammar sources, among which she mentions grammar books and style manuals; lexicographic sources, consisting of general dictionaries; terminological sources, with information about the speciality field; and specialized sources, such as standards, manuals and handbooks that provide specific information about a particular field of knowledge.

After the collection of the reference material, the third stage in the organization of the documents and corpus design consists of the classification and selection of the material according to certain specific criteria. At this stage, the texts collected were classified into four different corpora: two *Monolingual* corpora, one with the English texts and another one with the Spanish texts; one *Parallel* corpus, with aligned segments of the texts in both languages obtained from translated texts or from texts published in both languages; and finally, a fourth corpus, called *Reference* corpus, which included other source texts that did not meet the requirements to be included in one of the other types of corpora but which contained relevant information for the terminological work. For example, common reference documents in R&D reports like Frascati's Manual or Oslo Manual, which are not specifically reports on R&D management, were included in the reference corpus, as they were useful for the final task of terminology validation, because they are key documents that established the basis for the management of research activities.

The next step is the development of the electronic corpus for further data processing using computer tools. Simultaneously to corpus design, the different documents are identified using codes that indicate the date, language and source institution or author of the document. This code, labelled in the terminological file entry as *source,* will serve to identify the authority, reliability and validity of the term.

For the research on R&D management reports, the main source was found in the original document itself, and in the bibliographic references used by the author, documents which range from general handbooks to specialized research papers. This initial framework served to classify the texts. In other cases, the list of contents of a R&D report also helped to establish a general outline of the field of study.

Secondly, the references most frequently quoted in the documents of the corpus were consulted, and the table of contents or the glossaries found at the end of some of the documents were used to complete the database.

From the perspective of Library Science and documentation studies, Cordón *et al.* (1999:39) propose three steps as starting point for bibliographical search: the first stage consists of defining the objectives and application of the search; synthesizing our knowledge about the issue; highlighting those aspects we may be particularly interested in and rejecting non-relevant aspects; and indicating any relationship of the topic with other scientific fields which may help to differentiate the text and avoid confusion and errors. The second stage consists of defining the search criteria, e.g. date, language, type of document. Finally, the

third stage consists of formulating the search strategy, from the identification of a number of concepts to the definition of the search tool.

Corpus design started with visits to the websites of the most important research institutions with texts in Spanish and/or English and an exhaustive search in R&D documents (see Table 1).

Table 1: Text sources.

Language	**Institution**
EN	Committee for Economic Development (CED)
EN	National Science Foundation (NSF)
ES/EN	Fundación Cotec (COTEC)
ES/EN	Oficina de Ciencia y Tecnología (OCYT)
EN/ES	Observatorio de Prospectiva Tecnológica Industrial (OPTI)
ES/EN	Centro para el Desarrollo Tecnológico Industrial (CDTI)
ES/EN	Oficina de Ciencia y Tecnología (OCYT)
ES/EN	Madrid+d
ES	Federación Española de Entidades de Innovación y Tecnología
ES/EN	Institute for Prospective Technological Studies (IPTS)
EN/ES	Centro Común de Investigación
ES/EN	CORDIS
EN	United Nations Industrial Development Organization
EN	UK Research Office (UKRO)

In many occasions, the problem lies in deciding which factors confer the expected quality to the texts selected, i.e., which elements contribute to the decision-making process during the stage of corpus design.

Therefore, after the identification of the main information sources, we strictly applied a number of criteria for the selection of texts and inclusion in the corpus. These criteria can be summarized as follows: to be representative of the subject matter, updated, sufficiently explicit, and with proven competence of the author.

The final criterion for text selection was availability of the document in electronic format. This greatly simplifies the subsequent phase of text processing with computer tools for the development of the database. All the electronic publications selected had their printed version, which is a quality indicator of the material used.

In a first stage, 357 documents were collected from the Internet, which after selection and critical analysis were reduced to 139 documents, of which 58 were in Spanish and 81 in English. Some of the reasons for rejecting the documents were technical as certain documents could not be processed electronically because they were published as image; in other cases, the documents were protected with a password; an important reason was adequacy to the

topic, because after a fast revision it was observed that the topic suggested in the text title or abstract did not correspond with the scope of this research work or the selection criteria mentioned above. However, this group of texts was stored and included in the reference corpus.

Depending on the translator's needs and the translation project specified in this study, after text selection, the documents were classified into four different corpora:

- Spanish monolingual corpus
- English monolingual corpus
- English-Spanish aligned corpus
- Reference corpus

All the documents were classified in files and identified with the most relevant data for further processing. The bibliographical data were classified into two classes: administrative data, for corpus processing management purposes; and bibliographical data, with information about the document.

The selection of the tools and strategies for corpus processing was based on one hand on the results expected; for that end it was necessary to know the frequency of terms in English and Spanish, and the location of the terminological and idiomatic data in a fast and reliable way. On the other hand, as we were dealing with translation-oriented corpus design management, the tool outputs should be capable of being integrated within the translation process, that is, in a translation memory. The tool selected for the design and processing of the parallel corpus was TRADOS' WinAlign, since this software is available with the same set of applications as the translation memory and processing of exiting translations and further inclusion in the translation memory.

5. CONCLUSION

The delimitation of the conceptual scope and the careful selection of keywords becomes one of the most essential steps in corpus design, as this will determine the efficiency of the search for reference material and will simplify the analysis of the collected texts, which in turn results in a better efficiency of the translation process. The texts collected are used for the design of a bilingual corpus based on strict selection criteria, which provides a first approach for the extraction of terms and terminological data that will serve as starting point for the design of the terminology database

The corpus contributes to accurately identify technical terms, collocations, and standard text segments commonly used in the target language.

The terminological work is the only step in the process which demands total attention from the translator, and requires to interrupt the translation process momentarily. However, the translator's efforts to design a reliable database becomes compensated when he/she incorporates the database to the computer-assisted translation process, since it will help to solve most of the terminology problems.

We may conclude that the need for the systematization of translation-oriented terminological work becomes evident in view of the high number of inconsistencies found in the texts during the process of term validation.

The exhaustive analysis of the bilingual corpus for the identification of the inconsistencies about the use of the terms goes beyond the scope of this work, but opens a line of research for future works on translation-oriented terminology.

REFERENCES

Cabré, M. T (1993). *La terminología. Teoría, metodología, aplicaciones*. Barcelona: Antártida/Empúries.

Cabré, M. T. (1999). Fuentes de información terminológica para el traductor. In M. Pinto and J. Cordón (ed.). *Técnicas documentales aplicadas a la traducción.* Pp. 19-39. Madrid: Síntesis.

Cordón García, J., López Lucas, J. and Vaquero Pulido, J. R. (1999). *Manual de búsqueda documental y práctica bibliográfica*. Madrid: Pirámide.

Cornella, A. (2000). *Infonomia! Com. La empresa es información*. Bilbao: Deusto.

Esselink, B. (2000). *A Practical Guide to Localization.* Philadelphia: John Benjamins Publishing.

Gommlich, K. (1993). Text Typology and Translation-Oriented Text Analysis. In S. E. Wright and L. Wright (eds). *Scientific and Technical Translation American Translators Association Scholarly Monograph Series. Volume VI. Pp.* 175-184. Amsterdam: John Benjamins Publishing.

Hatim, B. y Mason, I. (1995). *Teoría de la traducción: Una aproximación al discurso.* Barcelona: Ariel.

Mayoral, R. (1994). La explicitación de información en la traducción intercultural. In A. Hurtado Albir (ed). *Estudis sobre la traducció. Pp.* 73-96. Castellón: Publicacions de la Universitat Jaume I.

POINTER Project. (1998). Proposals for an Operational Infrastructure for Terminology in Europe. Downloaded June 2000, from http://www.computing.surrey.ac.uk/ai/pointer /app/index.html.

Spalink, K. (2000). Improving Cost-Effectiveness in the Documentation Development through Integrated Translation. In P. J. Hager, and H.J. Schreiber (eds.). *Managing Global Communication in Science and Technology. Pp.* 153-178. New York: John Wiley & Sons, Inc.

Wright, S. E. and Budin, G. (eds). (2001). *Handbook of Terminology Management, Volume 2. Application-Oriented Terminology Management*. Amsterdam: John Benjamins Publishing.

Zampolli, A. (1991). Hacia bases multifuncionales de datos léxicos. In J. Vidal Beneyto. *Las industrias de la lengua*. Pp. 185-202. Madrid: Pirámide.

Optimizing readability indexes: an experiment on reading ease in English FL textbooks

PASCUAL CANTOS GÓMEZ
ÁNGELA ALMELA SÁNCHEZ-LAFUENTE
Universidad de Murcia

Abstract
From their creation in the 1950s, readability indexes have been widely used in order to measure textual difficulty. In the current paper, we examine the accuracy of the six most commonly used measures, namely Flesch Reading Ease Score, Flesch–Kincaid Grade Level, Gunning Fog, Automated Readability Index, SMOG, and Coleman-Liau Index. These measures have been applied to a number of English FL-texts, already graded into elementary, pre-intermediate, intermediate and upper-intermediate, according to the textbooks they appear in. Subsequently, by means of the data obtained, we have come up with a new optimized measure which attempts to bring considerable benefit to the area of language teaching.
Keywords: readability indexes, textual difficulty, FL English, language teaching

Resumen
Desde su creación en los años 50, los índices de legibilidad han sido usados con frecuencia para determinar la dificultad textual. En el presente trabajo estudia la precisión de medidas más usadas en la actualidad: la escala de legibilidad de Flesch, el índice de nivel de Flesch-Kincaid, el índice Gunning Fog, el índice de legibilidad automática, SMOG y el índice Coleman-Liau. Se han aplicado estas medidas a una selección de textos de inglés como lengua extranjera, previamente clasificados como pertenecientes al nivel elemental, pre-intermedio, intermedio y pre-avanzado, de acuerdo con los libros de texto en los que aparecen. Con los datos obtenidos, hemos diseñado una nueva medida optimizada que pretende contribuir especialmente al área de la enseñanza de lenguas.
Palabras clave: índices de legibilidad, dificultad textual, inglés como lengua extranjera, enseñanza de lenguas

1. INTRODUCTION

Readability indexes have been widely used in order to measure textual difficulty. In the 1950s, they became increasingly popular, and researchers in the field devoted great effort to devising a substantial number of new formulae. As this study is not intended to provide an extensive review of all the readability formulae, we will offer a brief overview and description of the most commonly used readability indexes.

First of all, the Flesch/Flesch–Kincaid readability tests include two indexes: the *Flesch Reading Easiness Score* and the *Flesch–Kincaid Grade Level*. The first system was devised by Rudolf Flesch in 1948. After several attempts at simplification (Farr, Jenkins & Paterson, 1951; Kincaid, *Fishburne, Rogers & Chissom,* 1975), this is the resulting formula:

$$FRE = 206.835 - 1.015 * \left(\frac{total\_words}{total\_sentences} \right) - 84.6 * \left(\frac{total\_syllables}{total\_words} \right)$$

In 1976, a revision commissioned by the U.S. Navy ended up in a modification of this index to generate a grade-level score, enabling the translation from the 0–100 score to a U.S. grade level. Nowadays, the ensuing formula is known as the *Flesch–Kincaid Grade Level*:

$$F\_KGL = 0.39 * \left(\frac{total\_words}{total\_sentences} \right) + 11.8 * \left(\frac{total\_syllables}{total\_words} \right) - 15.59$$

As can be seen, it uses the same core measures as the Reading Easiness test, namely word length and sentence length. However, the weighting factors are different, and the results of the two tests hence correlate inversely. In this way, a text with a relatively high score on the former test normally achieves a lower score on the latter.

A further index whose score corresponds to U.S. grade level is the *Gunning Fog Index*, or simply *Fog Index*. It was developed by Robert Gunning (1952), becoming particularly popular owing to its easy calculation without a calculator (DuBay, 2004). *GFI* gets its index from mean sentence length (in words) and average number of complex words (words with three and more syllables):

$$GFI = 0.4 * \left(\frac{words}{sentence} \right) + 100 * \left(\frac{complex\_words}{words} \right)$$

Subsequently, *Automated Readability Index (ARI)* was worked out by Smith and Senter (1967) for the U.S. Army, and its validity on technical materials was proved by Smith and Kincaid (1970). The formula uses mean word length (in characters) and mean sentence length (in words):

$$ARI = 4.71 * \left(\frac{characters}{word} \right) + 0.5 * \left(\frac{words}{sentences} \right) - 21.43$$

In 1969, G. H. McLaughlin published *SMOG (Simple Measure of Gobbledygook)* in an attempt to make *Gunning Fog Index* calculation even easier. Indeed, in his work the author describes it as “laughably simple” (McLaughlin, 1969: 639). It is based upon the conviction that the word length and sentence length are to be multiplied rather than added. The formula used at present is the following one:

$$SMOG\_grade = 1.043\sqrt{30 \times \frac{number\_of\_polysyllables}{number\_of\_sentences}} + 3.1291$$

where polysyllable count refers to the number of words of more than two syllables. The resulting score corresponds to the years of education needed to thoroughly understand a piece of writing.

Finally, the *Coleman-Liau Index* was devised by M. Coleman and T. L. Liau (1975). Like the *ARI*, this measure relies on characters instead of syllables per word, which, as commented on above, is not the trend in readability indexes. A further point of similarity between the *ARI* and the *CLI* which is also shared by the Flesch–Kincaid readability tests and the *GFI* is that the ensuing score stands for U.S. grade level. The *CLI* is calculated with the following formula:

$$C\_LI = 5.89 * \left(\frac{characters}{words}\right) + 29.5 * \left(\frac{sentences}{words}\right) - 15.8$$

It is certainly true that the limitations of these indexes have provoked much discussion and debate, and that in the last decades of the 20th century there were serious criticism on their extensive use in areas such as law, journalism or health care. Some representative instances of this scholarly controversy are Maxwell (1978) and Connaster (1999), who offered some reasonable alternatives to readability indexes like usability testing. Nevertheless, as DuBay puts it, "although the alternatives are useful and even necessary, they fail to do what the formulas do: provide an objective prediction of text difficulty" (2004: 3).

2. Research Goal

The aim of this investigation is twofold: first, examining the accuracy of six of the most commonly used readability indexes: Flesch Reading Ease, Flesch–Kincaid Grade Level, Gunning Fog, Automated Readability Index, SMOG, and Coleman-Liau; and second, by means of the data obtained, trying to come up with a new optimized measure.

3. Methodology

3.1. Tasks and procedures

In order to test the accuracy of the six readability indexes mentioned above, we shall calculate their indexes on 20 already graded texts, five texts for each linguistic level from the

coursebook series Innovations (Dellar & Walkley, 2005a; Dellar & Walkley, 2005b; *Dellar, Walkley & Hocking, 2004; Dellar, Hocking & Walkley, 2004*). For each linguistic level (elementary, pre-intermediate, intermediate and upper-intermediate), 5 text samples were randomly taken.

3.2. Data analysis

The preliminary data on the texts are given in Table 1 below. Note that the order of the text samples in Table 1 corresponds to its order of appearance in the various text books. Therefore, we might assume that text sample 1 (elementary, sample 1, LL-code 1) is easier to read than text sample 2 (elementary, sample 2, LL-code 1), and so on.

Table 2: Presentation of the most representative topics and categories of the proposed classification

Linguistic Level (LL)	**Sample**	**LL-Code**	**Tokens**	**Characters**	**Sentences**	**Syllables**	**Complex words**
Elementary	1	1	289	1161	37	312	5
Elementary	2	1	278	1112	25	330	7
Elementary	3	1	322	1426	29	399	12
Elementary	4	1	233	1014	41	270	4
Elementary	5	1	268	1104	21	320	3
Pre-intermediate	1	2	306	1174	24	347	12
Pre-intermediate	2	2	564	2089	43	608	6
Pre-intermediate	3	2	444	1772	27	482	8
Pre-intermediate	4	2	608	2453	44	676	15
Pre-intermediate	5	2	661	3062	44	854	32
Intermediate	1	3	543	2249	39	631	16
Intermediate	2	3	648	2771	41	773	16
Intermediate	3	3	291	1196	19	347	5
Intermediate	4	3	606	2548	29	755	28
Intermediate	5	3	506	2267	28	653	17
Upper-intermediate	1	4	408	1930	29	561	25
Upper-intermediate	2	4	383	1774	30	508	22
Upper-intermediate	3	4	596	2661	32	746	27
Upper-intermediate	4	4	555	2665	26	744	34
Upper-intermediate	5	4	564	2695	23	782	28

We shall first calculate all readability indexes for each text. Subsequently, an initial correlation analysis (*Pearson Correlation Test)* shall be performed so as to determine major similarities and divergences among them. Furthermore, in order to find out possible deviations between the textbook placing of the texts and the readability indexes, the mean divergences (*MD*) of all texts shall be calculated.

4. RESULTS AND DISCUSSION

4.1. Examining the accuracy of the readability indexes

First, all readability indexes for each text were calculated (see Table 2).

Table 2: Readability indexes

Linguistic Level (LL)	**Sample**	**ARI**	**C-LI**	**FRE**	**F-KGL**	**GFI**	**SMOG**
Elementary	1	2,1780	4,0851	107,5742	0,1953	3,8164	5,2291
Elementary	2	4,0820	5,1071	95,1237	2,7540	5,4552	6,1520
Elementary	3	6,0906	7,6275	90,7346	3,3621	5,9321	6,8039
Elementary	4	2,4774	4,6419	103,0325	0,3002	2,9599	4,9135
Elementary	5	5,6295	6,1517	92,8667	3,4767	5,5525	5,2883
Pre-intermediate	1	4,2904	4,4839	97,9585	2,7635	6,6686	7,1686
Pre-intermediate	2	3,8851	3,7669	102,3220	2,2459	5,6720	5,2631
Pre-intermediate	3	7,2342	5,9130	98,3033	3,6332	7,2985	6,2387
Pre-intermediate	4	5,8636	5,8286	98,7477	2,9188	6,5141	6,4646
Pre-intermediate	5	9,4021	9,5210	82,2853	5,5142	7,9456	8,0009
Intermediate	1	6,4317	6,4764	94,3926	3,5523	6,7479	6,7882
Intermediate	2	8,1940	7,5205	89,8736	4,6501	7,3096	6,6978
Intermediate	3	7,1174	6,4816	90,4091	4,4539	6,8136	6,0597
Intermediate	4	10,9117	7,5535	80,2240	7,2610	10,2068	8,7425
Intermediate	5	10,5148	8,9562	79,3150	6,6859	8,5724	7,5804
Upper-intermediate	1	9,2915	9,9652	76,2300	6,1219	8,0786	8,4333
Upper-intermediate	2	8,0460	9,1709	81,6659	5,0402	7,4043	8,0212
Upper-intermediate	3	10,7740	8,9136	82,0387	6,4435	9,2621	8,3766
Upper-intermediate	4	13,9942	11,1006	71,7589	8,5534	10,9889	9,6619
Upper-intermediate	5	15,7892	11,1416	64,6454	10,3345	11,7945	9,4323

An initial correlation analysis (*Pearson Correlation Test*; Table 3) reveals strong similarities between ARI and F-KGL (0.987), and F-KGL and GFI (0.9721). These similarities contrast with the divergences found between C-LI and GFI (0.8224). After that, we ordered the texts according to their reading ease in the textbook (see column *Textbook*) and according to the readability indexes obtained (columns *ARI* to *SMOG*; Table 4).

Table 3: Readability index correlations (*Pearson Correlation Test*)

	C-LI	**FRE**	**F-KGL**	**GFI**	**SMOG**
ARI	0.9132	-0.9469	0.9870	0.9658	0.9023
C-LI		-0.9547	0.8962	0.8224	0.8820
FRE			-0.9655	-0.8945	-0.9119
F-KGL				0.9721	0.9083
GFI					0.9236

Table 4: Text ordered according to readability ease: textbook and readability indexes

Linguistic Level	Sample	LL-Code	Text book	ARI	C-LI	FRE	F-KGL	GFI	SMOG
Elementary	1	1	1	1	2	1	1	2	2
Elementary	2	1	2	4	5	7	4	6	6
Elementary	3	1	3	8	13	10	7	11	11
Elementary	4	1	4	2	4	2	2	1	1
Elementary	5	1	5	6	8	9	8	4	4
Pre-interm.	1	2	6	5	3	6	5	12	12
Pre-interm.	2	2	7	3	1	3	3	3	3
Pre-interm.	3	2	8	11	7	5	10	7	7
Pre-interm.	4	2	9	7	6	4	6	8	8
Pre-interm.	5	2	10	15	17	13	14	14	14
Intermed.	1	3	11	9	9	8	9	10	10
Intermed.	2	3	12	13	11	12	12	9	9
Intermed.	3	3	13	10	10	11	11	5	5
Intermed.	4	3	14	18	12	16	18	18	18
Intermed.e	5	3	15	16	15	17	17	13	13
Up.-interm.	1	4	16	14	18	18	15	17	17
Up.-interm.	2	4	17	12	16	15	13	15	15
Up.-interm.	3	4	18	17	14	14	16	16	16
Up.-interm.	4	4	19	19	19	19	19	20	20
Up.-interm.	5	4	20	20	20	20	20	19	19

A brief examination of the data above (Table 4) reveals that textbook sample 1 is typified by the readability indexes as the easiest one to read (*ARI, FRE* and *F-KGL*) or the second easiest one (*C-LI, GFI* and *SMOG*). In contrast, textbook sample 3 is considered by the readability indexes as being the 8^{th}, 13^{th}, 10^{th}, 7^{th}, 6^{th} or 11^{th} most difficult text to read. This is a striking case, as this text seems clearly too difficult to read, and it is placed at the very beginning of the elementary textbook. According to the indexes, this text should not have been placed in the elementary book, but in a more advanced level: pre-intermediate (*ARI, FRE, F-KGL* and *GFI*) or even intermediate (*C-LI* and *SMOG*).

In order to determine the divergences between the textbook placing of the texts and the readability indexes, we have calculated the mean divergences (*MD*) of all texts:

$$MD = \frac{\sum RI}{6} - TB$$

where $\sum RI$ is the sum of all ordinal transformations of the readability indexes (*ARI+C-LI+FRE+F-KGL+GFI+SMOG*) and *TB* the ordinal textbook placing of a text (Table 5). Positive *MD*s indicate that, on average, the text might be too difficult to read for learners at

that stage, whereas negative values might be a sign of excessive reading ease. In addition, MDs higher/lower than 2 might be considered as *abnormal*[1], as they behave significantly different to what might be considered as *normal*.

Table 5: *MD*s indexes

Text book	ARI	C-LI	FRE	F-KGL	GFI	SMOG	Text book	Index Mean	MD
1	1	2	1	1	2	2	1	1,50	0,50
2	4	5	7	4	6	6	2	5,33	3,33
3	8	13	10	7	11	11	3	10,00	7,00
4	2	4	2	2	1	1	4	2,00	-2,00
5	6	8	9	8	4	4	5	6,50	1,50
6	5	3	6	5	12	12	6	7,17	1,17
7	3	1	3	3	3	3	7	2,67	-4,33
8	11	7	5	10	7	7	8	7,83	-0,17
9	7	6	4	6	8	8	9	6,50	-2,50
10	15	17	13	14	14	14	10	14,50	4,50
11	9	9	8	9	10	10	11	9,17	-1,83
12	13	11	12	12	9	9	12	11,00	-1,00
13	10	10	11	11	5	5	13	8,67	-4,33
14	18	12	16	18	18	18	14	16,67	2,67
15	16	15	17	17	13	13	15	15,17	0,17
16	14	18	18	15	17	17	16	16,50	0,50
17	12	16	15	13	15	15	17	14,33	-2,67
18	17	14	14	16	16	16	18	15,50	-2,50
19	19	19	19	19	20	20	19	19,33	0,33
20	20	20	20	20	19	19	20	19,67	-0,33

According to the MDs, we find four texts which are presented to students that have not already reached the required linguistic proficiency to read them:

- Text 3; MD: 7.00
- Text 10; MD: 4.50
- Text 2; MD: 3.33
- Text 14; MD: 2.67

Similarly, also some linguistically less demanding texts are offered to the students:

- Text 4; MD: -2.00
- Text 11; MD: -1.83
- Text 18; MD: -2.50

[1] It is a *z-score* like transformation; note that the mean of all *MD*s is zero.

- Text 9; MD: -2.50
- Text 17; MD: -2.67
- Text 13; MD: -4.33
- Text 7; MD: -4.33

Data also reveal that some texts seem to have been improperly placed, as their linguistic demand is higher/lower for the textbook they appear in:

- Text 3 – *elementary*; should be *pre-intermediate*
- Text 14 – *intermediate*; should be *upper-intermediate*
- Text 15 – *intermediate*; should be *upper-intermediate*
- Text 11 – *intermediate*; should be *pre-intermediate*
- Text 17 – *upper-intermediate*; should be *intermediate*
- Text 13 – *intermediate*; should be *pre-intermediate*
- Text 7 – *pre-intermediate*; should be *elementary*

In order to determine the accuracy of the readability indexes, we shall first order the texts according to their *Index Means* (IMs) and re-typify them as being elementary (IM ≤5), pre-intermediate (IM ≥ 5 and ≤ 10), intermediate (IM ≥ 10 and ≤ 15) and upper-intermediate (IM ≥ 15). The re-typification (*New LL-Code*) is given in Table 6. A correlation test (*Spearman Correlation Test*) evidences that the readability indexes that best fit with the new linguistic level coding (*New LL-Code*) are *ARI*, *F-KGL* and *GFI* (Table 7). *SMOG* and *C-LI* are the least precise ones, though their correlation values are highly significant.

Table 6: Texts re-typified according to *IM*s

Text book	Index Mean	Linguistic Level	New LL-Code
1	1,5	Elementary	1
4	2	Elementary	1
7	2,67	Elementary	1
2	5,33	Pre-interm.	2
3	10	Pre-interm.	2
5	6,5	Pre-interm.	2
6	7,17	Pre-interm.	2
8	7,83	Pre-interm.	2
9	6,5	Pre-interm.	2
11	9,17	Pre-interm.	2
13	8,67	Pre-interm.	2
10	14,5	Intermed.	3
12	11	Intermed.	3
17	14,33	Up.-interm.	3
14	16,67	Up.-interm.	4
15	15,17	Up.-interm.	4
16	16,5	Up.-interm.	4
18	15,5	Up.-interm.	4
19	19,33	Up.-interm.	4
20	19,67	Up.-interm.	4

Table 7: Correlation analysis of *New LL-Code* with readability indexes

	ARI	C-LI	FRE	F-KGL	GFI	SMOG
New LL-Code	0.941	0.869	0.897	0.941	0.941	0.857

Regarding wrong linguistic level assignment, *ARI* and *F-KGL* accounted for five errors, *C-LI* for six errors, although text 13 was two-level wrongly assigned to *upper-intermediate* instead of *pre-intermediate* (see Table 8); *GFI* for seven errors; *SMOG* for ten errors (and a two-level wrong assignment); and *FRE* for eleven errors.

Surprisingly enough, the three indexes that best adjust to the *New LL-Code* use different measures. As commented on above, *ARI* uses mean word length and mean sentence length, and to obtain the *F-KGL* index, we need mean sentence length and mean syllable per word. On the contrary, *GFI* gets its index from mean sentence length and average number of complex words. In this way, the calculation of the *ARI* and of *CLI* is straightforward; some easy text processing by means of any standard concordance program will output the information required to calculate this index (i.e. *WordSmith*[2]). Nonetheless, *F-KGL* and *GFI*

[2] http://www.lexically.net/wordsmith

are more demanding as we need reliable software syllable counting (i.e. *WordCalc*[3] or *Syllable Counter*[4]). These applications are less consistent and result data might vary significantly.

Table 8: LL-assignment errors

Text book	**New LL-Code**	**ARI**	**C-LI**	**FRE**	**F-KGL**	**GFI**	**SMOG**
1	1	Correct	Correct	Correct	Correct	Correct	Correct
2	1	Correct	Correct	Correct	Correct	Correct	Correct
3	2	Incorrect (1)	Incorrect (1)	Incorrect (1)	Incorrect (1)	Incorrect (1)	Incorrect (1)
4	1	Correct	Correct	Incorrect (1)	Correct	Correct	Incorrect (1)
5	2	Correct	Correct	Incorrect (1)	Correct	Correct	Correct
6	2	Correct	Correct	Correct	Correct	Incorrect (1)	Incorrect (1)
7	1	Correct	Correct	Incorrect (1)	Correct	Incorrect (1)	Incorrect (2)
8	2	Incorrect (1)	Correct	Incorrect (1)	Correct	Incorrect (1)	Correct
9	2	Correct	Correct	Correct	Correct	Correct	Correct
10	3	Incorrect (1)	Correct	Incorrect (1)	Incorrect (1)	Incorrect (1)	Correct
11	2	Correct	Correct	Incorrect (1)	Incorrect (1)	Correct	Incorrect (1)
12	3	Correct	Correct	Correct	Correct	Correct	Incorrect (1)
13	2	Incorrect (1)	Incorrect (2)	Incorrect (1)	Incorrect (1)	Incorrect (1)	Incorrect (1)
14	4	Incorrect (1)	Correct	Incorrect (1)	Incorrect (1)	Incorrect (1)	Incorrect (1)
15	4	Correct	Incorrect (1)	Correct	Correct	Correct	Incorrect (1)
16	4	Correct	Incorrect (1)	Incorrect (1)	Correct	Correct	Correct
17	3	Correct	Incorrect (1)	Incorrect (1)	Correct	Correct	Incorrect (1)
18	4	Correct	Incorrect (1)	Correct	Correct	Correct	Correct
19	4	Correct	Correct	Correct	Correct	Correct	Correct
20	4	Correct	Correct	Correct	Correct	Correct	Correct
Total errors		*5 (5)*	*6 (7)*	*11 (11)*	*5 (5)*	*7 (7)*	*10 (11)*

Regarding complex word count (words with three and more syllables), we performed some preliminary experimenting and evidenced that 95% of all English words with 8 or more characters do entail at least three syllables; this is the measure which have been used to calculate the GFI index.

4.2. Modelling a new index

To attempt the modelling of a new readability index able to classify text samples according to reading ease, we shall take:

- The data on the various texts analyzed (Table 1), entailing all the distinct measures required by the individual readability indexes examined, and
- The *New LL-Code*, as this is a sort of average measure of all individual readability indexes we have considered.

[3] http://www.wordcalc.com
[4] http://*www.wordscount.info/hw/syllable.jsp*

We shall try to model an index by means of *Discriminant Function Analysis* (*DFA*, hereafter). *DFA* is concerned with the problem of assigning individuals, for whom several variables have been measured, to certain groups that have already been identified in the sample. It is used to determine the variables that discriminate between two or more naturally occurring groups. Thus, our aim is not just to measure and model reading ease, but also to look at the dataset that best describes it.

The *DFA*, using all variables (tokens, characters, sentences, syllables and complex words) outputs very promising results: only two errors. One elementary text has been assigned to pre-intermediate (text 2) and an upper-intermediate one has been classified as an intermediate one (text 15). This gives an overall precision of 90% compared to the best precision scores of two readability indexes above (*ARI* and *F-KGL*) of 75%.

Table 9: Preliminary *DFA*

New LL-Code			**Predicted Group Membership**				**Total**
			Elementary	Pre-Intermed.	Intermediate	Upper-Intermed.	
Original	Count	Elementary	3	1	0	0	4
		Pre-Interm.	0	7	0	0	7
		Intermediate	0	0	3	0	3
		Upper-Interm.	0	0	1	5	6
	%	Elementary	75.0	25.0	0.0	0.0	100.0
		Pre-Interm.	0.0	100.0	0.0	0.0	100.0
		Intermediate	0.0	0.0	100.0	0.0	100.0
		Upper-Interm.	0.0	0.0	16.7	83.3	100.0

A further use of DFA is that, if it has turned out to be positive, it is possible to generate a predictive discriminant model to classify new cases. By means of the *Fisher Coefficients*, we are given a table (Table 10) with a constant value and a number of coefficients for each of the variables (tokens, characters, sentences, syllables and complex words) with reference to each readability-ease level.

Table 10: Fisher Coefficients

	Readability-ease level			
	Elementary	Pre-intermed.	Intermediate	Up.-intermed.
Tokens	-0.149	-0.116	-0.266	-0.261
Characters	-0.068	-0.043	-0.055	-0.059
Sentences	1.361	0.795	1.019	0.638
Syllables	0.356	0.260	0.442	0.464
Complex words	-0.434	-0.207	-0.186	0.009
(Constant)	-21.862	-13.091	-31.817	-31.438

This gives four equations, one for each readability-ease level:

- *Elementary = –21.862 + (–0.149*Tokens) + (–0.068*Characters) + (1.361*Sentences) + (0.356*Syllables) + (–0.434*Complex-words)*
- *Pre-intermediate = –13.091 + (–0.116*Tokens) + (–0.043*Characters) + (0.795*Sentences) + (0.260*Syllables) + (–0.207*Complex-words)*
- *Intermediate = –31.817 + (–0.266*Tokens) + (–0.055*Characters) + (1.019*Sentences) + (0.442*Syllables) + (–0.186*Complex-words)*
- *Upper-intermediate = –31.438 + (–0.261*Tokens) + (–0.059*Characters) + (0.638*Sentences) + (0.464*Syllables) + (0.009*Complex-words)*

To illustrate the potential applicability of the equations above, we can take, for example, a randomly chosen text. Let us imagine that when we compute it, we get the following values:

- Tokens = 300
- Characters = 1,200
- Sentences = 40
- Syllables = 400
- Complex words = 10

Using the above discriminant equations and instantiating the values for tokens, characters, sentences, syllables and complex words, we can calculate the scores of the four discriminant functions:

- *Elementary = –21.862 + (–0.149*300) + (–0.068*1,200) + (1.361*40) + (0.356*400) + (–0.434*10) = 44.338*
- *Pre-intermediate = –13.091 + (–0.116*300) + (–0.043*1,200) + (0.795*40) + (0.260*400) + (–0.207*10) = 34.239*
- *Intermediate = –31.817 + (–0.266*300) + (–0.055*1,200) + (1.019*40) + (0.442*400) + (–0.186*10) = 38.083*
- *Upper-intermediate = –31.438 + (–0.261*300) + (–0.059*1,200) + (0.638*40) + (0.464*400) + (0.009*10) = 30.672*

The randomly chosen text with: tokens = 300; characters = 1,200; sentences = 40; syllables = 400; and complex words = 10, will be assigned to the readability-ease level with the largest resulting value according to the four functions above. Thus, maximising the four coefficients we find that this text is most likely to be an elementary text, as Elementary is the highest resulting coefficient (44.338); in second place, it would be classified under Intermediate (34.239). The least likely group membership would be Upper-intermediate (30.672), as the coefficient obtained in the corresponding equation is the lowest one.

5. CONCLUSIONS

Readability indexes can be really useful for the automatic classification of texts, especially within the language teaching discipline. Among other applications, they allow for the previous determination of the difficulty level of texts directly extracted from the Internet. The problem is that these indexes may offer disparity, and this is precisely what has motivated our attempt to unite their potentiality, utilizing each variable used by them. A discriminant analysis of all the variables under examination has enabled the creation of a much more precise model, improving the previous best results in 15%. It is also worth noting that errors or disparities in the difficulty level of the analyzed texts have been detected.

Our intention is to go more deeply into the refinement and use of readability indexes for areas such as automatic classification of texts, especially within the area of language teaching, comparing different languages and confirming whether these indexes offer a similar degree of precision or if they require any adjustment for its calculation as far as variables are concerned.

REFERENCES

Coleman, M., Liau, T.L. (1975). A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2). (pp. 283-284).

Connaster, B. F. (1999). Last rites for readability formulas in technical communication. *Journal of technical writing and communication*, 29(3). (pp. 271-287).

Dellar, H. & Walkley, A. (2005). *Innovations Elementary*. Coursebook. Boston: Thomson ELT.

Dellar, H. & Walkley, A. (2005). *Innovations Pre-Intermediate*. Coursebook. Boston: Thomson ELT.

Dellar, H., Walkley, A. & Hocking, D. (2004). *Innovations Intermediate*. Coursebook. Boston: Thomson ELT.

Dellar, H., Hocking, D. & Walkley, A. (2004). *Innovations Upper-Intermediate.* Coursebook. (2nd ed.). Boston: Thomson ELT.

DuBay, W. H. (2004). *The Principles of Readability. CA, Impact Information. Available at* http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/bf/46.pdf

Farr, J. N., Jenkins, J. J. & Paterson, D. G. (1951). Simplification of the Flesch Reading Ease Formula. *Journal of Applied Psychology*, 35(5). (pp. 333-357).

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32. (pp. 221–233).

Gunning, R. (1952). *The technique of clear writing*. New York: McGraw-Hill.

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L. & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel.* Research Branch Report 8-75, Millington, TN: Naval Technical Training, U. S. Naval Air Station.

McLaughlin, G. H. (1969). *SMOG grading - a new readability formula.* Journal of reading, 22. (pp. 639-646).

Maxwell, M. (1978). Readability: Have we gone too far? *Journal of reading*, 21. (pp. 525-530).

Smith, E. A. & Senter, R. J. (1967). *Automated Readability Index. Defense Technical Information Center. Available at* http://stinet.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=AD0667273.

Smith, E. A. & Kincaid, P. (1970). *Derivation and validation of the Automated Readability Index for use with technical materials.* Human Factors*, 12(1). (pp. 457-464).*

The conceptual implications in a corpus of lexical errors

MARÍA LUISA CARRIÓ PASTOR
EVA MARÍA MESTRE MESTRE
Universidad Politécnica de Valencia

Abstract
To obtain a precise and exact knowledge of the specific terms the researchers want to use in a text aimed at a specific (peer) reader requires not only knowledge of the grammar of the foreign language, but also a thorough knowledge of the adequate terminology. There are often errors in texts written in a second language. In the case of lexical errors, most are a result of confusion over terms. In this study, the lexical errors produced in technical texts written in English will be determined to identify the most significant features and improve written production in order to obtain a more fluent communication among writers who use the English language as a lingua franca. In addition, identifying the causes of these errors is useful for establishing the necessary guidelines to correct them and to pinpoint the recurrent patterns of these errors. Most errors found in this study are due to mother tongue interference, the incorrect spelling of the term, or an erroneous word choice. These data offer information that we can apply to the learning processes of new terms by non-native speakers of English language, in order to avoid errors and identify the mechanisms which permit the correct production of the specialised lexicon.
Keywords: technical terms, English language, linguistic errors, ESP

Resumen
Llegar al conocimiento preciso y exacto de los términos específicos que se quieren utilizar en un texto dedicado a un lector específico, requiere no solo el conocimiento de la gramática de la lengua extranjera, sino también de la terminología adecuada. A menudo se producen errores en la producción de textos en una segunda lengua, y en el caso del léxico, la mayoría se deben a causas de confusión de términos. En este estudio se van a determinar los errores léxicos que se producen en los textos técnicos escritos en inglés, para con ello, identificar sus características más significativas y mejorar la producción escrita, logrando una comunicación más fluida entre escritores que utilizan la lengua inglesa como lengua vehicular. Por otro lado, discernir las causas de estos errores nos aporta las pautas necesarias para corregirlos y poder establecer los patrones por los que se cometen. La mayoría de los errores encontrados en este estudio se deben a la interferencia de la lengua materna, a una escritura incorrecta del término o a la elección errónea de una palabra. Estos datos nos aportan información que aplicamos a los procesos de aprendizaje de nuevos términos por parte de hablantes no nativos de la lengua inglesa, con el fin de evitar los errores e identificar los mecanismos que permiten una producción correcta del léxico especializado.
Palabras clave: términos técnicos, lengua inglesa, errores lingüísticos, ESP

1. INTRODUCTION

When written communication is directed at an international reader, the establishment of a set of guidelines by the writers is vital for the development and correct understanding of the genre, since communicating in a previously established manner helps the exchange of ideas between writers from different cultural backgrounds and/or mother tongues. One of the genres with a greater international projection is academic writing, and, in particular, the

scientific production, which permits worldwide communication between scientists by means of research articles published in international journals.

Due to its intrinsic nature, scientific language has a series of implicit features which render it completely different from any other genre. It lends itself to a series of characteristics inherent to scientific thought and expression. Also, this hinders its production and understanding to the layperson, with a varied range of specialised terms, depending on the audience being addressed.

A crucial requirement in specific writing is that the authors must be aware of the target audience of the genre employed to enable them to decide upon the type of lexicon and structures to be used in order to appropriately transmit their thoughts, as Dudley-Evans & St. John explain (1998: 115):

> Knowledge of genre involves an understanding of the expectations of the discourse community that reads the text and of the conventions that have developed over time about the structure, the language and the rhetoric of the genre.

In particular, Alcaraz Varó (2000: 138-9) specifies the features of scientific-technical terms in English when the means of communication is the research article:

> *"La alta densidad sémica o conceptual de las unidades léxicas compuestas; el empeño por la precisión expresiva, materializado en los sintagmas nominales largos"*.

Eggins & Martin (2000: 336) enumerate the lexical characteristics of scientific writing in:

a) use of standard syntax without abbreviations;
b) use of structures of lexically dense nominal sentences, with heavy postmodification;
c) use of nominalised vocabulary, with action words expressed by means of nouns;
d) use of elaborate vocabulary;
e) paucity of adverbs;
f) and, finally, the use of terms which possess specialised technical meanings in the academic field.

Furthermore, Cabré (2003: 183) points to essential concepts to determine the features characteristic of specialised language:

> The multifaceted terminological units are at one and the same time units of knowledge, units of language and units of communication. Based on this approach, the description of a terminological unit must necessarily cover these three components: a cognitive component, a linguistic component and a socio-communicative component. […] they fulfil restricted conditions in each of their cognitive, grammatical and pragmatic constituent components.

The abovementioned features can be summarised by confirming that the most common lexical feature is the high density of specialised lexicon. In addition, another important feature of scientific lexical units is that the author does not explain the precise or specialised terms as it is presumed that the reader is already familiar with these. It can even be said that they are some types of code, shared by both reader and writer, since the reader profile is very constrained. Arden-Close (1993) and Mudraya (2006) are aware of the importance of lexicon in the learning of a second language as it causes errors. They indicate that more attention should be paid to this, and that meaning should be explained from a lexical perspective.

Corder (1967) exhaustively studied errors made by second language speakers (L2) from a contrastive point of view, as can be witnessed in his many publications. However, the trend of Error Analysis (EA) re-emerged inspired by Chomsky's generative grammar, which questioned psychological behaviourism - the basis of contrastive analysis - and directed the study towards a new treatment of errors from a more tolerant perspective (Richards, 1971; 1974). However, not until the beginning of the 1970s was there a demand for an analysis capable of explaining errors, which stemmed from other confirmed causes and which were not linked to mother tongue interference.

Several years later, the concept of linguistic relativity as applied to errors has been rekindled. The theory proposed is as follows: if speakers of different languages do not understand each other this is not because their languages do not lend themselves to translation, (which is obviously possible), but because they do not share a common ground with regards to observing and interpreting reality. Therefore values, which signify words, are not represented in the same way; that is, the understanding of another language does not depend on equivalent structures but on equivalence between the concepts emerging from reality and the method of expressing these. As Yoshii (2006: 88) remarks, "[…] L2 learners rely on word-to-word links in early stages, but as their L2 proficiency develops, they link L2 directly to concepts (conceptual links)".

Cabré (2008:15) illustrates this in her reference to terms: *"La razón parece clara: se admite que los términos, como todo signo, además de denominar significan en sí mismos, con lo que la Terminología asumiría uno de los fundamentos esenciales de la lingüística cognitiva: la motivación de los signos del lenguaje"*. In this sense, in order to unravel the mechanisms behind language, an essential aspect is the cognitive interpretation of language.

Error analysis has helped in the understanding of error not as merely an unwanted phenomenon in language, but as a source of information which helps improve learning and production in an L2. Errors detected in the written production facilitate knowledge of the production and help us to understand the mechanisms that the speaker of a foreign language adopts. As a result, by interpreting these error patterns, several strategies can be designed to improve written production in an L2 and several aspects should be considered when analysing errors.

The first important aspect of EA is identification. The correct identification of errors helps to define whether these have arisen due to cognitive, linguistic or socio-communicative lapses. Many studies concerning error have focused on the nature of error, but very few have analysed the ability to identify and interpret error in a second language (Carrió Pastor, 2004; Hamid, 2007; Rifkin & Roberts, 1995). The second important aspect of EA is that errors produced in a second language are a result of different causes. These can be divided into three categories: *interlingual errors*, which are due to first language interference upon the second (learnt) language; *intralingual errors*, which are produced regardless of the mother tongue and are due to a deficiency in the learning process (James, 1998:179; Larsen-Freeman and Long, 1992:58) and finally *conceptual errors*, caused by the failure of the speaker to match the correct idea with the particular expression, that is, a lack of relationship between concept-term.

Webber (1993) states that the most common causes of error in non-native English speakers (NNES) are found in lexicon, due to L1 interference. As Cabré (2003:178) indicates, we could be of the opinion that terms are systematic and equivalent among languages, and that their concepts are semantically precise. However, this author warns us *"[...] with variation according to the different functional registers of specialised communications, the data are less systematic, less unambiguous, less universal than the others"*. Terms help us to communicate and express our thoughts, and consequently change with these, hence preventing total equivalences between languages. The compilation of a lexical error corpus can help in determining why the concept is not universal and therefore depends on cultural

conceptions. Language proves that we have multiple conceptions of reality and it is expressed accordingly to our convictions, culture and linguistic conventions.

The third important aspect in EA is error classification. Lexical errors have traditionally been classified according to formal aspects of vocabulary or to semantic aspects related to meaning. The most frequent formal classification of lexical errors (James, 1998:145) is in: *misselection* (wrong word choice), *misformation* (words that are non-existent in the L2 but exist in L1) and *distortion* (words that are non-existent in both the L2 and the L1). With regards to the semantic errors in lexis, there are two main types: *confusion of sense relations* (a word being used in contexts where a similar word should be used) and *collocational errors* (the choice of a word to accompany another is inappropriate). However, in this study we do not follow this formal classification. Errors are classified according to the reasons which have caused them. By doing this, their very reason for being is determined, as well as the factor which causes their appearance, and whether this is cognitive, linguistic or socio-communicative.

Corpora have been used in English for Specific Purposes since the 1980s, mainly for identifying language characteristics, but it can also be used to detect language errors. The compilation of an error corpus would facilitate the understanding of the conceptual implications in second language learning, student progression and development and also course and material design (Belz, 2004; Chapelle, 2004; Hunston and Francis, 2000; Krishnamurthy and Kosem, 2007; Nelson, 2006).

The objectives of this work are, firstly, to elaborate a corpus of specialised lexicon errors which appear in scientific texts produced by writers with a B2 English level. Secondly, to define the most significant causes of such errors in order to generate guidelines which can help improve written production, thus promoting a more fluent communication between writers using technical English as a language of international communication. Finally, as a result of the aforementioned objectives, the third objective is to propose a new classification of lexical error data, including the conceptual component, in order to determine the relations established by the L2 writer when the error is produced. This can provide the necessary guidelines to determine the linguistic and cognitive components which cause the errors.

2. METHODOLOGY

From the outset of the study the type of corpus that would be used in order to achieve the results and the final reflection upon the lexical errors made in scientific-technical English was

restricted. Through the correction and translation service of the Universidad Politécnica de Valencia we obtained several scientific papers in their original version and the version corrected by native proofreaders. These articles had initially been written by Spanish researchers holding an intermediate (B2) level of English, according to the European Common Framework of Reference for Languages. They were intended for publication in international journals, and had thus previously been sent to the correction service offered by the university.

Once the papers had been obtained, all tables, graphs, bibliography and references were removed and the documents were saved in a text format, to enable the data to be analysed using the computer programme *Wordsmith Tools 5.0* (Scott, 2009). Finally, and after eliminating five papers, which did not have an appropriate level of language due to the expressions and language used, a corpus of thirty original papers written in English by Spanish researchers, together with the corrected versions of these papers, which had been reviewed by native proofreaders, was gathered.

Next, and since currently there is no computer tool capable of detecting lexical errors made in the texts, the lining tool in the *Wordsmith Tools* programme was used to identify errors by comparing the original sentence and the sentence after correction. Once the errors were identified, they were classified depending on the causes that had produced them.

Finally the frequency of error occurrence was calculated in order to obtain information about their relevance in the final corpus, the different frequencies were compared and several conclusions were drawn.

3. RESULTS

The corpus which supports the study presented in this paper for the detection and categorisation of errors displays the following features (see Table 1 below):

Table 1: Statistical data of the articles which integrate the corpus of original and corrected texts.

	SENTENCE FEATURES	FREQUENCY ORIGINAL TEXTS (%)	FREQUENCY CORRECTED TEXTS (%)
STATISTICAL DATA	**Words in the corpus**	**110,154 (50.37%)**	**108,535 (49.63%)**
	List of words	**8,110 (51.68%)**	**7,583 (48.32%)**
	Number of sentences	**5,468 (50.24%)**	**5,416 (49.76%)**
	Average number of words per sentence	**20.1 (50.01%)**	**20 (49.99%)**
	Number of paragraphs	**1,755 (50.78%)**	**1,701 (49.22%)**
	Number of sentences per paragraph	**3.1 (49.63%)**	**3.2 (49.37%)**

It can be seen that the number of sentences, paragraphs, words and lists of words diminishes in the corpus corrected by the native reviewers of the articles. The proofreaders shortened the texts during the revision process, since the Spanish authors used more words than necessary to express themselves in English language. This is a direct consequence of conciseness in the expression of scientific-technical language.

The results have been divided according to the underlying causes of these errors in order to finally design strategies which can facilitate their correction. The results have been grouped as errors caused by mother tongue interference, errors arising due to lexical distortions, errors caused by incorrect word choice and errors due to concept confusions.

The lexicon identified in the corpus of this study was academic, and therefore very concise and circumscribed, since it was employed within a scientific-technical context. The standard speaker would not necessarily be familiar with this type of lexicon as it is directed at a very specific sector within the field of scientific research. Its objective is communication to other groups of researchers, and therefore word choice must be very precise and the expressions used must clearly reflect the sense of what is intended. Often, the results obtained in a research work require explanation in order to be understood. For this reason, the features found when the specific corpus was analysed clearly differ from those which can be found in other, more general, corpus types. This factor was taken into account when the classification of 577 errors found in lexis was further developed. Subsequently, the results found following analysis were distributed into groups according to the reasons behind the errors. Exhaustive knowledge of the reasons which cause errors provides relevant information to the linguist; in as far as they can help determine the relations and conceptual associations of the speakers of

a second language. This information can also assist the educator of a second language, since these results can highlight those aspects which need reinforcement during the learning process. Next, the results obtained will be listed in detail. They will be ordered according to the less frequent causes to the most frequent causes of error.

First of all, the least frequent cause of lexical error was the misspelling of certain words or *lexical distortions*, to use James's term (1998). These were classified as shown in Table 2.

Table 2: Errors due to lexical distortions

	TYPES OF ERRORS	FREQUENCIES (%)
LEXICAL DISTORTIONS	1. Omission of part of the word	11 (68.75%)
	2. Overinclusion within the word	3 (18.75%)
	3. Misordering in the word	2 (12.50%)
	Total	16 (100.00%)

This table shows that the frequent types of error in this group were due to the omission of some part of the word, although the cases detected within this classification were not numerous when considering the overall results. The other two sources of errors had insignificant occurrences, as demonstrated by the percentages. Some examples of these errors are:

1. Omission: *overhead*/ overheads; *modeled*/ modelled; *specially*/ especially; *valuation*/evaluation; *current*/ currently.
2. Overinclusion: *by means*/ following; *ones*/ one; *functioning*/ function.
3. Misordering: *fulfli*/ fulfil.

Next appear errors which arise due to L1 interference (Spanish), since the mental structure, as well as word formation, present a pattern created based on the first language spoken. As such, the usual trend is to copy that pattern in all spoken linguistic expressions. Table 3 shows the results obtained following the analysis of the corpus:

Table 3: Formation errors due to L1 influence

	CAUSES	FREQUENCIES (%)
ERRONEOUS FORMATION L1 INFLUENCE	Calques	53 (65.43%)
	Adaptation of words of L1 to words which do not exist in L2	15 (18.52%)
	Borrowing	13 (16.05%)
	Total	81 (100.00%)

The most significant datum is that almost half the errors found consist of the use of linguistic calques (54%), that is, L1 greatly influenced the choice of vocabulary used in the texts. However, linguistic borrowings and word adaptations from L1, which do not exist in L2, are scarce and therefore were not statistically relevant in the global result of errors due to L1 influence. Some examples of this type of errors are:

1. Linguistic borrowing: *in the casos*/in the cases; *traduced*/ translated; *millones*/ millions.
2. Word adaptation from L1 to L2: *centered*/ focused; *of the order*/ in the range; *bidimensional meshed*/ two- dimensional meshed.
3. Linguistic calques: *realized*/ performed; *more*/ a larger number; *botanic*/ botanical; *precision adjusted*/ precisely adjusted*; the more used*/ the usual; *all the other*/ the remaining.

Thirdly, another type of error was discovered, the choice of an incorrect word. One word was used when a different one was intended. General terms were used within a specific context; several words were coined by the authors, resulting in a set of errors which caused confusion in the receptor. The classification proposed for this type of error, and therefore the results obtained are detailed in Table 4:

Table 4: Errors caused by wrong word choice

	TYPES OF ERRORS	FREQUENCIES (%)
ERRORS DUE TO WRONG CHOICES	Use of a general word instead of a specific term	120 (73.17%)
	Erroneous collocation	32 (19.51%)
	Coinages (invention of new words)	12 (7.32%)
	Total	164 (100.00%)

The most frequent errors were produced as a result of selecting a general word sufficient to the context of the paper instead of employing a specific term. The second category in which errors were found was the erroneous collocation of two words. Some examples of this type of error found in the corpus are:

Word invention (coinage): *obtention/* obtaining*; capacity saying/* stating*; three typlets/* triplets*; nonprofit/* not- for- profit*; noncomplete/* incomplete.

The use of one word with a general meaning instead of a more specific word: *fabrication/* manufacturing*; tried/* tested*; specific/* particular*; happens/* takes place*; stay still/* remain*; use/* employ*; direction/* path.

Erroneous collocation: *reproduce well/ reproduce accurately; are agreed/ agree; whatever branched/ any other branched; no penetration/ zero penetration; principle resources/ main resources; great amount/ a high amount; wrong results/ inaccurate*.

Finally, in the fourth place, we can see in Table 5 those errors produced confusion between concept and term. As can be observed, these are the most numerous in the corpus:

Table 5: Conceptual errors.

	TYPES OF ERRORS	FREQUENCIES (%)
LEXICAL-CONCEPTUAL ERRORS	Choice of a word with a similar meaning to a another	161 (50.95%)
	Error due to confusion between words formally similar to the writer	102 (32.28%)
	Use of a word due to confusion of meaning	53 (16.77%)
	Total	316 (100.00%)

The erroneous choice of a word due to confusion over meaning is not due to L1 influence, but arises because the Spanish writer associates it with the literal meaning, choosing one instead of the other because of their similar forms. As far as error as a result of confusion over meaning, the Spanish writer chose a word because for him or her it represented the same meaning as the other word. Hence, confusion was due to wrong word choice. Examples of this type of error can be observed:

1. Words that are formally similar *such as*/ as; *like*/ such as; *show*/ illustrate; *sensibility*/ sensitivity; *how reliable*/ the reliability; *with success*/ successfully.
2. Choice of a word because it is similar to another word: *grill*/ grid; *conveniently*/ carefully; *to choose*/ to chose; *eventual*/ eventually; *lightened*/ lessened; *produced*/ processed; *able*/ capable.

The use of a word instead of another due to confusion over meaning: *related*/ connected; *institution*/ organisation; *remove*/ eliminate; *goodness*/ degree of success; *foreseeable*/ forecast; *solving*/ resolution; *important rise*/ large increase; second/ secondly; *concordance*/ agreement; *connected*/ related; *pronounced*/ significant; *work*/ paper.

4. CONCLUSION

As mentioned in the previous sections of this paper, the initial hypothesis, which motivated this piece of research, was the determination of errors produced by Spanish writers when using English as lingua franca for their international publications. Traditionally, errors have been divided into grammatical or semantic errors taking into account a linguistic perspective or into *interlingual* or *intralingual* errors from a didactic perspective. However, in this study another variant in the interpretation of errors is studied, a cognitive classification; that is, conceptual errors.

The existing relationship between object-concept-term is not universal and inalterable in language. Several concepts have forms of identification, which vary depending on the cultural background of the speakers, who select certain terms depending on their specialised knowledge of the subject matter, the socio-communicative components inherent in each communication act and the learning processes they may have experienced in that second language. Thus, as Cabré (2003; 2008) states, the relationship object-term is not lineal but implies several nuances. Evidence that these concepts are not associated to specific terms, and that there is no universal form by which these can be labelled centres on the fact that certain errors exist due to the influence of the mother tongue, to lexical distortion or incorrect spelling, to erroneous choice of a term or to an inadequate conceptual association.

As detailed in the results section of this paper, conceptual errors were the most prolific cause of error in the English language production analysed in the present research work; in particular, the most frequent cause of error was the subcategory of choice of a word with a similar meaning to another. This shows that non-native speakers of English with an intermediate (B2) level of English did not achieve a conceptual equivalence between the term and the object despite the fact that their grammatical proficiency is sufficiently adequate to enable communication. These results led to the following questions: what exactly is the linguistic process that helps us relate equivalent terms in two languages? Can we completely equate terms, which refer to concepts or ideas in two different languages? The teaching of a second language is at present being carried out using a communicative approach. Concepts, and not terms, are being taught following this perspective. This may be the reason for the errors found in the corpus. The Spanish writers know conceptually the specific term in their L1. The problem arises when this concept has to be associated to the L2. Therefore, the process would be: object-concept-two terms (L1/L2). According to our results, the writer with an intermediate (B2) level of language proficiency has not absorbed this process, and

simply relates object-concept-L1 term/L2 term. The learning strategies of foreign languages should transmit that the relationship between concept and term is multiple, and is not unidirectional, but travels in as many directions as the languages known to the speaker.

These issues are also reinforced by the results which can be observed in Table 4. Although these errors have been classified as lexical distortions in accordance with James (1998), the erroneous production is also related to a lack of contextualisation, or location within the adequate lexicon the author should use. As can be observed in that table, the most frequent error in that category is the use of a general word instead of a specific term. This shows that the author has not linked the specific concept, which is known to him or her in L1, to the specialised concept in L2, or that the idea has not been identified as a concept.

It may be of interest to point to the need to expand the study concerning the implications of the relationship between object/idea-concept-term, since these are vital for an error-free production in a second language. Not only must the linguistic processes, but also the cognitive and socio-communicative processes of the speakers of a second language be improved in order to ensure the accurate expression of ideas. This can be obtained by a compilation of a corpus of lexical errors, determining their causes and the cognitive relationships that second language writers establish when incorporating a new term.

Finally, it is not our intention to centre upon conclusions that refer to error correction, which can be found in articles purely dedicated to didactics, and which focuses on the learning of a foreign language (Gaskell and Cobb, 2004; Lee, 2004; Salem, 2007). However, we are well aware that the conclusions expressed in this study can be applied to the field of foreign language teaching, both from the point of view of the design of learning strategies and from the elaboration of materials aimed at practising and correcting the aspects which most frequently cause errors for non-native writers of the English language.

REFERENCES

Alcaraz Varó, E. (2000). *El inglés profesional y académico.* Madrid: Alianza.

Arden-Close, C. (1993). Language problems in science lectures to non-native speakers. *English for Specific Purposes, 12*(3). (pp. 251-261).

Belz, J. A. (2004). Learner corpus analysis and the development of foreign language proficiency. *System*, 32. (pp. 577-591).

Cabré, M. T. (2003). Theories of terminology. Their description, prescription and explanation. *Terminology, 9*(2). (pp. 163-199).

Cabré, M. T. (2008). El principio de la poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en terminología. *Ibérica, 16.* (pp. 9-36).

Carrió Pastor, M. L. (2004). *Contrastive analysis of scientific-technical discourse: Common writing errors and variations in the use of English as a non-native language.* Ann Arbour: UMI.

Corder, S. P. (1967). The significance of learners errors. *International Review of Applied Linguistics, 5.* (pp. 161-170).

Chapelle, C. A. (2004). Technology and second language learning: expanding methods and agendas. *System, 32.* (pp. 593-601).

Dudley- Evans, A. & St. John, M. J. (1998). *Developments in English for Specific Purposes: A Multidisciplinary Approach.* Cambridge: Cambridge University Press.

Eggins , S. & Martin, J. R. (2000). Géneros y registros del discurso. In T. A. Van Dijk [ed.] *El discurso como estructura y proceso* (pp. 335-371). Barcelona: Gedisa Editoral.

Gaskell, D. and Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System, 32.* (pp. 301-319).

Hamid, O. (2007). Identifying second language errors: how plausible are plausible reconstructions? *ELT Journal, 61*(2). (pp. 107-116).

Hunston, S. and Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English.* Amsterdam: John Benjamins.

James, C. (1998). *Errors in Language Learning and Use*. London: Longman.

Krishnamurthy, R. and Kosem, I. (2007). Issues in creating a corpus for EAP pedagogy and research. *Journal of English for Academic Purposes, 6.* (pp. 356-373).

Larsen-Freeman, D. & Long, M. H. (1992). *An Introduction to Second Language Acquisition Research.* London: Longman.

Lee, I. (2004). Error correction in L2 secondary writing classrooms: the case of Hong Kong. *Journal of Second Language Writing, 13.* (pp. 285-312).

Mudraya, O. (2006). Engineering English: a lexical frequency instructional model. *English for Specific Purposes, 25.* (pp. 235-256).

Nelson, M. (2006). Semantic associations in Business English: A corpus-based analysis. *English for Specific Purposes, 25.* (pp. 217-234).

Richards, J. (1971). A non-contrastive approach to error analysis. *English Language Teaching Journal, 25.* (pp. 204-19).

Richards, J. (1974). *Error Analysis*. London: Longman.

Rifkin, B. and Roberts, F. D. (1995). Error gravity: a critical review of research design (review article). *Language Learning, 45*(3). (pp. 511-37).

Salem. I. (2007). The lexical-grammatical continuum viewed through student error. *ELT Journal, 61*(3). (pp. 211-219).

Scott, M. (2009). *WordSmith Tools 5.0.* Liverpool: Lexical Analysis Software.

Webber, P. (1993) Writing medical articles: a discussion of common errors made by L2 authors and some particular features of discourse. *UNESCO-ALSED LSP Newsletter*, *15*(2). (pp.38-49).

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology, 10*(3). (pp. 85-101).

La voz pasiva en textos médicos en inglés y español: análisis basado en corpus

SABELA CEBRO BARREIRO
Universidad de Vigo

Resumen
En la lingüística actual, los corpus son una herramienta de trabajo fundamental. Se ha desarrollado un corpus bilingüe de textos médicos en inglés y en español para analizar la voz pasiva en ambos idiomas. Tras la consulta a diferentes fuentes teóricas sobre el uso de la pasiva en inglés y en español, se constata que en este tipo de textos, los textos médicos, se podrían encontrar formas pasivas en ambos idiomas. Mediante la observación de ejemplos concretos extraídos del corpus creado ad hoc se determina si en los textos traducidos al español se mantienen las formas pasivas o si se adaptan de alguna manera en la lengua de llegada.
Palabras clave: corpus, corpus bilingües, textos médicos, pasiva, traducción

Abstract
Corpora are an essential tool in studying and researching in current linguistics. It has been developed a bilingual corpus of medical texts in English and Spanish to analyze the passive voice in both languages. After consulting different theoretical sources on the use of the passive voice in English and Spanish, it has been confirmed that in this type of texts, in medical texts, we could find passive forms in both languages. Observing concrete occurrences drawn from the corpus built ad hoc we determine if in the texts translated into Spanish passive forms are kept or if these forms are somehow adapted in the target language.
Keywords: corpus, bilingual corpus, medical texts, passive, translation

1. INTRODUCCIÓN

En el panorama lingüístico actual, los corpus son una herramienta de trabajo muy utilizada, además de un recurso de consulta casi imprescindible en estudios lingüísticos y de traducción. Debido a este hecho, ante el interés por analizar en profundidad, y de forma comparada, textos médicos escritos en inglés y en español, hemos visto la necesidad de crear un corpus *ad hoc* que podamos utilizar para llevar a cabo este análisis, puesto que en la actualidad no existe un recurso similar sobre el que se pueda trabajar. Las nuevas tecnologías y la sociedad de la información en la que actualmente vivimos ha facilitado esta tarea sobremanera ya que la obtención de textos puede llegar a ser relativamente sencilla si se sabe dónde buscar.

Para este estudio en concreto, y puesto que la temática preferida son los textos sobre medicina, hemos recurrido al organismo sanitario más internacional: la Organización Mundial de la Salud (OMS). La OMS publica mensualmente un boletín con artículos

clasificados en diferentes secciones, entre las que se encuentra la sección "Investigación" que es de la que hemos extraído todos los textos. El resto de secciones, que pueden variar de un mes a otro, se centran en editoriales y en temas relacionados con la salud pero que se acercan más a cuestiones de políticas sanitarias, que se quedan fuera de nuestro objeto de estudio, ya que nuestra intención es analizar textos médicos propiamente dichos.

El Boletín de la OMS se publica en inglés y a continuación se hace un compendio de los artículos más interesantes o relevantes que se traducen a diferentes idiomas, entre ellos el español. Para elaborar este corpus hemos elegido todos los artículos de investigación publicados en la recopilación en español n.º 1 del año 1999. Son doce textos originariamente escritos en inglés y traducidos posteriormente al español. Este corpus tiene un total de 79.783 palabras: 36.029 palabras en inglés y 43.754 palabras en español. En el Apéndice 1 se detallan los títulos de los artículos y el número de palabras desglosado por artículo, tanto en inglés como su correspondiente traducción al español.

El objetivo concreto de este artículo es analizar el uso de la pasiva en los textos en inglés y su forma verbal correspondiente en español. Para ello, en primer lugar, ofreceremos una breve explicación teórica sobre el uso de esta forma verbal en inglés y en español y, en segundo lugar, presentaremos evidencias concretas de estas formas que aparecen en nuestro corpus. Para llevar a cabo este análisis hemos utilizado herramientas lingüísticas informáticas tales como *ParaConc* y *AntConc*, además de software adicional no propiamente lingüístico para poder convertir los textos, originariamente en formato PDF, a formato de texto plano para poder utilizarlos con las herramientas lingüísticas informáticas.

2. La voz pasiva en inglés y en español

La voz pasiva en inglés se forma a partir de una forma conjugada del verbo *to be* seguido del participio pasado (*past participle*) y seguido, si procede, de la frase preposicional introducida por *by*. Asimismo, en inglés existe una pasiva construida con el verbo *to get* en vez de con el verbo *to be*. Sin embargo, al pertenecer esta última estructura el inglés informal (Wardhaugh, 1995: 122), no la tendremos en cuenta en este estudio dado que asumimos que todos los textos publicados en el Boletín de la OMS presentarán un registro formal.

En cuanto al uso de las pasivas en inglés, Hewings lo explica de la siguiente manera en su libro *Advanced Grammar in Use*:

> Here are some situations where we typically choose a passive rather than an active.
>
> - When the agent is not known, is 'people in general', is unimportant, or is obvious, we prefer passives. In an active sentence we need to include the agent as subject; using a passive allows us to omit the agent by leaving out the prepositional phrase with **by** [...].
>
> - In factual writing, particularly in describing procedures or processes, we often wish to omit the agent and use passives [...].
>
> - In spoken English we often use a subject such as **people, somebody, they, we,** or **you** even when we do not know who the agent is. In formal English, particularly writing, we often prefer to use a passive. [...]
>
> - In English we usually prefer to put old information at the beginning of a sentence (or clause) and new information at the end. Choosing the passive often allows us to do this. [...]
>
> - It is often more natural to put agents (subjects) which consist of long expressions at the end of a sentence. Using the passive allows us to do this. (Hewings, M., 1999: 60).

Huddleston *et al.* (2002: 1427) afirman que una oración está en voz pasiva cuando el sujeto se asocia con un papel pasivo, el papel del sujeto paciente. Quirk *et al.* añaden que el cambio de activa a pasiva puede suponer más diferencias que el mero énfasis y que puede entrañar un cambio de significado aunque tan sólo sea en pequeños matices.

Sánchez Benedito clarifica en su libro *Gramática inglesa* que, aunque, en principio, la elección entre activa y pasiva es indiferente, existen pequeños matices que hacen que nos inclinemos por una u otra forma:

> Se debe tener en cuenta, desde luego, que la diferencia entre la voz activa y la voz pasiva es más que nada cuestión de forma externa, ya que, en el fondo, el significado es el mismo. [...]
>
> Sin embargo, aunque en principio la diferencia entre la activa y la pasiva sea cuestión de forma externa, no siempre son intercambiables ambas formas, y, a menudo, es el matiz que se quiere dar el que condiciona la elección. Frecuentemente, el empleo de la voz pasiva comporta matices de impersonalidad, formalismo, etc., de los que carecería la correspondiente oración activa. La voz pasiva está, por consiguiente, más indicada en lenguaje oficial, estilo periodístico, inglés científico, etc., que en conversación corriente. (Sánchez Benedito, F., 1999: 189)

En español, la pasiva se forma con las formas conjugadas del verbo *ser* seguido del participio pasado. Asimismo es muy frecuente la «pasiva refleja» que se construye anteponiendo la partícula *se* al verbo léxico conjugado en voz activa.

La intuición básica de los hablantes nativos es que la forma preferida en la activa, como advierte Ávila: «Una idea se puede expresar tanto en voz activa como en voz pasiva. [...] La voz activa [...] es más propia del español que la pasiva [...].» (Ávila, F., 2006: 240). Sin embargo, Lunsford y Lunn defienden el uso de la pasiva apelando a la necesidad de impersonalidad en determinados casos:

En términos muy sencillos, la voz pasiva es una opción gramatical que permite hablar de una acción sin identificar a la persona que la lleva a cabo.

La voz pasiva responde a una necesidad comunicativa. En muchos casos, queremos hablar de una acción independientemente del autor de ella (el "agente" en términos gramaticales). [...] En otros casos, se utiliza la voz pasiva porque no se quiere, o no se puede, comunicar la identidad de alguien [...] Y muchas veces la identidad del agente es sencillamente irrelevante o de poco interés. (Lunsford, E. J. y Lunn, P. V., 2003: 49)

Más concretamente, Lunsford y Lunn ratifican el uso de esta estructura en el lenguaje médico:

> Con la voz pasiva se puede hablar de una situación verbal sin identificar a la persona que la lleva a cabo. Esta estructura gramatical es sumamente útil porque permite que nos expresemos en términos impersonales. El habla impersonal se ve mucho en el lenguaje de la medicina para difundir información de interés general sobre las enfermedades y los medicamentos. (Lunsford, E. J. y Lunn, P. V., 2003: 99)

Atendiendo a lo expuesto anteriormente, podemos partir de la hipótesis de que las formas pasivas propiamente dichas pueden aparecer con frecuencia en ambos idiomas.

3. EVIDENCIAS DE LA VOZ PASIVA EN INGLÉS Y SU TRADUCCIÓN AL ESPAÑOL

A continuación presentaremos casos concretos de pasivas en inglés e indicaremos la opción de traducción elegida en español.

En inglés se han detectado 652 casos de formas verbales en voz pasiva. En el Apéndice 2 presentamos algunos ejemplos que hemos extraído del corpus y que analizaremos en esta sección. Los ejemplos expuestos en el Apéndice 2, aunque pocos, son representativos del corpus completo ya que hemos escogido casos de los diferentes textos.

Tras un análisis pormenorizado de las evidencias de la voz pasiva en inglés con su correspondiente traducción al español, se ha detectado un uso muy extendido en español de la pasiva refleja, uso que representa un 68% de las pasivas inglesas traducidas como pasivas reflejas en español, como en «is routinely used» que se ha traducido como «se utiliza

sistemáticamente» o «are used» cuya traducción es «se utilizaban». La pasiva refleja es una construcción española mucho más natural y utilizada que la pasiva perifrástica, sobre todo en la lengua oral (López Fernández, 1998). Por este motivo, un texto con pasivas reflejas en vez de pasivas perifrásticas resulta mucho más legible en español.

El 32% restante de pasivas en inglés se ha traducido por diferentes estructuras sin que predomine ninguna sobre otra. Una de estas fórmulas es la utilización de una perífrasis verbal para expresar la pasiva inglesa tal y como se puede observar en «are involved» traducida como «se ven afectadas» o «were also required to return» traducida como «debían enviar».

Por último, hemos observado, en general, que se han empleado recursos para evitar utilizar una pasiva perifrástica en español, llegando, a veces, a reformular toda la oración y omitiendo una forma verbal en español, que expresaría la pasiva inglesa, e incluyendo, por ejemplo, una frase preposicional. Una muestra de este fenómeno es «*The incidence of mumps encephalitis* ***is reported*** *to range from 1 in 6000 mumps case*» que se ha traducido como «Según las notificaciones, la incidencia de encefalitis parotidítica oscila entre 1 por 6000 casos».

En otros ejemplos se ha evitado la utilización de la pasiva en español empleando una forma verbal, aunque en voz activa, para expresar la pasiva inglesa, como es el caso de «*A variety of other clinical symptoms* ***are seen*** *with mumps*» que se ha traducido como «Con la parotiditis **aparecen** otros varios síntomas clínicos».

4. Conclusión

A la luz de los datos que hemos observado en los ejemplos extraídos del corpus, podemos establecer que la pasiva es una forma muy utilizada en los textos médicos ingleses. Sin embargo, no ocurre lo mismo en los textos en español de nuestro corpus. En español, la forma de pasiva propiamente dicha, es decir, el verbo *ser* conjugado seguido del participio pasado, apenas se utiliza, ni siquiera en este tipo de textos, formales y/o académicos.

Como se ha visto, en español se recurren a diferentes fórmulas que evitan incluir la pasiva perifrástica y que son las siguientes: i) pasiva refleja (partícula *se* seguida del verbo léxico conjugado en voz activa); ii) perífrasis verbal; iii) reconstrucción completa de la oración y omisión de la forma verbal; iv) forma verbal en voz activa. Aunque en español se utilicen otras formas distintas de la voz pasiva, se mantiene el mismo significado que en inglés, por lo que podemos adscribir este cambio de estructura a una estrategia lingüística que hace que el texto sea más «ligero» y más «legible» en la lengua de llegada.

APÉNDICES

Apéndice 1: Relación de textos del corpus

Nº Pal.	Inglés	Cód. EN	Cód. ES	Español	Nº Pal.
6.099	Mumps and mumps vaccine: a global review	0101 EN	0101 ES	Parotiditis y vacuna antiparotidítica: situación mundial	7.541
2.587	Measles: effect of a two-dose vaccination programme in Catalonia, Spain	0102 EN	0102 ES	Sarampión: efecto de un programa de vacunación con dos dosis en Cataluña (España)	2.917
2.275	The antibacterial paradox: essential drugs, effectiveness, and cost	0103 EN	0103 ES	La paradoja de los antibacterianos: medicamentos esenciales, eficacia y costo	2.786
2.323	Pathobiological determinants of atherosclerosis in youth (PBDAY Study), 1986-96	0104 EN	0104 ES	Determinantes biopatológicos de la aterosclerosis en la juventud, (Estudio PBDAY) 1986-1996	2.673
2.068	Use of illicit drugs among high-school students in Jamaica	0105 EN	0105 ES	Consumo de drogas ilícitas entre los estudiantes de secundaria en Jamaica	4.443
3.512	Smoking: attitudes of Costa Rican physicians and opportunities for intervention	0106 EN	0106 ES	Tabaquismo: actitudes de los médicos de Costa Rica y oportunidades de intervención	4.278
3.291	Rapid ethnographic assessment of breastfeeding practices in periurban Mexico City	0107 EN	0107 ES	Evaluación etnográfica rápida de la práctica de la lactancia natural en una zona periurbana de Ciudad de México	4.954
4.235	Chagas disease vector control through different intervention modalities in endemic localities of Paraguay	0108 EN	0108 ES	Lucha contra los vectores de la enfermedad de Chagas mediante distintas modalidades de intervención en localidades endémicas del Paraguay	3.688
3.146	Epidemiology of human fascioliasis: a review and proposed new classification	0109 EN	0109 ES	Epidemiología de la fascioliasis humana: revisión y propuesta de nueva clasificación	2.293
2.042	Seasonal diarrhoeal mortality among Mexican children	0110 EN	0110 ES	Mortalidad estacional por diarrea entre los niños mexicanos	2.502
2.078	True status of smear-positive pulmonary tuberculosis defaulters in Malawi	0111 EN	0111 ES	Situación real de los pacientes con tuberculosis pulmonar y frotis positivo inobservantes del tratamiento en Malawi	2.956
2.373	Tobacco smoking among Portuguese high-school students	0112 EN	0112 ES	Consumo de tabaco entre estudiantes de secundaria portugueses	2.723
36.029	**Total palabras en inglés**			**Total palabras en español**	**43.754**

Apéndice 2: Ejemplos de voz pasiva en inglés con su traducción al español

a reported to WHO up to Apri1 1998, mumps vaccine **is routinely used** by national immunization programme	0101 EN.txt
OMS hasta abril de 1998, la vacuna contra la parotiditis **se utiliza sistemáticamente** en los programas nacionales de inmunización	
ance with investigation of outbreaks. Where mumps **is targeted** for elimination, countries need to add a	0101 EN.txt
Donde **se haya fijado** la meta de eliminar la parotiditis, los países habrán de añadir una segunda dosis de vacuna para los niños,	
tries in different regions of the world, Guidance **is provided** for countries contemplating the introduc	0101 EN.txt
Se dan orientaciones a los países que tienen previsto introducir la vacuna y a los que ya la utilizan.	
are the only natural hosts for mumps virus, which **is usually spread** by respiratory droplets. The incub	0101 EN.txt
es el único huésped natural del virus de la parotiditis, que **normalmente se propaga** mediante microgotas respiratorias.	
ead systemic involvement (Table 1). Classic mumps **is characterized** by enlargement of the parotid and o	0101 EN.txt
La parotiditis clásica **se caracteriza** por una inflamación de la glándula parótida y de otras glándulas salivales	
ng pregnancy has not been found (6). Pancreatitis **is seen** in about 4% of patients with mumps (7). Ther	0101 EN.txt
Se ha detectado pancreatitis en alrededor del 4% de los pacientes con paperas	
als (8). In mumps case the central nervous system **is frequently infected** and about 50% of asymptomatic	0101 EN.txt
En los casos de parotiditis con frecuencia **se produce** una infección del sistema nervioso central	
ability (11). The incidence of mumps encephalitis **is reported** to range from 1 in 6000 mumps cases (0.0	0101 EN.txt
Según las notificaciones, la incidencia de encefalitis parotidítica oscila entre 1 por 6000 casos (0,02%)	
21). Death due to mumps is exceedingly rare, and **is mostly caused** by mumps encephalitis. In the USA,	0101 EN.txt
La muerte por parotiditis es muy rara y **casi siempre se debe** a la encefalitis asociada.	
oduction of vaccine. Protective maternal antibody **is passively transferred** to the infant and its half-	0101 EN.txt
Se transmiten pasivamente al lactante anticuerpos maternos protectores cuya semivida es de 35-40 días	
three-quarters of cases and other salivary glands **are involved** in 10% of cases (1). Epididymo-orchitis	0101 EN.txt
la enfermedad es bilateral en las tres cuartas partes de los casos; en el 10% **se ven afectadas** otras glándulas salivales	
e deaf (17). A variety of other clinical symptoms **are seen** with mumps. Mild renal function abnormalitie	0101 EN.txt
Con la parotiditis **aparecen** otros varios síntomas clínicos	
n higher amounts. Sorbitol and hydrolysed gelatin **are used** as stabilizers in mumps vaccine, and neomyci	0101 EN.txt
En las vacunas antiparotidíticas **se utilizan** sorbitol y gelatina hidrolizada como estabilizantes y neomicina	
d with human immunodeficiency virus (HIV) and who **are not severely immunocompromised** (35). Mumps vaccin	0101 EN.txt

virus de la inmunodeficiencia humana (VIH), sintomáticas o no, y a quienes **no sufran trastornos inmunitarios grave**s (35).	
ndings are reassuring further prospective studies **are planned**. In Germany, the Jeryl Lynn strain was as	0101 EN.txt
Si bien estos resultados son alentadores, **está prevista** la realización de nuevos estudios prospectivos	
y 8-11 million doses of Leningrad-3 mumps vaccine **are produced** annually (34). Studies have shOW1 89- 98	0101 EN.txt
Su producción anual es de unos 8-11 millones de dosis de vacuna (34).	
ted by vaccination. As of mid-1998, mumps vaccine **was routinely used** by national childhood immunization	0101 EN.txt
A mediados de 1998, la vacuna contra la parotiditis **se utilizaba sistemáticamente** en los programas nacionales de inmunización de 82 países	
illness, In the pre-vaccine era in Sweden, mumps **was estimated** to cause about 1000 cases of meningitis	0101 EN.txt
En Suecia, **se estimaba** que en la época anterior a la vacuna la parotiditis causaba unos 1000 casos de meningitis cada año,	
preventable cases. Associations between variables **were determined** using odds ratios (OR). The incidence	0102 EN.txt
Las asociaciones entre variables **se determinaron** calculando la razón de posibilidades («odds ratio», OR).	
in 1971 to 30.9 in 1995. A total of 50 outbreaks **were investigated**. The outbreaks that occurred in the	0102 EN.txt
Se investigaron en total 50 brotes. En los brotes declarados en los dos últimos años	
heads of health care centres (in- and outpatient) **were also required to return** a form showing the number	0102 EN.txt
los jefes de los centros de atención sanitaria (hospitales y ambulatorios) **debían enviar** un formulario señalando el número de	
endemicity of human fascioliasis is low, habitats **have been identified** where lymnaeids are infected but which	0109 EN.txt
En Córcega, donde el nivel de endemicidad de la fascioliasis humana es bajo, **se han descubierto** hábitats en	
bes capable of detecting F. hepatica in lymnaeids **have been developed** (47, 48, 53-56). One such assay detects	0109 EN.txt
Se han obtenido varias sondas de AD N capaces de detectar F hepática en limneidos (47) 48) 53-56). Un ensayo de ese tipo	
n response to the cholera epidemic in 1991. • It **has been well documented** that rotavirus diarrhoeas are oft	0110 EN.txt
cólera de 1991. • **Está bien documentado** que las diarreas por rotavirus causan a menudo deshidratación grave	
e of male smokers (6) 7). Smoking by adolescents **has been studied** in several developed countries (1) 2) 8)	0112 EN.txt
El tabaquismo de los adolescentes **se ha estudiado** en varios países desarrollados (1) 2) 8)	

Alcántara Plá, M. (2007). *Introducción al análisis de estructuras lingüísticas en corpus.* Madrid: Ediciones UAM.

Anthony, L. (2007). *AntConc 3.2.1w (Windows)*. Japan.

Ávila, F. (2006). *Español correcto para dummies.* Barcelona: Granica.

Barlow, M. (1996-2001). *ParaConc*. Houston: Athelstan.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma.* Frankfurt: Peter Lang.

De Irazazábal, A., Fernández P. y De Andrés, I. (2001). Traducción y nuevas tecnologías. Herramientas auxiliares del traductor. En C. Valero Garcés e I. De la Cruz Cabanillas (Eds.), *Traducción y nuevas tecnologías. Herramientas auxiliares del traductor* (pp.707-712). Alcalá: Universidad de Alcalá, Servicio de Publicaciones.

Gonzalo Claros, M. (2005). MedTrad: un foro de traducción médica en internet. *Trans. Revista de Traductología, 9*, 151-159.

Hewings, M. (1999). *Advanced Grammar in Use*. Cambridge: Cambridge University Press.

Huddleston, R. *et al.* (2002). *The Cambridge Grammar of the English Lnaguage*. Cambridge: Cambridge University Press.

Johansson, S. (2007). *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies.* Ámsterdam/Filadelfia: John Benjamins Publishing Co.

Llorach, E. A. (1994). *Gramática de la lengua española.* Madrid: Espasa-Calpe.

López Fernández, J. (1998). La voz pasiva y la construcción impersonal en español: dos maneras de presentar, manipular y seleccionar información. En *Actas del IX Congreso Internacional de ASELE*. Santiago de Compostela: ASELE.

Lunsford, E. J. y Lunn, P. V. (2003). *En otras palabras: Perfeccionamiento del español por medio de la traducción.* Washington D.C.: Georgetown University Press.

McEnery, T., Xiao, R. y Tono Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book.* Nueva York: Routledge.

Orozco Jutorán, M. (2002). Revisión de investigaciones empíricas en traducción escrita. *Trans. Revista de Traductología, 6*, 63-85.

Quirk, R. *et al.* (1985). *A Comprehensive Grammar of the English Language*. Londres: Longman.

Sánchez Benedito, F. (1999). *Gramática inglesa*. Madrid: Alhambra Longman.

Sánchez Trigo, E. (2005). Investigación traductológica en la traducción científica y técnica. *Trans. Revista de Traductología, 9*, 131-148.

Wardhaugh, R. (1995). *Understanding English Grammar: A linguistic approach.* Oxford: Blackwell.

A corpus based approach on event structure: simple and complex predicates of Spanish

MARTA COLL-FLORIT
Universitat Oberta de Catalunya
JUAN APARICIO
IRENE CASTELLÓN
Universitat de Barcelona

Abstract
In order to modelize the behavior of event structure for the simple and complex verbal predicates of Spanish, we use corpora to obtain empirical evidence of verbal behavior. We assume the hypothesis that the Aktionsart is compositional. However, unlike the traditional approach, we start from the basic hypothesis that the verbal units have different lexical weights or degrees of interaction between the lexicon and the grammatical construction. The data show that aspectual categories impose different contextual restrictions and that there exist different degrees of prototypicity within each category –some members are more stable / flexible than others-. The gradual internal structuring of each category is not arbitrary; it is due to the semantic characteristics shared by the different aspectual verbal subsets.
Keywords: Aktionsart, prototypicity, corpus linguistics

Resumen
Para poder modelizar el comportamiento eventivo de los predicados verbales simples y complejos del español, utilizamos corpus para obtener evidencia empírica del comportamiento verbal. Asumimos la hipótesis de que el Aktionsart es composicional, sin embargo, no como en la aproximación tradicional, nuestra hipótesis de partida básica es que las unidades verbales presentan diferentes pesos léxicos o grados de interacción entre el léxico y la construcción gramatical. Los datos muestran que las categorías aspectuales imponen diferentes restricciones contextuales y que existen diferentes grados de prototipicidad dentro de cada categoría –algunos miembros son más estables / flexibles que otros-. La estructura gradual interna de cada categoría no es arbitraria, sino que se debe a las características semánticas que comparten los diferentes subconjuntos verbales aspectuales.
Palabras claves: Aktionsart, prototipicidad, lingüística de corpus

1. INTRODUCTION

The main goal of this paper is to present the research[1] that we have carried out in order to modelize a representation of event structure for the simple and complex predicates of Spanish. In the different existing computational resources (FrameNet (Subirats, 2005), Adesse (García-Miguel, Costas and Martinez, 2005), AnCora (Aparicio, Taulé and Martí, 2008), WordNet (Miller, Beckwith, Fellbaum, Gross and Miller, 1990)), either this information is not present or the analysis tends to be quite simplified. Moreover, periphrases

[1]This research has been carried out within the KNOW2 project granted by Ministry of Education and Science, Spain Government (Semantic Knowledge Representation (SKR), TIN2009-14715-C0403).

are not represented at all or, if they are, the representation is not made from an event structure perspective.

From a methodological point of view, it is difficult to find aspectual studies that contain a significant amount of predicates; generally, the analysis is reduced to a set of predicates (Dowty, 1979; Verkuyl, 1993; among others). Accordingly, our interest is to work with a wide and varied number of predicates in order to obtain a representation system based on empirical methods. With this idea in mind, we use corpora to obtain empirical evidence of verbal behavior, which is the main goal of this paper. In addition to that, we carry out psycholinguistic experiments to determine the basic aspectual properties that human beings distinguish in language processing (Coll-Florit and Gennari, 2009). These experiments will allow us to use these properties as basic elements in the predicates representation, as well as to validate the data obtained from corpora.

Our starting points are the four ontological types of events (states, activities, accomplishments and achievements), which correspond with the classification of Vendler (1967) and Dowty (1979). We also assume the hypothesis that the Aktionsart is compositional. However, unlike the traditional approach (Verkuyl, 1993), we start from the basic hypothesis that the verbal units have different lexical weights or degrees of interaction between the lexicon and the grammatical construction. In particular, depending on the degree of aspectual flexibility that the verbs accept, we assume that three types of predicates can be identified, following Coll-Florit (2009): stable monosemic verbs (e.g. equivaler 'to be equivalent'), flexible monosemic verbs (e.g. *perder* 'to lose'), and polysemic verbs (e.g. *contener* 'to include / to hold back'), a typology that is reflected in different degrees of syntactic, semantic and morphological stability -from less to more flexible, respectively-.

Another theoretical difficulty that we are now analyzing refers to verbs participating in complex predication and their effects on the event structure. Aspectual periphrases are sensitive to and may produce effects on the Aktionsart of the predication (Dick, 1989). For example, the Egressive Aspect, such as the one expressed by means of terminar de 'to finish' can only modify dynamic predications and it renders the predication telic. Therefore, the context can provoke that some verbal predicates may move towards other aspectual classes.

2. STUDY OF CORPORA

We focused on monosemic verbs. The study was reduced to two basic aspectual categories: states (non-dynamic and durative events) and achievements (dynamic and punctual events). The aims of this study were:

1. To test whether these aspectual categories imposed different contextual restrictions;
2. To check out if there were different degrees of prototypicity within each category -some members are more stable / flexible than others-, which may imply movement to another aspectual categories

The procedure of the study was divided into two phases. In the first one, a total of 60 simple Spanish predicates were analyzed: 30 belonging to states and 30 to achievements. We have considered a total of 14 grammatical constructions that, in the bibliography on Aktionsart, are used in a regular way to identify parameters such as dynamism, delimitation and / or duration. In the second phase, in order to confirm the results, we extended the study to exactly the same group of verbs when appearing as complex predicates, specifically as aspectual periphrases.

The basic methodology consisted on analysing the frequencies of appearance of each verb for the total number of constructions. In particular, with reference to simple predicates, the study was based on a subcorpus of 81 million words from the Corpus de Referencia del *Español Actual* by the *Real Academia Española*[2]. As for complex predicates, the study was based on a subcorpus of 23 million words from the *Corpus del Español*[3].

The procedure of the analysis was divided into two stages. In the first one, an *intercategorial* analysis was carried out in order to check out if the two selected aspectual categories (states and achievements) presented different patterns of use. In the second stage, an *intracategorial* analysis was carried out so as to identify which was the set of verbs of a category that appeared with the highest and lowest frequency for each construction.

The results of the intercategorial analysis clearly show different patterns of morphosyntactic use for states and achievements. For instance, a clear interaction has been

[2] http://www.rae.es
[3] http://www.corpusdelespanol.org

noticed between the lexical aspectual type and the frequency of appearance with a determined verbal tense: stative predicates, inherently non-delimited, are more frequent with imperfective tenses, whereas predicates that express achievements, inherently delimited, present the highest frequency rate with perfective tenses (table 1).

Table 1: Intercategorial analysis of the simple predicates: verbal tense

	Present	Imperfect	Past Simple
States	40 %	15%	3,6%
Achievements	16 %	4%	16%

Moreover, the results show equivalent patterns of frequency among constructions that imply the same aspectual parameter, as well as inverse patterns for those contexts that imply opposite parameters (table 2).

Table 2: Intercategorial analysis of the simple predicates: the durative parameter

		States	Achievements
Punctual constructions	*De repente* *'S*uddenly'	0,003%	0,14%
	A las X horas 'At X'	0,0001%	0,07%
Durative constructions	*Desde hace X tiempo* 'X time ago'	0,4%	0,009%
	Durante X tiempo 'For X time'	0,7%	0,05%

According to the results of the intracategorial analysis, aspectual categories show a gradual internal structuring, which is not arbitrary; it is due to the semantic characteristics shared by the different aspectual verbal subsets.

Within the states, three subtypes differing in aspectual flexibility have been identified: permanent, transitory and psychological states. The predicates that express permanent states (e.g. *equivaler* 'to be equivalent'*, caber* 'to fit') are the most prototypical of the category, since they do not accept either dynamic or punctual constructions. Regarding the transitory states (e.g. *estar preocupado* 'to be worried'*, estar enfermo* 'to be ill'), they occupy an intermediate position since they admit constructions that delimit the temporary period in which the situation takes place. Finally, the psychological states (e.g. *conocer* 'to know'*, creer* 'to believe', *gustar* 'to like') are the most flexible ones. Specifically, they are closer to achievements when they appear in the simple past: *Joaquín, de 25 años, conoció a la empresaria* ['Joaquín, 25 years old, met the businesswoman'], or in the immediate

prospective periphrasis: e*stán a punto de conocerse* [‘They are about to know each other’]. However, when they appear with an ingressive periphrasis, they are closer to processes: *a Oliveira le empezó a gustar más el cigarillo* [‘Oliveira began to like more the cigarette’].

Regarding achievements, they also present an internal gradation. On the one hand, there are prototypical punctual predicates that do not accept durative constructions (e.g. *atrapar*, ‘to catch’, *detectar*, ‘to detect’) and, on the other hand, there are more flexible predicates that accept changes of aspectual interpretation (e.g. *perder*, ‘to lose’, *cerrar*, ‘to close’). More precisely, this last verbal group accepts durative constructions focusing on the resulting state of the achievement: *cerraron las instalaciones durante una semana* [‘They closed the facilities during a week’]. When this happens, they are closer to states. This subset of verbs also accepts ingressive periphrases: *comenzó a perder la vista a los diez años* [‘He began to lose his sight when he was ten years old’]. In this case, they profile the ingression of a process. Finally, this group of verbs also admits egressive periphrases: *ya ha terminado de cerrar la puerta,* [‘She has already finished closing the door’]. In this case, they are closer to accomplishments since they focus on a delimited durative event.

We have just seen how verbal predicates may move towards other aspectual classes, with unequal different lexical strength -some elements are more prototypical than others-.

3. TOWARDS A REPRESENTATION OF EVENT STRUCTURE

Several resources such as *WordNet*, *SenSem* or *AnCora* include event structure information in the semantic characterization of verbal predicates. However, these lexical resources present three basic problems. First of all, each resource adopts a different typology of classification, which makes it difficult to establish equivalence relations between them. Secondly, neither the movement between aspectual categories, nor the different degrees of prototypicity within each category are represented. Finally, and in a related way, the different existing lexical resources codify the verbal senses ‘per se’, without considering the emergence of periphrastic combination units.

Given these lacks, and considering the results of the present study, we understand that the computational representation of the event structure has to include, at least, the interaction of three basic phenomena for each verbal sense:

1. Inherent event structure configuration. This factor will have to reflect the association prototypicity of the predicates with the different classes;

2. Description of the incompatible constructions and/or constructions that imply a change of aspectual interpretation (where the verbal periphrases would be included);
3. Semantic type of the complements, a key factor when representing aspectual polysemic verbs, as long as each sense imposes different selection restrictions (Coll-Florit, 2009).

4. CONCLUSIONS AND FUTURE WORK

We have presented the research that we have carried out in order to reach, in a near future, a representation of event structure based on empirical methodology. We have also presented the theoretical problems we have come across and the solutions that we propose. According to this framework we have presented empirical studies based on corpora and the results obtained for simple as well as for complex predicates. The data show that aspectual categories impose different contextual restrictions and that there exist different degrees of prototypicity within each category. Finally, we have pointed out the main characteristics that our representation system is going to have.

For the future work, it is necessary to describe this representation system in detail, to decide the formal language that will be used, and to define in which way the properties associated with the different classes and with the predicates will be implemented. This representation system must reflect the degree of prototypicity of a predicate in relation to its class. Moreover, it must be related to the capacity of a predicate to move to other aspectual classes depending on the context.

REFERENCES

Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2007). The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish. N.Nikolov, K. Bontcheva, G.Angelova and R. Mitkov. (ed.), *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005.* John Benjamins Publishing Co.

Aparicio, J. M. Taulé, M.A. Martí (2008). Ancora-Verb: A Lexical Resource for the Semantic Annotation of Corpora, *Proceedings of 6th International Conference on Language Resources and Evaluation*. Marrakesh.

Coll-Florit, M. (2009). La modalitat de l'acció. Anàlisi empírica, reformulació teòrica irepresentació computacional. PhD Dissertation. IN3/UOC.

Coll-Florit, M., S. Gennari (2009). Time in language: event duration in language comprehension, *Proceedings of the 22nd Annual* CUNY *Conference on Human Sentence Processing*. University of California.

Dick, S.C. (1989). *The Theory of Functional Grammar*. *Part 1*.: The structure of the Clause. Dordrecht: Foris.

Dowty, D. (1979). *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Dordrecht: Reidel.

García-Miguel, J. M., L. Costas, S. Martínez (2005). Diatesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. Wotjak, Gerd, & Juan Cuartero Otal (eds.), *Entre semántica léxica, teoría del léxico y sintaxis*. Frankfurt am Main: Peter Lang.

Miller G.A., R. Beckwith, C. Fellbaum, D. Gross, K. Miller (1990). *Five Papers on Wordnet*. CSL Report 43. Cognitive Science Laboratory, Princeton University.

Subirats Rüggeberg, C. (2005). FrameNet español. Una red semántica de marcos conceptuales. E. Serra, G. Wotjak, (eds.), *Cognición y percepción lingüísticas.* Valencia: Universidad de Valencia and Universidad de Leipzig (pp. 182-196).

Vendler, Z. (1967). *Linguistics in Philosophy*. Ithaca, N.Y.: Cornell University Press.

Verkuyl, H.J.(1993). *A Theory of Aspectuality: The Interaction between Temporal and Atemporal Structure*. Cambridge: Cambridge University Press.

Clause pattern DB: a corpus-based tool

ELISABET COMELLES
NATÀLIA JUDITH LASO
ISABEL VERDAGUER
EVA GIMÉNEZ
Universitat de Barcelona

Abstract
The recent proliferation of studies on corpus linguistics is undeniable and thus its contribution to the field of applied linguistics is remarkable. Most corpus linguistic studies carried out up to now deal with the description of language from an empirical perspective, yet there are few empirical studies exploring the benefits of following this approach in an L2 and EFL teaching context.
As part of a project consisting in the compilation of specific sub-corpora from mystery novels, the GReLiC group at the University of Barcelona is currently working in the production of new teaching materials based on corpora. So as to organise and manipulate in a more effective way the vast amount of linguistic data provided by the compiled corpus, the Clause Pattern DB has been developed. This paper aims at presenting the potential of such a tool for the design and creation of class materials.
keywords: corpus-based grammatical studies, clause pattern database, L2 & EFL teaching

Resumen
La abundante literatura en lingüística de corpus evidencia su innegable contribución al campo de la lingüística aplicada. A pesar de la existencia de numerosos estudios empíricos basados en corpus, los beneficios del uso de ejemplos de lengua real en la enseñanza de segundas lenguas y lenguas extranjeras han sido escasamente estudiados.
A partir del análisis de un corpus de novelas de misterio, nuestro grupo de investigación (GReLiC, Universitat de Barcelona) está actualmente trabajando en el diseño de materiales ilustrativos de los contenidos tratados en el aula. Con el objetivo de organizar y manipular la gran cantidad de datos lingüísticos proporcionados por el mencionado corpus hemos diseñado una base de datos de patrones sintácticos (Clause Pattern DB). Este artículo presenta el potencial de dicha herramienta para el diseño y creación de recursos y materiales docentes.
Palabras clave: estudios gramaticales basados en corpus, base de datos de patrones sintácticos, enseñanza de segundas lenguas y lenguas extranjeras

1. INTRODUCTION[1]

The recent proliferation of studies on corpus linguistics is undeniable (Ädel & Reppen, 2008; Lüdeling & Kytö, 2008; Scott & Tribble, 2006; Tognini-Bonelli 2001) and thus its contribution to the field of applied linguistics is remarkable (Aijmer, 2009; Granger & Meunier, 2008, Gavioli, 2005; Hunston, 2002). Most corpus linguistic studies carried out up to now deal with the description of language from an empirical perspective, yet there are few

[1] The authors acknowledge the support of the AGAUR (2008MQD00020) and the *Programa d'Innovació Docent* of the University of Barcelona (2009PID-UB/06) and the technical support provided by *temuevo.com*.

empirical studies exploring the benefits of following this approach in an L2 teaching context (Laso & Giménez, 2007 & 2008). Despite the fact that those benefits have been described a priori and seem consistent with language learning theory, little research has been carried out trying to assess the effectiveness of using corpus-based materials in an EFL classroom.

Retrieving information from a corpus has proven to be a fundamental resource for any language user, as it provides them with new insights into language structure and use. According to Tsui (2004: 42), EFL learners are not sufficiently exposed to the target language so as to be able to acquire that language efficiently on their own. Thus, the use of corpora in an EFL context contributes to fill that gap since it allows learners to focus on naturally-occurring utterances as well as highly-frequent combinations of words.

In line with such a methodological perspective and as part of a project consisting in the compilation of specific sub-corpora from mystery novels, a group of instructors involved in the teaching of Descriptive English Grammar and English Lexicology and Morphology at the University of Barcelona has been working over the last two years in the production of new teaching materials based on corpora and the use of concordancing programmes. These newly designed materials are aimed at providing students with a wide representative sample of clause patterns in context. So as to organise and manipulate in a more effective way the vast amount of linguistic data provided by the compiled corpus, we thought it would be wise to design a database which would organise and systematize the contents of the corpus data. Such a database would thus improve the teaching and learning of English lexicology and grammar.

The choice of mystery novels to compile our corpus was motivated by two main purposes; on the one hand, show large quantities of real language in use and on the other, offer contemporary texts by authors students could be familiar with. It seemed to us that well-known mystery novels could draw learners' attention more easily.

In this paper we aim at presenting the Clause Pattern DB designed by the GReLiC[2] group. This tool will simplify and automate the creation of teaching/learning materials for the subjects we are currently teaching at the English Department of the University of Barcelona. More specifically, the Clause Pattern database will serve the following purposes:

1. Systematize and organize the information provided by the compiled corpus.

[2] *Grup de Recerca en Lexicologia i Lingüística de Corpus* (GReLiC) [Lexicology and Corpus Linguistics Research Group], University of Barcelona.

2. Design teaching/learning materials, such as supplementary exercises illustrative of class contents, learning evaluation tools, in-class and self-learning activities.
3. Interrelate the contents of different subjects (e.g. English Lexicology and Morphology, Descriptive English Grammar).
4. Enhance collaborative work between discussion groups, aimed at promoting good teaching experiences/practice. This will certainly help to make the corpus maximally used as well as design corpus-based teaching materials.

This study also attempts at illustrating a lexical approach to grammar which entails a new descriptive perspective; that is, we aim at throwing some light on the usefulness of such an approach. Under this umbrella, clear meanings are associated with certain syntactic patterns. In Francis' words,

> syntax is driven by lexis: lexis is communicatively prior. As communicators we do not proceed by selecting syntactic structures and independently choosing lexis to slot into them. Instead, we have concepts to convey and communicative choices to make which require central lexical items, and these choices find themselves syntactic structures in which they can be said comfortably and grammatically (1993: 143)

Traditionally, language has been described in language classrooms by means of setting rules and offering some examples to reinforce such pre-established rules. However, these rules hardly reflect real instances of language. Therefore, our research team considers that a new approach that emphasises the close interrelationship between lexis and syntax and offers examples of real language in context will be extremely beneficial for language teaching. In this context, corpus linguistics can help a great deal to compensate the existing mismatch between traditional descriptions and real language instances.

2. METHODOLOGY

The GReLiC group compiled a corpus based on mystery novels containing 7,675181 words from general contemporary English. The compilation of this corpus met the need of using real language in class and in the design of class materials, as we tend to use a corpus-based approach in our classes. To this respect, the design of a database which would help us to store all the data extracted from the corpus seemed to be a useful tool so as to make the corpus maximally used and design teaching materials.

The starting point to exploit the corpus and create the database was the subcategorization or valence of lexical verbs, being this one of the main contents looked at in the course of Descriptive English Grammar when dealing with Clause Complementation and Adjuncts. A selection of prototypical verbs, previously used in class, and illustrative of the 5 canonical patterns (i.e. SV, SVCs, SVO, SVOCo and SVOO) established by Huddleston & Pullum (2002), was used to perform several searches by means of *WordSmith Tools*[3].

Once explored those verbs and patterns, further searches were made with semantically & syntactically-related verbs. For instance, verbs which behave similarly from a syntactic point of view (e.g. *buy*/*sell* which both can undergo a dative alternation) or others which, despite being semantically related, have a different syntactic behaviour (e.g. *give*/*donate* are semantically related but only *give* can undergo a dative alternation), were taken into account.

New examples drawn from the corpus were then analysed and introduced in the database by means of registers. Each register contains the following fields to be filled: the sentence extracted from the corpus, the pattern of the sentence (by means of a scroll down menu the appropriate pattern can be selected), the lexeme (dictionary entry) of the main verb and the verb valence (only the phrasal categories of the complements are included).

As shown in Figure 1, the lexical verb analysed is *defend* in the sentence *The squatters were defended by a group of solicitors*. The syntactic pattern of this sentence appears in the Patterns box, SV[pass]Obl[by], as the analysed verb in this sentence is a passive verb which subcategorizes for a Subject and an Oblique introduced by the preposition *by*. The lexeme of the main verb under analysis is also taken into account in the Lexeme box (*defend*). Finally, the valence of the verb is stated in the Valence box by means of phrasal categories. In the current example, the Subject is realised by a NP and the Oblique by a PP introduced by the preposition *by*.

[3] http://www.lexically.net/wordsmith/

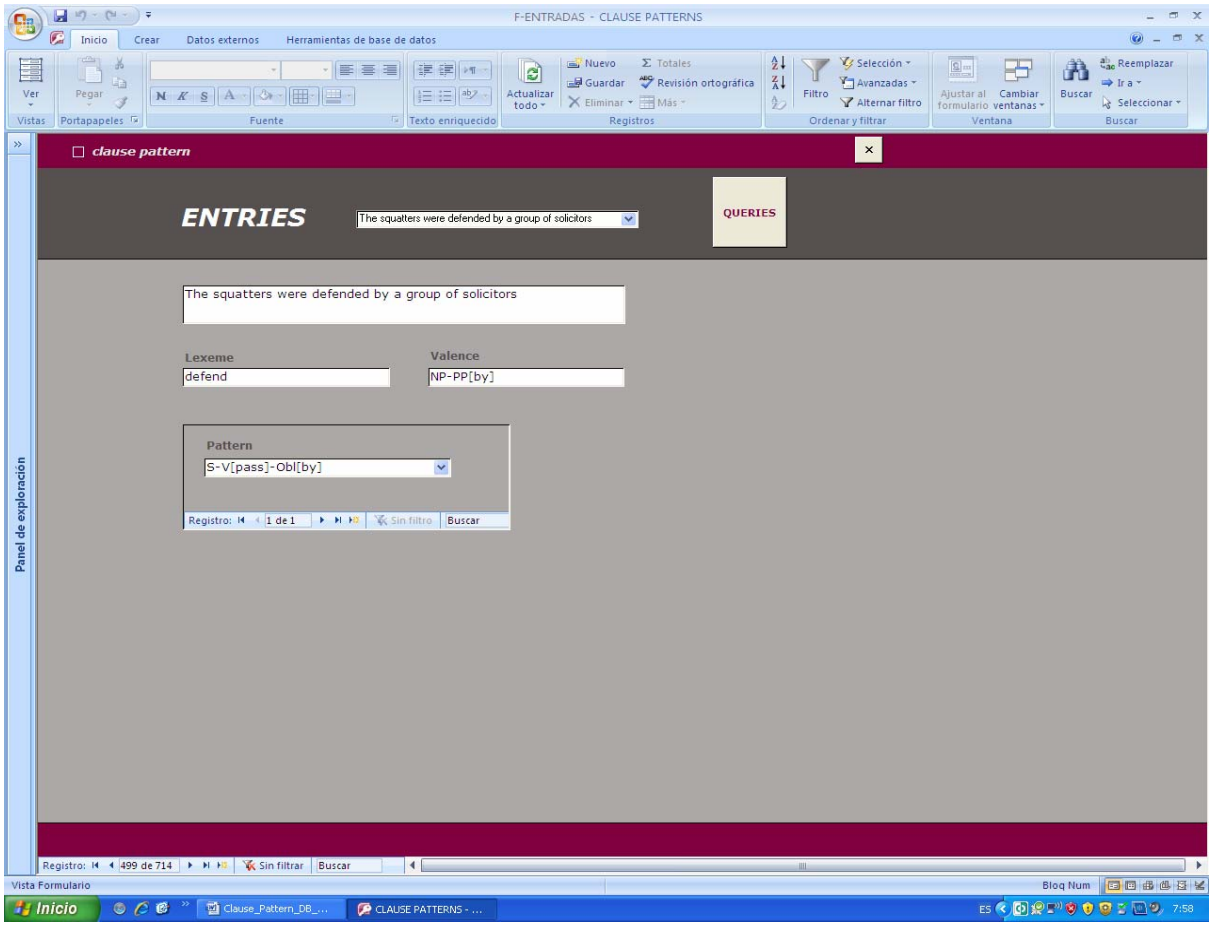


Figure 1: Entries

3. THE DATABASE

The database used has been developed by means of Microsoft Access. This database enables both database entries and pattern searches, which means that the user cannot only enter new entries and the information related to them but also perform queries in the same database. Thus, when entering the database, the user decides whether to include new information in the database or search for a pattern or lexeme by clicking on one of the two buttons on the main screen (Figure 2).

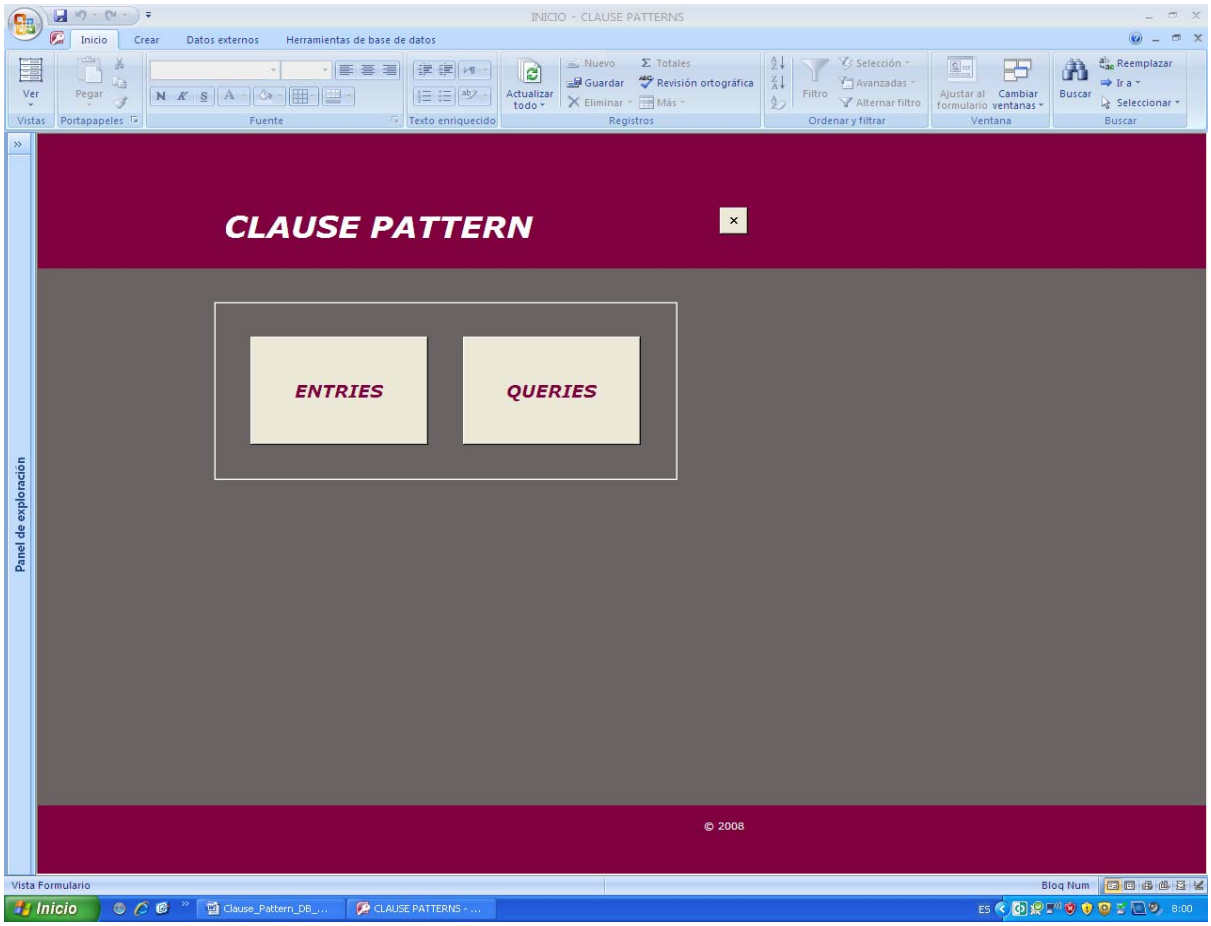


Figure 2: Main screen

If the user opts for the former, they will have to click on “Entries” and open a new register. In this new register the user must enter the sentence extracted from the corpus, the lexeme of the verb, the pattern exemplified in that sentence, with the possibility of identifying passive constructions, and the valence of the verb by means of the phrasal categories of the complements (Figure 1). The prepositional phrases allow the user to specify between brackets the preposition heading the phrase.

If the user decides to perform a query, they will have to click on “Queries” and a new screen will be displayed allowing them to search by “Lexeme” or “Pattern” (Figure 3).

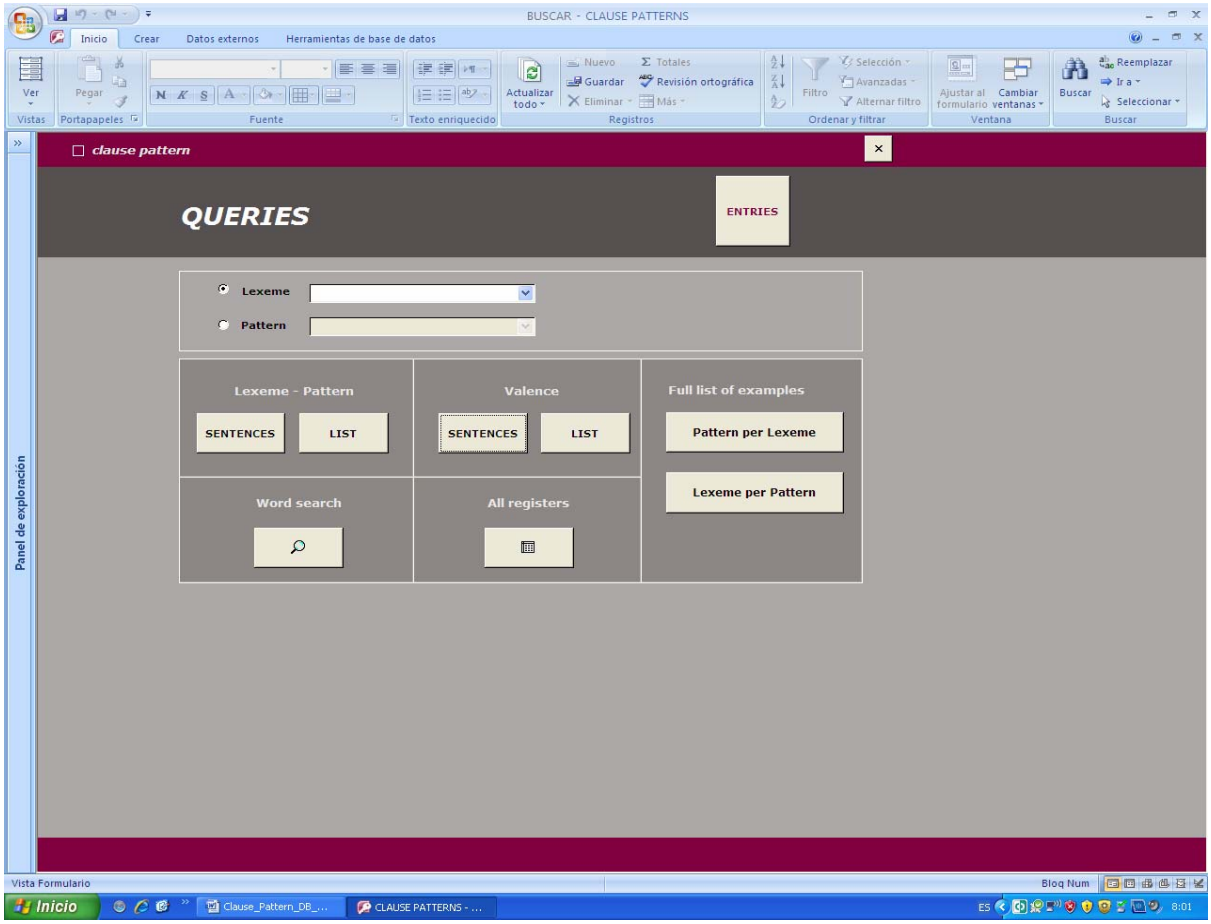


Figure 3: Queries

When searching by "Lexeme", a scroll down menu opens up with all the lexemes contained in the database. The user selects one lexeme and decides whether to look for the several patterns linked to this lexeme or its valence. When searching for the patterns linked to a specific lexeme, a list of the patterns and the corresponding sentences linked to the lexeme stored is displayed in the database (Figures 4 and 5) by clicking on the "Sentences" button. Alternatively, only a list of the patterns which belong to that lexeme stored in the database (Figures 6 and 7) can be obtained by clicking on the "List" button.

In order to exemplify these searches, a couple of semantically-related lexemes have been chosen: *give* and *donate*. From a semantic point of view these two verbs are closely related as both are classified as verbs of change of possession (Levin 1993). Even in recent resources aimed at computational applications such as the latest version of the lexical database Wordnet (Wordnet 3.0[4]), *give* is encoded as the direct hypernym of *donate*. However, from a syntactic point of view, they show a completely different behaviour in relation to the syntactic alternations that both verbs can take. Thus, *donate* cannot be found in

[4] http://wordnet.princeton.edu/

the double object construction, whereas *give* can display the dative alternation. As stated by Levin (1993), the failure of the former to show the SVOO pattern is due to its latinate origin.

As shown in Figures 4 and 6 the search performed with verb *give* shows 4 different patterns; SV[pass]OObl[by], SVO, SVOO, SVOObl and their corresponding examples:

1. SV[pass]OObl[by], a passive structure with an oblique introduced by the preposition *by*: *I was given this watch by my father.*

2. SVO, a monotransitive pattern: *She gave the lecture despite her illness*.

3. SVOO, a double object construction: *You've given Harry plenty of information*.

4. SVOObl, a dative structure: *Pat gave your phone number to Martin*.

However, the results for the verb *donate* (Figures 5 and 7) fail in showing the ditransitive alternation because this verb only accepts the SVOObl pattern. Here are some illustrative examples of the patterns in which the verb *donate* occurs:

1. SV[pass]Obl[to], a passive structure with an oblique headed by the preposition *by*: *Bodies are donated to The Body Farm*.

2. SVO, a monotransitive construction: *You never should have donated defective sperm in the first place*.

3. SVOObl, a dative structure: *The old professor donated all his books to the library*.

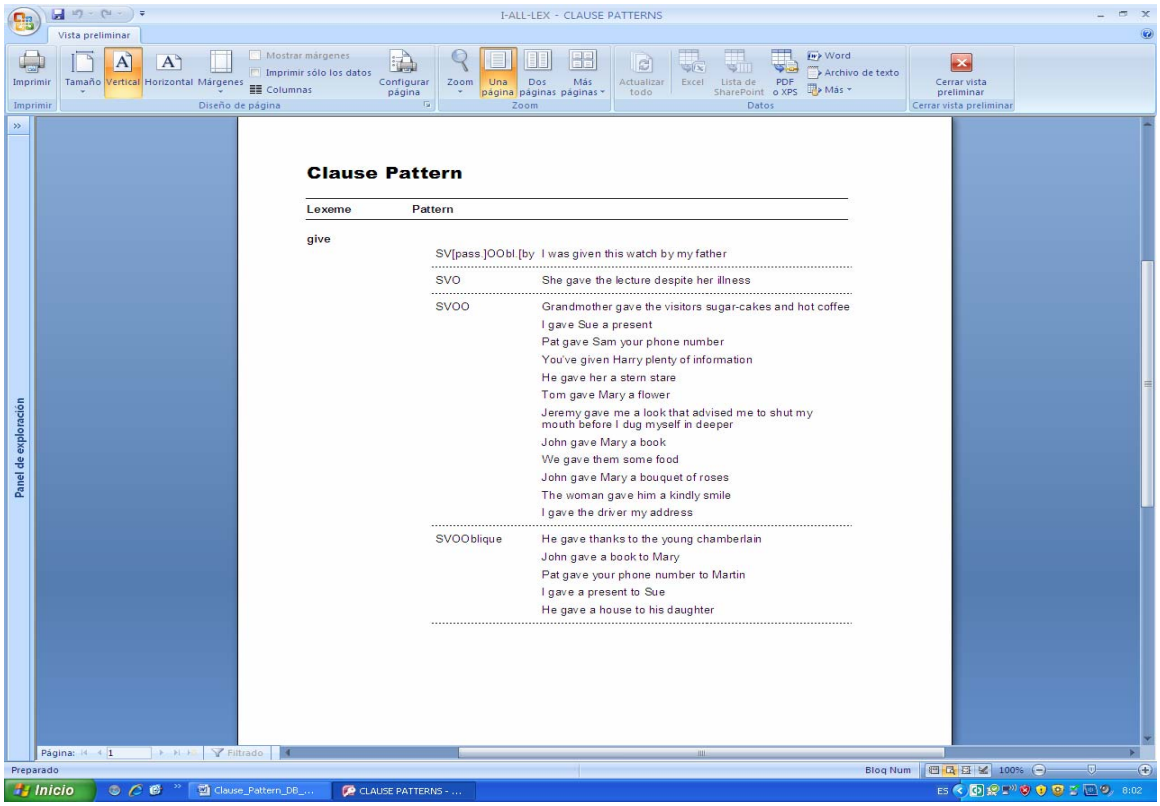


Figure 4: Patterns and examples for the lexeme *give*

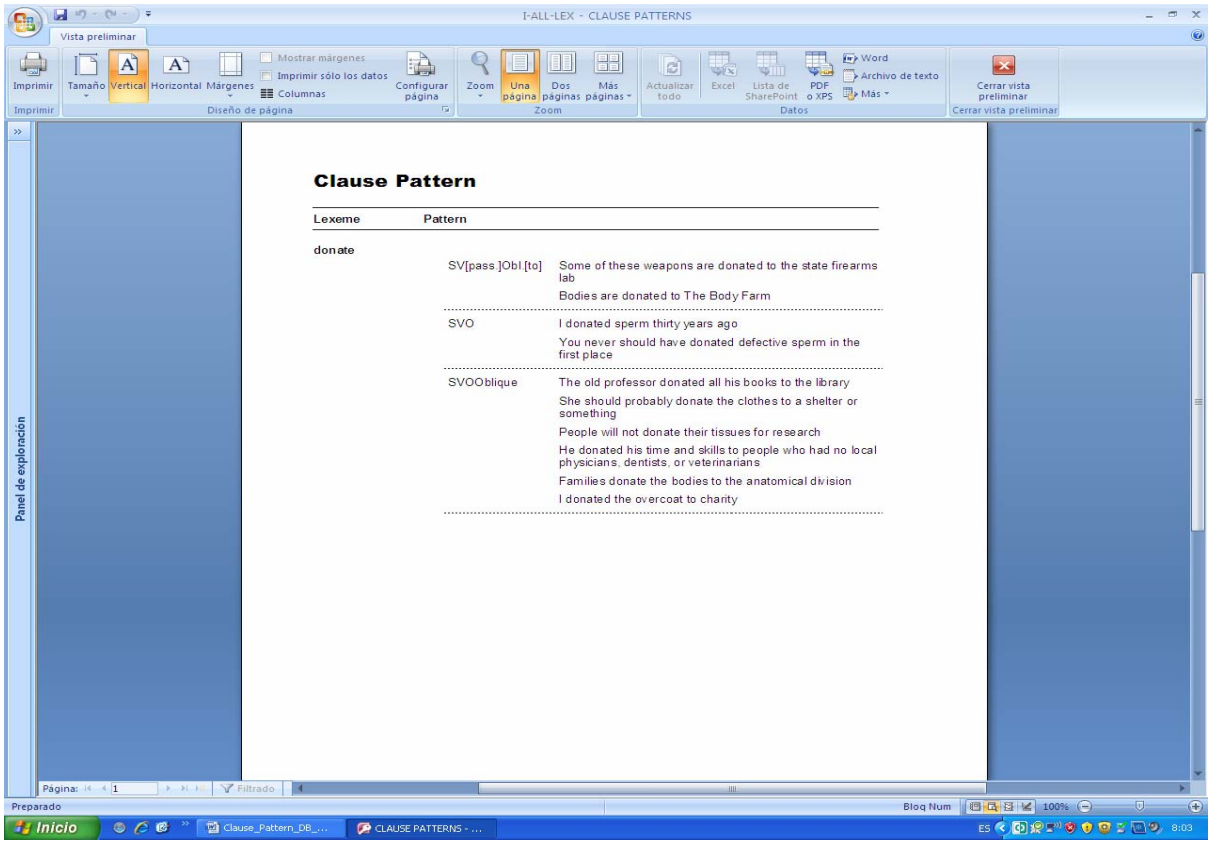


Figure 5: Patterns and examples for the lexeme *donate*

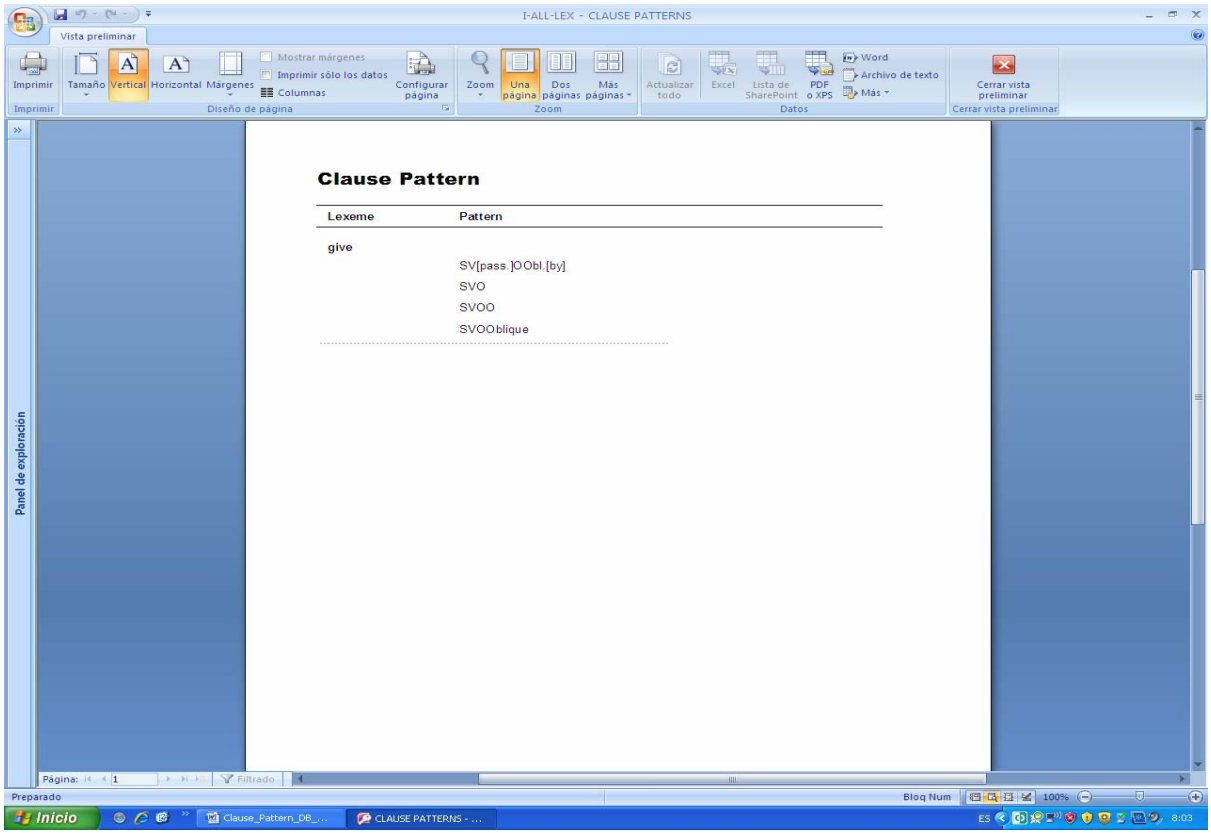


Figure 6: Patterns of the lexeme *give*

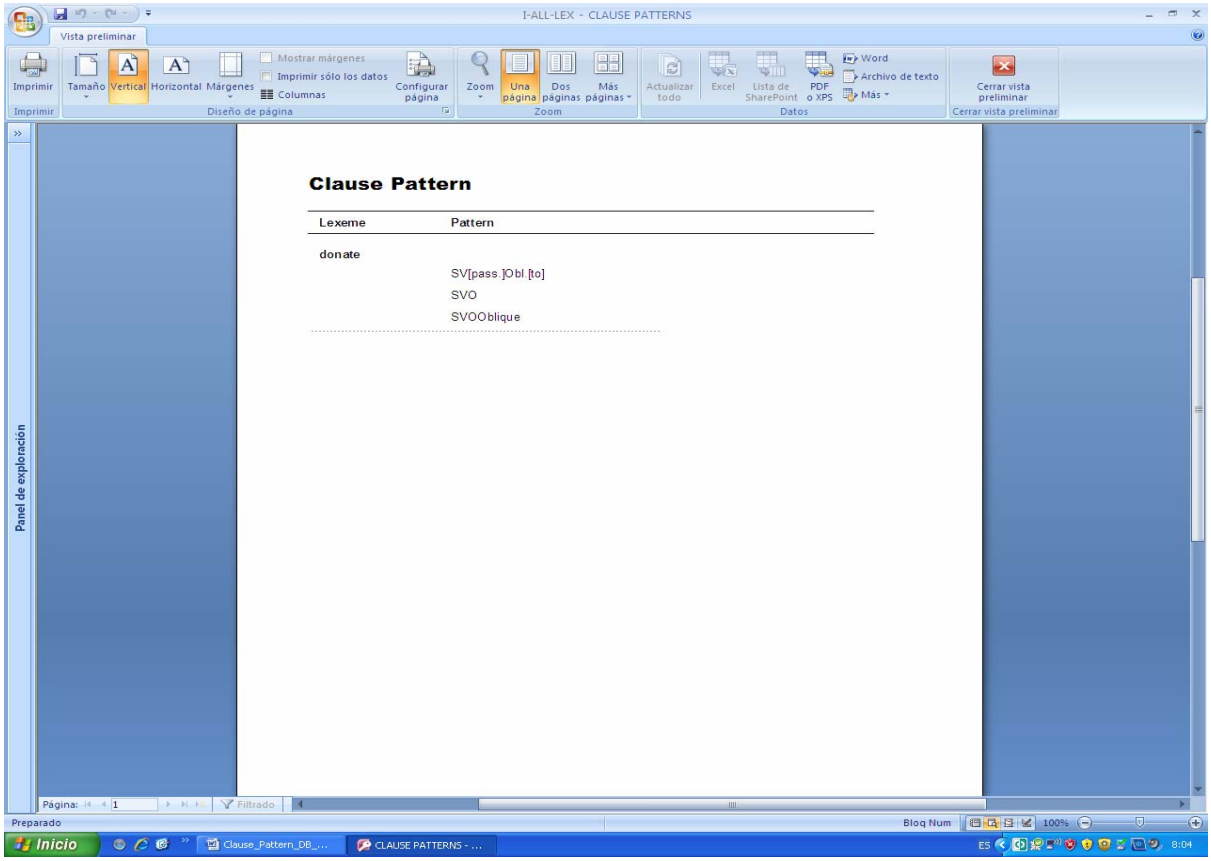


Figure 7: Patterns of the lexeme *donate*

When searching by valence, similar queries can be performed. Thus, the user can search for the valence and the sentences associated to one lexeme by clicking on the "Sentences" button (Figures 8 and 9). For instance, as can bee seen in Figure 8, the verb *give* has 4 possible valences:

1. NP-NP, two Noun Phrases, linked to a monotransitive construction: *She gave the lecture despite her illness*.

2. NP-NP-NP, three Noun Phrases related to a ditransitive construction: *The woman gave him a kindly smile*.

3. NP-NP-PP[by], two Noun Phrases and a *by*-Prepositional Phrase, which exemplifies a passive structure: *I was given this watch by my father*.

4. NP-NP-PP[to], two Noun Phrases and a *to*-Prepositional Phrase, which illustrates the dative alternation: *He gave a house to his daughter*.

On the other hand, the verb *donate* shows a wide range of patterns, being none of them a NP-NP-NP pattern. Instead, this verb is characterized by a dative construction, taking a *to*-Prepositional Phrase as part of its complementation:

1. NP-NP, a monotransitive structure realised by two Noun Phrases: *I donated sperm thirty years ago*.

2. NP-NP-PP[for], a dative construction realised by a *for*-Prepositional Phrase: *People will not donate their tissues for research*.

3. NP-NP-PP[to], a dative structure with a *to*-Prepositional Phrase: *The old professor donated all his books to the library*.

4. NP-PP[to], a passive structure with a *to*-Prepositional Phrase: *Some of these weapons are donated to the state firearms lab*.

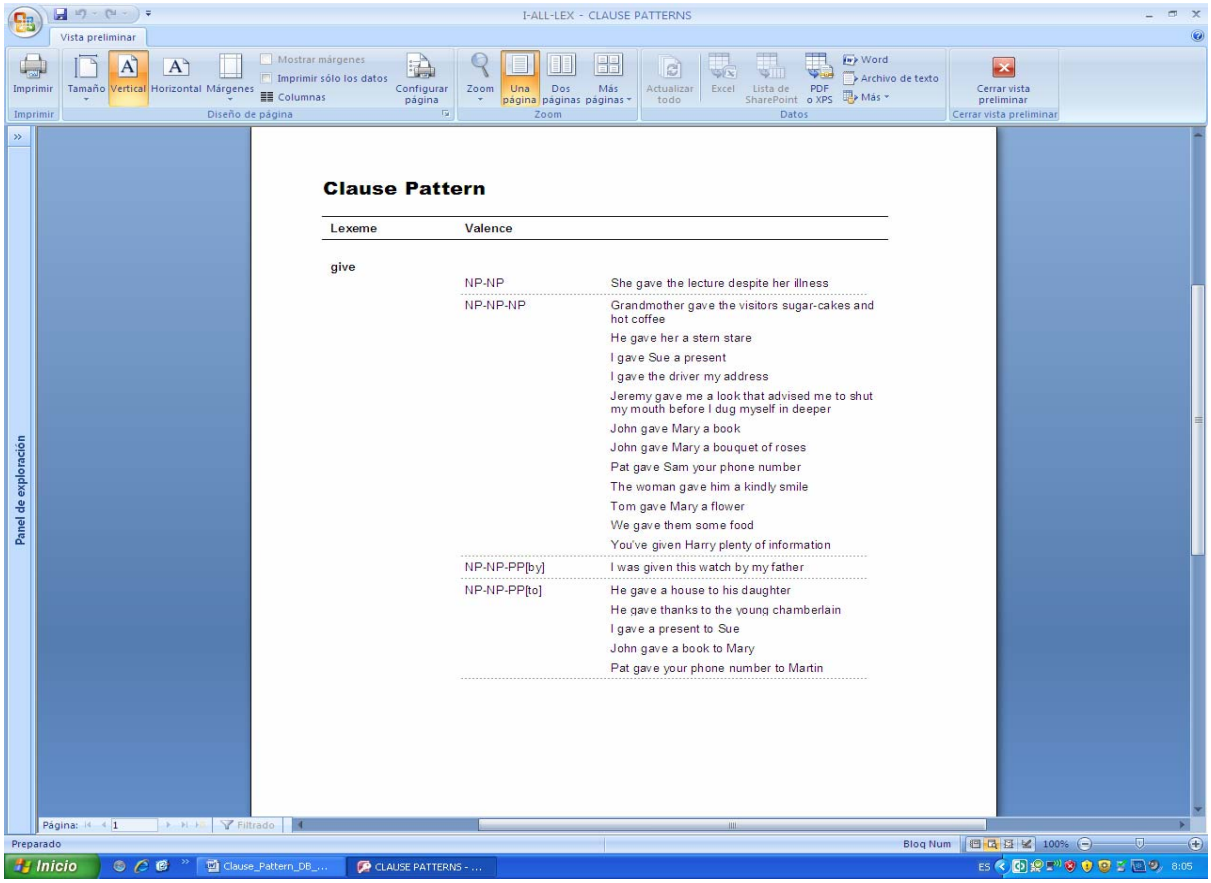


Figure 8: Valence and examples of the lexeme *give*

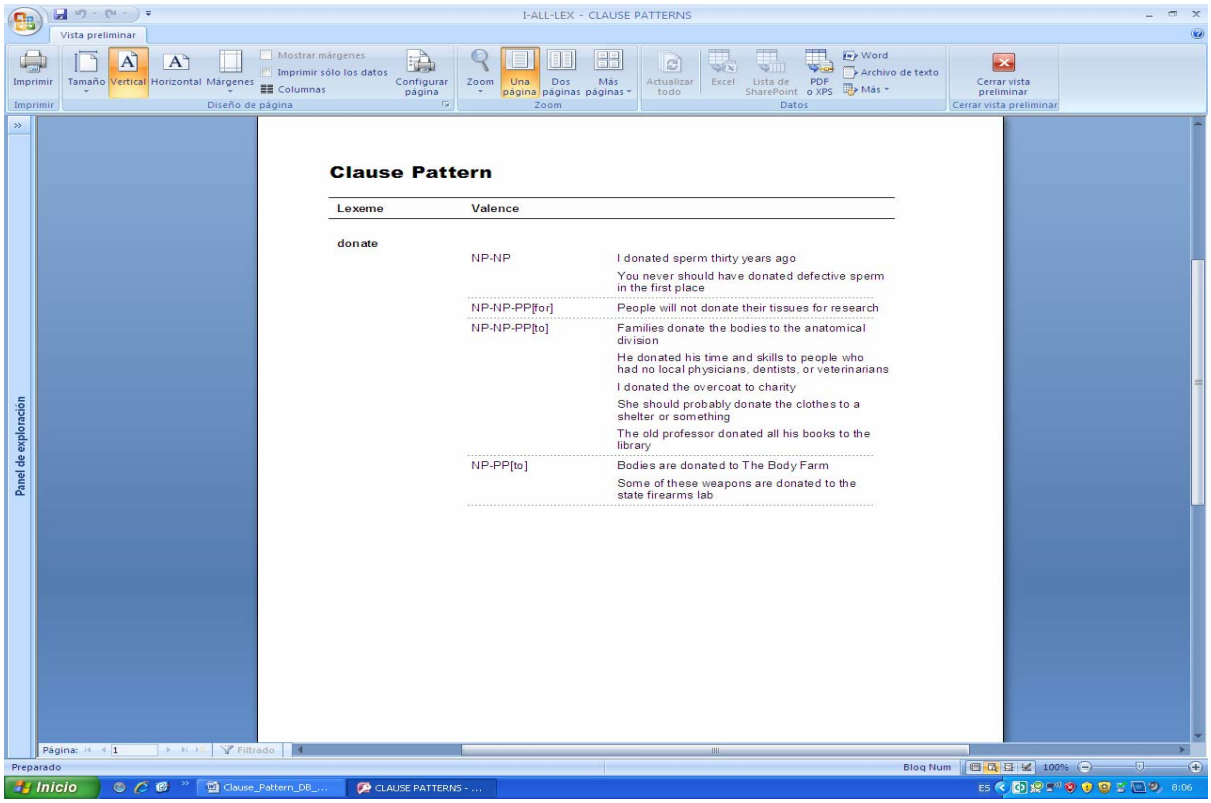


Figure 9: Valence and examples of the lexeme *donate*

As occurs when searching for the patterns which belong to a given verb, here the user also has the possibility to get only one list with the different valences (Figures 10 and 11) by clicking on the "List" button.

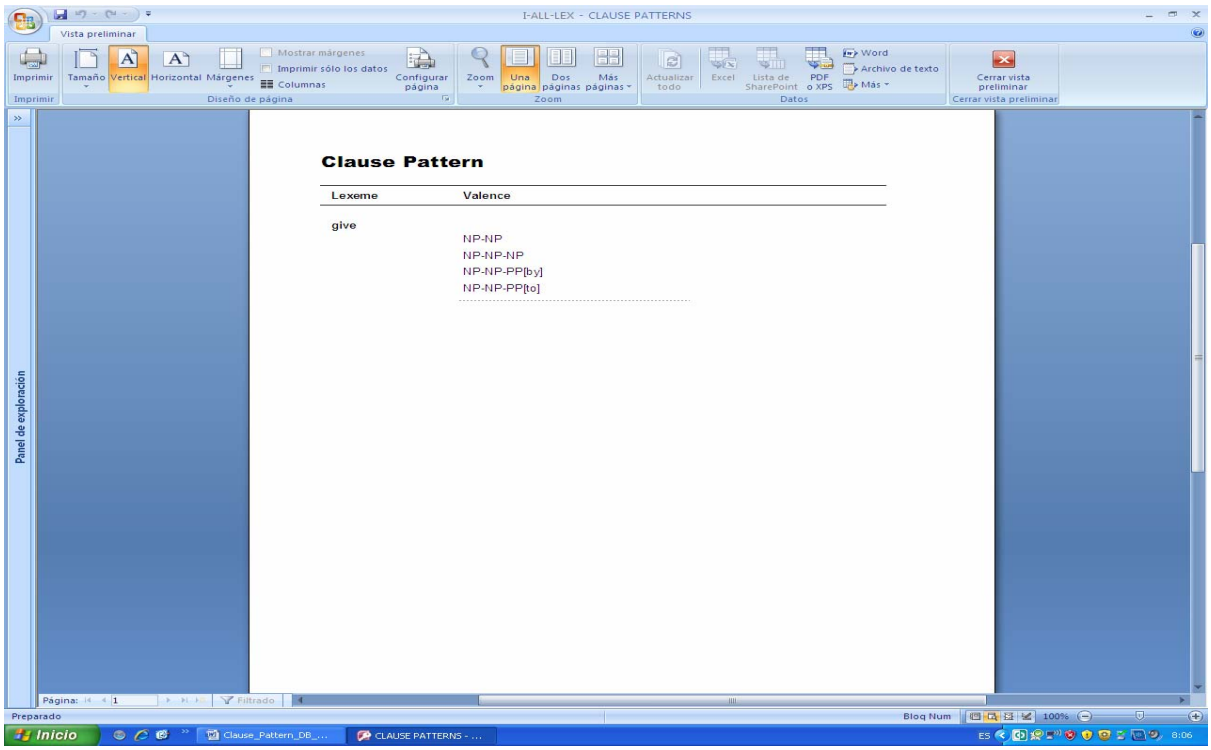


Figure 10: Valences of the lexeme *give*

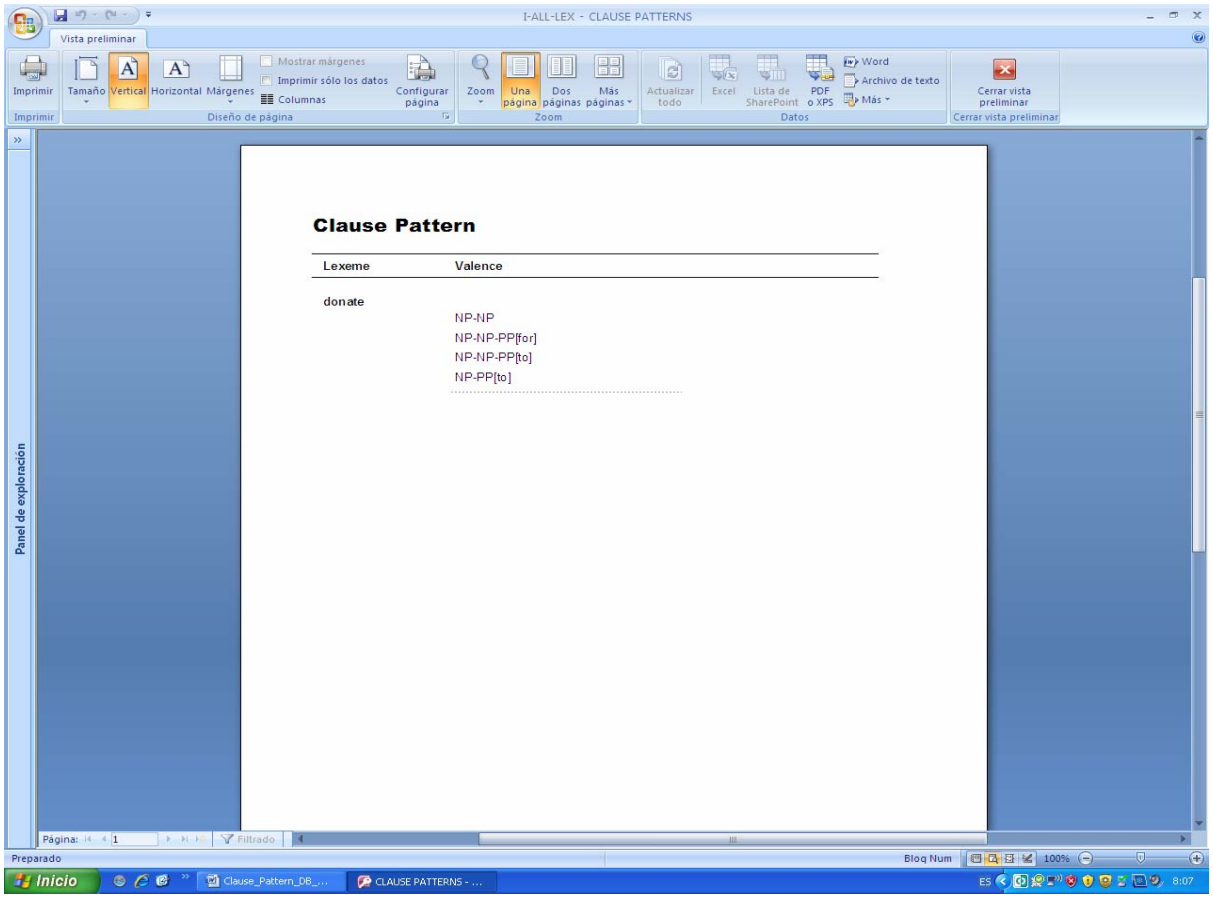


Figure 11: Valences of the lexeme *donate*

As well as searching by lexeme, there is also the possibility of searching by pattern. In order to perform this type of search, the user has to select one of the patterns appearing in the scroll down menu (Figure 12) and decide whether they want a list of lexemes and sentences containing this pattern (Figure 13), by clicking on the "Sentences" button under "Pattern", or only the lexemes related to this pattern (Figure 14), by clicking on the "List" button.

As shown in Figures 13 and 14 only those verbs that accept the SVOO pattern, such as *give*, are displayed, whereas other verbs which do not accept this pattern, such as *donate*, are omitted.

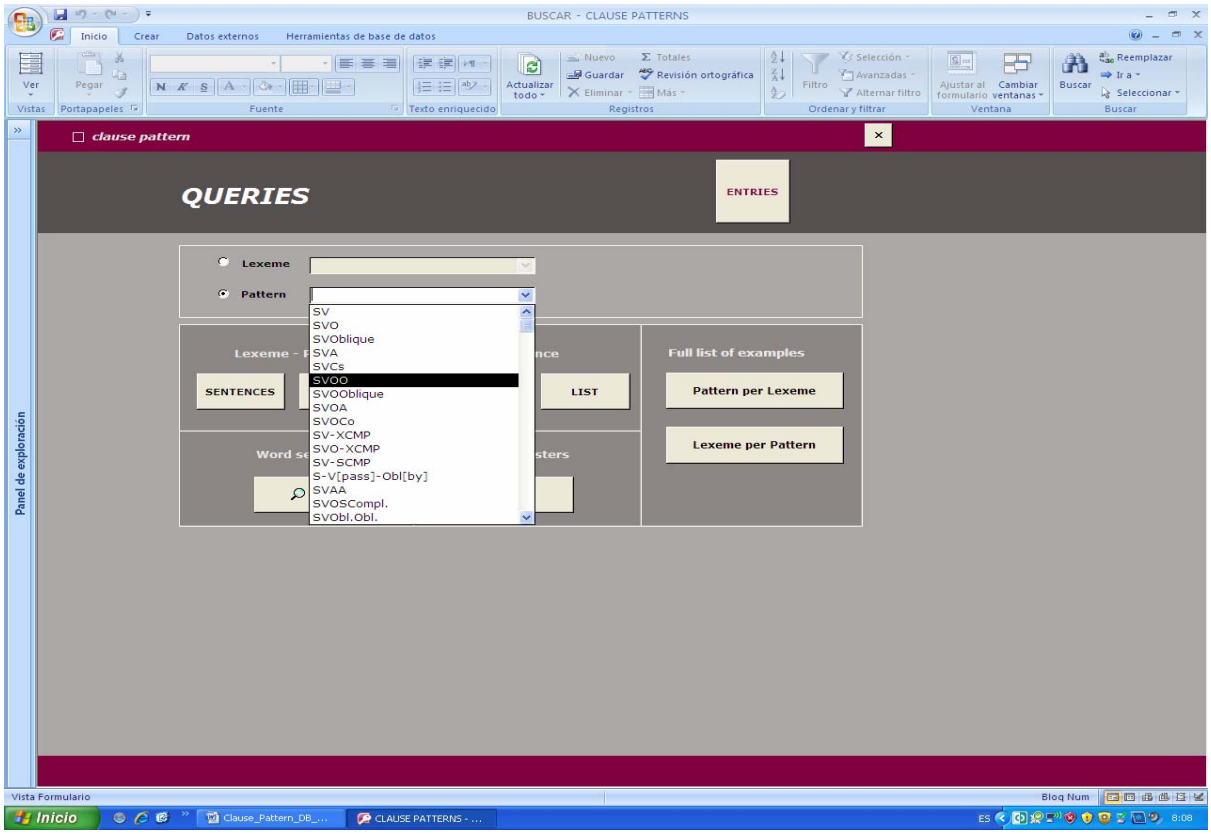


Figure 12: Pattern selection in the database entries

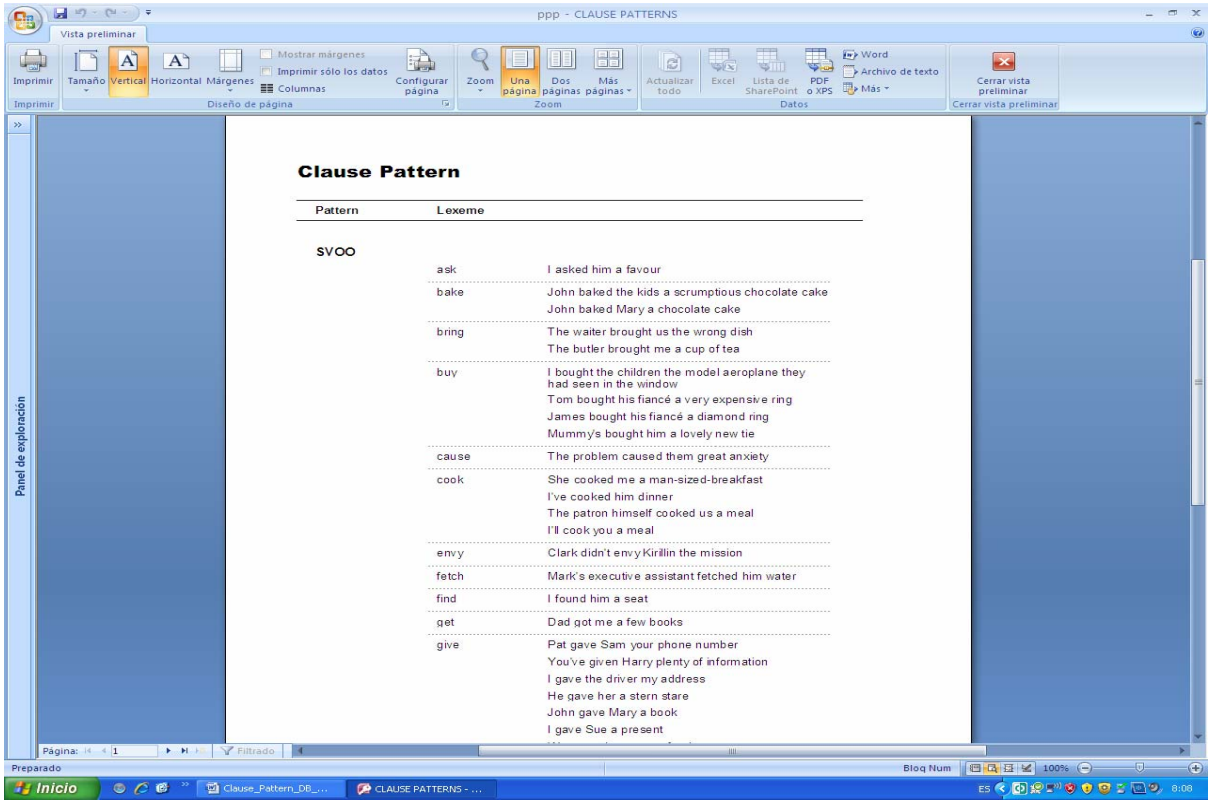


Figure 13: List of lexemes and sentences containing the SVOO pattern

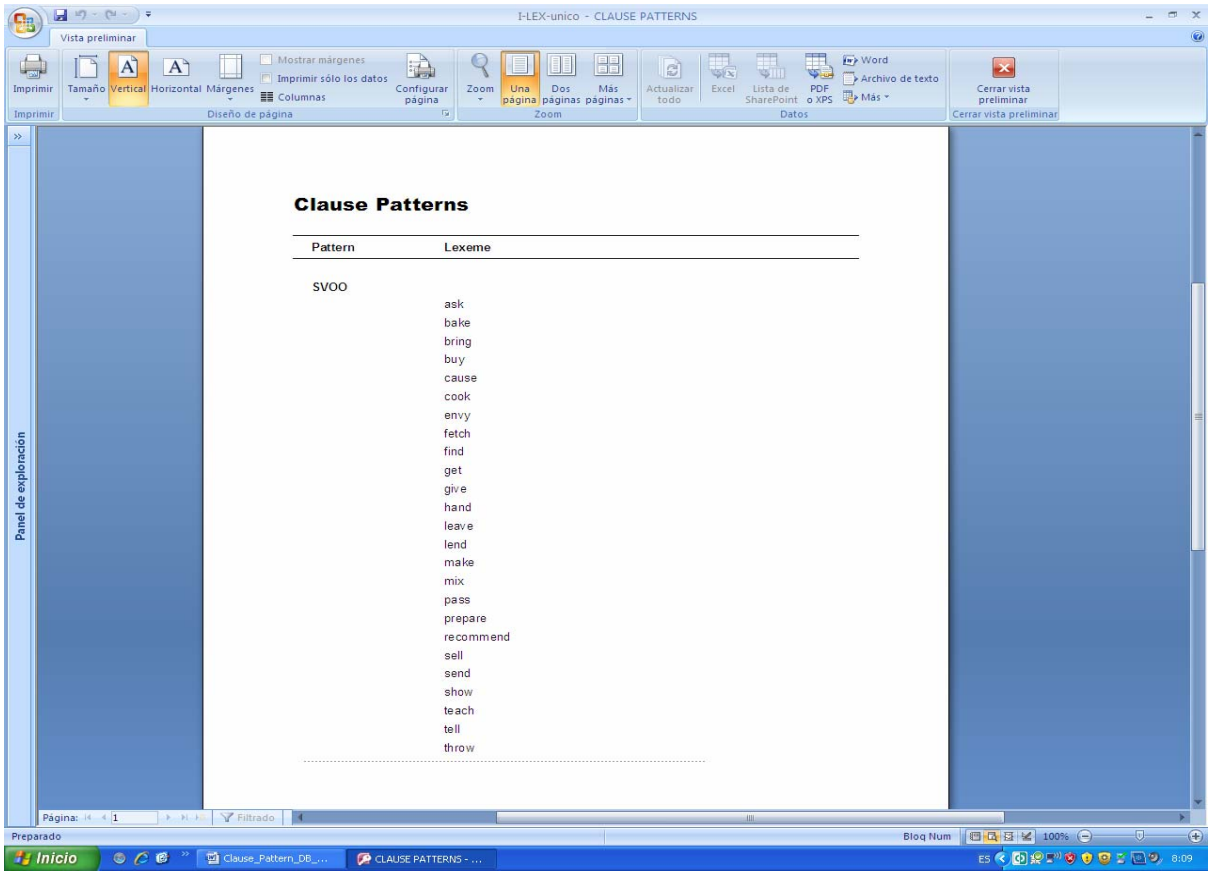


Figure 14: List of lexemes containing the SVOO pattern

Finally, the user can also obtain a full list of lexemes in two different formats. On the one hand, a complete list of all the lexemes stored in the database organised by pattern can be obtained by clicking on the "Pattern x Lexeme" button under "Full List of examples" (Figure 15). On the other hand, by clicking on the "Lexeme x Pattern" button a full list of all the lexemes in alphabetical order along with the patterns they accept is displayed (Figure 16). This has a great potential not only as a teaching resource but also as a research one. It displays lists that help the teacher find good examples and relevant sentences as well as help the researcher make some relationships between lexemes/patterns much more easily identifiable and therefore much more effective than intuition.

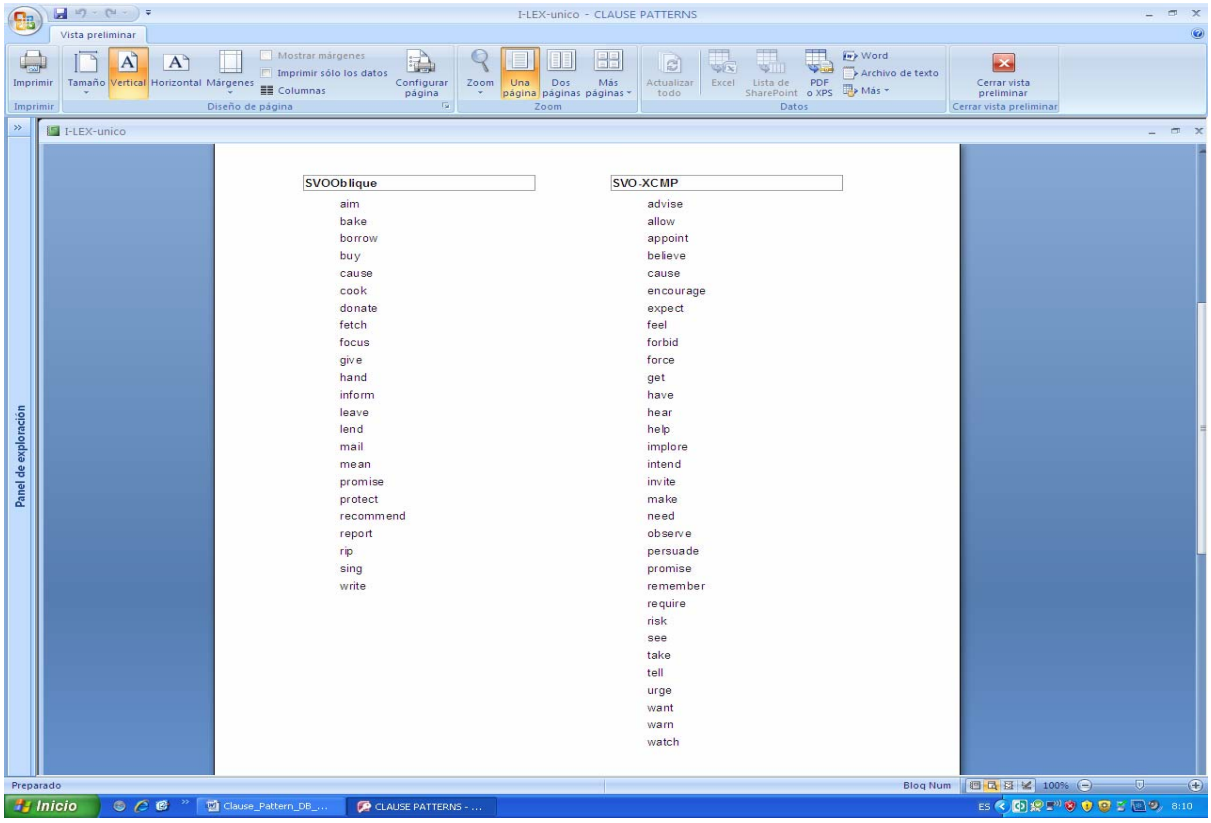


Figure 15: Full list of lexemes organised by pattern

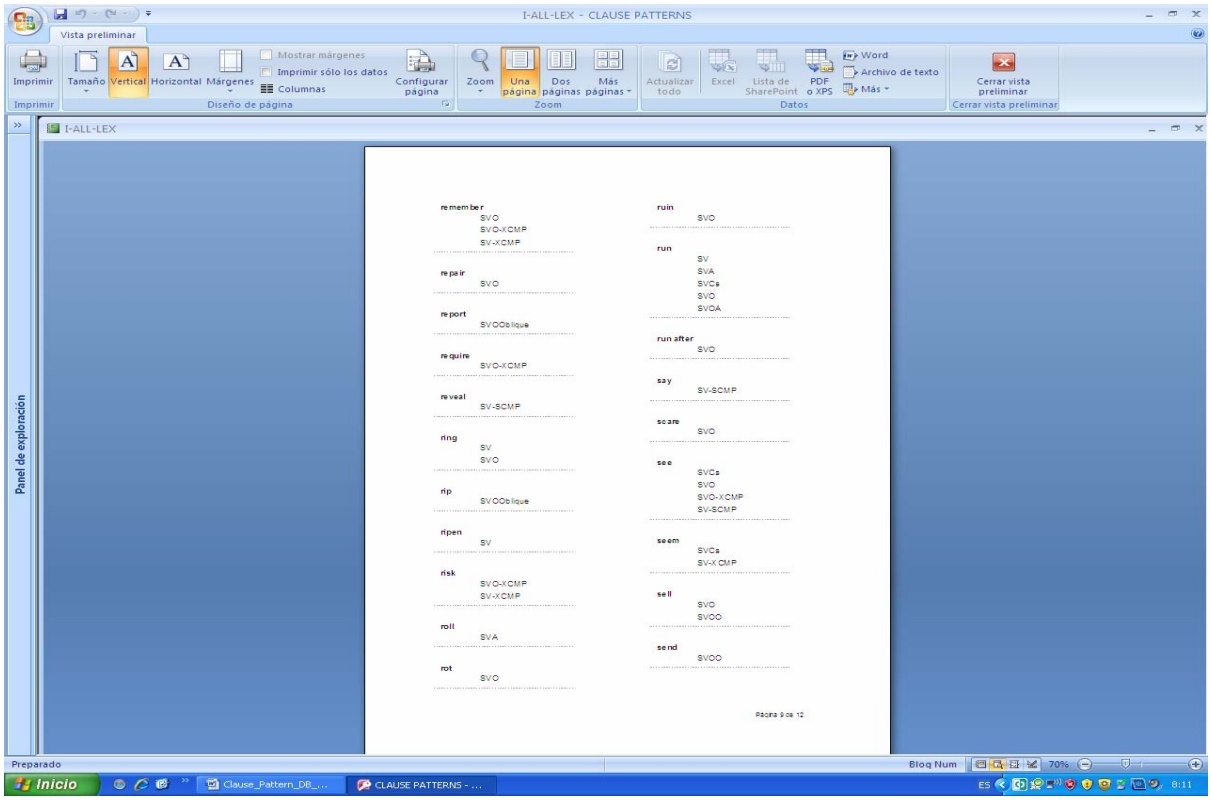


Figure 16: Full list of patterns organised by lexemes

The options presented here are the ones available so far, but the *Clause Pattern DB* has been conceived as a dynamic tool, allowing the introduction of new functions. At this moment we are currently working on planning to supplement this tool with the addition of tree structures connected to the sentences of the database. In addition, we are also considering the possibility of adding information about the frequency of occurrence of the different patterns of a given lexemes. This information would optimise the usefulness of the tool, as we are trying to provide learners not only with real language but also with the language they will need to use in real contexts.

4. CONCLUSION

As has been shown, the purpose and the utility of this tool is twofold: On the one hand, it allows users to organize and systematize the great amount of information which can be derived from a corpus, concentrating on certain aspects. On the other hand, it allows them to query and obtain information. The queries can be designed so as to draw the students' attention not only on the different syntactic patterns in English or the valence of verbs, but also on the interrelation of syntax and semantics, making them aware of the fact that semantically-related verbs tend to be associated with certain grammatical constructions. Thus, if searches are performed on lexemes, we can focus their attention on the difference in complementation patterns of the different meanings of polysemous words, but if we perform searches on patterns, we can also show them, in line with Francis and Hunston (2000), a different perspective, and concentrate on the classes of words which are selected by the different patterns.

Although much of the literature published on the use of corpora in the language classroom has stressed its effectiveness, Conrad (2005: 404) complains about the lack of empirical studies that demonstrate to what extent the application of corpus for the designing of classroom activities is helpful to second language acquisition or, at least, whether this approach is more effective than the use of traditional activities. To this respect, the materials we are currently designing as part of our teaching innovation project aim at bridging the gap between corpus research and language learning. We firmly believe that the former is a useful tool for the latter.

REFERENCES

Ädel, A. & Reppen, R. (Eds.). (2008). *Corpora and Discourse. The challenges of different settings*. Amsterdam/Philadelphia: John Benjamins.

Aijmer, K. (Ed.). (2009). *Corpora and Language Teaching*. Amsterdam/ Philadelphia: John Benjamins.

Conrad, S. (2005). Corpus Linguistics and L2 Teaching. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (pp. 393-409). New York: Lawrence Erlbaum Associates.

Francis, G. (1993). A Corpus-Driven Approach to Grammar. Principles, Methods and Examples. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and Technology. In Honour of John Sinclair* (pp. 137-156). Amsterdam/ Philadelphia: John Benjamins.

Gavioli, L. (2005) *Exploring Corpora for ESP Learning*. Amsterdam/Philadelphia: John Benjamins.

Granger, S. & Meunier, F. (Eds.). (2008). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam/Philadelphia: John Benjamins.

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

Hunston, S. & Francis, G. (2000). *Pattern Grammar.* Amsterdam/Philadelphia: John Benjamins.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Laso, N. J. & Giménez, E. (2007). Bridging the Gap between Corpus Research and Language Teaching. In C. Periñán (Ed.), *Revisiting Language Learning Resources*.(pp. 49-64). Newcastle: Cambridge Scholars Publishing.

Laso, N. J. & Giménez, E. (2008). Exploring the Impact of Concordancers in the Teaching of English Lexicology and Morphology. In L. Pérez, I. Pizarro & E. González (Eds.), *Estudios de Metodología de la Lengua Inglesa (IV)* (pp. 281-290). Valladolid: Universidad de Valladolid, Secretariado de Publicaciones e Intercambio Editorial.

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

Lüdeling, A & Kytö, M. (Eds.). (2008). *Corpus Linguistics. An International Handbook.* (Vol. 1). Berlin: Mouton de Gruyter.

Scott, M. & Tribble, C. (2006). *Textual patterns*. Amsterdam/Philadelphia: John Benjamins.

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

Tsui, A. B. M. (2004). What Teachers Have Always Wanted to Know – and How Corpora Can Help. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 39-61). Amsterdam/Philadelphia: John Benjamins.

Data-driven analysis on the weight of the explicit and implicit construct in ELT textbooks

RAQUEL CRIADO SÁNCHEZ

AQUILINO SÁNCHEZ PÉREZ

Universidad de Murcia

Abstract

It is generally accepted that there is a close relationship between explicit and declarative knowledge, and between implicit and procedural knowledge. Explicit is also associated with learning in which consciousness is engaged, while implicit is associated with acquisition, which takes place without a conscious control of the process. Research in explicit and implicit knowledge should have a counterpart in real language teaching and in the classroom, if research is to be the trigger of innovation and open new frontiers in ELT. The promotion of explicit or implicit knowledge (learning or acquisition of language) is necessarily present in the teaching materials and/or in the classroom action. As regards teaching materials, the minimal teaching unit par excellence, the exercise or activity, should reveal the nature of the type of knowledge aimed at. In this study a corpus with the activities of an ELT textbook was compiled. This was then analysed and systematised from the perspective of its explicit and implicit potential. The results show whether the textbook complies or not (and how much) with the requirements leading to explicit or implicit learning.

Keywords: explicit knowledge, implicit knowledge, explicit learning, implicit learning, acquisition, explicit teaching, implicit teaching, language teaching materials

Resumen

Existe un acuerdo generalizado sobre la estrecha relación existente entre conocimiento declarativo y explícito, y entre conocimiento procedimental e implícito. Explícito se asocia a aprendizaje, en el que se activa la consciencia, mientras que implícito se vincula a adquisición, cuyo proceso tiene lugar sin que intervenga la consciencia. Si partimos de la premisa de que la investigación anteriormente mencionada debe ser el motor de la innovación y abrir nuevas fronteras en la enseñanza de idiomas, es de esperar que los estudios resultantes se vean reflejados en el aula real y la enseñanza de lenguas extranjeras. El desarrollo del conocimiento explícito e implícito (aprendizaje o adquisición) se refleja ineludiblemente en los materiales para la enseñanza de lenguas y/o en la acción docente en el aula. Respecto a los primeros, la unidad mínima por excelencia es la actividad, la cual debe revelar la naturaleza del tipo de conocimiento que se pretende fomentar. En el presente trabajo hemos compilado un corpus de las actividades de un libro para la enseñanza del inglés como lengua extranjera. Posteriormente se analizaron y sistematizaron dichas actividades desde la perspectiva de su potencial para promover la enseñanza/aprendizaje de lo explícito o implícito. Los resultados muestran hasta qué punto el manual se ajusta a los requisitos necesarios para posibilitar el aprendizaje explícito e implícito.

Palabras clave: conocimiento explícito e implícito, aprendizaje explícito, aprendizaje implícito, adquisición, enseñanza explícita e implícita, materiales para la enseñanza de lenguas extranjeras

1. INTRODUCTION[1]

The terms *explicit* and *implicit* (knowledge) are nowadays at the centre of the paradigms in SLA. They correlate to two other key terms, *declarative* and *procedural* (knowledge), widely used in psycholinguistics and neurolinguistics, and are close to two other well-known terms in the Western tradition, *rationalism* and *empiricism* (Criado-Sánchez & Sánchez, 2009). *Explicit* and *implicit* are also heavily indebted to Krashen (1981). His dichotomy, *learning* vs. *acquisition*, is at the heart of SLA studies, initiated in the early 80s of the last century. Learning is typically associated with explicit, and acquisition with *implicit* (DeKeyser, 2003; Ellis, 2005; Hulstjin, 2005; Robinson, 1996; Schmidt, 1990, 1994, among others). During the last 30 years or so, research and discussion on L2 teaching and learning cannot be adequately understood unless we take into consideration the concepts and ideas developed around those pairs of words.

2. EXPLICIT AND IMPLICIT KNOWLEDGE AND LEARNING

There has been much discussion on the adequacy of claiming two kinds of knowledge, explicit and implicit. Some scholars (Shanks, 2003) argue that knowledge is a single entity and a single source of a varied performance rooted in the way retrieval takes place, while others (Anderson, 2005; Wallach & Lebiere, 2003) are strongly in favour of such a dichotomy. Anderson's model in particular has exerted a strong influence in SLA studies. His model claims a tight interaction and interplay between both types of knowledge and strengthened the strong-interface position, which considers declarative knowledge as a springboard towards implicit or proceduralised knowledge. The debate on this issue is still undecided, but the experience of adult language learning cannot but support the view that conscious learning is an important element in the acquisition of knowledge, even if the details on how this takes place are still blurred.

Dörnyei (2009) warns about the profuse coverage of the terms *explicit/implicit* and the confusion that may derive from it, since they are applied to different concepts: knowledge (*implicit/explicit knowledge*), learning (*explicit/implicit learning*) and memory or information storing (*explicit/implicit memory*). The meaning of *explicit/implicit* keeps the core features in

[1] This research is financially supported by the Spanish Ministry of Science and Innovation, research project (Ref.: FFI2009-07722), funded by the *Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica.* Dr. A. Sánchez is professor and Dr. R. Criado-Sánchez associate professor at the University of Murcia, Spain (*Lacell* Research Group, http://www.um.es/grupos/grupo-lacell/index.php). E-mail addresses: asanchez@um.es; rcriado@um.es.

the three uses, particularly that concerning the role of consciousness, but its application to knowledge, learning or memory results in important differences regarding the end-product.

With this caveat in mind, we will use the term *explicit* as implying consciousness, awareness and reasoning, while *implicit* excludes consciousness, or conscious control of the processes performed. More specifically, explicit learning generates explicit knowledge and facilitates explicit memory, and implicit learning generates implicit knowledge and facilitates implicit memory. *Explicit knowledge* is the kind of knowledge we may have access to, knowledge we can *declare*, we are aware and conscious of. This type of knowledge can therefore be verbalised and consciously controlled (Ellis, 1994; Paradis, 2009; Schmidt, 1990). *Implicit knowledge*, on the other hand, is the kind of knowledge we are not conscious about. This kind of knowledge does not require our conscious collaboration as it is automatic and proceduralised; once triggered, it proceeds automatically until it comes to a final goal (Anderson, 2005; Hulstjin, 2005; Hulstjin & De Graff, 1994; Schmidt, 1993a, 1993b).

Finally, we assume here the parallelism between explicit-declarative and implicit-procedural. Accordingly, *declarative knowledge* (DEC) or *knowledge-that*, is believed to keep the same properties as explicit knowledge, and *procedural knowledge* (PRO) or *knowledge-how*, the same properties as implicit knowledge.

3. THE TEACHING MATERIALS: THE IMPLICIT AND EXPLICIT CONSTRUCT

Matching the way of teaching with effective learning will first require that both processes agree in that they are guided by the same principles and run somehow parallel. If we focus our analysis on the role of explicit and implicit teaching and their probable influence on explicit and implicit learning, it is meaningful to analyze the degree of explicitness and implicitness of the teaching materials. Even though it is well known that students quite often learn what has not been taught, while sometimes they may actually learn what they have been instructed to learn (Lewis, 1996; Willis & Willis, 2001), the expectations are that learning by adult students in instructed acquisition will run parallel to the teaching action deployed. We may reasonably assume that explicit materials will favour explicit learning and implicit materials will result in more implicit learning. We do not attempt, however, to take a stand for explicit or implicit teaching. We just aim to offer data based on real teaching materials regarding the amount of explicit or implicit learning they may promote through the activities they offer.

The goal we pursue here requires the identification of the features and characteristics through which we will be able to decide on whether a specific activity promotes implicit or explicit learning/teaching. For that purpose, a reliable diagnosis of the *implicit/explicit* construct must count first with the tools necessary to perform such a task. The description of explicitness and implicitness in the previous sections guides us in this task.

Since the teaching action in textbooks is typically based on activities or exercises, they can be taken as the units for analysis. Activities have their own structure and the potential for promoting explicit or implicit knowledge depends on the nature of their constituent elements, which are (i) the goal they aim at, and (ii) the activated means in order to reach such a goal, that is, the strategies deployed. Accordingly, the identification and analysis of the goal and the strategies of each activity are decisive in detecting if they have been designed for promoting implicit or explicit learning.

The explicit and implicit constructs are not neatly shaped and delimited. This brings with it an additional problem: the question of whether it is necessary or convenient to use a scale in detecting the degree of explicitness or implicitness. From our point of view, the convenience of such a scale derives from the fact that the activities analysed will probably subscribe to all, none or some of the features defining explicit or implicit teaching. The pedagogical action *per se* (instructed acquisition) tends to introduce some kind of explicitness, whether deductive or inductive. On the other hand, whenever implicitness is pursued in the classroom, the materials are very often pedagogically arranged to match the (implicit) goals previously defined. In that case, the input is previously 'manipulated' and includes many instances of a particular lexical or grammatical point even though no explicit information is given on the underlying target forms. Moreover, explicit teaching is also accompanied by abundant practice, which favours proceduralisation, that is, implicit learning.

The scale applied here is organised along a continuum from 0 to 10, with 10 being the maximum and 0 the minimum (total absence of implicit or explicit character in the activity). In case both constructs are present to some extent in the same activity, the total of the unit is always 10.

With this in mind, the scheme of the analysis and assessment of each activity regarding its potential for favouring explicitness or implicitness is adjusted to the study of the following characteristics:

Table 1: Features of implicit and explicit learning in activity goals and strategies

Type of learning promoted/favoured	***Features of activities (promoting/favouring each type of learning)***
Implicit learning	**1. Activity goals,** 1.1. do not require awareness of nor demands attention to form 1.2. offer linguistic input or triggers output aiming at proceduralisation 1.3. favour focus on meaning (and not on the form) 1.4. aim at fluent and efficient communication (productive and receptive) 1.5. offer genuine and authentic materials **2. Activity strategies,** 2.1. centre on the transmission of meaning (meaning centred/oriented) 2.2. do not require controlled and conscious processing of language (spontaneous practice/speech, varied responses, free use of target forms) 2.3. require the use of the target language 2.4. are interactive 2.5. 2.5. favour proceduralisation through meaningful repetition and frequent instances of use
Explicit learning	**3. Activity goals,** 3.1. look for awareness on formal aspects of language 3.2. offer declarative knowledge on the language 3.3. demand explicit attention to specific forms 3.4. aim at accuracy in the use of the language, or at the explicit understanding of language use 3.5. offer materials artificially arranged around grammar/structural aspects/vocabulary items/pronunciation **4. Activity strategies,** 4.1. centre on learning of form 4.2. demand explicit knowledge or controlled processing for performance (controlled practice, target structures, unvarying responses) 4.3. do not require the use of native language 4.4. do not aim at really interactive communication 4.5. favour declarativisation: ask for/give/points at memorisation of grammatical explanations/rules/vocabulary items/pronunciation

The implicit and explicit constructs are built around two axes: the **goals** and the **strategies** guiding each activity. We also try to detect and identify each type of knowledge through opposing features, that is, marking the presence or absence of specific features, as is for example the case of 'awareness'. This procedure allows for a clear and functional analysis.

Regarding the goals, five questions are examined:

(i) Awareness: explicit knowledge requires awareness of the linguistic elements being introduced or practiced, while implicit knowledge does not.

(ii) Activities offer some kind of input, be it for presentation or for practice. If the input is explicit, declarativisation prevails; if not, the input will favour proceduralisation.

(iii) Meaning centred activities leave form aside and attention to linguistic elements is not favoured. Hence, implicit learning is more likely to occur.

(iv) Fluency in communication is appropriate for natural language use and therefore favours proceduralisation (implicit learning). Accuracy in the use of forms usually implies explicit information of the language and is connected to declarative knowledge (explicit).

(v) Authentic and genuine materials are guided by communicative goals, because this is the natural use of language. Implicitness is the most likely outcome. Materials selected according to formal criteria emphasise specific formal elements, so explicitness is therefore be favoured.

Regarding the **strategies** through which the goals may be attained, five questions are also analysed:

(i) Strategies requiring attention to meaning (what to say) promote implicit learning, since the formal elements are not emphasised.

(ii) Strategies which require conscious control of the linguistic elements used in communication favour declarative knowledge (explicit).

(iii) The use of the target language is a necessary ingredient for intensive exposure to the language and proceduralisation.

(iv) Interactive events are appropriate for real communication, hence implicit learning is favoured. Communicative events centred on accuracy and form can sometimes be de-contextualised (explicit).

(v) Practice (be it repetitive or not) when it is meaningful, offers instances of input for proceduralisation (see Sánchez and Criado-Sánchez, in press). Mechanical practice leads to declarativisation because it is centred on form (structures).

4. THE ANALYSIS OF A TEACHING UNIT

The teaching materials were taken from a textbook for teaching English as a foreign language: *New English File Elementary Student's Book*. The textbook is structured in 9 files or units, each of which contains 4 subfiles (A, B, C and D). We selected a unit from the second half of the book, file 7. The selection was done at random.

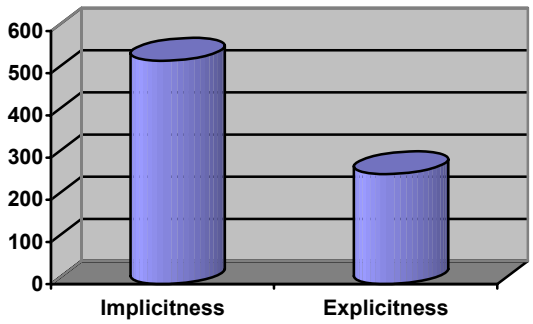
This unit contains several sections focused on the four skills and the grammar, vocabulary and pronunciation sub-skills, with a total of 79 activities. The activities were carefully analysed following the criteria mentioned in section 3. Due to space restrictions, a sample of the analysis carried out in one of the four subfiles of the unit, subfile 7B, is shown in Table 2. The results below refer to the data resulting from the analysis of the 79 activities of the whole unit. The weight of explicitness and implicitness in each activity is given in numbers within the scale 0-10. The sum of the figures in each column represents the weight of each one of the constructs in the unit.

Table 2: Sample analysis (subfile 7B)

Activity: promotes/favours….	***Implicit (…/10)***	***Explicit (…/10)***
B Pronunciation **18 1.a.** Listen and repeat the sounds and words. [Features met: (1.2; 2.3) (3.1, 3.3, 3.4, 3.5; 4.1, 4.2, 4.4, 4.5)	2	8
19 1.b. Listen and practice the dialogue [Features met: (1. 2; 2.3) (3.1, 3.3, 3.4, 3.5; 4.1, 4.2, 4.4, 4.5)]	2	8
Speaking **20 2.a.** Read the introduction and the questionnaire. [Features met: (1,1, 1.2, 1.3, 1.4, 1.5; 2.1, 2.2, 2.3, 2.3, 2.4, 2.5)]	10	0
21 2.b. In pairs, interview your partner. Who drinks more water? [Features met: (1.1, 1.2, 1.3, 1.4, 1.5; 2.1, 2.2, 2.3, 2.3, 2.4, 2.5)]	10	0
Grammar **22 3.a.** Complete the questions with *how much* or *how many.* [Features met: (1.2; 2.3) (3.1, 3.3, 3.4, 3.5; 4.1, 4.2, 4.4, 4.5)]	2	8
23 3.b. Match the sentences and pictures. [Features met: (1.2; 2.3) (3.1, 3.3, 3.4, 3.5; 4.1, 4.2, 4.4, 4.5)	2	8
24 3.c. Grammar Bank. a) Complete with How much/How many. [Features met: (2.3) (3.1, 3.2, 3.4, 3.5; 4.1, 4.2, 4.3, 4.4, 4.5)]	1	9
25. 3.c Grammar Bank. b) Cross out the wrong words. [Features met: (2.3) (3.1, 3.2, 3.4, 3.5; 4.1, 4.2, 4.3, 4.4, 4.5)]	1	9
26 3.d. Complete the questions with *how much* or *how many.* [Features met: (2.3) (3.1, 3.2, 3.4, 3.5; 4.1, 4.2, 4.3, 4.4, 4.5)]	1	9
27 3.e. In pairs, ask and answer. Answer with an expression from d or a number. [Features met: (1.2, 1.3; 2.1, 2.3, 2.4, 2.5) (3.1, 3.4, 3.5; 4.2)]	6	4
Reading **28 4.a.** Cover the magazine article *Water – facts and myths*. In pairs, look at these questions. Can you answer any of them? [Features met: (1,1, 1.2, 1.3, 1.4, 1.5; 2.1, 2.2, 2.3, 2.3, 2.4, 2.5)]	10	0
29 4.b. Read the article. Put the questions in ***a*** in the gaps [Features met: ((1,1, 1.2, 1.3, 1.4, 1.5; 2.1, 2.2, 2.3, 2.3, 2.4, 2.5)]	10	0
30 4.c. Read the article again. Match the highlighted words with these phrases. [Features met: (1.1, 1.2, 1.4, 1.5; 2.1, 2.3, 2.3, 2.4) (3.3; 4.5)]	8	2
31 4.d. Look at the questions in ***a*** again. In pairs, answer them from memory [Features met: (1,1, 1.2, 1.3, 1.4, 1.5; 2.1, 2.2, 2.3, 2.3, 2.4, 2.5)]	10	0
32 4.e. Is there anything in the article you don't agree with? [Features met: (1,1, 1.2, 1.3, 1.4, 1.5; 2.1, 2.2, 2.3, 2.3, 2.4, 2.5)]	10	0
Weight of explicitness/implicitness in subfile 7B	**85**	**65**
TOTAL weight of explicitness/implicitness in the unit	**529**	**261**

5. ANALYSIS OF DATA

The prevailing implicit character of the activities in the whole unit analysed in the previous section is evident: out of a total of 790 features, 529 favour implicit learning, while only 261 favour explicit learning. The weighting given to implicitness almost doubles that granted to explicitness (Graph 1).

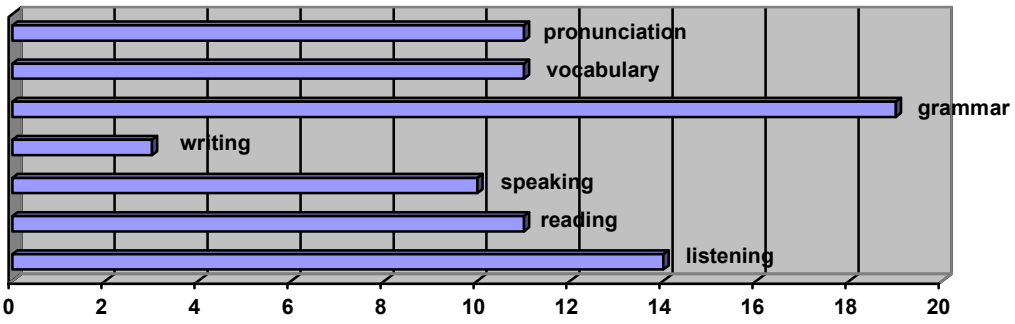


Graph 1: Total weight of implicitness vs. explicitness as revealed in the activities

This fact is not innocuous from a methodological point of view. It reveals that teaching is clearly biased towards a 'natural approach', that is, exposure to and practice with the input as a way of learning is given more emphasis than learning about the language or just practising with (artificially arranged) phrases or forms. Additionally, the input and practice associated with it are most often centred on meaning, close to authentic materials and real use of language, and aim at communicative functions appropriate to language within communicative settings or situations. On the other hand, grammar, form or information on the linguistic system, rules, abstract and explicit explanations of the language are not avoided but are given a secondary role in the overall work promoted by the activities. The result of this global appraisal is that the prevailing method is clearly communicative (CLT), even though elements from other methods emphasising the formal aspects of language are also present.

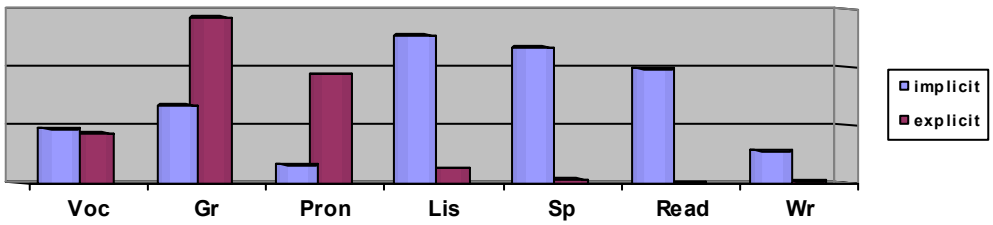
This global conclusion should be nuanced, however. Since the file is structured in partially autonomous sections and each section promotes a specific skill or sub-skill, the weight of the construct explicit/implicit in the sections by necessity reveals whether their distribution of this construct is homogeneous or varies depending on the nature of each one of the sections. To begin with, Graph 2 shows the total of activities assigned to each one of the sections or skills/sub-skills. It is important to take this information into account. The data

shows a clear imbalance in the amount of activities reserved for listening (14), for example, versus the activities reserved for writing (3), while the activities for teaching grammar (19) are significantly higher than those assigned to vocabulary teaching (11) and pronunciation (11).



Graph 2: Number of activities per skill and sub-skill

The proportion of the number of activities devoted to each skill or sub-skill does not correspond however to the weight of explicit or implicit features they actually favour. There are important differences regarding the presence of explicitness and implicitness in each section, regardless of the number of activities in each section. A case in point is the section devoted to grammar: the percentage of activities centred on grammar within the whole file is 24.05%. However, within this same section the weight of the implicit construct reaches only 13.37% of the features within the whole subfile, while the promotion of explicitness reaches 44.08%. The four skill sections show a homogeneous distribution though, with a neat and clear prevalence of the implicit construct in all of the skills (see Graph 3, where Voc, Gr, Pro, Lis, Sp, Read and Wr respectively stand for Vocabulary, Grammar, Pronunciation, Listening, Speaking Reading and Writing). In the same graph, the data shows that teaching of the three sub-skills (grammar, vocabulary and pronunciation) abound in explicitness, while the teaching of the four skills is very poor in this respect. Within the three sub-skills, only vocabulary shows a balance in the weight of implicitness and explicitness. The unbalance in favour of explicitness is evident in the teaching of grammar and pronunciation. Similarly, the imbalance in favour of implicitness is evident in the sections of listening, speaking, reading and writing.



Graph 3: Weight of the features of the implicit and explicit constructs per skill and sub-skill

The distribution of the explicit/implicit constructs across skills and sub-skills shows what should probably be considered a traditional belief in the role of skills and sub-skills. Skills are approached from a global and comprehensive perspective. Skills are conceived with meaning playing a leading role, since they are by necessity holistic and meaning centred. In that sense, it can be affirmed that the prevalence of implicitness is a logical outcome, since explicitness tends to stress specific linguistic elements (form) at the expense of meaning, which would be assigned a secondary role.

Sub-skills show a more complex picture. Grammar has been traditionally associated with methods displaying a heavy emphasis on the teaching of form. The higher index of explicitness is therefore to be expected in the grammar section, as is the case in the unit analysed here. Pronunciation, however, can be said to have been present throughout the history of teaching. The traditional grammar-translation method emphasised correct pronunciation, and mostly emphasised correct grammar and vocabulary use. But the same stress is found in natural methods, from the Direct to the Communicative Method. Vocabulary knowledge shares a similar role across all methods. In the unit analysed, however, only the vocabulary section keeps a balance in the role assigned to explicitness and implicitness. Pronunciation is taught here with a heavy emphasis on the role of explicit knowledge.

6. CONCLUDING REMARKS

Learning is a complex process, a fact which is clearly acknowledged in SLA. The analysis of this unit from the perspective of implicit/explicit teaching reveals a similar situation. Activities throughout the unit favour both the acquisition of explicit and implicit knowledge, although the weight of implicitness almost doubles that of explicitness. The unbalance of implicit vs. explicit load pervades seven of the eight sections structuring the unit. The

unbalance in favour of implicitness is condensed in the activities for teaching the four skills, while the sections reserved for teaching the sub-skills are unbalanced toward explicitness, particularly in the case of grammar and pronunciation. Research and findings in favour of implicit teaching appear in this unit in close association with traditional practices and convictions.

No doubt, recent trends in SLA may have exerted a decisive influence on the methodological and pedagogical design of this unit. In this case, it proves that research in language acquisition and learning does affect teaching and the elaboration of teaching materials. If this is so, the kind of analysis carried out here deserves more attention, in the sense that it brings together the application of SLA insights to FLT research and practice.

REFERENCES

Anderson, J. R. (2005). *Cognitive Psychology and its Implications*. (6th ed.). New York: Worth Publishers.

Criado-Sánchez, R. & Sánchez, A. (2009). The Universal Character of the *DEC−>PRO* Cognitive Sequence in Language Learning and Teaching Materials. *RESLA, 22,* 89-106.

DeKeyser, R.M. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313-348). Cambridge, Mass.: Blackwell.

Dörnyei, Z. (2009). *The Psychology of Second Language Acquisition*. Oxford: Oxford University Press.

Ellis, R. (1994). *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. A psychometric Study. *SSLA*, 27, 141-172.

Krashen, S. (1981). *Second language acquisition and second language learning.* Oxford, England: Pergamon.

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second language learning: Introduction. *Studies in Second Language Acquisition*, *27,*(2), 129-140.

Hulstijn, J. H. & De Graff, R. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review, 11,* 97-110.

Lewis, M. (1996). Implications of a lexical view of language. In D. Willis & J. Willis (Eds.), *Challenge and Change in Language Teaching* (pp. 10-16). Oxford: Heinemann.

Oxenden, C., Latham-Koening, C. & Seligson, P. (2004). *New English File Elementary Student's Book.* Oxford: Oxford University Press.

Paradis, M. (2009). *Declarative and Procedural Determinants of Second Languages*. Amsterdam: John Benjamins.

Robinson, P. (1996). *Consciousness, Rules and Instructed Second Language Acquisition*. New York: Peter Lang.

Sánchez, A. & Criado-Sánchez, R. (in press). Cognitive underpinnings of repetitive practice in the learning of EFL. *Proceedings of the XXVII AESLA International Conference*. Ciudad Real, 2009.

Schmidt, R. W. (1990). Consciousness in second language learning. *Applied Linguistics, 11*(2), 129-158.

Schmidt, R. (1993a). Awareness and second language acquisition. *Annual Review of Applied* Linguistics,13, 206-226.

Schmidt, R. (1993b). Consciousness, learning, and interlanguage pragmatics. In G. Kasper & S. Blum-Kulka (Eds.), *Interlanguage pragmatics* (pp. 21-42). Oxford: Oxford University Press.

Schmidt, R. (1994). Implicit learning and the cognitive unconscious: Of artificial grammars and SLA. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 165-209). London: Academic Press.

Shanks, D. (2003). Attention and awareness in 'implicit' sequence learning. In L. Jiménez (Ed.), *Attention and Implicit Learning* (pp. 11-42). Philadephia. PA: John Benjamins.

Wallach, D. & Lebiere, C. (2003). Implicit and explicit learning in the unified architecture of cognition. In L. Jiménez (Ed.) *Attention and Implicit Learning* (pp. 215-252). Philadelphia. PA: John Benjamins.

Willis, D. & Willis, J. (2007). *Doing Task-based Teaching*. Oxford: Oxford University Press.

On the history of the Old English prefix sam-

ISABEL DE LA CRUZ CABANILLAS

Universidad de Alcalá

Abstract

The objective of the present paper is to trace the historical development of the Old English prefix sam-, which declined in productivity and eventually ceased to be productive at all in Present-Day English. For the Old English period, the data have been obtained from the database Nerthus, a comprehensive tool that allows the retrieval of records taking as basis different criteria. All in all, the information extracted from this database will be complemented by consulting various relevant lexicographic works. Likewise, different Middle English dictionaries as well as the Oxford English Dictionary will be used to examine the evolution of the prefix through the later periods of the history of English. Finally, the Helsinki Corpus will be searched as well.

Keywords: Old English prefixes, affixation, productivity, competition

Resumen

El objetivo del presente trabajo es estudiar la evolución histórica del prefijo sam- en ingles antiguo. Dicho afijo vio reducida su productividad paulatinamente y finalmente ha dejado de ser productivo en inglés contemporáneo. Para el periodo de inglés antiguo, los datos se han obtenido de la base de datos Nerthus, una herramienta completa que permite la extracción de registros mediante criterios diversos. Con todo, la información obtenida de esta base de datos se complementa con la consulta de varias fuentes lexicográficas relevantes para el periodo. Igualmente se recurre a diferentes diccionarios de inglés medio, así como al Oxford English Dictionary, que ayudará a trazar la evolución experimentada por el prefijo a través de los periodos posteriores de la historia de la lengua inglesa. Finalmente, se consultará también el Corpus de Helsinki.

Keywords: Prefijos en ingles antiguo, afijación, productividad

1. INTRODUCTION[1]

The present paper discusses the evolution undergone by the Old English prefix sam- up to the present moment. Some Old English prefixes were widely used in the first stages of the history of English, but with the passing of time they declined in productivity and eventually ceased to be productive at all in Present-Day English. Several prefixes have reduced their productivity, but have already been treated by other scholars (De la Cruz, 1975; Fraser, 1985; Hiltunen, 1983; Martín Arista, 2005-08, among others). Thus, for the present study, we concentrate on the analysis of one of the prefixes which has attracted little attention, namely sam-.

To carry out the analysis various sources have been consulted. For the Old English period, the data have been obtained from the database *Nerthus* and complemented with

[1] The research reported here has been funded through the project FFI2008-04448/FILO. My thanks are due to Dr Martín Arista who provided me with the information retrieved from the database *Nerthus*.

information from some relevant lexicographic works. The database *Nerthus* has been implemented by Martín Arista and his team at University of La Rioja in order to provide a detailed description of the Old English lexicon. It allows the retrieval of records taking as basis the criterion of prefix, which is an obvious advantage, as the researcher does not need to filter the data provided by traditional dictionaries. The lexicographic works used for its compilation were originally Clark Hall's *A Concise Anglo-Saxon Dictionary* and Sweet's *The Student's Dictionary of Anglo-Saxon.* It is now being enlarged by adding information from Bosworth-Toller's *An Anglo-Saxon Dictionary,* both the 1898 edition and the additional *Supplement* (1921).

As Martín Arista explains (2005-08: 211), the database provides a full description of the derivational morphology of the predicates along with the inflectional morphology relevant for derivation:

Nerthus analyses each headword in terms of more than sixty variables, which are grouped in three blocks of information: predicate (including category, form, variants, translation, inflectional morphology, type of predicate and morphological process), derivation and compounding (both including canonical, non-canonical, inflective, and phonologically-modified base and adjunct).

As the implementation of the database is still in progress, the number of records will vary depending on the moment when the reference was written. One of the latest figures was provided by Torre, Martín, Ibañez, González and Caballero (2008) and by Martín Arista (forthcoming). For the present study we take into account those data offered by the latter scholar. Thus, currently, *Nerthus* contains 29,987 predicates. Out of these, 16,690 are nouns, 5,785 adjectives, 5,618 verbs, 1,654 adverbs and 240 records correspond to other minor categories, such as adpositions,[2] conjunctions, pronouns, interjections, numerals, possessives and demonstratives/articles. If one puts into relation the figures of *Nerthus* with the ones of the published letters of the *Dictionary of Old English* (*DOE* hereafter), the following comparison can be established:

[2] The research team implementing *Nerthus* prefers the term *adposition* to *preposition*, as some of the words within this category may appear in the postfield, rather than preceding the noun or pronoun.

Table 1: *Nerthus* predicates by letter compared to entries in *DOE*

A 1,376 (D O E 1,505)	N 568
æ 581 (D O E 617)	O 1,288
B 1,853 (D O E 2,202)	P 296
C 1,074 (D O E 1,637)	R 565
D 634 (D O E 89 7)	S 2,866
E 1,155 (D O E 1,450)	T 990
F 2,537 (D O E 3,014)	Þ 747
G 2,629 (GE-) 1,457	U 1,90
H 2,489	V 1
I 370	W 2,224
L 974	Y 255
M 1,268	

Taking into consideration the fact that the *Dictionary of Old English* includes just the first letters of the alphabet, *Nerthus* proves to be a valuable instrument for the linguistic analysis of the period. All in all, it shows shortcomings, most of which are encountered by any scholar that decides to carry out any kind of research on historical periods of the English language. On the one hand, even its compilers admit that "*Nerthus* requires more formalised meaning definitions of Old English lexical items" on which the research team is working now (Martin Arista, 2005-08: 228). It is also true that *Nerthus* is a database that is not linked so far to a textual corpus. And thirdly, it is still in progress, which means some data may be revised, changed and enlarged in the near future. On the other hand, corpora on Old English do not abound. Even if the achievements of the *Helsinki Corpus* were magnificent at the time it was compiled, its size turns out to be insufficient for some specific linguistic analyses, as will be shown below.[3] Somehow, Bosworth and Toller (1898 and 1921) tried to compensate this lack by providing sentences to illustrate a given entry. Likewise, the *Thesaurus of Old English* serves to establish the semantic networks of the lexical units, but offers no information about their definition, use or inclusion in texts. Nonetheless, as long as the *Dictionary of Old English* is not finished, the research has to be carried out by making use of the tools available at the moment.

For the Middle English period, the information is retrieved from the *Middle English Dictionary* (*MED*, henceforth) and Stratmann's *Dictionary*, while the *Oxford English Dictionary* (*OED*, hereafter) is consulted as a supplementary guide all throughout the history of English, from its beginnings up to the present state. As long as the academic community cannot resource to a textual corpus that comprises all the texts available for a given period as

[3] See De la Cruz Cabanillas (2007) on this particular issue.

well as electronic tools that can retrieve the information easily, researchers must constraint to the material at their disposal.

2. DISCUSSION OF THE RESULTS

The chosen prefix was present in Old English in various word classes. According to the data provided by Nerthus, the prefix distributes over the following lexical categories: adjective, adverb, noun and verb, as shown in figure 1.

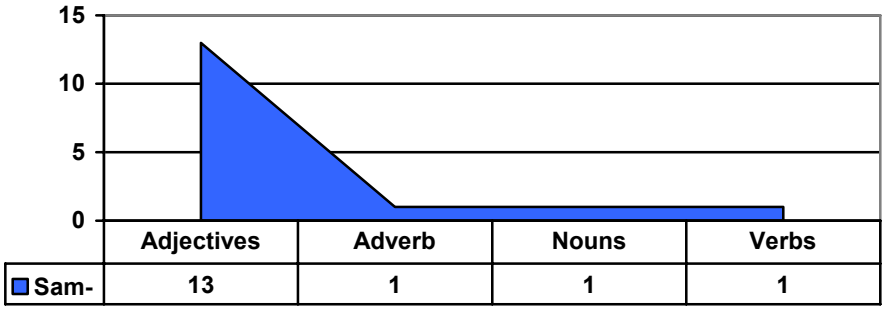


Figure 1: Predicates in *Nerthus* distributed by word class

If we have a look at the kind of word class, one can observe that it is mainly found in adjectives whose meaning has been modified by the affix in the sense of 'half, partial'. In fact, in the etymological section of the *OED* for the entry *half*, it is explained that in "OTeut. *halþ-* appears to have been a later substitute for the original *sāmi-*, OE. *sam-*, as in OHG. *sâmiquec*, OS. *sâmquic*, OE. {*samcwic*} half alive, so *sambærnd* half-burnt, *sambrice* a half-breach; = L. *sēmi-* in *sēmidoctus*, *sēmivīvus*, *sēmicoctus*, *sēmideus*, *sēmihomo*, etc.;" Thus, it seems as if the prefix *sam-* had gradually been replaced by *half*. That could explain why the survival of these elements in Middle English is scarce, as, out of the 16 elements, just 3 are recorded in the *Middle English Dictionary*, namely, *samehale*, *samlered*, *samsoden*.[4] However, *samlæred* is included with a single quotation dating from 1150 from the *Early English Homilies from the Twelfth Century*. All in all, there are new derivatives in the ME period, namely, *sam-crisp*, *samded*, *samred*, *samripe*, *samcloth.* It follows from here that the affix *sam-* is still productive in this period, although the only predicate that survives into Early Modern English is *samsodden*.

[4] None of the terms beginning with the prefix *sam-* is attested in Stratmann (1891).

Table 2: Old English prefix *sam-* across the centuries

Sam- in *Nerthus*	Dates of quotations in *MED*	Dates of quotations in *OED*	Meaning
sambaerned			half-burned
samboren			Born out of due time
sambryce			Partial breach
samcwic, *samcucu*, *somcwic*			Half-dead
samgrene		1000	Half-green, immature
samgung			young
samhal	1400	1000-1023	unwell
samlæred	1150		half-taught
samlocen			Half-closed
sammelt			Half-digested
samsoden	1450	*Samsodden*, 1000-1891	Half-cooked
samstorfen			Half-dead
samswæled			Half-burned
samweaxen			Half-grown
samwis			Stupid, dull
samwyrcan			To half do a thing
	samcloth, 1450-1500		Loin cloth, apron
	sam-crisp, 1500	1425	Somewhat curly
	samded, 1325	1297	Half-dead
	samred, 1400-1525	1393-1440	Half-red
	samripe, 1400		Partly ripe

The evidence provided by the *Helsinki Corpus* has not been included in the table for not being considered significant. In fact, out of the 16 predicates retrieved from *Nerthus*, just 4 items were found in the Old English period: *samcwic* (with 2 tokens) and 1 occurrence of each of the following: *samlæred*, *sammelt* and *samwis*. No occurrences at all were found in the Middle English or Early Modern English sections of the corpus. These findings have

much to do with the usual deficiencies adduced for the study of historical texts. Thus, Görlach (1990: 164) warns about the representativeness of the historical corpora, because

1. They just include written texts with their limitations, that is, written texts are not representative of all registers, genres, age, sex, or social condition of speakers.
2. Modern readers do not have access to every text produced at a specific period. They are restricted to some types of texts that contemporary readers considered it was worth copying.
3. The survival of the original texts is often arbitrary and by chance, which implies the data are not complete as a consequence of the random preservation. This means that there is a random selection of texts.

The validity of the data has also been questioned on the part of some scholars, as the written sources compiled in corpora are limited in size. It follows from here that the sample may be considered invalid because of its provenance and reduced size (Schneider, 2002: 81-90).

If we have a glance at the *OED* data, the dictionary claims that the prefix *sam-* is obsolete except dialectally. There is not a single entry for any of the above terms in it. However, some units are found in quotations illustrating other entries. For instance, in the entry for *sam-*, it explains that the prefix is found "in various adjs. as *sam-crisp*, *sam-dead*, *sam-red*, *sam-ripe*; *sam-hale*, 'half-whole', in poor health; {*sam-sodden*}, half cooked, half done; also fig. 'half baked', stupid."

With the exception of *samsodden*, no other derivative of *sam-* is found later than the 16th century. In fact, apart from the native rivalry between *sam-* and *half-*, from the 17th century onwards the prefix *semi-*, cognate of *sam-*, is extensively used for technical vocabulary. Even if there were already some words including this affix in Middle English and Early Modern English, as the *OED* claims,

In the 16th–18th c., the number of permanent compounds was increased mainly by the accession of terms more or less technical (many of them adapted or imitated from Latin), such as *semibreve*, *semicircle*, *semidiameter*, *semilunar*, *semi-Pelagian*, *semivowel*. At the same time there was gradual enlargement of the scope of the prefix in the formation of general nonce-compounds, which became very frequent in the 19th c., and of which it is possible to illustrate but a small proportion in the present article.

Apart from the wide use of *semi-*, another competitor appeared. For instance, the French prefix *demi-*, from Latin *dimidium* 'half', started being used in English in the 15th century. According to the *OED* it was "attrib. in Heraldry, and in the 16th c. in names of

cannon, and soon passed to other uses […] and it has become to a large extent a living element, capable of being prefixed to almost any n. (often also to adjs., and sometimes to verbs)."

Kastovski's (2009: 168) words also seem to support this idea of the competition between the Old English native resources and the French and Latin affixes:

The real productivity of Romance and Latin derivational patterns would seem to have started during the Early Modern English period, when a certain critical mass of borrowings and analogical formations had accumulated to get the derivational processes going.

To complete the scene, the Greek variant of a former *sami-*, cognate with Latin *semi-* and OE *sam-*, is introduced in the form of *hemi-*, as several Greek words containing this element were in use as technical terms in later Latin, e.g. *hemicyclium*, *hemisphærium*, *hemistichium*. As the *OED* states, "in the modern langs. they are very numerous, not only in terms adopted or adapted from Gr. (directly or through L.), but in new formations, scientific or technical, from Greek, or on Greek analogies." These terms were mainly coined from the 18th century onwards.

Thus, even it cannot categorically be affirmed, it is my contention that the competition among *sam-*, *half*, *semi-*, *demi-* and *hemi-* could have played a role in the disappearance of the native Old English prefix.

Before reaching our final conclusions, it is worth mentioning a case of popular etymology or false association undergone by the adjective *sandblind*. There seems to have existed a term corresponding to OE **samblind*, which is not attested until the 15th century, according to the *OED*, neither is recorded by any of our sources in Old English. The present form is probably a perversion of the original prefix *sam-*, after the noun *sand*, for the meaning of the word was explained by Johnson as 'having a defect in the eyes, by which small particles appear to fly before them' (*OED*).

3. CONCLUSIONS

In the preceding pages a historical development of Old English prefix sam- was traced. Firstly, concerning the main tool used, even if Nerthus presents some deficiencies that need solving, it has been demonstrated to be a suitable descriptive database for the analysis of derivational morphology of Old English both on qualitative and quantitative grounds. In fact, the research work that is being carried out by the team implementing the database, which includes information on word class, meaning, variants, derivation possibilities and other

features, facilitates the task of filtering the data according to the required parameters. Some of the other sources consulted have not proven so resourceful, but they also help to provide evidence that the prefix was still productive in the Middle English period. Nonetheless, the only predicate that survived into Modern English and is documented until the 19th century is samsodden, according to the OED.

Secondly, regarding the reasons that may account for the declining in productivity of Old English prefixes *sam*-, it is hard to determine them with certainty, but, in the light of the data, one may think that competition was a factor to be taken into consideration. Further research is needed in the future about this and other prefixes, which have also reduced their productive, to see if any conclusive ideas can be drawn.

REFERENCES

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

Bosworth, J., & Toller, T. N. (1921). *An Anglo-Saxon Dictionary Supplement*. Oxford: Oxford University Press.

Clark Hall, J. R. (1931). *A Concise Anglo-Saxon Dictionary*. Cambridge: Cambridge University Press.

The Dictionary of Old English Corpus. Online version available at: http://ets.umdl.umich.edu/o/oec/

De la Cruz, J. M. (1975). Old English Pure Prefixes: Structure and Function. *Linguistics*, 47-81.

De la Cruz Cabanillas, I. (2007). Semantic Primes in Old English: A Preliminary Study of Descriptors. *SELIM*, 14: 37-58.

Fraser, T. (1985). Etymology and the lexical semantics of Old English preverb *be*-. In Fisiak, J. (Ed.) *Historical Semantics. Historical Word-Formation* (pp. 113-153). Berlin: Mouton de Gruyter.

Görlach, M. (1990). Corpus problems of text collections: linguistic aspects of the canon. *Studies in the History of the English Language*. Carl Winter, Heidelberg: 163-78.

Hiltunen, R. (1983). *The Decline of the Prefixes and the Beginnings of English Phrasal Verb*. Turku: Turun Yliopisto.

Middle English Dictionary. Online at http://ets.umdl.umich.edu/m/med/.

Martín Arista, J. (2005-2008). Old English *Ge-* and the Descriptive Power of *Nerthus*. *Journal of English Studies*, 5-6: 209-232.

Martín Arista, J. (Forthcoming). Building a lexical database for Old English: issues and landmarks. In J. Considine (Ed.) *Selected papers from the 2008 ICHLL Conference*. Cambridge: Cambridge Scholars.

Martín Arista, J., Caballero González, L., González Torres, E., Ibañez Moreno, A. & Torre Alonso, R. (Forthcoming). *Nerthus: An Online Lexical Database of Old English*.

Rissanen, M., & Ihalainen, O. (1991). *The Helsinki Corpus of English Texts. Diachronic and Dialectal*. Helsinki: University of Helsinki.

Roberts, J., & Kay, C. (1995). *A Thesaurus of Old English.* London: King's College London Medieval Studies.

Schneider, E. (2002). Investigating Variation and Change in Written Documents. In Chambers, J. K., Trudgil, P. T., & N. Schilling-Estes (Eds.) *The Handbook of Language Variation and Change* (pp. 67-96). Blackwell: Oxford.

Sweet, H. (1896). *The Student's Dictionary of Anglo-Saxon*. Oxford: Clarendon Press.

Simpson, J. A., & Weiner, E. S. C. (eds.) (1989). *The Oxford English Dictionary*. Oxford: Oxford University Press. (2nd ed. on CD-ROM, Version 4.0)

Stratmann, F. H. (1891). *A Middle English Dictionary*. London: Oxford Clarendon Press.

Torre Alonso, R., Martín Arista, J., Ibañez Moreno, A., González Torres, E., & Caballero González, L. (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.

Visualizations for exploratory corpus and text analysis

CHRIS CULY

VERENA LYDING

Institute for Specialised Communication and Multilingualism, EURAC research

Abstract
In this paper we are concerned with how visualizations can support exploratory corpus and text analysis. We start by giving an overview of previous work in information visualization that is based on language data. We discuss how existing approaches differ from our approach . The core of the present article consists of a detailed presentation of five visualization components that we have designed for supporting the undirected exploratory analysis of text material. For each component we point out possible application contexts and motivations for the design choices and how they are related to established visualization principles. We conclude with a short discussion on future needs for the field of linguistic information visualization.
Keywords: corpus linguistics, visualization, exploratory analysis

Resumen
En este artículo nos ocupamos de cómo las visualizaciones pueden ayudar al análisis exploratorio del corpora y al análisis de texto. Comenzamos por dar una visión general de lo que se ha hecho previamente en visualización de información basada en datos del lenguaje. Discutimos cómo los enfoques existentes se diferencian del nuestro. El núcleo del articulo consiste en una presentación detallada de cinco componentes de visualización que hemos diseñado para ayudar al análisis exploratorio de datos de texto. Para cada componente señalamos posibles contextos donde pueden ser aplicados, las razones por las cuales tomamos las diferentes decisiones de diseño, y como se relacionan con los principios de visualización establecidos. Concluimos con una breve discusión de las necesidades futuras en el campo de visualización de información lingüística.
Palabras clave: lingüística de corpus, visualización, análisis exploratorio

1. INTRODUCTION

The field of information visualization is concerned with "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" (Card et al., 1999), making use of the special capabilities for pattern recognition of the human visual system. The field of information visualization has been around for about 20 years, but it has recently started to mature. Major media outlets such as the New York Times regularly publish information visualizations, and basic tools to create visualizations easily are becoming increasingly available (e.g. Google Visualization API[1], Many Eyes[2]).

[1] http://code.google.com/apis/visualization/
[2] http://manyeyes.alphaworks.ibm.com/

At the same time, the multitude and extent of available corpora and text collections call for new methods to present and access this semi-structured data. Visualizations "use graphics to organize information, highlight important information, allow for visual comparisons, and reveal patterns, trends, and outliers in the data" (Hearst, 2009: ch. 10), and thus are particularly valuable for understanding the nature of a text collection in a broad way, without having a strong hypothesis in mind. These *exploratory* phases of text inspection can be considered a recurring part of most corpus-based studies (cf. e.g. Gilquin & Gries, 2009). While the directed search aspect of corpus and text analysis is well supported by common text analysis and query tools, there is currently little targeted support for exploratory search. Visualizations, especially interactive ones, are particularly suited for exploratory search. In fact, we see visualization as the future for exploratory linguistic analysis.

2. PREVIOUS WORK

Information visualization has mainly been concerned with numeric data. Until recently, language related visualizations had largely been concerned with presenting search results, especially for intelligence or business analysts (e.g. ThemeRiver in Havre et al., 2000). As well, visualizations of semantic information of document content, either at the lexical level (DocuBurst (Collins, 2007), Leximancer (Smith, 2000)), or at the document level (Rohrer et al., 1998) have been around for some time and have more recently been complemented by visualizations of thesauri (e.g. Visual Wordnet[3]).

Over the past few years, "clouds" have become a popular way to represent thematic or textual popularity, and have recently been included into corpus interfaces (cf. e.g. the beta interface to the DWDS corpus[4] and Monk[5]). Other word level visualizations include TileBars (Hearst, 1995), which presents document length and frequency for specific query terms together with their distribution across the text, and TextArc (Paley, 2002), an innovate alternative to standard concordances.

There have been some efforts to explore structure in texts. Arc Diagrams (Wattenberg, 2002) and work by Ruecker et al. (2008) are different visualization methods for representing patterns of repetition. Word Trees (Wattenberg & Viégas, 2008) are a technique for the visualization and interactive exploration of keyword in context lines as tree structures, and Phrase Nets (van Ham et al., 2009) visualize phrasal patterns.

[3] http://kylescholz.com/projects/wordnet/
[4] http://beta.dwds.de/
[5] http://www.monkproject.org/

Visualizations found in today's corpus query tools proper, are largely limited to dispersion plots showing the distribution of search terms over text (cf. WordSmith Tools, (Scott, 2004)), charts indicating frequency distributions of words over text types or over time, and networks for the display of co-occurrences. Additionally, corpus tools occasionally make use of color for highlighting, or size to indicate frequencies (in TAPoR[6]).

Summing up, a great part of the language related work is concerned with visualizing information derived from texts (e.g. automatically extracted key words) or visualizing entire documents. Much less consideration is given to visualizing linguistic features. Also, many of the language related visualizations lack a linguistic foundation. To give a concrete example, Word Trees provides an interesting and inspiring visualization of textual data, but does not make use of linguistic information, and thus lacks some of the options that it could provide (e.g. distinguishing words by their part of speech to treat homographs appropriately). Furthermore, visualizations for the linguistic analysis of textual data, and more specifically, visualizations targeted to the exploratory corpus/text analysis are extremely rare.

3. VISUALIZATIONS IN THE NEAR FUTURE

The trend in information visualization is to provide toolkits or components that are reusable in different contexts, rather than building visualizations that are application specific. While there have been recent calls for increasing the efforts to create visualization applications[7], we believe that the component approach is appropriate for linguistic visualizations, where we mainly find prototypical application examples, practically no toolkits and few components. In this section we present some of our visualization components, still under active development, to give an idea of the kinds of visualizations that are relevant to exploratory search.

When visualizing textual data, we have to decide what parts of the data are crucial to be displayed, how the data can be condensed, what abstractions are sensible and how different views of the data can be combined. Established visualization principles (cf. Card et al., 1999; Hearst, 2009) provide guidance on how best to create visualizations that meet the context-specific information aims. Thus, the starting point for constructing visualizations is understanding the task of the user. For each of our visualization components, we give the user's task and explain the general principles and techniques employed and how the component helps the user accomplish the task.

[6] http://portal.tapor.ca/
[7] E.g. Enrico Bertini: http://diuf.unifr.ch/people/bertinie/visuale/2009/06/im_sick_and_tired_so_many_libr.html

3.1. Corpus Clouds

One aspect of exploratory corpus inquiry is getting an idea of what is frequent and how phenomena are distributed across the corpus. Corpus Clouds is a small program which provides visualizations of different types of frequency and distribution information for dynamic queries via a standard query system, integrated with a KWIC display. The overall design is inspired by Schneiderman's (1996) information visualization mantra overview first, zoom and filter, then details on demand, along with the idea of multiple views of the data, implemented in Corpus Clouds as four parallel views on query results [Fig. 1, from top]:

1. A distribution graph, showing the distribution of tokens and results over the corpus
2. A results pane displaying all strings that match the query
3. A KWIC display for a selected result type
4. A pane showing the extended context for one KWIC line

The different panels are coordinated by a technique called *brushing and linking*. Changes in one panel will update the other panels accordingly, for example selecting a specific result in the results pane, causes its distribution to be displayed as a graph in panel 1 and its concordance lines shown in the KWIC display panel. The KWIC display further provides for a view in which small bars indicate the frequency of each word in context, similar to the *sparklines* technique of Tufte (2006).

Clouds have been criticized in the human computer interface literature as not being good interfaces for web sites (Hearst & Rosner, 2008). However, the cloud view in Corpus Clouds is designed to meet the needs of corpus users, who are interested in frequency (of phrases as well as of words), by being interactive and flexible in its display order and scaling, as well as allowing for a simple list view instead of the cloud. By taking seriously the user's task (discovering information about frequency and distribution), we can repurpose a technique that is not optimal in other situations.

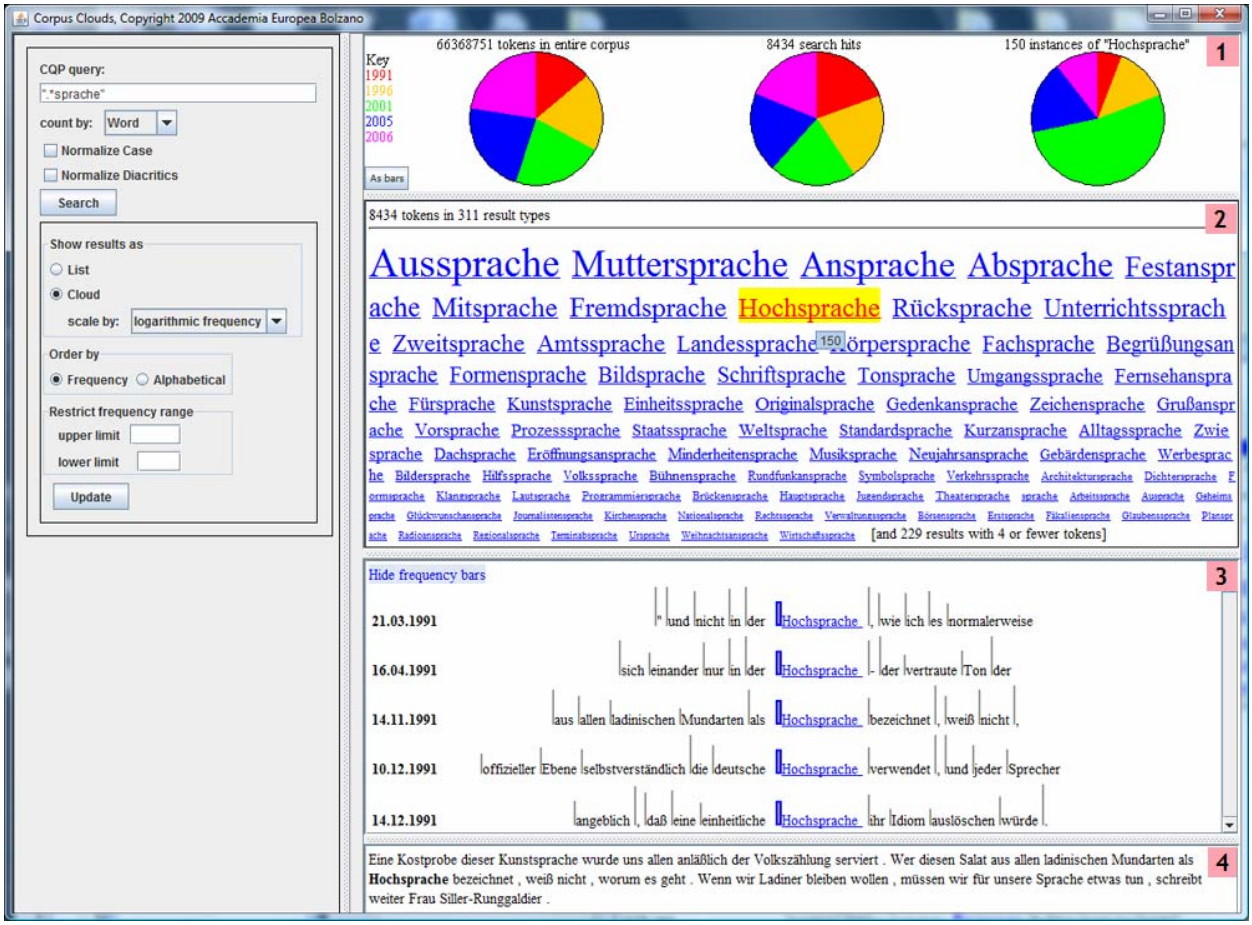


Figure 5: Corpus Clouds

3.2. Double Tree

Concordances with their KWIC display predate computers by centuries. They help the user in the task of discovering the linguistic context of words. One problem with the KWIC display is that it is not easy to make sense of a large number of results for a single term – it is difficult to detect regularities and differences in contexts. A second problem with standard KWIC displays is that they can only be sorted by left OR by right context, making it difficult to detect examples with a context of interest on both sides of the target term, without doing a new query. Wattenberg and Viégas (2008) provide Word Trees to help with the first problem. Word Trees collapse identical left or right contexts into a single line, giving a branching tree structure when contexts diverge. Two problems with Word Trees for corpus linguists is that infrequent results are suppressed without any notice, and only one side of context is visible at a time.

To overcome the problems with standard KWIC displays and to make up for the one-sidedness of Word Trees, we have made a new visualization, Double Tree (see Fig. 2), a two

sided word tree, which displays both left and right contexts. Initially a Double Tree shows one word of context on each side. For each context word, color shade indicates the number of distinct words that precede/follow that node, while the total number of instances of the word (in that specific context) is shown when the mouse is over the node. Selecting a context word expands the context by one level and dynamically colors all the paths on the opposite side for the results containing the context word.

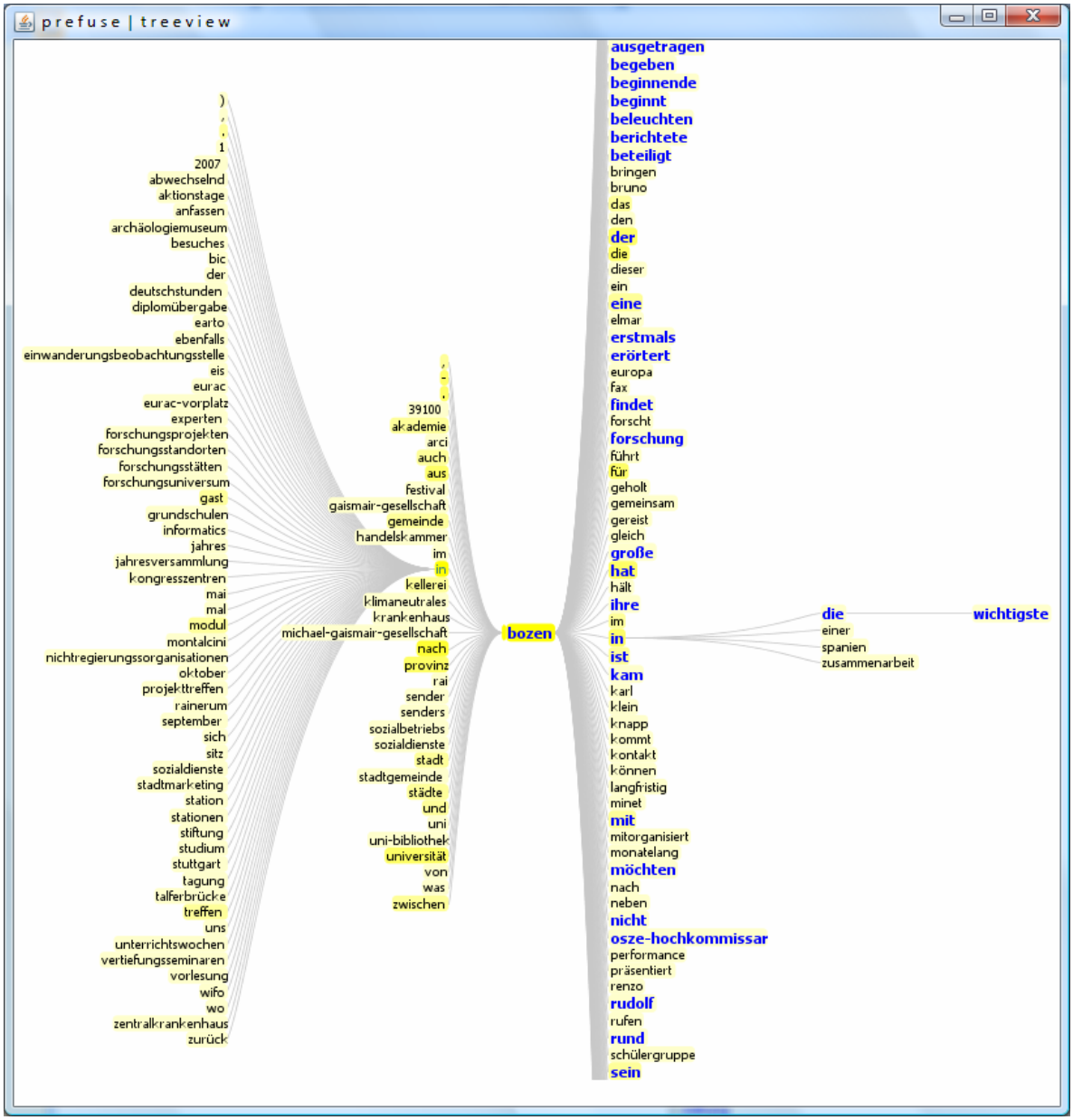


Figure 6: Double Tree

As with Word Trees, Double Trees implement a general goal of visualization to present more information in less space. Expanding only a given node is a limited example of a

fisheye view (Furnas, 1981) according to the degree of interest. In addition, since the relatedness of left and right contexts is conveyed visually by assigning a unitary color to the words of one result phrase, Double Trees also implement Tufte's (2006: 70) principle of *sameness*.

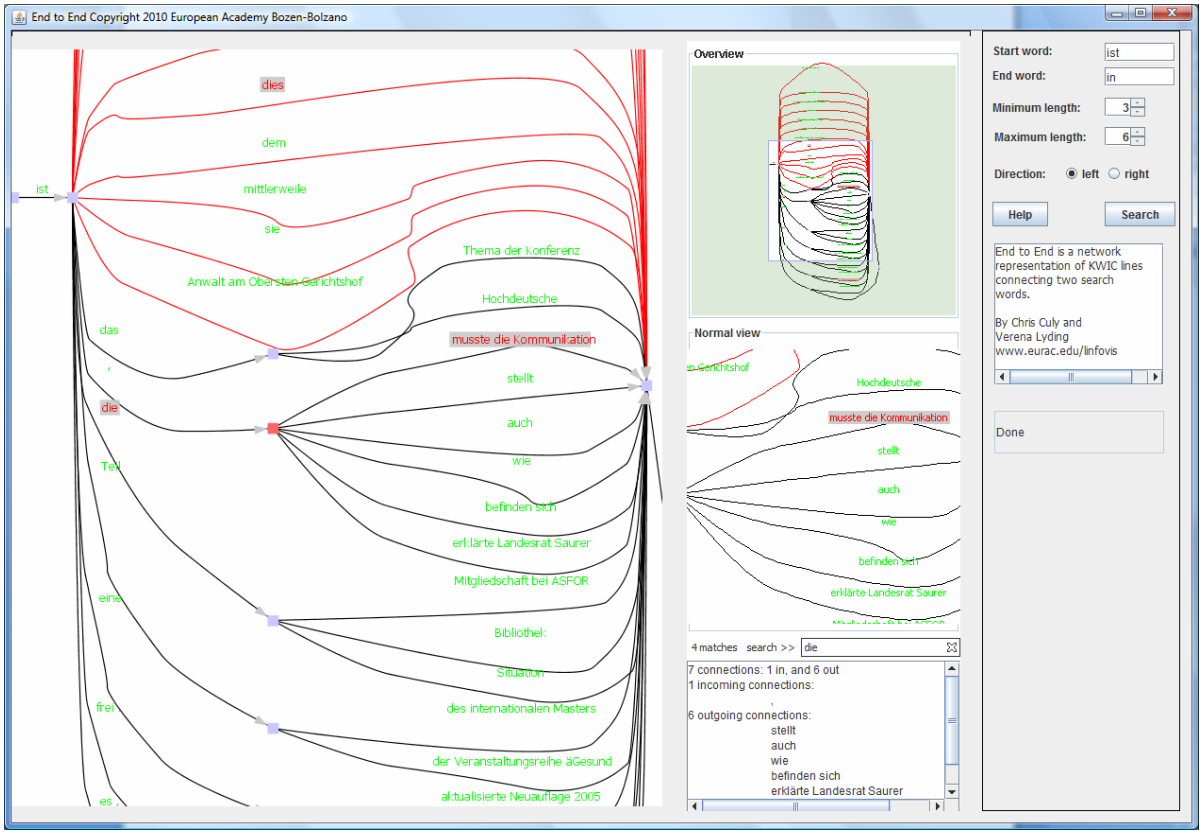


Figure 7: End-to-end for search "ist" "in"

3.3. End-to-end

Another type of exploratory search involves looking for variation and repetitions in connections between words, e.g. as collocations. End-to-end (see Fig. 3) is a component that creates a network of phrases connecting two search terms. In this way, the user can see at a glance all the connections between the terms. The user can also drill down by searching the network for particular words in between the search terms, or get a more detailed report of the context of a given word, i.e. providing details on demand. Other options allow the specification of the range of the lengths of the phrases, as well as optimizing the network from either a left to right perspective or a right to left perspective, thus providing multiple views of the data.

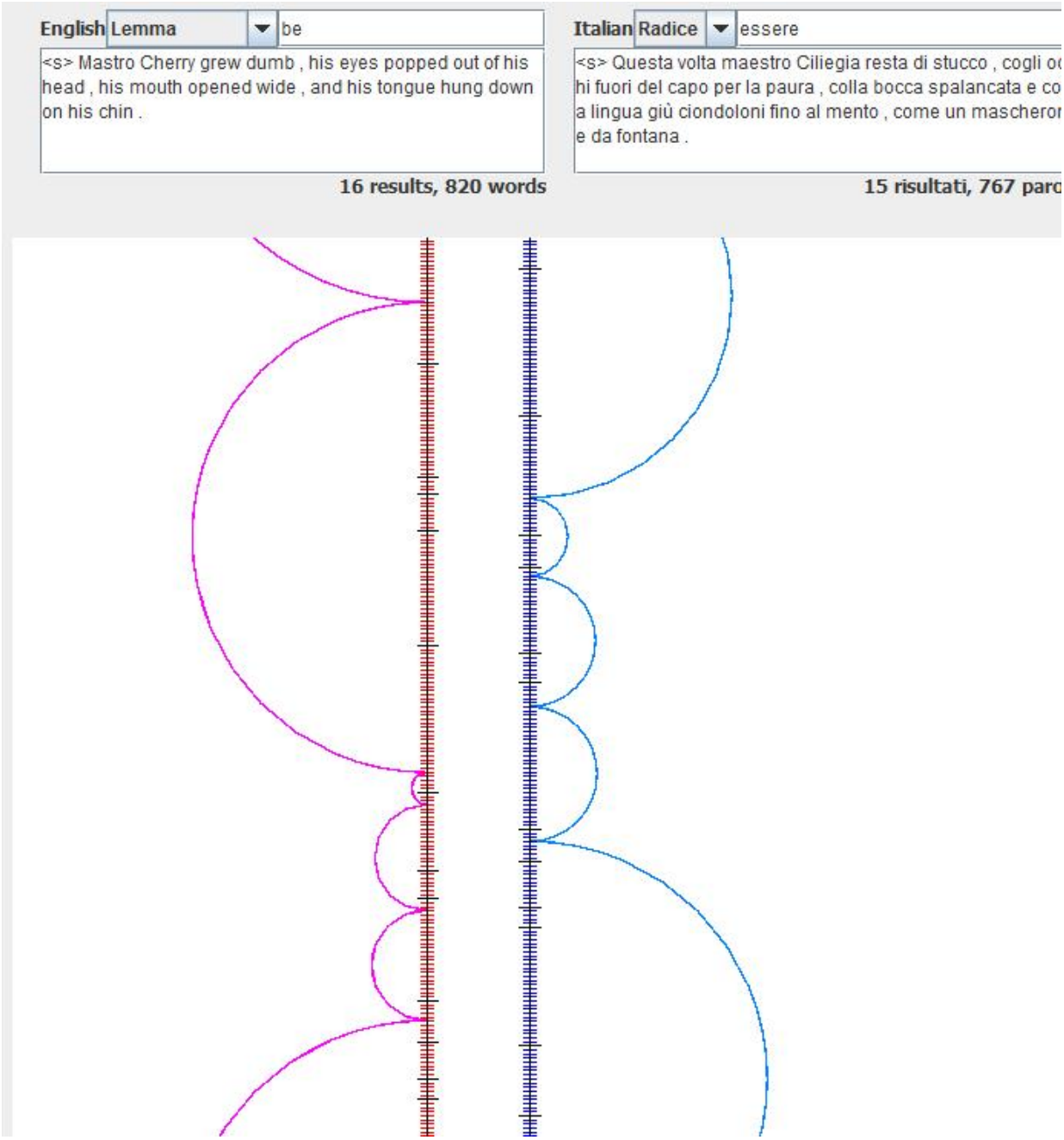


Figure 8: Comparison Arcs

3.4. Comparison Arcs

Another aspect of exploratory search is looking for positional patterns of occurrences. Wattenberg's (2002) Arc Diagrams provide a visualization of occurrence patterns by representing sequences of units as tokens on a line and connecting identical units with arcs. With Comparison Arcs (see Fig. 4), we expand the basic idea of Arc Diagrams in several ways. First, we allow the display of more than one sequence (in our case texts in the same or different languages) in parallel. Second, the visualization of Comparison Arcs is dynamic,

visualizing the results of the user's search. Third, Comparison Arcs is user-selective, only showing the search results rather than showing all (automatically chosen) correspondences. Fourth, we incorporate linguistic information by allowing the user to search for lemmas and parts of speech in addition to words.

With respect to visualization principles, we give the user information on demand by providing information about tokens and sentence boundaries by moving the mouse over the diagram.

3.5. Distribution Viewer

While Comparison Arcs visualizes co-occurrences of a particular type, Distributional Viewer (see Fig. 5) focuses on a different aspect of distribution by visualizing occurrences of a particular category. In our test example, we visualize the parts of speech of the initial words of each sentence in a small corpus. Notice that in contrast to Comparison Arcs, Distributional Viewer can handle corpora as well as individual texts.

We use two different visualization techniques, which together provide multiple views of the data. One visualization technique is essentially a *starfield* (Ahlberg & Shneiderman, 1994), where each part of speech is given a different color and plotted on a grid with sentence position on the horizontal access and text on the vertical axis. This allows the user to see broad patterns in the distribution.

To allow the user to follow up on initial observations, we provide two other views, one which shows for each part of speech its distribution across sentence document position, and the other which shows for each sentence document position, the distribution of parts of speech in that position. In both cases, we use Tufte's (1999) technique of *small multiples*, which shows separate bar graphs for each case (part of speech or sentence document position). As well, as with the other visualizations, appropriate additional information is provided by moving the mouse over the diagrams.

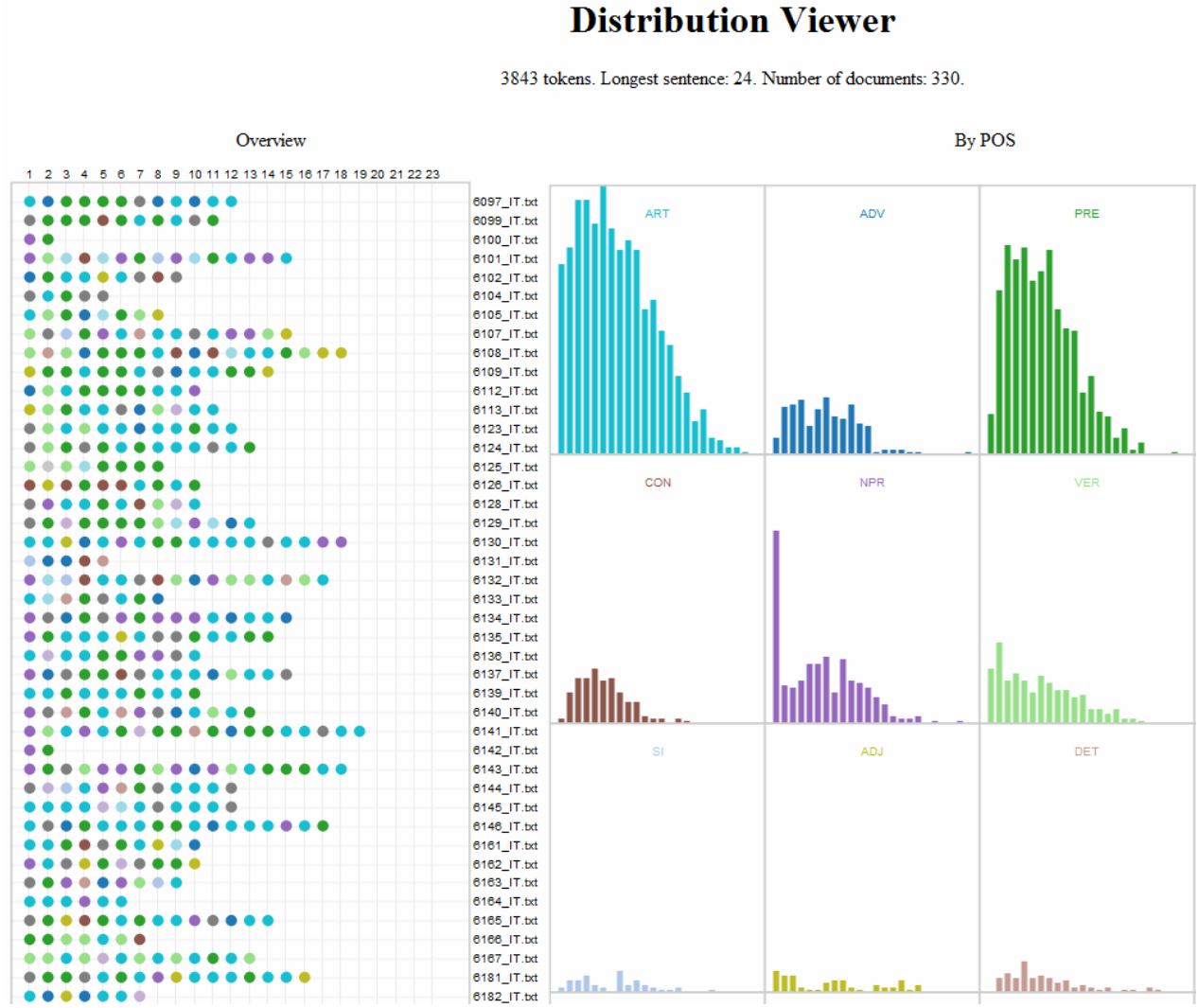


Figure 9: Distribution Viewer

Both Comparison Arcs and Distribution Viewer emphasize the point that we can visualize not only words, but lemmas, parts of speech and more. Almost all language-related visualizations that we are aware of visualize only words.

4. Discussion

In this paper we have presented a collection of visualization components for the exploratory analysis of corpus and textual data. All these visualizations build on established visualization principles to best integrate textual data and connected linguistic information into concise displays.

To extend these initial efforts, we need more experience with linguistic information visualization in general, and the user's demands with respect to particular tasks. On the theoretical side, we need to determine what visualization techniques are applicable to language data. On the practical side, we need to evaluate what kinds of visualizations can benefit what users in what kinds of tasks, and what visualization alternatives are favored over others in specific usage contexts. Furthermore, we need to find ways to guarantee efficient and flexible interoperability of all tools and components that aid the work of the language analyst especially since we see visualization tools becoming a central aspect of the next generations of corpus and text analysis tools.

REFERENCES

Ahlberg, C., & Shneiderman, B. (1994). Visual information seeking: tight coupling of dynamic query filters with starfield displays. In B. Adelson, S. Dumais, & J. Olson (Eds.). *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating interdependence*. CHI '94. ACM, New York. (pp. 313-317).

Card, S., Mackinlay, J., & Shneiderman, B. (Eds.). (1999). *Readings in Information Visualization: Using Vision to Think.*. San Diego: Academic Press.

Collins, C. (2007). Docuburst: Radial space-filling visualization of document content. Knowledge Media Design Institute, University of Toronto, Technical Report KMDI-TR-2007-1.

Furnas, G. (1981). The FISHEYE view: A new look at structured files. Bell Laboratories Technical Memorandum. No 81-11221-9.

Gilquin, G., & Gries, S. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory, 5*(1). (pp. 1-26).

Havre, S., Hetzler, B., & Nowell, L. (2000). ThemeRiver: Visualizing Theme Changes over Time. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*. INFOVIS. IEEE Computer Society, Washington, DC, 115.

Hearst, M. (1995). Tilebars: Visualization of term distribution information in full text information access. In *Proceedings CHI'95*, Denver, Colorado. (pp 56–66).

Hearst, M., & Rosner, D. (2008). Tag Clouds: Data Analysis Tool or Social Signaller? In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*. HICSS. IEEE Computer Society, Washington, DC, 160.

Hearst, M. (2009). *Search User Interfaces*. Cambridge: Cambridge University Press.

van Ham, F., Wattenberg, M., & Viégas, F. (2009). Mapping Text with Phrase Nets. *IEEE Transactions on Visualization and Computer Graphics* 15, 6. (pp. 1169-1176).

Paley, W. (2002). TextArc: Revealing Word Associations, Distributions and Frequency. Interactive Poster presented at the IEEE INFOVIS'02.

Rohrer, R., Sibert, J. & Ebert, D. (1998). The shape of Shakespeare: Visualizing text using implicit surfaces. In *Proceedings of the IEEE Symposium on Information Visualization,* Washington: IEEE Computer Society Press. (pp. 121–129).

Ruecker, S., Radzikowska, M., Michura, P., Fiorentino, C., & Clement, T. (2008). Visualizing Repetition in Text. CHWP. Retrieved from http://www.chass.utoronto.ca /epc/chwp/CHC2007/Ruecker_etal/Ruecker_etal.htm.

Scott, M. (2004). WordSmith Tools. Liverpool: Lexical Analysis Software.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, Washington: IEEE Computer Society Press. (pp. 336–343).

Smith, A. (2000). Machine Mapping of Document Collections – the Leximancer System. In *Proceedings of the Fifth Australasian Document Computing Symposium, DSTC, Sunshine Coast, Australia*.

Tufte, E. (2006). *Beautiful Evidence*. Cheshire, Connecticut: Graphics Press LLC.

Tufte, E. (1999). *Envisioning Information*. Cheshire, Connecticut: Graphics Press LLC.

Wattenberg, M. (2002). Arc diagrams: visualizing structure in strings. In *Proceedings of the IEEE Symposium on Information Visualization,* Washington: IEEE Computer Society Press. (pp. 110–116).

Wattenberg, M., & Viégas, F. (2008). The word tree, an interactive visual concordance. *IEEE Trans. on Visualization and Computer Graphics, 14*(6). (pp. 1221–1228).

The challenges of introducing corpora and their software in the English lexicology classroom: some factors

MIGUEL FUSTER MÁRQUEZ

IULMA, Universitat de València

Abstract

This paper sets out to identify the local teaching conditions in our tertiary educational system which might influence the decision of introducing corpus tools in the classroom. Different factors, which are claimed to be challenging for those who intend to explore corpora with students in the subject of English lexicology, will be looked into. To illustrate these factors I will analyze my own teaching experience with corpus tools. Some of these relate to the specific kind of student in the course I teach; other factors have to do with the students' lack of training with corpus and its tools. However, I consider that corpus based activities are highly beneficial to students in a TEFL and content learning in a foreign language environment and point to the usefulness of its implementation in lectures dealing with English lexicology. Regrettably, there is still a lot of work to be done before corpus-based approaches become a reality in descriptive linguistics courses.

Keywords: corpus, lexicology, teaching and learning.

Resumen

Esta contribución se propone identificar las condiciones educativas propias de nuestro sistema universitario que pueden resultar decisivas a la hora de adoptar herramientas de corpus en el aula. Se abordan varios factores que podrían dificultar la intención de explorar corpus con estudiantes en la asignatura de Lexicología Inglesa. Estos factores se examinan a través de la experiencia del autor con herramientas de corpus. Algunos de estos factores tienen que ver con la tipología de estudiante en estos cursos, otros con la falta de preparación en el empleo de estas herramientas. Sin embargo, el autor considera que los ejercicios basados en corpus son muy beneficiosos para los estudiantes en un entorno de lengua extranjera al tiempo que se posiciona a favor de su implementación en clases de Lexicología Inglesa. Desafortunadamente todavía nos aguarda una gran labor hasta lograr que la introducción de propuestas basadas en corpus sea una realidad en cursos descriptivos de lengua.

Palabras clave: corpus, lexicología, teaching and learning.

1. INTRODUCTION

Corpus Linguistics (CL) fits perfectly into new trends in tertiary education teaching and learning, particularly by fostering learner autonomy (Boulton, 2009: 37). Many studies have emphasized that the adoption of a corpus approach enhances the students' role as active agents of their own learning, whereas the teacher becomes a mediator (McEnery & Wilson, 1997: 6). However Boulton has recently noted that there is a lack of empirical evidence to support claims about its effectiveness (2009: 38).

For some years I have made use of corpus in the *English lexicology classroom* at the University of Valencia. In my view, in order to address its suitability for teaching English lexicology or any other similar discipline we need to consider various local contextual factors. The volumes of Biber *et al.* (1998), or McEnery and Wilson (1996) have emphasized, to a greater or lesser extent, the resourcefulness of corpus technology in descriptive language courses. Despite some resistance, according to McEnery and Wilson (1997: 5), there seems to be a growing influence of corpus linguistics in mainstream linguistics. Unfortunately, very few studies have examined the usefulness and effectiveness of corpus technology in relation to local teaching conditions and context. This paper attempts to examine some issues that teachers have to face when confronted with the challenge of using a corpus in the classroom or when proposing autonomous learning through it.

2. THE STUDENT'S BACKGROUND

Firstly, we need to have an assessment of students' competence in English as the target language. On average, students who embark on our degree at the University of Valencia are placed between B1 and B2, only a few may show either lower or higher competence. A frequent claim in the literature is that corpus linguistics is more effective with advanced learners, although Boulton (2009: 51) also reports positive results with intermediate students. Although it has been emphasized that learners' understanding of authentic language shown in concordances depends greatly on their language competence, this does not have to be the case in our course. Indeed, teachers can benefit from corpus activities which require less competence than the ability of handling concordances as will be shown later.

A second relevant issue mentioned in the literature is whether students have had any prior exposure to corpus technology. In this respect, when my students were asked during one of our first sessions, they invariably acknowledged no earlier experience of any kind with corpus. Perhaps this was only to be expected since the implementation of CL in secondary education, so to speak through *Data Driven Learning* (DDL) activities, is non-existent in secondary schools in the Valencian Community. This might be closely related, not just to the probable lack of more innovative teacher training courses, but to the insufficiently proficient competence of learners of EFL in secondary education, which renders DDL quite ineffective in the eyes of English teachers.

Thirdly, our decision to use corpus technology may depend upon the availability of felicitous reports in descriptive language courses. In this case, outside British and American

Universities, corpus use is not very firmly established. This situation contrasts greatly with the more widespread use of corpus, not just for research purposes, but in teaching *English for Specific Purposes* (ESP) or *English for Academic Purposes* (EAP) (Krishnamurthy and Kosem, 2007: 357), and also in translation studies. Greater presence of corpus in ESP could be partly explained by the extreme facility ESP and EAP teachers have in developing a suitable corpus that meets the needs of students in those domains. Carefully targeted corpora, with no need of large amounts of data, have also been highly recommended for material writers in ESP since they offer the possibility of focusing on the specific discoursal phraseology that ESP students will encounter in their fields (see O'Keeffe, McCarthy & Carter, 2007: 182). Software developments, such as *Antconc*, *Monoconc* or *Wordsmith* together with free online access in academic institutions to relevant scholarly publications are major factors in the success of corpora exploitation in ESP. Further, instructors may benefit greatly from the extensive research carried out in both ESP and EAP. However, recent studies seem to contradict this assumption. Notably, Aijmer claims that "the direct exploitation of corpora in the EFL classroom is unusual and the impact of corpora on syllabus and materials design has been slight" (2009: 2); and Krishnamurthy and Kosem (2007: 369) argue that much of the software technology is pedagogically inadequate.

3. THE CHOICE OF CORPUS AND INTERFACE

A major difference also sets apart *English as a foreign language* (EFL) courses and the linguistics courses we teach, such as English lexicology. Whereas the pedagogical goal of EFL is that of teaching English, the basic aim of an English lexicology course is that of *describing* the lexicon, 'not teaching' it, although as Pennock (2009) has pointed out, a secondary goal in these courses is to attempt to improve students competence in English. The kind of English that is described and/or learned also sets apart the goals of ESP and General English courses. This establishes priorities as to the selection of an appropriate corpus. Hoffmann *et al.* (2008: 14-5) have rightly observed different corpora are suitable for different kinds of linguistic analysis. While the choice of a corpus in ESP or EAP courses is relatively unproblematic, a descriptive course aiming to embrace something approaching General Standard English demands a more complex sampling of language. In other words, we need a corpus that faithfully mirrors the entire English language, although we are warned that absolute criteria of corpus representativeness do not exist (Hoffmann *et al*. 2008: 18; see also Stubbs, 2009: 118).

While more specific corpora can be shown and exploited in the ESP or EAP classroom, suitable large corpora of a general type are beyond the reach of any individual teacher. It is well known that the vast majority of those which have been developed belong to large private publishing companies (see O'Keefe *et al.*, 2007: 17). To date, permission to use these corpora is granted almost exclusively to authors working for those publishing houses. Thus, we are left with extremely few publicly available reference corpora which may comply with our highly demanding requirements: *The Bank of English*, or *The British National Corpus*, *the Corpus of Contemporary American English*. However, there is no limit to the different kind of corpora that might be used for more specific tasks, so one might add some academic corpora, such as the the *Michigan Corpus of Academic Spoken English* (*MICASE*), and a few others (O'Keefe *et al.*, 2007: 204). These can be accessed freely online through different interfaces, each one offering their own search options.

A final general problem mentioned in the literature refers to the amount of class time needed to teach students the contents of these corpora or how to use the tools. It has also been suggested that some interfaces seem more adequate for research than for the classroom (Krishnamurthy & Kosem, 2007: 368). A case in point is the *BNCWeb*, which has an enormous potential (Leech, 2008: xiii) but training students in its use would hardly be worth the effort, especially if they are absolute beginners. Nevertheless, *BNCWeb* offers teachers the option of extremely sophisticated guided classroom searches which cannot be performed through other interfaces.

4. NATIVE VS. NON-NATIVE CLASSROOM CONTEXTS

Our students are non-native speakers, who cannot aspire to be what Prodromou (2003) labels as *Successful Users of English* in the near future. Further, they are not immersed in an L2 context. Foreign students cannot be trustworthy judges of correctness, grammaticality or the acceptability of lexical choices. It has been acknowledged that even very advanced students may have problems concerning delicate linguistic questions. Therefore, the empirical principles of corpus compilation and corpus linguistics are perhaps the best option at our disposal. This means that the long-standing controversy among theoretical linguists as to whether one should rely on introspection or adopt an empirical approach is pointless in a foreign language context. The only valid option in our descriptive explanations is that of examining the attested data contained in collections of native production. This production can be accessed through corpus searches. Incidentally, Sinclair (1991: 112; also Yallop, 2004: 28)

has also called into question the validity of introspection performed by natives for instance in discussions of meaning, especially with reference to habitual words.

In this respect, a central issue which should be of concern to descriptive teachers is the fact that native production consists of large numbers of repeated word sequences in ordinary discourse. Pawley and Syder have argued, for instance, that native adults know and make use of hundreds of thousands of prefabricated word combinations (1983: 192; also more recently Conklin & Schmitt, 2008). A foreign learner's only way of knowing more precisely about them is through corpus consultation. Mastering these typical sequences is a lifelong challenge for foreign language learners (Bogaards, 2000: 492). In sum, descriptive language teachers who teach foreign students should be aware that although their aim is to describe the language, they are addressing non-native speakers whose competence in the target language is limited. Here corpus tools may be crucial, since non-native students lack the innate competence natives have to test many language matters. Corpus linguistics uncovers for us what is frequent and typical in language, and therefore it deserves priority in our teaching agenda (see Fuster & Pennock, 2008; Lee, 2001; Stubbs, 2001: 151).

5. SOME TASKS IN ENGLISH LEXICOLOGY

There is a wide range of issues where corpus searches are undoubtedly useful. The following is a selection of some common uses where concordancing is not a requirement. For all the cases I have made use of *Collins Concordancer Sampler* [http://www.collins.co.uk/corpus/CorpusSearch.aspx], and Mark Davies interface with BNC [http://corpus.byu.edu/bnc]:

a. Variation in word spelling (also contrasts between varieties)
b. Variation in the spelling of word compounds within a single or across varieties (frequencies could be noted).
c. Inflectional morphological searches, noting irregularities.
d. Word-formation patterns, to examine the productivity of affixes (through the use of wild cards).
e. Word-families and the relative frequency of members.

These are fairly straightforward morphological searches which focus on form. If reference is made to meaning, it seems more sensible to design activities which require the display of concordances either through KWIC (*Key Word in Context*) or sentence view. This

approach is adequate when examining phraseology, collocations, or sense relations like synonymy or antonymy. There is extensive literature on the subject (see Stubbs, 2002; Biber *et al.*, 1998: 51; McCarthy, 1990: 16-7). Particular activities can be designed around phraseology, collocability, and differences of usage between the student's L1 and L2. For instance, even in cases of close equivalents, as in the English sentence stem *to look (sb) in the eye*, a Spanish speaker has to pay attention to encoding by (1) using a different preposition and (2) a different number for the noun, because his/her own language has *mirar a los ojos*. Other options are simply wrong. Thousands of constructions in both these languages show differences were the fact of being judged transparent or compositional is irrelevant. These and many other combinations follow what Sinclair (1991: 110) labelled as the *idiom principle*. Moreover, scholars have deemed it pedagogically more useful to rely on authentic language use shown in corpora than abstract representations of polysemy when one has the foreign learner in mind (Almela and Sánchez, 2007: 30). Table 1 below shows some activities where students are invited to explore the relationship between vocabulary, variation and choice in Present-Day British English through their own corpus investigation.

Extremely valuable discussions of meaning could turn around the relationship between corpus and lexicographical work. A frequent observation in the literature is the heavy dependence of contemporary *lexicography* on corpus research; therefore the option of dealing with both could be appropriate. Halliday (2004: 20) has claimed that "The best source of information about lexicology is the dictionary or thesaurus itself. It is important to become familiar with these works, which are now fairly common within the household.". A detailed analysis of activities which relate dictionary making and corpus would be beyond the scope of this paper. However, we would like to make clear that there is practical side to *English lexicology*, in which knowledge of the guiding principles of both modern lexicography and those of corpus linguistics are extremely useful for our students. Rewarding classroom activities and discussions about compilation principles can be proposed through corpus searches which could be checked against the contents of entries in students own dictionaries. Corpus-based activities and lexicographical work should be conceived as complementing each other for teaching purposes.

Table 1: Vocabulary, variation and choice in Contemporary British English

1. **Spelling variation in Br E**:
   Search for verbs ending in *–ize* or *–ise* and note down their frequencies.
   Find the past forms of the verbs *focus, dream, spell*. Which forms are particularly common in BrE?
2. **Variation in the spelling of compounds:**
   Check the variants of these two compounds and note down their frequencies:
   dropouts, girlfriend. Search for other common compounds with spelling variants. Which of the three options –open, hyphenated or solid– is the most usual? Did your findings lead you to any general rule?
3. **Variation and choice in the use of prepositions:**
   Search for common prepositions with a similar "meaning" which fit in these contexts and determine their relative frequencies: *knock __ the door; __ the street*; *__ the corner*.
4. **Variation of affixes:**
   It is well-known that *-ic/-ical* shows variation as a derivational suffix used to build adjectives. Search for the *-ic/*-ical adjectives which correspond to these English nouns: *analysis, geography, lexicography, poem, syntax*.
5. **Variation and fixedness of English expressions and idioms:**
   Examine common collocates which fill the gaps in these two expressions, and decide about possible variants or lack of them. Please, restrict your searches to the specific word classes indicated in square brackets.
   to swim against the ____, [nouns]
   the tip of ___ iceberg. [articles]
   ___foot and nail. [verbs]
6. **Collocations and choice**:
   Here are some words followed by a common collocate in brackets. Check these collocations and determine whether the collocate in round brackets competes with any other common collocate: *insights (useful), appetite (whet), bluntly (put), complained (bitterly), issue (raise), currency (gain), objections (raised), claims (make), interest (vested), reckeless (driving), blithering (idiot), conclusions (draw).*
7. **Irregular inflectional morphology and variation:**
   Examine the nouns listed below and decide through your linguistic knowledge why they are irregular: *contents*, *news*, *data*, *means, formulae, appendix, criteria, referendum, frescoes,* etc. Did you find alternative forms for any of these words?
8. **Variation and choice of appropriate synonym:**
   How could we decide if certain common words are synonyms? Making use of concordances find two or three contextualised examples in which these pairs are not synonymous in English and two or three more examples where they are synonymous

6. CONCLUSIONS AND FURTHER RESEARCH

In this paper we set out to describe some of the challenges and implications of corpus use in a descriptive language course like *English Lexicology*. Biber *et al.* have suggested that stimulating corpus-based activities can be designed for practically any language discipline (Biber *et* al., 1998: 12). However, a thorough knowledge of the specific teaching circumstances is a requirement. We are not claiming that the corpus and its tools should be substitutes for other familiar resources, such as grammars or dictionaries or any specific textbook. Ideally students should be introduced to them in very early stages, perhaps by

proposing simple awareness-raising activities. It has been claimed that the use of corpus and its tools is far from being established in tertiary education. There is little doubt that teachers would be more willing to embark on a corpus experience if there was a guarantee that their effort was worth it. To make corpus linguistics more attractive to faculty teachers in descriptive courses we need free access to more updated online reference corpora and the development of more attractive interfaces (see Aijmer, K. 2009: 1; O`Keefe, McCarthy and Carter, 2007: 246-247). A good example of this might be Mark Davies website [http://davies-linguistics.byu.edu/personal]. Lexicologists are also reminded that recent progress in their discipline is directly related to corpus investigation and development (Halliday, 2004: 16).

REFERENCES

Aijmer, J. (Ed.). (2009). *Corpora and Language Teaching*. Amsterdam; Philadelphia: John Benjamins.

Almela, M. & Sánchez, A. (2007). Words as Lexical Units. Learning/Teaching Vocabulary. *International Journal of English Studies*, *7*(2), 21-40.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bogaards, P. (2000). Testing L2 Vocabulary Knowledge at a high level: the Case of the Euralex French Tests. *Applied Linguistics,* 21(4), 490-516.

Boulton, A. (2009). Testing the limits of data-driven learning: language proficiency and training. *ReCALL,* 21(1), 37-54.

Conklin, K. & Schmitt, N. (2008). Formulaic Sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers?. *Aplied Linguistics*, 29(1), 72-89.

Fuster, M. & Pennock, B. (2008). The spoken core of British English: a diachronic analysis based on the BNC. *Miscelánea: A Journal of English and American Studies,* 37*,* 53-74.

Halliday, M.A.K. (2004). Lexicology. In M.A.K. Halliday, W. Teubert, C. Yallop, A. Čermakova & R. Fawcett (Eds.), *Lexicology and Corpus Linguistics*. *An Introduction* (pp. 1-22). London and New York: Continuum.

Halliday, M.A.K., Teubert, W., Yallop, C. & Čermáková, A. (2004). *Lexicology and Corpus Linguistics*. *An Introduction*. London and New York: Continuum.

Hoffmann, S., Evert, S., Smith, N., Lee, D. & Berglund Prytz, Y. (2008). *Corpus Linguistics with BNCweb-a Practical Guide*. Frankfurt am Mein: Peter Lang.

Krishnamurthy, R., & Kosem, I. (2007). Issues in creating a corpus for EAP pedagogy and research. *Journal of English for Academic Purposes,* 6, 356-373.

Lee, D. (2001). Defining Core Vocabulary and Tracking its Distribution across Spoken and Written Genres: Evidence of a Gradience of Variation from the British National Corpus. *Journal of English Linguistics, 29*(3), 250-278.

Leech, G. (2008). Foreword. In S. Hoffmann, S. Evert, N. Smith, D. Lee, D. & Y. Berglund Prytz (Eds.), *Corpus Linguistics with BNCweb-a Practical Guide* (pp. xiii-xvi). Frankfurt am Mein: Peter Lang.

McCarthy, M. (1990). *Vocabulary*. Oxford: Oxford University Press.

McEnery, T. & Wilson, A. (1997). Teaching and language corpora. *ReCALL,* 9(1), 5-14.

McEnery, T. & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

O'Keeffe, A., McCarthy, M. & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Pennock-Speck, B. (2009). European Convergence and the Role of ICT in English Studies at the Universitat de València: Lessons Learned and Prospects for the Future. In M. L. Pérez Cañado, M. L. (Ed.), *English Language Teaching in the European Credit Transfer System: Facing the Challenge* (pp. 169-186). Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien: Peter Lang Publishing.

Pawley, A., & Syder, F.H. (1983). Two puzzles for linguistics theory: natively selection and native like influence. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication* (pp. 191-226). London: Longman.

Prodromou, L. (2003). In search of the successful user of English. *Modern English Teacher,* 12(2), 5-14.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, M. (2002). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell Publishing.

Stubbs, M. (2001). Text, Corpora, and Problems of Interpretation: A Response to Widdowson. *Applied Linguistics,* 22(2), 149-172.

Stubbs, M. (2009). The Search for Units of Meaning: Sinclair on Empirical Semantics. *Applied Linguistics,* 30(1), 115-137.

Yallop, C. (2004). Words and meaning. In M.A.K. Halliday, W. Teubert, C. Yallop, A. Čermakova & R. Fawcett (Eds.). *Lexicology and Corpus Linguistics*. *An Introduction* (pp. 23-71). London and New York: Continuum.

Compilación del CoDiECan: Subcorpus de Viera y Clavijo

VICTORIA GALVÁN GONZÁLEZ

ELENA QUINTANA TOLEDO

Universidad de Las Palmas de Gran Canaria

Resumen

El objetivo de este trabajo es presentar el Corpus Diacrónico del Español de Canarias y, concretamente, el Subcorpus de Viera y Clavijo. Este proyecto de investigación se está desarrollando actualmente en la Universidad de Las Palmas de Gran Canaria a cargo de los miembros de la División de Tecnologías Emergentes aplicadas a la Lengua y la Literatura del Centro Tecnológico para la Innovación en Comunicaciones. Se ofrece una descripción de los textos que se están compilando en relación a los géneros que se incluyen, la localización y selección de los mismos, y las convenciones editoriales que guían la transcripción. El CoDiECan contará, además, con una herramienta informática específica que facilitará la gestión y explotación del corpus en investigación, de la que también se realizará una breve descripción.

Palabras clave: corpus diacrónico, español de Canarias, Viera y Clavijo, herramienta informática

Abstract

This paper aims at presenting the Diachronic Corpus of Canarian Spanish and, specifically, the Subcorpus of Viera y Clavijo. This project is currently underway in the University of Las Palmas de Gran Canaria in the research unit División de Tecnologías Emergentes aplicadas a la Lengua y la Literatura of the Technological Centre for Innovation in Communications. The presentation will touch upon the genres which are included, the localisation and selection of texts, and the editorial conventions used in the transcription. Researchers will be able to exploit this corpus thanks to a corpus management tool which is being implemented as well. Some notes regarding its design will be also provided.

Keywords: diachronic corpus, Canarian Spanish, Viera y Clavijo, corpus management tool

1. INTRODUCCIÓN

La creación de un corpus diacrónico de una lengua o de alguna de sus variedades es esencial para realizar análisis lingüísticos desde el punto de vista histórico. Este tipo de corpus no sólo proporciona información acerca de las peculiaridades de una lengua y su evolución en el tiempo, sino que también permite conocer aspectos contextuales necesarios para la interpretación de los datos. En el caso del español de Canarias, ya se ha hecho notar la necesidad de investigar en esta dirección. Medina López (1994-1995: 221) señala lo siguiente:

> investigar los textos redactados en las Islas ayudará a situar la lengua en un contexto histórico y social determinado; aportará información sobre la fecha de algunos fenómenos como el seseo o ceceo […] En el terreno léxico, por ejemplo, se ofrecerían datos de gran interés, sobre todo en la línea de los arcaísmos […], al igual que la introducción de portuguesismos y occidentalismos.

El objetivo de este trabajo es mostrar la labor realizada por el momento en la compilación del *Corpus Diacrónico del Español de Canarias* (CoDiECan) y, específicamente, en el *Subcorpus de Viera y Clavijo*. Este proyecto se está desarrollando actualmente en la Universidad de Las Palmas de Gran Canaria en el seno de la División Tecnologías Emergentes aplicadas a la Lengua y la Literatura (DTELL) del Centro Tecnológico para la Innovación en Comunicaciones (CeTIC).

En general, nuestra motivación parte de la necesidad de preservar el patrimonio cultural de las Islas y nuestro propósito es crear una compilación electrónica de buena parte del material histórico que se encuentra custodiado en diferentes archivos de Canarias comenzando por la obra literaria de José de Viera y Clavijo. El CoDiECan podrá ser accesible desde un sitio web y gestionado a través de una herramienta informática específicamente diseñada para ello.

2. ANTECEDENTES DE CORPUS SIMILARES

La lingüística de corpus ha experimentado un auge considerable en las últimas décadas gracias al desarrollo de proyectos como *Survey of English Usage Corpus*, *Cobuild Corpus* o *Brown Corpus*, por mencionar algunos, favorecidos en buena medida por los avances de la tecnología computacional. En lo que a la lengua española se refiere, existen compilaciones electrónicas de gran importancia como el *Corpus Diacrónico del Español* (CORDE) (http://corpus.rae.es/cordenet.html). Éste es un corpus textual que contiene muestras de todas las épocas del idioma hasta el año 1975, y cuenta ya con 300 millones de palabras aproximadamente. En él se puede encontrar una amplia variedad de géneros con textos procedentes mayoritariamente de España e Hispanoamérica, organizados en tres períodos: desde los orígenes del idioma hasta 1491, desde 1492 hasta 1712, y desde 1713 hasta 1974.

En el panorama dialectal concretamente, hay que destacar el trabajo que se está llevando a cabo para la construcción de corpora electrónicos como el *Corpus Diacrónico del Español de Chile* (CorDECh) (Contreras Seitz, 2009) o el *Corpus Histórico del Español de México* (CHEM) (http://www.iling.unam.mx/chem/). Ambos están siendo estructurados atendiendo a cuestiones cronológicas, geográficas y genéricas, y, tal y como lo ha hecho la Real Academia Española para la computerización del CORDE, están siendo etiquetados siguiendo el estándar internacional SGML (*Standard General Markup Language*) y las recomendaciones de TEI (*Text Encoding Initiative*).

En el ámbito dialectal que pretende cubrir el CoDiECan, es preciso mencionar la labor que está desempeñando un grupo de investigadores de las Universidades de La Laguna y de Las Palmas de Gran Canaria. Dicha labor ha quedado materializada en la publicación de numerosas obras como el *Atlas Lingüístico y Etnográfico de las Islas Canarias* (ALEICan) (Alvar, 1975-1978), un punto de referencia básico que proporciona información fonética, gramatical y léxica fundamental para el desarrollo de los estudios de dialectología de esta variedad del español. En el terreno de la lexicografía, destaca el *Tesoro Lexicográfico del Español de Canarias* (Corrales Zumbado, Corbella Díaz y Álvarez Martínez, 1996), un trabajo verdaderamente novedoso en cuya edición participó la Real Academia Española y tras cuya publicación empezaron a editarse otros diccionarios como el *Diccionario de Canarismos* (Lorenzo, Morera y Ortega, 1994), el *Gran Diccionario del Habla Canaria* (O'Shanahan, 1995) y el *Diccionario Diferencial del Español de Canarias* (Corrales Zumbado, Corbella Díaz y Álvarez Martínez, 1996). Sin embargo, no existe un corpus que tenga las características que aquí se plantea, esto es, que cubra distintas etapas de la historia del español en las Islas, que esté informatizado, que sea accesible a través de Internet y que se pueda gestionar mediante software específico.

3. DESCRIPCIÓN DEL CORPUS

Un corpus que reproduzca los textos más relevantes de la historia canaria ha de incluir la obra de un clásico como José de Viera y Clavijo (1731-1813), inserto en el contexto histórico y literario de la Ilustración. Ello explica la diversidad genérico-textual de su producción y el cultivo de las formas didáctico-ensayísticas, dada la voluntad comunicativa aneja al código ilustrado. Destaca como rasgo definitorio de la escritura dieciochesca la necesidad de transferir las novedades del conocimiento, desde la ciencia a la religión, la superación de los prejuicios que lastran el avance humano, el destierro de toda suerte de supercherías y la configuración de un modelo de lector estimulado por estos intereses. De todo ello es un ejemplo señero la obra del benedictino Fray Benito Jerónimo Feijoo y Montenegro por ser el creador del ensayo moderno en España, medio eficaz en la adopción de nuevas perspectivas en la transmisión de los varios campos del saber. El espacio diacrónico al que pertenece la obra de Viera, deudor de la obra de Feijoo, ofrece amplias posibilidades para la confección de un corpus que permita investigaciones en el plano lingüístico y literario.

3.1. Géneros textuales

El subcorpus que presentamos, según la tipología al uso, tiene un carácter textual, monolingüe, general, canónico, simple y diacrónico. Se propone mostrar un volumen de textos enteros en el contexto de la obra de un autor del siglo XVIII en el ámbito canario con las consiguientes utilidades para la investigación. Los géneros que integran su obra, de acuerdo a una estructuración interna temática, pueden clasificarse del siguiente modo:

3.3.1. Prosa:

- Obra historiográfica, con la obra por la que es un clásico de las letras canarias, en la que renueva los principios de la historiografía insular, *Noticias de la historia de las Islas Canarias*, en cuatro tomos. Su edición motiva su desplazamiento a Madrid en 1770 por la imposibilidad de materializar su proyecto literario más ambicioso en las Islas. Allí publicará los cuatro tomos (1772, 1772, 1776, 1783).

- Prosa científica, que incluye diálogos (*Librito de la Doctrina Rural para que se aficionen los jóvenes al estudio de la agricultura, propio del hombre*, impreso en 1807; *Noticias de la tierra o geografía para niños*, impreso en 1807; *Tratado sobre la barrilla dispuesto en forma de diálogo*, impreso en 1810), un diccionario (*Diccionario de Historia Natural de las Islas Canarias o Índice alfabético descriptivo de sus tres reinos Animal, Vegetal y Mineral*, impreso en 1866), memorias, exámenes científicos (que redactó para la Económica de Gran Canaria, a partir de 1785, institución de la que fue su director) y una carta (*Carta filosófica sobre la Aurora Boreal observada en la ciudad de La Laguna de Tenerife en la noche del 18 de enero de 1770*).

- Literatura de viajes: *Diarios de viaje por Francia y Flandes*, que realizó entre 1777 y 1778; *Alemania e Italia*, entre 1780 y 1781; y *por el Viso de la Mancha en Ciudad Real* en 1774. En el origen están los desplazamientos del marqués de Santa Cruz, en cuya casa vivía en calidad de ayo de su hijo.

- Novela: *Vida del noticioso Jorge Sargo: novela picaresca*, escrita en sus primeros años en Tenerife, quizá hacia 1744.

- Memorias: En el marco del género autobiográfico redactó sus memorias literarias, a instancias de Juan de Sempere y Guarinos, para incluirlo en *Los mejores escritores del reinado de Carlos III*, imprescindible para el seguimiento de la producción literaria que escribió en su dilatada biografía.

- Correspondencia: Cartas familiares. En un siglo tan proclive a la comunicación abundan los epistolarios. Viera escribió numerosas cartas a distintas personalidades del

ámbito científico, literario, político, religioso o social. Relevante es la nómina de los destinatarios que incluye, todos ellos actores y testigos privilegiados de la vida cultural y literaria de la centuria.

- Periodismo manuscrito: Obras ligadas a la tertulia de Nava, de la que fue su principal artífice: *Gacetas de Daute* y *Memoriales del Síndico Personero* (compuestas durante la década de 1760).

3.3.2. Teatro

Tragedias traducidas francesas, clásicas e italianas y piezas breves. Básicamente su obra dramática está conformada por traducciones de autores de moda como *Junio Bruto*, de Voltaire; *La Mérope*, del italiano Scipion Maffei, que tradujo también Voltaire; *Los Barmecidas*, de M. de la Harpe; o *Mitrídate* y *Berenice* de Jean Racine, entre otras.

3.3.3. Poesía

- Poesías originales: Cultivó varios géneros poéticos en boga durante la centuria en la línea estético-literaria del Neoclasicismo y de la Ilustración, con pervivencia del barroquismo propio de los primeros decenios del siglo XVIII en la literatura española. Así su corpus poético incluye poesía satírico-burlesca, amorosa, didascálico-científica, patriótica o épico-heroica, circunstancial y religiosa.

- Poesías traducidas: El interés de Viera por la traducción es manifiesto a tenor del corpus que recoge su obra. Se intensificó esta tarea en especial en los años finales de su existencia. Muestra una marcada preferencia por los autores franceses, pero no están ausentes los italianos, ingleses y suizos. Los asuntos elegidos están en función de las tendencias del gusto literario adscrito al código ilustrado. Destacamos sus traducciones de la *Sátira V*, sobre la nobleza, de Boileau; *Geórgicas*, a partir de la obra de Virgilio y *Los Jardines o el Arte de hermosear paisajes*, del abate Delille; *La Enriada*, de Voltaire; o *El hombre. Poema moral*, de Alexander Pope, según la traducción francesa del abate du Resnel.

- Poesías imitadas: Con este apartado nos referimos a la clasificación que lleva a cabo el autor al distinguir entre imitación y traducción, que se explica por la conocida diversidad tipológica traductora del siglo XVIII (Urzainqui, 1991). Entre otras muestras de esta actividad del autor está el poema *Los Meses. Poema en doce cantos*, escrito en 1795, a partir del poema homónimo del francés Roucher, que publicó en 1779.

3.2. Localización y selección de los textos

La producción textual de Viera se halla básicamente en las bibliotecas públicas del Archipiélago. Es cierto que también hay testimonios manuscritos en bibliotecas privadas,

pero en este apartado sólo se puede contar con los archivos particulares de los herederos de Álvarez Rixo en el Puerto de la Cruz de Tenerife o el de los herederos del marqués de Acialcázar en Las Palmas de Gran Canaria. En principio, trabajaremos en los espacios públicos por las evidentes facilidades para el proceso investigador. Hay que señalar que, aunque las fuentes que nos interesan son manuscritas y primeras ediciones, la obra del autor cuenta también con una trayectoria impresa notable, aunque todavía insuficiente. Para toda esta información es una herramienta imprescindible la obra de Agustín Millares Carlo y Manuel Hernández Suárez (1992).

Las bibliotecas insulares que albergan los textos de Viera son las siguientes:

Biblioteca Municipal de Santa Cruz de Tenerife: Cuenta con un catálogo de manuscritos que están en formato de microfilm. La obra de Viera está en entradas referidas a títulos suyos o en colecciones de fondos procedentes de bibliotecas particulares, como el de Antonio Pereira Pacheco y Ruiz o Francisco María de León. Destacamos por su valor documental los manuscritos autógrafos de las tragedias, por ejemplo, *Berenice* de Jean Racine o *La Mérope* de Maffei, entre otras; el único manuscrito de su *Diccionario de Historia Natural de las Islas Canarias*, de 1799; la *Colección de algunos opúsculos poéticos* o *Fruta del tiempo en el Parnaso*; o también, de su puño y letra, los cuatro tomos de sus *Cartas familiares escritas a varias personalidades, por sus dignidades, clases, empleos, literatura o buen carácter de amistad y virtud*.

Biblioteca de la Universidad de La Laguna (Fondo Canario): Atesora un volumen importante de manuscritos canarios y entre ellos se hallan algunos títulos de Viera. Hay un catálogo de manuscritos del fondo antiguo (Fernández Palomeque y Morales Ayala, 2002), que se puede consultar *online*. Además, durante el mes de febrero de 2010 la biblioteca ha presentado su catálogo de manuscritos digitalizados.

Biblioteca de El Museo Canario de Las Palmas de Gran Canaria: Se trata del centro de documentación bibliográfico más antiguo de las Islas. De Viera se halla aquí el mayor volumen bibliográfico. Destaca la colección de obras copiadas por Agustín Millares Torres y Juan Padilla. Entre otras "joyas" se conserva un manuscrito autógrafo de Viera de *Los Meses* con dibujos del autor.

Biblioteca de la Sociedad Económica de Amigos del País de Tenerife. Conserva todo el fondo del marqués de Nava, amigo de Viera, que incluye obras relevantes como un manuscrito del autor del poema *Las Bodas de las Plantas*.

3.3. Transcripción de los textos

La mayoría de los textos se transcribirán directamente de los archivos o, en su defecto, utilizando manuscritos microfilmados o digitalizados. La transcripción se llevará a cabo respetando en todo momento la grafía original del texto manteniendo los signos de puntuación originales y conservando las abreviaturas, la duplicación de letras, las contracciones, el uso de mayúsculas y minúsculas, y la acentuación tal y como figuran en el original.

3.4. Codificación de los textos

En cuanto a la codificación de los textos, en los archivos aparecerá un encabezado que contendrá la información bibliográfica del documento, el tamaño del archivo y el número de palabras. Las notas marginales, la numeración de las páginas, la presencia de caracteres especiales, o el uso de idiomas distintos del español como pueden ser el inglés o el portugués, vendrán indicados mediante la codificación en XML (*eXtensible Markup Language*), utilizando para ello las recomendaciones propuestas por TEI, de forma que puedan ser eventualmente volcados en la red, y que navegadores como *Internet Explorer*, *Mozilla Firefox*, *Netscape* o *Safari* puedan interpretar la codificación y mostrar los textos limpios.

4. USO DE SOFTWARE ESPECÍFICO

Actualmente, estamos trabajando en el desarrollo de una herramienta informática específica para el tratamiento y análisis de los textos, y que, como hemos apuntado anteriormente, podrá ser accesible a través de Internet. Las primeras pruebas que se han hecho con el motor de búsqueda arrojan buenos resultados, si bien aún quedan muchos aspectos por tratar, como el indexado del corpus y la velocidad de respuesta.

5. IMPLEMENTACIÓN Y EXPLOTACIÓN DEL CORPUS EN INVESTIGACIÓN

El CoDiECan resulta atractivo por múltiples razones. Obviamente, desde un punto de vista lingüístico permitiría realizar análisis diacrónicos de fenómenos propios del habla de las Islas, lo que vendría facilitado por la posibilidad de cuantificar los datos y hacerlo de forma más rápida gracias a la herramienta informática en cuestión. Permitiría, además, acceder a la información codicológica de los manuscritos y textos impresos, así como otra relativa a la encuadernación, los escribas o las afiliaciones textuales, entre otras. Dado que los textos se encontrarán transcritos en su totalidad, permitirán un estudio pormenorizado de sus

contenidos, lo que a su vez proporcionará información de tipo social, cultural e histórica, de gran importancia para los estudios pragmáticos de los documentos. Sin embargo, el uso de este corpus no quedaría restringido únicamente a lingüistas e investigadores interesados en analizar aspectos literarios, sino que otras disciplinas como la historia, la etnografía o la antropología podrían igualmente verse beneficiadas.

6. CONCLUSIÓN

El CoDiECan llena un hueco en el estudio de la variedad del español de Canarias, sobre todo, porque no existe una compilación electrónica con textos representativos de las Islas que sean susceptibles de ser manejados mediante una herramienta informática específicamente diseñada para ello. El principal objetivo de este corpus es facilitar la realización de análisis lingüísticos y literarios, aunque también permitiría el estudio de los textos desde otras disciplinas como la historia, la antropología o la etnografía. Hasta ahora se han transcrito un buen número de textos pertenecientes al Subcorpus de Viera y Clavijo, algunos de los cuales se han codificado parcialmente para realizar las primeras pruebas del motor de búsqueda. Queda ampliar la nómina de textos transcritos para este subcorpus, codificarlos en su totalidad e introducirlos en la base de datos de la herramienta informática que se está implementando.

REFERENCIAS

Alvar, M. (1975). *Atlas lingüístico y etnográfico de las islas Canarias* (Tomo I 1975, Tomo II 1976 y Tomo III 1978). Las Palmas de Gran Canaria: Ediciones del Excelentísimo Cabildo Insular de Gran Canaria.

Contreras Seitz, M. (2009). *Hacia la constitución de un corpus diacrónico del español de Chile*. Revista de Lingüística Teórica y Aplicada 47(2), 111-134.

Corrales Zumbado, C., D. Corbella Díaz y Mª. A. Álvarez Martínez. (1992). *Tesoro lexicográfico del español de Canarias*. Madrid: Real Academia Española – Gobierno de Canarias.

Corrales Zumbado, C., D. Corbella Díaz y Mª. A. Álvarez Martínez. (1996). *Diccionario diferencial del español de Canarias*. Arco Libros.

Fernández Palomeque, P. y Mª. L. Morales Ayala. (2002). *Catálogo de manuscritos de la Biblioteca Universitaria de La Laguna*. La Laguna: Universidad de La Laguna.

Lorenzo, A., M. Morera y G. Ortega. (1994). *Diccionario de canarismos*. La Laguna: Francisco Lemus (ed.).

Medina López, J. (1994-1995). Dialectología y diacronía en el español de Canarias: Perspectivas futuras. *Revista de Filología Románica* 11-12, 217-236.

Millares Carlo, A. y M. Hernández Suárez. (1992). *Biobibliografía de escritores canarios (Siglos XVI, XVII y XVIII)* (Tomo VI). Madrid: Ediciones del Cabildo Insular de Gran Canaria.

O'Shanahan, A. (1995). *Gran diccionario del habla canaria. Más de 13000 voces y frases isleñas, de utilidad para propios y ajenos, recogidas de la tradición oral y escrita*. La Laguna: Centro de Cultura Popular Canaria.

Urzainqui, I. (1991). Hacia una tipología de la traducción en el siglo XVIII: los horizontes del traductor. En Mª. L. Donaire y F. Lafarga (Eds.), *Traducción y adaptación cultural. España-Francia* (pp. 623-638). Oviedo: Universidad de Oviedo, Servicio de Publicaciones.

La gramaticalización de cierto y semejante como determinantes. Estudio con datos de corpus

MARCOS GARCÍA SALIDO

Universidade de Santiago de Compostela

Resumen

Este trabajo estudia la gramaticalización de cierto y semejante como determinantes. En ambos la forma de partida es un adjetivo que, tras una serie de cambios semánticos y sintagmáticos, da origen a una forma con valor determinativo y con restricciones en cuanto a su combinabilidad con otros determinantes. El proceso de gramaticalización de estas dos formas ilustra la relación existente entre las categorías adjetivo y determinante, por un lado, y el hecho de que el desarrollo de un valor determinativo no está necesariamente relacionado con la naturaleza pronominal de una unidad concreta, por otro. Antes al contrario, que un determinante cuente con una forma pronominal relacionada es consecuencia de la naturaleza de su étimo.

Palabras clave: cierto, semejante, gramaticalización, determinantes

Abstract

This paper deals with the grammaticalization of cierto and semejante as determiners. Both develop from an adjective that, after a series of semantic and syntagmatic changes, gives rise to a form with determinative meaning which exhibits some restrictions regarding its combinability with other determiners. The grammaticalization of cierto and semejante shows, on the one hand, the relationship between adjectives and determiners and, on the other hand, the fact that the development of a determinative function is not necessarily related to the pronominal character of a linguistic unit. On the contrary, the existence of determiners with pronominal cognates seems to be a consequence of the nature of its source.

Keywords: cierto, semejante, grammaticalization, determiners

1. INTRODUCCIÓN[1]

A partir de la obra de Abney (1987), dentro de la lingüística chomskyana se ha asumido mayoritariamente que el determinante es el núcleo de lo que hasta entonces se venía llamando frase nominal y que pronombres y determinantes pertenecen a la misma clase léxica[2]: los determinantes son, según esto, una subclase de pronombres necesariamente transitivos que aparecen con predicados que restringen su referencia. Salvando las posibles diferencias,

[1] El autor es beneficiario de una beca para la Formación del Profesorado Universitario concedida por el MEC, referencia AP 2006-02002.

[2] La identificación pronombre-determinante no es una novedad surgida a partir de Abney. Ya Postal (1970 [1966]), dentro de la gramática generativo-transformacional defendía que se trataban de la misma clase de palabras, si bien no los consideraba núcleos de su frase –los pronombres eran frases nominales con un núcleo borrado en la estructura superficial-. Jespersen (1975 [1924]) o Fernández Ramírez (1987 [1951]) también los consideran dentro de una misma clase, aunque puedan funcionar como término primario o secundario de una determinada construcción. Mucho antes, Andrés Bello (1988 [1847]: 205, 277 y ss.) señalaba las similitudes de artículos y pronombres, que no eran sino nombres en función adjetiva o sustantiva, respectivamente.

varios han sido los que se han mostrado partidarios de aplicar un análisis similar al español (así Bosque y Moreno, 1991; Leonetti, 1999a y b, o Luján, 2004).

Existen varios argumentos que se podrían argüir a favor de esta posición o contra ciertas alternativas: por un lado está la combinatoria parcialmente diferente de frases nominales con determinante y sustantivos y sus distintas interpretaciones referenciales (lo que cuestiona la nuclearidad del sustantivo en estas construcciones) y por otro las similitudes semánticas, referenciales y distribucionales entre pronombres y frases nominales. Algunas de estas razones, llevan a Rojo y Jiménez Juliá (1989) a postular una distinción entre frases nominales (frases con determinante, estructuras exocéntricas) y frases sustantivas (estructuras con sustantivos como núcleo), lo que supone una alternativa a la primera postura mencionada.

Jiménez Juliá (2006, 2007) ha reelaborado su concepción acerca de la estructura de la frase nominal. Según el autor, los determinantes no son unidades que funcionen a un nivel sintáctico[3]. Son elementos que se han adquirido estatus de cuasi morfemas tras un proceso de gramaticalización y por ello han perdido los rasgos combinatorios de las unidades que operan a nivel sintáctico, entre ellos la recursividad (Jiménez Juliá, 2007: 24-25): muestra de ello es que los determinantes que han experimentado este proceso de gramaticalización hasta sus últimas consecuencias están en distribución complementaria (**la mi casa, *un mi amigo*). Estos son para el citado autor los determinantes propiamente dichos. Por supuesto, existen unidades que cumplen una función determinativa y que están libres de esta restricción. Por sus características morfológicas, Jiménez Juliá asimila estos últimos a los adjetivos y conserva así la etiqueta tradicional de 'adjetivo determinativo'.

La propuesta de Jiménez Juliá reviste el atractivo de permitir encuadrar a los determinantes en un proceso de gramaticalización que los relaciona con su origen adjetivo, al tiempo que ofrece claves sobre la relación distribucional entre sustantivos y frases nominales. Es cierto que estos dos elementos poseen distribuciones distintas, pero no lo es menos que existen recursos distintos de los determinantes que hacen aceptables sustantivos en contextos típicos de frases nominales (i. e. recursos que hacen de los sustantivos elementos nombradores, que los actualizan, en términos coserianos). Uno de ellos es la pluralización (cf. el famoso ejemplo de Emilio Alarcos **pasa vaca / pasan vacas*). Se han citado también la modificación y la coordinación como recursos para conseguir este tipo de efecto:

[3] Por supuesto, tampoco aquí está solo el autor (cf. Givón 2001: 97), aunque otros critican que se trate a los determinantes como meros morfemas sobre la base de argumentos tales como que se unen a sustantivos que han sido sometidos a operaciones sintácticas como modificación y complementación o que entre ellos y el sustantivo aparezcan constituyentes sintácticos como en *el a veces denostado hábito de…* (cf. Leonetti, 1999b: 807-808).

(1) *Vagos rumores* lo alcanzarían un día por sendas indirectas (BDS)

(2) […] donde *asesores científicos y directores de sociedades vinculadas a los Laboratorios Hermes* aguardaban a Abreu en fechas y horas prefijadas. (BDS)

Con la pluralización conseguimos una interpretación similar a la que se consigue con un determinante indefinido (un conjunto inespecífico de entidades)[4], pero por sí sola no habilita al sustantivo para ocupar la posición temática, cosa que sí parecen hacer los modificadores de (1) y (2).

En definitiva, aunque modificación y determinación son procedimientos distintos[5] (mientras en el primero se añaden notas semánticas a un núcleo mediante la subordinación de unidades léxicas, en el segundo, típicamente, se dan indicaciones acerca de la interpretación referencial de un nominal a través de unidades pertenecientes a clases léxicas menores) existe cierta relación entre ellos que se manifiesta de manera sincrónica (ejemplos 1 y 2) y diacrónica (ilustrada por la gramaticalización de los determinantes del español).

Los determinantes del español constituyen una innovación romance, puesto que tal clase no existía en latín, y proceden en su mayoría de adjetivos determinativos latinos[6], muchas veces homónimos con respecto a formas pronominales (relación que probablemente algunos de los autores arriba citados clasificarían como identidad más que homonimia). Tenemos pues un proceso de gramaticalización que se podría esquematizar así:

(3) adjetivo determinativo > determinante

Ahora bien, algunos casos siguen una ruta un tanto distinta, que se podría esquematizar de la manera siguiente:

(4) adjetivo calificativo > adjetivo determinativo (> determinante)

A diferencia de lo que sucede con los elementos de (3), los de (4) carecen de usos pronominales y no parecen adquirirlos como resultado de desarrollar un valor determinativo. *Cierto* y *semejante*, ilustran este patrón. A su gramaticalización dedicaré las páginas que siguen. Para ello me basaré en datos procedentes del *Corpus diacrónico del español* (*CORDE*).

[4] Aunque el indefinido *un, -a,* etc. podría recibir también una interpretación específica.
[5] Como nota Coseriu (1967 [1956]: 304-306) al separar actualización y discriminación, por un lado, de delimitación.
[6] O de ciertas combinaciones: *aliquis+unus, nec+unus,* etc. (cf. Penny, 1993: 149).

2. GRAMATICALIZACIÓN DE *CIERTO* Y *SEMEJANTE*

Recientemente han ganado protagonismo en lingüística los estudios sobre gramaticalización, cambio según el cual una palabra autónoma deviene elemento gramatical o un elemento gramatical sigue su curso de gramaticalización. Este tipo de proceso engloba una serie de cambios de orden fónico, semántico y sintáctico. Desde el punto de vista fónico se observa un desgaste que tiene su contrapartida en el plano semántico (Lehmann 2002: 112 y ss.). Además, semánticamente pueden producirse cambios de contenido descriptivo a procedimental, subjetivización, intersubjetivización, etc. Desde el punto de vista sintáctico se ha señalado que la gramaticalización generalmente reduce el ámbito en el que funciona la unidad gramaticalizada (así, por ejemplo, un modificador puede acabar siendo un morfema), aunque también se ha observado el efecto contrario (Traugott 1995).

No todos estos cambios pueden verificarse en la gramaticalización de *cierto* y *semejante*. Por ejemplo, no parece que el cuerpo fónico de estas formas haya sufrido un gran desgaste, frente a lo que sucede en el caso del artículo donde la erosión de *el* o *la* frente a las formas latinas *ille, illa* es evidente. Tal desgaste es especialmente acusado si se compara con la forma pronominal (*la*/*ella, los/ellos*) procedente del mismo étimo pero con una libertad sintáctica mayor. Así pues, me centraré a partir de aquí en dos tipos de cambios solamente: semánticos y sintáctico-distribucionales.

3. CAMBIOS SEMÁNTICOS

Las diferencias semánticas entre el adjetivo calificativo *cierto* y su contrapartida determinativa son evidentes. Cierto con valor determinativo ya no predica de una entidad determinada que es verdadera o segura, sino que se convierte en una indicación por parte del hablante a su interlocutor de que una determinada entidad no le resulta identificable pero que la descripción que le está ofreciendo no es adecuada para cualquier miembro de la clase descrita (mientras que *un* permite tanto esta lectura como una interpretación inespecífica). Hay por tanto un salto de lo que Halliday denomina función ideacional del lenguaje a un valor textual (con el valor de no anafórico) o acaso interpersonal (indicando el conocimiento que de un referente tienen los interlocutores). Estos dos valores de *cierto* aparecen ya consignados en Autoridades (s. v. cierto) y en el corpus manejado se documenta relativamente temprano una variante determinativa de cierto. En los siguientes ejemplos de finales del s. XV *cierto* parece presentar ya valor de indefinido.

(5) Como el dicho general thesorero, de nuestro expresso y verbal mandamiento... haya dado y realmente pagado a Pere Joan Spinello, criado y continuo del illustre rey don Fernando de Napoles, cinquenta y quatro mil ciento y ochenta y quatro marauedis, moneda del reyno nuestro de Castilla, los quales le hauemos mandado dar, porque truxo *cierto presente*, que nos enuio el dicho rey de Napoles (CORDE, 1492, Anónimo, Isabel Ordena, [*Documentos sobre relaciones internacionales de los Reyes Católicos I*])

(6) Muy santo y bienaventurado Padre... sepa vuestra beatitut como de presente en el mi reyno de Sicilia se a innouado *cierto litigio* entre don Johan de Luna de una parte, y Anthonio Allata y su mujer, dona Eleonor, de otra (CORDE, 1497, Anónimo, Don Fernando al papa, sobre el pleito de Juan de Luna [*Documentos sobre relaciones internacionales de los Reyes Católicos*])

El precedente de este valor determinativo podría estar en ciertas construcciones que aparecen en el corpus de forma más o menos recurrente, muchas veces en textos jurídicos, en el s. XIV. El adjetivo *cierto* se usa acompañando sustantivos como *día* o *tiempo* indicando un límite temporal del que el emisor parece seguro pero cuya situación en el futuro no se especifica:

(7) astrologos sabidores ca el buen asrologo quando viere que se ha de afogar en *algun dia cierto*: estonces guardarse ha de caualgar enesse dia (CORDE, ca. 1381 – 1418, Anónimo, Sevillana medicina de Juan de Aviñón).

(8) el que es aplazado por carta del rey. a *dia cierto* demas del dia del plazo que les fue puesto (CORDE, ca. 1310, Anónimo, *Leyes de estilo*).

(9) […] provar la paga que fiziere fasta *tiempo cierto*, salvo si la renunciare la parte que […] (CORDE, 1301, Anónimo, Carta de venta de doña Juana).

Un uso parecido se da con el sustantivo *lugar*:

(10) […] estos receptores que se ayunten en lugar cierto & que den plazos segunt fuero (CORDE, ca. 1310, Anónimo, *Leyes de estilo*)

Las diferencias semánticas entre el determinativo *semejante* y el calificativo homónimo son menos llamativas. De hecho, en *Autoridades* (s. v. *semejante*) aparece solo una acepción para esta voz, a pesar de que contextos de uso similares a los actuales se encuentran desde muy pronto en el corpus manejado. Junto al adjetivo (11) y al sustantivo (12) se pueden encontrar empleos parecidos a los del determinante actual (13).

(11) Ca *semejante* es el prinçipado ala tutoria (CORDE, 1293, Anónimo, *Castigos e documentos para bien vivir ordenados por el rey*)

(12) todas las cosas segund materia quieren su *semejante* (CORDE, 1293, Anónimo, *Castigos e documentos para bien vivir ordenados por el rey*)

(13) Pues assí, no siendo ignorante desto el auctor desta obra, cuyo nombre sub silentio iacet, e considerando ser onesto exercicio & provechoso a los que se exercitan en el arte militar, quiso ocuparse en *semejante obra,* de la qual no menor provecho alcançarán los lectores que de otras (CORDE, 1300-1305, *Libro del Cavallero Cifar*)

En (13) *semejante* aparece como un elemento que indica una relación anáforica con respecto al referente de la frase de la que es constituyente. Este valor anafórico es el que menciona García-Romero (2006), que mantiene que se trata de una forma definida, como propio del determinante *semejante* —en Jiménez Juliá (2007) aparece dentro de la nómina de determinantes indefinidos—. García-Romero señala un par de restricciones que afectan a este determinante y que, desde mi punto de vista, están relacionadas con su carácter anafórico: señala el autor su incompatibilidad con sustantivos modificados por una relativa (**Nunca he visto semejante libro que tiene Pepe en su biblioteca*) y la extrañeza de *semejante* determinando a un sustantivo con algún tipo de complemento. A mi juicio, este tipo de restricciones está relacionado con el contexto de uso de *semejante*: este determinante suele aparecer con sustantivos que mantienen una relación de sinonimia o hiperonimia con su antecedente. Se busca pues una expresión nominal con una especifidad igual o menor que la de su antecedente, que es precisamente el efecto contrario que se obtendría de insertar algún modificador de tipo especificativo.

Tampoco está claro que tenga que existir una relación de correferencia estricta entre la frase con determinante *semejante* y su antecedente. Este tipo de frase recuerda al siguiente ejemplo de Gundel *et al*. (1993):

(14) Dr. Smith told me that exercise helps. Since I heard it from A DOCTOR, I'm inclined to believe it.

En este ejemplo no hay una relación de identidad entre las frases *Dr. Smith* y *A DOCTOR*. Si existe algún tipo de anáfora, es con respecto a la clase designada por el antecedente, tal como interpretan Gundel *et al.* (1993: 296): "[…] it is the property of being a doctor, and not the identity of this particular doctor, which is relevant here". Con *semejante* se pueden conseguir las mismas interpretaciones que con ciertos indefinidos, de ahí que su inclusión entre este tipo de determinantes esté justificada.

Por otra parte, *semejante* con valor determinativo no se usa solo para indicar una relación anafórica, sino introduciendo información nueva en construcciones consecutivas, como la ejemplificada a continuación:

(15) Los bancos funcionan defectuosamente, y los notarios, con sus oficiosidades, con sus precipitaciones, echan los pies por alto antes de tiempo y organizan *semejante desbarajuste* que después no hay quien se entienda (CORDE, 1951-69, Camilo José Cela, *La colmena*).

4. CAMBIOS SINTÁCTICO-DISTRIBUCIONALES

Si en el caso de cierto la semántica es reveladora de cambios en su estatus —al menos del desarrollo de un uso determinativo ya en el s. XV—, hemos visto que en lo que se refiere a semejante este criterio es menos útil y solo su uso contextualizado como marca anafórica da ciertos indicios de su gramaticalización como elemento determinativo. En lo que sigue se revisan ciertas características sintáctico-distribucionales que pueden ayudar a completar el retrato de este proceso de gramaticalización.

Uno de los criterios citados como índice de gramaticalización es la obligatorización de las formas gramaticalizadas. Los determinantes son necesarios para la aparición de nominales en ciertos contextos y para ciertas interpretaciones referenciales. Ahora bien, hemos visto —ejemplos (1) y (2)— que la lengua dispone de otros recursos para este fin y en última instancia este criterio no sirve para deslindar adjetivos determinativos de determinantes –en caso de que se distingan estas dos clases, tal como hace Jiménez Juliá (2006, 2007).

Otras restricciones que se han relacionado con procesos de gramaticalización son la resistencia de los elementos gramaticalizados a recibir modificaciones o a ser negados. En efecto, si tenemos una unidad como *cierto* modificada por elementos como *muy, poco, absolutamente,* etc. parece que la única interpretación posible es la de *cierto* como sinónimo de "verdadero" o "seguro". Como en *semejante* la diferencia semántica entre el adjetivo calificativo y la unidad determinativa es menos clara, la presencia de modificadores o complementos indica que estamos ante una unidad poco gramaticalizada, con posibilidades sintagmáticas amplias –esto es, ante *semejante* adjetivo— pero no destierra una interpretación semántica determinada. Algo parecido ocurre con la negación.

(16) fundando la osadía presente en *el no cierto suceso* por venir (CORDE, 1579, Jerónimo Zurita, *Anales de la corona de Aragón*).

(17) parte del cuerpo animal compuesto de semejantes o *no semejantes partes* (CORDE, 1491, Fray Vicente de Burgos, *Traducción de El Libro de Propietatibus Rerum*).

De nuevo la diferencia semántica es más evidente con respecto a *cierto*: (16) es el único caso documentado de la secuencia "no cierto" precediendo a un sustantivo en CORDE. Además, la presencia del artículo hace que la interpretación indefinida que tiene esta unidad

cuando funciona con valor determinativo sea imposible[7]. De nuevo, con *semejante* las diferencias semánticas entre su uso determinativo y su uso calificativo son más difíciles de establecer. Con todo, en (17) "partes semejantes" parece que ha de parafrasearse como "partes similares entre sí" y que *semejante,* en consecuencia, no es índice de ninguna relación anafórica[8].

El rechazo a la negación que experimentan ciertas unidades tras un proceso de gramaticalización se ha relacionado con la pérdida de contenido proposicional. Así ocurre con los marcadores discursivos y con ciertos adjetivos que pasan de ser calificativos a expresar ciertas actitudes de un hablante, como los que estudian De Smet y Verstraete (2006). En los casos aquí estudiados también hay una pérdida de contenido proposicional, pero en lugar de adquirir valores actitudinales *cierto* y *semejante* pasan a comunicar indicaciones informativas.

El último criterio que revisaré será la combinabilidad de las unidades estudiadas con otros determinantes. Jiménez Juliá (2007) señala que los determinantes propiamente dichos han llegado a un punto en su proceso de gramaticalización en el cual no es posible su combinación con otros elementos de su mismo paradigma. Según este criterio, *semejante* sería un elemento determinativo más gramaticalizado que *cierto* pues su combinación con otros determinantes (en distribución prenominal) no es posible actualmente, mientras que *cierto* con valor determinativo se combina sin problemas con determinantes indefinidos. Los datos del corpus plantean un panorama algo más complejo.

Por una parte, *semejante,* la unidad en teoría más gramaticalizada de las dos estudiadas, aparece en posición prenominal precedido por artículos e indefinidos hasta fechas relativamente recientes (dos primeras décadas del s. XX), si bien este tipo de ejemplos es muy infrecuente: dos ocurrencias con artículos y ocho con el indefinido *un.*

(18) Si no lleva trazas de desaparecer la supervivencia de los oráculos entre los blancos, tan potente todavía, puede deducirse una tenacidad igual, por lo menos, en *las semejantes mateotecnias de los negros* (*CORDE*, 1906, Fernando Ortiz, *Los negros brujos*).

(19) Y así, en vez de la desbandada de antes, el teatro español conoció una fuerte continuidad de gusto y de orientación, que se mantuvo firme durante más de un siglo, período considerable, si recordamos el teatro inglés que tuvo también *una semejante orientación nacional*, pero sostenida sólo en tiempo de Shakespeare. (*CORDE*, 1910-1945, Ramón Menéndez Pidal, *La epopeya castellana a través de la literatura española*).

[7] La forma *incierto* carece evidentemente de valor determinativo y no me ocupo de ella aquí.
[8] Este ejemplo es también el único documentado en *CORDE* de *semejante(s)* negado y precediendo a un sustantivo. Hay algún otro de la secuencia *no semejante(s)*+sustantivo en los que el ámbito de la negación no es la forma *semejante(s)*.

El primero de los ejemplos plantea un problema con respecto a la caracterización semántica de *semejante* que se ofrece más arriba. Su compatibilidad con determinantes definidos parece desmentir el valor indefinido de esta unidad. Sin embargo, los únicos dos ejemplos de *semejante* en concurrencia con un determinante definido presentan características atípicas: ambos aparecen con sustantivos modificados, contexto que el determinante *semejante* rechaza (cf. García-Romero 2006). Esto hace pensar que en ejemplos como (18) no estamos ante una forma determinativa, algo que la similitud semántica entre *semejante* calificativo y determinativo dificulta apreciar.

En el caso de *cierto*, la presencia de un determinante definido provoca inmediatamente una lectura no determinativa. Ahora bien, su combinación con indefinidos está bastante restringida en la actualidad a la forma *un, -a,* etc. En *CREA* solamente hay dos casos de combinación de *cierto* con la forma *algún, -a, etc.* lo cual sugiere que *un cierto* es un determinante complejo con un alto grado de fijación y que efectivamente existe distribución complementaria entre *cierto* y otros determinantes. Asimismo, estos datos revelan que *cierto* ha alcanzado también un avanzado estado de gramaticalización.

5. Conclusión

Semejante y *cierto* son dos unidades que adquieren pronto usos determinativos: tales usos se documentan en CORDE antes de 1400 para el primero y a finales del s. XV en el caso del segundo. Sin embargo, los dos no experimentan cambios de igual magnitud. Mientras que el significado original de *cierto* está bastante alejado de la semántica que normalmente poseen los elementos determinativos, *semejante* apenas tiene que sufrir variaciones con respecto a su significado original para trasladar indicaciones de naturaleza informativa (un cierto tipo de anaforicidad en unos casos o la expresión de una magnitud imprecisa en contextos consecutivos).

En la actualidad el grado de especialización de las dos unidades estudiadas es tal que, en contexto prenominal, sus posibilidades de combinación con otros determinantes son nulas o muy restringidas.

Por otro lado, el estudio diacrónico de los determinantes es interesante para comprender la relación entre las clases de palabras adjetivo, determinante y pronombre. Los casos de *cierto* y *semejante* hablan de la dificultad de trazar una frontera entre determinantes y adjetivos durante el curso de la gramaticalización de los primeros: es a veces difícil distinguir entre las ocurrencias de las formas estudiadas que presentan valor determinativo y las que no,

pues morfológicamente son idénticas y hasta fechas recientes sus posibilidades sintagmáticas son similares. Además, en el caso de *semejante,* la semántica de la unidad determinativa es prácticamente igual a la de la no determinativa. Por otra parte, *cierto* y *semejante* a pesar de haberse gramaticalizado como determinantes no han desarrollado usos pronominales y esto tiene que ver con su origen exclusivamente adjetivo (cf. con los indefinidos pronominales *nada* o *nadie,* que carecen de correlatos determinantes y cuyas formas originales, los participios latinos *nata* y *nati* —"las cosas nacidas", "los nacidos"— probablemente se usaban de manera autónoma y no modificando a ningún sustantivo en aquellos contextos que dieron origen a su interpretación indefinida negativa).

REFERENCIAS BIBLIOGRÁFICAS

Abney, S. P. (1987). The English noun phrase in its sentential aspect. Tesis doctoral.

Bello, A. (1988). *Gramática de la lengua castellana destinada al uso de los americanos* (Ed. de Ramón Trujillo). Madrid: Arco Libros (Publicado originalmente en 1847).

Bosque, I. y J. C. Moreno (1990). Las construcciones con lo y la denotación del neutro. *Lingüística,* 2. (pp. 5-50).

Coseriu, E. (1967). Determinación y entorno. Reed. en E. Coseriu. *Lecciones de lingüística general* (pp. 282-323). Madrid: Gredos (Trabajo original publicado en Romanistisches Jahrbuch VII, 29-54, 1956).

De Smet, H. y J. C. Verstraete (2006). Coming to terms with subjectivity. *Cognitive Linguistics,* 17/3. (pp. 365-392).

Fernández Ramírez, S. (1987). *Gramática española. El pronombre* (Ed. de José Polo) (Vol. 3/2). Madrid: Arco Libros (Original publicado en 1951).

García-Romero, S. (2006). Determinando adjetivos: el caso de Semejante [Resumen]. Disponible en http://www.uned.es/sel/36Simposio/resumenes/Garcia-Romero.pdf.

Givón, T. (2001). *Syntax. An Introduction* (Vol. I). Ámsterdam: John Benjamins.

Gundel, J. et al. (1993). Cognitive status and the form of referring expressions. *Language*, 69/2. (pp. 274-307).

Jespersen, O. (1975). *La filosofía de la gramática* (Carlos Manzano, trad.). Barcelona: Anagrama (Original publicado en 1924).

Jiménez Juliá, T. (2006). *El paradigma determinante en español. Origen nominativo, formación y características*. Santiago de Compostela: Universidade de Santiago de Compostela (Anejo 56 de Verba).

Jiménez Juliá, T. (2007). *Aspectos gramaticales de la frase nominal en español*. Santiago de Compostela: Universidade de Santiago de Compostela (Anejo 60 de Verba).

Lehmann, C. (2002). Thoughts on Grammaticalization (2ª ed.). Disponible en http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-2058/ASSidUE09.pdf.

Leonetti, M. (1999a). *Los determinantes*. Madrid: Arco Libros.

Leonetti, M. (1999b). El artículo. En I. Bosque y V. Demonte (dirs.). *Gramática descriptiva de la lengua española* (Vol. I) (pp. 787-890). Madrid: Espasa.

Luján, M (2004). Determiners as Pronouns. En A. Castro et al. (eds.). Collected Papers on Romance Syntax (pp. 129-148). Cambridge, Mass.: MIT.

Penny, Ralph (1993). *Gramática histórica del español*. Barcelona: Ariel.

Postal, P. (1970). On so-called pronouns in English. En R. A. Jakobson y P. S. Rosenbaum (eds.). *Readings in English Transformational Grammar* (pp. 56-82). Waltham, Mass.: Ginn & Co. (Trabajo original publicado en 1966).

Rojo, G. y T. Jiménez Juliá (1989). *Fundamentos del análisis sintáctico funcional*. Santiago de Compostela: Universidad de Santiago de Compostela.

Traugott, E. C. (1995). The Role of the Development of Discourse Markers in a Theory of Grammaticalization. Paper presented at *ICHL XII*, Manchester 1995, Version of 11/97. Disponible en http://www.stanford.edu/~traugott/ect-papersonline.html.

Corpus:

BDS: Base de Datos Sintácticos, http://www.bds.usc.es/.

CORDE: *Corpus diacrónico del español*, http://corpus.rae.es/cordenet.html.

CREA: *Corpus de referencia del español actual*, http://corpus.rae.es/creanet.html.

Diccionarios:

Autoridades: Real Academia española (1726-1739). *Diccionario de autoridades*. Disponible en http://buscon.rae.es/ntlle/SrvltGUISalirNtlle.

La traducción de las unidades fraseológicas en la novela Lolita de Vladimir Nabokov

NAILYA GARIPOVA
Universidad de Almería

Resumen
Las obras de Nabokov han despertado el interés de los críticos literarios desde su publicación hasta nuestros días. Sin embargo, la lengua de este escritor está aun por explorar, especialmente en lo que a la fraseología se refiere. El presente trabajo analiza los métodos de traducción de las unidades fraseológicas en la novela Lolita en inglés y en ruso. Debido a su naturaleza específica, su riqueza semántica y su concentración emocional, los elementos fraseológicos presentan varias dificultades a la hora de traducirlos. Por esta razón, el análisis de los métodos de la traducción realizada por Nabokov aporta un buen campo de investigación tanto para la teoría de la traducción, como para los estudios contrastivos.
Palabras claves: fraseología, traducción, análisis contrastivo

Abstract
Nabokov's literary legacy has so far received a lot of attention from literary critics all around the world. However, little has been written about his language, particularly, about phraseology. The aim of this paper is to specify and illustrate the main methods of Nabokov's translation of phraseological units in the English and Russian versions of Lolita. Due to the specific nature, the semantic richness and the emotional concentration, phraseology seems to present a range of difficulties when it comes to translation. Therefore, a research analysis into the main translation strategies regarding the different phraseological units made by a bilingual writer contributes useful data to the theory of translation and contrastive studies in general.
Keywords: phraseology, translation, contrastive studies

1. INTRODUCCIÓN

Las obras de Nabokov han despertado el interés de los críticos literarios desde su publicación hasta nuestros días. En los EEUU ya hace décadas que se publican revistas anuales dedicadas a este escritor, como The Nabokovian y Nabokov Studies. En Rusia aparecieron dos ediciones Pro et Contra, en los años 1997 y 2001. Hoy se celebran conferencias internacionales sobre Nabokov: en 2008 en los EEUU, en 2009 en Rusia y en 2010 en Japón. El interés por sus obras y la magnitud de los estudios de su legado desde una perspectiva literaria resulta evidente. Sin embargo, en cuanto a su lengua queda mucho por explorar, como su fraseología, pues, existen muy pocos trabajos sobre este aspecto.

El presente artículo analiza los métodos de traducción de las unidades fraseológicas en la novela *Lolita* en inglés y en ruso. Como sabemos, Nabokov escribió esta obra primero en inglés y después la tradujo él mismo al ruso.

Debido a su naturaleza específica, su riqueza semántica y su concentración emocional, los fraseologismos presentan varias dificultades a la hora de traducirlos. Por esta razón, el análisis de los métodos de la traducción realizada por Nabokov aporta un buen campo de investigación tanto para la teoría de la traducción, como para los estudios contrastivos. Hemos escogido los siguientes métodos para nuestro estudio: el método contrastivo-tipológico y el análisis fraseológico desarrollado por Kunin (1986), el análisis de las definiciones lexicográficas y el análisis contextual. El corpus del trabajo contiene más de 250 unidades fraseológicas en inglés y sus traducciones correspondientes al ruso.

En la primera parte del trabajo nos ocupamos brevemente de los métodos tradicionales de traducción de los fraseologismos en la literatura lingüística, establecemos una distinción entre las unidades fraseológicas no transformadas y transformadas (las últimas abundan en la prosa de Nabokov). En la segunda parte, analizamos los métodos que Nabokov utiliza para traducir los fraseologismos del inglés al ruso y comentamos las peculiaridades de esta traducción.

2. La traducción de la fraseología en la literatura lingüística

Los lingüistas, como Jovanskaya, Retsker, Comisarov, entre otros, señalan que el mejor modo de traducir las unidades fraseológicas consiste en buscar equivalentes fraseológicos. Por esta razón, la identificación de unos determinados criterios para distinguir, por un lado, entre los equivalentes completos y parciales y, por otro, los análogos, presenta un problema acuciante en la fraseología contemporánea. Así, Retsker (1974) y Comisarov (1976) entienden por equivalentes fraseológicos aquellas unidades fraseológicas que tienen una correspondencia total a nivel significativo y que se basan en una misma imagen (*to shed crocodile tears* – 'llorar lágrimas de cocodrilo', OSD: 1111[1]). Los críticos rusos definen como análogos las unidades fraseológicas que coinciden en el nivel significativo, pero están basadas en imágenes diferentes (*to get out of bed* on the wrong side – 'levantarse con el pie izquierdo', OSD: 982). Teniendo en cuenta el nivel semántico, estructural y gramatical del inglés y el ruso, Arsentieva (1989) distingue entre equivalentes, análogos y unidades fraseológicas sin equivalencias. Los equivalentes fraseológicos son unidades con una semántica y una organización estructural y gramatical idénticas, con los componentes idénticos (*a man of his word* – 'dueño de sus palabras', OSD: 1938). Por analogías

[1] De aquí en adelante, citaremos *The Oxford Spanish Dictionary* y las páginas correspondientes. Véanse las "Referencias bibliográficas".

fraseológicas, Arsentieva entiende aquellos fraseologismos que tienen un significado idéntico o parecido, pero que se distinguen por su configuración formal (*one's blood turns to ice* – 'se ponen los pelos de punta'). Por último, distingue los fraseologismos sin equivalencias en otra lengua. Cuando la lengua a la que se traduce no ofrece equivalentes o análogos fraseológicos, las unidades fraseológicas se trasladan a través de una traducción no fraseológica. Entre este tipo de traducción distinguimos la traducción a través del calco (traducción literal), traducción léxica (con un lexema) y traducción descriptiva (utilizando estructuras gramáticas libres).

También hay que mencionar otro tipo de unidades fraseológicas: los fraseologismos transformados por el escritor. Éstos abundan en la prosa de Nabokov. Varios lingüistas, como Kunin, Fraser, Bairamova y Naciscione, aplican diferentes términos para identificar distintos tipos de estos fraseologismos transformados. En este estudio adoptamos la clasificación ofrecida por Naciscione (2001) y desarrollada mas tarde por Jovanskaya (2005). Así, distinguimos los siguientes tipos de fraseologismos transformados: la ampliación del contenido fraseológico, la sustitución de los componentes, el juego de palabras, el enriquecimiento fraseológico y los neologismos fraseológicos. En la auto-traducción de *Lolita,* Nabokov utiliza todos los métodos de traducción de las unidades fraseológicas que hemos mencionado. Procedamos, pues, a analizarlos detalladamente por separado.

3. LA TRADUCCIÓN DE LAS UNIDADES FRASEOLÓGICAS EN *LOLITA*

Nabokov a menudo emplea la traducción lexicográfica. En este caso, se trata de unidades fraseológicas que presentan "unas permanentes concordancias lexicográficas" (Jovanskaya, 2005: 114). Así, encontramos por un lado, la traducción con los equivalentes completos, como por ejemplo: *I put an end to this – Я положил конец его; With a heavy heart I – с тяжелым сердцем я; o as I laid my hands upon her – как только наложу на нее руки*. Por otro lado, Nabokov también traduce las unidades fraseológicas con los equivalentes parciales, como: Lo made a scene – Ло закатила сцену (véase el "Apéndice" (números: 1, 3, 4, 7, 11 y 22)).

Otro método de traducción utilizado por Nabokov consiste en traducir las unidades fraseológicas buscando análogos, como: *it was a daughter of a big shot – Ее отец - большая шишка* (véase el "Apéndice" (números: 9, 17, 19, 23 y 24)). Nabokov también traduce los fraseologismos no transformados a través de la traducción descriptiva: *You talk*

like a book, Dad – Ты что-то очень книжно выражаешься, милый папаша, como ejemplo (véase el "Apéndice" (números ejemplos 13, 16 y 20)).

Por último, en cuanto a traducción lexicográfica de las unidades fraseológicas no transformadas, no olvidemos que Nabokov emplea unos lexemas separados. Aunque no hay muchos ejemplos de este método de traducción en la novela (véase el "Apéndice" (números: 2 y 10)): *to lie low for a couple of years – притаиться годика на два.*

También, encontramos unidades fraseológicas traducidas mediante el método contextual. Los casos en los que Nabokov rechaza las equivalencias tradicionales y nos presenta sus versiones propias son los más interesantes, pues, podemos ver las cualidades del escritor como traductor, al mismo tiempo que observar cómo él percibe su obra. Debemos destacar que cuando la lengua rusa ofrece más de un equivalente de la unidad fraseológica en inglés, Nabokov escoge un fraseologismo con las particularidades rusas, tales como arcaísmos y elementos folklóricos. Veamos algunos ejemplos por separado. Nabokov traduce así el fraseologismo inglés: *Minnesota, to whom I take off my hat...– Миннесота, которой низко кланяюсь.* La unidad fraseológica *to take one's hat* tiene un equivalente completo, *снимать шляпу перед кем-то*, que conlleva una connotación neutral. La expresión *низко кланяюсь,* en cambio, es un arcaísmo. El fraseologismo inglés y su equivalente contextual ruso significan lo mismo y tienen la misma connotación subjetivo-evaluativa. Sin embargo, sus aspectos funcionales y estilísticos no coinciden y además se basan en imágenes diferentes.

En otro ejemplo, Nabokov no utiliza un equivalente completo del ruso sino que introduce uno nuevo para traducir el fraseologismo inglés: *to show me the door – как будто собиралась меня прогнать*. Aquí tenemos un fraseologismo nuevo que forma parte del texto en ruso, sin resultar ajeno al estilo de Nabokov. Estos fraseologismos tienen un significado parecido, aunque sus componentes y su organización estructural y gramatical son diferentes.

Veamos otro ejemplo de este tipo de fraseologismos: "Hey", she cried "take it easy"– "Эй", крикнула она, "легче на поворотах". La expresión *take it easy* se utiliza aquí como respuesta a una reacción negativa del interlocutor para que éste no se enfade. El fraseologismo ruso, *легче на поворотах*, se utiliza como una amenaza. Según la definición del diccionario de Ozhegov (2007: 480), esta expresión aparece en la segunda mitad del siglo XX en el registro de chóferes y significa: "ten cuidado, reduce la velocidad del vehículo en

las curvas"[2]. Además del tipo señalado, encontramos otro fraseologismo que recibe un significado diferente tras su traducción. En el ruso esta unidad fraseológica significa 'tener la cara descompuesta' (*DRE*:[3] 424): *the poor guy looked like his own ghost – на бедняге просто лица не было*.

Nabokov también emplea una traducción libre, no fraseológica. En el siguiente ejemplo, la unidad fraseológica en inglés se traduce al ruso con un lexema: *She did a double take – а потом как спохватится*. *Do a double take* significa 'reaccionar tardíamente' (*OSD*: 1164), sin embargo en la versión rusa *спохватится* quiere decir 'acordarse repentinamente de' (*DRE*: 813). Vemos que la versión traducida no se parece a la unidad fraseológica original.

También hay que señalar un amplio grupo de ejemplos en los que las expresiones inglesas tras su traducción al ruso se convierten en fraseologismos, aunque no lo son en la versión original. Pongamos por caso: *knocking out my poor heart out of its groove – выбили мое сердце из колеи*. La expresión *to knock one's poor heart out of its grove* se basa en una metáfora y, en realidad, presenta un calco del fraseologismo ruso *выбить из колеи* ('descaminar, sacar de rutina', *DRE:* 390). En inglés la expresión *to knock out of its grove* no es un fraseologismo, aunque *groove* significa también 'rutina', 'ranura'*; poor heart* enriquece la imagen. Tras su traducción al ruso aparece una unidad fraseológica que mantiene su elemento *poor heart*. En el ejemplo siguiente, la expresión en el original también resulta ser un calco del ruso. La lengua inglesa no tiene fraseologismo en *blind faith in something*. Sin embargo, en ruso este fraseologismo sí existe*: with her blind faith in the wisdom of her church – с ее слепой верой в мудрость своей религии.*

Como se ha comentado, Nabokov a menudo transforma las unidades fraseológicas sustituyendo componentes del fraseologismo, ampliando el contenido fraseológico y creando neologismos y juegos de palabras. El éxito de traducir un fraseologismo transformado depende de la posibilidad de encontrar en la lengua a la que se traduce un fraseologismo idéntico. Cuando la lengua de traducción ofrece equivalentes o análogos fraseológicos se puede transmitir el método del uso del fraseologismo y su efecto estilístico. Veamos estos métodos por separado.

El traductor amplía el contenido de un fraseologismo para aumentar sus posibilidades expresivas, bien para especificar su semántica o para intensificar su carga emocional-expresiva. Nabokov utiliza un equivalente completo para la traducción introduciendo el

[2] La traducción del ruso es nuestra.
[3] De aquí en adelante, utilizaremos *DRE* para referirnos al *Diccionario ruso-español*. Véase las "Referencias bibliográficas".

elemento añadido en la versión inglesa y traduciéndolo literalmente: *putting my hand on my ailing heart – положа руку на мое больное сердце.* Para traducir otro fraseologismo, transformado por ampliación, él emplea dos métodos: ampliación y aliteración. En la versión rusa utiliza un equivalente fraseológico *сильный как бык* ('fuerte como un toro', *DRE*: 200) y además introduce *штык* ('bayoneta') para conseguir el efecto de aliteración en vez de la palabra *топор* ('hacha'). Para mayor juego fónico añade una *-с* del habla popular. Así, transmite dos aspectos del fraseologismo inglés: la transformación por ampliación y la aliteración. Y además, mantiene el efecto estilístico de estos aspectos: *strong as an ox on ax* (122) — *сильным как бык-с или штык-с* (181).

En otras ocasiones, Nabokov traduce los fraseologismos transformados por repetición de los componentes, repitiendo los mismos en el ruso: *It was love at first sight, at last sight, at ever and ever sight – Это была любовь с первого взгляда, с последнего взгляда, с извечного взгляда*. Claro, a veces no se consigue reconstruir un fraseologismo transformado tras su traducción al ruso. Pongamos por caso: *Be a cake, in fact – Словом, быть паинькой*. La unidad fraseológica *to be a piece of cake* se introduce en la versión inglesa parcialmente, pero en el texto ruso Nabokov se decanta por una traducción descriptiva: *быть паинькой* ('ser un niño bueno', *DRE*: 574).

Otro método de transformación fraseológica empleada con frecuencia por Nabokov consiste en sustituir los componentes de los fraseologismos. El efecto final de esta transformación fraseológica depende de las características del componente sustituido. Encontramos fraseologismos en los que la sustitución de alguno de sus elementos cambia la semántica del fraseologismo; en otros, esta sustitución conlleva una variación en la intensidad del significado. Para traducir este tipo de unidades fraseológicas Nabokov utiliza equivalentes completos, sustituyendo el mismo elemento y traduciéndolo literalmente, como vemos en: *a tempest in a test tube – сводится к буре в пробирке.*

Otro tipo de fraseologismos transformados, que llama la atención, son los neologismos fraseológicos: Nabokov no sólo escoge las unidades fraseológicas más coloridas y especificas, y no sólo utiliza varios métodos para transformarlas, sino que además, crea sus propios fraseologismos. Lo hace principalmente para reproducir el habla original e irrepetible de sus personajes, presentando sus neologismos fraseológicos entre comillas, dejando así claro que la expresión pertenece a un personaje concreto. Como en: *'in the ebony' (as John had quipped*) – "*в чем ночь родила" (как сострил Джон)*. El fraseologismo original es *в чем мать родила*, aunque tras la sustitución de la palabra *мать* por *ночь*, el significado también se reconoce: 'sin ropa', 'desnudo'. La palabra *ebony* significa 'negro', 'oscuro' y la

expresión *take a dip*: 'darse un chapuzón' (*OSD:* 1147). Así, podemos decir que en la novela en inglés la frase significa 'darse un chapuzón en el agua oscura', mientras que en la versión rusa tenemos: 'bañarse sin ropa' o 'bañarse desnudo'.

En otras ocasiones Nabokov traduce sus neologismos a través del calco: *Aurora had hardly "warmed her hands", as the pickers of lavender say in the country of my birth – Аврора едва «согрела руки», как говорят сборщики лаванды у меня на родине*. Para traducir el siguiente neologismo Nabokov escoge un fraseologismo ruso y lo modifica para conseguir una aliteración y así transmitir la originalidad de la expresión inglesa en ruso: *dull bulb – ты тоже глуп как пуп*. *Dull* en inglés significa 'torpe'. Quizás Nabokov consideraba que los lectores rusos no captarían la imagen y la connotación del fraseologismo inglés. Por eso selecciona un fraseologismo ruso *глуп как пробка* ('torpe como un alcornoque', *DRE*: 259) y sustituye *пробка* por *пуп* ('ombligo').

Otro tipo de fraseologismos empleados y traducidos por Nabokov es el de unidades fraseológicas transformadas para conseguir un juego de palabras. Aquí se rompe la estabilidad del fraseologismo a nivel semántico, debido a una actualización del significado léxico de uno de los componentes. Así, distinguimos diferentes tipos de juegos de palabras: por un lado, fraseologismos constituidos a base de una combinación de los sentidos propios y figurados, como: *ignored stepfathers with motherless girls on their hands and knees – игнорировали отчимов с сиротками, оставшимися у них на руках и коленях*. En este caso la lengua rusa ofrece un equivalente. Si bien la traducción del juego de palabras resulta fácil, la ausencia de equivalentes o análogos no le impide en modo alguno a Nabokov reproducir el juego de palabras en el original: *Let me follow a train of thought - Was I on that train? – Дай мне продумать одну комбинацию. А я участвовала в этой комбинации?*

En otras ocasiones los fraseologismos transformados en juegos de palabras se traducen a través de un análogo parcial: *going round and round... like a God-damn mulberry moth – вертясь как проклятая бабочка в колесе*. El fraseologismo inglés *go round in a circle* tiene dos significados: *to work hard at something without making any progress* y *to be busy with something without achieving anything important.* Nabokov lo traduce al ruso empleando un fraseologismo *вертеться как белка в колесе* ('como una ardilla', *DRE*: 184), cambiando *белка* por *бабочка* ('mariposa') e introduciendo los elementos calcados de la versión inglesa. De esta manera consigue una combinación de los fraseologismos parecidos: 'como ardilla' y 'como un maldito' y, además, mantiene las características de su fraseologismo inglés.

Como hemos visto, en la versión rusa de la novela Nabokov reproduce todos los métodos de transformación fraseológica de su *Lolita* inglesa. Tras el estudio comparado de

estas novelas, se observa un mayor número de fraseologismos en la obra rusa. Esto demuestra que Nabokov crea una nueva versión de la novela y no una simple traducción que se adapte al lector ruso[4]. Al introducir en su traducción nuevos fraseologismos, Nabokov no sólo reproduce una de las *trademarks* de *Lolita*, sino que compensa otras peculiaridades del habla conversacional. Por esta razón, nos resulta extraño escuchar a Humbert o a Lolita decir las expresiones del folklore ruso o los arcaísmos rusos, véase el "Apéndice" (números: 5, 6, 12, 14, 18, 21 y 25).

4. CONCLUSIONES

La auto-traducción de Lolita realizada por Nabokov cumple todas las características de una traducción literaria. La comparación de las unidades fraseológicas del texto inglés y ruso nos permite distinguir los métodos principales de esta traducción. Al traducir los fraseologismos transformados, Nabokov los adapta al lector ruso. Su objetivo principal consiste no sólo en transmitir las características formales y sustanciales del original sino que además, mantiene su valor emocional y estético. Nabokov traslada la mayor parte de las unidades fraseológicas a través de la traducción contextual y en particular con los análogos parciales. A veces nos ofrece una traducción libre. Otras veces hace que la versión rusa sea diferente del original, aunque esto no afecta a la transmisión del contenido ideológico o conceptual. Cuando la lengua rusa presenta varios equivalentes fraseológicos, Nabokov escoge el fraseologismo con el contenido específico ruso con sus elementos folklóricos y con los arcaísmos o las imágenes de la cultura rusa. De todos los métodos de transformación fraseológica, él siempre mantiene los juegos de palabras.

A nuestro entender, el análisis comparado de la auto-traducción de Nabokov presenta un material útil para los que traducen sus obras, ya que permite identificar y emplear los métodos y recursos de traducción utilizados por el propio autor. Y esto supone una gran ventaja.

[4] Para más detalle véase Garipova, N. 2008. "The Russian *Lolita*: a Comparative Analysis of Nabokov's Translation".

APÉNDICE

1. …hand in hand with his child-wife. (28) –… рука об руку с малюткой женой (56).
2. .... on whom Lo has a crush (43) –… которым бредит Ло (76).
3. An old spinster (55) – старую деву (93).
4. ... a caree girl at heart... (55) – конторщица по натуре (93).
5. ... than to mope on a suburban lawn (63) –... чем бить баклуши на пригородном газоне (102).
6. ... certain motions pertaining to the business *in* hand – if I may coin an expression (69) –... что некоторые действия, относившиеся так сказать - простите за выражение - до синицы в руке (111).
7. Paying my tribute to (74) – ...отдавая дань (118).
8. She would commit suicide. (74) –... она бы покончила с собой (118).
9. It gave me the creeps (75) – у меня прошел мороз по коже (118).
10. …put her to death (87) – убить (135).
11. The "studio-bed" ... had long been converted into the sofa it had been at hear (91) – Кровать-кушетка... была превращена просто в кушетку, чем ... всегда оставалась в душе (140)**.**
12. I had had to be careful with him. (92) – Между прочим, мне приходилось быть с ним на чеку. (145)
13. You talk like a book, Dad. (114) – Ты что-то очень книжно выражаешься, милый папаша (170)**.**
14. ... Boyd was quite a boy ... (125) – ... что Пар - парень на ять ... (185).
15. ... checking with the assistance of Vienna... (125) –... проверяя гульфик... (185).
16. ... pulled at the odds and ends of sheets... (128) –... потянул к себе концы и края простынь... (189).
17. I shook in my shoes ... (189) – ... у меня дрожали поджилки ... (203).
18. Cold spiders of panic crawled down my back. (140) – Холодные пауки ползали у меня по спине (212).

19. ... spick and span… (159) –... одетый с иголочки... (227).
20. ...I have learned a few odds and ends. (171) –... кое-какие случайные сведения до меня дошли (243).
21. ... this was a joke... (191) –... ходячая шутка... (270).
22. .. Lo was playing a double game ... (243) – ... Лолита вела двойную игру ... (342)
23. I am going nuts… (266) – Я схожу с ума... (373).
24. Rite had still been dead to the world. (267) –... Рита спала мертвым сном (373).
25. ... and very tight hunter... (294) –... и вдрызг пьяный охотник... (410).

REFERENCIAS BIBLIOGRÁFICAS

Arsentieva, E. (1989). *Сопоставительный аналисиз фразеологических еденищ*. Kazán: Universidad de Kazán.

Bairamova, L. (2004). *Введение в контрастивную лингвистику*. Kazán: Servicio de publicaciones de la universidad de Kazán.

Comisarov, V. (1976). Тетради переводчика. *Высшая школа, 13.* (pp. 167-189).

Fraser, B. (1970). Idioms within a Transformational Grammar. *Foundation of Language*, 6. Disponible en http://www.jstor.org/stable/25000426

Garipova, N. (2008). The Russian *Lolita*: a Comparative Analysis of Nabokov's Translation. En J.J. Torres Núñez y S. Nicolás Román (Eds.). *Estudios de literatura norteamericana: Nabokov y otros autores contemporáneos.* (pp.75-95). Almería: Editorial Universidad de Almería.

Jarman, B., y Russel, R. (Eds.). (2003). *The Oxford Spanish Dictionary*. Oxford University Press.

Jovanskaya, C. (2005). *Фразеологические единицы в произведениях Набокова*. Kazán: Universidad de Kazán.

Kunin, A. (1986). *Курс фразеологи*и совренного английского языка. Moscú.

Martínez Calvo, L. (2005). *Diccionario Ruso-Español*. Barcelona: Editorial Ramón Sopena.

Naciscione, A. (2001). *Phraseological Units in Discourse: Towards Applied Stylistics*. Riga: Latvian Academy of Culture.

Ozhegov, S. (2007). *Толковый словарь русского языка*. Moscú.

Retsker, Y. (1993). The Theory and Practice of Translation. En P. Zlateva (Ed.). *Translation as Social action: Russian and Bulgarian perspective.s* (pp. 18-31). Londres: Routledge.

A Corpus-based Study of Applied Linguistics Research Articles: A Multidimensional Analysis

KUNYARUT GETKHAM

National Institute of Development Administration (NIDA)

Abstract

This paper employed a multidimensional analysis (Biber, 1995; Biber et al. 2004) to investigate co-occurring patterns of linguistic features and compared how they are used across research sections. The corpus came from 60 research articles (RAs) published in five leading Applied Linguistics Journals based on the ranking of journals in Journal Citation Reports: Science Edition (2007). Twelve articles were selected to represent each journal covering the one-year period of 2006. Data were collected from the introduction, methodology, results, and discussion parts of research articles.

Findings indicated some interesting co-occurring patterns of linguistic features and multidimensional differences across research sections. Such knowledge may help non-native English research writers better understand the use of linguistic features in Applied Linguistics RAs and may help these writers produce English-medium Applied Linguistics RAs or related fields that would be more likely to be accepted by scholarly journals. The findings also provided significant implications for teaching research or academic writing in English for Academic Purposes (EAP) or English for Specific Purposes (ESP) classrooms.

Keywords: corpus-based study, multidimensional analysis, linguistic features, academic writing, research writing, Applied Linguistics, ESP, EAP

Resumen

Este trabajo empleó un análisis multidimensional (Biber, 1995; Biber et al. 2004) para investigar patrones de co-ocurrencia de rasgos lingüísticos e hizo comparaciones de sus utilidades en secciones de investigación. El corpus procedió de 60 artículos de investigación (AIs) publicados en cinco principales Diarios de Lingüísticas Aplicadas basados en el rango de diarios en Journal Citation Reports: Science Edition (2007). Doce artículos fueron seleccionados para representar cada diario durante todo el año 2006. Los datos fueron acumulados de las siguientes partes de artículos de investigación: introducción, metodología, resultados y discusión.

Los resultados mostraron unos interesantes patrones de co-ocurrencia de rasgos lingüísticos y diferencias multidimensionales encontrados en las secciones de investigación. Dicho conocimiento podría ayudar a los autores de investigación no nativos de habla inglesa a tener una mejor comprensión del uso de rasgos lingüísticos en los artículos de investigación en Lingüística Aplicada. Asimismo, podría ayudar a esos autores a producir los artículos de investigación en Lingüística Aplicada o en campos relacionados en lengua inglesa con más posibilidades de ser admitidos por revistas académicas. Los resultados también proporcionaron implicaciones importantes para la enseñanza de escritura de investigación o escritura académica en clases de Inglés para Fines Académicos (IFA) o Inglés para Fines Específicos (IFE).

Palabras clave: estudio basado en corpus, análisis multidimensional, rasgos lingüísticos, escritura académica, escritura de investigación, Lingüística Aplicada, IFE, IFA

1. INTRODUCTION

The development of computer-based approaches to discourse analysis has facilitated numerous corpus-based studies investigating linguistic features. These corpus-based studies have been conducted to investigate the use of linguistic features such as hedging (Burrough-Boenisch, 2005; Falahati, 2009; Hyland, 1994; Jensen, 2008; Lin & Liou ,2006; Salager-Meyer, 2002;Varttala, 1999;Vassileva, 2001), verb tenses (Gredhill, 2000; Li & Ge) voices (Tarone et.al, 1998), first person pronouns (Hyland, 2002), stance (Auria, 2008; Hyland& Tse, 2004; Groom, 2005), moves (Connor & Maurenen, 1999; Kanoksilapatham, 2003) and the use of corpora as a powerful tool for non-native language learning (Cobb,1997; Chambers, 2005; Gaskell& Cobb, 2004; Kennedy & Miceli,2001; Sun, 2003; Samad, 2004; Yoon & Hirvela , 2004).

Researchers have employed several methodologies to conduct corpus-based studies. One of the more effective tools utilized by Biber and his successors (Biber, 1995; Conrad, 1996; Conrad & Biber, 2001; Kanoksilapatham, 2003; Rappen, 2001) in corpus studies is a statistical method called a multidimensional analysis which was originally developed by Biber (1988) to analyze the range of spoken and written registers in English.

To my knowledge, studies employing multidimensional analysis to investigate linguistic features in applied linguistics research articles and comparing them across sections have been scarce. It is therefore the aim of this study to employ a multidimensional analysis to investigate the co-occurring patterns of linguistic features and compare them across research sections.

2. METHODOLOGY

A corpus of 60 research articles drawn from five leading applied linguistics journals was analyzed using a multidimensional analysis to investigate the co-occurring patterns of 38 linguistic features. Twelve articles were selected from each of the five journals covering the one-year period of 2006. Data were collected from the introduction, methodology, results, and discussion parts of research articles. The corpus was automatically tagged by a POS tagger called CLAWS 7 (Rayson, 2009) and automatically counted by Mono Conc Pro 2.2 (Barlow, 2004). The raw frequencies were normalized per 100 words. Results were analyzed by means of a factor analysis, an ANOVA test, and a post hoc Scheffé test. Table 1 presents the descriptive statistics of the 38 linguistic features.

Table 1: Descriptive Statistics of the 38 Linguistic Features

Features	Mean	Minimum	Maximum	Range	Std. Deviation
Past tense verbs	2.65	.00	6.38	6.38	1.61
Perfect aspect verbs	.34	.00	2.87	2.87	.35
Present tense verbs	2.37	.09	5.70	5.61	1.34
First person pronoun	.32	.00	2.19	2.19	.46
Extraposed *IT*	.28	.00	1.89	1.89	.29
Place adverbials	.11	.00	.72	.72	.12
Time adverbials	.10	.00	.66	.66	.12
Noun	28.00	18.25	44.98	26.73	2.95
Cause connectors	.18	.00	.96	.96	.17
Concessive connectors	.38	.00	1.53	1.53	.27
Whether/If	.48	.00	2.24	2.24	.56
Result connectors	.04	.00	.39	.39	.06
Other connectors	.11	.00	1.05	1.05	.13
Preposition	11.91	6.53	18.32	11.79	1.67
Attributive adjective	7.13	.54	13.17	12.64	1.99
Predicative adjective	.61	.00	2.06	2.06	.32
Adverbs	3.15	.22	47.77	47.55	3.11
Hedges	1.21	.00	8.52	8.52	.92
Public verbs	.13	.00	.82	.82	.14
Features	Mean	Minimum	Maximum	Range	Std. Deviation
Private verbs	.46	.00	3.63	3.63	.38
Suasive verbs	.08	.00	.56	.56	.10
Synthetic negation	.12	.00	1.09	1.09	.15
Analytic negation	.48	.00	1.54	1.54	.28
Pointer	.32	.00	1.69	1.69	.38
Reference	.79	.00	3.65	3.65	.72
Demonstratives	1.11	.00	2.22	2.22	.42
Nominalization/gerunds	13.39	1.54	37.20	35.66	8.60
Passive	1.31	.00	4.56	4.56	.91
Participial modifier	.92	.00	3.25	3.25	.46
Coordination	6.21	.36	12.57	12.21	2.04
TO infinitive	2.78	.32	7.34	7.02	1.46
Th/wh relatives	1.15	.22	3.29	3.07	.56
Amplifiers	.16	.00	3.67	3.67	.32
Type/token ratio	29.64	6.37	64.12	57.75	8.09
Word length	5.38	3.60	6.45	2.85	.31
that clause controlled by a verb	.55	.00	1.56	1.56	.35
that clause controlled by an adjective	.06	.00	1.05	1.05	.11
that clause controlled by a noun	.59	.00	2.06	2.06	.37

As shown in the table, the mean scores range from .04 to 29.64. The feature that occurs most frequently is nouns (28.00 per 100 words) and the feature that occurs least frequently is

result connector. The frequency of features across the corpus varies; nominalization/gerund has a maximum frequency of 37.20 per 100 words and a minimum frequency of 1.54 per 100 words.

3. FINDINGS

3.1. Co-occurring Patterns of Linguistic Features

Findings indicated that there were six factors/dimensions of co-occurring linguistics features. A summary of the factorial structure is provided in Figure 1. In this figure, features having the largest loadings on other factors were put in parentheses and were not used in the computation of factor scores.

Factor 1

Features	**Loadings**
Whether/If	.630
Passives	.548
References	.507
Prepositions	.481
Type/token ratio	.397
……	
Nominalization/gerunds	-.853
First person pronoun	-.639
Analytic negation	-.439
Amplifiers	-.348
(Pointers	-.339)
(All coordination	-.312)

Factor 2

Features	**Loadings**
To infinitives	.729
All coordination	.719
Concessive connectors	.565
(Whether/if	.547)
Perfect	.529
(References	.487)
(*That* cl. con. by a verb	.470)
(Extraposed *it*	.315)
……	
(Passives	-.435)

Factor 3

Features	**Loadings**
Suasive verbs	.623
That cl.con. by an adj.	.553
Public verbs	.549
That cl. con. by a verb	.498
Th/wh relatives	.497
Predicative adj	.486
(*That* cl.con by a noun	.361)
(Extraposed *it*	.344)
……	

Pointers	-.471

Factor 4

Features	**Loadings**
Present tense verbs	.784
Extraposed *it*	.504
That cl.con. by a noun	.503
(First person pronouns	.445)
Place adverbials	.426
Result connectors	.350
……	
Past tense verbs	-.673

Factor 5

Features	**Loadings**
Private verbs	.777
Hedges	.746
Other connectors	.686
Public verbs	.405
Cause connectors	.367
……..	
No negative features	

Factor 6

Features	**Loadings**
Nouns	.669
Word length	.574
Participial modifiers.	.572
Attributive adjective	.499
……..	
Synthetic negation	-.318

Figure 1: Factorial Structure of the 6-factor model of the 38 linguistic features

As seen in Figure 1, six dimensions emerged. The interpretation of the functions of the co-occurring features reflects dimensional functions as follows:

Dimension 1: Established Knowledge/Expression of Ownership, Dimension 2: Expression of Purpose, Dimension 3: Evaluative Stance, Dimension 4: Expression of Generality, Dimension 5: Framing Claims, and Dimension 6: Conceptual Complexity.

3.2. Dimensional Differences across Research Sections

The comparison of dimension scores demonstrated some interesting differences in the use of the co-occurring patterns across research sections. The descriptive statistics of dimension scores are presented in Table 2 and Figure 2. The results of a post hoc Scheffé test are presented in Table 3.

Table 2: Mean and Standard Deviation for Six Dimension Scores of RA sections

Dimension	Section	Mean	SD
1	Introduction	33.8723	7.7134
	Methodology	30.6341	7.6848
	Results	7.9414	11.0311
	Discussion	7.7990	8.6808
2	Introduction	5.2895	2.6650
	Methodology	-1.9519	1.1418
	Results	2.1105	1.4711
	Discussion	4.0412	1.8079
3	Introduction	-4.5107	.9731
	Methodology	-5.0553	.8314
	Results	-5.4206	.9429
	Discussion	-3.8203	1.1449
4	Introduction	-1.2045	1.8936
	Methodology	-5.4778	2.9822
	Results	-5.5255	2.6867
	Discussion	-1.9208	1.9163
5	Introduction	-3.7503	.6892
	Methodology	-2.8528	2.1561
	Results	-3.7213	.6958
	Discussion	-2.7659	.9552
6	Introduction	12.0682	4.1083
	Methodology	8.3267	3.2030
	Results	7.2988	4.0568
	Discussion	11.0062	3.4064

Dimension 1: Established Knowledge/Ownership Expression
Dimension 2: Expression of Purposes
Dimension 3: Evaluative Stance
Dimension 4: Expression of Generality
Dimension 5: Framing Claims
Dimension 6: Conceptual Complexity

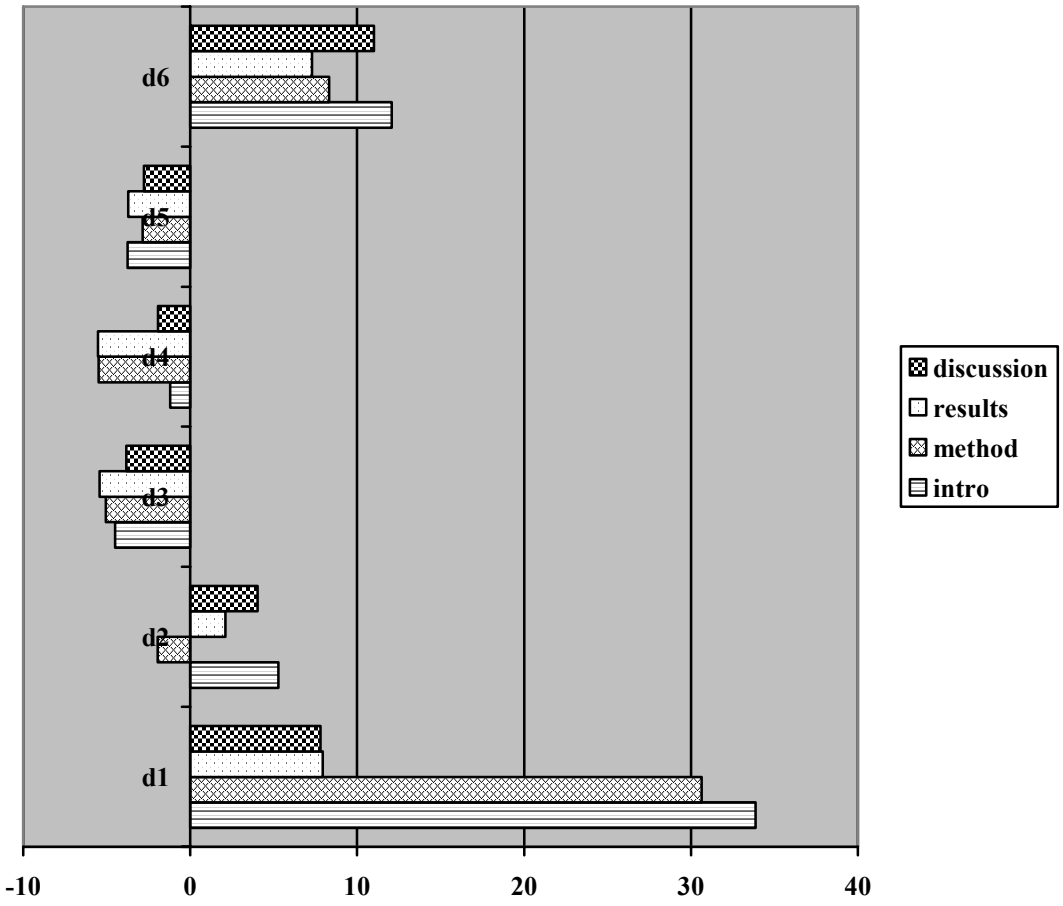


Figure 2: Means of Dimension Scores of Research Sections

Table 3: Summary of Multidimensional Differences across Sections

Dimension	Section	Section	Mean Difference	P value
1	Introduction	Methodology	3.2382	.266
		Results	25.9309	<.01
		Discussion	26.0733	<.01
	Methodology	Results	22.6927	<.01
		discussion	22.8351	<.01
	Results	Discussion	.1424	1.000
2	Introduction	Methodology	7.2414	<.01
		Results	3.1789	<.01
		Discussion	1.2483	<.01
	Methodology	Results	-4.0625	<.01
		Discussion	-5.9931	<.01
	Results	Discussion	-1.9306	<.01
3	Introduction	Methodology	.5446	<.05
		Results	.9100	<.01
		Discussion	.6903	<.01
	Methodology	Results	.3653	.246
		discussion	-1.2350	<.01
	Results	Discussion	-1.6003	<.01
4	Introduction	Methodology	4.2733	<.01
		Results	4.3210	<.01
		Discussion	.6983	.476
	Methodology	Results	.0477	1.000
		Discussion	-3.5750	<.01
	Results	Discussion	-3.6227	<.01
5	Introduction	Methodology	-.8975	<.01
		Results	.0290	.999
		Discussion	-.9844	<.01
	Methodology	Results	.8685	<.01
		Discussion	-.0869	.987
	Results	Discussion	-.9554	<.01
6	Introduction	Methodology	3.7415	<.01
		Results	4.7693	<.01
		Discussion	1.0620	.485
	Methodology	Results	1.0278	.514
		Discussion	-2.6795	<.01
	Results	Discussion	-3.7074	<.01

As shown in the tables and in the figure, differences occurred across most sections. There were similarities between some pairs.

4. DISCUSSION

On Dimension 1: Established Knowledge/Expression of Purpose, it seemed that authors tended to use condition connectors, passive voice, prepositional phrases, and citations to provide readers with established knowledge in the field. Authors appeared to use first person pronoun, participial modifiers, analytic negation, and amplifiers to express their ownership of the content. On this dimension, dimensional differences were found among most research

sections. However, there were no differences between Introduction and Method sections as well as between Results and Discussion sections.

The similarities between Introduction and Method sections regarding the styles of writing could be seen in presenting established knowledge whereas the similarities between Results and Discussion sections could be seen when authors express their ownership of the content. Presumably, in the Introduction section, it is significant for authors to refer to established knowledge as background for the readers. In addition, in the Method section, it is typical to give credits to creators of standard procedures (Kanoksilapatham, 2003).

Regarding expression of ownership, differences were not found in either Results or Discussion sections. In both sections, authors express their ownership of the content by using similar patterns.

On Dimension 2: Expression of Purposes, significant differences were found among all sections. The highest mean score occurred in the Introduction section suggesting that authors focused on expression of purpose. Presumably, researchers state the purpose of the study in the Introduction sections by employing *infinitive to*, coordination, concessive connectors, whether/if, perfect aspect verbs, *that* clause controlled by a verb, extraposed *it* and some citations.

On Dimension 3: Evaluative stance, authors framed their evaluation by using stance "that" (*that* clause controlled by an adjective, *that* clause controlled by a noun) including suasive verbs and public verbs, th/wh relatives, predicative adjectives with extraposed *it*. In Biber's (2006) study this type of discourse is labeled "Stance focused discourse".

On this dimension, there were stylistic similarities between the Method and Results sections in evaluative stance. In the Method sections, authors evaluate the methods, models, or theories they had drawn on in the research and in the Results sections, authors evaluated their own findings (Hyland and Tse, 2005).

On Dimension 4: Expression of Generality, authors tended to use present tense verbs, extraposed *it*, first person pronoun, *that* clause controlled by a noun, place adverbials and result connectors to express generality. There were stylistic similarities between Introduction and Discussion sections as well as between Method and Results sections. Finding that stylistic similarities occur between Introduction and Discussion sections is consistent with Li and Ge, (2000)'s study in that in the Introduction sections authors used this style for reference to established knowledge or universal truth and in the Discussion section to emphasize the generality of specific findings. However, similarities between the Method and Results

sections did not indicate the expression of generality since the mean scores of both sections suggested that both sections were less concerned about the expression of generality.

On Dimension 5: Framing Claims, authors frame their claims by using private and public verbs, hedges, cause connectors and other connectors. There were stylistic similarities between Introduction and Results sections as well as between Method and Discussion sections. However, the mean scores of Introduction and Results sections suggested that both were less concerned with framing claims than the Method and Discussion sections. Presumably, authors made claims about their methods and their findings in the Method and Discussion sections.

On Dimension 6: Conceptual Complexity, authors conveyed their concepts by using nouns modified by either attributive adjectives or participial modifiers resulting in more complex concepts. Authors similarly used these linguistic features in the Introduction and Discussion sections. Mean scores of Introduction and Discussion sections were relatively higher than those of Method and Results sections. The high mean scores of both Introduction and Discussion sections suggested that these two sections focused more on concepts. In contrast, the low mean scores of both Method and Results sections indicated that there was less focus on in these two sections.

5. CONCLUSION

A multidimensional analysis is a powerful tool to investigate co-occurring patterns of linguistic features in Applied Linguistics RAs. The analysis reveals that the corpus has high density of information. Applied Linguistics research writers tended to employ six patterns of co-occurring features to convey their messages to readers. Such knowledge may help not only non-native English speaking students and research article writers better understand the use of linguistic features in Applied linguistics RAs but may help these writers produce English-medium RAs in Applied Linguistics or related fields that are more likely to be accepted by scholarly journals. The findings also yield significant implications for teaching research and academic reading or writing in English for Academic Purposes or English for Specific Purposes courses.

APPENDICES

Appendix 1: Corpus of Research Articles Included in the Study

Journals	Impact Factors
1. Journal of Memory and Language (JOM)	2.83
2. Studies in Second Language Acquisition (SLA)	2.42
3. Brain and Language (B&L)	2.32
4. Journal of Speech, Language, and Hearing Research (JOS)	1.80
5. Computational Linguistics (COMLING)	1.80

Journal of Memory and Language (JOM)

JOM 1 Creel, S.C., Aslin, R.N., & Tanenhaus, M.K . (2006). Acquiring an artificial lexicon: Segment type and order information in early lexical entries. *Journal of Memory and Language*, 54(1): 1-19.

JOM 2 Salthouse, T.A., Siedlecki, K.L., & Krueger, L.E. (2006). An individual differences analysis of memory control. *Journal of Memory and Language***,** 55(1): 102-125.

JOM 3 Lozano,S.C., & Tversky, B. (2006). Communicative gestures facilitate problem solving for both communicators and recipients. *Journal of Memory and Language***,** 55(1): 47-63.

JOM 4 Arndt, J. (2006). Distinctive information and false recognition: The contribution of encoding and retrieval factors. *Journal of Memory and Language, 54(1):* 113-130**.**

JOM 5 Cleland, A.A. & Pickering, M.J. (2006). Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language, 54(2): 185-198.*

JOM 6 Staub, A., Clifton, C. Jr., Frazier, L. (2006). Heavy NP shift is the parser's last resort: Evidence from eye movements. *Journal of Memory and Language,* 54(3): 389-406**.**

JOM7 Jefferies, E.; Frankish, C.; & Lambon Ralph, M. (2006). Lexical and semantic binding in verbal short-term memory. *Journal of Memory and Language, 54(1): 81-98.*

JOM 8 Kensinger, E.A., Garoff-Eaton, R.J., & Schacter, D.L. (2006). Memory for specific visual details can be enhanced by negative arousing content. *Journal of Memory and Language, 54(1):* 99-112**.**

JOM 9 Jones, L. & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language, 55(1): 18-32.*

JOM 10 Unsworth, N., & Engel, R.W. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language, 54(1): 68-80.*

JOM 11 Zevin, J. D., & Seidenberg, M.S. (2006). Simulating consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language, 54(2): 145-160.*

JOM 12 Richard Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language, 55(1): 64-88.*

••

Studies in Second Language Acquisition (SLA)

SLA 1 Sunderman, G., & Kroll, J. F. (2006). First language activation during second language lexical processing: an investigation of lexical form, meaning, and grammatical class. *Studies in Second Language Acquisition, 28(3): 387-422*

SLA 2 Ellis, R., Erlam, R, & Loewen, S. (2006). Implicit and explicit corrective feedback and the acquisition of l2 grammar. *Studies in Second Language Acquisition, 28(2): 339-368.*

SLA 3 McDonough, K. (2006). Interaction and syntactic priming: English l2 speakers' production of dative constructions**.** *Studies in Second Language Acquisition,* 28(2): 179-207**.**

SLA 4 Lyster, R., & Mori, H. (2006). Interactional feedback and instructional counterbalance. *Studies in Second Language Acquisition, 28(2): 269-300.*

SLA 5 Carpenter, H., Seon Jeon, K., MacGregor, D., & Mackey, A. (2006). Learners' interpretations of recasts. *Studies in Second Language Acquisition, 28(2): 209-236.*

SLA 6 Lieberman, M., Aoshima, S., & Phillips, C. (2006). Nativelike biases in generation of wh-questions by nonnative *speakers of Japanese. Studies in Second Language Acquisition, 28(3): 423-448.*

SLA 7 Ammar, A., & Spada, N. (2006). One size fits all?: recasts, prompts, and L2 learning. *Studies in Second Language Acquisition, 28(4): 543-574.*

SLA 8 Morgan-Short, K., & Bowden, H.W. (2006). Processing instruction and meaningful output-based instruction: effects on second language development. *Studies in Second Language Acquisition, 28(1): 31-65.*

SLA 9 Harada, T. (2006). The acquisition of single and geminate stops by english-speaking children in a Japanese immersion program. *Studies in Second Language Acquisition, 28(4): 601-632.*

SLA 10 Munro, M.J., Derwing, T.M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28(1): 111-131.*

SLA 11 Zyzik, E. (2006). Transitivity alternations and sequence learning: insights from L2 Spanish production data. *Studies in Second Language Acquisition, 28(3): 449-485.*

SLA 12 Polio, C., Gass, S., & Chapin, L.(2006). Using stimulated recall to investigate native speaker perceptions in native-nonnative speaker interaction. *Studies in Second Language Acquisition, 28(2): 237-267.*

Brain and Language (B&L)

B&L 1 Hamilton, R.H., & Shenton, J. T. & Coslett, H. B.(2006). An acquired deficit of audiovisual speech processing. *Brain and Language, 98(1), 66-73.*

B&L 2 Watson, P., & Montgomery, E. B. (2006). The relationship of neuronal activity within the sensori-motor region of the subthalamic nucleus to speech. *Brain and Language, 97(2), 233-240.*

B&L 3 Francis, A. L. & Driscoll, C. (2006). Training to use voice onset time as a cue to talker identification induces a left-ear/right-hemisphere processing advantage. *Brain and Language, 98(3), 310-318.*

B&L 4 Plante, E., Holland, S. K., & Schmithorst, V. J. (2006). Prosodic processing by children: An fMRI study. *Brain and Language, 97(3), 332-342.*

B&L 5 Pell, M. D., Cheang, H. S., & Leonard, C. L. (2006). The impact of Parkinson's disease on vocal-prosodic communication from the perspective of listeners. *Brain and Language, 97(2), 123-134.*

B&L 6 Eckert, M.A., Leonard, C. M., Possing, E. T., & Binder, J. R. (2006). Uncoupled leftward asymmetries for planum morphology and functional language processing. *Brain and Language, 98(1), 102-111.*

B&L 7 Barde, L., Schwartz, M. F., & Boronat, C. B. (2006). Semantic weight and verb retrieval in aphasia. *Brain and Language, 97(3), 266-278.*

B&L 8 Halliday, L. F., & Bishop, D.V.M. (2006). Is poor frequency modulation detection linked to literacy problems? A comparison of specific reading disability and mild to moderate sensorineural hearing loss. *Brain and Language, 97(2), 200-213.*

B&L 9 Weber-Fox, C., Hart, L. J., & Spruill, J. E.III (2006). Effects of grammatical categories on children's visual language processing: Evidence from event-related brain potentials. *Brain and Language, 98(1), 26-39.*

B&L 10 Chiarello, C., Lombardino, L. J., Kacinik, N. A., Otto, R., & Leonard, C. M. (2006). Neuroanatomical and behavioral asymmetry in an adult compensated dyslexic. *Brain and Language, 98(2), 169-181.*

B&L 11 Weems, S., & Reggia, J. (2006). Simulating single word processing in the classic aphasia syndromes based on the Wernicke–Lichtheim–Geschwind theory. *Brain and Language, 98(3), 291-309.*

B&L 12 Wible, C.G., Han, S.D., Spencer, M.H., Kubicki, M., Niznikiewicz, M.H., Jolesz, F.A., McCarley, R.W., & Nestor, P.G. (2006). Connectivity among semantic associates: An fMRI study of semantic priming. *Brain and Language, 97(3), 294-305.*

Journal of Speech, Language, and Hearing Research (JOS)

JOS 1 Richardson, J., Harris,L., Plante, E , & Gerken, L. A. (2006). Subcategory

Learning in Normal and Language Learning-Disabled Adults: How Much Information Do They Need? *Journal of Speech, Language, and Hearing Research, 49(6), 1257-1266.*

JOS 2 Mainela-Arnold, E., Evans, J.L.,& Alibali, M.W (2006). Understanding Conservation Delays in Children With Specific Language Impairment: Task Representations Revealed in Speech and Gesture. *Journal of Speech, Language, and Hearing Research, 49(6), 1267-1279.*

JOS 3 Luinge, M. R., Post W. J., Wit, H. P., & Goorhuis-Brouwer, S. M. (2006). The Ordering of Milestones in Language Development for Children From 1 to 6 Years of Age. *Journal of Speech, Language, and Hearing Research, 49(5),923-940.*

JOS 4 Gray, S. (2006). The Relationship Between Phonological Memory, Receptive Vocabulary, and Fast Mapping in Young Children With Specific Language Impairment. *Journal of Speech, Language, and Hearing Research, 49(5), 955-969*.

JOS 5 Newman, R.M., & McGregor, K.K. (2006). Teachers and Laypersons Discern Quality Differences Between Narratives Produced by Children With or Without SLI. *Journal of Speech, Language, and Hearing Research, 49(5), 1022-1036.*

JOS 6 Yoder, P., & Stone, W. L. (2006). A Randomized Comparison of the Effect of Two Prelinguistic Communication Interventions on the Acquisition of Spoken Communication in Preschoolers with ASD. *Journal of Speech, Language, and Hearing Research, 49(4), 698-711.*

JOS 7 Leonard, L., Camarata, S., Pawtowska, M., Brown, B., & Camarata, M. (2006).Tense and Agreement Morphemes in the Speech of Children With Specific Language Impairment During Intervention: Phase 2. *Journal of Speech, Language, and Hearing Research, 49(4), 749-770.*

JOS 8 Connor, C. M., & Craig, H.K. (2006). African American Preschoolers' Language, Emergent Literacy Skills, and Use of African American English: A Complex Relation. *Journal of Speech, Language, and Hearing Research, 49(4), 771-792***.**

JOS 9 Kashinath, S., Woods, J., & Goldstein, H. (2006). Enhancing Generalized Teaching Strategy Use in Daily Routines by Parents of Children With Autism. *Journal of Speech, Language, and Hearing Research, 49(3), 466-485.*

JOS 10 Shriberg, L. D., Ballard, K. J., Tomblin, J. B., Duffy, J. R., Odell, K. H., & Williams, C. A. (2006).Speech, Prosody, and Voice Characteristics of a Mother and Daughter With a 7;13 Translocation Affecting FOXP2. *Journal of Speech, Language, and Hearing Research, 49(3), 500-525.*

JOS 11 Martin, J.S., Jerger, J.F., Ulatowska, H.K., & Mehta, J.A. (2006). Complementing Behavioral Measures with Electrophysiological Measures in Diagnostic Evaluation: A Case Study in Two Languages. *Journal of Speech, Language, and Hearing Research, 49(3), 603-615.*

JOS 12 Plyler, P. N., & Fleck, E.L. (2006). The Effects of High-Frequency Amplification on the Objective and Subjective Performance of Hearing Instrument Users With Varying Degrees of High-Frequency Hearing Loss. *Journal of Speech, Language, and Hearing Research, 49(3), 616-627.*

Computational Linguistics (COMLING)

COMLING 1 Lapata, M. (2006). Automatic Evaluation of Information Ordering: Kendall's Tau. *Computational Linguistics, 32(4): 471-484.*

COMLING 2 Bestgen, Y. (2006). Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics, 32(1): 5-12.*

COMLING 3 Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics, 32(1): 13-47.*

COMLING 4 Merlo, P., & Ferrer, E.E.(2006). The Notion of Argument in Prepositional Phrase Attachment. *Computational Linguistics, 32(3): 341-377.*

COMLING 5 Kiss,T., & Strunk, J. (2006). Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics, 32(4): 485-525.*

COMLING 6 Girju, R.., Badulescu, A., & Moldovan, D. (2006). Automatic Discovery of Part–Whole Relations. *Computational Linguistics, 32(1): 83-135.*

COMLING 7 Inkpen, D., & Hirst, G., (2006). Building and Using a Lexical Knowledge Base of Near-Synonym Differences. *Computational Linguistics, 32(2): 223-262.*

COMLING 8 Ringlstetter, C., Schulz, K.U., & Mihov, S.(2006). Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. *Computational Linguistics, 32(3): 295-340.*

COMLING 9 Turney, P. D. (2006). Similarity of Semantic Relations. *Computational Linguistics, 32(3):* 379-416.

COMLING 10 Litman, D., Swerts, M., & Hirschberg, J.(2006). Characterizing and Predicting Corrections in Spoken Dialogue Systems. *Computational Linguistics, 32(3): 417-438.*

COMLING 11 Mariño, J.B., Banchs, R.E., Crego, J.M., de Gispert,A., Lambert, P., Fonollosa, J.A.R., & Costa-jussà, M.R.(2006). N-gram-based Machine Translation. *Computational Linguistics, 32(4):* 527-549.

COMLING 12 Navigli,R. (2006). Consistent Validation of Manual and Automatic Sense Annotations with the Aid of Semantic Graphs. *Computational Linguistics, 32(2): 273-281.*

Appendix 2: The Initial List of 60 Linguistic Features (Biber, 1995)

Linguistic Features	Explanations or Examples
past tense	Any past tense form that occurs in the dictionary
perfect aspect verbs	Perfect aspect forms mark actions in past time with current relevance
present tense	All VB (base form) or VBZ (third person singular present) verb forms in the dictionary,

	excluding infinitives.
Place adverbials	Aboard, above, across, ahead, behind etc.
Time adverbials	Afterwards, again, earlier, recently, previously, subsequently, etc.
First person pronoun	I, me, we, us, my, our, myself, ourselves.
Second person pronoun	You, your, yourself, yourselves (plus contracted forms)
Third person pronoun	She, he, they, her, him, them, his, their, himself, herself, themselves (plus contracted forms)
Pronoun IT	it
Demonstrative pronoun	this, that, these, those
Indefinite pronouns	Anybody, anyone, everybody, everyone, somebody, someone, etc.
DO as proverbs	Do as proverb substitutes for an entire clause (e.g. the subject did it.)
WH questions	>what<|>which<|>where<|>when<|>who
Nominalizations	All words ending in tion+ *ment+ *ness + *ity
gerunds	All participle forms serving nominal functions
Common nouns	All common nouns
Agent less passive	Verb to be+ VBN- (BY passives)
BY passive	Verb to be + VBN + by - phrase
BE as main verbs	Verb to be used as main verb
Existential THERE	There+be+nouns
THAT verb complements	e.g. I said that he came.
THAT adjective complements	I'm glad that you won.
WH clauses	I believed what you said.
Infinitives	To + base form of verb
past participial WHIZ deletion relatives object	e.g. The study conducted last year was approved.
present participial WHIZ deletion relatives	e.g.The event causing this decline is…
THAT relatives: subj position	e.g. The dog that bit me.
THAT relatives: obj positions	e.g. The questionnaire that I sent.
WH relatives: sub positions	e.g. The participant who has low reading proficiency,
WH relatives: obj positions	e.g. The man who I saw
WH relatives: pied pipes	e.g. the manner in which he was told
sentence relatives	e.g. He likes fired mangoes, which is the most disgusting thing I've ever heard of .
adv. Subordinator - cause	because
adv. sub. - concession	although, though
adv. sub. - condition	if, unless
adv. sub. - other	since, while, whilst, whereupon, whereas, whereby, such that, so that, as long as, as soon as
prepositions	all prepositions such as against, at, besides, by, despite,etc.
attributive adjectives	adjectives located in front of nouns (e.g. the important issue)

predicative adjectives	adjectives located after all linking verbs (e.g. the issue is important.)
adverbs	any adverb form occurring in the dictionary
type/token ratio	the number of lexical items in a text, dividing by the total numbers of words in the text, and multiplying by 100
word length	The number of characters in a text dividing by the total numbers of words in the text.
conjuncts	e.g. alternatively, althgether, consequently, conversely, furthermore, etc.
down toners	almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly, somewhat
General hedges	at about , something like, more or less, almost, maybe, sort of, kind of
amplifiers	absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensively, perfectly, strongly, thoroughly, totally, utterly very
emphatics	for sure, a lot, such a, real, just, really, most, more
discourse particles	well, now, anyway, anyhow, anyways
demonstratives	that, this, these, those
possibility modals	may, might, could, can
necessity modals	ought, should, must
predictive modals	will, would, shall
public verbs	acknowledge, admit, agree, assert , claim, complain, declare, deny, explain, hint, insist, mention, proclaim, promise, protest, remark, reply, report ,say, suggest , swear, write
private verbs	anticipate, assume, believe, conclude, decide, demonstrate, determine, discover, doubt, estimate, fear, feel, find, forget, guess, hear, hope, imagine, imply, indicate, infer , etc.
suasive verbs	agree, arrange, as, beg, command, decide, demand, grant, insist, etc.
SEEM/APPEAR	seem, appear
split infinitives	e.g. he wants to convincingly prove that…
split auxiliaries	e.g. they are objectively shown to…..
phrasal coordination	e.g. the participants were tested and asked to complete the questionnaires.
synthetic negation	no, neither, nor
analytic negation	not
pointer	The term used instead of text to direct readers to visual presentations (e.g. see Figure 2).
reference	The term used instead of parenthetical citations or non-integral citations (e.g. Hovy and Lin, 2003).

REFERENCES

Auria, M.P. (2008). Stance and Academic Promotionalism: A Cross-disciplinary Comparison in the Soft Sciences. *Journal of the Spanish Association of Anglo-American studies* 30 (1). (pp. 129-145).

Ayers,G. (2008). The evolutionary nature of genre: An investigation of short texts accompanying research articles in the scientific journal Nature. *English for Specific Purposes* 27(1). (pp. 22-41).

Barlow, M. (2004). *MonoConc Pro 2.2*. Texas: Athelstan.

Beatty, K. (2003). *Teaching and Researching Computer-assisted Language Learning.* Longman: Pearson.

Biber, D. (1995). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (2004). Modal use across registers and time. In Anne Curzan and Kimberly Emmons (eds.). *Studies in the history of the English language II: Unfolding conversations*. (pp. 189-216). Berlin: Mouton de Gruyter.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes* 5 (2). (pp. 97-116).

Biber, D.& Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*. Published by Elsevier Ltd. available online Jan 27, 2010.

Biber, D., Conrad, S, Reppen, R. Byrd, P., Helt,M., Clark, V., Cortes, V., Csomay, E.C. and Urzua, A. (2004). Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus**.** *ETS TOEFL Monograph Series, MS-25*. Princeton, NJ: Educational Testing Service.

Biber, D., S. Conrad, R. Reppen, P. Byrd, and M. Helt. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly*, 36. (pp. 9-48).

Burrough-Boenisch, J. (2005). NS and NNS scientists' amendments of Dutch scientific English and their impact on hedging. *English for Specific Purposes,* 24 (1). (pp. 25-39).

Carter-Thomas, S. & Rowley-Jolivet, E. (2008). If-conditionals in medical discourse: From theory to disciplinary practice. *English for Specific Purposes*, 7 (3). (pp. 191-205).

Chambers, A. (2005). Integrating corpus consultation in language studies. *Language and* Technology, 9(2). (pp. 111-125).

Charles, M. (2006). Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes* 25(3): 310-331.

Cobb, T. (2003). Is there any measurable learning from hands-on concordancing? *System*, 25 (3): 301-315.

Conrad, S. (1996). Investigating academic texts with corpus based techniques: An example from Biology. *Linguistics and Education*, 8. (pp. 229-326).

Dietsch, B.M., (2006). *Reasoning &Writing Well: A Rhetoric, Research Guide, Reader, and Handbook* 4th edition. McGraw Hill, New York.

Falahati, R. (2007). The use of hedging across different disciplines and rhetorical sections of research articles. In Nicole Carter, Loreley Hadic Zabala, Anne Rimrott & Dennis Storoshenko (Eds.). *Proceedings of the 22nd Northwest Linguistics Conference*

(NWLC) at Simon Fraser University (pp. 99 - 112). Burnaby, Canada: Linguistics Graduate Student Association.

Field, A. (2000). *Discovering Statistics using SPSS for Windows*. London – Thousand Oaks, New Delhi: Sage publications.

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing Errors? *System* 32(3). (pp. 301-319).

Gredhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, 19. (pp. 115-135).

Grieve, J. Biber, D., Friginal, E. and Nekrasova, T. Variation among blogs: a multidimensional analysis. In Mehler, Sharoff, Rehm and Santni (eds.). *Genres on the Web: Corpus Studies and Computational Models*. New York: Springer-Verlag.

Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *English for Academic Purposes*, 4(3). (pp. 257-277).

Halliday, M.(1994). *Introduction to functional grammar* (2nd ed.). London: Arnold.

Harwood, N. (2005) We Do Not Seem to Have a Theory . . . The Theory I Present Here Attempts to Fill This Gap': Inclusive and Exclusive Pronouns in Academic Writing. *Applied Linguistics*, 26, 3. (pp. 343–375).

Heffernan, J.A.W.,Linclon, J.E., Atwill, J. (2001). *Writing: A College Handbook*. (5th edition). W.W.Norton & Company. New York. (pp. 387-391).

Hunston, S. & Thompson, G. (2000). *Evaluation in Text: Authorial Stance and the Construction of Discourse***.** Oxford: OUP.

Hyland, K (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27 (1). (pp. 4-21).

Hyland, K. (1994) Hedging in academic writing and EAP textbooks. *English for Specific Purposes* 13 (3). (pp. 239-256).

Hyland, K. (1999). Disciplinary discourse: writer stance in research articles. In C. Candlin and K. Hyland (Eds.). *Writing: texts, processes and practices.* (pp. 99-121). Harlow: Addison -Wesley Longman.

Hyland, K. (2002). Authority and invisibility: Authorial identity in academic Writing. *Journal of Pragmatics*, 34: 109–112.

Hyland, K.& Tse, P. (2005). Hooking the reader: a corpus study of evaluative "that" in abstracts. *English for Specific Purposes*, 24. (pp. 123 – 139).

Kanoksilapatham B. (2003). *A Corpus-based Investigation of Biochemistry Research Articles: Linking Move Analysis with Multidimensional Analysis*. Unpublished Ph.D. thesis, Georgetown University, Washington, DC.

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approachesto corpus investigation. *Language Learning & Technology*, 5(3). (pp. 77-90).

Kettemann, B., & Marko, G. (Eds.). (2002). *Teaching and learning by doing corpus analysis.* Amsterdam, New York: Rodopi.

Lau, H. H. (2004). The structure of academic journal abstracts written by Taiwanese PhD students. *Taiwan Journal of TESOL*, 1(1). (pp. 1-25).

Li, L.J. & Ge, G.C. (2009). Genre analysis: Structural and linguistic evolution of the English-medium medical research articles (1995-2004). *English for Specific Purposes,* 28 (2). (pp. 93-104).

Lin, M. C. & Liou, H. C. (2006). Development of online materials for academic English writing: Contribution of text analysis on the discussion section of research articles. *Proceedings of the 23rd International Conference on English Teaching and Learning in the ROC* V. 2. (pp. 862-875).

Martinez, I. (2005). Native and non-native writers' use of first person pronouns in the different sections of biology research articles in English. *Journal of Second language Writing* 14 (1). (pp. 174-190).

Rayson, P.(2009). *CLAWS7* UCREL available from http://www.comp.lancs.ac.uk/ucrel/clasws7tags.html

Rietveld, T. & Van Hout, R. (1993). *Statistical Techniques for the Study of Language and Language Behavior*. Berlin, New York: Mouton de Gruyter.

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes*, 13 (2). (pp. 149-170).

Samad, A. (2004). Beyond concordance lines: Using concordances to investigate language development. *Internet Journal of e-Language Learning and Teaching,* 1 (1). (pp. 44-50).

Sun, Y.C. (2003). Learning process, strategies and Web-based concordancers: A case-study. *British Journal of Educational Technology*, 34(5). (pp. 601-613).

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Tang, R. & John., S. (1999). The ‘‘I’’ in identity: Exploring writer identity in student academic writing through the first person pronoun. *English for Specific Purposes* 18. (pp. 23-39).

Tarone, E., Dwyer, S., Gillette, S., & Icke, V. (1998). On the use of the passive and active voice in astrophysics journal papers: With extensions to other languages and other fields. *English for Specific Purposes*, 17 (1). (pp. 113-132).

Tutin, A. (2008). *Evaluative adjectives in academic writing in the humanities and social sciences.* Retrieved January 12, 2010 from http://w3u- grenoble3.fr/lidilem/lobo/file/ evalative_adjectives 2008_tutin.pdf..

Varttala, T. (1999). Remarks on the Communicative Functions of Hedging in Popular Scientific and Specialist Research Articles on Medicine. *English for Specific Purposes,* 18 (2). (pp. 177-200).

Vassileva, I. (2001). Commitment and detachment in English and Bulgarian academic writing. *English for Specific Purposes*, 20 (1). (pp. 83-102).

Wang, S. (1991). A corpus study of English conditionals. Unpublished MA thesis, Victoria University of Wellington.

Yoon, H., & Hirvela, A. (2004). ESL student attitudes towards corpus use in L2 writing. *Journal of Second Language Writing*, 13. (pp. 257-283).

Using translation corpora as a discovery procedure. The case of *discourse deictic retrospective labelling*

PATRICK GOETHALS
University of Ghent

Abstract
In this paper I show how the use of bidirectional translation corpora works as a discovery procedure. Quantitative data on translation shifts show us where to look for possible contrastive differences. This methodology is applied to Spanish and Dutch demonstratives. The data show that Dutch and Spanish behave differently in the specific context of discourse deictic uses of demonstrative modifiers. Whereas the Dutch distal may be used even to refer discourse-deictically to the last-mentioned segment in the preceding co-text, Spanish avoids the medial-distal ese in discourse deixis, although in other contexts it is the prototypically equivalent translation of Dutch distal die. The data show that the prototypical meaning of ese is not deictic.
Keywords: demonstratives, translation corpora, discourse deixis

Resumen
En esta contribución abogaré por el uso de los corpus bidireccionales de textos traducidos como procedimiento de descubrimiento. Los datos cuantitativos sobre los cambios de traducción sugieren pistas de investigación para el análisis contrastivo. La metodología se aplica a los demostrativos en español y neerlandés. Los datos muestran que los modificadores demostrativos españoles y neerlandeses se comportan de manera diferente en el contexto específico de la deixis discursiva (discourse deixis). En neerlandés, la forma para marcar distancia incluso puede usarse para referirse deíctico-discursivamente al último segmento mencionado en el co-texto. En español, en cambio, se evita casi sistemáticamente la forma ese en el contexto de la deixis discursiva, aunque en otros muchos contextos es la traducción prototípica de la forma de distancia en neerlandés. Los datos muestran que el significado prototípico de ese no es deíctico.
Palabras clave: demostrativos, corpus de textos traducidos, deixis discursiva

1. INTRODUCTION

In recent years, several authors have argued in favour of integrating translation corpora in linguistic or cross-linguistic research (Da Milano, 2006; Johansson, 2007). In this study I will use data from a Dutch (NL) – Spanish (ES) bidirectional corpus of translated texts, in order to show that Dutch and Spanish demonstratives behave differently in 'discourse deictic' contexts as exemplified below. In (1), the Dutch distal marker *die* is translated into Spanish by the proximal *esta*. The corpus analysis will reveal that this happens almost systematically.

(1) *NL Tijd is het systeem dat ervoor moet zorgen dat niet alles tegelijk gebeurt, die zin had ik een keer in een flits uit de radio gehoord. (Verhaal, §20)*

"Time is the system that must prevent everything from happening at once." I had caught that sentence in a flash, from the radio one day.

ES «El tiempo es el sistema que debe cuidar de que no ocurra todo al mismo tiempo»; esta frase la oí una vez captada al azar en la radio.

"Time is the system that must prevent everything from happening at once." I had caught this sentence in a flash, from the radio one day.

2. USING TRANSLATION CORPORA AS A DISCOVERY PROCEDURE

Empirical corpus research on translated texts is methodologically complex, since differences (or similarities) between ST and TT may be due to systemic contrastive differences, the translation process, or individual factors (e.g. personal preferences of the translator). Therefore, translation corpus research is a discovery procedure, or a way of generating hypotheses. When translated texts are used in the context of contrastive research, this may reveal certain facts in a faster and easier way than monolingual methodologies (Johansson, 2007), but these data will need to be complemented by further analysis.

In this paper I will show how the 'discovery procedure' works in the case of the distribution of Spanish and Dutch demonstrative modifiers. The translation corpus is a bidirectional translation corpus with 6 large samples (the average number of words is 24.000). This methodology has two major advantages. First, the fact that the corpus is bidirectional allows distinguishing between the effect of contrastive differences between the two language systems, and the effect of the translation process. Second, the fact that the corpus consists of large samples allows distinguishing between homogeneous effects, which occur in all samples of one translation direction, and heterogeneous effects, when there are important differences between the samples, and a more fine-grained analysis is required (Goethals, 2007).

The corpus includes three Dutch STs and their Spanish TTs, and three Spanish STs and their Dutch TTs (see annex). All the examples with a demonstrative modifier (this/that/… + NP) in ST and/or TT were extracted. In NL-ES there is a demonstrative modifier in both ST and TT in 61% of the examples. In 24% there is only a demonstrative modifier in the NL ST and in 15% only in the ES TT. In ES-NL there is a demonstrative modifier in both texts in 56% of the examples, in 10% only in the ES ST and in 34% only in the NL TT.

Table 1: Overall distribution of the demonstrative modifiers in the bidirectional corpus.

NL Source Text	**ES Target Text**				**ES Source Text**	**NL Target Text**		
dem. modif.	dem. modif.	367	61%		dem. modif.	dem. modif.	372	56%
dem. modif.	other	143	24%		dem. modif.	other	64	10%
other	dem. modif.	90	15%		other	dem. modif.	225	34%
		600					661	

In Goethals (2007), I focused mainly on the asymmetric examples, where there is a demonstrative modifier in only one of both texts. I found that frequently Dutch demonstrative modifiers are translated with or are the translation of a Spanish definite article. Moreover, this trend was found in all samples of the corpus. This means that it is a general trend that does not depend on the translation process, but that is presumably an overall contrastive difference between NL and ES. It can be seen as evidence for the hypothesis that NL demonstratives are more grammaticized than ES demonstratives, and therefore take over some of the roles that are assumed by the definite article in ES (see also Maes & Noordman, 1995: 279). On the other hand, I also found that demonstrative *pronouns* are translated by a demonstrative modifier + noun more often than vice versa, independently of which language was ST or TT. This trend was also found in all samples, and could be linked to the *translation* process, which would favour a more explicit rendering of the meaning (assuming that a demonstrative modifier + noun is more explicit than a demonstrative pronoun).

However, the shifts within the demonstrative modifier paradigm (i.e. between proximal and distal forms), seem to be far more complex, amongst other reasons because the tendencies are not homogeneous (i.e. contradictory tendencies are found in different samples). In fact, in Goethals (2007) I could only suggest that we should pay attention to the specific characteristics of certain samples of the corpus, or to specific uses of the demonstratives. The first objective was realized, in Goethals & De Wilde (2009), where we explain the deviant behaviour of one sample by its narratological characteristics. In this paper I will now try to make a first step in the analysis of specific uses of the demonstratives.

Let us look in detail at the distribution of the different demonstratives in the subgroup of examples where both ST and TT use a demonstrative modifier.

Table 2: Shifts within the demonstrative paradigm.

NL-ES				ES TT					
				este (proximal)		ese (medial-distal)		aquel (marked distal)	
					%		%NL		%NL
	NL ST	proximal deze/dit	75	55	**73%**	12	16%	8	11%
		distal die/dat	292	88	30%	169	**58%**	35	12%
			367						
ES-NL				ES ST					
				este (proximal)		ese (medial-distal)		aquel (marked distal)	
					%NL		%NL		%NL
	NL TT	proximal deze/dit	133	114	**93%**	15	9%	4	5%
		distal die/dat	239	8	7%	161	**91%**	70	95%
			372						

First, it is clear that *aquel* is a marked construction that is not a prototypical equivalent for one of the Dutch demonstratives (11% of the proximal and 12% of the distal). Secondly, it is clear that there are prototypical equivalences between the Dutch and the Spanish proximal forms, and between the Dutch distal and the Spanish medial-distal forms. Nevertheless, these prototypical equivalences differ considerably when we look at the exact frequencies, in decreasing order:

➢ proximal ES > proximal NL	93%
➢ medial-distal ES > distal NL	91%
➢ proximal NL > proximal ES	73%
➢ distal NL > medial-distal ES	58%

The prototypical equivalence with the lowest score is distal NL > medial-distal ES. This means that the Dutch distal form is frequently translated by other forms than the most prototypical equivalent, the Spanish medial-distal. Interestingly, the translation by the marked distal *aquel* does not compensate this (see table 2). Indeed, when we look at the following list of non-prototypically equivalent translations, we find that the translation of distal NL by proximal ES obtaines a high score:

➢ proximal ES > distal NL	7%
➢ medial-distal ES > proximal NL	9%
➢ proximal NL > medial-distal ES	16%
➢ distal NL > proximal ES	30%

It is important to acknowledge that, although the quantitative data reveal a shift from Dutch distal towards Spanish proximal, they do not explain it. If it were a general contrastive difference (for example that, in general, Dutch distals are more 'proximal' than Spanish medial-distals), we would also expect a higher score in the opposite translation direction (i.e. proximal ES > distal NL), but this is not the case (7%). Or if it were a general translational phenomenon (for example, that translations are typically more 'proximal' than source texts), we would expect to find a similar 'proximization' in the subcorpus ES-NL. Again, this is not the case (9%). As a conclusion, we can say that the 'discovery procedure' shows us where to look (the subgroup Distal NL > proximal ES), but does not reveal yet what we will find.

This is finally how examples such as (1) called our attention, for the corpus analysis revealed that this was not an isolated case. Examples (2) and (3) are similar cases of discourse deictics, with a distal in Dutch and a proximal in Spanish:

(2) *NL* *de verbindingen […], het bouwsel, de schoonheid. Alweer dat woord, maar er is niets aan te doen. (Verhaal §13)*
[…] the connections, the structure, the beauty. That word again, but it cannot be helped.
ES *las conexiones, la construcción, la belleza. De nuevo esta palabra; pero no hay nada que hacer.*
[…] the connections, the structure, the beauty. This word again, but it cannot be helped.

(3) *NL* *'Laat jij je cremeren?' vroeg ik. Met die vraag kun je in elk gezelschap terecht. (Verhaal §39)*
"Do you want to be cremated?" I asked. That question works wonders in any company.
ES *-¿Vas a hacer que te incineren? -pregunté. Con esta pregunta tienes éxito asegurado en todas las reuniones.*
"Do you want to be cremated?" I asked. This question works wonders in any company.

The translational data suggest that Spanish avoids the medial-distal *ese* in this particular kind of examples. In the following section, I will try to delineate as precisely as possible this group of examples.

3. DISCOURSE DEICTIC RETROSPECTIVE LABELLING: AT THE INTERSECTION BETWEEN DISCOURSE DEIXIS AND RETROSPECTIVE LABELLING.

Two terms have been used to describe examples such as (1)-(3): "textual/discourse deixis" (a.o. Levinson, 2004; Lyons, 1977), and "retrospective metalinguistic labelling" (Francis, 1994; Botley, 2006). Since both terms have a broader scope, I will propose a blend, namely

"discourse deictic retrospective labelling". Since the terminology is often very complex and even confusing , I will briefly discuss these notions[1].

The first notion is "discourse or textual deixis". (4) and (5) are typical examples given in the literature:

(4) '*You are wrong'. That's exactly what she said.* (Levinson, 2004: 23)

(5) I *bet you haven't heard this story: ...* (Botley, 2006: 80)

"Discourse deictics" constitute a special group because they are situated at the intersection between exophores (deictics) and endophores. Like endophores, discourse deictics presuppose the presence of a co-text. This is not the case with prototypical exophores such as *I give you this book*, where the book is physically present. But the relation that discourse deictics establish with the co-text differs from the endophoric relation with the co-text. Endophores establish a correferential link: the demonstrative refers to a concept or a referent that has previously been referred to (Lyons, 1977: 668). Instead of this correferential link with the co-text, discourse deictics establish a deictic link. They do not refer to a previously evoked referent, but to a part of the physically present co-text as such (see also Macías Villalobos, 2000: 55-6).These examples bring Levinson (2004) to the claim that "it is clearly not sufficient to distinguish simply between exophoric (deictic) and endophoric (non-deictic) […] since discourse deixis is intra-text but deictic" (Levinson, 2004).

There are three uses[2] that all authors seem to include in the category of "textual/discourse deixis": references to another discourse segment (4-5, and 6), to the form of what is said ("pure textual deixis", Lyons, 1977: 667-8) (7), and references to the speech act that is realized in the preceding co-text, or to the mental process that sustains it (8):

(6) 'Projects are also introducing changes in teaching styles. […]' That quotation comes not from the Plowden report, but from […]. (Francis, 1994: 93)

(7) – That's a rhinoceros.- A what? How do you spell that? (Lyons, 1977: 637)

(8) A> I disagree.
U> dumbass
A> That's not very nice of you. I did nothing to deserve that kind of language. (internet chat)

[1] I will use the examples given by the authors, since the confusing complexity of the field is not only terminological. Often, similar examples are grouped in different ways, or similar taxonomic categories are applied to very different examples.

[2] I will not consider self-referential expressions (e.g. *This sentence has five words*) because they exclude by definition the use of non-proximal forms. So, it would be rather trivial to say that Spanish avoids the use of *ese* in these contexts.

Other examples in my opinion do not fulfil the restrictive definition of "discourse deixis". In my opinion, examples such as (9)-(10) are correferential relations:

> (9) —I've never seen him.
> — That's a lie. (Lyons 1977:668) (see Eguren, 1999: 937 for a similar example)
>
> (10) Right at that moment the three boys come walking … (example quoted in Doiz-Bienzobas, 2003)

Lyons (1977:668) described (9) as "*impure* textual deixis", but I think that it is better to exclude them from this category: what we find is a correferential relation with an idea that is evoked previously, in the proposition as a whole. In (10) the demonstrative refers to the temporal setting, which was not explicitly expressed in the previous co-text. However, the example does not fulfil the definition of discourse deixis. What we find is a correferential relation with the conceptual world that was (implicitly) evoked in the co-text. Summarizing, in my view, (9-10) do not establish a deictic link. They establish a correferential link, a rather complex correferential link indeed, because it concerns an idea expressed by the whole proposition and not a referent expressed by a nominal syntagm, or the implicitly evoked temporal frame rather than the explicit content of what was said. However, this more complex endophoric nature should not leave us to put them in the same category as discourse deictics, which establish a link with an element that is physically present, as a linguistic signifier, or really taking place, as a speech act.

The term *retrospective labelling* is borrowed from Francis (1994), who uses the term to describe examples such as (6), quoted above.

4. EMPIRICAL DATA

Of the 88 non-prototypically equivalent distal NL > proximal ES translations (table 2), 12 correspond to the definition of discourse deictic retrospective labelling, which is an important number, given the very specific character of this use. In order to evaluate these data, I extracted from the corpus all other examples that correspond to the definition (another 19 examples). This gives the following distribution:

Table 3: Discourse deictic retrospective labelling.

		proximal ES (*este*)	medial-distal ES (*ese*)	marked distal ES (*aquel*)
NL-ES	proximal NL	3	0	1
	distal NL	12	2	0
ES-NL	proximal NL	5	0	1
	distal NL	2	3	2
Total	proximal NL	8	0	2
	distal NL	14	5	2

Comparing these data with the overall frequencies of the demonstrative forms, we find a clear overrepresentation of *este* in the group of discourse deictic retrospective labelling (71%, against the overall 35,9% result). *Ese* is clearly underrepresented (16,1% versus overall 48,3%).

Table 4: Frequencies of *este*, *ese* and *aquel*.

	discourse deictic retrospective labelling		total	
este	22	71%	265	35,9%
ese	5	16.1%	357	48,3%
aquel	4	12.9%	117	15,8%
	31		739	

The overrepresentation of *este* is not due to an interference with Dutch, since an important number of examples are translations of Dutch distals. This means that the inherent semantic restriction of the Spanish *ese* is important enough to provoke a translational shift.

A special kind of examples are those where the discourse deictic function is realized by the addition of a complement such as *laatste/último* (*last, the former*). In Dutch, even in these contexts it is possible to use the distal form. In Spanish, only the proximal form is used:

> (11) *NL Er zijn boeken om te verkopen en boeken om te houden. Wat die laatste betreft, […].*
> There are books to sell, and books to collect. Regarding those last ones (the latter), […].
> *ES Hay libros para vender y libros para guardar. En cuanto a estos últimos, […]. (Dumas 108)*
> There are books to sell, and books to collect. Regarding these last ones (the latter), […].

The following example is similar, although strictly speaking it is a pronominal use (therefore, it was not included in the quantitave data above):

(12) *NL* *Ik kon zelfs betalen met echt geld. Dat laatste was wat mij betreft het overtuigendste. (Verhaal §2)*
And I was even able to pay with real money. That last, as far as I was concerned, was the most convincing evidence of all.
ES *Podía pagar incluso con dinero de curso legal. Esto último fue, por lo que a mi respecta, lo más convincente.*
And I was even able to pay with real money. This last, as far as I was concerned, was the most convincing evidence of all.

In table 3, we also found 5 cases of medial-distal *ese*. However, all 5 'counter-examples' are complex cases where the link with the preceding co-text seems is both deictic and anaphoric. Example (13) is a typical example:

(13) *NL* *En in die stilte klonk dat ene idiote woord waarmee de leerlingen mij altijd benoemden. 'Sokrates.'*
Het wilde iets, dat woord. (Verhaal §114)
ES *Y en ese silencio sonó aquella palabra idiota con la que los alumnos siempre me nombraban. Sócrates.*
Quería algo, esa palabra.
The silence was broken by that idiotic name my pupils always called me. "Socrates."
It wanted something, that word.

If we only consider the link *Sócrates-esa palabra* (*Socrates-that word*), this example is similar to the previous examples. But if we look at the broader context, we find that the metalinguistic reference was already thematized and evoked in the segment *aquella palabra idiota*. This means that *esa palabra* not only referes discourse-deictically to *Sócrates* but also establishes a purely correferential relation with the segment *aquella palabra idiota*.

5. Discussion

Let us now consider which conclusions can be drawn from the restriction on the use of *ese*. Frequently, it is said that *ese* is neutralized and does not longer express a strong contrast with *este* or *aquel* (Charaudeau, 1992; De Kock, 1992). This would explain why *ese* cannot be used in the prototypical contrastive uses of *este* and *aquel* such as (14):

(14) En julio de 1968, el precio del café […] había bajado un treinta por ciento […]. Sin embargo, el consumidor norteamericano no pagaba más barato su café, sino un trece por ciento más caro. Los intermediarios se quedaron, pues […] con este trece y con aquel treinta: ganaron a dos puntas. (Venas §193)

But does neutralization also explain why Spanish avoids *ese* in examples (1-3)? This does not seem to be correct. In the examples with a shift between the Dutch distal and the Spanish proximal, a contrastive reading is not relevant, for the Dutch distal does not allow it either. Moreover, it makes not much sense to say that it is a matter of neutralization, since the Dutch distal is more neutralized than the Spanish medial-distal, up to the point that it can even be combined with *laatste* (*último/the latter*), or that it is frequently translated by a definite article in Spanish (Goethals, 2007; or Maes & Noordman, 1995 on the grammaticization of the Dutch distal).

Rather, the data seem to suggest that the semantics of *ese* conflicts with the *deictic* character of the relation as such (see also Gómez Díez, 2009 or Jungbluth, 2005). The contrast with the Dutch distal, nevertheless, suggests that the non-deictic nature of *ese* is not necessarily a kind of neutralization, but that perhaps it is better understood as full-fledged semantic meaning. Following this interpretation, *ese* orients us towards *conceptualizations* which were evoked through the co-text (or conceptualizations tout court, through *deixis ad phantasma*) and not to referents which are identified deictically.

6. CORPUS[3]

[*Mede*] de Vries, A. 1984. *Medeplichtig.* Lemniscaat. [Translated by M. Monteagudo Romero & M. García Sendón: *Cómplice*. SM]

[*Venas*] Galeano, E. 1973. *Las venas abiertas*. Siglo Veintiuno. [Translated by M. Sabarte Belacortu: *De aderlating van een continent*. Van Gennep]

[*Verhaal*] Nooteboom, C. 1991. *Het volgende verhaal.* [Translated by J. Grande: *La historia siguiente*. Siruela]

[*Omweg*] Nooteboom, C. 1992. *De omweg naar Santiago*. [Translated by J. Grande: *El desvío a Santiago.* Siruela]

[*Dumas*] Pérez Reverte, A. 1993. *El Club Dumas*. Alfaguara. [Translated by J. Schalekamp: *De Club Dumas*. De Prom]

[*Fiesta*] Vargas Llosa, M. 2000. *La fiesta del Chivo.* Alfaguara. [Translated by A. van der Wal: *Het feest van de Bok*. Meulenhoff]

[3] Since it is an electronic corpus, the number of the paragraph is given instead of the page number. Please send an e-mail to the author if you want to make use of the corpus.

REFERENCES

Botley, S. P. (2006). Indirect anaphora. Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics,* 11(1). (pp. 73-112).

Charaudeau, P. (1992). *Grammaire du sens et de l'expression*. Paris: Hachette Éducation.

Da Milano, F. (2007). Demonstratives in parallel texts: a case study. Parallel Texts: Using translational equivalents in linguistic typology. *Special Issue of Sprachtypologie und Universalienforschung,* 60(2). (pp. 135-147).

De Kock, J. (1992). Los pronombres demostrativos en registros análogos y diferentes. In J. De Kock, C. Gómez Molina & R. Verdonk (Eds.), *Los pronombres demostrativos y relativos* (pp. 11-90). Salamanca: Ediciones Universidad de Salamanca.

Diessel, H. (1999). *Demonstratives: Form, Function, and Grammaticalization*. Amsterdam, Netherlands: Benjamins.

Doiz-Bienzobas, A. (2003). An Analysis of English, Spanish and Basque Demonstratives in Narrative: A Matter of Viewpoint. *International Journal of English Studies (IJES),* 3(2), (pp. 63-84).

Eguren, L. (1999). Pronombres y adverbios demostrativos. Las relaciones deícticas. In I. Bosque & V. Demonte (Eds.), *Gramática Descriptiva de la Lengua Española* (pp. 929-972). Madrid: Espasa.

Francis, G. (1994). Labelling discourse: an aspect of nominal-group lexical cohesion. In M. Coulthard (Ed.), *Advances in Written Text Analysis* (pp. 83-101). London: Routledge.

Goethals, P. (2007). Corpus-driven Hypothesis Generation in Translation Studies, Contrastive Linguistics and Text Linguistics. A case study of demonstratives in Spanish and Dutch parallel texts. *Belgian Journal of Linguistics,* 21, (pp. 87-104).

Goethals, P., & De Wilde, J. (2009). Deictic center shifts in literary translation: the Spanish translation of Nooteboom's *Het Volgende Verhaal*. *Meta,* 54(4), (pp. 770-794).

Gómez Díez, I. (2009). Los demonstrativos en español: ¿un sistema ternario? Análisis cuantitativo de un corpus de teatro español contemporáneo. In R. de Maeseneer & e. al. (Eds.), *El hispanismo omnipresente: homenaje a Robert Verdonk* (pp. 183-197). Brussel: University Press Antwerp.

Gutiérrez-Rexach, J. (2002). Demonstratives in context. In J. Gutiérrez-Rexach (Ed.), *From Words to Discourse: Trends in Spanish Semantics and Pragmatics*. Amsterdam: Elsevier.

Johansson, S. (2007). *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Jungbluth, K. (2005). *Pragmatik der Demonstrativpronomina in den iberoromanischen Sprachen*. Tübingen: Niemeyer.

Levinson, S. (2004). Deixis. In L. Horn & G. Ward (Eds.), *The Handbook of Pragmatics* (pp. 97-121): Blackwell.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Macías Villalobos, C. (2006). *El demostrativo en Miguel Delibes*. San Vicente de Raspeig: Taller Digital de Establecimiento de Textos Literarios y Científicos.

Maes, A., & Noordman, L. (1995). Demonstrative nominal anaphoras: a case of nonidentificational markedness. *Linguistics,* 33, (pp. 255-282).

“estas minhas limitadas cifras tenham a felicidade de acharem a VMce. desfrutando aquela saúde espiritual e corporal tão feliz como lhe deseja o meu afecto” - Different perspectives on correspondance conventionalities

MARIANA GOMES
LEONOR TAVARES
ANA RITA GUILHERME
University of Lisbon

Abstrac
This paper arises from our current experience of making the critical electronic edition of Portuguese private letters involving more than 2,000 men and women from all social strata who lived in the Modern era (16th to 19th century). The letters are being treated as sources for various intellectual approaches, including Historical Pragmatics (Jacobs and Jucker, 1995), the approach that is relevant for this paper. The final aim of the project, though, is to articulate different methodologies in analysing texts in an effort to conjure the old “murmuring voices of societies”, as Michel de Certeau called the creative behaviour of individuals in everyday life.
Keywords: Historical Pragmatics, Discourse Corpus, formulaic expressions, private letters, philology, electronic edition

Resumen
Este artículo presenta nuestra actual experiencia de edición crítica electrónica de cartas privadas portuguesas incluyendo más de 2.000 hombres y mujeres de todos los estratos sociales y que han vivido en la época Moderna (siglos XVI a XIX). Las cartas son tratadas como fuentes para varias perspectivas intelectuales, incluyendo la Pragmática Histórica (Jacobs and Jucker, 1995), la perspectiva más relevante en este artículo. El objetivo último de nuestro proyecto, sin embargo, es articular metodologías diferentes en el análisis de textos intentando evocar "murmuring voices of societies" ("voces murmuradoras de las sociedades"), como Michel de Certau ha llamado al comportamiento creativo de los individuos en su vida cotidiana.
Palabras clave: Pragmática Histórica, corpus del discurso, expresiones estereotipadas, cartas privadas, filología, edición electrónica

1. SOURCES FOR THE STUDY OF HISTORICAL PRAGMATICS

Historical Pragmatics has been attracting those scholars who show an interest on the situated uses of language in history; as Fitzmaurice and Taavitsainen (2007:1)put it, “˜it˜ offers insights into earlier communicative practices, registers, and linguistic functions as gleaned from historical discourse”. The discipline follows, very naturally, the current trends in linguistic inquiry, so it has adopted the same cautious method of dealing with large quantities of data – i.e. computer-based corpora – since this posture manifestly allows for linguistic

analyses to present themselves as empirically reliable (Taavitsainen and Fitzmaurice, 2007:16–17).

However, as all pragmaticists promptly recognize, a linguistic corpus with encoded contextual information – hence language usage information – is no easy object to obtain (Kohnen, 2007). In the depths of a corpus marked-up for parts of speech, for instance, the unity of the texts to which language utterances once belonged becomes totally lost. Lost is also, in this way, the meaning language usage got from the discursive and extra-linguistic surroundings it belonged to.

So we should ask: – When was it that the historical study of language had a systematic way of accounting for texts in context? Answer: – When it was called Philology and didn't diverge from literary studies or textual criticism. Now, the subsequent question would be: – How did those old partners react to computerized technology and the subsequent facility of accounting for large quantities of data? – Well, literary studies can rely today on their own electronic corpora, i.e. hyper textual editions of literary texts prepared by textual critics who gained computer science competences (*cf*., for instance, Robinson, 1997). It becomes clear that the methodology used today on digital critical editions should be useful, in principle, to Historical Pragmatics research.

Reasoning like this, we prepared a diachronic corpus of Portuguese letters following conventions currently used within Textual Criticism to produce digital literary editions: these conventions belong to the well-known project TEI, *Text Encoding Initiative*, and to a variant of it, DALF, *Digital Archive of Letters in Flanders*, especially conceived for the letters' genre critical edition.

2. THE CORPUS FOR THIS STUDY

Our sources were carefully chosen as relevant sources from the point of view of Historical Pragmatics: the speakers who wrote them came from several social backgrounds and the letters were written on several different local contexts. That kind of complete representativeness was firstly made possible because the manuscripts we use were largely kept within judicial court-files from the Portuguese Inquisition (1536-1822) and the Portuguese Royal Appeal Court (late 18th century-1836), surrounded with documented testimonies on the accused individuals who wrote or received the written correspondence.

The information on social backgrounds and local contexts, being thus partially recoverable, gets also encoded in the electronic critical edition. By following Textual

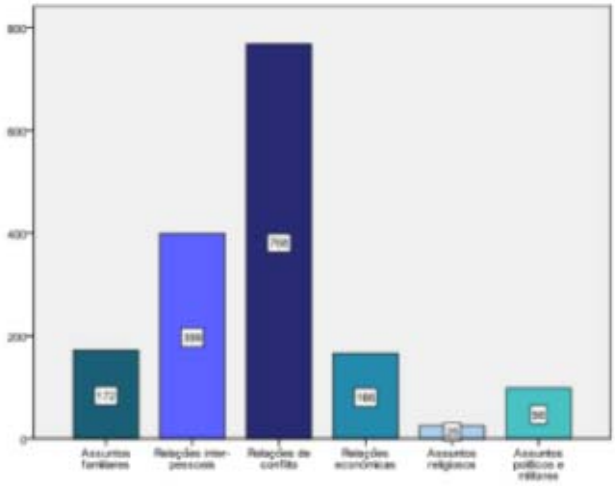
Criticism methodologies while keeping in mind social analyses preoccupations, we manage to connect such diverse information as manuscript's physical layout, original writing, authorial emendations, editorial conjectures, information on variants (when copies or comments also survived), information on the context of the letter's composition (an event within a social context), and information on the participants' biographies (*cf*. Table 1).

Table 1: Abbreviated description of the corpus

The *corpus* includes the following main fields (with several sub-divisions): - Letter title - Sponsor and Funder - Project Identification - Archival Identification - Letter's Identifier (author, receiver, scribe, annotator, time and place) - Letter's synopsis - Letter's context - Letter's transcription The manuscripts' transcription is quasi-diplomatic, allowing only for regular word separation. Authorial strikethroughs and additions, line-breaks, lacunae, difficult deciphering, abbreviations and non-orthographic uses are respected in the transcription by means of XML tags specifically developed by TEI to fit primary sources idiosyncrasies. External to this electronic support, but linked to it by anchors, there is a database containing biographic information on authors and addressees. As for the corpus dimensions, it contains, for now, 450.000 words, more than 1.600 letters and information on more than 2.000 participants (several letters were written or received by the same person).

Of course the themes of letters apprehended by court officials, as can be seen on Table 2, were different, in terms of frequency, from the ones figuring in the total number of letters that actually circulated in those times: letters related to conflict relations, for instance, are much more frequent in our corpus than the banal example of letters on family affairs:

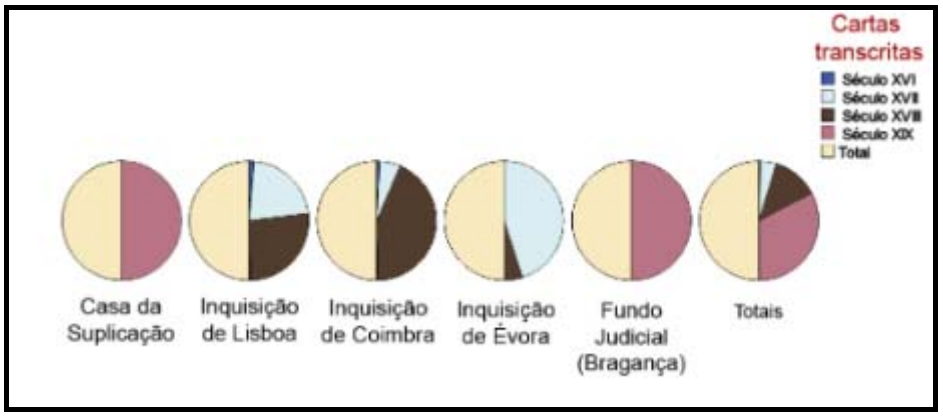
Table 2: Thematic distribution of letters within the *corpus*



172 – Family affairs 399 – Interpersonal relations 768 – Conflict relations 166 – Economic relations 25 – Religion affairs 98 – Politics and military affairs
Total – 1.628 (these results go back to October 2009)

On the other hand, the sample has a large geographical representation, along with a fairly good chronological one (Table 3). The Royal Appeal Court (*Casa da Suplicação*) judged lawsuits from all over the reign and from the overseas empire; the Lisbon Inquisition covered the reign's central area and the overseas territories; the Coimbra Inquisition covered the reign's north, and the Évora Inquisition covered the south. We used also some few lawsuits from the late XIXth century coming from the north-eastern part of the reign (Bragança):

Table 3: A real and chronological distribution of the transcribed letters



3. AN ANALYSIS OF FORMULAIC EXCERPTS WITHIN THE LETTERS

One of the most important groups of tags we use in our corpus, considering the Historical Pragmatic point of view, is the protocol/eschatocol group. Letters can be highly conventionalized in matters of textual openings and closings. Although the ones from the Modern period didn't keep exactly the mediaeval structure codified in the ars dictaminis tradition (Guillén, 1986), one can still find a large amount of fossilized expressions surrounding the expository text in those 16th to 19th century letters. This means that for matters of automatic search, looking for linguistic structures, for instance, within the letters' text, it is important to hide either the formulaic part of the corpus, or its expository one.

The system we adopted was partly dictated by DALF conventions, which allow for a mark-up of the external part (in visual terms) of letters – date, place, addressing formula – and partly dictated by TEI conventions, which has a tag for text segments that can receive an arbitrary attribute; the chosen attributes for the segments were precisely "harengue", for the formulaic introduction, and "peroration", for the formulaic conclusion within le letters' visual body. An example goes like this (Tables 4 and 5):

Table 4: Protocol (external opener+harengue) from a letter sent to her mother by a Brazilian farmer's daughter, 18th century

```
                    <opener>
        <salute>Minha May e<abbr>Snra</abbr></salute>
                    </opener>
<p><seg type="formulaicText" n="harengue">estimarei que
estas limitadas regra<lb/>a axhem a <abbr>vme</abbr>
asistida di perfeita saude em companhia<lb/>de Meu Pay e
<abbr>Snro</abbr> e de toda a nobre caza para disporem da
minha<lb/> que o prezente he boa seja Deus
louvado<lb/></seg>
```

My mother and lady
I hope these few lines will find you in perfect health in the company of my father and lord and of all the noble house, so that you can dispose of mine that is good at present, praise the Lord

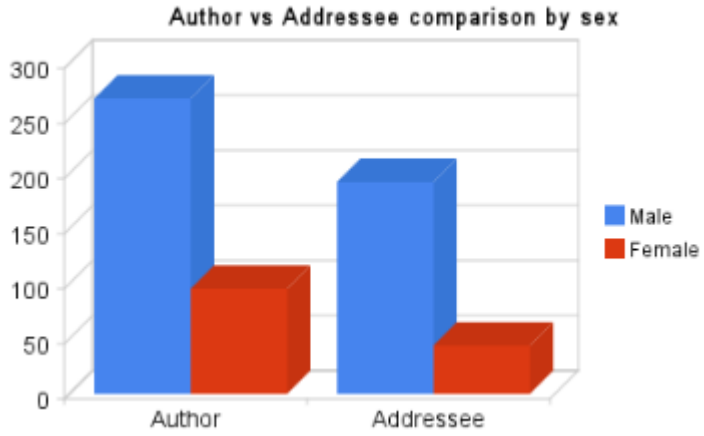
Table 5: Eschatocol (peroration+external closer) from a letter sent from a Portuguese architect to a mathematician, 18th century

```
<seg type="formulaicText" n="peroration">que Deus Guarde
por felices annos <seg></p>
<closer>
<salute>De <abbr>VMce</abbr><lb/> Muito attento Venerador,
e Verdadeiro amigo</salute>
<signed>Mathias Francisco de Pazos</signed>
</closer>
```

... may the Lord guard (Your Mercy) for happy years
The venerating attentive and true friend of Your Mercy
Mathias Francisco de Passos

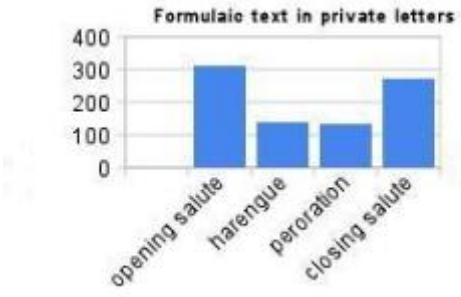
This sort of mark-up allows for an automatic distinction between letters presenting formulaic utterances, along with letters missing them. We have analyzed a sample of 182.000 words, 362 letters, from the late 16th (11), 17th (100), 18th (188) and early 19th century (63). Besides from the chronological one, the sample has also some balanced representativeness in terms of gender (Table 6). Male authors (268) and receivers (281) are more present than female ones, but the feminine universe is nevertheless there, with 94 women-authors and 44 women-receivers.

Table 6



Looking for the repetitive expressions in their beginnings and endings, we arrived at the following results (Table 7): 449 protocols *vs* 423 eschatocols; 309 opening salutations *vs* 289 closing salutations; 138 harengues *vs* 132 perorations.

Table 7



So far, we could conclude that, in a socially balanced sample of private letters written in Early Modern and Late Modern Portuguese history, i) the presence of formulaic utterances is the rule, while its absence is the exception; ii) opening formulaic utterances are more frequent than closing ones; iii) given the letter body, external pieces of those repetitive

expressions (salutations) are more frequent than the internal ones (harengues and perorations). A cross-examination of the authors' biographies gives a good indication of people from low social strata apparently haranguing and perorating a lot in letters, while *élite* letter writers largely dispensed the strategy. (Precise numbers on this point will be given at the Conference).

Although it goes without saying that we are dealing with a narrow sample, some temporary interpretations can be made regarding the formulaic behaviour of our letter writers.

On the one hand, physical factors must not be discarded: the blank area while beginning a letter is forcefully wider than the available one when ending it; practical writers must put up with that, finishing the letter inside the available space, and thinking no more of it. But besides that chance factor pending on spontaneous reactions, those letter writers were also social actors engaged in verbal and non verbal deferred interaction, so the formulaic parts in their texts must also have been partly meaningful to them. This means that the semantic fields covered by the utterances from the protocols and eschatocols in the sample must also have been relevant in the pragmatic behaviours of those time.

Semantic classification of historical texts can be very biased, so we followed the same choice made by Archer and Culpeper (2009) in their study of language usage in a close-to-spoken *corpus* similar to our own, namely dialogues in plays and trial proceedings of the 17th and 18th centuries; this means we also applied the *UCREL Semantic Analysis System* (USAS), and selected from it the following relevant categories:

Semantic Group	**Sub-classification**
B: The Body and the individual	**B2:** Health and diseases
E: Emotional actions, states and processes	**E2**: Liking
Q: Linguistic actions, states and processes	**Q1.1.:** Communication in general
S: Social actions, states and processes	**S1.2.4.:**Politeness **S3.1**.:Relationship in general **S4:** Kin **S7.2**.: Respect **S9:** Religion
X: Psycological actions, states and processes	**X1**: General

The presence of these semantic fields in the sample's formulaic utterances is summed-up in tables 9 and 10:

Table 9: Semantic fields in private letters' protocols

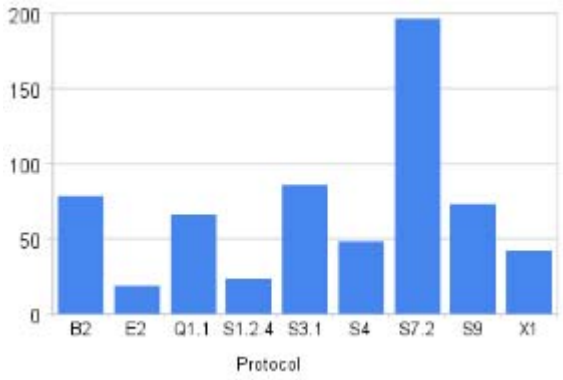
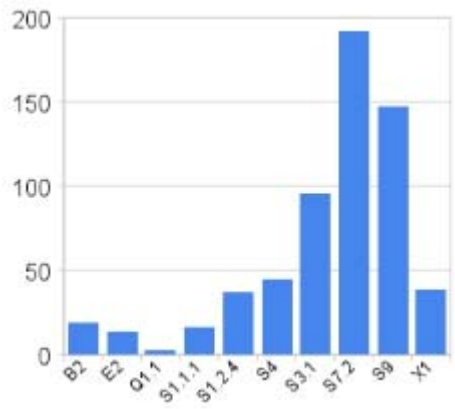


Table 10: Semantic fields in private letters' eschatocols



Description:

(i) In the textual segment protocol, the semantic field of "respect" stands out – it appears 196 times - followed by terms related with general relations such as friendship – 86 times. The "health" semantic field is also very prominent – it appears 79 times in the harengue segment-, as does the "religious" semantic world – 73 times.

(ii) Other semantic fields found in the protocol are: "communication in general" (66); "kin" (48); general expressions related with "psycological actions, states and processes" (42); "politeness" (23); and "liking emotions" (19).

(iii) All the above mentioned semantic fields were also found in the closing formulaic textual segment- the eschatocol. And, similar to what happened in protocols, the semantic field of "respect" is the most expressive – it appears

192 times –, followed by the "religion" semantics – 147 times. Reference to "health" is less expressive than in the harangue – only 19 times. Also not very expressive are expressions concerning "communication in general" (3 times).

(iv) The other semantic fields in the eschatocol have the following distribution: terms related with general relations such as friendship – 96 times; "kin" words – 45 times; references to "psycological actions, states and processes" appears 39 times; "politeness" expressions such as "não enfado mais" (*I won't bother you longer*) appear 37 times in the *corpus*; "goodbye" expressions as "adeus, adeus" were counted 16 times and "liking emotions", 14 times.

(v) Taking into account all the semantic fields observed in the formulaic segments of the letters, the most expressive group is the one concerned with Social Actions, States and Processes: the most prominent semantic fields found in our corpus, such as "respect", "religion", "kin" words and terms expressing general relations fall under the scope of that larger group.

Although this is a strictly descriptive study, we think this new kind of data can be very informative to Historical Pragmatics in terms of textual linguistics (Adam, 1992/2005), politeness in language (Brown and Levinson, 1987) and discourse analysis (Fairclough, 2003, Van Dijk, 1997). We will bring to the Conference some final remarks on those theoretical implications.

REFERENCES

Adam, Jean-Michel (1992/2005). *Les textes: types et prototypes. Récit, description, argumentation, explication et dialogue.* Paris: Nathan.

Archer, Dawn Archer and Jonathan Culpeper (2009). Identifying key sociophilological usage in plays and trial proceedings (1640–1760): an empirical approach via corpus annotation. *Journal of Historical Pragmatics,* 10:2. pp. 286–309.

Brown, Penélope and Stephen C. Levinson (1987). *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press (2nd ed.)

Certeau, Michel (1990/2002). *L'Invention du quotidien, 1. Arts de faire*. Paris: Gallimard.

DALF, http://www.kantl.be/ctb/project/dalf/dalfnl.htm (accessed Feb. 15, 2010).

Fairclough, Norman (2003). *Analysing Discourse: Textual Analysis for Social Research*. Londres: Routledge.

Fitzmaurice, Susan, and Irma Taavitsainen (2007). Introduction. In Susan M. Fitzmaurice and Irma Taavitsainen (eds.). *Methods in Historical Pragmatics*. Berlin/New York: Mouton de Gruyter, (pp. 1–10).

Guillén, Claudio (1986). Notes towards the study of the Renaissance letter. In Barbara Kiefer Lewalski (ed.). *Renaissance Genres: Essays on Theory, History and Interpretation*. Cambridge, MA: Harvard University Press, (pp. 70–101).

Jacobs, Andreas, and Andreas Jucker (1995). The historical perspective in pragmatics. In Andreas H. Jucker (ed.). *Historical Pragmatics: Pragmatic Developments in the History of English*. Amsterdam/Philadelphia: John Benjamins, (pp. 3–33).

Kohnen, Thomas (2007). Text types and the methodology of diachronic speech act analysis. In Susan M. Fitzmaurice and Irma Taavitsainen (eds.). *Methods in Historical Pragmatics*. Berlin/New York: Mouton de Gruyter, (pp. 139–166).

Robinson, Peter (1997). A stemmatic snalysis if the fifteenth-century witness to The Wife of Bath's prologue. In Blake and Robinson (eds.). *The Canterbury Tales Project's Occasional Papers*. Oxford: Office For Humanities Communication, (pp. 69–132).

Taavitsainen, Irma and Susan Fitzmaurice (2007). Historical pragmatics: what is it and how to do it?. In Susan M. Fitzmaurice and Irma Taavitsainen (eds.). *Methods in Historical Pragmatics*. Berlin/New York: Mouton de Gruyter, (pp. 11–36).

TEI, http://www.tei-c.org/index.xml (accessed Feb. 15, 2010).

UCREL Semantic Analysis Systems, http://ucrel.lancs.ac.uk/wmatrix2.html (accessed Feb. 15, 2010).

Van Dijk, Teun A. (1997). The study of discourse. In Teun A. van Dijk (ed.). *Discourse as Structure and Process*. Londres: Sage, (pp. 1–34).

Los usos de *tocar* en la construcción transitiva

FITA GONZÁLEZ DOMÍNGUEZ

Universidade de Vigo

Resumen

En este breve estudio haremos un análisis de los significados del verbo español tocar *en la construcción transitiva SD, dentro de la base de datos ADESSE. Como veremos, el análisis de las combinaciones léxicas sintagmáticas del verbo será la clave para la identificación de nuevos significados, los cuales, en este caso, no dependerán tanto de la naturaleza del sujeto, como de la del objeto. Por lo tanto, el significado estará motivado en gran medida dependiendo si se toca una* bombilla*,* fondo*, un determinado* tema *o* asunto*,* madera*, un* instrumento musical *o un* timbre*, etc. Aparte de la combinatoria léxica, para la identificación de los significados tendremos en cuenta el significado léxico del verbo y de la construcción, además de factores contextuales y pragmáticos.*

Palabras clave: combinatoria léxica, esquemas sintáctico-semánticos, el verbo tocar, esquema transitivo SD

Abstract

I*n this brief study we will make an analysis of the meanings of the Spanish verb* tocar *in the transitive construction SD, by means of the ADESSE database. The analysis of the lexical syntagmatic combinations of the verb will be the key for the identification of new meanings, which, in this case, will depend more on the nature of the object than on that of the subject. Therefore, the meaning will depend to a great extent on the type of objects touched:* bombilla*, the* fondo, *a certain* tema *or* asunto*,* madera*, an* instrumento musical *or a* timbre*, etc. Apart from the lexical combinatory, for the identification of the meanings we will take into account the lexical meaning of the verb and the construction, as well as contextual and pragmatic factors.*

Keywords: lexical combinatory, syntactic - semantic schemes, the verb to touch, SD transitive scheme

1. INTRODUCCIÓN

Lejos de pretender poner límites a los significados, nuestro objetivo es hacer un análisis riguroso de los diferentes sentidos que puede adoptar el verbo español *tocar* cuando se combina con el esquema transitivo SD. Consideramos insuficiente la mera enumeración de los sentidos o subsentidos de un verbo o el saber identificarlos, por eso, pondremos énfasis en la tarea de interrelacionarlos, con el propósito de ir tejiendo la red conceptual que nos rodea. No se trata, únicamente, de establecer lazos entre los usos de un lema, sino con los sentidos de otras formas verbales, los cuales, en ocasiones, se aproximan más que entre los propios usos.

1.1 La base de datos ADESSE

Este análisis nace dentro del proyecto ALEXSYS (Anotación Léxica, Sintáctica y Semántica de corpus del español)[1], el cual posee la base de datos ADESSE (Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español) con anotación sintáctico-semántica de los verbos del español del corpus ARTHUS de 1,5 millones de palabras. Las casi 159.000 cláusulas del corpus han sido anotadas sintácticamente en la BDS (Base de Datos Sintácticos del español) y la labor en ADESSE ha sido la de añadir información semántica[2]:

-Esquemas sintáctico-semánticos de cada verbo

-Clasificación semántica de sentidos verbales

-Identificación de papeles participantes

al mismo tiempo que introdujo las frecuencias de uso, información que permite obtener datos empíricos sobre la interacción entre léxico y construcciones.

Además de la riqueza de anotación que presenta el corpus, en ALEXSYS, a la vez que nos ocupamos de la parte puramente lexicográfica, estamos añadiendo los núcleos léxicos de cada uno de los argumentos verbales. Todo ello con la finalidad de obtener para cada verbo una completa caracterización de su significado y de su combinatoria sintagmática.

2. ANÁLISIS DE LOS DATOS DEL CORPUS: SIGNIFICADOS DE TOCAR

2.1. Tocar y el esquema transitivo SD

El esquema más frecuente con el que se combina *tocar* en el corpus es el transitivo SD. De los 262 ejemplos analizados, 119 se combinan con este esquema, es decir, aproximadamente un 73.4%.

La construcción transitiva implica la existencia de dos participantes que presentan una relación asimétrica donde el sujeto es el participante más activo y el objeto es la entidad sobre la que recae la acción del verbo. La acción es unidireccional: se orienta del sujeto al objeto. (Langacker, 1991a:309-313). Si a la información semántica del esquema le añadimos cómo son los argumentos prototípicos exigidos por el verbo, inferimos que generalmente el sujeto es una entidad animada y que el objeto es prototípicamente concreto, ya que *tocar* implica el *contacto físico entre dos entidades*. Los datos de la tabla muestran dichas expectativas,

[1] El proyecto ALEXSYS recibe financiación de Ministerio de Ciencia e Innovación (FFI2008-01953/FILO) [2009-2011]. Puede ser de interés el resumen general del diseño de la base de datos ADESSE en García-Miguel y Albertuz 2005 o su página Web http://adesse.uvigo.es/.

aunque hay ejemplos que, como veremos con detalle, no se ajustan a los "cánones prototípicos" y será de interés estudiarlos de manera pormenorizada, puesto que posiblemente este hecho sea un indicio de cambio semántico.

Tabla 1: Rasgos semánticos del esquema SD con *tocar*

Sujeto	Objeto Directo
Animado 113 Concreto 6	*Concreto 93* Animado 15 Abstracto 11

2.2. Tocar: un acercamiento a sus usos transitivos

Para análisis de los significados hemos seguido dos perspectivas que se complementan:

-En primer lugar, hemos agrupado los ejemplos con un nivel de abstracción general.

-En segundo lugar, el proceso es el contrario, no se trata de generalizar, sino de especificar los diferentes usos concretos que adquiere el verbo en una determinada construcción.

La información léxica y sintáctica anotada en el corpus ha sido la base de este estudio.

Tabla 2: Perfil combinatorio de *tocar1* en el esquema SD

Función sintáctica	SUJETO	OBJETO DIRECTO		
Tipo semántico	Animado 113/Concreto 6	Concreto 93/Animado 15/Abstracto 11		
Realizaciones léxicas frecuentes (no animados)	*mano* 2 *rayo* 2 *sol* 1 *hierro* 1	*piano 14* *bolsa* 4 *tema* 4 *hoja* 2 *timbre* 2 *fondo* 2 *arpa* 1 *fuente* 1	*pito 10* *guitarra* 4 *tierra* 3 *puerta* 2 *mueble* 2 *reloj* 1 *balón* 1 *madera1*	*tambor* 1 *música* 1 *violín* 1 *armónica*1 *flauta* 1 *órgano* 1 *bombilla* 1 *bocina* 1

Esta información nos ofrece una visión de la combinatoria léxica de *tocar*, que nos permite de forma orientativa identificar los usos más generales, pero como el significado no lo aporta únicamente la semántica léxica del verbo, sino un conjunto de factores circunstanciales, para determinar todos los significados es fundamental el estudio del uso en

el contexto en el que se produce: *The basic units are occurrences of the word in context* (Kilgarriff, 1997: 108).

Tras la primera fase de agrupación y generalización de sentidos verbales, hemos diferenciado:

-*Tocar* como verbo que implica *contacto* (*Tocar 1*)

- *Tocar* como verbo que implica *emisión* de sonido (*Tocar 2*)

El primer grupo focaliza la relación física entre dos entidades, mientras que en el segundo lo relevante es la emisión de sonido y el tipo de sonido. En el caso de *tocar 2*, la noción de contacto pasa a ser un componente opcional dentro del *frame* (Fillmore, 2003) activado. Cuando alguien dice que *sabe tocar la guitarra*, nosotros interpretamos que conoce la técnica para hacerla sonar de manera melódica; cuando *alguien toca el timbre* no significa que lo esté toqueteando sin más, sino que lo hace sonar, posiblemente para que le abran la puerta.

Una vez que conocemos el perfil combinatorio de *tocar* con el esquema transitivo SD, teniendo en cuenta las características semánticas de los dos grupos sabemos que:

Tocar 1 presenta las siguientes posibilidades combinatorias:

-Alguien toca a alguien /algo (concreto o abstracto)

-Algo toca algo concreto

Donde el sujeto siempre se va a corresponder con el argumento responsable de la acción verbal, en este caso, al ser de la clase de contacto, el *contactante (A1)* y el objeto con el *contactado (A2)*.

Y *tocar 2*:

-Alguien toca algo (concreto o abstracto)

En este caso el sujeto también se corresponde siempre con el papel de *iniciador, pero la* selección de argumentos del objeto directo varía dependiendo de si es concreto o abstracto, es decir, si focalizamos el instrumento (*toca la guitarra*) o el sonido (*toca música*). Cuando perfilamos el "instrumento" el papel semántico se corresponde con el de emisor, mientras que si a lo que damos relevancia es al "sonido" estamos ante el argumento emisión. Por lo tanto, la construcción SD con *tocar 2* presenta dos realizaciones argumentales diferentes: S=iniciador (A0) y D=emisor (A1) o emisión (A2)

A continuación con el análisis de los ejemplos identificaremos los significados que puede tener cada construcción en el corpus.

2.3. Análisis de los significados de tocar 1

Alguien toca a alguien:

Esta construcción es una generalización o abstracción de ejemplos como:

> -¿Puedo tocarlo (se refiere a un animal)? A condición de que me dejes tocarte a ti [PAI:58,10]
> - Aunque Dorothy fuera la única mujer disponible, yo no podría tocarla. Eres mi hermano [CIN:97,10]
> -Con sus cincuenta y cinco años sin haber tocado mujer [SON:219,11]
> -Tocándole, acariciándole. [ZOR:35,29]
> -¡No estoy aquí, en Florencia! ¡No me toquéis, soltad! [COA:16,21]

Todos ellos se ajustan a la definición del DEA: "Llegar con las manos, o con otra parte del cuerpo, o con un objeto que uno sostiene, a una persona o cosa de la manera que se pueda sentir su presencia o alguna cualidad física", donde *tocar* manifiesta el *contacto* entre dos entidades. Pero de cada uno de ellos hacemos una interpretación diferente, por lo tanto, estamos ante usos particulares, donde *tocar* adquiere otros matices de significado.

En este caso, como los participantes del evento son animados, el significado del verbo dependerá del *tipo de contacto físico* que mantenga el *contactante* con el *contactado*. Por nuestra experiencia sabemos que el contacto entre seres animados puede ser:

-cariñoso. Próximo a la semántica de verbos como *acariciar*, *besar*, etc.

> -(Tocándole, acariciándole.) Estás fuerte, porque te alimentas como ellos [ZOR:35,29]
> -¿Puedo tocarlo (se refiere e un animal)? A condición de que me dejes tocarte a ti [PAI:58,10]

Incluso algunos ejemplos tienen claras *connotaciones sexuales*:

> -Aunque Dorothy fuera la única mujer disponible, yo no podría tocarla. Eres mi hermano, Javi... Mi hermano. [CIN:97,10]
> -Con sus cincuenta y cinco años sin haber tocado mujer, pero con aquellos ojos azules que te miraban y te adivinaban, te sacaban los pensamientos [SON:219,11]

-violento. Próximo al significado de golpear, etc.

> -¡No me toquéis, soltad! ¿Por qué, por qué me sujetáis así? [COA:16,21]

En este ejemplo se observa un contacto más agresivo, en donde se interpreta brusquedad en la acción. Concretamente, el evento muestra violencia y está presente a la vez el significado de *agarrar*.

Esto pone de manifiesto que el verbo *tocar* en la construcción transitiva, con sujeto y objeto animados, puede adquirir matices de significado propio de otros verbos como: *acariciar, golpear, agarrar*. Estas diferencias semánticas están en parte motivadas por aspectos extralingüísticos, como puede ser, el tipo de relación existente entre los participantes.

Alguien toca algo concreto:

Este esquema es el más frecuente, puesto que los participantes tienen las características semánticas exigidas tanto por el verbo como por la construcción (sujeto: animado, objeto: inanimado-concreto).

El corpus nos ofrece una amplia variedad de objetos que son tocados por personas, pero *tocar* no siempre se limita a establecer el contacto entre dos entidades con la finalidad de comprobar o percibir algo a través del sentido del tacto (característica exclusiva de los sujetos humanos), como en:

> -Toca su bolsita de amuletos y vuelve a mirarse al espejo [SON:98,10]
> -Toqué las bombillas y comprobé que estaban frías [LAB:92,20]

Sino que, una vez más, hay ejemplos en los que el significado adquiere matices diferentes:

> -Tu relojito. ¡No habrá que tocarlo ni en veinte años! [CAI:48,10]
> -Porque el zapatero no puede tocar clases con su lezna: sólo toca zapatos individuales [LIN:67,10]

En estos usos *tocar* no implica únicamente: relación física de contacto entre dos entidades, sino que interpretamos que *el reloj no será necesario tocarlo* con el significado de *arreglarlo*. Y lo mismo le sucede al segundo ejemplo, ya que sabemos que lo *que le hace el zapatero a los zapatos en la lezna* es *arreglarlos*.

> -Toca madera [HOT:20,17]

En el caso de *tocar madera*, a pesar de que la acción de la entidad animada se limita a acercar su mano a algún objeto de madera, por nuestro conocimiento del mundo, sabemos que la interpretación que hacemos del evento es más compleja.

Según María Moliner, *tocamos madera:*

> Para desvirtuar un supuesto maleficio; como hacen, por ejemplo, las personas supersticiosas cuando se pronuncia en su presencia la palabra culebra u otra equivalente. Se usa mucho simbólicamente como comentario cuando se menciona algo que puede significar mala suerte o puede tener malas consecuencias.

A pesar de inferir todas esta información, el verbo *tocar* no varía su significado, puesto que cuando utilizamos esta expresión, realmente la acompañamos de la acción de poner la mano sobre algo de madera.

Este significado complejo lo inferimos si pertenecemos a una determinada cultura y en una situación determinada. Pero no viene dada ni por la semántica del verbo ni por la de la construcción.

-Toca fondo [JOV:14, 29]

Estamos ente una construcción locucional, donde, a diferencia de lo que pasa con *tocar madera*, la idea de *contacto* entre dos entidades es metafórica.

Tocar fondo el DUE define como: "Llegar a su peor momento una persona, situación, etc., a partir del cual puede empezar su recuperación"

Esto sería un ejemplo de *metáfora orientacional* (Lakoff, 1991: 50), donde nuestra concepción occidental del mundo nos hace ver que las cosas buenas y positivas están arriba, en un eje vertical, y las malas abajo. Esta construcción es fija y su significado es cultural y está basado en una manera determinada de interpretar la realidad. Significado que relacionamos con la idea de *contacto*, pero ya no físico y desde luego el concepto de *fondo* nos sitúa en una parte concreta del espacio, a la que en este caso, atribuimos connotaciones negativas.

Alguien toca algo abstracto:

Nos situamos en una esfera diferente cuando lo que *tocamos* son *cosas abstractas*. Ya no estamos ante un significado literal, sino que se trata de una extensión metafórica.

-En medio de la noche fueron la señal de que habíamos llegado al final del viaje, de que habíamos tocado fondo [CAR:174,14]
-Alguna vez, al tocar este tema, me pregunté: ¿es realmente moderna la literatura latinoamericana? [TIE:162,9]

El significado de *tocar un tema* o *asunto* hace referencia a la relación que mantiene una persona cuando *habla* o *escribe* sobre *algo*. Nos referimos a una relación intelectual, puesto que no hay contacto físico, pero la relación con el significado literal viene dada porque el sujeto establece un tipo de vínculo abstracto con el objeto. Estamos ante un nuevo significado figurado: Tratar o hablar de [un tema].

Algo toca algo concreto:

-A lo lejos, el sol está casi tocando una fuente [1INF: 071, 17]
-Mira, los rayos del sol ya casi tocan las hojas del último árbol. [1INF: 024, 07]

En estos ejemplos *tocar* presenta el significado más general, definido por el DEA como: "Estar una cosa al lado de otra de manera que en algún punto no queda ningún espacio de separación".

En los usos del corpus se trata en todos los casos de ejemplos metafóricos, puesto que el *sol* se podría interpretar como una entidad animada y los *rayos* se identificarían con los *brazos*. Pero, a pesar de ello, se trata de una relación estática y durativa, que la diferencia de los ejemplos vistos en los que el sujeto era humano. Cuando dos *cosas concretas se tocan* se pierden los matices de significado como el de *manipulación*, *control*, el de *tocar* para *apreciar* o *sentir*, de ahí que sea una significado más general.

2.4. Análisis de los significados de tocar 2

Alguien toca algo concreto:

El significado de tocar 2 también varía dependiendo de si el objeto es un instrumento musical u otro objeto con el que se puedan emitir sonidos, pero no melódicos. Estamos, pues, antes significados diferentes:

Hacer sonar un objeto como un *timbre*, un *pito*, una *bocina* o algo similar.

-Suponte que era un tipo que hubiera agarrado y hubiera tocado el timbre [BAI:51,10]

Hacer sonar un *instrumento musical*.

-Mi hermano Luis Enrique, que entonces tocaba la guitarra como un profesional, improvisó [CRO:69,11]
-Y él toca los platillos [BAI:486,31]
-Tocaba el arpa maravillosamente [DIE:130,25]

Alguien toca algo abstracto:

El único con objeto abstracto que se *toca* presente en el corpus es la *música*. A diferencia de *tocar un timbre*, o *tocar la guitarra* donde se focaliza el objeto con el que se produce el sonido. En el caso de *tocar música*, todos sabemos que el referente real que se *toca* es un instrumento y que estamos ante una relación de metonimia, donde se da prominencia al resultado por el instrumento. Este elemento referencial (música) es lo que llama Langacker (1991b: 189-201) *zona activa*.

Tocar música presenta un significado relacionado con el de emisión (*hacer sonar*), pero añade el matiz de "interpretar una pieza musical con un instrumento".

-Tocábamos música también cuando nos reuníamos en casa [MAD:245,21]

3. CONCLUSIONES

Tocar en la construcción transitiva SD presenta, dentro del corpus de trabajo, dos sentidos diferentes: contacto (tocar1) y emisión (tocar2).

Dentro de *tocar1* la estructura transitiva SD se corresponde siempre con una misma estructura semántica, donde el sujeto es el contactante y el objeto la entidad contactada. Pero a diferencia esto, el esquema transitivo SD se corresponde con dos esquemas semánticos, dependiendo de si el objeto es el instrumento (emisor) o el sonido (emisión).

Teniendo en cuenta la semántica del verbo y de la construcción, así como, la naturaleza de los participantes y el contexto de uso, observamos que tanto *tocar1* (contacto) como tocar2 (emisión) presentan diferentes significados.

Como verbo de contacto *tocar* presenta los siguientes usos ene l corpus:

-*Contacto físico entre dos entidades concretas*. Es el sentido más general del verbo.

Este significado se hace más específico en los casos en los que la entidad que desempeña la acción verbal es animada. Diferenciándose los siguientes usos:

-Llegar con la mano u otra parte del cuerpo hasta [algo o alguien] de manera que se pueda sentir su presencia o alguna cualidad física

Cuando las entidades son animadas el significado del evento puede tener *connotaciones de cariño o de violencia*.

Y, en ocasiones también se aproxima al significado de verbos propios del campo conceptual del *control* como *agarrar*.

-Alterar, modificar o cambiar el estado o condición [de algo o alguien]

-Tocar pasa a tener el significado de *tratar* o *escribir*, cuando lo que se *tocan* son *temas*, *asuntos*, etc. es un significado metafórico creado a partir del significado literal de *contacto físico*.

A partir del significado literal de de tocar como verbos de contacto se derivan usos metafóricos, dando lugar a locuciones como *tocar fondo,* en donde la asociación es figurada.

Y como verbo de emisión *tocar* presenta los siguientes usos, motivados fundamentalmente por la naturaleza del objeto:

-Hacer sonar un timbre, bocina o algo similar.

-Hacer sonar un instrumento musical de manera melódica.

-Interpretar una pieza musical.

Los significados, por lo tanto, son abstracciones que hacemos de escenas de la realidad, que están motivados por aspectos sintáctico, léxicos, contextuales y pragmáticos. Y todos ellos se relacionan de manera gradual, a la vez que se entrelazan con significados propios de otros verbos.

REFERENCIAS BIBLIOGRÁFICAS

ADESSE. *Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español* (Universidad de Vigo): http://adesse.uvigo.es/

DEA (1999). *Diccionario del Español Actual*. Madrid, Aguilar.

DUE (2007). *Diccionario de Uso del español*. Madrid, Gredos.

Fillmore Ch. J., Johnson Ch., Petruck M. (2003). "Background to FrameNet", *International Journal of Lexicography* 16/3: 235-250.

Langacker, R. W. (1991a). *Foundations of Cognitive Grammar*, Vol. II: Descriptive Application. Stanford, Stanford Univ. Press: 282-329.

Langacker, R. W. (1991b). “Active Zones”, *Concept, Image and Symbol. The cognitive Basis of Grammar,* Berlin: Mouton de Gruyter, 189-201.

Kilgarriff, A. (1997). "I don't believe in word senses". *Computers and the Humanities,* 31, pp. 91–113.

Lakoff, George y Johnson M. (1991). *Metáforas de la vida cotidiana*, Madrid: Cátedra.

García-Miguel, J.M. y Albertuz F.J. (2005). "Verbs, semantic classes and semantic roles in the ADESSE project", en Erk, Katrin; Alissa Melinger & Sabine Schulte in Walde (eds): *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbrücken, 28 febrero - 1 marzo 2005, pp. 50-55.

La Wikipedia como fuente multilingüe de corpus comparables

ISAAC GONZÁLEZ LÓPEZ
PABLO GAMALLO OTERO
Universidade de Santiago de Compostela

Resumen
En este artículo se describe un método automático de selección de corpus comparables a partir de la Wikipedia, utilizando categorías temáticas como elementos restrictivos. Nuestra estrategia se fundamenta en dos propiedades de la Wikipedia: el ser un recurso multilingüe y el tratarse de una enciclopedia libre disponible para descarga en formato XML. Las herramientas y los corpus generados dispondrán de licencia libre GPL (General Public License).
Palabras clave: Corpus comparables, extracción de información multilingüe

Abstract
This article describes an automatic method to select comparable corpora from Wikipedia using categories as topic restrictions. Our strategy is based on two properties of Wikipedia: to be a multilingual resource and to be a free encyclopedia available in a XML file. Tools and corpus will be distributed under GPL license (General Public License).
Key words: Comparable corpora, multilingual information extraction

1. INTRODUCCIÓN

La Wikipedia es una enciclopedia libre multilingüe *online* y colaborativa con entradas para alrededor de 300 lenguas, de las cuales, el inglés es la más representativa con casi 3 millones de artículos. Como se puede observar en la tabla 1, el número de entradas/artículos en las lenguas más usadas de la Wikipedia alcanza ya un nivel más que suficiente para poder llevar a cabo investigación multilingüe con solvencia. La tabla muestra que el español se encuentra en noveno lugar, con 460 mil entradas, muy cerca del portugués, que alcanza las 470 mil. Nuestras primeras experiencias se realizarán con estas dos lenguas.

En consonancia con la enorme expansión y rápido crecimiento de la Wikipedia, en los últimos años han surgido numerosos trabajos que explotan este recurso para diferentes objetivos multilingües: extracción de diccionarios bilingües (Yu & Tsujii, 2009; Tyers & Pienaar, 2008), alineamiento, paralelización y traducción automática (Adafre & Rijke, 2006; Tomás et al. 2008), recuperación de información multilingüe (Potthast et al., 2008). Por último, comienzan a aparecer trabajos sobre la comparabilidad de los artículos en diferentes

lenguas de la Wikipedia, con la posibilidad de elaborar, a partir de ellos, corpus comparables (Filatova, 2009).

Tabla 1: *Presencia de las lenguas en la Wikipedia. Las 10 primeras lenguas ordenadas en función del número de artículos (datos de abril 2009).*

Lenguas	**Número de artículos**
Inglés	2.826.000
Alemán	888.000
Francés	786.000
Polaco	593.000
Italiano	556.000
Japonés	576.000
Holandés	528,000
Portugués	470.000
Español	460.000
Ruso	376.000

Un corpus comparable se compone de textos en diferentes lenguas con una temática similar (McEnery & Xiao, 2007). Una selección de artículos de periódicos en diferentes lenguas, tratando del mismo tema y en la misma época supone un buen ejemplo de corpus comparable. Este tipo de corpus es útil en múltiples y variadas líneas de investigación. Por citar sólo algunas, puede servir para ayudar a realizar estudios contrastivos, o bien como base para la extracción automática de léxicos y terminologías bilingües, así como fuente de entrenamiento de sistemas de traducción automática (cuando hay falta de corpus paralelos en el par de lenguas objetivo del sistema). Una de las principales características de los corpus comparables es su enorme expansión. A diferencia de los corpus paralelos, que exigen la traducción de una lengua a otra, los corpus comparables se encuentran, de forma natural, en la web, a medida que aumenta el número de lenguas presentes en este medio. De hecho, como ya ha sido mencionado antes, la Wikipedia es una fuente natural de corpus comparables. Sólo es preciso construir las herramientas adecuadas para extraerlos.

Aprovechando las potencialidades multilingües de la Wikipedia, el objetivo de este artículo es describir un método para extraer de ella corpus comparables en función de dos parámetros de variabilidad: las lenguas concretas que se quieran escoger y el tema seleccionado. En concreto, dadas dos lenguas y un tema, nuestra estrategia crea un corpus con textos en las dos lenguas escogidas que versan sobre la temática seleccionada. Las herramientas y los corpus generados con ellas dispondrán de licencia libre GPL (*General Public License*) y estarán disponibles para descarga en: *http://gramatica.usc.es/pln*

Este artículo se organiza del siguiente modo. La sección 2 describe el modo en que transformamos la Wikipedia en un corpus codificado, que llamamos CorpusPedia. La sección 3 presenta las estrategias de construcción de corpus comparables a partir de la CorpusPedia. En la sección 4, se ofrecen datos empíricos de la CorpusPedia, así como de algunas experiencias realizadas a partir de las estrategias definidas en 3. El artículo se cierra con algunas consideraciones acerca de las nuevas tareas que pretendemos llevar a cabo para ampliar y mejorar nuestras herramientas.

2. CORPUSPEDIA

La primera parte de nuestro método consiste en transformar los ficheros fuente de la Wikipedia en un conjunto de ficheros con un formato de fácil manipulación: la CorpusPedia. Para ello desarrollamos herramientas que permiten descargar automáticamente la Wikipedia en los idiomas requeridos y aplicar después un proceso de conversión del XML descargado al XML formato del corpus.

2.1. Formato de la Wikipedia

La Wikipedia es descargable en su totalidad en ficheros XML, en distintas versiones en la que se puede escoger que cantidad de metadatos descargar. Como se puede ver en el ejemplo de una entrada de la Wikipedia que se muestra en la figura 1; la forma de acceso a los datos es también simple y muy eficiente. La diferencia respecto a la web al uso es que el texto en vez de tener un formato plano, html o xhtml, tiene un formato propio de la Wikipedia, que ha de ser convertido a texto sin formato.

En la figura 1 podemos ver el extracto de una entrada de la Wikipedia en formato XML, si nos fijamos en el campo text, podremos observar el formato específico del que hablábamos.

```
<page>
    <title>Arqueoloxía</title>
    <id>3</id>
    <revision>
      <id>1310468</id>
      <timestamp>2009-10-06T02:42:14Z</timestamp>
      <contributor>
        <username>SieBot</username>
        <id>2109</id>
      </contributor>
      <minor />
      <comment>bot Engadido: [[ku:Arkeolojî]]</comment>
      <text xml:space="preserve">{{Historia en progreso}}

A '"arqueoloxía"' é a [[ciencia]] que estuda as [[arte|artes]], [[monumento|monumentos]] e [[obxecto]]s da [[antigüidade|antigüidade]], especialmente a través dos seus restos. O nome ven do [[lingua grega|grego]] "archaios", &quot;vello&quot; ou &quot;antigo&quot;, e "logos", &quot;ciencia&quot;, &quot;saber&quot;.

[...]

[[zh:考古学]]
[[zh-yue:考古]]</text>
    </revision>
  </page>
```

Figura 1*:* Ejemplo del formato XML de la Wikipedia (artículo gallego "Arqueoloxía").

2.2. Formato de la CorpusPedia

El formato de la CorpusPedia consiste, esencialmente, en un título y el texto de cada entrada, junto con otras informaciones que se extraen gracias al formato semiestructurado de la Wikipedia y de ciertas convenciones entre los editores (Clark et al., 2009).

En la figura 2 podemos observar el código XML del corpus generado a partir de la Wikipedia. Los campos *title*, *category* y *plaintext* son los necesarios relativos al uso de este fichero como corpus comparable, siendo el apartado *category* el que puede aportar información sobre la temática del texto y así agruparlo con otros textos de categoría similar en caso de búsqueda de alta comparabilidad. El campo *wikitext* contiene el formato original de la Wikipedia, que puede ser útil para futuras extracciones y del cual, aplicando un parser, se obtiene el campo *plaintext* (i.e., texto plano sin codificación).

```
<article>
  <title>Arqueoloxía</title>
  <category>Arqueoloxía</category>
  <related>Antropoloxía, Arqueoloxía industrial, Arqueoloxía submarina</related>
  <links>ciencia, arte|artes, monumento|monumentos, obxecto, antigüidade|antigüidade, lingua grega|grego, cultura, estudo, psicolóxico, condutistas, antropoloxía, idade de pedra, Idade Media, Arqueoloxía industrial, Antropoloxía, Arqueoloxía industrial, Arqueoloxía submarina</links>
  <translations># Arqueologia Arqueología Archaeology Archéologie Arqueologia Arkeologia # Archeologia Archeologie Археология Αρχαιολογία</translations>
  <plaintext>A arqueoloxía é a ciencia que estuda as artes, monumentos e obxectos da antigüidade, [...] o que se coñece como Arqueoloxía industrial.</plaintext>
  <wikitext>{{Historia en progreso}}

A '''arqueoloxía''' é a [[ciencia]] que estuda as [[arte|artes]], [[monumento|monumentos]] e [[obxecto]]s da [[antigüidade|antigüidade]], [...]
[...]
[[yi:ארכעאלאגיע]]
[[zh:考古学]]
[[zh-yue:考古]]</wikitext>
</article>
```

Figura 2: Formato del corpus generado a partir de la Wikipedia

El campo *translations* es una lista de los enlaces *interlanguage*, es decir un enlace a esa misma entrada pero en otro idioma; lo que aporta una herramienta muy útil para crear comparabilidad como veremos más adelante. Esta lista de enlaces siempre está ordenada del mismo modo (gl pt es en fr ca eu al it cs bg el). Además, en caso de no existir el enlace, se coloca "#" para indicar explícitamente la no existencia.

Existen otros dos campos que aportan más información y más relaciones con otras entradas de Wikipedia. El campo *related* aporta el título de otras entradas relacionadas con la actual, que han sido así explicitadas en Wikipedia. Por otro lado, *links* es el conjunto de enlaces salientes a otras entradas.

2.3. Estrategias para crear corpus comparables

Dada la estructura y la información contenida en la CorpusPedia, es factible, no sólo, agrupar artículos sobre una misma temática, sino también relacionarlos con artículos que traten la misma temática en otras lenguas. Por ello, la estructura de la CorpusPedia permite construir con cierta comodidad corpus comparables. Para llevar a cabo esta tarea, creamos varias herramientas orientadas a extraer corpus con diferentes grados de comparabilidad. Estas tres herramientas, que se corresponden con tres estrategias, se describen a continuación.

2.4. Comparables sin alinear

Esta estrategia extrae los artículos en dos lenguas que tratan de un mismo tema, donde el tema es sugerido por medio de una categoría y su traducción (por ejemplo el par *Arqueología-Arqueologia*, en castellano y portugués). En concreto, el algoritmo es el siguiente:

> Dadas dos lenguas, L1 y L2, y dos categorías, C1 y C2, donde C2 es una traducción de C1 en L2, se procede a:
>
> 1. extraer todos los artículos de la corpuspedia de L1, siempre y cuando C1 esté dentro de la lista de categorías de cada artículo procesado;
> 2. Repetir el proceso partiendo de los artículos de L2.

El resultado es un corpus comparable no-alineado, compuesto por textos en dos lenguas (L1 y L2) que abordan la misma temática: C1 o C2. Es un corpus no-alineado porque el título de un artículo en una lengua puede o no tener su traducción en la otra lengua, es decir, puede o no tener un enlace *interlanguage* a un artículo de la otra lengua.

2.5. Alineamiento estricto

El corpus resultado del anterior proceso puede ser visto como demasiado heterogéneo, pues incluye artículos en una lengua que no tienen su correspondiente versión en la otra. Por ejemplo, puede haber un artículo español titulado "Arqueología de España" que no tiene un enlace *interlanguage* en portugués, es decir un artículo sin su correspondiente versión "Arqueologia de Espanha". El método que utilizamos para construir un corpus alineado a nivel de los artículos, permitiendo sólo recoger aquellos que tienen enlaces *interlanguage* en la otra lengua, es el siguiente:

> Dadas dos lenguas, L1 y L2, y dos categorías, C1 y C2, donde C2 es una traducción de C1 en L2, se procede a:
>
> 1. extraer todos los artículos de L1 de la corpuspedia, siempre y cuando: i) C1 esté dentro de la lista de categorías de cada artículo procesado, ii) cada artículo contenga un enlace *interlanguage* a un artículo de L2 conteniendo la categoría C2.
> 2. Repetir el proceso partiendo de los artículos de L2 y eliminar las inconsistencias.

El resultado es un corpus comparable alineado de manera muy estricta, ya que no sólo cada artículo en una lengua tiene su artículo correspondiente en la otra, sino que ambos artículos comparten la misma restricción categorial.

2.6. Alineamiento laxo

El alineamiento estricto puede dejar fuera artículos relevantes, por ejemplo, aquellos que, aun teniendo un enlace *interlanguage* a un artículo de la otra lengua, no cumplen la restricción

categorial. En concreto, puede haber artículos categorizados en la Wikipedia española como siendo de "Arqueología", pero que no fueron categorizados en portugués por su correspondiente bilingüe "Arqueologia". De hecho, la versión portuguesa de la Wikipedia está menos categorizada que la española. Esta escasez categorial resta cobertura a la estrategia descrita en la subsección anterior (3.2). Para subsanar esta circunstancia, proponemos otro método de alineamiento más laxo. El objetivo es extraer todos los artículos que tienen enlaces *interlangue* conteniendo la categoría requerida en, al menos, una de las dos lenguas. El algoritmo es el siguiente:

> Dadas dos lenguas, L1 y L2, y dos categorías, C1 y C2, donde C2 es una traducción de C1 en L2, se procede a:
> 1. extraer todos los artículos de L1 de la corpuspedia, siempre y cuando: i) C1 esté dentro de la lista de categorías de cada artículo procesado, ii) cada artículo contenga un enlace *interlanguage* a un artículo de L2.
> 2. extraer todos los artículos de L2 que tienen enlace con los artículos extraídos en el proceso anterior.
> Repetir los dos procesos desde L2 y eliminar los artículos duplicados y las inconsistencias.

El resultado es un corpus alineado artículo a artículo, aunque de temática no tan específica como la conseguida por la estrategia más restrictiva.

2.7. Experiencias y Resultados

En el momento actual, el texto plano (*plaintext*) de la CorpusPedia es de aproximadamente 20 millones de palabras para la versión en gallego de la Wikipedia, 120 millones en la versión portuguesa y 180 en la versión en español. El hecho de que la versión española tenga más palabras que la portuguesa contrasta con el mayor número de artículos introducidos en esta última (ver tabla 1 en la introducción). De ello se infiere que los artículos de la versión portuguesa tienden a ser más pequeños que los de la española.

Tomando como base la CorpusPedia, realizamos una experiencia para construir corpus comparables español-portugués sobre una temática específica, arqueología, utilizando las tres estrategias descritas en la sección anterior. La temática específica de los textos españoles fue seleccionada por medio de la categoría "Arqueología", y se usó su traducción, "Arqueologia", para seleccionar los textos portugueses. La tabla 2 muestra los resultados numéricos de la experiencia.

Tabla 2: Tres estrategias para construir corpus comparables español-portugués usando la categoría "Arqueología-Arqueologia".

Estrategia	Tamaño (en palabras)	Número de artículos
es/pt no-alineado	344.000 / 64.000	420 / 100
es/pt alineado estricto	27.000 / 11.000	19 / 19
es/pt alineado laxo	132.000 / 64.000	119 / 119

La tabla 2 deja entrever que hay bastante disparidad en el tamaño de los corpus. Por ejemplo, usando la estrategia básica sin alineamiento, el corpus español alcanza las 344 mil palabras frente a las 64 mil del portugués. Esto se debe, sobre todo, a que se han encontrado 420 artículos en español restringidos por la categoría "Arqueología", frente a los 100 en portugués asociados a la categoría "Arqueologia". Además, el tamaño de los artículos es también significativamente mayor en el corpus español. Usando la estrategia de alineamiento laxo, para los mismos artículos (119), el corpus español alcanza las 130 mil palabras, frente a las 64 mil del portugués. De aquí se deduce que los artículos en portugués sobre arqueología tienden a contener la mitad de información que los españoles.

Finalmente, la tabla 3 muestra, a modo de ejemplo, 5 pares bilingües de títulos correspondientes a la lista de artículos extraídos usando la estrategia "alineado estricto". Esto nos permite observar el grado de comparabilidad de los corpus extraídos, si bien una cuantificación del grado de comparabilidad será objeto de estudio de posteriores experimentos.

Tabla 3: 5 primeros artículos extraídos usando la estrategia de alineado estricto.

Artículos en español	Artículos en portugués
Arqueoastronomía	Arqueoastronomia
Arqueología	Arqueologia
Arqueología bíblica	Arqueologia bíblica
Arqueología procesual	Arqueologia processual
Arqueología subacuática	Arqueologia subaquática

2.8. Conclusiones y trabajo futuro

La aparición de recursos multilingües, como la Wikipedia, posibilitan nuevos métodos de creación de corpus a partir de la web, que son más eficientes y potentes que los tradicionales. Asimismo permiten crear corpus comparables a partir de la extracción de la(s) temática(s) de cada porción de texto. En la actualidad, estamos buscando cómo mejorar las estrategias de extracción ampliando la cobertura de los artículos seleccionados sin perder precisión. Para

ello, estamos evaluando dos técnicas: por un lado, el uso del campo *related* para aumentar el número de categorías restrictivas, por otro, la expansión automática por hipónimos de la categoría escogida. Esta última estrategia sólo se podrá llevar a cabo si se dispone de una ontología de categorías previamente construida. Una de nuestras tareas, en la actualidad, es construir una ontología de categorías a partir de información estructurada de la Wikipedia.

Para la ampliación de cobertura hemos formulado algunas estrategias: el uso de los enlaces internos de wikipedia (no sólo los enlaces *interlanguage*) que ya son extraídos y el uso de enlaces externos que generarán corpus con alguna de las técnicas de *web as corpus*. Si bien estas estrategias ampliarían la cobertura, probablemente disminuirían la comparabilidad, por lo que esos valores deberán ser evaluados y posteriormente incorporado o no el nuevo corpus, dependiendo de las necesidades de comparabilidad y cobertura.

Por último, realizaremos evaluaciones sobre el grado de comparabilidad de los corpus generados. Para ello, utilizaremos métodos inspirados en los trabajos de Saralegi & Alegria (2007).

REFERENCIAS BIBLIOGRÁFICAS

Adafre, S.F. & de Rijke, M. (2006). Finding Similar Sentences acorss Multiple Languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 62-69).

Clark M, Ian Ruthven & Patrik O'Brian Holt (2009). The Evolution of Genre in Wikipedia. *JLCL,* vol 24 (1), 1-22.

Filatova, Elena (2009). Directions for Exploiting Asymmetries in Multilingual Wikipedia. *Proceedings of CLEAWS3*. Boulder, Colorado, (pp. 30-37).

González, I. and Gamallo, P. (2010). Estrategias para la elaboración de corpus comparables a partir de la web. At *XXXIX Simposio internacional de la SEL*.

McEnery, A. M. and Xiao, R. Z. (2007). Paralell and comparable corpora: What are they up to? In: James, G. and Anderman, G., (eds.) *Incorporating Corpora: Translation and the Linguist.* Translating Europe . Multilingual Matters: Clevedon, UK.

Potthast, M. Stein, B. and Anderka, M. (2008). A Wikipeda-Based Multilingual Retrieval Model.

Saralegi X. and Alegria I. (2007). Similitud entre documentos multilíngües de carácter científico-técnico en un entorno Web. *Procesamiento del Lenguaje Natural,* 39.

Tomás, J., Bataller, J. and Casacuberta, F. (2008). Mining Wikipedia as a Parallel and Comparable Corpus. *Language Forum*, 34(1).

Tyers, M.F. and Pieanaar, J.A. (2008). Extracting bilingual word pairs from Wikipedia. At *LRE SALTMIL Workshop*. Marrakech, Marroco.

Yu, Kun and Tsujii, Junichi (2009). Bilingual Dictionary Extraction from Wikipedia. *Proceedings of MT Summit XII*, Ottawa, Canada.

STATE: a multimodal tool for assisted creation of corpora

SERGIO GRAGERA, DAVID LLORENS, ANDRÉS MARZAL,
FEDERICO PRAT and JUAN MIGUEL VILAR
Universitat Jaume I

Abstract
We present a complete assisted transcription system for ancient documents: State. The system consists of three different applications that offer a pen-based interactive environment to assist humans in transcribing ancient documents for creating corpora. One application creates projects from a set of images of document pages, other application controls an automatic transcription system, and the third application allows the user to interact with the transcriptions and easily correct them as needed. This division of labor allows great flexibility for organizing the work in a team of transcribers.
Keywords: corpus compilation tools, assisted transcription, ancient documents

Resumen
Presentamos un sistema completo de transcripción asistida para documentos antiguos: STATE. El sistema consta de tres aplicaciones diferentes que ofrecen un entorno interactivo manejable con lápiz electrónico para asistir a los humanos en la transcripción de documentos antiguos y la creación de corpus. Una aplicación crea proyectos a partir de imágenes de páginas, otra controla un sistema de transcripción automático y la tercera permite al usuario interactuar con las transcripciones y corregirlas fácilmente donde sea necesario. Esta división permite gran flexibilidad en la organización del trabajo dentro de un equipo de transcriptores.
Palabras clave: herramientas de compilación de corpus, transcripción asistida, documentos antiguos

1. INTRODUCTION[1]

The creation of large corpora is a labor intensive task in which transcription errors have potentially catastrophic consequences. This makes it an ideal candidate for automatization or at least computer support. Although conventional Optical Character Recognition (OCR) systems offer an appropriate performance when dealing with modern documents, their high error rates when recognizing unusual fonts make them almost useless for handwritten or ancient documents. Besides, ancient documents are usually affected by many sources of noise (moisture, patches, holes, etc) that make them harder to transcribe.

We present here the STATE system, which assumes that OCR performance will be frequently poor on handwritten or ancient documents and human help will be needed in order to accurately transcribe them; therefore, it is designed as a multimodal, adaptive, extensible system that provides a set of user-centered tools. STATE is the acronym for *Sistema de*

[1] This work has been partially funded by the Spanish *Ministerio de Educación y Ciencia* (*Consolider Ingenio 2010* CSD2007-00018) and *Fundació Caixa Castelló-Bancaixa* (P1·1B2009-48).

Transcripción Asistida de Texto Escrito, which is the Spanish for *Written-Text Assisted Transcription System*. The current version supports multimodal interaction via mouse, keyboard, and stylus.

STATE has been designed with extensibility and flexibility in mind, which is achieved by providing different applications that are loosely coupled. One application is responsible for creating projects from sets of document pages, other for automatically obtaining transcriptions, and a third one for correcting them. STATE assists the transcription process in a scenario where each document or series of documents to be transcribed can also train a remotely accessible recognition engine used simultaneously by all the users. Each user can interactively manipulate the images (for instance to eliminate stains or increase their contrast) and detect or edit their text layout before transcribing one or more lines of text. Text transcriptions are shown line by line, together with their original images. When the transcription of a line contains errors, the user can correct them with the keyboard and a pen-based, user-friendly interface. The corrected transcription and associated image line can then be sent to the recognition engine in order to adapt it to the peculiarities of the task (rare fonts or glyphs, unusual words, ancient syntax, etc). A privileged user (in the following, the "administrator") supervises these corrections with a graphical front end and she enhances the engine with those corrected transcriptions. The enhanced recognizer is immediately available to all the users, since it is accessed as a web service.

The paper is organized as follows. In Section 2 the STATE architecture is described. Section 3 presents the Transcription Assistant, an application which integrates an image conditioner, a layout manager, and an assisted line transcriber. Section 4 introduces the Recognition Engine application. Finally, Section 5 is devoted to drawing some conclusions and presenting future lines of development.

2. THE STATE SYSTEM ARCHITECTURE

STATE is not a single application, but a set of tools to assist the transcription of documents. In its current version, it consists of three applications. Human transcribers use the main application, *StateTA*, to produce the text version of each page in a document. An application, *StatePC*, is used to prepare the scanned images of the documents in *projects*, groups of pages that will be used by *StateTA*. The other application, *StateRE*, is a recognition engine concurrently accessible by several instances of *StateTA*. Since this application is accessible using the http protocol, it can be easily replaced by a different one. *StateRE* offers a

graphical, pen-based interface to control its parameters and to edit corpora of character samples.

The user interface of these applications has been designed to be comfortably used with a stylus. An on-line character recognition engine has been built in order to have a flexible, extensible pen-based text introduction system. This engine is described in detail in (Prat, Marzal, Martín, Ramos-Garijo, & Castro, 2009).

3. The Transcription Assistant Application

StateTA has been designed to work with projects. Initially, a project is simply a set of images, one for each of the pages to be transcribed. The projects can be created using *StatePC*, whose main window can be seen in 0. As the user proceeds, more information is added to the project, including the image processing operations needed to clear the pages, their layout, and their transcription. For flexibility and ease of later processing, all this information is stored in an XML file.

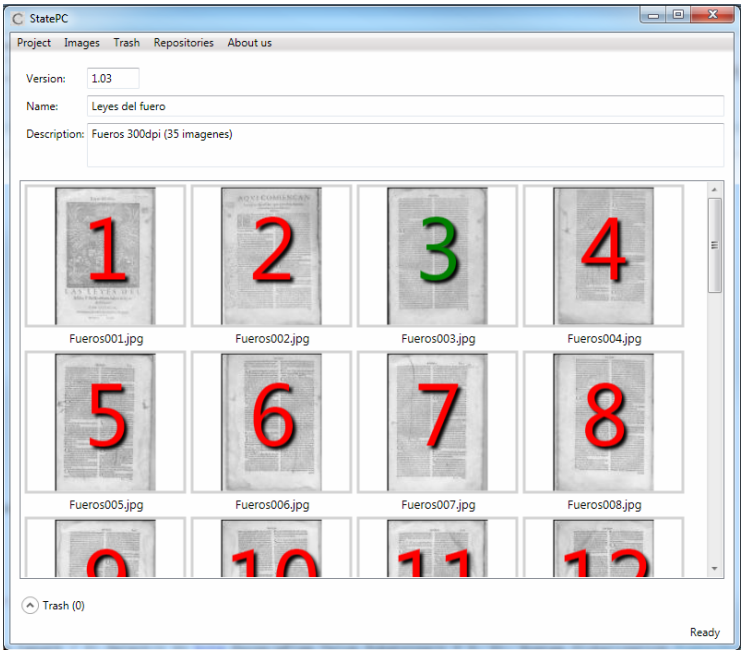


Figure 1: The *StatePC* main window.

The user selects one project and can open simultaneously as many different pages as needed. Each page passes through a sequence of three stages: the Image Conditioner, the Layout Manager, and the Assisted Line Transcriber. Screenshots of these stages can be seen in Figures 2 to 4, they are accessible through the tabs on the left of the window. The tabs on the top correspond to different pages.

The order of the stage tabs is designed after the order of usual workflow on a page: first, the page is conditioned (a set of image processing tools are applied to it); second, its layout is detected; finally, the user transcribes each line in the page.

The Image Conditioner and the Layout Manager offer a similar interaction behavior: both show a view of the page and let the user act on it by executing commands. The Image Conditioner (0) offers commands to remove noise and to enhance the text in the image. The Layout Manager (0) offers commands to automatically detect and to interactively edit the layout (lines, columns, and text flows). Finally, the Assisted Line Transcriber (0) allows the user to obtain transcriptions of the lines using a recognition engine and to manually correct these transcriptions. If the user feels that certain samples (line images and their corresponding transcriptions) may help to improve recognition accuracy, she can send them to the recognition engine.

The user is free to alter this workflow: for instance, she could condition all the pages first, then detect the layout of the pages and, finally, proceed to transcribe them in any order. To keep track of the progress, the pages can be marked with status information (cleaned, layout found, etc), this is graphically shown by a color code in the tab of the page.

3.1. The Image Conditioner

Ancient documents are usually in poor condition due to moisture, torn pages, fading ink, staining, hand-made annotations, etc. It is difficult to automatically enhance them in order to successfully feed an OCR. The goal of this stage is to offer a comfortable interface for image cleaning and a set of tools to appropriately restore the ancient document text. The screen is divided into several regions (0): on the left, the image of the page after the commands have been applied to it, this image has overlaid controls for zooming and selecting; on the right, an area divided into two parts: in the upper part, the available commands are arranged as a tree and, in the lower part, a list shows the commands applied to the page.

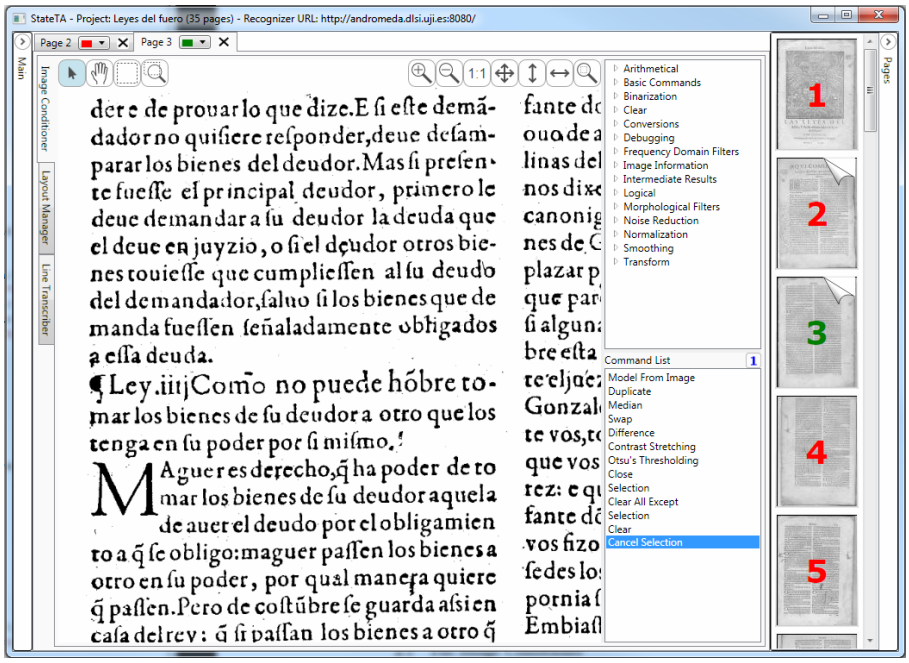


Figure 2: The *StateTA* Image Conditioner.

The command selection tree offers a hierarchical catalog of commands, ie actions performing some kind of processing on all or part of a page. The application provides a large set of commands for image processing: binarizers, morphological operators, noise reducers, smoothers, etc. This set can be extended by defining new commands as needed.

At any moment, the user can add a command to the list from the command selection tree. Initially, the command is executed with default values for the parameters that can be changed at will. When a command of the list is selected, the page view shows the effect of executing all the commands up to (and including) it, providing the opportunity to change its parameters. The first command always yields the original image, thus providing unlimited "undo". The command list can be edited, ie commands can be deleted, replaced, or added at any place. An interesting feature is the possibility of copying the current command list to every page in a given status. That way it is possible to automate repetitive tasks when processing batches of pages.

3.2. The Layout Manager

The Layout Manager offers a set of layout detection commands and an interactive environment to edit the layout. A page layout is a hierarchical structure whose higher level components are text flows, which are ordered sequences of text columns possibly continuing in other pages. For instance, the main text of a book is a text flow and the set of footnotes

may be defined as a different text flow. Each text flow is identified with a user-defined label and it contains zero or more columns, each one comprising zero or more lines.

Automatic layout detection tools usually need human assistance. The interfaces of the Layout Manager and of the Image Conditioner are similar, as can be seen in 0 (actually, both share most of their implementation). The command selection tree offers commands related to line and column detection and editing operations (such as deletion and merging). Some commands are introduced implicitly by user interactions with the mouse or stylus, for instance to redefine the limits of a line by moving its handles. The set of commands is also extensible with user-defined commands. The command list is linked to the command list in the Image Conditioner so changes in the conditioning of the page force the execution of the Layout Manager command list. The user can copy the commands from the current page to others and save time, just as with the Image Conditioner. Once the user is satisfied with the page layout, she can proceed to transcribe the text using the Assisted Line Transcriber.

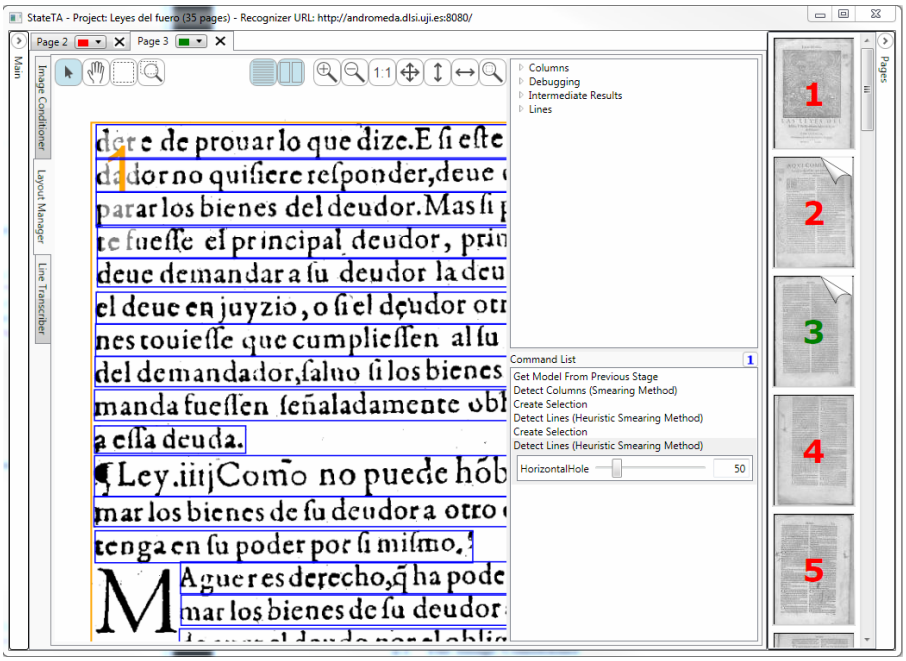


Figure 3: The *StateTA* Layout Manager.

3.3. The Assisted Line Transcriber

This stage helps the user to transcribe the lines in the text flows of a page. It uses a recognition engine available as a web service to provide automatic transcriptions that can be edited. The user can either start a recognition engine or connect to a running one by providing its URL. The OCR engine is expected to perform poorly when transcribing the first pages of

a new kind of document. To allow for this, the user of the Line Transcriber can send corrected transcriptions to *StateRE*, which can use them to better adapt itself to the task.

The Line Transcriber can be used by pen, keyboard, and mouse. It can be comfortably used with a Tablet PC, digitizing tablet, or pen-sensitive screen. 0 shows the GUI of the Line Transcriber. When the user selects a text flow (top left box), all its lines are shown on the right panel. Each line (or all lines consecutively) can be sent to the recognition engine to obtain a transcription. The user can also send a corrected transcription and the corresponding line to inform *StateRE* that they can be used for improving the decoder. *StateRE* can be adapted in this way to a specific type of documents.

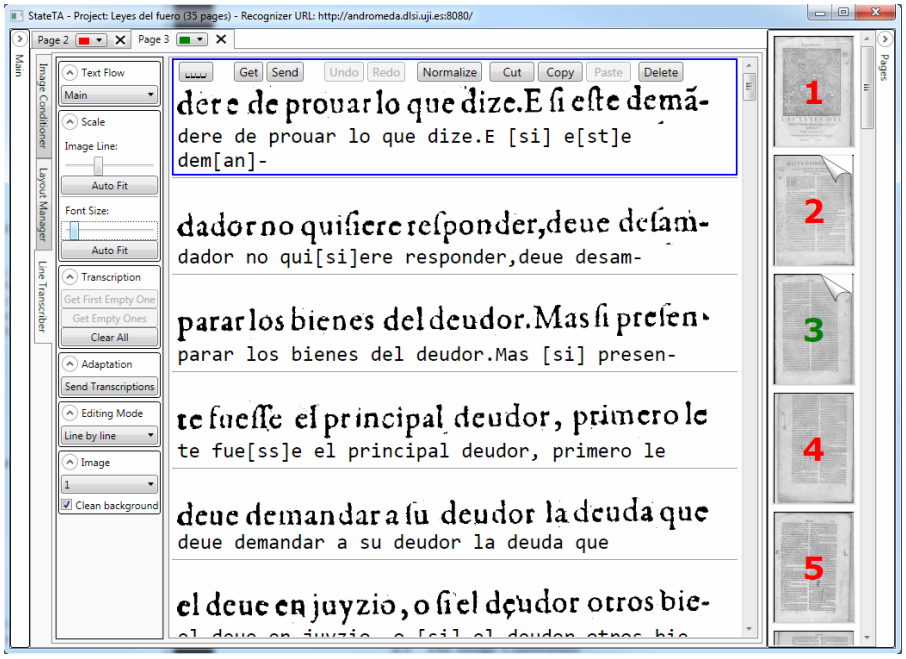


Figure 4: The *StateTA* Line Transcriber.

The transcription result is stored as XML text in the project file. It reflects the layout hierarchy: text flows, columns, and lines. Each line is coded as a sequence of (possibly multi-character) symbols and stores its coordinates on the original image as attributes.

In this stage, the text transcription is shown under the graphical view of each line, so that its validation is fast and easy. When the line or the transcription is clicked, a pen-input panel appears in-place with the current transcription. The user can edit the text in the input panel with gestures and ink. Since we use the .NET 3.5 platform, the Microsoft Ink Input panel could have been used for this pen-based interaction, but we decided to build our own on-line isolated character recognition engine, thoroughly described in (Prat et al., 2009), and input widget for the sake of flexibility and extensibility. It should be taken into account that

old fonts include some glyphs which are not recognized by standard on-line handwritten text recognizers. For instance, in Spanish texts from the 16th or 17th centuries there are ligatures or special characters, such as ﬁ (“si”), ꝗ (“que”), ﬅ (“st”) or ō (“on”), which are not part of modern standard writing. Our pen-input panel can learn new symbols from user input.

It should be noted that on-line handwritten text can also contain errors. These can be corrected with pen gestures, with the keyboard, or by selecting the right character on a menu of alternatives that automatically appears under each character.

Pen-based input is appropriate when the user is checking a transcription and must correct a few errors, but it can be tiresome when the user must provide a full transcription or there are many errors in the automatic transcription. *StateTA* has been designed so that it can be effectively used with the keyboard: the pen-based input panel is also sensitive to keyboard. The user can move a cursor, navigate from line to line with key gestures (arrow keys, enter key, etc), copy and paste text, etc.

If the user so whishes, she can edit the lines while seeing the image of the document without splitting it into lines, as seen in 0. This mode is designed to help those corrections that may need more context to find the appropriate transcription of a line.

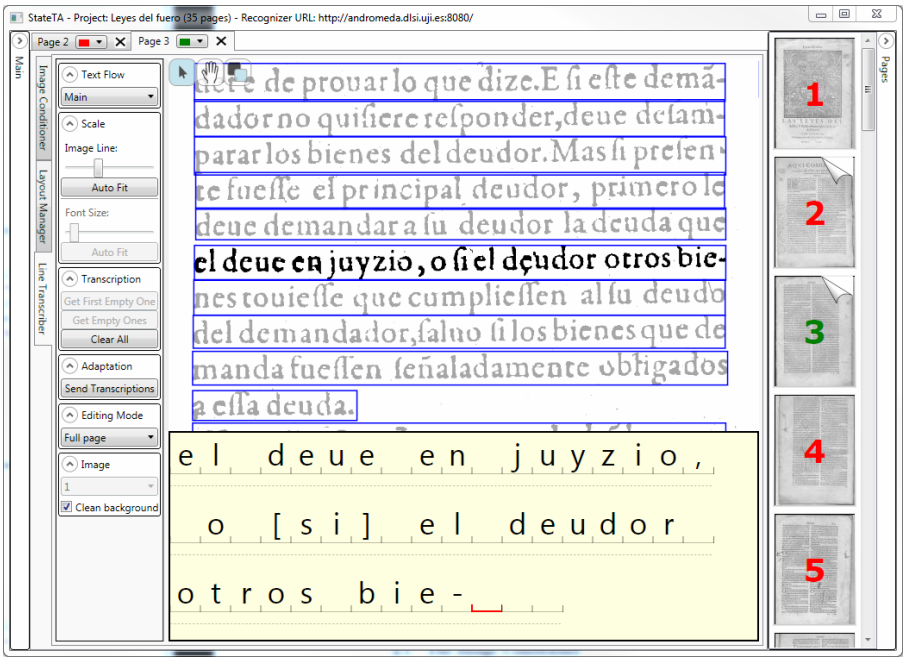


Figure 5: The full page mode for editing transcriptions.

4. The Recognition Engine Application

StateRE, the recognition engine application, is a different application from *StateTA*. It offers transcription services for images representing lines of text and can learn from these images

when accompanied by their respective transcriptions. The engine can be accessed from *StateTA* via http. *StateRE* offers a web service interface and a GUI front end to help the administrator in validating new learning samples, editing the corpus, setting training parameters, etc.

The recognition engine is integrated in an interactive application that lets the administrator manage the lines sent by transcribers, edit the corpus, and re-train the engine. Let us introduce the main interactive forms of this tool.

4.1. Main tab

With this tab, the user can control different parameters of the recognition engine and train it. This tab can be seen in 0.

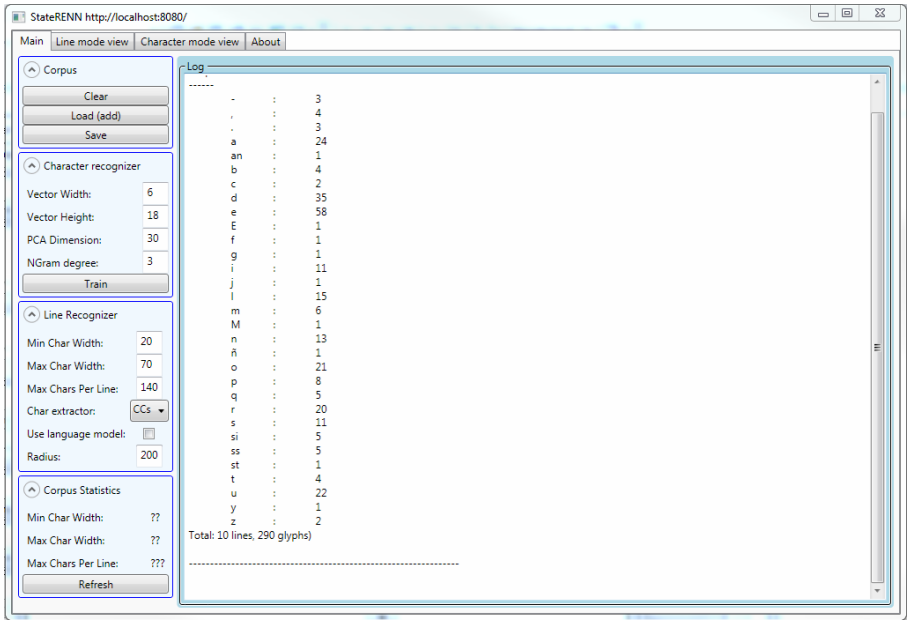


Figure 6: The main tab of *StateRE*.

4.2. Line mode view

All the transcriptions sent from the transcribers to the recognition engine are shown here with their respective bitmaps (see 0). The administrator can select them to validate the transcription and to edit an automatic segmentation of the line, so that each character in the transcription is correctly placed on the line. This task can be performed with a pen-based interface: vertical strokes add segmentation marks, horizontal strokes delete them, on-line handwritten text modifies the transcription, etc. Once a line has been validated, their characters are added to the engine.

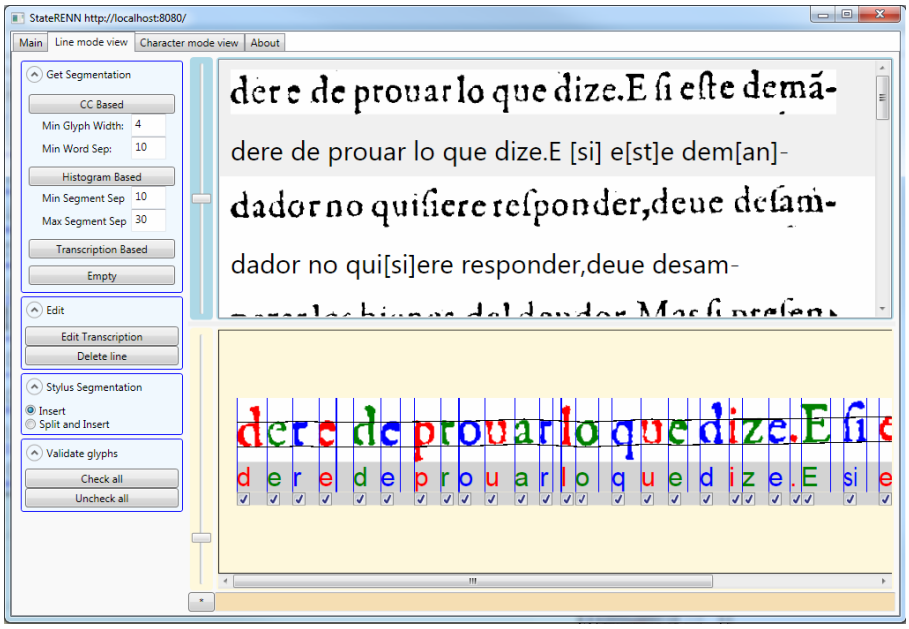


Figure 7: Transcription validation in *StateRE*.

4.3. Corpus editor

The recognition engine uses a labeled set of character samples to learn. These characters are obtained from those lines sent by transcribers to *StateRE* and validated by the administrator. 0 shows the GUI of the corpus editor. The administrator can see all the samples (by category or arrival time) and perform some basic operations: deleting inappropriate samples (for instance, those having a noisy bitmap or defective glyphs), moving samples between characters sets, adding new labels, etc. The edited samples can be saved for future reuse.

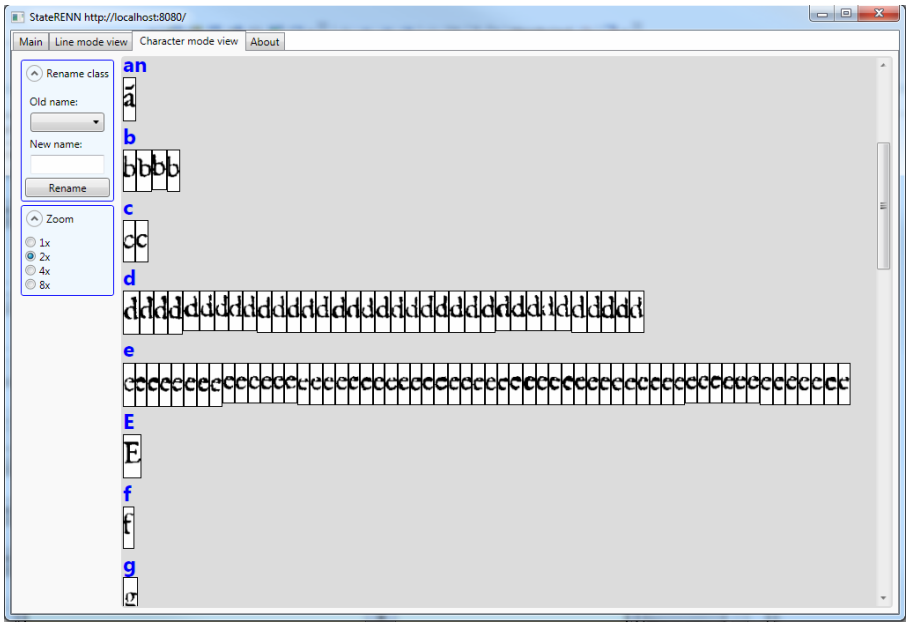


Figure 8: Corpus editing in *StateRE*.

5. Conclusions and Future Work

We have presented the STATE system for assisted transcription of ancient documents. It consists of thres applications: *StateTA*, an interactive application to condition images, detect and edit the layout, and interactively edit transcriptions; *StateRE*, a recognition engine; and *StatePC*, an application to manage transcription projects. The engine can be easily adapted to new documents, since it learns from samples obtained from the documents to be transcribed. A lot of effort has been put in usability aspects and the application can be controlled with mouse, keyboard, and stylus.

We are now enhancing the STATE system in different directions. We plan to include recognition confidence information for each character and to show it graphically in the Line Transcriber, so that the user can be quickly led to places where corrections may be needed. We also plan to improve the recognition engine using language models and HMM-based decoders.

References

Prat, F., Marzal, A., Martín, S., Ramos-Garijo, R., & Castro, M. J. (2009). A Template-based Recognition System for On-line Handwritten Characters. *Journal of Information Science and Engineering , 25* (3). (pp. 779—791).

Análisis preliminar de rasgos de definiciones de categorías semánticas del Corpus lingüístico de sujetos sanos y con Enfermedad de Alzheimer: Una investigación transcultural hispano-argentina

LINA GRASSO

Centro Interdisciplinario de Investigaciones en Psicología Matemática y Experimental

CONICET (Buenos Aires)

MARÍA DEL CARMEN DÍAZ MARDOMINGO

HERMINIA PERAITA ADRADOS

UNED

Resumen

La investigación sobre la memoria semántica y la representación del conocimiento ha constatado diferencias cuantitativas en la producción de rasgos entre personas sanas y enfermos de Alzheimer. La literatura señala un deterioro semántico diferencial entre las categorías de seres vivos y seres no vivos y entre determinados tipos de rasgos. Este trabajo, derivado del Corpus Lingüístico de Peraita y Grasso (2009) con 1266 definiciones y 79.665 palabras de una muestra española y otra argentina, analiza la producción de rasgos del Corpus de definiciones de seis categorías semánticas de seres vivos y no vivos. Los rasgos se clasificaron en función del modelo de rasgos de Peraita, Elosúa y Linares (1992).En este trabajo se analizan las frecuencias de producción de sujetos mayores sanos y enfermos de Alzheimer de la muestra argentina, y las diferencias en las categorías en determinados rasgos. Se halló que los sanos producen, proporcionalmente y con mayor frecuencia rasgos evaluativos y los enfermos un número significativamente menor de rasgos, pero proporcionalmente, mayor número de rasgos funcionales.

Palabras clave: déficit semántico; deterioro diferencial de categorías; definiciones verbales; análisis de atributos semánticos; corpus lingüistico; neuropsicología cognitiva

Abstract

The investigation on semantic memory and the representation of knowledge Alzheimer's disease, evidence a lower production of semantic features compared to healthy controls. It has also been observed a semantic deterioration between certain categories belonging to animate beings and inanimate objects, and between certain kinds of features. The purpose of this study, derived from the Linguistic Corpora made by Peraita and Grasso (2009) with 1266 definitions and 79.665 words ot two samples comes from Spain and Argentina, is the analysis of the feature production belonging to the Corpora definitions of six semantic categories of animate beings and inanimate objects. The study examines the logic underlying the proposed analysis of features (Peraita, Elosúa & Linares, 1992). The argentine sample is composed of healthy old aged participant, and patients with Alzheimer's disease. We analyzed the frequency of production and were found to produce healthy, proportionally, more often evaluative features and patients significantly fewer features, but a proportionately higher number of functional features.

Keywords: deficit semantic; differential impairment categories, definitions verbal; semantic attribute analysis, corpus linguistics, cognitive neuropsychology

1. INTRODUCCIÓN[1]

Los Corpus de definiciones verbales constituyen un instrumento teórico-metodológico de primer orden para el estudio de patologías como el deterioro semántico que cursa con un déficit léxico-semántico y conceptual en la Enfermedad de Alzheimer. Mediante éstos se posibilita el acceso y consulta para su uso en investigación y en el ámbito clínico, en la medida que la base de datos sea accesible en función de una serie de variables: enfermo/sano, hombre/mujer, seres vivos/no vivos y tipos de rasgos o a partir de combinaciones de éstas.

El presente trabajo, inscripto en el área de la Lingüística clínica y la Lingüística de corpus, presenta un análisis preliminar de la producción de rasgos o atributos semánticos del Corpus de definiciones de seis categorías semánticas elaborado por Grasso y Peraita (2009). Se analiza la frecuencia de producción de rasgos entre participantes sanos y enfermos de Alzheimer y se indaga sobre la existencia de diferencias en las categorías de seres vivos (SVs) y seres no vivos (SNVs). Este proyecto se inscribe además en el actual marco de la Neuropsicología Cognitiva y en su interés teórico por analizar el deterioro semántico en determinadas patologías.

2. ANTECEDENTES

Uno de los temas que se debaten en el marco actual de la Neuropsicología Cognitiva, es el estudio de cómo se representan mentalmente determinadas categorías semánticas mediante modelos teóricos de rasgos o atributos semánticos.

Las categorías semánticas son clasificaciones que se llevan a cabo en el mundo que nos rodea y que permiten tratar como equivalentes objetos que son diferentes entre sí. La memoria semántica se organiza mediante categorías, y por ello pueden realizarse una serie de funciones cognitivas como hacer inferencias, establecer relaciones entre ejemplares, atribuir propiedades a objetos que no conocemos, etc.

A partir de ponerse en evidencia la existencia de un deterioro semántico diferencial entre categorías SVs y SNVs en pacientes con lesiones cerebrales, en los últimos años se ha renovado el interés en el estudio de este tipo de problemática. Se han identificados múltiples casos clínicos en los que como consecuencia de un daño degenerativo o no degenerativo del

[1] En el marco de un proyecto de investigación originariamente financiado por el MEC, Referencia PB94/1573, pero cuya financiación actual procede de la Fundación BBVA.
En el marco de una beca postdoctoral, en el CIIPME- CONICET (Centro Interdisciplinario de Investigaciones en Psicología Matemática y Experimental- Consejo Nacional de Investigaciones Científicas y Técnicas de la República Argentina)

sistema nervioso central (SNC) el conocimiento de algunas categorías semánticas se deteriora o se pierde diferencialmente (Rogers y Plaut, 2002; Capitani, Laiacona, Mahon & Caramazza, 2003).

Los *déficits específicos de categoría* se evidencian en la ejecución de tareas donde existe una peor ejecución con categorías del dominio de SVs, con una relativa o total preservación con las categorías del dominio de SNVs. Si bien este patrón –mayor preservación del dominio de SVs- es el que se ha observado con mayor frecuencia (61 de 79 pacientes revisados por Capitani, Laiacona, Mahon & Caramazza, 2003), existe un pequeño número de casos (18 de 79 según Capitani, Laiacona, Mahon & Caramazza, 2003) en los que se da el patrón contrario: hay un mayor deterioro del dominio de los SNVs o artefactos, mientras que el dominio de los seres vivos está, en su mayor parte, preservado. A esto es lo que se denomina una doble disociación.

No todos los casos de déficits específicos de categoría reflejan esta doble disociación. En otras ocasiones el déficit afecta a la mayor parte de las categorías de un dominio y a alguna categoría perteneciente al otro dominio (Borgo & Shallice, 2001); Warrington & Shallice, 1984)-.

En los últimos 20 años, se ha renovado el estudio de estos déficits específicos de categoría, ya que las regularidades en los patrones de deterioro pueden utilizarse para contrastar diferentes teorías sobre la estructura y organización de la memoria semántica.
Sin embargo, aún existe un debate teórico muy amplio sobre los modelos de representación que subyacen a estos campos semánticos categoriales -modelos de memoria semántica- y sobre la metodología más adecuada para abordarlos.

En el modelo de rasgos descrito por Peraita, Elosúa y Linares (1992) Peraita & Moreno (2006) (figura 1) se asume que los conceptos y las categorías en general, y, más específicamente, las categorías semánticas de SVs y SNVs, están constituidas por un conjunto de rasgos, atributos o propiedades semánticas que, a manera de componentes básicos, determinan su núcleo o estructura conceptual. La pérdida gradual de dichos rasgos y, por tanto, el deterioro del núcleo conceptual, acarrea problemas semánticos de identificación y reconocimiento, denominación, clasificación y uso en una palabra, entre otras habilidades cognitivo-lingüísticas.

En este modelo se proponen 11 bloques conceptuales básicos, considerados como *componentes conceptuales* que subyacerían a toda organización conceptual. Cada uno de ellos posee una etiqueta identificatoria (“*funcional”, “clasificatorio”, “evaluativo”, “destinatario*”...) y una gramática o enunciado con la cual, por lo general, se los introduce

lingüísticamente ("sirve para…"; "es un….", "es...."; "es para…", etc). Los componentes conceptuales, se refieren tanto a la categoría genérica de inclusión (ej: la silla es un mueble) —*componente taxonómico*—, como a las partes que la forman o configuran (ej.: la silla tiene respaldo, asiento y patas) —*componente parte-todo*—, a la función o uso (ej: sirve para sentarse) — *componente funcional*—, al *lugar/hábitat* donde suele encontrarse (ej: se encuentra en las distintas habitaciones de la casa), a las dimensiones de evaluación tanto físicas (perceptuales: forma, color, tamaño, textura*)*, como sociales y afectivas (bondad, simpatía) —*componente evaluativo*—, como a los *tipos o ejemplares* que pertenecen a la misma (ej: hay sillas de cocina, de despacho, de bar, etc.), al agente que las *produce o genera* ej: las hace el carpintero)— *componente causal*—, y procedimiento de uso —*componente procedimental*.

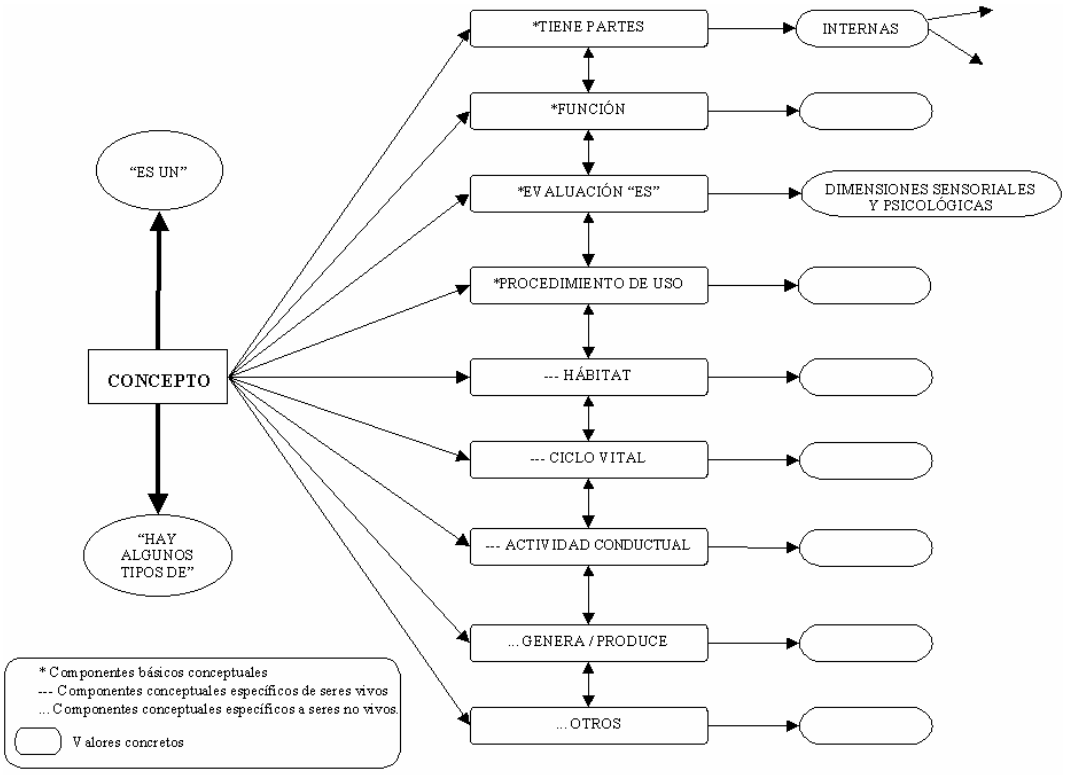


Figura 1. Modelo de rasgos semánticos para la representación conceptual de elementos de categorías semánticas de seres vivos y no vivos.

El análisis teórico de la estructura conceptual de las categorías semánticas que propone este modelo, toma en consideración la enorme riqueza subyacente a dicho conocimiento, y considera que para su estudio debe incluirse información de tipo contextual. Esta propuesta amplía las primeras explicaciones teóricas sobre la estructura que subyace a la representación conceptual, como la Teoría Sensorial Funcional (TSF) (Warrington & Shallice, 1984;

Warrington & McCarthy 1987) ya que propone una estructura de rasgos rica, es decir, no reducida a dos núcleos conceptuales: perceptual y funcional y propone una justificación teórica que considera los distintos tipos de información con los que nos representamos el mundo que nos rodea. El apoyo teórico a esta estructura, parte de una metodología basada en datos obtenidos a partir de muestras de sujetos, y no de definiciones de diccionarios.

Pese a las dificultades metodológicas que se presentan para este tipo de análisis, algunos autores, a partir de definiciones conceptuales de categorías, postulan un concepto que consideran básico para el abordaje de la estructura conceptual: la *relevancia* de los rasgos. Este concepto se define como el conocimiento que los sujetos dicen tener de un objeto y se obtiene en base a la frecuencia empírica en tareas de producción (Garrard, Lambon Ralph, Patterson, Pratt & Hodges, 2005; Sartori & Lombardi, 2004). Este conocimiento se manifiesta de manera diferente en los rasgos de las categorías de SVs y de SNVs, ya que algunos rasgos tienen representación en un solo dominio y no en el otro. Asimismo, el perfil representacional de los ancianos sanos presenta diferencias con los enfermos de Alzheimer, encontrándose variaciones en las distintas fases de la enfermedad (Peraita & Moreno, 2006).

Partiendo del supuesto que la representación conceptual que los sujetos poseen sobre las entidades del mundo que los rodea, presenta rasgos que tienen más relevancia o peso que otros, la elaboración de dos corpus paralelos, aún estando ambos en lengua española, proporciona datos sobre la forma en que las diferencias de ambos contextos culturales como son el de España y Argentina, podría expresarse en las definiciones categoriales. Esto es así, dado que en ellas subyace una estructura contextual y un conocimiento del mundo específico que podría proporcionar elementos que enriquecen el marco de análisis teórico de la estructura conceptual.

El objetivo del presente trabajo es analizar la frecuencia de producción de rasgos entre: sanos y enfermos de Alzheimer, así como la frecuencia de algunos rasgos en las diferentes categorías de SVs y SNVs.

Estos planteamientos, que hasta hace unos años tenían una importancia o relevancia meramente teórica, ahora la tienen aplicada, como posible herramienta diagnóstica, pronóstica y terapéutica en los campos científicos citados.

En ese contexto, se inscribe este estudio, tanto con el objetivo de refinar conceptualmente el modelo que presentamos en este trabajo, como por su posible aplicación a programas de diagnóstico, evaluación y tratamiento.

3. MÉTODO

El trabajo realizado ha consistido en la revisión de cada uno de los protocolos verbales obtenidos procedentes de la aplicación de la Batería de Evaluación de la Memoria Semántica en enfermos con demencia tipo Alzheimer (EMSDA) (Peraita, González Labra, Sánchez Bernardos, & Galeote, 2000), a una muestra de sujetos controles mayores sanos y a otra de enfermos de Alzheimer en grado leve y moderado (1,2).

La muestra está compuesta por 142 participantes: mayores cognitivamente sanos (N = 77) y enfermos de Alzheimer (N = 65) comprendidos en un rango de edad de 60 a 90 años. Este trabajo se centra sólo en la muestra argentina de las dos de las que se compone el Corpus.

Se ha llevado a cabo en primer lugar la transcripción de las cintas que contienen las grabaciones obtenidas a partir de una de las tareas de las ocho de que consta dicha Batería: la prueba de Definición de categorías semánticas o semántico-conceptuales, realizada como tarea oral de producción verbal libre, con restricción temporal (2 minutos). En segundo lugar, el análisis semántico de cada una de las definiciones, estructurado u organizado en rasgos o atributos, en función de un modelo de rasgos citado anteriormente. Se abordó la lógica implícita que subyace al análisis de rasgos que se propone basado en el modelo citado de Peraita, Elosúa & Linares (1992) y Peraita & Moreno (2006).

3. 1. Participantes y tamaño del Corpus

Las características sociodemográficas de la muestra pueden verse en la tabla 1.

Tabla 1. Resumen de los datos sociodemográficos de la muestra

Muestra N = 142	Control N = 77 Media (Dt)	Alzheimer N = 66 Media (Dt)
Edad	70 (6,5)	74 (7,5)
MMSE	28,6 (1,59)	20 (5,1)
Educación	10,93 (4,1)	7,82 (4,5)

El material lingüístico analizado consta de 852 definiciones de categorías semánticas, procedentes de 142 sujetos.

Definiciones de categorías de seres vivos (*perro, manzana y árbol*): 426 y de seres no vivos (*coche, pantalón y silla*): 426.

3. 2. Procedimiento

Se calcularon las frecuencias producidas para tres tipos de rasgos: *evaluativos*, *funcionales* y *taxonómicos*. Los rasgos *evaluativos* se refieren a las dimensiones de evaluación tanto físicas (perceptuales: forma, color, tamaño, textura*)*, como sociales y afectivas (bondad, simpatía); los rasgos o atributos *funcionales* a la función o uso y los *taxonómicos* se refieren a la inclusión en una categoría genérica.

Es importante destacar que dado que para este estudio se analizaron sólo los rasgos *evaluativos, funcionales y taxonómicos*, las frecuencias de otros componentes conceptuales fueron tabuladas en "otros rasgos", pero fueron tomadas en cuenta para el análisis total.

En base a la sumatoria de las frecuencias de atributos para cada categoría, se obtuvo el puntaje total por sujeto y posteriormente por grupo diagnóstico.

4. ANÁLISIS DE RASGOS SEMÁNTICO-CONCEPTUALES

Se estimaron los porcentajes de rasgos producidos por los participantes sanos y los enfermos de Alzheimer. Mediante la comparación de proporciones para muestras independientes (utilizando el paquete estadístico EPI DAT 3.1), se analizó si existían diferencias en la proporción de rasgos evaluativos, funcionales y taxonómicos producidos por el grupo de participantes sanos y enfermos de Alzheimer.

Tabla 2: Porcentajes de rasgos obtenidos por tipo de atributo para el grupo de sanos y enfermos de Alzheimer

Rasgos	Sanos	%	Enfermos	%	IC	z	p	
Evaluativos*	2251	43,3	878	37,4	0,035 0,083	4,78	0,0001	sig
Funcionales*	261	5,01	152	6,48	-0,026 -0,003	2,51	0,01	sig
Taxonómicos	142	2,72	85	3,61	-0,018 0,000	2,02	0,04	ns
Otrosrasgos*	2550	43,2	1235	52,6	-0,118 -0,069	7,4	0,0001	Sig

Se observó que los sujetos sanos obtuvieron una mayor producción de rasgos que los enfermos. Sin embargo, la proporción de dichos rasgos, valorada en función del total de rasgos producidos por cada grupo, presenta variaciones.

Los participantes sanos produjeron mayor proporción de atributos evaluativos que los enfermos, y éstos produjeron proporcionalmente más atributos funcionales y otro tipo de rasgos (no evaluados en este trabajo preliminar).

Para los dominios de SVs y SNVs los participantes sanos, obtuvieron proporcionalmente más rasgos en el dominio de SNVs. En el grupo de enfermos de Alzheimer, tomando como referencia el total de atributos producidos, fue mayor la proporción de rasgos que pertenecían al dominio de SVs.

Tabla 3: Porcentajes de atributos obtenidos por categoría y dominio para el grupo de sanos y enfermos de Alzheimer

Categorías:	Sanos	%	Enfermos	%	IC	z	p	
Manzana	984	18,9	440	18,7	-0,017- 0,021	0,15	0,87	Ns
Perro	996	19,1	411	17,4	1,67 0,0943	1,67	0,09	Ns
Pino*	617	11,9	396	16,9	-0,06 -0,032	5,86	0,0001	Sig
Coche*	837	16,1	297	12,6	0,017 0,052	3,84	0,0001	Sig
Pantalón	908	17,4	383	16,3	- 0,007 0,030	1,196	0,23	Ns
Silla	862	16,5	423	18	- 0,033 0,004	1,504	0,13	Ns
Dominios:								
SV*	2597	49,9	1247	53,06	-0,05 – 0,07	2,518	0,01	Sig
SNV*	2607	50,9	1103	46,93	0,007 0,056	2,518	0,01	Sig
Total atributos N= 7554	5204	68,9	2350	31,1	0,363 - 0,393	46,42	0,01	Sig

5. CONCLUSIÓN

La motivación de este trabajo es realizar una aportación específica a partir de los modelos de rasgos o atributos –en el marco de los modelos de redes de la memoria semántica-, con la finalidad de poder validar algunas de las hipótesis actuales sobre el procesamiento diferencial de categorías semánticas de seres vivos y entes no vivos, tal y como se presenta en algunas patologías degenerativas y no degenerativas del SNC.

En nuestra investigación, el grupo de enfermos de Alzheimer, presentó, como era de esperar, menor producción de rasgos que los ancianos sanos. Estos hallazgos dan cuenta de las diferencias en el perfil representacional de los ancianos sanos y los enfermos de Alzheimer. Los primeros, obtuvieron mayor frecuencia en la producción de rasgos y mayor variedad, poniendo en evidencia una mayor riqueza de propiedades semánticas. En el grupo

de enfermos de Alzheimer, esta disminución supone el déficit semántico-conceptual presente en esta patología neurodegenerativa, pero también su desestructuración.

En cuanto al análisis de la estructura conceptual subyacente a las representaciones mentales de las categorías, se observó que los sujetos sanos producen proporcionalmente y con más frecuencia rasgos *evaluativos* - referidos a las características percibidas por cualquiera de las modalidades sensoriales-. El grupo de enfermos, obtuvo un número significativamente menor, pero proporcionalmente mayor de atributos *funcionales* –referidos a la función de un objeto, pero también de los seres vivos-. La proporción diferencial de rasgos producidos por ambos grupos en estos componentes conceptuales, indica la especificidad y diferenciación de estos tipos de rasgos. Si bien en este análisis preliminar no se consideró la relevancia de otros componentes conceptuales, estos hallazgos corroboran la existencia de –al menos- dos núcleos conceptuales, que en la línea de la *TSF* (Warrington & Shallice, 1984; Warrington & McCarthy, 1987; Borgo & Shallice, 2001) corresponderían a los rasgos perceptuales y los rasgos funcionales. Estos resultados también se observaron en otras investigaciones, que postulan cierta idiosincrasia de este tipo de rasgos que subyacen a la estructura conceptual (Peraita y Moreno, 2006).

El enfoque de la *TSF* antes mencionado, propone que en los rasgos funcionales se clasifica prácticamente toda la información no-perceptual. En nuestro modelo el componente *funcional* refleja exclusivamente la función –del objeto y de los SVs-. Ha de observarse que existe un significativo número de rasgos que no han sido objeto de análisis, y que en este trabajo han sido codificados globalmente como *otros rasgos,* que da cuenta de la existencia de otros componentes y la enorme riqueza que subyace a la representación conceptual.

En cuanto a los hallazgos de las categorías de SVs y SNVs, no se ha ratificado la hipótesis de un mayor deterioro de los SVs, tal y como mantiene la literatura, es decir, que las categorías de SVs se deterioren antes en los enfermos de Alzheimer, y la tendencia indicaría que los rasgos producidos en este dominio tienden a mantenerse. Sin embargo, para obtener datos más precisos, esta valoración debería hacerse también tomando en cuenta la fase de la enfermedad.

Los resultados encontrados se generalizarán cuando el análisis llevado a cabo sobre la muestra argentina pueda llevarse a cabo en la muestra española.

Borgo, F. & Shallice, T. (2001). "When living things and other sensory quality categories go together: a novel category-specific effect", en: *Neurocase*, 7. 201-220.

Butman, J., Arizaga, R.L., Harris, P., Drake, M., Baumann, D., de Pascale, A., Allegri, R. F, Mangone, C.A. & Ollari, J.A. (2001). "El Mini Mental State Examination en Español Normas para Buenos Aires", en: *Revista Neurológica Argentina* 26 (1). 11-15.

Capitani, E., Laiacona, M., Mahon, B. & Caramazza, A. (2003). "What are the facts of semantic category- specific deficits? A critical review of the clinical evidence". *Cognitive Neuropsychology,* 49. 213- 261.

Cree, G. S. & McRae, K. (2003). "Analyzing the Factors Underlying the Structure and Computation of the Meaning of *Chipmunk, Cherry, Chisel, Cheese,* and *Cello* (and Many Other Such Concrete Nouns)". *Journal of Experimental Psychology: General,* 132. 163–201.

Crosson, B., Moberg, P. J., Boone, J. R., Rothi, L. J. G., & Raymer, A. (1997). Category-specific naming deficit for medical terms after dominant thalamic/capsular hemorrhage. *Brain and Language*, 60. 407-442.

Folstein, M.F., Folstein, S.E. & McHugh, P.R. (1975). „Mini-mental state. A practical method for grading the cognitive state of patients for the clinician". *Journal of Psychiatric Research,* 12. 189-198.

Garrard, P., Lambon Ralph, M. A., Patterson, K., Pratt, K. H., Hodges, J. R. (2005) "Semantic feature knowledge and picture naming in dementia of Alzheimer's type: A new approach". Brain & Language, 93. 79-94

Hart, J., Berndt, R. S. & Caramazza, A. (1985). "Category specific naming deficit following cerebral infarction". *Nature*, 316. 439-440.

Humphreys, G.W., & Forde, E.M.E. (Eds.). (2001). *Category specificity in brain and mind.* Hove, U.K.: Psychology Press.

Lobo, A., Saz, P. & Marcos, G. (2002). *MMSE- Examen Cognoscitivo Mini Mental*. Madrid: Ed. TEA.

McRae, K., Cree, G. S., Seidenberg, M. S. & McNorman, C. (2005). "Semantic feature production norms for a large set of living and non living things". *Behaviour Research Methods,* 37. 547-559. Base de datos completa en *www. psychonomic. org.*

Peraita, H. & Moreno, F. J. (2006). ”Análisis de la estructura conceptual de categorías semánticas naturales y artificiales en una muestra de pacientes de Alzheimer”. *Psicothema,* 18, (3). 492 – 500.

Peraita, H. & Grasso, L. (2009). “Corpus lingüístico de definiciones de categorías semánticas de sujetos ancianos sanos y con la enfermedad de Alzheimer. Una investigación transcultural hispano-argentina”. Artículo presentado en el I Congreso Internacional de Lingüística de Corpus de la Asociación Española de Lingüística de Corpus (AELINCO), Murcia, España.

Peraita, H. & González Labra, M.J., Sánchez Bernardos, M.L, Galeote, M. (2000). “Batería de evaluación del deterioro de la memoria semántica en Alzheimer”. *Psicothema*, 12. 192-200.

Peraita, H. & Moreno, F. J., Díaz, C. (2001). “Is the dichotomy between functional/ associative and visual/ perceptual features conceptually adequate to address the topic of categorical Dissociations?”, en *T.Shallice (chair): The connection between working memory and long-term memory.* Presentado en el Symposium at *III Congreso de Memoria ICON. Valencia,* Julio de 2001.

Peraita, H. & Elosúa, R., & Linares, P. (1992). Representación de categorías naturales en niños ciegos de nacimiento. Madrid: Editorial Trotta.

Rogers, T. T. & Plaut, D. C. (2002). “Connectionist perspectives on category-specific deficits”, en E. M. E. Forde & G. W. Humphreys (Eds.), *Category-specificity in brain and mind*. East Sussex, England: Psychology Press.

Sartori, G. & Lombardi, L. (2004). “Semantic relevance and semantic disorders”. *Journal of Cognitive Neuroscience, 16* (3). 439-452.

Warrington, E. K. & Shallice, T. (1984). “Category specific semantic impairments”. *Brain*, 107. 829-854.

Warrington, E. K. & McCarthy, R. (1987). “Categories of Knowledge. Further fractionations and an attemped integration, *Brain, 110, 1273-1296.*

NeoDet – in search of patterns of creative language use

MARTA GROCHOCKA

Adam Mickiewicz University

Abstract
Language development is reflected, inter alia, in the addition of neologisms to the lexicon. First, a definition of neologism is provided, which seems crucial in view of the fact that the concept of "newness" is relative. For the purpose of neologism excerption, a corpus of British newspaper articles has been compiled and a web-based tool called NeoDet has been designed to make the process as automatic as possible. Once a list of neologisms has been extracted, the Internet is accessed as a corpus, using WebCorp, with the aim of eliciting the meaning of the neologisms by looking at the ways they are used in context. The analysis of neologism types is hoped to shed light on the current tendencies in the productive morphology of English since the final goal of the project is to draw conclusions regarding the treatment of creative patterns of usage in pedagogical dictionaries of English.
Keywords: neologism, neologism typology, semiautomatic detection, monitor corpus, exclusion corpus

Resumen
El desarrollo de la lengua está reflejado, entre otros aspectos, en la incorporación de los neologismos al lexicón. En primer lugar, se proporciona la definición de "neologismo", lo cual parece crucial si uno tiene en cuenta la relatividad del concepto "novedad". Con el fin de localizar los neologismos, se ha recopilado un corpus lingüístico a base de la prensa británica, y se ha creado NeoDet, una herramienta en línea, para automatizar al máximo el análisis. Una vez obtenida la lista de neologismos, se accede a Internet a modo de corpus mediante WebCorp, con vistas a dilucidar su significado tras interpretar sus contextos de uso. El análisis de varios tipos de neologismos pretende arrojar luz sobre las tendencias actuales de la morfología productiva inglesa, puesto que el objetivo final del presente proyecto es sacar conclusiones en cuanto al tratamiento de patrones de uso en los diccionarios pedagógicos del inglés.
Palabras clave: neologismo, tipología de neologismos, detección semiautomática, corpus de estudio, corpus de exclusión

1. THE NOTION OF NEOLOGISM

In order to be able to study neologisms, it is crucial to specify what exactly is meant by a new word. Firstly, it is vital to specify if a neologism is only a new word or maybe also a new phrase or expression, or a new grammatical construction, not to mention a new sense of an already existing word. Furthermore, the notion of "being new" is vague and relative as there is no other way to define it but to juxtapose it with "being old". Hence, the question that may be raised is how young a word should be in order to be considered a neologism. One may also wonder what frequency and range of usage it should exhibit to be finally incorporated in a dictionary. Any attempt to provide answers to these questions will unavoidably involve decisions that are at least to some extent arbitrary.

As aptly put by Metcalf (2002: VII), the definition of a neologism should not be restricted solely to new words. Instead, it should also encompass phrases, acronyms, and affixes, so it seems more plausible to use the term "a new vocabulary item", rather than "a new word". The rationale behind it is the simple fact that "units of meaning do not always correspond to single words". To illustrate the problem, he offers such examples as the acronym "PC" or the suffix "-gate" to signify a scandal.

Also for Hartman and James (2008: 99) a neologism can be either a word or a phrase which has entered the language "relatively recently" and is "often commented on and collected in specialized dictionaries". Borrowing, coinage or semantic change are listed as the processes which lead to the appearance of a neologism. The trouble is that "relatively recently" is a very relative term and it can be interpreted in a variety of ways.

The problem was first raised by Rey (1975, as cited in: Janssen 2009: 3-4) who provided a definition of a neologism which has been frequently cited and referred to. According to him, we can speak of a neologism when a word is regarded as new by the language community. Not only is the notion of novelty highly subjective, but also the reliability of ordinary language users as a source of information may prove doubtful. This type of definition is of little use when confronted with the practical task of neologism detection in the monitor corpus.

Cabré (1999: 205) lists four parameters which can be used in order to define a neologism, i.e.: diachrony (a neologism is a unit which has recently entered the lexicon), lexicography (a neologism is a unit which cannot be found in a dictionary), systematic instability (a neologism is a unit whose form or/and meaning is/are not stable yet), and psychology (a neologism is a unit which is considered new by the speakers)[1]. As she points out, neologists give preference to the lexicographic parameter as being the most systematic one, that is to say, a lexical item is regarded as a neologism on condition that it does not occur in the reference corpus (also called the exclusion corpus or the lexicographic corpus).

Janssen (2009) makes a clear distinction between a lexicographic definition and a corpus-based definition of a neologism, contingent on exclusion from a dictionary and a reference corpus respectively. He discards the two definitions as unsatisfactory for the practical task of semi-automatic neologism tracking, and instead proposes a combination of the two, which he terms the extended lexicographic diachronic definition, or a controlled corpus-based definition. The method consists in using a morphological database created on

[1] When applying Cabré's criteria, Rey's definition of a neologism can be said to be based on psychology.

the basis of a dictionary as the exclusion list. However, absence from the database is not a sufficient condition for a word to be regarded as a neologism, as words may be lacking from the database due to other reasons, such as their semantic transparency or spatial restrictions. Therefore, the next step is to manually analyze the corpus in order to verify if the morphological database lacks a word because it is indeed a very recent coinage, or maybe it was a lexicographer's conscious choice to exclude it from the dictionary. Only words which are not present in the morphological database because of their newness are deemed neologisms. Furthermore, in order to succeed in passing through the corpus verification process, the neologism candidate must be a correct lexical item (e.g. not a typo or a proper name) and the frequency of its occurrence cannot exceed a certain previously established frequency threshold. Also a threshold period is settled specifying how old a word should be to cease to be treated as a neologism, i.e. words from texts which are more than 3 years old are discarded as already established in the language. As Janssen (2009) himself points out, there are a few arbitrary parameters in the method which need to be decided upon (i.e. the content of both the morphological database and the exclusion corpus, the threshold frequency and the threshold period). Nevertheless, having delineated these parameters, the definition provides clear enough criteria to be applicable in practice. In fact, the extended lexicographic diachronic definition has already been applied by the Observatório de Neología de Português (ONP) to track neologisms in European Portuguese with the use of *NeoTrack*, an on-line neologism tracker integrated with *MorDebe*, a morphological database of Portuguese (Janssen 2005).

2. In search of neologisms

2.1. Aims of the study

The goal of this study is to extract a list of neologism which can later be analyzed in order to corroborate or reject the hypothesis formulated by Moon in her article "Lexicography and Linguistic Creativity" (2008). According to Moon, creativity in language is frequently systematic. She conducted a study on four types of lexical creativity: figurative meaning, word formation (or affixation), idioms and spelling. Additionally, she discussed their treatment in three monolingual pedagogical dictionaries. She drew the data from the 450-million word Bank of English, and though she claims her examples are representative, she also admits that her approach is not truly scientific as she investigated only a few selected

cases which came to her mind. No systematic methodology was applied as far as the selection of the sample is concerned. Finally, she also points to the fact that learners' dictionaries fail to provide adequate treatment of such creative but systematic language phenomena.

Therefore, the aim of the present study is to uncover patterns of creative language use using a genuinely scientific method. Trends and regularities may be observed on various levels: phonological, morphological, syntactic, semantic, or pragmatic. The present study aims at tracking systematicity only on the morphological level, so that figurative meaning and idioms are excluded from the scope of interest.

In order to be able to achieve this aim, it is necessary to compile a corpus and work out a systematic methodology of neologism extraction.

2.2. The monitor corpus

Contemporary lexicography is inextricably linked with corpus research. Hence, if the object of the study is to make the process of neologism excerption more scientific, a monitor corpus has to be compiled. Moreover, in order to keep up with the aspirations to create a balanced corpus, clear selection criteria need to be established and all the subjective choices regarding its design, such as the content and the size of the corpus, should be determined by its intended function (Atkins & Rundell 2008: 54-55). The monitor corpus which is being compiled for the purpose of this study consists of newspaper articles. It is a lexicographic corpus as it has been created with a view to generating a list of neologisms that could later be used to create a dictionary of neologisms. An attempt has been made to make the corpus representative by employing a stratified sampling method (Atkins & Rundell 2008: 64-68). In crude terms, it boils down to including many different samples from various sources. Collecting fewer articles but from various sources seems a better strategy than collecting a larger number of articles from one source only.

The monitor corpus consists of newspaper articles published after 1st January 2010. The newspapers that have been selected for inclusion are the three most widely read British daily broadsheets, i.e. *The Daily Telegraph, The Times* and *The Guardian* – with the readership of 1.8m, 1.8m and 1.2m respectively – as well as two tabloids, *The Sun* and *The Daily Mail*, with the readership of 7.8m and 4.8m[2].

The monitor corpus is thus a monolingual one. It is comprised of articles taken from the British online newspapers so the language variety it represents is journalistic text. It is a

[2] According to the National Readership Survey (the figures represent the twelve months to June 2009), source: http://www.mediauk.com

synchronic corpus as the constituent texts are very recent, published in a specified short period of time. The corpus is quite homogenous as the text types it incorporates include only newspaper articles and blogs.

Journalistic texts have been chosen for two reasons. Firstly, the availability of online newspaper articles makes access to such data relatively easy. Secondly, I would venture a guess that journalistic texts seem to exhibit sensitivity to creative language use and are prone to resort to neologisms in order to be catchy and more attractive to potential readers. Newspapers constitute a plausible choice for one more reason, namely, they touch upon different subject matters, hence, the monitor corpus is hoped to provide evidence for a lot of uses of different neologisms in a variety of subject domains.

As far as the proportion of the texts in the corpus is concerned, it has been decided to take into account the number of words from all the articles taken from a given newspaper and to represent each of the newspapers in equal proportions, i.e. having more or less the same number of words. Words rather than articles have been chosen as a measure due to the fact that, depending on the newspaper, the articles have very different average lengths.

As for the size of the corpus, it needs to be noted that neologisms tend to be low frequency items. They may even occur as hapax legomena, i.e. occur only once in the monitor corpus (Paryzek 2008: 164). Bearing this in mind and in the light of Zipf's law pointing to the very low frequency of occurrence of less common words and rarer usages (Atkins & Rundell 2008: 59-61), it is crucial to have a huge amount of data to fish out as many neologisms as possible. Concluding, the bigger the corpus the better, as it will yield more neologism candidates.

2.3. NeoDet – neologism detector

With an aim to make the process of neologism extraction as automatic and objective as possible, a computer program called NeoDet has been developed. It is language independent and works on a corpus not marked morphologically. It works on the exclusion list principle, just as most of such projects do, for example NeoTrack (Janssen 2005) or Buscaneo (Cabré & Bagot 2009). In other words, it operates according to the lexicographic definition and the corpus-based definition of a neologism, that is to say, a word is regarded as a neologism (or rather a neologism candidate) if it occurs neither in the corpus nor in the existing dictionaries. In a nutshell, the program first creates a list of all the words occurring in the monitor corpus and then checks if the words appear in the exclusion corpus which is comprised of the British

National Corpus and a few online English monolingual dictionaries. Only words which are not found in any of the sources are classified as neologism candidates for further analysis.

NeoDet is a web-based software which consists of four major components: *Words*, *Add article*, *Articles* and *Statistics*. The Words section provides access to a list of all the words appearing in the database, a list of words from the exclusion list (i.e. dictionary words), neologisms candidates, and words already assigned neologism status. It is also equipped with the Exceptions option which contains all the non-neologisms that have not been found in the exclusion list but cannot be classified as neologisms due to being a proper name, a typo, a part of a citation in a foreign language or a non-word such as an e-mail address.

The *Add article* section serves the purpose of uploading articles to the monitor corpus. Information concerning the title, the date of publication, the name of the newspaper and the newspaper section the article has been copied from is saved along with the article itself. All these options are to be chosen from drop-down menus. A lot of sections employed by different newspapers have been reduced to seven: *Blog, Business and Finance, Entertainment, Environment, News, Science and Technology, Sport*. If the name of the section is different from the ones mentioned (e.g. *Life & Style*), it is indicated in the *Notes* section.

The *Article* section is where all the articles can be accessed individually and the search can be filtered by the title, the newspaper, the newspaper section, the date of publication and the date the article was added. Each and every article can also be downloaded or deleted from the database. The articles are saved in the *.txt format so that the database can be manipulated with another piece of software if needed. For instance, the articles can be uploaded to one of the concordance programs, like *MonoConc Pro* or *WordSmith*, in order to leave the possibility of applying another method of neologism detection, e.g. the examination of lexical discriminants.

In the *Statistics* section information is provided on the total number of words occurring in the corpus, the number of unique words as well as the number of words that have been classified as neologisms. From this section one can also learn how many articles have been uploaded and what the figures are regarding the number of words per newspaper and per section. In addition, the average article length for each newspaper is calculated. All the figures are also presented in the form of bar charts.

2.4. The exclusion corpus

The exclusion corpus is comprised of the British National Corpus (1991-94) and the following online dictionaries: LDOCE5 (2009), CALD3 (2008), MEDAL2 (2007), OALD7 (2005), but the list is open-ended and there are plans to add more dictionaries, such as dictionaries of slang. It is also possible to add to the exclusion list some of the word lists of proper names and geographical names available on the Internet. This would make the filtration process more efficient, yielding much less noisy output.

The words from the monitor corpus are looked up in the BNC and the online dictionaries in the same way as if they were typed in the dictionaries' search engines by a dictionary user. For example, if the plural form of a noun or the past participle of a verb is looked up, the dictionaries recognize the form and redirect the user to the basic form of the word. Hence, it can be said that the dictionaries are equivalent to the morphological database used by the already mentioned *NeoTrack* tool.

The advantage of this approach is that it is possible to find out how successful each dictionary from the exclusion corpus is in dealing with neologisms by looking at the number of words which were absent from the dictionaries (i.e. neologism candidates).

2.5. Neologism management

Once the neologism candidates have been analyzed and all the false candidates have been eliminated, a list of neologisms can be generated. The NeoDet application enables a researcher to manage the database of neologisms by indicating the citation form of each new word, its possible typographic variations, the syntactic category and the neologism type (the last two to be chosen from a drop-down menu). In addition, the meaning of a neologism can be provided in the Definition section. Sometimes the meaning of a new word is explicitly explained in the article/s it comes from which is actually to be expected, especially when dealing with a very recent neologism. However, it is not always the case and sometimes the meaning has to be elicited from the context, following Hanks's advice to map meaning onto use (Hanks, 2002). As the context in which a given neologism appears in the monitor corpus may prove insufficient to be able to arrive at a satisfactory definition, a link is provided to the WebCorp tool, a linguistic search engine that can be used to get access to the World Wide Web as a corpus. Thanks to it, it is possible to establish the meaning of a neologism by means of looking at more ways in which it is used on the Internet.

Last but not least, in the *Neologism management* section of *NeoDet* one can find statistical information regarding the parts of speech and the neologism types in the database.

Information is also provided on the number of occurrences of neologisms as well as the number of sources in which they occur.

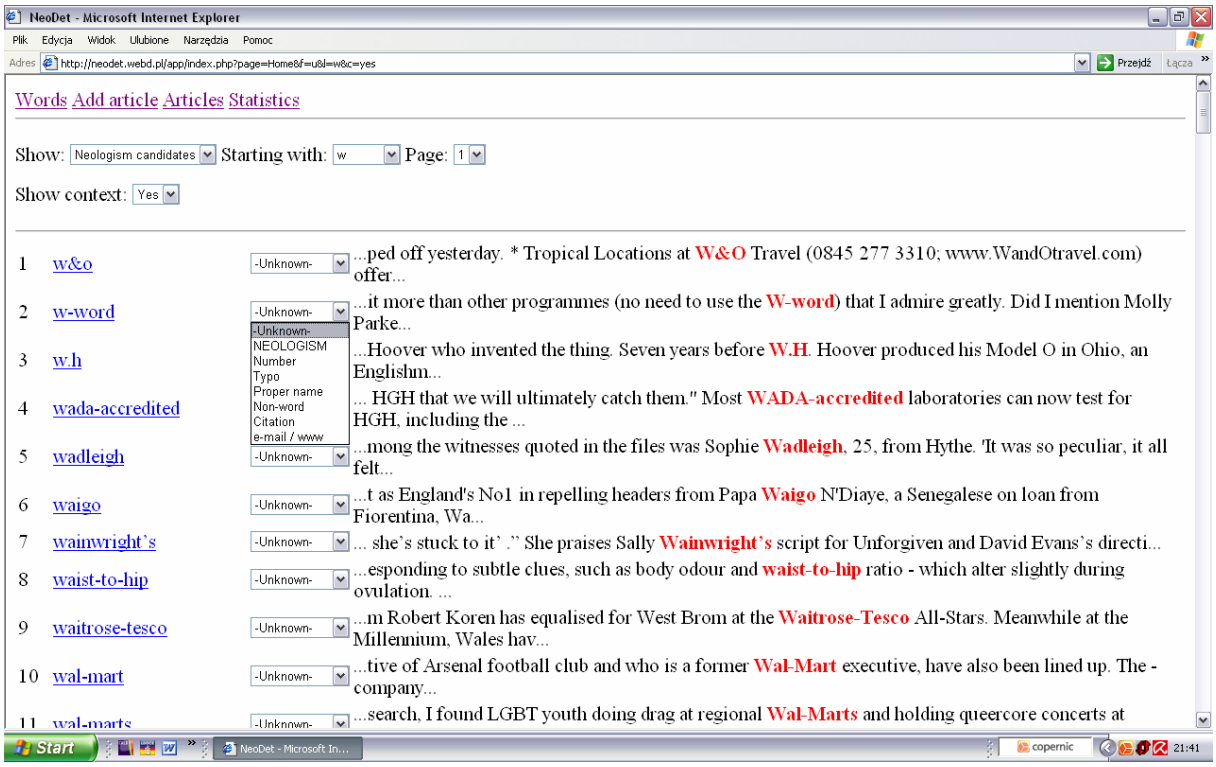


Figure 1: Neologism candidates management in *NeoDet*.

3. ONGOING PROJECT

The study is an ongoing project and both the monitor corpus and the exclusion corpus are expanding. Certain adjustments are still being made to the NeoDet application as well, such as adding additional statistical information, making it a more and more powerful tool. As has already been mentioned, the aims of the study are definitely not restricted to creating a database of English neologisms. The analysis of the collection of neologisms is hoped to shed light on the current trends and tendencies in the productive morphology of English which is expected to carry practical implications for lexicography, in particular, it is hoped to contribute to a better treatment of creative patterns of usage in pedagogical dictionaries of English. This is in accordance with Moon (2008), who postulates that creative aspects of lexis and the systematicity of creative usage are ignored or neglected by dictionaries. Such a state

of affairs may be quite frustrating, especially to language learners who are left to their own devices in the search for trustworthy information if a dictionary fails to satisfy their needs.

REFERENCES

Atkins, B. T. S. & Rundell, M. (2008). *The Oxford guide to practical lexicography.* Oxford: Oxford University Press.

Cabré, M. T. (1999). *Terminology: Theory, methods, and applications.* Amsterdam: John Benjamins Publishing Co.

Cabré, M. T. & Bagot, R. E. (2009). Trabajar en neologia con un entorno integrado en linea: La estacion de trabajo OBNEO. *Revista de Investigación Lingüística 12*: 17-38.

Hanks, P. (2002). Mapping meaning onto use. In M. H. Corréard (Ed.). *Lexicography and natural language processing: A festschrift in honour of B. T. S. Atkins.* (pp. 156-198). Grenoble: EURALEX.

Janssen, M. (2005). NeoTrack: Semiautomatic neologism detection. Lisboa: ILTEC. Retrieved from http://maarten.janssenweb.net/Papers/APL2005-mjanssen.pdf

Janssen, M. (2009). Orthographic neologisms: Selection criteria and semi-automatic detection. Retrieved from: http://maarten.janssenweb.net/Papers/neologisms.pdf

Metcalf, A. (2002). *Predicting new words: The secrets of their success.* Boston: Houghton Mifflin Company.

MonoConc Pro. (2009). Retrieved from: http://www.athel.com/mono.html

Moon, R. (2008). Lexicography and linguistic creativity. *Lexicos 2008*: 131-153.

Paryzek, P. (2008). Comparison of selected methods for the retrieval of neologisms. *Investigationes Linguisicae 16*: 163-181.

Rey, A. (1995). The concept of neologism and the evolution of terminologies in individual languages. In J. C. Sager (Ed.). *Essays on terminology.* (pp. 9-28). Amsterdam: John Benjamins Publishing.

WebCorp Live. (2010). Retrieved from: http://www.webcorp.org.uk/index.html

WordSmith Tools. (2009). Retrieved from: http://www.lexically.net/wordsmith/.

Errores en la traducción de textos veterinarios del inglés al español: los corpus lingüísticos y los recursos *web* como asistentes en la traducción

FRANCISCO GUTIÉRREZ
PASCUAL CANTOS
Universidad de Murcia

Resumen
Entre los recursos tecnológicos que se vienen utilizando en el proceso de traducción destacan las memorias de traducción y las bases de datos terminológicas. En cambio la incorporación de los corpus lingüísticos y los recursos de corpus en la web es escasa, especialmente en la traducción de textos especializados, caso de la traducción veterinaria. En esta comunicación se alude a la necesidad de contar con dichos recursos. Se ilustran los distintos tipos de errores y la posibilidad de subsanarlos y/o evitarlos con la ayuda de corpus generales o el uso de la propia web como corpus. Se constata el potencial de estas herramientas y se discute la precisión en la traducción del lenguaje veterinario que se podría alcanzar si existieran corpus lingüísticos específicos y bases de datos propias de dicho lenguaje.
Palabras clave: lenguaje veterinario, traducción, bases terminológicas, corpus lingüísticos específicos, errores de traducción, web como corpus.

Abstract
Among the technological aids that are most commonly used in translation, we have translation memories and terminology databases. However the integration of linguistic corpora and web-corpus resources are scarce, especially in the translation process of specialized texts, such as veterinary ones. This paper aims at stressing the need for such resources. We shall illustrate different types of common translation errors and how to solve and/or avoid them by means of general corpora or by using the web as a corpus. We shall also highlight the potential of these tools and discuss the improvement in accuracy of veterinary text translations if we had at our disposal domain specific linguistic corpora and databases on animal health.
Keywords: veterinary language, translation, terminology data bases, specific linguistic corpora, translation errors, web as corpus.

1. INTRODUCCIÓN

Son varios los autores que han resaltado las bondades de los corpus lingüísticos y los recursos de corpus en la web para la práctica profesional de la traducción (Olohan, 2004; Beeby et al., 2009). El valor de los corpus en la traducción viene dado por las evidencias objetivas (datos cuantitativos y cualitativos) que estos nos pueden proporcionar a la hora de la toma de decisiones (Baker, 2004: 184; Shreve, 2006: 311), bien durante el proceso de traducción o en la evaluación posterior de las propias traducciones. Sin embargo, la integración del uso de los corpus lingüísticos en la práctica profesional de la traducción de lenguajes especiales no se ha producido de la misma medida (Bernardini y Castagnoli, 2008, Bowker y Barlow, 2008).

Esta tendencia se hace especialmente patente en el ámbito de la traducción del lenguaje veterinario. Esta parcela prácticamente no ha recibido atención alguna (Gutiérrez y Crespo 2004), si bien pueden serle de aplicación las consideraciones vertidas hasta el momento sobre la traducción del lenguaje biosanitario en general y la traducción de textos médicos en particular. En cualquier caso, si bien el lenguaje veterinario se enmarca dentro del lenguaje biosanitario, sigue vigente su especificidad terminología con respecto a otros lenguajes del mismo ámbito (como el biológico o el médico). Una forma de apreciar o dirimir lo que es común a la traducción de los distintos lenguajes biosanitarios y lo que es específico de la traducción de cada uno de ellos, se conseguiría precisamente mediante la compilación de un corpus específico y una base de datos terminológica derivada del mismo. En el caso del lenguaje veterinario, no existe ninguna base terminológica que sirva de herramienta para la práctica de la traducción y la evaluación de las traducciones.

Algunas de las principales razones que justifican el escaso impacto de los corpus en la traducción especializada veterinaria están relacionadas con: (i) los recursos y el esfuerzo que requiere la compilación de un corpus de estas características; y (2) el hecho de que el resultado que el uso de los corpus tienen sobre la productividad y calidad de las traducciones no sea tan inmediata, como ocurre con otros recursos (p. ej. las bases terminológicas, las memorias de traducción, etc.).

Kilgariff y Grefenstette (2003) proponen en cambio el propio uso de Internet como corpus, en vez de la compilación de corpus *ad hoc*. A priori, esta propuesta de uso sistemático de la *web como corpus* podría conllevar algunos riesgos, a la vez que entrañar dificultades a la hora de la toma decisiones traductológicas ya que muchos de los textos en Internet vienen plagados de anglicismos léxicos, sintácticos y estilísticos, traducciones erróneas o inadecuadas (House, 2001).

En el presente caso, se trata de ilustrar una tipología de errores ampliamente aceptada en la literatura traslaticia y describir cómo subsanarlos mediante el uso y manipulación de distintos tipos de corpus en el proceso de traducción. En dicha tarea utilizaremos ejemplos de errores encontrados en varios documentos de la Organización Mundial de Sanidad Animal[1], traducidos del inglés al español. Dichos errores emanan del uso de estrategias inadecuadas para la solución de los problemas traslaticios. Esos problemas se plantean a nivel lingüístico (morfosintáctico y léxico) y textual. Los errores del nivel textual pueden afectar a la

[1] http://www.oie.int/ESP/ES_INDEX.HTM

dimensión comunicativa (sobre todo a los registros lingüísticos), a los usos pragmáticos y al componente semiótico del contexto textual.

2. ERRORES GRAMATICALES: LOS CORPUS MONOLINGÜES

Entre los distintos tipos de corpus que pueden utilizarse en el campo de la traducción están los corpus monolingües. Son corpus con textos en una sola lengua (Kenny, 2001: 58), bien con muestras textuales generalistas o textos especializados. Los corpus monolingües resultan especialmente útiles como asistentes en la cuantificación de la calidad de las traducciones (Bowker, 2002) y/o para la extracción de terminología (Pearson, 1999).

En cuanto a la cuantificación de la calidad de las traducciones, los corpus monolingües ofrecen un sustento empírico especialmente valiosos para subsanar errores gramaticales: (i) calcos incorrectos; caso de *"gamma interferón" del inglés *gamma interferon*; (ii) usos incorrectos del gerundio: utilización de un gerundio en lugar de una cláusula de relativo con el verbo en forma finita y un pronombre relativo como sujeto; como en *"conteniendo cefixima y telurito potásico"; (iii) uso indebido de la premodificación; por ejemplo, *"un detallado protocolo" (a *detailed protocol*), en lugar de "un protocolo detallado"; o (iv) contraste en el orden de los elementos de la cláusula: en inglés, la posición inicial del sujeto suele ser invariable, mientras que en español hay una mayor flexibilidad al respecto: sin variar el contenido léxico-semántico de la frase, es común que el sujeto pueda ir en posición inicial o en posición post-verbal; la posposición del sujeto en español puede venir exigida por la idiosincrasia del verbo utilizado, con lo que estaríamos ante un uso exigido por la pragmática del castellano (en contraste con la del inglés).

Veamos la utilidad de los corpus monolingües aplicada a los calcos incorrectos. Además de la justificación lingüística, podemos verificar con datos objetivos el uso más aceptado/correcto para la traducción al español del término compuesto inglés *gamma interferon.* Una búsqueda en *WebCorp*[2] (Figura 1) de las traducciones potenciales de *gamma interferon* al castellano nos ofrecen la siguiente estadística:

[2] http://www.webcorp.org.uk

Tabla 1: Traducciones y porcentajes de *gamma/beta/alfa interferon*

Inglés	Español	Ocurrencias	%
gamma interferon	gamma interferón	54	30,86
	interferón gamma	121	69,14
	interferón de gamma	0	0,00
Total		*175*	*100,00*
beta interferon	beta interferón	56	30,77
	interferón beta	126	69,23
	interferón de beta	0	0,00
Total		*182*	*100,00*
alfa interferon	alfa interferón	100	23,70
	interferón alfa	322	77,30
	interferón de alfa	0	0,00
Total		*422*	*100,00*

Los datos son concluyentes, la traducción más usada y aceptada, y por ende la debería adoptarse para *alfa/beta/gamma interferon* (adjetivo/nombre 1 + nombre 2) es mediante la estructura de postmodificación (interferón alfa/beta/gamma; nombre 2 + adjetivo/nombre 1): 73,04 % frente al casi 27 % de traducción como resultado del calco incorrecto inglés. Por el contrario, no hemos encontrado ningún ejemplo de traducción con la estructura de postmodificación que incluya la preposición relacional *de* (nombre 2 + de + adjetivo/nombre 1).

Todavía más abrumadores son los datos que obtenemos en el *Corpus del Español*[3] en donde todas las 119 ocurrencias de "gamma" aparecen en función postmodificadora.

No obstante, un corpus monolingüe con textos solamente (palabras) no siempre es suficiente. A veces precisamos de corpus monolingües más sofisticados que permitan hacer búsquedas no simplemente de palabras o formas, sino también de palabras o formas asociadas de una determinada parte de la oración. Por ejemplo, buscar *que* cuando aparece como pronombre de relativo y no como nexo completivo. Es decir precisamos de un corpus monolingüe etiquetado morfológicamente. Uno de estos corpus es el *Spanish Web Corpus*[4] disponible con *Sketch Engine*[5].

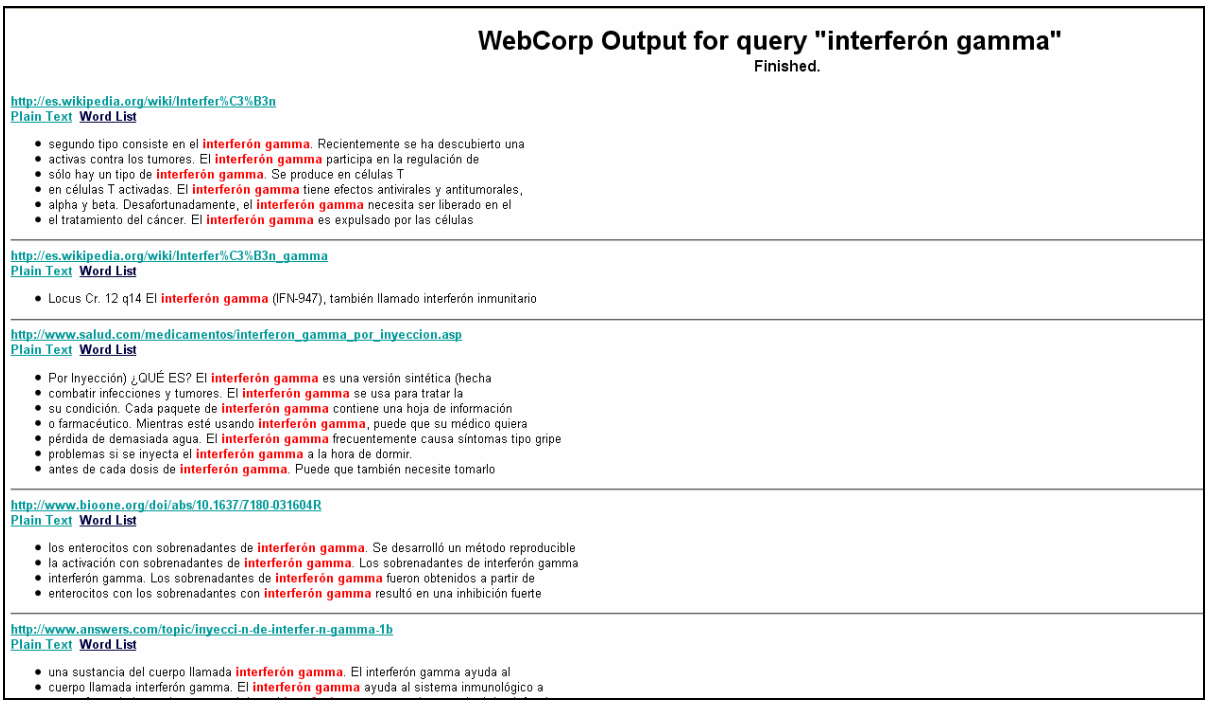
El análisis sobre los usos incorrectos del gerundio en español, arroja datos que no dejan lugar a dudas. El gerundio *conteniendo* aparece un total de 381 veces, de las cuales solamente en 47 ocasiones introduce una cláusula de relativo (Tabla 2). En cambio, la construcción

[3] http://www.corpusdelespanol.org
[4] Spanish Web Corpus es un corpus de español actual compilado con textos de páginas web; contiene 116.900.060 palabras.
[5] http://www.sketchengine.co.uk

pronombre de relativo + contener (en todas sus formas) sale en el *Spanish Web Corpus* 4.678 veces. O lo que es lo mismo, si eligiéramos al azar una frase en castellano que contenga el gerundio *conteniendo*, la probabilidad de que se use introduciendo una oración de relativo prescindiendo del pronombre de relativo, calcando la estructura del inglés conocida como *reduced relative clause,* es de únicamente el 1 %.

Tabla 2: Usos y porcentajes de conteniendo/que contener

conteniendo	introduce cláusula de relativo	47	12,34
	otros usos	334	87,66
	Total	*381*	*100,00*
conteniendo	introduce cláusula de relativo	47	1,00
pronombre de relativo + contener	introduce cláusula de relativo	4.678	99,00
	Total	*4.715*	*100,00*



WebCorp Output for query "interferón gamma"
Finished.

http://es.wikipedia.org/wiki/Interfer%C3%B3n
Plain Text Word List

- segundo tipo consiste en el **interferón gamma**. Recientemente se ha descubierto una
- activas contra los tumores. El **interferón gamma** participa en la regulación de
- sólo hay un tipo de **interferón gamma**. Se produce en células T
- en células T activadas. El **interferón gamma** tiene efectos antivirales y antitumorales,
- alpha y beta. Desafortunadamente, el **interferón gamma** necesita ser liberado en el
- el tratamiento del cáncer. El **interferón gamma** es expulsado por las células

http://es.wikipedia.org/wiki/Interfer%C3%B3n_gamma
Plain Text Word List

- Locus Cr. 12 q14 El **interferón gamma** (IFN-947), también llamado interferón inmunitario

http://www.salud.com/medicamentos/interferon_gamma_por_inyeccion.asp
Plain Text Word List

- Por Inyección) ¿QUÉ ES? El **interferón gamma** es una versión sintética (hecha
- combatir infecciones y tumores. El **interferón gamma** se usa para tratar la
- su condición. Cada paquete de **interferón gamma** contiene una hoja de información
- o farmacéutico. Mientras esté usando **interferón gamma**, puede que su médico quiera
- pérdida de demasiada agua. El **interferón gamma** frecuentemente causa síntomas tipo gripe
- problemas si se inyecta el **interferón gamma** a la hora de dormir.
- antes de cada dosis de **interferón gamma**. Puede que también necesite tomarlo

http://www.bioone.org/doi/abs/10.1637/7180-031604R
Plain Text Word List

- los enterocitos con sobrenadantes de **interferón gamma**. Se desarrolló un método reproducible
- la activación con sobrenadantes de **interferón gamma**. Los sobrenadantes de interferón gamma
- interferón gamma. Los sobrenadantes de **interferón gamma** fueron obtenidos a partir de
- enterocitos con los sobrenadantes con **interferón gamma** resultó en una inhibición fuerte

http://www.answers.com/topic/inyecci-n-de-interfer-n-gamma-1b
Plain Text Word List

- una sustancia del cuerpo llamada **interferón gamma**. El interferón gamma ayuda al
- cuerpo llamada interferón gamma. El **interferón gamma** ayuda al sistema inmunológico a

Figura 1: Ejemplo de búsqueda en *WebCorp*

3. ERRORES LÉXICOS: LOS BUSCADORES EN LA *WEB*

Según los autores del Proyecto MeLLANGE[6] la herramienta web más recurrida entre los traductores es el buscador Google. Se trata, pues, de utilizar la propia web como si de un

[6] http://mellange.upf.edu

corpus se tratara. Las búsquedas se realizan así en la mayor base de datos textual de que disponemos, compuesta de cientos de miles de textos reales.

La aplicabilidad de este recurso resulta inestimable especialmente para determinar y encontrar errores de tipo léxicos tan frecuentes como evitables. En esta tipología de errores de traducción abundan: (i) los anglicismos léxicos. Son bastantes los anglicismos léxicos firmemente establecidos en la literatura biosanitaria: *"tipado" (*typing*), en lugar de "tipificación"; "serotipado" (*serotyping*) en lugar de "serotipificación"; *"homogenado" (*homogenate*), en lugar de "homogeneizado"; y (ii) los falsos amigos, una tentación permanente para el traductor: *"bacterial" (*bacterial*), en lugar de "bacteriano/a".

Bastaría tomar como referencia los datos que la web nos ofrece al respecto para evitar muchos de estos anglicismos léxicos (Tabla 3).

Tabla 3: Traducciones y porcentajes de errores léxicos

Inglés	**Traducción español**	**Ocurrencias en la *web***	**%**
typing	tipado/a	20.200	3,02
	tipificado/a	649.000	96,98
Total		*669.200*	*100,00*
homogenate	homogenado/a	3.829	5,65
	homogeneizado/a	64.000	94,35
Total		*67.829*	*100,00*
baceterial	bacterial	331.000	19,01
	bacteriano/a	1.410.000	80,99
Total		*1.741.000*	*100,00*

Respecto a *serotyping*, la traducción más adecuada debería ser "serotipificación", por analogía con *typing* (tipificación), tanto por homogeneidad como por sistematicidad.

4. Errores léxicos, pragmáticos, semióticos y de registro lingüístico: los corpus paralelos

Los corpus paralelos están compuestos por textos originales en una lengua y las traducciones de los mismos en otra o varias lenguas (Laviosa, 2002: 37). Este tipo de corpus son de especial utilidad para contrastar y comprobar cómo otros han traducido previamente un término, unidad fraseológica, oración, párrafo, etc.; una funcionalidad común a las memorias de traducción. No obstante, la manipulación de estos corpus mediante programas de concordancias bilingües (véase ParaConc[7]) permiten búsquedas de una sola palabra o

[7] http://athel.com

segmentos oracionales, además de permitir examinar párrafos completos o la traducción del texto completo en lugar de pares de oraciones independientes (Bowker y Barlow, 2008).

En la Figura 2 se ilustra el uso de un corpus paralelo para comprobar y conocer cómo otros traductores han traducido al inglés la sigla "FA" del término "fiebre aftosa". Nótese que el programa (*ParaConc*) alinea los párrafos, es decir el primer párrafo del texto 1 con el primer párrafo del texto 2, etc. Otros programas más sofisticados alinean las oraciones (Moore 2002), con el consiguiente riesgo de posibles alteraciones de orden y/o agrupaciones de más de una oración en la traducción.

Los corpus paralelos, conocidos también como bi-textos, pueden resultar de gran utilidad para algunos problemas traductológicos: (i) de tipo léxico; (ii) utilización de términos propios de un registro informal en un registro formal; (iii) de tipo pragmático; y (iv) de corte semiótico.

Entre los errores léxicos que más fácilmente pueden subsanarse mediante los corpus paralelos están los falsos amigos; *"aparente" (*apparent*), en lugar de "manifiesto/visible/evidente"; *"inaparente" (*inapparent*), en lugar de "latente/asintomático/a".

Los errores relacionados con el registro lingüístico también pueden paliarse en gran medida con este recurso, especialmente los relacionados con la utilización de términos propios de un registro informal en un registro formal, o de una variedad técnica en una no técnica (o general), o a la inversa; tal es el caso del término "agallas" (propio de un registro no técnico) en lugar de "branquias" (propio de un registro técnico) o "descarga nasal" (no técnico) en lugar de "rinorrea" (técnico).

El uso sistemático de corpus paralelos puede asistir a los profesionales de la traducción veterinaria, también, en cuestiones de corte pragmático: (i) animismo vs. inanimismo: en inglés es común que las entidades inanimadas "ejecuten" acciones propias de las entidades animadas. En español no suele ser este el caso, salvo por la contaminación propiciada por las traducciones erróneas, como la del ejemplo que sigue: *A modification of the method of Witte ... uses ...*; por *"El método modificado de Witte … utiliza …"; en lugar de "En el método modificado de Witte … se utilizan …"; en la línea nada sutil que separa a los animales y los humanos, la pragmática del inglés y del castellano no siempre coinciden. Es frecuente en la traducción de los textos veterinarios el error consistente en asignar a los animales algunos rasgos propios de los humanos, como "embarazo/embarazada", en lugar de "preñez/preñada" (distinción inexistente en inglés, donde *pregnancy/pregnant* sería la opción común para los animales y los humanos). Otros ejemplos son "pie" (*foot*) en lugar de "pata" (*leg* o *foot*); (ii)

sustancia cuantificada + cuantificador vs. cuantificador + sustancia cuantificada: constituye un error la posposición de un cuantificador con respecto a la sustancia cuantificada: * "Se añade glutamina 2 mM" (*glutamine 2 mM is added…*); traducción alternativa: "Se añaden 2 mM de glutamina"; la posposición de la unidad de medida solo es aceptable en listados de sustancias o ingredientes que entran en una combinación; pero en ese caso ha de haber una coma o dos puntos entre la sustancia medida y la unidad de medida, o esta última debe colocarse entre paréntesis: "Ingredientes: glutamina, 2 mM"; o bien "Ingredientes: glutamina (2 mM)"; (iii) sujetos largos en inglés vs. sujetos cortos en español; en español, cuando el predicado verbal es corto y el sujeto es largo, aquél suele ocupar la posición frontal en la cláusula: *After harvest of the stock from the tank, all loose objects and large-sized organic debris such as algae, faeces and left-over feed should be removed*; *"Tras recolectar la existencia del tanque, todos los objetos sueltos y los restos orgánicos de tamaño grande, tales como algas, heces, y restos alimenticios deben eliminarse"; traducción alternativa: "Tras recolectar la existencia del tanque, deben eliminarse todos los objetos sueltos y los restos orgánicos de tamaño grande, tales como algas, heces, y restos alimenticios". A menudo, dependiendo de la idiosincrasia del verbo –por ejemplo cuando es claramente temático- debe ocupar la posición frontal según el esquema tema-rema: *An antigen-capture enzyme-linked immunosorbent assay (ELISA) for the detection of camelpox virus has been described*; *"Una prueba ELISA de captura de antígeno para la detección del virus de la viruela del camello ha sido descrita"; la alternativa obligatoria para esa versión errónea es: "Se ha descrito una prueba ELISA de captura de antígeno"; y (iv) también son de orden pragmático los errores consistentes en la utilización de denominaciones inadecuadas: "peces salvajes" (*wild fish*), en lugar de "peces libres" o "peces silvestres", que son distintos de los peces de cultivo (*cultured fish*) o de criadero (*farmed fish*).

En los errores de tipo semiótico o de falta de conocimiento enciclopédico se pone de manifiesto un déficit en el conocimiento del contenido especializado por parte del traductor: *... and sheep erythrocytes, sensitised by the anti-sheep erythrocyte serum, are added*; *"…y se añaden eritrocitos de oveja que están sensibilizados frente a anti-suero contra eritrocitos de oveja"; traducción alternativa: "…y se añaden eritrocitos de oveja que están sensibilizados por el suero anti-eritrocitos de oveja". A veces la traducción literal puede cambiar el significado de la expresión. Tal es el caso de la traducción de *antimicrobial resistance* como *"resistencia antimicrobiana" (es decir, resistencia a los microbios), en lugar de la forma correcta "resistencia a los antimicrobianos" (es decir, resistencia de los microbios a los productos biológicos destinados a combatirlos).

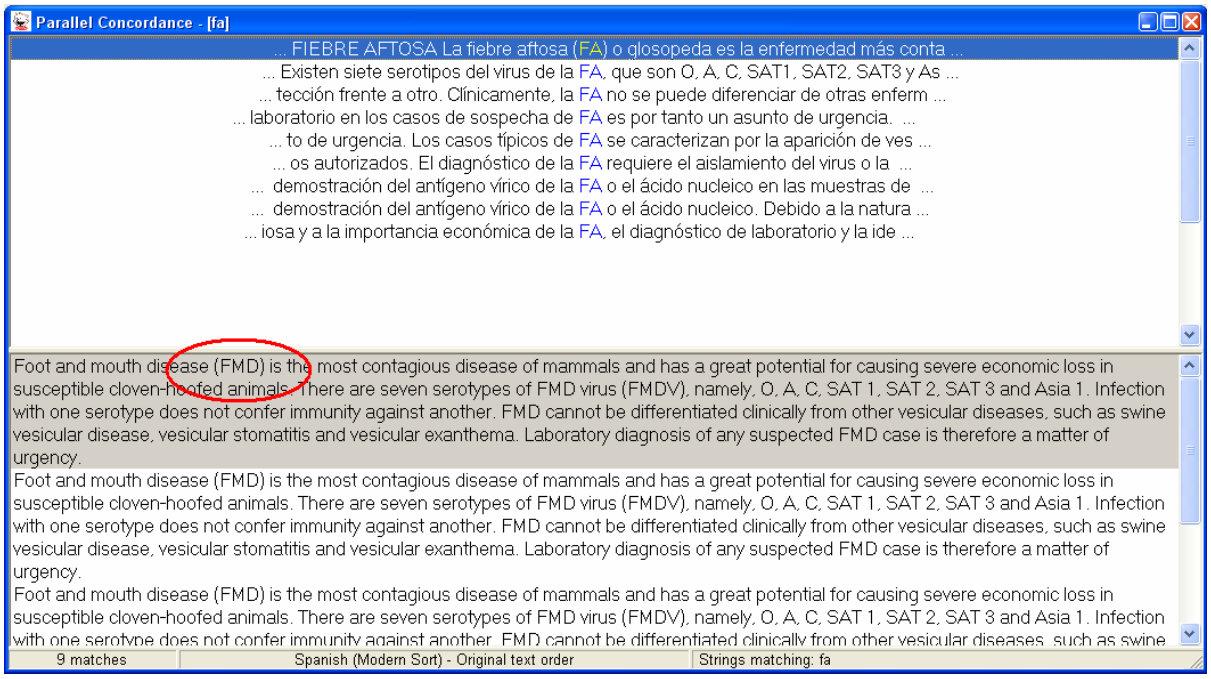


Figura 2: Exemplo de *ParaConc*

5. Compilación de corpus monolingües y bilingües específicos sobre lenguaje veterinario

Los ejemplos y muestras anteriores han dejado patente la utilidad de los corpus en las tareas de traducción de textos especializados, incluso los corpus generalistas no especializados. El valor añadido que aportan los corpus lingüísticos como asistentes en la cuantificación de la calidad de las traducciones no es comparable con ningún otro medio. No obstante, el panorama se susceptible de mejorar aún más si disponemos de corpus lingüísticos especializados tanto monolingües como bilingües. Para ello es preciso erradicar la idea que algunos traductores tienen sobre el alto coste en recursos y esfuerzo que requieren la compilación de un corpus de estas características. Es cierto que los primeros corpus precisaban de presupuestos y recursos importantes para su diseño y compilación. No obstante, existen en la actualidad medios tecnológicos que facilitan enormemente esta tarea, y la inversión en tiempo y medios es ínfima.

Una de estas aplicaciones tecnológicas para la compilación de corpus es *BootCat*[8] (Baroni et al., 2006). Sucintamente, basta con definir: (i) la temática del corpus (mediante

[8] http://sslmit.unibo.it/~baroni/bootcat.html

palabras clave), (ii) el idioma de los textos, (iii) si deseamos etiquetar morfológicamente el corpus; y (iv) el nombre del corpus (*Animal Health*; Figura 3). *BootCat* rastreará la *web* en busca de textos que satisfagan los criterios de temática e idioma, para a continuación presentarnos las páginas en donde ha localizado textos relevantes. El usuario solamente tiene que cotejar las direcciones y/o textos preseleccionados y confirmar aquéllos que mejor estime (Figure 4), y el corpus monolingüe sobre lenguaje de sanidad animal quedará listo para el investigador/traductor, véase la consulta de la sigla *FMD* (*Foot and Mouth Disease*; Fiebre Aftosa; Figura 5). El tiempo que nos ha llevado crear este corpus han sido escasos dos minutos.

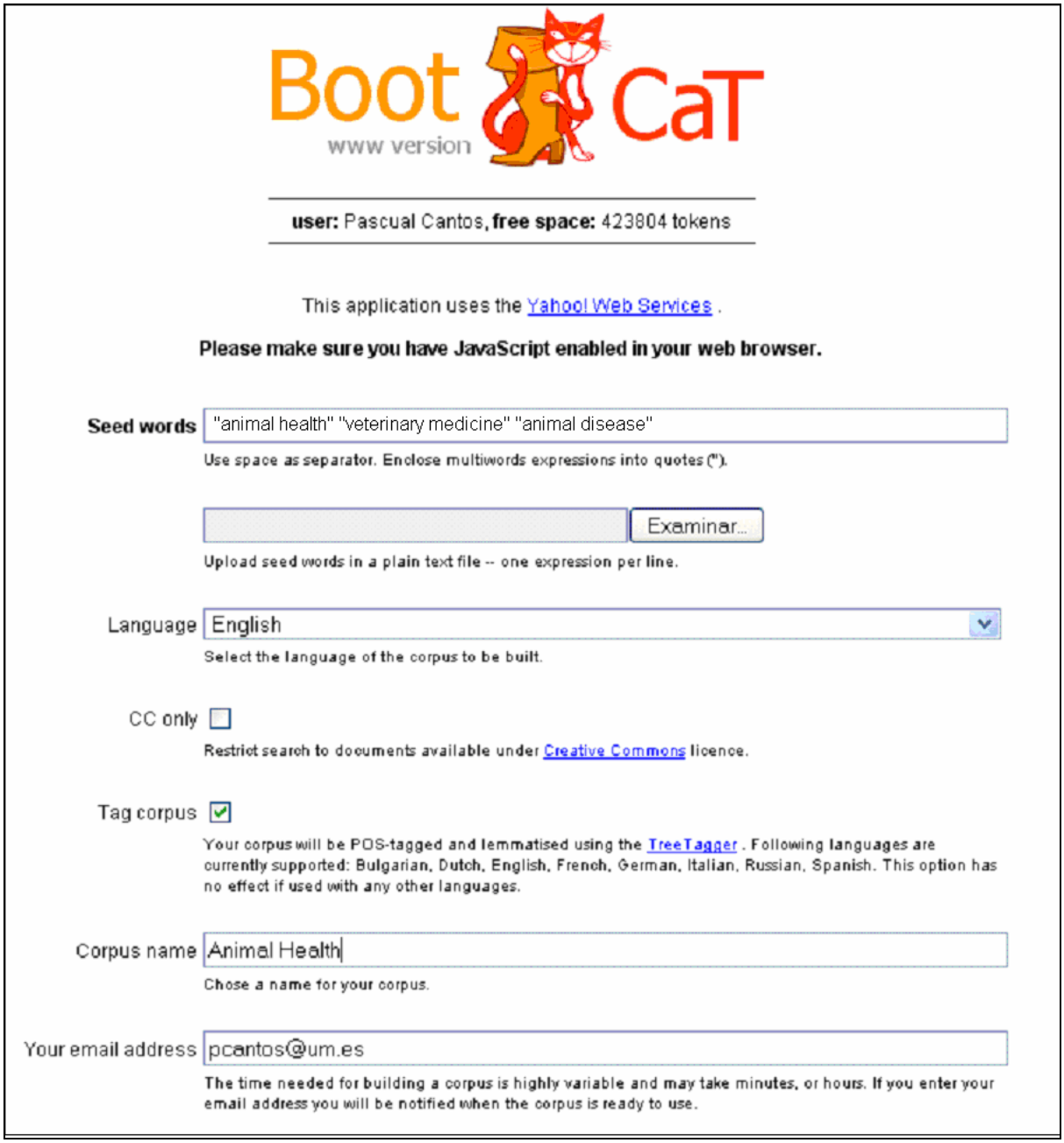


Figura 3: Creación de un corpus con *BootCat*

Para la compilación de un corpus paralelo, el proceso es más laborioso además de entrañar otras dificultades añadidas, algunas de las cuales son de muy difícil solución. Una primera aproximación consiste en buscar páginas *web* bilingües o multilingües, caso de la página Organización Mundial de Sanidad Animal (Figura 6) e intentar bajar documentos de libre acceso en inglés y sus traducciones al español con el fin de ir compilando dos corpus monolingües específicos de idéntico contenido, para luego utilizarlas de forma conjunta con un herramienta de concordancias bilingües que de forma automática alinee los textos fuente con las traducciones, bien párrafo con párrafo u oración con oración (por ejemplo, *ParaConc*).

Las problemas mayores están en encontrar páginas *web* bilingües/multilingües de libre acceso que ofrezcan una cantidad importante de textos validados y actualizados (cuantos más mejor) y que las traducciones estén revisadas y validadas por traductores y especialistas en el tema.

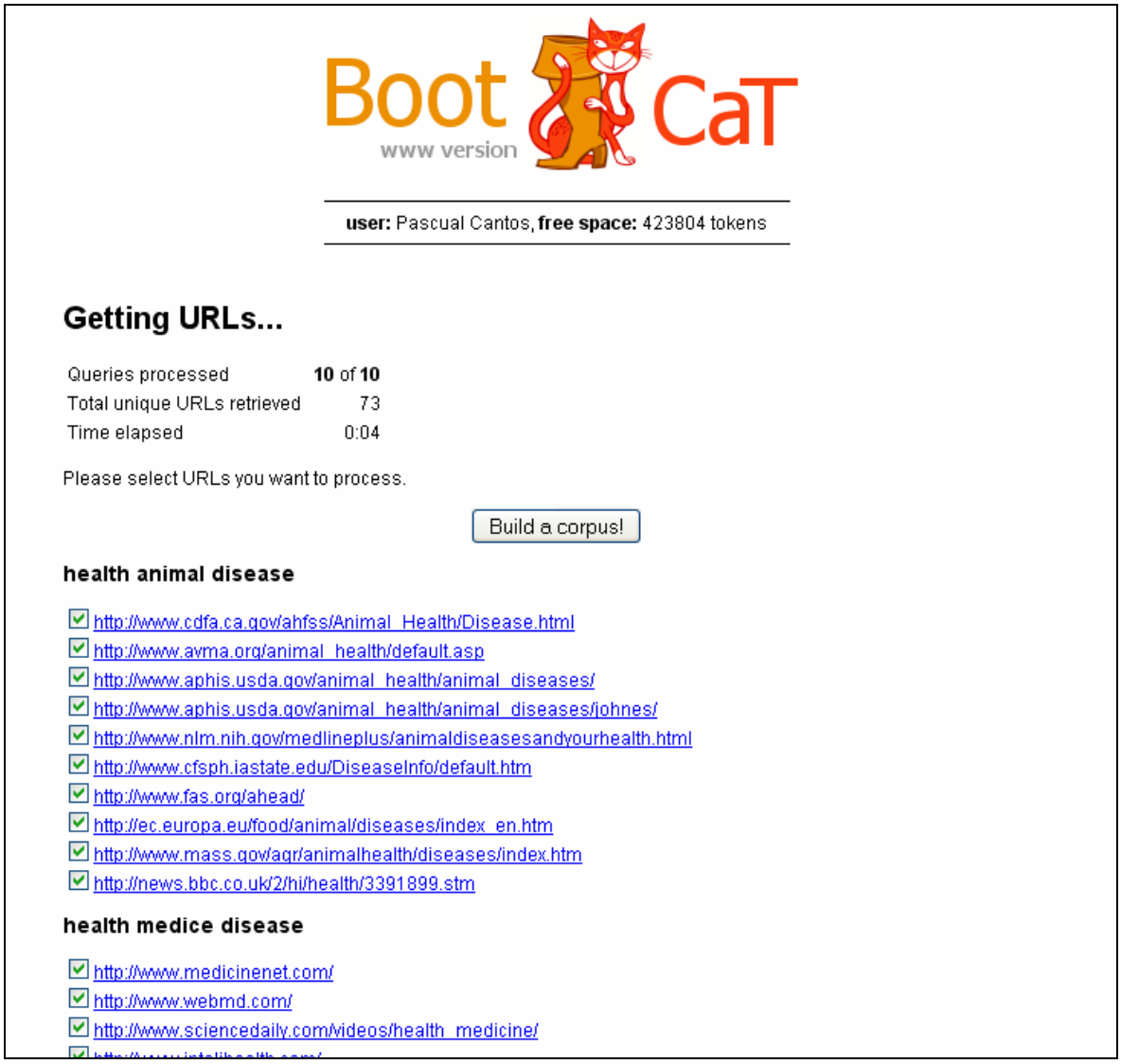


Figura 4: Creación de un corpus con *BootCat*: páginas *web* encontradas

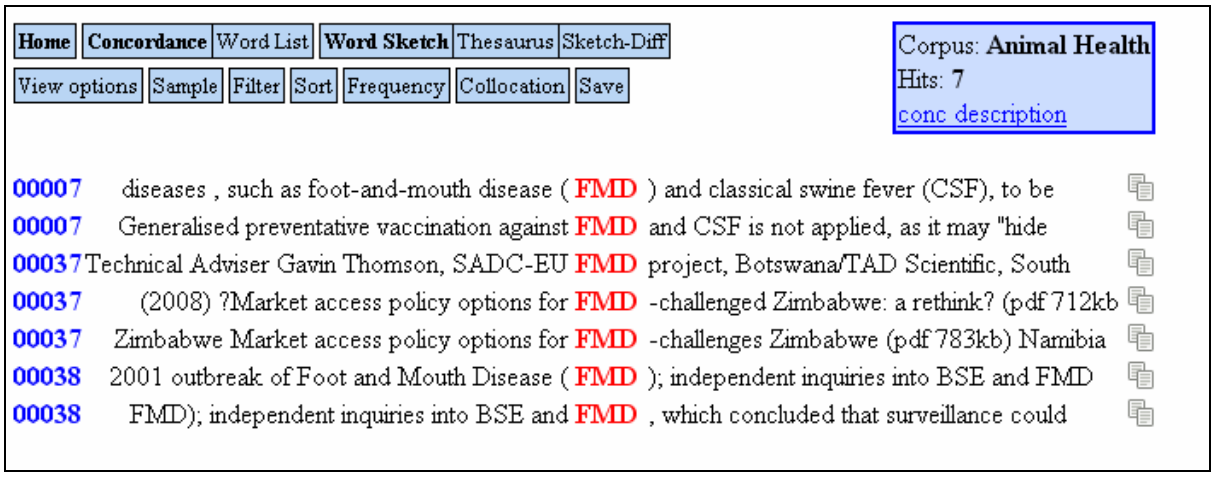


Figura 5: Ejemplo de consulta del corpus monolingüe sobre sanidad animal

Figura 6: Página de la OIE

6. CONCLUSIÓN

Hemos ilustrado los principales tipos de error traslaticio que se suelen encontrar en los textos veterinarios, utilizando varios documentos producidos por la OIE, y cómo con la integración de corpus lingüísticos y herramientas de corpus en la web en el proceso global de traducción se pueden subsanar gran parte de los errores descritos, además de propiciar la investigación empírica de los tipos y frecuencia de errores traslaticios y la verificación y/o elaboración de hipótesis explicativas de tales errores en el ámbito de la traducción veterinaria (inglés-español).

El camino se ha iniciado pero queda mucho por andar. Es preciso y urgente la elaboración de corpus lingüísticos del lenguaje veterinario, así como la de bases terminológicas bilingües de inglés-español emanadas de los mismos, que sirvan de herramientas para los siguientes fines: estandarización de la terminología veterinaria; filtración de términos y usos erróneos en las traducciones al español; y evaluación de la calidad de traducciones. La viabilidad de un proyecto de esta magnitud pasa necesariamente por la colaboración y sinergia de lingüistas, expertos en veterinaria e instituciones públicas (OIE, Ministerios de Agricultura, etc.).

Baker, M. (2004). A Corpus-based View of Similarity and Difference in Translation. *International Journal of Corpus Linguistics* 9 (1): 167-193.

Baroni, M., Kilgarriff, A. Pomikalek, J. y P. Rychly (2006) WebBootCaT: Instant Domain-specific Corpora to Support Human Translators. *Proceedings of EAMT-2006*, 247-252.

Beeby, A., Rodríguez Inés, P. y Sánchez-Gijón, P. (2009). *Corpus Use and Translation.* Ámsterdam-Filadelfia: John Benjamins.

Bernardini, S. y Castagnoli, S. (2008). Corpora for Translation Education and Translation Practice. En E. Yuste Trigo (ed.) *Topics in Language Resources for Translation and Localization*. Pp. 39-55 Ámsterdam-Filadelfia: John Benjamins.

Bowker, L. (2002). *Computer Aided Translation Technology: A Practical Introduction* Ottawa: University of Ottawa Press.

Bowker, L. y Barlow, M. (2008). A Comparative Evaluation of Bilingual Concordancers and Translation Memory Systems. En E. Yuste Trigo (ed.) *Topics in Language Resources for Translation and Localization*, (pp. 1-22). Ámsterdam-Filadelfia: John Benjamins.

García Yebra, V. (1982). *Teoría y práctica de la traducción*. Madrid: Gredos.

Gutiérrez F. y Crespo, F. (2004). La traducción al español del *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals*: soluciones presentes y futuras. *Revista Científica y Técnica*, 23 (3): 1023-1031. París: OIE

Hatim, B. y Mason, I. (1990). *Discourse and the Translator.* Londres: Longman.

House, J. (2001). Translation Quality Assessment: Linguistic Description versus Social Evaluation. *Meta* 46 (2), 243-257.

Kenny, D. (2001). *Lexis and Creativity in Translation. A corpus-based Study*. Manchester: St. Jerome.

Kilgarriff, A. y Grefensttete, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29 (3), 333-347.

Laviosa, S. (2002). *Corpus-based Translation Studies*. Ámsterdam: Rodopi.

Moore, L. (2002) Fast and Accurate Sentence Alignment of Bilingual Corpora. *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users,* (pp. 35-144). Londres: Springer Verlag.

Olohan, M. (2004). *Introducing Corpora in Translation Studies.* Londres: Routledge.

Pearson, J. (1998). *Terms in Context.* Ámsterdam: John Benjamins.

Puerta, J.L. y Mauri, A. (1995). *Manual para la redacción y publicación de textos médicos.* Barcelona: Masson.

Rabadán, R., Labrador, B. y Ramón, N. (2009). Corpus-based Contrastive Analysis and Translation Universals. *Babel***,** 55 (4): 303-327**.**

Shreve, G. M. (2006). Corpus Enhancement and Localization. En K. Dunne (ed.) *Perspectives on Localization*, (pp. 309-331). Ámsterdam-Filadelfia: John Benjamins.

Vázquez-Ayora, G. (1977). *Introducción a la traductología.* Washington: Georgetown University Press.

Towards a corpus of early American literature: on the challenges of compiling a comparable diachronic corpus

MIKKO HÖGLUND
KAJ SYRJÄNEN
University of Tampere

Abstract
This paper outlines a diachronic work-in-progress corpus of American English literature. The project is inspired by the Corpus of Late Modern English Texts (CLMET), comprising British English texts from 1710 to 1920. Our corpus (the Corpus of Early American Literature or CEAL) aims to cover roughly the same period of American English as the CLMET, using texts available in the public domain. The topic of our paper is essentially two-fold. First, an overview of American English history is provided, outlining linguistic development during the time span covered by our corpus. Second, the suitability of the CLMET's compilation criteria are evaluated against the linguistic history of American English on the one hand, and the available text material on the other.
Keywords: empirical linguistics, corpus compiling, American English

Resumen
Este trabajo perfila un corpus diacrónico en curso sobre la literatura inglesa americana. El proyecto está inspirado en el Corpus of Late Modern English Texts (CLMET), que consta de textos de inglés británico de 1710 a 1920. Nuestro corpus (el Corpus of Early American Literature, o CEAL) tiene por objetivo cubrir aproximadamente el mismo período del inglés americano que el CLMET, utilizando textos disponibles en el dominio público. El tema de nuestro trabajo está fundamentalmente dividido en dos partes. En primer lugar, se ofrece un resumen de la historia del inglés americano, trazando el desarrollo lingüístico durante el período cubierto por nuestro corpus. En segundo lugar, la idoneidad de los criterios utilizados en la compilación del CLMET son evaluados en cuanto a la historia lingüística del inglés americano por un lado, y el material de texto disponible por el otro.
Palabras clave: lingüística empírica, compilación de corpus, inglés americano

1. INTRODUCTION[1]

The use of text corpora in linguistic research has become widespread in recent decades. As one of the most used languages today, English has been in the forefront of corpus linguistics not only as a popular object of linguistic research but also as a language that most of the pioneering linguistic corpora have attempted to encapsulate. These pioneering corpora range from the 1-million word Brown and LOB of the 1960s to the later, more comprehensive corpora such as the 100-million word British National Corpus.

Different English variants, however, have not been evenly documented. For instance, British English text corpora are available to a greater extent than American English corpora.

1 The authors are grateful to Ian Gurney, of the University of Tampere, for proofreading the manuscript of this article and providing helpful comments. The authors are solely responsible for any remaining errors.

The availability of American material has improved recently thanks to large corpus projects such as the synchronic 400-million-word *Corpus of Contemporary American English* (Davies, 2009), but the currently available resources are still far from comprehensive, diachronic coverage. This is somewhat surprising, as a considerable amount of public domain American text material is relatively easily available in electronic format.

Some historical American English corpora exist or are being compiled – for instance, the *Corpus of Early American English*, a component of the *Helsinki Corpus* (Kytö, 1994), and Mark Davies' *Corpus of Historical American English*, a large, work-in-progress corpus covering American English from the early 19th century to the present day. However, historical language research, as well as the study of language change, would benefit from additional corpora documenting the historical stages of American English.

This paper introduces a diachronic work-in-progress corpus of American English literature. The project is inspired by Hendrik De Smet's freely available *Corpus of Late Modern English Texts* (*CLMET*) which comprises British English texts from 1710 to 1920, divided into 70-year subperiods. The corpus (which currently goes by the name *Corpus of Early American Literature*, or *CEAL* for short) aims to cover the same time span of written American English and ensure as much compatibility with the CLMET as possible, using texts available in large Internet text repositories such as *Project Gutenberg* and the *Oxford Text Archive*. The aim is to make the corpus available for research with both online and offline corpus tools.

Compiling an American counterpart to the British English CLMET is not a straightforward task. Aspects such as the socio-cultural differences between the two speaker communities, the nature of the available text material, and the required modifications to the original compilation criteria employed in the CLMET must all be taken into account. These matters are the primary focus of the present paper. In addition to this, the CLMET must be closely examined to ensure maximal compatibility between the two concrete corpora, not just their compilation criteria.

The paper is divided as follows: first, a brief historical account of American English is provided. This sets the stage for the following section, which reviews the compilation criteria of the CLMET and their suitability to the compilation of CEAL. This is followed by a brief section that focuses on problems that arise in applying the CLMET's compilation criteria in the context of American English.

2. THE EMERGENCE OF AMERICAN ENGLISH

The purpose of this section is to review some key points in the development of American English and establish the cultural and linguistic environment of the new corpus. This basic information is necessary for the compiling of the corpus, as socio-cultural differences between Britain and America affect the applicability of the corpus compilation criteria, making some criteria partially or wholly incompatible. This overview begins somewhat earlier than our corpus (the early 17th century) as it flows more naturally into standard historical accounts of American English.

2.1. General information

Nowadays it is fairly common to refer to the languages spoken in the United States and in the United Kingdom as American English (AmE) and British English (BrE), respectively. Generally they are considered to represent mutually comprehensible variants of the same language, English. It should be noted, however, that the definition of 'language' is fundamentally non-objective, and occasionally extralinguistic factors have caused more or less mutually comprehensible variants to be considered separate languages. Such is the case with Swedish, Norwegian and Danish, for instance (Chambers & Trudgill, 1980: 3–4). Consequently it is not surprising that BrE and AmE, too, have occasionally been called separate languages (for instance, Strevens, 1972: 21–22; Mencken 1936). In fact, it has been noted that variants of English are currently driven by opposite change pressures: one which drives Englishes towards mutual intelligibility and a unified 'Global English', and another which strives to foster separate national and linguistic identities (Crystal, 2003: 113).

At least in the time span we are interested in (1710–1920), several distinguishing features between AmE and BrE are encountered in an established or evolving state. These differences are found at various linguistic levels, e. g. lexicon, orthography, pronunciation and – to some extent – grammar. Differences are also encountered beyond in broader linguistic contexts, such as the realm of culture-specific idioms. For instance, the baseball-related idiom 'to be caught off base' in the meaning of 'to be caught in an embarrassing situation' (Tottie, 2002: 137) is generally considered to be primarily American, whereas the idiom 'to grasp the nettle' in the sense of 'to bravely deal with' is primarily British.

It is not straightforward to specify the exact time when the gap between American and British English had deepened to the extent that they assumed distinct linguistic identities. However, it is generally agreed that the separation of the two Englishes began already at the

turn of the 16th and 17th centuries when the first people migrated to the new continent of America in search of a better life (Strevens, 1972: 27; Crystal, 2003: 92).

2.2. The Colonial period

The first settlers founded the settlement of Jamestown in 1607 and the first permanent colony was Plymouth, founded by the Pilgrims in 1620. The first high tide of colonization started around 1630, when 20,000 Puritans from eastern England migrated to America. Fischer (1989) has suggested that the colonization of America proceeded in four major stages. Even though his proposal is an oversimplification, it presents a good general overview of the early stages of emigration to America:

1629-1641	Puritans from East Anglia (20,000)
1642-1675	Cavaliers and their servants from southern England (40,000)
1675-1725	Quakers from the North Midlands and evangelicals from Wales, Germany, Holland and France (23,000)
1717-1775	Common people from northern England, Scotland and northern Ireland (275,000)

More than 95 % of the settlers in the original colonies were from Great Britain, making English the obvious choice of primary language in Colonial America (Fisher, 2001).

There are many reasons as to why American English began to diverge from British English. First and foremost was geographical distance. With the ocean separating the American settlers from the British, communication between the two speaker communities was not frequent. In addition to this, traveling was time-consuming and expensive. Also, the new environment and distinct social and economic conditions in America imposed new language requirements. New words were borrowed from Amerindian languages and the languages of other settlers: French, Spanish, Dutch, German, and so on. New words were also coined and old words put to new uses.

As a whole, the English-speaking community of Colonial America was heterogeneous, with the colonies scattered across the American coast and the colonists originating from different dialect areas of England. However, communication between the colonies eased thanks to new roads, stagecoaches, weekly newspapers – following the establishment of the first press in Cambridge in 1638-9 (Venezky, 2001) – and postal services. These changes brought the colonies closer together in every aspect of life, not least in language (Algeo, 2001). Consequently, it is clear that these Colonial period developments were the first steps towards a unique and unified American English.

Read (2002), in his lectures first delivered in 1979, lists several milestones in the linguistic progress in America. With respect to the Colonial period, these are occasions when different people, usually British travelers, teachers, and the like, pointed out linguistic changes in the language in America. In 1621 Alexander Gill, a schoolmaster in London, wrote that there are certain words borrowed from Amerindian languages, such as *maiz* and *kanoa*. These two, *maize* and *canoe*, are the earliest recorded Americanisms according to Read, although Bryson (1994: 29) mentions several earlier cases. In 1663 John Josselyn paid attention to the remarkable way the colonials used the word *ordinary*. Its semantic scope had narrowed from the British use 'a place where meals could be obtained' to 'a place primarily for drinking'. The next milestone is found in an account of Pennsylvania by Gabriel Thomas at the end of the 17th century. He mentions a case in which the British and Americans have different terms for the same entity. Writing about plants in Pennsylvania, he found out that a certain plant was called *poke-root* which in England was called *jallop*. In the 18th century the American language was apparently quite distinguishable, because on different occasions it began to be described as barbaric by English writers. In 1735 Francis Moore wrote about the town of Savannah: "it stands upon the flat of a hill, the Bank of the River (which they in barbarous *English* call a *bluff*) is steep and about 45 Foot perpendicular." Benjamin Franklin in his letter to Jared Eliot in 1752 noted that there were different dialects in the colonies. In 1754 Richard Owen Cambridge came up with the idea of a glossary of American expressions and one was produced by Jonathan Boucher who started working on it in 1774 and continued up to his death in 1804 (although his material was not printed until 1832). These single occasions, no matter how trivial they may seem, together draw the timeline of the first stages of American English in the Colonial period.

2.3. After the Revolution

Fisher (2001) describes the situation after the American Revolutionary War (1775-1783) as 'schizophrenic': on one hand, there was hostility towards the English regime, but on the other hand, 'acute nostalgia' for the English heritage. In any case, the Revolution sparked a need to redefine the language of the new nation as 'American English'. At the time, it seemed that British and American English were going to evolve into two separate languages, as Italian and Spanish 1,500 years earlier (Algeo, 2001: 36-37), but this course of events was interrupted by the progress of more efficient telecommunications and transportation.

Noah Webster (1758-1843) was in the vanguard of promoting AmE, and his *American Spelling Book* (1783) provided a standard for American orthography. In the 18th century,

English spelling was largely standardized, but there were still some alternatives, for example *colour*/*color*, *travelling*/*traveling* and *realize*/*realise*. Webster chose the alternatives that were in his opinion "simpler, more historical or analogous", promoting these spellings in his 'Blue-Backed Speller' (as his spelling book was called) and thus established the American spelling standard (Algeo, 2001: 34). Perhaps as a result of the declined overall relations between Britain and America, in many cases the alternative spelling version eventually became the standard in British English, further deepening the gap between the two Englishes. Many of the spelling differences that even today distinguish AmE from BrE originate from Webster's work, although it is not clear whether he was consciously attempting to create a new variant or simply aiming to standardize English spelling in general. Be that as it may, his *American Spelling Book* was used to teach English to many generations of early Americans and provided the basis for American English (Ibid., 2001).

Read's (2002) linguistic milestones after the Declaration of Independence begin with the coinage of the term "Americanism" in 1781 by John Witherspoon to refer to concepts unique to the American speech community. Witherspoon was Scottish and being used to avoiding "Scotticisms", the coinage of "Americanism" was only natural. The next milestones are dictionaries: The first dictionary of Americanisms, *Vocabulary*, by John Pickering was published in 1816 and *An American Dictionary of the English Language* by Webster in 1828. Already in 1800 Webster wrote:

> It is found that a work of this kind [dictionary of the American Language] is absolutely necessary, on account of considerable differences between the American and English language. New circumstances, new modes of life, new laws, new ideas of various kinds give rise to new words, and have already made many material differences between the language of England and America.

This quotation, taken from Read (2002: 17), shows that by the late 18th century, a clearly distinguishable American English had emerged, acknowledged even by Webster, perhaps the most influential authority in early American linguistics. The next milestone was John Russell Bartlett's dictionary of Americanisms in 1848. Whereas Pickering's *Vocabulary* 30 years earlier had been, in Read's words, apologetic, Bartlett's work praised Americanisms and the American way of expression. The last milestone in the 19th century was the founding of the American Dialect Society in 1889 that strengthened the prestige of American English even further. Coming to the 20th century, Read mentions two important milestones that sealed the status of American English as a recognized and prominent variant of English: *The*

American Language by Mencken in 1919 and the *Dictionary of American English* (*DAE*, 1938-44).

The illustration below (Figure 1) places the most important milestones and occasions in the development of American identity and American English on the timeline of the corpus. The figure is partially reproduced from Hendrik De Smet's outline of the CLMET. The details of the illustration are further elucidated in the following section.

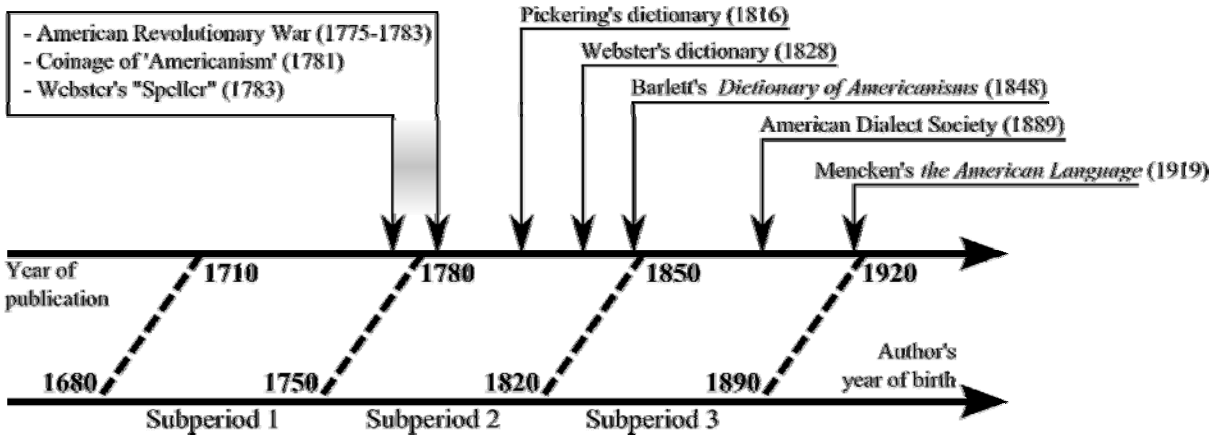


Figure 1*:* American English milestones placed on the timeline of the corpus

The emergence of American English began already when the first English speaking settlers landed on the American continent in the early 17th century. Because language is not a stable entity but an ever evolving one, American English is still changing in the 21st century. The corpus only captures a fragment from the middle of American English development, beginning a hundred years after the arrival of the first immigrants and ending 210 years later, not reaching the present day. But even though American English began to emerge before the time span of the corpus and continues to evolve after it, the 210-year period it covers is perhaps the most essential in the history of American English. In Dixon's (1997) terms, the corpus centralizes on a period of *punctuation*, a time of rapid language change which is often motivated by extralinguistic factors (in this case, the establishment of the United States as an independent nation). However, it also covers the surrounding periods of relative *equilibrium*, the times of gradual and stable linguistic change that generally follow and precede the period of rapid change.

The outline presented above is not a comprehensive image of the linguistic history of American English. It merely offers a short description of the early stages of American

English on its path to the position it has today. However, it should be sufficient for the purposes of this paper and the compilation of CEAL.

3. PRINCIPLES OF COMPILATION

As the aim of this project is to compile an American English corpus that will be compatible with the CLMET, the prerequisite is that it is compiled in accordance with the principles of the CLMET compilation. These principles are presented in De Smet (2005), and are reproduced here with additional points that are unique to this corpus and may therefore differ in some aspects from De Smet's outline.

De Smet introduces four main principles he utilized in compiling the CLMET. First, the material is divided into three 70-year subperiods according to their publication dates. In addition, the ages of the authors are controlled, and their birth dates must match accordingly restricted 70-year periods. These 70-year periods that classify the authors by their birth dates precede the text subperiods by 30 years. Thus, the texts that fall into the first subperiod 1710-1780 have authors born between 1680 and 1750, the texts in the second subperiod 1780-1850 have authors born between 1750 and 1820 and lastly, the authors of the third subperiod (1850-1920) were born between 1820 and 1890. The reason for this precise author and text selection is to maximize linguistic differences between the different chronological subsections. The author birth – publication time relationship is depicted in Figure 2 below, which is essentially a simplified version of Figure 1.

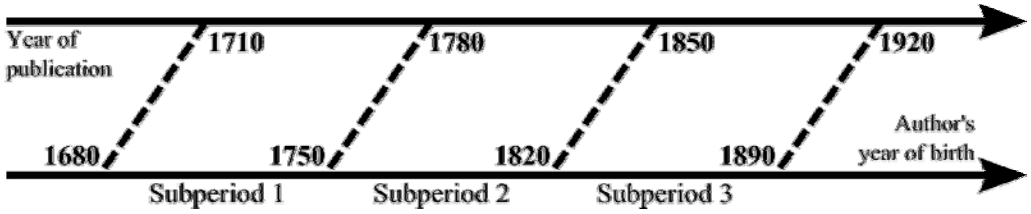


Figure 2: Corpus subperiods, reproduced from De Smet (2005)

De Smet's second principle is that all authors are British and native speakers of English. In the corpus under construction, this principle is taken into account only with the modification that the authors must be American. From this arises the first implementation problem: which authors may be considered "American" especially during the Colonial period of America. This is discussed in greater detail in the following section.

The third principle refers to the amount of text per author. De Smet drew the line at 200,000 words. No apparent reason exists for modifying this principle for the present corpus. De Smet's corpus is designed to be easily expandable by increasing the amount of texts per section. This has resulted in two different versions of the corpus: the standard version, documented in De Smet (2005) and a later, extended version called the CLMETEV, documented on De Smet's website (2010). Bearing this in mind, the size of the original standard version of the CLMET (approximately 10 million words) acts as the first milestone for CEAL, but it will very likely expand even beyond this.

The fourth and last principle concerns the quality of the texts. De Smet remarks that he has attempted to favor non-literary texts, lower registers and woman writers as much as possible. In spite of this, these texts still remain in the minority, making the CLMET biased to literary texts written by higher class male adults. In the pursuit of making the two corpora compatible for reliable comparison, the principle of favoring non-literary text genres, along with the problems prominent in the CLMET, complicate the compilation of CEAL. This issue is discussed in greater detail below.

4. Problems and special cases to consider

The preparation of a diachronic corpus of American English with an aim of making it comparable with a British English counterpart is not without problems. Many of the problems at hand are commonly encountered in corpus design, and are also often referenced in corpus literature. Hunston (2002: 25-32), for instance, distinguishes four aspects that surface in linguistic corpora, affecting their design and purpose: size, content, balance (or representativeness), and permanence. Size is the amount of text included in the corpus. Content includes matters such as text selection criteria, copyright issues and so on. Balance and representativeness addresses the amount or the lack of heterogeneity of the corpus. Permanence deals with the general instability of language: a corpus is not a stable representation of a language but one fixed to a specific spatio-temporal realization of language. Similar conceptual divisions of corpora are also found in other literature (for instance, Biber et al., 1998: 248–249).

The four aspects mentioned above are not entirely discrete but largely intertwined. For example, size is not only reflected in the overall size of the corpus, but in the overall size of texts per author, texts per genre and the like. These selections in turn affect the balance and representativeness of the corpus. Permanence is especially firmly connected to all the other

aspects in the compilation of a diachronic corpus. Texts faithful to the original linguistic form (that is, text that remain linguistically as unrevised as possible), a heterogeneous selection of contents (author-wise, text genre-wise and time-wise) from any given era, and sufficient size for all subperiods of the corpus all contribute to the permanence of the corpus.

The corpus compilation criteria of the CLMET (De Smet 2005, discussed above) lay the foundation for CEAL. Owing to this shared foundation, some compilation problems are shared by the two corpora. For instance, De Smet (2005: 71) points out that material has made its way into the large Internet text repositories such as *Project Gutenberg* and the *Oxford Text Archive* for reasons other than linguistic interest. This has resulted in a somewhat generally biased author selection. In the case of British English, texts written by male writers belonging to the better-off layers of the 18th and 19th centuries are abundant, whereas others are less readily available. This appears to be a tendency expanding beyond British English, as it also affects the selection of American texts.

However, there are also matters that affect solely the compilation of CEAL that have to do with the differences between the divergent linguistic and cultural evolution of Britain and America. These mostly affect the first subperiod of CEAL (1710–1780) especially due to the available text selection: fewer texts exist in the public domain from the Colonial American period than the period following the Revolution, making it more difficult to maintain the period as textually heterogeneous as possible by using an alternative selection of texts. After the Revolution, however, the nation was ever growing, and fiction and prose started to gain more ground in the literary field. Charles Brockden Brown (1771-1810), Washington Irving (1783-1859) and James Fenimore Cooper (1789-1851) were among the trailblazers of this development. Their works, as well as those of their many successors, are well represented in the public domain.

Two major implementation issues, the disproportion of text genres and the problem of defining the American writer, are discussed briefly below.

4.1. Disproportionate text genres

Looking back to the Colonial America and what kind of people emigrated to America (see above), it is no surprise that early prose and fiction is scarce. Many of the first settlers were Puritans, Pilgrims and Quakers with firm religious beliefs whose literary contributions were likewise ecclesiastical: sermons, religious pamphlets, and the like. Strong Christianity prevailed even after the initial waves of settlement, and re-emerged, for instance, in the Great

Awakening of the early 18th century, a wave of evangelism that has been called "America's first mass movement" (Tindall & Shi 1989: 66).

In the first half of the 18th century, one encounters another genre that becomes more and more prominent alongside religious texts: political writing. There were disputes, feuds and wars between the colonies, provoking many writings of the social and political situation in America. These writings also foreshadowed the American Revolution of the late 18th century.

Religious texts are usually written in a specific style (generally archaic or biblical), and consequently, are not the best choice to represent the general American English of the Colonial period. Similar problems are apparent with political writings, although perhaps to a lesser extent. Fortunately there are enough texts available in the public domain to allow some screening. In order to retain as much genre variation as possible and alleviate genre bias, texts that are primarily non-political or non-religious will be favored. In addition to this, political texts will be favored over religious texts (in order to avoid the excessive presence of archaisms).

However, the endeavor to make CEAL maximally compatible with the existing version of the CLMET complicates the problem of text genres even further. As was already noted in the Principles of Compilation section above, the CLMET consists mostly of imaginative prose and other literary texts despite the compilation criteria that stipulate the favoring of non-fiction and a more heterogeneous selection of texts. It is difficult to compensate for this problem in an American context, since prose – as a whole – is scarce in the Colonial period of America. In fact, the favoring of non-literary texts appears to be easier in the more genre-biased early American texts than it is in the respective period of the CLMET. In any case, it is very likely that there will be a partial genre-wise disproportion between the first subparts of the CLMET and CEAL due to the nature of the available texts from the Colonial period.

4.2. Defining the 'American' writer

During the Colonial period (and up until the end of the 19th century) British literature was the major authority in America. Books were imported from England and British authors were read and their style imitated. Similarly to the disproportion of text genres in early American writing, this also leads to possible text selection problems, as early political writers, for instance, were strongly influenced by English models. Even Benjamin Franklin (1706-1790), who is thought to represent the typical American writer of the 18th century, wrote in the style of English writers (Venezky, 2001). It may well be that the early American English spoken in

the colonies never reached the status of "proper" language with sufficient prestige to be the written standard.

Further issues in determining which texts may be considered 'American' have to do with the writers' background. Many prolific early American writers were not only influenced by British writing, but also received education on the other side of the Atlantic. For example, Thomas Paine (1737-1809), one of the most influential revolutionary authors and one of the Founding Fathers, is generally cited as an American writer. However, from a linguistic perspective he was British; he was born, raised and educated in England and did not emigrate to America until the age of 37. It is thus reasonable to assume that his writing style – at least from a purely linguistic perspective – was essentially British.

It is difficult, if not impossible, to locate an objective and straightforward solution to specifying the early American writer, especially as it is unclear whether a clear-cut American English existed during this period (as debate on the subject continues to the present day). In order to at least minimize the interference of British influence, we have decided to favor authors that were born, raised and educated in America in the text selection process as much as possible. Be that as it may, prolific 'non-American' authors such as Paine may find their way into the final corpus not only due to the possible lack of alternative material but also due to their strong influence on the American authorship as a whole.

In the case of the two later subparts of the corpus (1780–1850 and 1850–1920) the specification of a 'genuinely American' author becomes considerably easier. As noted above, by the 19th century American English was already an established variant, with school books such as Webster's "Speller" used in language teaching. Quite surprisingly, the CLMET's first cut-off point between the subparts of the corpus (1780) flows quite naturally into the linguistic history of America. As was already illustrated above, it is situated around the Revolution, and precedes Webster's "Speller" by only three years.

5. Conclusion

In this paper we have illustrated possible problematic issues that arise in the compilation of an American counterpart to the Corpus of Late Modern English Texts. First and foremost of these is the division of American history into two periods: the Colonial period, during which American culture remained largely connected to British culture, and the period following the American Revolution, during which a distinct American culture emerged. It has also been

pointed out that the emergence of American English takes place gradually from the 17th century onwards, covering both these periods.

The division between Colonial and post-Revolutionary America causes problematic issues in implementing the time span and the compilation criteria of the CLMET into an American context. Problems affect especially the period of Colonial America, covered by the first subpart of CEAL. Especially problematic is the question of which authors of this era qualify as 'American' not only by birth but also in language usage.

Another problem affecting the first subpart of CEAL is the availability of heterogeneous text material. Owing in part to the cultural history of America, text selection from this period is partly biased towards ecclesiastical and political writings. In addition to narrowing down text variation, it also makes it difficult to assure maximal comparability with the CLMET, which is partly biased towards narrative fiction. In the second and third subparts of CEAL this problem evens out as more American prose is available.

REFERENCES

Algeo, J. (2001). External History. In J. Algeo (Ed.), *The Cambridge History of the English Language (vol. 6): English in North America* (pp. 1–58). Cambridge: Cambridge University Press.

Biber, D., Conrad S., & Reppen R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Bryson, B. (1994). *Made in America*. London: Secker & Warburg.

Chambers, J. K., & Trudgill, P. (1980). *Dialectology.* Cambridge: Cambridge University Press.

Crystal, D. (2003). *The Cambridge Encyclopedia of the English Language.* 2nd edition. Cambridge: Cambridge University Press.

Davies, M. (2009). The 385+ million word *Corpus of Contemporary American English* (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2). (pp. 159–190).

De Smet, H. (2005). A Corpus of Late Modern English. *ICAME-Journal 29*. (pp. 69–82).

De Smet, H. (2010). The Corpus of Late Modern English Texts (extended version). Retrieved from https://perswww.kuleuven.be/~u0044428/clmetev.htm (accessed: February 24, 2010)

Dixon, R. M. W. (1997). *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

Fischer, D. H. (1989). *Albion's Seed: Four British Folkways in America*. New York: Oxford University Press.

Fisher, J. H. (2001). British And American, Continuity And Divergence. In J. Algeo (Ed.), *The Cambridge History of the English Language (Vol. 6): English in North America* (pp. 59-85). Cambridge: Cambridge University Press.

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Kytö, M. (1994). Towards a corpus of early American English. In M. Kytö, M. Rissanen, & S. Wright (Eds.), *Corpora Across the Centuries* (pp. 33–39). Amsterdam–Atlanta: Rodopi.

Mencken, H. L. (1936). *The American Language.* 4th edition. New York: Alfred A. Knopf.

Read, A. W. (2002). *Milestones in the history of English in America*. Durham, NC: Duke University Press.

Strevens, P. (1972). *British and American English.* London: Macmillan.

Tindall, G. B., & Shi D. E. (1989). *America: a Narrative History.* Brief Second Edition. New York - London: W. W. Norton & Company.

Tottie, G. (2002). *An Introduction to American English.* Oxford: Blackwell.

Venezky, R. L. (2001). Spelling. In J. Algeo (Ed.), *The Cambridge History of the English Language (vol. 6): English in North America* (pp. 340-357). Cambridge: Cambridge University Press.

Tipología semántica del nombre en colocaciones con el verbo *dar*

AMELIA HUZUM

Universidad de Vigo

Resumen
Para entender mejor el mecanismo de las colocaciones léxicas y su capacidad combinatoria es necesario saber qué sustantivos se combinan con qué verbos para expresar un significado dado. Por lo tanto, nos proponemos realizar una tipología semántica no exhaustiva de los nombres (bases colocacionales) que se combinan con el verbo dar dentro de colocaciones léxicas. Los casos obtenidos a partir de la base de datos ADESSE (Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español) serán contrastados con equivalentes rumanos para ejemplificar el carácter imprevisible y arbitrario de este tipo de construcciones.
Palabras clave: base colocacional, colocativo, clases semánticas

Abstract
In order to better understand the mechanism of lexical collocations and their combinatory capacity, it is necessary to know what nouns combine with certain verbs to express a given meaning. Therefore, it is our purpose to carry out a non-exhaustive semantic typology of the nouns (collocational bases) which combine with the verb dar within lexical collocations. The cases obtained from the ADESSE database (Spanish Alternations of Diathesis and Syntactic-Semantic Schemes) will be contrasted with their Romanian counterparts in order to illustrate the unpredictable and arbitrary character of this type of constructions.
Keywords: collocational base, collocative, semantic classes

1. INTRODUCCIÓN

Debido al carácter imprevisible de las colocaciones léxicas es difícil realizar una clasificación semántica de sus bases (nombres). Sin embargo, en el caso de las construcciones verbo-nominales, algunas obras lexicográficas han mostrado que se puede llegar a identificar clases de sustantivos que comparten ciertos rasgos semánticos y que seleccionan el mismo verbo dentro de la colocación[1].

En los estudios dedicados a la clasificación de los nombres que se combinan con el verbo *dar* se han propuesto varios criterios partiendo del significado global de la expresión (Koike, 1993; Herrero Ingelmo, 2002) y sus aportaciones reflejan una vez más la complejidad y la heterogeneidad de la tarea de sistematización semántica.

A continuación, trataré de dar cuenta de la versatilidad combinatoria del verbo *dar* a partir de una lista no exhaustiva de colocaciones de verbo + nombre en español y rumano. Se

[1] Es el caso del diccionario REDES de I. Bosque (2004) que intenta ofrecer clasificaciones representativas de combinaciones frecuentes entre palabras. Sin embargo, el método empleado es distinto del que se va a seguir: en el diccionario es el verbo el que selecciona los nombres para formar colocaciones.

han excluido las combinaciones metafóricas y las construcciones en las que *dar* pueda tener otros usos más que el de colocativo. Para el español, los casos se han recogido de la base de datos ADESSE[2] mientras que para el rumano se ha utilizado la información proporcionada por dos diccionarios monolingües, *Dicţionarul explicativ al limbii române* y *Noul dicţionar universal al limbii române*.

Para la selección y estructuración de los ejemplos incluidos en el trabajo se han tomado en cuenta la semántica del colocativo , las clases de nombres que funcionan como bases dentro de colocaciones y los tipos de situaciones[3] en las que aparecen: acciones (situaciones durativas que requieren un agente animado que pueda realizar de manera intencionada lo que el nombre sugiere), actos (situaciones puntuales que se desarrollan en un momento dado y que requieren una implicación motivada por parte de un agente animado), actividades (situaciones durativas realizadas de manera sistemática e intencionada por un agente animado) y acontecimientos (situaciones puntuales, espontáneas que tienen lugar de manera involuntaria).

Dentro de la información colocacional, en rumano aparecerá en primer lugar la expresión equivalente a la española junto con otras opciones sólo si ésta no representa la expresión más frecuente. En los casos en los que la equivalencia más habitual es la formada con el verbo *a da,* no se registrarán otras soluciones. Se añadirá también la traducción literal de la expresión rumana en los casos en los que las expresiones no son equivalentes, ya que la traducción podría esclarecer su uso.

2. ANÁLISIS DE LOS DATOS

Para facilitar la lectura, se han dividido las colocaciones encontradas en nueve grupos organizados según el significado semántico de los nombres (véase Tabla 1). A continuación se presentarán las características de los constituyentes de las expresiones de cada grupo, las posibilidades de traducción al rumano y las diferencias que se dan entre los dos idiomas.

2.1. Desplazamiento

En este primer grupo se han escogido ejemplos de construcciones cuyos nombres expresan la idea de desplazamiento. Las situaciones descritas son diversas, representando tanto

2. La base de datos ADESSE forma parte del proyecto ALEXSYS llevado a cabo en la Universidad de Vigo. Para una introducción general al proyecto, vid. García-Miguel et al. (2005) y http://adesse.uvigo.es/.
3. Para una descripción detallada de las etiquetas semánticas y tests para su comprobación, vid. Alonso Ramos (2004).

desplazamientos durativos intencionados (acciones – 1a) como desplazamientos puntuales (actos – 1b):

> 1a. Una vez te acompañé a dar una vuelta. [SUR: 033, 04]
> 1b. Leoncio da un paso hacia el árbol, se para, adelanta una oreja y queda a la expectativa […]. [1INF: 034, 03])

En todos los casos el verbo tiene el significado general de "hacer" o "realizar" la situación expresada por el nombre. No es de extrañar, por lo tanto, que en otros idiomas románicos como el rumano, francés o italiano el nombre se combine directamente con el verbo *hacer*: *a face un salt* (rum.), *faire un saut* (fr.), *fare un salto* (it.). La arbitrariedad con la cual el mismo nombre se combina con verbos diferentes en otros idiomas es una prueba más de la imprevisibilidad de las colocaciones léxicas.

2.2. Sonidos

El segundo grupo está representado por todo tipo de situaciones espontáneas, de corta duración, realizadas de manera más o menos intencionada por agentes animados (actos - 2a) o inanimados (acontecimientos - 2b):

> 2a. Ya sabía yo que andábais cerca. (Da un rugido de satisfacción.) [1INF: 010,18]
> 2b. El viejo coche *dio un chasquido*, empezó a echar humo, y se paró.

El verbo *dar* tiene, una vez más, el significado general "hacer" o "realizar". En las expresiones equivalentes rumanas, el verbo ayuda a distinguir entre las situaciones realizadas por agentes animados o inanimados. En el caso de construcciones como *dar un grito*, el rumano prefiere utilizar el verbo *a scoate* ("sacar") para enfatizar la idea de que algo animado emite o exterioriza cierto sonido. En español, se logra el mismo efecto estilístico con la ayuda de verbos como *emitir* o *soltar un grito.* En cambio, para las situaciones involuntarias, el rumano prefiere los verbos plenos (dar un estallido ~ *a crãpa*).

> 3. Juan bajó a la cocina y comió tanto jamón con pan que todos pensaron que iba a dar un estallido.[*Juan a coborât la bucatarie şi s-a îndopat cu atâta şuncã şi pâine cã toţi cei prezenţi au crezut cã o sã crape.]*

2.3. Acciones involuntarias

En el caso del tercer grupo se mantiene la tendencia de utilizar verbos plenos para "realizar" acontecimientos como *dar un respingo:*

4. El frenazo de un coche en la calle me hizo dar un respingo. [LAB: 145, 13]

No existe un modelo predeterminado para estas construcciones y la elección del verbo adecuado depende del contexto.

2.4. Contacto físico

El cuarto grupo está constituido por actos que implican contacto físico como los besos, los abrazos, todo tipo de golpes (5a) y acciones como *dar una paliza* (5b).

5a. Dio un bofetón al chino que tenía más cerca y lo envió rodando debajo de los lavabos. [LAB: 101, 12]
5b. Eso, estoy igual que si me hubieran dado una paliza. [AYE: 060, 23]

Si en los primeros tres grupos el verbo se podía sustituir por "hacer, realizar", en este grupo el verbo mantiene una cierta carga semántica de transferencia. Tanto las muestras de cariño (besos, abrazos) como los golpes (bofetadas, patadas, azotazos, empujones) implican la existencia de un emisor, de un objeto transmitido y un receptor como en el caso de una transferencia habitual.

Salvo unos pocos casos en los que se prefiere un verbo pleno, la equivalencia entre los dos idiomas es total. También cabe destacar que en rumano, para construir expresiones que designan golpes, es muy habitual que *dar* se sustituya por *a trage* ("tirar"), que se combina con nombres como *toque* (*a trage o loviturã uşoarã*), *golpe* (*a trage o loviturã*), *bofetada* (*a trage o palmã*), *patada* (*a trage un şut*), etc.

El grupo de los "golpes" ocupa un lugar importante dentro de las clases de nombres que se combinan con el verbo *dar*. Es más, el proceso de composición de nuevos nombres que designan golpes es muy productivo en español, sobre todo si se trata de nombres acabados en –azo, -ón, -ada. No pasa lo mismo con el rumano que recurre a un proceso de paráfrasis. Por ejemplo, cuando se intentan buscar equivalentes para nombres como *patada, puñalada, puntapié, codazo,* se prefiere utilizar el término general de golpe + el objeto con el que se realiza la acción: *loviturã cu piciorul* (´golpe con el pie´), *loviturã de cuţit* (´golpe de cuchillo´), *loviturã cu vârful piciorului* (´golpe con la punta del pie´)*, loviturã cu cotul* (´golpe con el codo´). También es frecuente suprimir el nombre *golpe* y formar directamente la construcción con la ayuda del objeto que se usa para realizar el golpe: *a da un picior* (´dar un pie´), *a da cu cuţitul* (´dar con el cuchillo´), *a da cu vârful piciorului* (´dar con la punta del pie´), *a da cu cotul* (´dar con el codo´).

Un caso especial lo representa el nombre *paliza* que, aunque pertenece a la clase de "golpes", no comparte las mismas características con los otros nombres, ya que representa una situación durativa (una acción), no un acto de corta intensidad como los demás.

2.5. Cuidado corporal

La quinta categoría agrupa acciones relacionadas con el cuidado corporal:

> 6. Pues deberías darte un baño y meterte en la cama. [OCH: 042, 08])

Como en el caso de las acciones que designaban actividades placenteras (*dar un paseo, dar un garbeo, dar una caminata,* etc.), al verbo *dar* elegido por la base colocacional le corresponde en rumano el verbo *hacer: a face duş* (´hacer ducha´), *a face (o) baie* (´hacer [un] baño´), etc.

2.6. Mensajes

El sexto grupo está dividido en tres tipos de mensajes: permisos (7a), instrucciones o avisos (7b) y sugerencias (7c).

> 7a. Me dan permiso de tomar café. [DIE: 152, 24]
>
> 7b. Mientras le daba instrucciones me iba dirigiendo a la puerta […] [LAB: 145, 22]
>
> 7c. Los dirigentes latinoamercianos le buscan para que les asesore y les dé consejos […] [1VOZ: 04, 3, 1, 023]

En los tres casos se trata de situaciones puntuales realizadas por un agente de manera intencionada (actos). El verbo mantiene una cierta carga semántica de transferencia. Al comparar las construcciones españolas con los equivalentes rumanos se observa una identidad total en cuanto al uso del verbo *dar* como colocativo.

2.7. Eventos públicos

En este séptimo caso se reunieron nombres que expresan la idea de actividad durativa que se realiza de manera intencionada y sistemática como clases, conferencias, charlas, etc.

> 8. Tendría que esperar a que tía Delia se marchara a dar sus clases de solfeo y piano. [SUR: 043, 04]

Los eventos públicos de este tipo se suelen *dar* en español y *sostener* en rumano. Se admite también el verbo *dar* pero su uso no es el más habitual. Es más frecuente que se

"sostenga una clase/una lección/ un curso/un discurso/ una conferencia/una charla/ un mitin" antes de que se dé. A veces, cuando se trata de nombres de actividades docentes, se suele emplear también el verbo *a preda* (enseñar): *a preda un obiect* (´enseñar una asignatura´), *a preda o materie* (´enseñar una materia´), *a preda un curs* (´enseñar un curso´), etc.

2.8. Acciones rápidas y/o a la ligera

El octavo grupo contiene nombres que representan acciones rápidas y/o a la ligera, es decir, actos.

9. Basta con dar una ojeada a la historia de las grandes conquistas *[...]* [TIE: 088, 30]

Es un grupo muy productivo en español en plena ascensión. Esta categoría representa otro caso de equivalencia total entre las dos lenguas románicas.

Como en el caso de los varios tipos de golpes, el rumano no presenta tanta variedad terminológica. Por lo tanto, se suele utilizar el objeto con el que se realiza la situación como término del acto propiamente dicho: *a da cu fierul* (´dar con la plancha´), *a da cu mãtura* (´dar con la escoba´), *a da cu sãpunul* (´dar con el jabón´), *a da cu peria* (´dar con la brocha´), etc.

2.9. Formas de trato social

Se han agrupado en el último grupo formas de trato social, actos indispensables en nuestra vida diaria. En muchos casos, se mantiene la equivalencia entre el verbo *dar* y su correspondiente rumano: *a da bunã ziua* (´dar buen día´), *a da bunã seara* (´dar buena tarde´), *a da noapte bunã* (´dar noche buena´).

3. CONCLUSIONES

A modo de conclusión, una primera aproximación a las colocaciones españolas pone de manifiesto una serie de fenómenos con respecto a la interpretación de *dar* como colocativo.

Por un lado, en la gran mayoría de los casos, el verbo conserva parte de su significado básico de transferencia (*dar un beso/una clase/un saludo*), mientras que el resto se puede traducir por el significado general de "hacer" o "realizar" una cosa especificada por el nombre (*dar una voltereta/un suspiro/un estirón*).

En cuanto a la tipología de situaciones ilustradas, cabe destacar el hecho de que predominan las situaciones dinámicas volitivas. De hecho, se registró un único caso en el que la situación tiene lugar de manera involuntaria (*dar un respingo/un chasquido*$_2$*, un estirón*).

A pesar del gran número de construcciones posibles con el verbo *dar* + nombre, no es totalmente imposible tratar de organizar las clases semánticas de los nombres que forman la base. Claro está que todo intento de sistematización está sujeto a muchas excepciones debido a la arbitrariedad combinatoria de las colocaciones. Aun así, se pueden distinguir grupos compactos de nombres que se combinan con el verbo *dar.* Como se puede ver en la Tabla 1, el grupo de "golpes" incluido en la clase de "contacto físico" ocupa un lugar muy importante entre todas las clases de nombres, tal vez por el gran número de nombres que lo conforma y por la productividad que demuestran los nombres acabados en –azo, -ón, -ada. Lo mismo pasa con los nombres que designan actividades placenteras de corta duración como los paseos, las caminatas, los garbeos, etc.

Si se evalúan comparativamente los dos listados de construcciones, se observa una clara similitud entre los dos idiomas. De las nueve clases semánticas, sólo tres aparecen en rumano bajo una construcción distinta a la de verbo *dar* + nombre. En muchos casos, la identidad es total. Esto sucede en los grupos de contacto físico, mensajes, eventos públicos, acciones rápidas y/o a la ligera y formas de trato. El uso puede ser ligeramente distinto presentando restricciones de orden sintáctico o de selección argumental, pero el significado de la expresión no se altera.

A veces, se da el caso de que las construcciones con el verbo *dar* no son las más utilizadas por los hablantes nativos de rumano: es más frecuente que se **tire una bofetada* o que se **sostenga una clase* antes de que se *dé*.

Existen también casos en los que la equivalencia no es posible. Se trata, sobre todo, de colocaciones en las que el significado semántico del verbo se puede parafrasear por "hacer", "realizar". Quizás sea esto una de las razones por las cuales en otros idiomas se prefieren construir las expresiones con la ayuda del verbo *hacer*. Tomemos el caso de *dar un paseo* que equivale en otras lenguas a *take a walk* (ingl. ´tomar un paseo´), *faire une promenade* (fr. ´hacer un paseo´), *fare una passeggiata* (it. ´hacer un paseo´), *a face o plimbare* (rum. ´hacer un paseo´).

En el caso del rumano, el uso del verbo *a face* (hacer) en construcciones como *a face o plimbare* (´hacer un paseo´) se puede deber, según el análisis sub-léxico (Rădulescu, 2009: 495), a la información semántica que contiene el verbo. En este sentido, el nombre en español parece contener información sobre una posible meta, lo que hace posible su ocurrencia con el verbo *dar* que también implica una cierta idea de trayecto. En rumano, sin embargo, el *paseo* carece de alusiones a la existencia de una meta y se entiende sólo como una acción. Por lo tanto, el nombre elige combinarse con un verbo que implica la idea de "realización".

Hay que añadir que la lista no es más que un esbozo del gran número de posibilidades combinatorias del verbo *dar.* Para poder llegar a una conclusión acerca de las equivalencias y las diferencias que se registran entre los dos idiomas, hay que analizar los datos también a partir de las colocaciones rumanas.

APPENDIX

Tabla 1. Clases semánticas del nombre como base colocacional

CLASE NOMBRE	EJEMPLO	EQUIVALENTE	TIPO SITUACIÓN	SIGNIFICADO COLOCATIVO
1. Desplazamiento de duración				
❖ indeterminada	DAR UN/A ➢ paseo ➢ caminata ➢ vuelta ➢ rodeo DAR UN/A ➢ garbeo	A FACE UN/O (hacer) ➢ plimbare ➢ excursie pe jos ➢ tur ➢ ocol A DA UN/O ➢ raitã	acción	"hacer, realizar"
❖ puntual	DAR UN/A ➢ curva ➢ salto ➢ paso ➢ zancada ➢ brinco ➢ voltereta ➢ pirueta DAR UN/A ➢ giro ➢ vuelco	A FACE UN/O (hacer) ➢ curba ➢ salt ➢ pas ➢ pas mare ➢ sãriturã ➢ tumbã ➢ piruetã VERBO PLENO ➢ a se întoarce (girarse) ➢ a se rãsturna(volcarse)	acto	"hacer, realizar"
2. Sonidos				
❖ voluntarios	DAR UN/A ➢ voz ➢ grito/chillido/alarido ➢ pitido/silbido ➢ gemido ➢ suspiro ➢ aullido ➢ rugido	A SCOATE UN/O (sacar) ➢ strigãt ➢ ţipãt ➢ fluierat/ a fluiera (silbar) ➢ geamãt ➢ suspin ➢ urlet ➢ rãget	acto	"hacer, realizar"

[4] Entendido como "ruido especial con la lengua, aplicándola al paladar y separándola bruscamente Ô chasquear" (DUE).

	➢ berrido ➢ ladrido ➢ gruñido DAR UN/A ➢ soplido ➢ resoplido ➢ chasquido[4]	➢ muget ➢ lãtrat (?)/ a lãtra (ladrar) ➢ grohãit (?)/ a grohãi (gruñir) VERBO PLENO ➢ a sufla (soplar) ➢ a rãsufla (resoplar) ➢ a plescãi (chascar)		
❖ involuntarios	DAR UN/A ➢ estallido ➢ chasquido[5]2	VERBO PLENO ➢ a crãpa (estallar) ➢ a trosni (chascar)	aconteci miento	"hacer, realizar"
3. Acciones involuntarias				
	DAR UN/A ➢ respingo ➢ estirón	PARÁFRASIS ➢ a avea o tresãrire (´tener un respingo´) ➢ a creşte repede (´crecer rápido´)	aconteci miento	"hacer, realizar"
4. Contacto físico				
	DAR UN/A ➢ beso ➢ abrazo ➢ toque DAR UN/A ➢ golpe/ azote/manotazo/torta zo/guantazo ➢ bofetada ➢ patada ➢ puntapié ➢ codazo ➢ puñetazo ➢ rodillazo	A DA UN/O ➢ sãrut ➢ îmbrãţişare ➢ o lovitură uşoară A DA, A TRAGE UN/O (dar, tirar) ➢ lovitură ➢ palmã ➢ şut ➢ picior/ (o lovitură) cu piciorul ("pié/ (un golpe) con el pié") ➢ cot/ (o lovitură) cu cotul ("codo/ (un golpe) con el codo") ➢ un pumn/ (o lovitură) cu pumnul ("puño/ (un golpe) con el puño") ➢ un genunchi/ (o lovitură) cu genunchiul ("rodilla/ (un golpe) con la rodilla")	acto	transferencia ("hacer pasar")

[5] Entendido como "ruido seco y repentino que se produce al resquebrajarse o romperse algo, esp. la madera" (CLAVE)

	➢ cornada ➢ puñalada/navajazo ➢ paliza	➢ corn/ (o lovitură) cu cornul (“cuerno/ (un golpe) con el cuerno”) ➢ a da/ a aplica o lovitură de cuţit (“dar/ asestar un golpe de navaja”) ➢ (o mamă de) bătaie	acción	
	DAR UN/A ➢ lamida ➢ mordisco ➢ empujón	VERBO PLENO ➢ a linge (lamer) ➢ a muşca (morder) ➢ a împinge	acto	
	DAR UN/A ➢ apretón	PARÁFRASIS ➢ a strânge în braţe (“apretar entre brazos”)		
5. Cuidado corporal				
	DAR(SE) UN/A ➢ ducha ➢ baño ➢ chapuzón/remojón	A FACE UN/O (hacer) ➢ duş ➢ baie ➢ baie scurtă	acción	“hacer, realizar”
6. Mensajes				
❖ permiso	DAR UN/A/- ➢ permiso ➢ autorización ➢ aprobación ➢ (su) consentimiento ➢ licencia	A DA UN/O ➢ permisiunea ➢ autorizaţie ➢ aprobare ➢ a(-şi) da consimţământul ➢ licenţă	acto	transferencia (“hacer pasar”)
❖ instrucción o aviso	DAR UN/A/- ➢ orden ➢ instrucción ➢ directriz ➢ aviso ➢ veredicto ➢ consigna	A DA UN/O ➢ ordin ➢ instrucţiune ➢ directivă ➢ aviz ➢ verdict ➢ dispoziţie		
❖ sugerencia	DAR UN/A/- ➢ asesoramiento ➢ consejo ➢ sugerencia	A DA UN/O/- ➢ consultanţă ➢ sfat ➢ a da/ a face o sugestie (“dar/hacer una sugerencia)		

7. Eventos públicos				
	DAR UN/A ➢ clase ➢ lección ➢ curso ➢ discurso ➢ rueda de prensa	A DA/ A (SUS)ŢINE (dar/sostener) ➢ orã ➢ lecţie ➢ curs ➢ discurs ➢ conferinţã de presã	actividad	transferencia ("hacer pasar")
	DAR UN/A ➢ conferencia ➢ charla ➢ mitin	A (SUS)ŢINE (sostener) ➢ conferinţã ➢ cuvântare ➢ miting		
	DAR UN/A ➢ asignatura ➢ materia	A PREDA (enseñar) ➢ obiect ➢ materia		
8. Acciones rápidas y/o a la ligera				
	DAR UN/A ➢ planchazo ➢ barrido ➢ jabonad ➢ brochazo ➢ pincelada DAR UN/A ➢ ojeada	A DA CU + ART DEF (dar con) ➢ fierul ➢ mãtura ➢ sãpunul ➢ peria ➢ pensula ECHAR UN/O ➢ privire	acto	"hacer, realizar"
9. Formas de trato social				
	DAR ➢ los buenos días ➢ las buenas tardes ➢ las buenas noches	A DA – ➢ a da bunã ziua ("dar buen día") ➢ a da bunã seara ("dar buena tarde") ➢ a da/a spune noapte bunã ("dar/decir noche buena")	acto	transferencia ("hacer pasar")

	DAR ➢ las gracias ➢ la bienvenida ➢ el pésame	OTROS COLOCATIVOS ➢ a spune mulţumesc (“decir gracias”) ➢ a ura bun venit (“desear bienvenida”) ➢ a exprima/a spune condoleanţe (“expresar/decir pésame)		
	DAR ➢ la enhorabuena ➢ saludo	VERBO PLENO ➢ a felicita (felicitar) ➢ a saluta (saludar)		

REFERENCIAS BIBLIOGRÁFICAS

Academia Românã, Institutul de Lingvisticã “Iorgu Iordan” (1998). *DEX: Dicţionarul explicativ al limbii române* (2ª ed.). Bucureşti: Univers Enciclopedic.

ADESSE. *Base de datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español)*. Disponible en http://adesse.uvigo.es/

Alonso Ramos, M. (2004). *Las construcciones con verbo de apoyo*. Madrid: Visor.

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

Maldonado, C. (Dir.) (1996). *Clave. Diccionario de uso del español.* Madrid: SM.

García-Miguel, J. M., Costas, L. y Martínez, S. (2005). Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. En G. Wotjak y J. Cuartero Otal (Eds.), *Entre semántica léxica, teoría del léxico y sintaxis* (pp. 373-384). Frankfurt am Main: Peter Lang.

Herrero, J. L. (2002). Los verbos soportes: el verbo *dar* en español. En A. Veiga, M. González Pereira, M. Souto Gómez (Eds.), *Léxico y gramática*. Lugo: Tris Tram.

Koike, K. (1993). *Dar* como verbo soporte. En *Actas del III Congreso de Hispanistas de Asia* (pp. 216-225).

Moliner, M. (2001). *Diccionario de uso del español.* Segunda edición en CD-ROM, versión 2.0. Madrid: Gredos.

Oprea, I., Pamfil, C. G., Radu, R., Zastroiu, V. (2007). *Noul dicţionar universal al limbii române.* Bucureşti-Chişinău: Litera Internacional.

Rădulescu, R. A. (2009). Sobre cuánto puede *dar de sí* el verbo *dar* en fraseologismos españoles y rumanos. En E. de Miguel, S. U. Sánchez Jiménez, A. Serradilla Castaño, R. A. Radulescu, O. Batiukova (Eds.), *Fronteras de un diccionario. Las palabras en* movimiento. San Millán de la Cogolla: Cilengua.

Alignment of un-annotated parallel texts

GALINA E. KEDROVA
SERGEY B. POTEMKIN
Moscow State University

Abstract
An alignment algorithm for aligning bilingual, linguistically un-annotated parallel corpora is presented. It is able to align at sentence level, using large bilingual lexicon and heuristic cues, along with linguistics-based rules. The system currently aligns Russian and English texts, requires no previous mark-up or other manual pre-processing of texts. Russian lemmas are obtained from the grammar dictionary. The flexible nature of the system allows experiments with a variety of fiction or scientific and juridical texts to find solutions to alignment problems like the correct alignment of one-to many sentences and omission of sentence, or how to align despite syntactic differences between two languages. First performance tests are promising, and we are going to develop further word and multiword alignment technique.
Keywords: Sentence alignment, unique words, dynamic programming, Russian classics

Resumen
Se presenta un algoritmo para la alineación de los cuerpos paralelos bilingües, lingüísticamente no anotados. El algoritmo ejecuta la alineación en el nivel de oraciones mediante la utilización de un gran léxico bilingüe y de las reglas heurísticas, así como de reglas gramaticales. Actualmente el sistema efectúa la alineación de textos ingleses y rusos, sin exigir un trazado preliminar o algún otro tratamiento manual de los textos. La lematización de las palabras rusas se realiza con la utilización de un diccionario gramatical. El carácter flexible del sistema permite realizar experimentos con diversos textos literarios, científicos y jurídicos, para hallar una solución a problemas tales como la comparación correcta entre una y varias oraciones, la omisión de oraciones en la traducción, o de cómo alinear a pesar de las diferencias sintácticas entre dos lenguas. Las apreciaciones preliminares son esperanzadoras y suponen un desarrollo ulterior de la técnica de alineación en el nivel de palabras y combinaciones de palabras.
Palabras clave: alineación de oraciones, palabras únicas, programación dinámica, obras clásicas rusas

1. INTRODUCTION

The scope of the paper is the problem of automatic processing of un-annotated parallel corpora. One of the most important and first to be solved task is the text alignment at the level of sentences. Sentences of parallel texts very often could not be mapped one-to-one, a sentence of the source text may correspond to several sentences of translation and vice versa, some sentences and even paragraphs may be omitted in translated text, the boundaries of the sentences may not be the same, etc. Application of the alignment method using lexical information at the level of sentences is discussed. The method is based on matching rare words in both texts. Such matched words are the points of reference, between which the

alignment is performed using dynamic programming. Experiments on alignment of Russian-English parallel texts show that these methods provide a high degree of accuracy even for the fiction texts.

Alignment of parallel texts, i.e., the automatic comparison of sentences or words in one text to their equivalents in the translation, is a very important step in pre-processing for many applications, including, without limitation, machine translation (Brown et al., 1993), information retrieval in different languages, compiling dictionaries (Smadja et al., 1996) and data acquisition for natural language processing (Kuhn, 2004). Very often there is no one-to-one correlation between items of the fiction texts and their translations because translation was performed within different cultural and historical environment. The most widely used statistical techniques that do not require advanced vocabulary base and can be used for rare languages, often give misleading results of alignment, requiring a subsequent costly manual inspection and correction. The bilingual dictionaries are used mainly for specialized texts (English and French records of the Canadian Parliament, the legal texts of EU, program specification, etc.).

The paper is structured as follows: A brief overview of standard approaches to alignment and discussion of their advantages and disadvantages (Section 1). Next, we describe our approach to alignment and key assumptions made during its development, the general scheme of the alignment algorithm and some features of its implementation (Section 2). Testing data and strategy for the results evaluation (Section 3). Finally, we sum up and give the perspectives for further work (Section 4).

2. RELATED RESEARCHES

2.1. Alignment of sentences.

Methods of alignment of the sentences are, roughly speaking, divided into three categories: - Based on the length (Gale and Church, 1993), assuming that the length of the sentences in the source text and in its translation are quasi equal. These methods are very sensitive to sentence omissions because a single lacuna may lead to improper subsequent alignment to the end of the text.

- Based on bilingual lexical information, such as obtained from corpora (Kay and Roescheisen, 1993; Fung and Church, 1994). This strategy is applicable only if there is a sufficiently large bilingual dictionary for a particular pair of languages (Gelbukh et al, 2006).

- Algorithms involving reference tags align the sentence on the basis of information contained in the annotated corpora or spelling similarity (Simard et al, 1992).
- Algorithms involving the annotated corpus also require a pre-prepared resource of aligned parallel texts the availability of which is often even less than the availability of bilingual dictionaries. Also hybrid methods (Schrader, 2006) that combine these standard approaches are used.

2.2. Alignment of the words.

Alignment at the lower level is usually performed using statistical models for machine translation (Brown et al, 1993; Vogel et al, 1999) where any word of the target language is considered as a possible translation for each word of the source language. The probability of some word of the target language being the translation of the source word depends on the frequency with which both words occur in the same or the nearest positions in the parallel texts. For each pair of words in the source and target text the number of segments (a) containing both words, (b) containing the word of the source language but not the word of the target language (c) containing the word of the target language, but not the word of the source language and (d) neither words - is recorded (Ribeiro et al, 2000). Found in such a way the most probable pairs are accepted as translation equivalents. This approach has several disadvantages associated with a large number of rare words in any text, different word order in different languages, and the presence of word combinations. Approximately half of the corpus dictionary consists of the so-called rare events, occurring in the text with a frequency less than 10. Rare events, obviously, do not give enough information for statistical analysis. On the other hand, from five to ten percent of the corpus dictionary consists of highly frequent words, i.e. words with frequencies of 100 or higher. These items appear frequently enough to perform statistical analysis, but because they occur in almost any position in text, they match anything, if the decision of the alignment is based solely on statistics. The N:M, mapping of words i.e. the alignment, in which phrases of N and M words are matched, was noted as a challenge for statistical alignment (Tiedemann, 1999) and again the methods based on the statistics does not work well for rare expressions.

To summarize, we note that alignment at the sentence level and at a lower level (words, phrases), needs to be improved: the existing models hardly cope with aligning texts with missing or mismatched sentences, rare words or phrases within the sentences and syntactic differences between the source and the target languages. However, if the alignment at the

level of words is complicated by different word order, at the sentence level, as a rule, the order of the sentences in the source and the target text is the same, allowing application of dynamic programming.

3. PROPOSED METHOD OF SENTENCES ALIGNMENT

The main problem in the automatic alignment at the level of sentences is the appearance of false pairs of sentences, which are not the translation equivalents. While developing the proposed method, we tried to minimize such effects taking into account the assumption that (a) order of the sentences in Russian and English texts usually is similar (b) there is no significant (more than 500 words) lacuna in the parallel texts, (c) the aligned texts are not too large - a story or a chapter of a novel, about 10 KWords. In contrast to statistical methods and methods based on the degree of proximity, we consider only the low-frequency words, namely, the words occurring only once in each text (*hapax legomena*). Initially, for each such (Russian) word we search in the dictionary a translation equivalent which also occurs only once in the (English) text. If there are a number of such equivalent words, they are excluded from consideration. Further, if the alignment of the sentences defined by hapax legomena violates the order of the sentences in the text, these words are also excluded. As a result we obtain a set of unique pairs of equivalents in two texts. These pairs form the primary structure of reference points or anchors, connecting some sentences of texts. At this stage we do not state the equivalence of the found pairs of sentences, but only say that such pairs of sentences have nonempty intersection. Then the source and the target text are divided into segments, bounded by the found pairs of sentences. These segments are considered as new parallel texts, and searching of reference points is repeated. The iterations continue while the new anchors appear. In practice, the number of iterations does not exceed 6.

After determining the reference points the critical path between them is defined. For each word of the Russian segment all translation equivalents in the relevant English segment are defined. The number of equivalents is calculated for each pair of sentences. This number is considered as the similarity measure between two sentences. All similarity measures are stored as the elements of the adjacency matrix. Then the search for the critical path was performed by standard methods of dynamic programming (Viterbi search).

4. EXPERIMENTAL ALGORITHM EVALUATION

We used several texts of Russian classics (Gogol, Dostoevsky, Chekhov) and its translation into English as the source and the target parts of parallel texts. Fig.1 illustrates the results of an experiment with the Chekhov's story "Anna on the neck" and its translation into English. Texts were not pre-processed. The boundaries of the sentence are determined by the point, the exclamation, the question mark, the ellipsis. The abbreviations such as (Mrs. Mr. Ms. Prof. Dr. Gen. Rep. Sen. St. etc. i.e.; e.g. and others) are recognized, and the point in this case is not considered as the end of the sentence. With this definition of sentence boundaries 223 sentences were allocated in the Russian text and 239 sentences in the English version. Lemmatizing for the Russian version is made with the help of a dictionary of lemma, (Krylov, Starostin 2003). As a result, the algorithm allocated 182 pairs of sentences (78% of texts). 165 sentences (90.5%) are the complete and accurate translation, 16 sentences (9%) are part of the translation of the original (or vice versa), and 1 sentence (0.5%) matched faulty. Similar statistics was observed for other texts ("The Overcoat" by Gogol, chapters of the novel "Crime and Punishment", Dostoevsky).

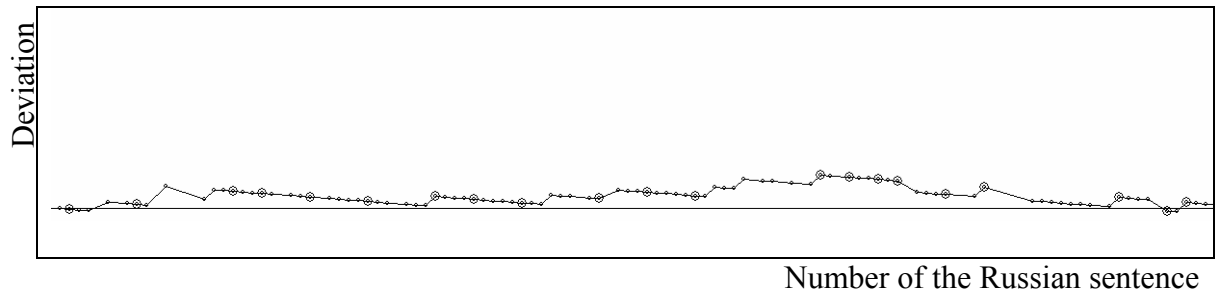


Figure 1: Alignment diagram

The points of the critical path are depicted. The reference points are marked by circles. The horizontal line represents the mean = number of English sentences / number of Russian sentences. Y-axis shows the deviation of an English sentence from the average.

Sometimes the fault matching of part of the sentence with the whole sentence is caused by different punctuation in Russian and English text. Such faults are fixed at the next stage of the algorithm. Most often it happens when 2 consecutive sentences (S_i, S_{i+1}) matches 2 non-sequential sentences (T_{j-1}, T_{j+1}) of the target text

It is assumed that sentence T_j corresponds to the part of S_i or S_{i+1}. To determine which of the sentences S_i, or S_{i+1} contains translation T_j we use the procedure similar to the initial

marking, i.e., unique words should be found in S_i or S_{i+1} and T_j, and translation is determined by the match, then either T_{j-1},T_j merges, or T_j,T_{j+1} merges. After several iterations of such merge throughout the text only about 3% of sentences remains unmatched. For these unmatched sentences we use less rigorous, quantitative methods such as assessment of similarity.

With regard to the erroneous matching like this:

> *Нужно назначить вам премию за красоту ...(*You should be awarded for your beauty…) <=> You must help us ....

some filter, also based on the degree of similarity is applicable (the filter should be used only to discard matching, not for making a decision about the matching).

5. Conclusion

The problem of parallel texts sentences alignment was discussed. It is noted that the one-to-one correspondence is violated in the translated text, i.e. one sentence may correspond to a number of translated sentences, some sentences are lost, etc. The method of automatic alignment at the level of sentences is based on searching *hapax legomena* in the source text and their unique translation in the target text. We describe one possible algorithm that implements this idea. The results of the of Russian-English texts alignment at the level of sentences showed the accuracy of more than 90%, with coverage of 78% of the text. Further improvement of the method allows matching a source sentence with 2 target sentences, or vice versa. In this case text coverage and accuracy of alignment reaches 97%. In the future we plan to use the results of alignment for allocation of equivalent words, phrases and fragments in parallel sentences.

References

Brown, Peter F. Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. (1993). The mathematics of machine translation: Parameter estimation. *Computational Linguistics, 19(2):* 263–311.

Fung, Pascale and Kathleen McKeown. (1994). Aligning noisy parallel corpora across language groups: word pair feature matching by dynamic time warping. *In*

Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-94), (pp. 81–88). Columbia, Maryland, USA.

Gale, William A. and Kenneth W. Church. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics, 19(1):* 75–102.

Kay, Martin and Martin Roescheisen. (1993). Texttranslation alignment. *Computational Linguistics, 19(1):* 121–142.

Kuhn, Jonas. (2004). Exploiting parallel corpora for monolingual grammar induction – a pilot study. In *Workshop proceedings of the 4th International Conference on Language Resources and Evaluation (LREC). Pp* 54–57. Lisbon, Portugal. At *LREC Workshop: The Amazing Utility of Parallel and Comparable Corpora.*

Ribeiro, António, Gabriel Lopes, and João Mexia (2000). A Self-Learning Method of Parallel Texts Alignment J.S. White (Ed.*): AMTA 2000, LNAI 1934,* (pp. 30−39). Springer-Verlag, Berlin, Heidelberg.

Simard, Michel, G. F. Foster, and P. Isabelle. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International conference on theoretical and methodological issues in Machine translation,* (pp. 67–81). Montreal, Canada.

Smadja, Frank, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics, 22(1):* 1–38.

Tiedemann, Joerg. (1999). Word alignment - step by step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics,* (pp. 216–227). Trondheim, Norway.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. (1999). HMM-based word alignment in statistical translation. In *Proceedings of the International Conference on Computational Linguistics,* (pp. 836–841). Copenhagen, Denmark.

А.Ф. Гельбух, Г.О. Сидоров, А. Вера-Феликс (2006). Словари в задачах автоматической обработки пар переводных текстов *Труды межд. конференции Диалог-2006, М., стр. 110-114. (in Russian)*

Крылов С.А. Старостин С.А. Актуальные задачи морфологического анализа и синтеза в интегрированной информационной среде STARLING *Труды международной конференции Диалог'2003, М.2003 (in Russian)*

Spelling-to-sound examples from an SMS corpus

ÚRSULA KIRSTEN TORRADO

Universidad de Vigo

Abstract

Undoubtedly, SMS communication has progressively widespread as a social phenomenon owing to its effectiveness and efficiency and because it is "a quick, cheap and easy to use" way of communication (Grinter & Eldridge, 2001:219). There is undeniably a number of adaptations of spelling when texting, such as contractions, G clippings and other types of clippings, acronyms, initialisms, letter/number homophones, "misspellings" and typos, non-conventional spellings, accent stylisation, smiley faces, lack of function words, lack of punctuation or over-punctuation (adapted from Thurlow, 2003a). Several of those imply a relationship between English pronunciation and spelling in SMS language. Nevertheless, texting is not the cause of bad spelling and it can improve the user's literacy (Crystal, 2008a, 2009). This paper outlines the features that point out sound-to-spelling relationships, providing examples from the SMS corpus at Netting-it.com, which contains 202 text messages, and explaining the adaptations people do when texting.

Keywords: text-messaging, SMS, corpus, new communication technologies, sound-to spelling

Resumen

Es indudable que los mensajes de texto se han extendido como fenómeno social debido a su eficacia y eficiencia y porque es una forma "rápida, barata y fácil" de comunicarse (Grinter & Eldridge, 2001:219). Es innegable que hay un número de adaptaciones ortográficas cuando se mandan SMS como: contracciones, omisión de la G, otras omisiones, acrónimos, inicialismos, homófonos de letras y números, faltas ortográficas y erratas, imitación de acentos, emoticonos, supresión de palabras gramaticales, supresión de signos de puntuación y exceso de puntuación (adaptación de Thurlow, 2003a). Algunas de estas características establecen una relación entre la pronunciación inglesa y la ortografía aunque mandar SMS no es la causa de una mala ortografía y puede mejorar la alfabetización de sus usuarios (Crystal, 2008a, 2009). Este artículo describe las características que establecen una relación entre sonido y ortografía y ofrece ejemplos de un corpus de 202 mensajes explicando las adaptaciones que los hablantes hacen.

Palabras clave: Mensajes de texto, corpus, comunicación a través de las nuevas tecnologías, relación entre pronunciación y ortografía

1. INTRODUCTION

It is a well known fact that we live in a new technology era which has an influence on the way people communicate with each other. Nowadays communication can be established by means of telephones, mobile phones, e-mails, chats and instant messages among others (see Baron, 1998; Crystal, 2001; Thurlow & McKay, 2003b). Undoubtedly, text-messaging or short-messaging services (hereafter, SMS) communication has progressively widespread as a social phenomenon owing to its effectiveness and efficiency and because it is "a quick, cheap and easy to use" way of communication (Grinter & Eldridge, 2001:219). According to the Mobile Data Association (hereinafter, MDA), UK sends 11 million text messages an hour. The report from the MDA covers text messages (SMS) and Picture and Video Messaging

(MMS) and states that there is a growth of messages with a daily average of 265 million text messages and 1.6 million picture messages by 2009. In 2009 a total of 96.8 billion text messages and over 600 million picture messages were sent throughout the whole year. With these data the statement "young and free but tied to the mobile" (Bryden-Brown, 2001: The Australian Newspaper) can be taken as true. It is undeniable that nowadays almost everybody has a mobile phone and it seems that teenagers prefer to let their fingers do the talking rather than to engage in a face-to-face conversation (Kunda, 2005).

Undoubtedly, SMS communication has progressively grown as a social phenomenon owing to its effectiveness and efficiency and because it is a fast and cheap way of communication. It may be thought that the 12-key keypad, the pocket-size screen and the small amount of 160 characters, including letters, numbers or symbols from the Latin alphabet, has caused a number of adaptations of spelling when texting. However, Thurlow's research (2003a) proves that the length of text messages is shorter than the one allowed when sending SMS (160 characters including spaces). Usually the space available is not used by the texters. Thus, the spelling adaptations would be caused by "the needs for speed, ease of typing and, perhaps, other symbolic concerns" (Thurlow, 2003a). The rise of this new type of SMS language is called "Textspeak" by Crystal (2003, 2008b).

Thurlow (2003a) classified the following spelling adaptations: shortenings, contractions, G clippings and other types of clippings, acronyms, initialisms, letter/number homophones, "misspellings" and typos, non-conventional spellings and accent stylisation. However, the list can be increased by adding other features such as smiley faces, lack of function words, lack of punctuation or over-punctuation[1]. Several of those features imply a relationship between English pronunciation and spelling in SMS language. Nevertheless, it must be born in mind that, as stated by Crystal (2008a, 2009), texting is not the cause of bad spelling and it can improve the user's literacy.

As suggested by Thurlow (2003a) and Thurlow and Poff (2009), SMS seems to follow three sociolinguistic maxims (cf. Grice, 1975):

1. brevity and speed;

[1] See Crystal (2009) for a classification of spelling adaptations into: pictograms and logograms (rebus abbreviation), initialisms, omitted letters (contractions and clippings), nonstandard spellings, shortenings and genuine novelties. López Rúa (2007) classifies the devices employed in text messages into: abbreviations, clippings, initializations (acronyms, analphabetisms and abbreviations composed of initials), phonetic respellings, letter and number homophones and symbols and onomatopoeic expressions.

2. paralinguistic restitution ["redress the apparent loss of such socio-emotional or prosodic features as stress and intonation (Thurlow and Poff, 2009: 14)]; and,
3. phonological approximation ["adds to paralinguistic restitution and engenders the kind of playful, informal register appropriate to the relational orientation of text messaging" (Thurlow and Poff 2009: 14)].

The purpose of this paper is to outline the features that point out a sound-to-spelling relationship providing examples from the SMS corpus at Netting-it.com and offering an explanation of the adaptations people do when texting. What distinguishes this research from previous ones (Thurlow, 2003a; Kunda, 2005) is that an electronic corpus has been employed for data collection rather than questionnaires or surveys.

2. DESCRIPTION OF THE CORPUS

For the purpose of this paper a free online SMS corpus has been used. The corpus can be found at Netting-it.com and contains 202 text messages. The word processor from Microsoft Word has been used to elaborate some statistics and patterns in the corpus with the "word count" function. The length of each individual message was calculated to obtain the overall results.

The corpus consists of 3957 words in total although the calculation is somewhat thorny as

(a) in some text messages there are two or more words combined which stand for two or more words but are counted as one (e.g. "2U" meaning "to you"; "tellmiss" instead of "tell miss"); Some of these combinations may be a mistake from the writer but the results vary with this type of examples.

(b) there are also examples of the opposite: several words which stand for just one word but count as several (e.g. "n e thin" meaning "anything");

(c) Sometimes spaces are left where they should not be, as for example in between exclamation marks or hyphens. Thus, they will be taken as two words instead of one (e.g. "mate! !" instead of "mate!!"). In addition to these three aspects, there are other situations that can be mentioned.

(d) Some apostrophes are missing from the contractions (e.g. "ill" instead of "I'll", meaning I will). These are devious examples which, as when contractions are used

with the apostrophe, also count as one word even if these contractions stand for at least two words.

(e) There are also repetitions of certain words which are probably mistakes from the writers, maybe due to forgetfulness when writing SMS or to non-revision of the message before sending it. There are also other types of mistakes within some words where the spelling is changed or another letter is added, maybe due to the previous reasons or maybe due to typing mistakes. Although the number of words may or not be increased, the number of characters per message will definitely change (e.g. "please dont please dont ignore my calls" instead of "Please don't ignore my calls" or "whill" instead of "will").

On this basis, the entire number of characters including spaces is 18328 and excluding spaces is 14565. There are two text messages which consist of only one word:

a. Message number 87: *ALRITE*
b. Message 128: *G.W.R*

The longest message is message number 161, which consists of 41 words:

> Hey! do u fancy meetin me at 4 at cha – hav a lil beverage on me. if not txt or ring me and we can meet up l8r. quite tired got in at 3 v.pist ;) love Pete x x x

The average message length in the corpus is 19.58911 words. The message with the smallest amount of characters excluding spaces consists of 3 characters (message 89: "*U 2.*"), and the message with the highest amount of characters excluding spaces consists of 139 characters (message 186: "*Si.como no?!listened2the plaid album-quite gd&the new air1 which is hilarious-also bought"braindance"a comp.ofstuff on aphex's ;abel,u hav2hear it!c u sn xxxx*"). The average number of characters per message is 72.10396.

The corpus is not tagged and, therefore, there is no information about the writers of those text messages, such as the age and gender. However, the type of messages and the language employed suggest that the writers are adults.

3. ANALYSIS OF DATA

The devices employed throughout this paper for the classification of the spelling adaptation used when texting follow Thurlow 2003a:

- Shortenings: dropping end letters.
- Contractions: dropping middle letters.
- G Clippings: loss of final *g*.
- Other types of clippings: missing final letter.
- Acronyms: words consisting of just initials pronounced one by one. Words will be classified into acronyms or initialisms depending on their orthographic shape and the way in which they are read. Therefore, if a word consists of just initials and they are pronounced one by one it will be considered as an initialism, but if it consists of more than just initials and it is pronounced as one word it will be considered as an acronym. It must be born in mind that these classifications may vary depending on the source used. Thus, some authors distinguish between acronyms, abbreviations, initialisms and alphabetisms (Cannon, 1989, López Rúa, 2002 & 2004)[2]
- Initialisms: words consisting of more than just initials pronounced as one word.
- Letter/number homophones: a word or part of a word is replaced by a number or letter with the same pronunciation.
- "Misspellings" and typos: misprints or spelling mistakes not done intentionally as a device or as a spelling adaptation.
- Non conventional spellings: abbreviation of a word respelling it phonetically.
- Accent stylization: "phonological approximation such as the 'regiolectal'"

[2] López Rúa (2004: 36) classifies initialisms into acronyms and alphabetisms, claiming that they "are 'degree of shortening', 'source form', and 'pronunciation': in this way, acronyms tend to be straightforwardly described as words formed from the initial letters or larger parts of the words in a phrase, and the parameter 'pronunciation' is often used to distinguish them from alphabetisms."
As stated by Cannon (1989: 107), acronyms are a subclass of abbreviations or vice versa. An acronym is described as:
"[A] word formed from the initial letter or letters of each of the successive parts or major parts of a compound term" (Cannon, 1989: 107). "An acronym pre-serves only the initial part(s) of a multiword source (…) traditional acronyms preserve at most only two initial letters/sounds (…) [A]n acronym must come from a source with at least three constituents, where a combining form can be a constituent (ASP 'Anglo-Saxon Protestant'). Not more than two initial letters/sounds of some or all of the constituents can be retained, though an exception of three or even four is permitted if the majority of the reduction typifies acronymy." (Cannon, 1989: 108)

Apart from the devices mentioned above, which are proposed by Thurlow 2003a, the following devices have also been added as they are also employed quite frequently in text messages:

- Smiley faces and other symbols
- Lack of function words: loss of grammatical words.
- Lack of punctuation: loss of punctuation.
- Over-punctuation: excessive use of punctuation.

The following table displays the results of the adaptations used by the writers of the text messages of the corpus:

Table1

	Shortenings	Contractions	G clippings	Other types of clippings	Acronyms	Initialisms	Letter/number homophones	"Misspellings" and typos	Non-conventional spellings	Accent stylisation	Smiley faces	Total
Total examples	89	67	76	69	0	30	463	31	154	118	125	1222
Percentage	7,3	5,5	6,2	5,6	0	2,5	38	2,5	13	9,7	10	100
Examples without repetition	49	34	36	40	0	20	82	26	66	58	24	435

As the table shows, the most common adaptation when texting are letter/number homophones, followed a long way behind by non-conventional spellings, smiley faces and other symbols and accent stylisation. Not surprisingly, the most frequent adaptations are the ones based on pronunciation. The main reason might be for communication purposes: if two words with different spellings are homophonous, and therefore have the same pronunciation, comprehension will be less problematic and complicated than if the text message is ambiguous. For instance, initialisms and acronyms are quite difficult to understand if the reader does not know what the message is about, and that may hinder understanding. The same may happen with shortenings, contractions and other types of clippings, which are

usually easy to understand but may sometimes become difficult. This factor may account for the low frequency of contractions and other types of clippings. Most of the instances of shortenings were names, which presupposes the sender's and addressee's shared knowledge of those people and, consequently, justifies the somewhat higher usage of shortenings. G clippings are also quite frequent and the reason may be that they are also phonologically based as the final *–g* is never pronounced in the gerund *–ing* even though it influences the way in which the *n* is pronounced.

It is presumably for these reasons that the adaptations based on the sound-to-spelling relationship, i.e., phonological approximation, are more frequent than the other ones as they generate no ambiguity and for reasons of brevity and speed. In addition, it might be claimed that native speakers, following the popular belief of spelling as odd and hilarious, have somehow developed the ability to find connections between English spelling and pronunciation. Hence, native speakers are able to find an alternative spelling for common words with the same pronunciation. This may show evidence that there is an intrinsic relationship between English pronunciation and spelling regardless of the independence that could be expected at first sight.

3.1. Shortenings

In the corpus employed for this research, there are 89 examples of shortenings, repeated words included. There are 49 non-repeated examples of shortenings throughout the text messages analysed. Out of the 49 words, 27 (59.55%) were shortened versions of proper names. There are also 6 examples of the days of the week which, if added, amount to 33 capitalised words, that is, 71.91% of the examples. (See table 1 in the appendix for examples of shortening.).

3.2. Contractions

67 cases of contraction could be found in the corpus, of which 34 are not repeated. Of those 34, 26 do not contain any vowel at all. Taking into account all the instances, 86.56% of the cases elide all the vowels but do keep most of the consonants, which facilitates the understanding of the words. This proves that "[t]exters seem to be aware of the high information value of consonants as opposed to vowels. It is fairly unusual to lose consonants, unless the words are likely to be easily recognized (…) but there are lots of instances where one vowel is dropped" (Crystal:2008b) (For a list of examples see table 2 in the appendix.)

3.3. G Clippings

G clippings consist in the loss of final g's. The loss of final g is quite a familiar practice in SMS language. G clippings are near homophones of the original words and their reduced counterparts. Although the pronunciation is not exactly the same when the g is pronounced or not, it does not give rise to misinterpretations. There are 76 instances of G clippings throughout the corpus, of which 40 are not repeated. In G Clippings, the final –g from the gerund ending is missing.

3.4. Other types of clippings

There are 69 examples of other types of clippings of which 40 are non-repeated. 23 out of those 40 miss just the last letter of the word. That means 73.91 of the words containing another type of clipping have been spelt without the last letter of the word. This is the most frequent technique, as opposed to other words where a letter from the beginning or middle of the word is missing. Understanding a word whose last letter is missing might be easier than if any other letter is missing, especially if it is a double letter or a silent –e.

3.5. Acronyms

There are no illustrations of acronyms in the corpus. It seems that the usage of acronyms is rather rare. It might be speculated that the sporadic use of acronyms may be due to a lack of generalized knowledge of a wide range of acronyms, apart from the most common ones. Consequently, using acronyms would render those text messages difficult to understand.

3.6. Initialisms

There are 30 instances of initialisms and 20 of them are non-repeated instances. 8 of those 20 are the initials of proper names and amount to 46.66% of all the instances of initialisms.

3.7. Letter /number homophones

Examples of letter/number homophones amount to 463, of which only 82 are non-repeated ones. This means that this is quite a recurrent spelling adaptation in the corpus. 57 examples out of the 82 are number homophones (that is, 42.11% of the 463 examples) and the rest are letter homophones. Crystal (2008c: 9) maintains that

> [a]lthough many texters like to be different and enjoy breaking the linguistic rules, they also know they need to be understood. There is no point in paying for a message if it breaks so many rules that it ceases to be intelligible. There is always an unconscious pressure to use the standard orthography. […] When messages are longer, containing more information, the amount of standard orthography actually increases. Many texters alter just the grammatical words.

Thus, the reason for the highly frequent number/letter homophones might be due to the fact that replacing several letters by one homophonous letter or number constitutes a quicker way of communication, and as both words are pronounced in the same way, they will not be misinterpreted. These letter/number homophones imply an evident relationship between English pronunciation and spelling, as even if the words are spelt differently, the words will be pronounced in exactly the same way. The most frequent examples are: *u* (*"you"*), *2* (either *"to"* or *"too"*), *4* (*"for"*), *r* (*"are"*), *ur* (either *"your"* or *"you're"*), *c* (*"see"*), and *b* (*"be"*).

3.8. "Misspellings" and typos

Misprints or spelling mistakes not done intentionally as a device or as a spelling adaptation will be considered as misspellings and typos. There is a total of 31 illustrations of misspellings and typos and 26 of them are not repeated. In this type of examples either one letter is added or one letter is missing, especially where double consonants should be spelt. The number of misspellings has been increased by message number 142 as there are 9 misspellings in only that message. The reason for this is that the writer pretends to imitate the language of a drunk person by adding letters and misspelling the words.

3.9. Non-conventional spellings

Non-conventional spellings are abbreviated phonetic respellings. They are based on homophony of the original items and their reduced counterparts. Out of the 154 instances of non-conventional spellings there are 66 instances of non-repeated examples. This type of writing adaptation is quite a frequent one. It consists in either replacing some letters for other ones basing the change on phonological criteria, or in respelling the word in such a way that the pronunciation of both words would be the same. Obviously, the new spellings are used in the situations where they are shorter than the words replaced. The most prevalent examples are luv ("love"), wot ("what"), thanx ("thanks"), nite ("night"), fone ("phone") wen ("when"), alrite ("alright") and mite ("might"). These examples illustrate a relationship between spelling and pronunciation inasmuch as English native speakers are able to respell the words, inventing nonexistent words which would be pronounced exactly the same. This suggests that native speakers are aware of the English orthographic system and possess a special ability to know how words are pronounced regardless of their spelling.

3.10. Accent stylisation

Accent stylisation is another type of phonetic respelling, but the words are not fully homophonous and the way in which the word is spelt tries to represent accents and regional

phonological features. The corpus features 118 illustrations of accent stylisation of which 58 are non-repeated examples. With this adaptation texters represent phonological approximations in the orthography as well as onomatopoeic and exclamatory expressions. The most prevailing examples are da or de ("the"), wanna ("want to"), ya ("you"), cos ("because"), bout ("about"), neva ("never"), ave ("have") and gonna ("going to"). Two striking examples of accent imitation would be for instance 3 ("free") and ifink ("I think"). Probably these words were written by a Londoner with a Cockney accent since this accent does not distinguish between the dental and the labiodental fricatives. Cockney speakers tend to pronounce <th> as a labiodental fricative. Accent stylisation examples highlight the connections that native speakers make between English orthography and pronunciation since the words used try to graphically represent the way in which they are pronounced.

3.11. Smiley faces and other symbols

There are 125 smiley faces and other symbols representing expressions, feelings or individual words. 24 out of those 125 are non-repeated ones. The most common symbol is x, which represents "kisses". It is frequent to find one, two or three xs although the variations are endless. The usage of & ("and") and of + (either "with" or "and") is also quite regular.

3.12. Lack of function words, lack of punctuation and over-punctuation

Out of the 202 SMS, there is a lack of function words in 69 messages (34.15%), a lack of punctuation marks in 98 text messages (48.51%) but over punctuation in 15 SMS (7.42%). Messages missing only one function word were categorized as lacking function words and those missing one apostrophe were classified as lacking punctuation. However, the last full stop of the SMS was missing in all messages and, thus, was not considered as lacking punctuation. Nevertheless, missing apostrophes, question marks, exclamation marks and commas were taken into account. It is possible to find messages with a lack of punctuation and over-punctuation at the same time, as it is possible to find two exclamation marks and no apostrophes in contractions. The following are examples of this:

(a) Y?WHERE U AT DOGBREATH? ITS JUST SOUNDING LIKE JAN C THAT'S AL!!!!!!!!!

(b) IM GONNAMISSU SO MUCH!!I WOULD SAY IL SEND U A POSTCARD BUTTHERES ABOUTAS MUCH CHANCE OF MEREMEMBERIN ASTHERE IS OFSI NOT BREAKIN HIS CONTRACT!! LUV Yaxx

(c) You stayin out of trouble stranger!!saw Dave the other day he's sorted now!still with me bloke when u gona get a girl MR!ur mum still Thinks we will get 2GETHA!

The lack of punctuation and of function words was expected to be greater in line with the maxims of brevity and speed. However, the results show that even though there is a wide range of examples, the frequency is not over 50%. The risk of ambiguity might be held responsible for this lack. By contrast, over-spelling is not as frequent an adaptation as might have been thought at first. Over-spelling may express attitude (Crystal, 2008b) and thus the low frequency can be considered unexpected.

4. CONCLUSION

The set of devices employed when sending text messages are popular and extensively used in SMS but it is presumably for reasons of lack of ambiguity, brevity and speed that the adaptations based on the sound-to-spelling relationship, i.e., phonological approximation, are the most frequent ones. The corpus is not tagged and, therefore, there is no information about the writers of those text messages, such as the age and gender. However, the type of messages and the language employed suggest that the writers are adults. Thus, further research should be carried out to compare these results with the ones from young teenagers.

APPENDICES

Table 1: Most frequent shortenings

Shortenings	Abs. Freq	Rel. Freq
kate = Katherine	11	12,36
Jen = Jennifer	7	7,87
mo = moment	6	6,74
sat = Saturday	5	5,62
jess = Jessica	4	4,49
Fran = Frances?	4	4,49

Table 2: Most frequent contractions

Contractions	Abs. Freq	Rel. Freq
txt = text	13	19,4029851
2mrw = tomorrow	5	7,46268657
sn = son	5	7,46268657
spk = speak	4	5,97014925
wrk = work	4	5,97014925

Table 3: most frequent G clippings

G clippings	Abs. Freq	Rel. Freq
darlin = darling	15	19,7368421
doin = doing	10	13,1578947
goin = going	9	11,8421053

Table 4: most frequent other types of clippings

Other types of clippings	Abs. Freq	Rel. Freq
Pete = Peter	11	15,942029
hav = have	9	13,0434783
Il = I'll	6	8,69565217
al = all	3	4,34782609
jus = just	3	4,34782609

Table 5: Most frequent initialisms

Initialisms	Abs. Freq	Rel. Freq
J = Name?	6	21,4285714
v = very	3	10,7142857
ps = Post Script	2	7,14285714
RV = name?	2	7,14285714
tb = text back	2	7,14285714

Table 6: Most frequent letter/number homophones

letter/number homophones	Abs. Freq	Rel. Freq
u = you	178	38,4449244
2 = to	58	12,5269978
4 = for	30	6,47948164
r = are	20	4,31965443
ur = your	12	2,59179266
b = be	11	2,37580994
c = see	11	2,37580994
2 = too	10	2,15982721
ur = you are	9	1,94384449

Table 7: Most frequent "misspellings" and typos

"Misspellings" and typos	Abs. Freq	Rel. Freq
bak = back	5	16,1290323
hey = eh	2	6,4516129

Table 8: Most frequent non-conventional spellings

Non-conventional spellings	Abs. Freq	Rel. Freq
luv = love	14	9,09090909
wot = what	13	8,44155844
thanx = thanks	9	5,84415584
nite = night	9	5,84415584
fone = pone	9	5,84415584
alrite = alright	7	4,54545455
wen = when	7	4,54545455
mite = might	5	3,24675325
2nite = tonight	5	3,24675325

Table 9: Most frequent accent stylisation

Accent Stylisation	Abs. Freq	Rel. Freq
da = the	11	9,3220339
Wanna = want to	10	8,47457627
ya = you	9	7,62711864
bout = about	6	5,08474576
cos = because	6	5,08474576
ave = have	4	3,38983051
de = the	4	3,38983051
neva = never	4	3,38983051

Table 10: Most frequent smiley faces and other symbols

Smiley faces	Abs. Freq	Rel. Freq
Xxx	31	24,8
& = and	20	16
X	18	14,4
Xx	18	14,4
+ = and	7	5,6
:)	5	4
Xxxx	5	4

REFERENCES

Baron, N.S. (1998). Letters by phone or speech by other means: The linguistics of email, *Language and Communication* 18, 133-170.

Cannon, G. 1989. Abbreviations and Acronyms in English Word-Formation. *American Speech* 64/2: 99-127

Crystal, D. (2001). *Language and the internet*. Cambridge: Cambridge University Press.

Crystal, D. (2003) TXT NE1? In Susan Tresman and Ann Cooke (eds). *The Dyslexia Handbook* (Reading: British Dyslexia Association, 2006), (pp. 179-83) [expansion of *New Statesman* 2003 article]

Crystal, D. (2008a). The joy of txt. *Spotlight*, November 2008, 16-21.

Crystal, D. (2008b). Texting. *ELT Journal* 62 (1), 2008, 77-83.

Crystal, D. (2008c). Txtng: frNd or foe? *The Linguist*, The Threlford Memorial Lecture 2008, 47 (6), December 2008-January 2009, 8-11.

Crystal, D. (2009). *Txting: the gr8 db8*. Oxford: OUP

Grice, H.P. (1975). Logic and conversation. In Peter Cole and Jerry Morgan (Eds). *Syntax and Semantics: Volume 3, Speech Acts*. New York: Academic Press.

Grinter, R. E., & Eldridge, M. A. (2001). y do tngrs luv 2 txt msg? In W. Prinz, M. Jarke, Y. Rogers, K. Schmidt & V. Wulf (eds). *Proceedings of the seventh European Conference on Computer-Supported Cooperative Work,* (pp. 219-238). Netherlands: Kluwer Academic Publishers.

Kunda, S. (2005). Teens Let Their Fingers Do the Talking, available at: http://www.txt2nite.com/forum/viewtopic.php?t=246

López Rúa, P. (2002). On the Structure of Acronyms and Neighbouring Categories: A Prototype-Based Account. *English Language and Linguistics* 6/1, 31-60.

López Rúa, P. (2004). Acronyms & Co.: A typology of typologies. *Estudios Ingleses de la Universidad Complutense,* 12: 109-129.

López Rúa, P. (2007). Teaching L2 vocabulary through SMS language: some didactic guidelines. *ELIA* 7, 165-188.

Mobile Data Association, *The Q4 2009 Uk Mobile Trends Report*, available at http://www.themda.org/mda-press-releases/the-q4-2009-uk-mobile-trends-report.php

Online SMS corpus available at: http://www.netting-it.com/txt8.html

Thurlow, C. (2003a). Generation Txt? The sociolinguistics of young people's text messaging. *Discourse Analysis Online (DAOL)* available at: http://extra.shu.ac.uk /daol/articles/v1/n1/a3/thurlow2002003.html

Thurlow, C., & McKay, S. (2003b). Profiling 'new' communication technologies in adolescence, *Journal of Language and Social Psychology* 22(1), 94-103.

Thurlow, C., & Poff, M. to appear (2009). The language of text-messaging in S. C. Herring, D. Stein & T. Virtanen (eds). *Handbook of the Pragmatics of CMC*. Berlin and New York: Mouton de Gruyte.