UNIVERSIDADE DA CORUÑA
Departamento de Computación

TESIS DOCTORAL

APLICACIONES DEL PROCESAMIENTO DEL LENGUAJE NATURAL EN LA RECUPERACIÓN DE INFORMACIÓN EN ESPAÑOL

AUTOR: Jesús Vilares Ferro

DIRECTORES: Miguel Ángel Alonso Pardo
José Luis Freire Nistal

A Coruña, Mayo de 2005

Tesis Doctoral: APLICACIONES DEL PROCESAMIENTO DEL LENGUAJE NATURAL EN LA RECUPERACIÓN DE INFORMACIÓN EN ESPAÑOL

Autor: D. Jesús Vilares Ferro

Directores: Dr. Miguel Ángel Alonso Pardo
Dr. José Luis Freire Nistal

Fecha:

Tribunal

Presidente:

Vocal 1º:

Vocal 2º:

Vocal 3º:

Secretario:

Calificación:

A mi familia

Agradecimientos

Quiero dar las gracias a todas aquellas personas que han hecho posible con su ayuda, ejemplo y amistad, que mi camino, paso a paso, día a día, me haya llevado hasta aquí.

En primer lugar a mis directores de tesis, Miguel Alonso Pardo y José Luis Freire Nistal. No todo el mundo tiene la suerte de contar con gente así guiando su trabajo, y no sólo guiándolo, sino compartiéndolo.

Desearía agradecer también a los miembros del tribunal el haber puesto a mi disposición su valioso tiempo y sabiduría para juzgar mi trabajo: Gabriel Pereira Lopes (Universidade Nova de Lisboa), John Irving Tait (University of Sunderland), Éric Villemonte de la Clergerie (Institut National de Recherche en Informatique et en Automatique - INRIA), Carlos Martín Vide (Universidad Rovira i Virgili), y Jorge Graña Gil (Universidade da Coruña). Asimismo, agradezco a los doctores Álvaro Barreiro García y Margarita Alonso Ramos (Universidade da Coruña) el haber aceptado formar parte del tribunal suplente.

Del mismo modo, quisiera mostrar mi agradecimiento a los profesores doctores que, con su suma premura y amabilidad, hicieron posible mi mención como *doctor europeo*: Jacques Farré (Universidad de Niza-Sophia Antípolis), Anne-Marie Hughes (Universidad de Niza-Sophia Antípolis), Vitor Rocio (Universidade Aberta de Portugal), y Paulo Quaresma (Universidade de Évora).

Mi familia ha sido siempre lo más importante para mí. Ellos me han apoyado y guiado siempre, y soy lo que soy gracias a ellos, ni más ni menos, ni menos ni más. Lo primero, gracias a mis padres, porque se esforzaron siempre para darme lo mejor y hacer de mí una buena persona. Gracias de todo corazón a mi hermano, porque sin su cariño, consejo, guía y ayuda me habría quedado a medio camino, aunque no siempre haya sabido expresar como debería lo mucho que se lo agradezco. Gracias a Susa, que buena paciencia debe haber tenido conmigo. Y gracias a Silvia y a David, por traer alegría (y alboroto también) a la casa.

No quiero dejar de agradecer todo el apoyo recibido, de una u otra forma, de todos los miembros del grupo COLE[1]: Jorge, David, Víctor, Fran y Mario, sin olvidarme de las gentes de la división UVigo (Juan y Leandro).

Siempre recordaré con agrado los buenos momentos que he pasado en Portugal durante mis estancias en el GLINT[2], en la Universidade Nova de Lisboa. Conocer a gente como Gabriel, Pablo, Joaquim y Tiago —sin dejar en absoluto de lado a Alexandre y a Greg— supone un acicate para seguir trabajando duro.

Durante el desarrollo del sistema de generación de familias morfológicas agradecí mucho la ayuda de Margarita Alonso, del Departamento de Gallego-Portugués, Francés y Lingüística, aunque todavía tenemos que sacarle todo el partido a ese diccionario de colocaciones. Igualmente, los sinónimos de Santi, llegados desde el Departamento de Lógica y Filosofía de la Moral de la Universidad de Santiago, me están esperando.

Finalmente, puede que mis amigos no hayan podido echarme una mano en mi trabajo

[1] COmpiladores y LEnguajes (`http://www.grupocole.org`)

[2] Grupo de Língua Natural (`http://terra.di.fct.unl.pt/glint/`)

propiamente dicho, pero siempre han estado ahí para darme ánimos. Muy buenos ratos he pasado en el Merlín con Pedro, Poley, Berto, Breo, Pablo, Rubén, Isaac... El apoyo a pie de trinchera ha corrido a cargo de estas gentes del laboratorio de Matemáticas —Javi, Dani, Joaquín, Cris, Elisa, Chus, Belén...— que me han tenido que soportar durante más tiempo del humanamente exigible. Otras personas también han puesto su granito de arena: Montse, Gloria, Rebo, Pili, Fran...; y hay quien, como Mónica, ha puesto dos o tres.

A todos y cada uno, gracias.

Resumen

Este trabajo de tesis se enmarca en los campos del Procesamiento del Lenguaje Natural, área de la ciencia y la tecnología encargada del procesamiento automático del lenguaje natural o lenguaje humano, y de la Recuperación de Información, cuyo objetivo es el de identificar, dada una colección de documentos, aquéllos que son relevantes a una necesidad de información del usuario.

Los sistemas convencionales de Recuperación de Información emplean técnicas estadísticas basadas en la distribución de los términos en el documento y en la colección para estimar la relevancia de un documento. Sin embargo, dado que un proceso de Recuperación de Información exige que el sistema comprenda en cierta medida el contenido del mismo, dicha tarea puede verse como perteneciente al ámbito del Procesamiento del Lenguaje Natural. Este razonamiento se ve apoyado por el hecho de que el mayor problema en la Recuperación de Información es la variación lingüística del idioma, consistente en que un mismo concepto se puede expresar de formas diferentes mediante modificaciones en la expresión.

El objetivo de esta tesis es el desarrollo de tecnología de base para el Procesamiento del Lenguaje Natural y el estudio de la viabilidad de su aplicación en sistemas de Recuperación de Información sobre documentos en español. Si bien existen estudios similares para otras lenguas, con un claro dominio del inglés, el español ha quedado relegado frecuentemente a un segundo plano. Además, la mayor complejidad lingüística del español frente al inglés en todos sus niveles no permite una extrapolación inmediata al español de los resultados obtenidos para el inglés, demandando la realización de experimentos específicos.

Por otra parte, hemos tenido que hacer frente a uno de los principales problemas en la investigación del procesamiento automático del español, la carencia de recursos lingüísticos libremente accesibles. La solución para minimizar este problema pasa por restringir la complejidad de las soluciones propuestas, centrándose en la utilización de información léxica, de obtención más sencilla. El hecho de acotar la complejidad de las aproximaciones planteadas permite que las técnicas desarrolladas sean fácilmente adaptables a otros idiomas de características y comportamiento similar, constituyendo de este modo una arquitectura general aplicable a otros idiomas mediante la introducción de las modificaciones oportunas, como sería el caso, por ejemplo, del gallego, el portugués o el catalán. Por otra parte, a la hora de minimizar el coste computacional de nuestras propuestas de cara a su aplicación en entornos prácticos, se ha hecho amplio uso de tecnología de estado finito.

En este contexto, en primer lugar, se ha desarrollado un preprocesador-segmentador avanzado de base lingüística para la *tokenización* y segmentación de textos en español, orientado principalmente a la etiquetación robusta del español, pero aplicable a otras tareas de análisis. Las modificaciones concretas para su aplicación en Recuperación de Información son también discutidas.

A nivel flexivo, se ha estudiado la utilización de técnicas de desambiguación-lematización para la normalización de términos simples, empleando como términos de indexación los lemas de las palabras con contenido del texto —nombres, adjetivos y verbos.

A nivel derivativo, se ha desarrollado una herramienta de generación automática de familias morfológicas —conjuntos de palabras ligadas derivativamente y que comparten la misma raíz— para su utilización también en tareas de normalización de términos simples. Esta propuesta, sin embargo, no es todavía por completo inmune a los problemas generados por la introducción de ruido por sobregeneración durante la creación de familias.

A nivel sintáctico se ha ensayado una aproximación basada en la utilización de dependencias sintácticas a modo de términos índice complejos más precisos y para tratar la variación sintáctica y morfosintáctica. Estas dependencias son generadas mediante análisis sintáctico superficial empleando dos analizadores de desarrollo propio: Patterns, basado en patrones, y Cascade, basado en cascadas de traductores finitos. Se ha estudiado también el empleo de diferentes fuentes de información sintáctica, consultas o documentos —siendo ésta última más efectiva—, y la restricción a dependencias nominales para la reducción de costes.

Finalmente, también a nivel sintáctico, se ha evaluado una nueva aproximación pseudo-sintáctica que emplea un modelo sustentado sobre similaridades en base a distancias para la reordenación de resultados obtenidos mediante una aproximación clásica basada en la indexación de lemas. De las dos propuestas planteadas, la primera basada en la mera reordenación conforme el nuevo modelo, y la segunda basada en una nueva aproximación a la fusión de datos mediante intersección de conjuntos, ésta última ha sido la más fructífera.

Abstract

This PhD. thesis belongs to both the fields of Natural Language Processing, the area of science and technology which deals with the automatic processing of natural or human language, and Information Retrieval, whose goal consists of identifying the documents in a collection which are relevant to a given information need of the user.

Conventional Information Retrieval systems employ statistical techniques based on the distribution of terms through the document and the collection to calculate the relevance of a document. Nevertheless, since an Information Retrieval process requires the system to understand, in some way, the content of the document, such a task can be also viewed as a Natural Language Processing task. This reasoning is supported by the fact that the major problem of Information Retrieval is linguistic variation of language, namely that the same concept can be expressed in different ways by changing the expression.

The aim of this PhD. thesis is the development of base technology for Natural Language Processing and the study of the viability of its application in Spanish Information Retrieval systems. Although similar works have been done for other languages, with English at the fore, Spanish has stayed in the background. Moreover, the greater linguistic complexity of Spanish with respect to English does not allow a direct extrapolation of the results obtained for English, demanding, in this way, that Spanish has its own experiments.

Furthermore, we have had to face one of the main problems in Natural Language Processing research for Spanish, the lack of freely available linguistic resources. The solution for minimizing this problem consists in restricting the complexity of the solutions proposed, by focusing on the employment of lexical information, which is easier to obtain. The fact of limiting the complexity of the approaches proposed allows the techniques developed to be easily adaptable to other languages with similar characteristics and behavior, resulting in a general architecture which can be applied to other languages by introducing the appropriate modifications, as in the case of Galician, Portuguese or Catalan, for example. On the other hand, in order to minimize the computational cost of our approaches for their application in practical environments, finite-state technology has been widely used. The general architecture proposed is described below.

Firstly, an advanced linguistically-based preprocessor has been developed for tokenization and segmentation of Spanish texts, mainly orientated to robust part-of-speech tagging of Spanish, but also applicable to other analysis tasks. Concrete modifications made for its application in Information Retrieval are also discussed.

At inflectional level, the employment of lemmatization-disambiguation techniques for single word term conflation has been studied, using as index terms the lemmas of content words — nouns, adjectives and verbs.

At derivational level, a tool for the automatic generation of morphological families —sets of derivationally related words that share the same root— has been developed for its employment in single word term conflation. Nevertheless, this approach is not wholly exempt from the problems caused by the noise introduced because of overgeneration during the generation of families.

At syntactic level, an approach based on the use of syntactic dependencies as complex index

terms has been tested in order to obtain more precise index terms and to manage syntactic and morpho-syntactic variation. These dependencies are obtained through shallow parsing by employing two parsers developed for this purpose: PATTERNS, based on patterns, and CASCADE, based on a cascade of finite-state transducers. The use of different sources of syntactic information, queries or documents, has been also studied —the latter being more effective—, as has the restriction of the dependencies employed to only those obtained from noun phrases in order to reduce costs.

Finally, also at syntactic level, a new pseudo-syntactic approach, which employs a retrieval model based on a similarity measure computed as a function of the distance between terms is used for reranking the documents obtained by a classical approach based on the indexing of lemmas. Two different approaches are proposed, the first one based on the mere reranking of the documents according to the new retrieval model, and the second one based on a new data fusion method which employs set intersection, this second approach being more fruitful.

Índice general

Índice de figuras

Índice de tablas

Lista de Abreviaturas

CLEF	*Cross-Language Evaluation Forum*
CLIR	Recuperación de Información Translingüe (*Cross-Lingual Information Retrieval*)
COLE	COmpiladores y LEnguajes
HMM	Modelo de Markov Oculto (*Hidden Markov Model*)
IE	Extracción de Información (*Information Extraction*)
IR	Recuperación de Información (*Information Retrieval*)
NLP	Procesamiento de Lenguaje Natural (*Natural Language Processing*)
QA	Búsqueda de Respuestas (*Question Answering*)
TREC	*Text REtrieval Conference*
WSD	Desambiguación del Sentido de las Palabras (*Word-Sense Disambiguation*)

Capítulo 1

Introducción

©2005 Google - Buscando 8.058.044.651 páginas web

Éste es el mensaje que muestra la página principal de *Google*[1], actualmente el buscador *web* más popular, mientras se redacta esta memoria. Esta cifra es un claro reflejo del increíble crecimiento que durante estos últimos años ha experimentado la web. Una cifra que debería llamarnos todavía más la atención si se tiene en cuenta que, aún siendo una suma más que considerable, únicamente representa una pequeña parte de la información disponible en soporte digital, a día de hoy, en intranets, bibliotecas digitales, etc. Es más, incluso restringiéndonos a la web, se estima que los buscadores más populares y de mayor cobertura —*Google*, *Altavista*[2] o *Yahoo*[3]— logran indexar, a lo sumo, entre un 30 % y un 40 % de la misma [34].

Quedan ya atrás los días en los que el problema venía dado casi siempre por la no disponibilidad de una información dada. Hoy en día la situación es prácticamente la contraria: todo tipo de información, tanto de carácter general como incluso específico, circula por la red; las empresas dedican ingentes recursos a crear bases de datos y repositorios electrónicos donde almacenar todo tipo de información de interés para su negocio; los estudiantes dedican cada vez menos tiempo a visitar la biblioteca para hacer un trabajo, optando en su lugar por Internet; incluso el ciudadano de a pie cuenta frecuentemente en su ordenador personal con una pequeña colección de documentos, de contenido personal o profesional —documentos, informes, páginas web, etc.

Sin embargo, a pesar de esta, podríamos decir, disponibilidad de información virtualmente ilimitada, la situación en cierto modo es similar. Hoy en día es tal la cantidad de información disponible, pero tan escasa su estructuración, que el acceso a una información determinada de nuestro interés puede llegar a convertirse en una tarea sumamente complicada. Como podemos observar, las causas son distintas, pero las consecuencias son las mismas.

Se hace necesario, por tanto, el desarrollo de una serie de técnicas especializadas que permitan hacer frente de modo adecuado a estas demandas de información.

1.1. Sistemas de Procesamiento Automático de la Información

Para dar respuesta a las necesidades de información de los usuarios, nacen los que podríamos denominar, de forma genérica, *sistemas de procesamiento automático de la información*. Estos sistemas tienen como objetivo el de suministrar, de forma eficaz y eficiente, aquella información

[1] `http://www.google.com`
[2] `http://www.altavista.com`
[3] `http://www.yahoo.com`

solicitada por los usuarios. Las diferencias entre los diferentes tipos de sistemas existentes radican principalmente en el enfoque que se da a este tratamiento automático de la información y, sobre todo, en el objetivo final que se desea conseguir, el cual viene dado por el tipo de necesidad que puede tener un usuario en un momento determinado.

Los tres grandes tipos de sistemas de búsqueda y extracción de información son:

1. Sistemas de *Recuperación de Información*.
2. Sistemas de *Extracción de Información*.
3. Sistemas de *Búsqueda de Respuestas*.

El objetivo de un sistema de *Recuperación de Información* (IR, *Information Retrieval*) [26] es el de identificar aquellos documentos de la colección que son relevantes a una necesidad de información del usuario planteada mediante una consulta formada por una serie de términos o palabras clave. Como resultado, estos sistemas devuelven una lista de documentos que suele presentarse ordenada en función de valores que tratan de reflejar en qué medida cada documento responde a dichas necesidades de información. El ejemplo más común de un sistema de este tipo son los motores de búsqueda en Internet.

Mientras los sistemas de Recuperación de Información localizan documentos que tratan el tema especificado por el usuario, los sistemas de *Extracción de Información* (IE, *Information Extraction*) [110] toman dichos documentos y localizan y extraen la información relevante que éstos contienen. Su finalidad consiste, por lo tanto, en detectar, extraer y presentar dicha información en un formato que sea susceptible de ser tratado posteriormente de forma automática. De esta forma el usuario deberá no sólo especificar el tema de interés, sino también especificar qué elementos de información son los que le interesan, generalmente mediante una plantilla que el sistema deberá rellenar de forma automática. Un ejemplo del mismo podría ser un sistema que partiendo de anuncios de venta de inmuebles, extraiga automáticamente el tipo de vivienda al que se refieren, su número de habitaciones, su superficie, su precio, etc., y a continuación almacene dichos datos en una base de datos creada a tal efecto.

Debido a la necesidad de definir detalladamente y con antelación la naturaleza de la información requerida, los sistemas de Extracción de Información presentan una dependencia respecto al dominio de aplicación de tal forma que impiden el tratamiento de preguntas arbitrarias formuladas por el usuario. Para cubrir este hueco surgen los sistemas de *Búsqueda de Respuestas* (QA, *Question Answering*) [236], que tienen como objetivo el de localizar y extraer respuestas concretas a necesidades de información puntuales y específicas por parte del usuario. De esta forma, un sistema de Búsqueda de Respuestas debería ser capaz de dar respuesta a preguntas como *"¿Cuál es la altura del Everest?"* o *"¿Quién compuso la ópera 'Carmen'?"*.

1.2. Procesamiento Automático del Lenguaje

A pesar de las diferencias de objetivos entre los distintos tipos de sistemas anteriormente descritos, éstos guardan ciertas interrelaciones y características comunes, siendo la principal el tipo de entrada de la que todos ellos parten. En ninguno de los casos el sistema toma como entrada un flujo de datos estructurados, sino que parten de textos, es decir, lenguaje humano escrito, con todo lo que ello supone: un documento desestructurado y que por lo general contiene múltiples errores y ambigüedades.

Puesto que el objetivo de estos sistemas es el de la búsqueda y/o extracción de información a partir de dichos documentos en base a una necesidad de información del usuario, es preciso que tales sistemas *"comprendan"*, hasta cierto punto, el lenguaje humano [229] o, como se

referencia en la literatura, el *lenguaje natural*. De esta forma, tales sistemas suelen integrar, en mayor o menor medida, algún tipo de técnica o aproximación al *procesamiento automático del lenguaje*, también llamado simplemente *Procesamiento de Lenguaje Natural* (NLP, *Natural Language Processing*) [59].

En el caso de los sistemas de Extracción de Información y Búsqueda de Respuestas, debido a la gran precisión que se requiere en los procesos de detección y extracción del tipo de información requerida, es deducible que estos sistemas han de aplicar complejas técnicas de Procesamiento de Lenguaje Natural para cumplir adecuadamente su objetivo. Por contra, en lo referente a los sistemas de Recuperación de Información, éste no suele ser el caso, ya que las aproximaciones actuales se fundamentan en la aplicación de técnicas estadísticas que permiten, en la mayoría de las ocasiones, decidir acerca de la relevancia de un documento respecto a una consulta con un grado de efectividad suficiente, dado el menor nivel de compromiso que esta tarea exige.

Sin embargo, el nivel de desarrollo de los sistemas de Recuperación de Información ha hecho patente que el modelo actual de procesamiento está llegando a los límites de su capacidad [229]. Tales límites vienen marcados por la propia naturaleza variable del lenguaje, el hecho de que un mismo concepto puede ser expresado de muy diferentes maneras empleando sinónimos, alterando la estructura sintáctica del enunciado, etc. Para hacer frente a esta *variación lingüística* [24] es preciso ir más allá de la mera estadística y aplicar técnicas de Procesamiento de Lenguaje Natural.

Las investigaciones llevadas a cabo hasta ahora en el campo de la aplicación del Procesamiento de Lenguaje Natural a la Recuperación de Información se han centrado mayoritariamente en el inglés, siendo muy escaso el trabajo realizado para el caso del español. No debemos olvidar que el español es, actualmente, una lengua en expansión a nivel mundial a todos los niveles, tanto meramente poblacional como económico, siendo ya, por ejemplo, la tercera lengua en la red[4] tras el inglés y el chino, y por delante del japonés, el alemán, y el francés —aun cuando su nivel de penetración en la red se considera todavía relativamente bajo—, y la segunda lengua en Estados Unidos, donde el 60 % de los universitarios la eligen como idioma extranjero a cursar. Por otra parte, la mayor riqueza lingüística del español frente al inglés —tanto a nivel morfológico como a nivel léxico, sintáctico y semántico—, no permite una extrapolación inmediata al español de los resultados obtenidos en las investigaciones para el inglés, demandando la realización de experimentos específicos.

Sin embargo, a la hora de trabajar en la aplicación del Procesamiento de Lenguaje Natural a la Recuperación de Información en español, debemos hacer frente a un grave problema, el de la falta de recursos lingüísticos libremente accesibles para este idioma. Deberemos, pues, apostar por técnicas que aborden el problema con un mínimo de recursos, dejando de este modo la puerta abierta a la investigación en otras lenguas próximas en similar o incluso peor situación, tales como el gallego, el portugués, o el catalán.

1.3. Motivación y Objetivos de la Tesis

Nuestro objetivo en esta tesis ha sido el de cubrir dicho hueco en el estudio de la aplicabilidad de técnicas de Procesamiento del Lenguaje Natural en *sistemas de procesamiento automático de la información* para el caso del español, desarrollando la tecnología de base adecuada de uso general o específico para el procesamiento del idioma y comprobando luego su viabilidad en dichos sistemas.

Por otra parte, teniendo presente la necesidad de desarrollar algoritmos y sistemas con capacidad real de integración, una de las premisas en nuestro trabajo ha sido el de la creación

[4] `http://www.glreach.com/globstats/`

de tecnología fácilmente adaptable a otros idiomas y el de la minimización de los costes computacionales, tanto a nivel espacial como temporal, empleando para ello, en la medida de lo posible, tecnología de estado finito. El otro punto tenido en cuenta, y ya anteriormente comentado, es el de abordar el problema con un mínimo de recursos lingüísticos, dada la carencia general de los mismos para el caso del español.

1.4. Ámbito de la Tesis

El trabajo desarrollado durante la tesis se enmarca dentro de dos áreas de investigación: la *Recuperación de Información*, cuya tarea es localizar, dentro de una colección o corpus de documentos, aquellos documentos que son relevantes a una necesidad de información de un usuario, y el *Procesamiento del Lenguaje Natural*, cuyo objetivo último es el de la comprensión del lenguaje humano por parte de la máquina. Asimismo, las técnicas de Procesamiento del Lenguaje Natural aquí desarrolladas entran dentro de los campos de la *Teoría de Autómatas y Lenguajes Formales*, por una parte, por el abundante empleo de autómatas y traductores finitos, y de los *Compiladores y Procesadores del Lenguaje* por otra, por el estudio de nuevas técnicas de análisis sintáctico.

Además, desde un punto de vista tecnológico, esta tesis se sitúa en el marco de las *Tecnologías de la Información y de las Comunicaciones*, apostando por los objetivos marcados en su momento por la iniciativa ESPRIT de Tecnologías de la Información, incluida en el V Programa Marco de la Unión Europea (1998-02). En efecto, el trabajo aquí presentado se refiere al menos a tres de las áreas consideradas en dicho programa:

1. *Sistemas y servicios para el ciudadano.* Entre los puntos considerados podemos destacar el desarrollo de interfaces hombre-máquina, y la concepción de sistemas multimedia y servicios que faciliten la accesibilidad a los datos de las administraciones publicas.

2. *Nuevos métodos de trabajo y comercio electrónico.* Entre los puntos incluidos en este programa destacamos el desarrollo de nuevos métodos que faciliten el acceso remoto a los centros de trabajo.

3. *Herramientas multimedia.* Los puntos de este apartado incluyen la edición interactiva de documentos electrónicos, el desarrollo de tecnologías que faciliten el acceso al entorno educativo, la concepción de estrategias de intercambio de información sobre documentos en diferentes lenguas, y el acceso, filtrado, análisis y proceso de información.

Ya a nivel nacional y tomando como referencia el Programa de Tecnologías de la Información y las Comunicaciones del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (2000-03) dentro del cual se ha desarrollado la mayor parte del trabajo de tesis, éste se enmarca en uno de sus objetivos prioritarios. Nos referimos al desarrollo de plataformas, herramientas y metodologías. Los puntos abordados incluyen las arquitecturas abiertas, los sistemas de información y gestión del conocimiento, los sistemas hipermedia, las interfaces hombre-máquina y los sistemas inteligentes.

En lo que respecta al contexto dentro del cual se ha desarrollado el trabajo de investigación de esta tesis, éste se ha llevado a cabo dentro de diferentes becas y proyectos que recogemos a continuación.

Becas de Investigación:

- Centro Ramón Piñeiro para la Investigación en Humanidades. *Beca Predoctoral de Investigación.*

- Programa de Promoción Xeral da Investigación, Secretaría Xeral de I+D, Xunta de Galicia. *Beca de Tercer Ciclo.*

- Ministerio de Educación y Ciencia. *Beca de Postgrado para la Formación de Profesorado Universitario (FPU).*

Contratos de I+D:

- Fundación Empresa/Universidad de Galicia (FEUGA) —dentro del proyecto *Smart Tulip* del programa *Innovation* de la Unión Europea). *Sistemas de Recuperación de Información a través de Técnicas de Lenguaje Natural para Evaluaciones Cognitivas de la Información.*

Proyectos de I+D de Ámbito Internacional:

- *Extracción y Recuperación de Información Aplicando Conocimiento Lingüístico* (FEDER 1FD97-0047-C04-02).

- *Análisis Sintáctico Robusto para Portugués, Gallego y Español* (acción integrada hispano-lusa HP2001-0044).

- *Analizadores Tabulares para el Lenguaje Natural 2* (acción integrada hispano-francesa HF2002-0081).

- *Leveraging Accesibility FOR Gentler Eclipse – LAForGE* (IBM Eclipse Innovations Awards).

Proyectos de I+D de Ámbito Nacional:

- *Recuperación de Información para la Búsqueda de Respuestas en Textos Económicos* (TIN2004-07246-C03-02).

Proyectos de I+D de Ámbito Autonómico:

- *Clúster de 30 Nodos con Arquitectura x86* (PGIDIT01PXI10501IF).

- *Aplicación de Inteligencia Artificial para la Extracción de Información Cognitiva y Cualitativa en Mercados Financieros* (PGIDIT02SIN01E).

- *Evaluación Interactiva de la Pertinencia en Entornos de Recuperación Automática de Información* (PGIDIT02PXIB30501PR).

Proyectos de I+D de Ámbito de la Universidad:

- *Extracción de Información de Noticias Bursátiles para la Evaluación del Sentimiento del Mercado.*

- *Aplicaciones de la Ingeniería Lingüística a los Sistemas Colaborativos y* Desktop Publishing.

1.5. Estructura de la Memoria

Esta memoria se estructura en cuatro partes. En la primera se realiza una introducción a los campos de la Recuperación de Información y del Procesamiento del Lenguaje Natural. En la segunda, que junto a la tercera conforman el núcleo de la memoria, se presentan diversas técnicas de procesamiento automático a nivel de palabra y su aplicación a la Recuperación de Información. El tercero de los bloques, partiendo de las técnicas de procesamiento a nivel de palabra, va un paso más allá e introduce una nueva serie de técnicas de procesamiento ya a nivel de frase. Finalmente, el cuarto bloque lo constituye una serie de apéndices en los que se presenta material que, si bien no es imprescindible, resulta de interés en el ámbito de la tesis. A continuación presentamos un breve resumen del contenido de cada uno de los capítulos.

Parte I. Conceptos Previos

Capítulo 2. En este capítulo se hace una introducción al campo de la *Recuperación de Información*. Tras presentar una serie de conceptos básicos y delimitar el ámbito de aplicación de este tipo de sistemas, se describe el funcionamiento de los mismos: los paradigmas de representación interna y comparación de documentos y consultas, así como el proceso de generación de dichas representaciones internas y su implementación. Finalmente se describen las medidas de evaluación a emplear, así como las colecciones de prueba utilizadas a tal efecto y la metodología seguida.

Capítulo 3. En este capítulo se presenta el otro campo dentro del cual se encuadra esta tesis, el *Procesamiento del Lenguaje Natural*. Tras realizar una introducción inicial a los diferentes niveles de procesamiento y a la naturaleza de las técnicas a emplear, se realiza una descripción más detallada de dichos niveles así como de las técnicas y herramientas empleadas en cada uno de ellos. Estos niveles son: el nivel morfológico, el nivel sintáctico, el nivel semántico y el nivel pragmático.

Parte II. Normalización de Términos Simples

Capítulo 4. En este capítulo se explica nuestra aproximación al problema de la *tokenización* y segmentación de textos. Describimos nuestro preprocesador avanzado de base lingüística diseñado para la *tokenización*, segmentación y preprocesamiento de textos en español como fase previa a un proceso de desambiguación-etiquetación. Seguidamente se describen las adaptaciones realizadas en el preprocesador original para su empleo en tareas de Recuperación de Información.

Capítulo 5. En este capítulo abordamos el problema de la eliminación de la variación morfológica de carácter flexivo mediante el empleo de la lematización como técnica de normalización. En primer lugar se describe el funcionamiento del etiquetador-lematizador empleado, MrTagoo, así como los recursos lingüísticos empleados. Posteriormente se describe el método de normalización, basado en la indexación de los lemas de las palabras con contenido del texto. Finalmente se evalúa la aproximación propuesta en los corpus de evaluación al efecto.

Capítulo 6. En este último capítulo del bloque dedicado al tratamiento de términos simples, extendemos el estudio del tratamiento de la variación morfológica al caso de la variación morfológica derivativa. Tras una introducción a los aspectos lingüísticos involucrados en la generación de nuevo léxico en la lengua española, se presenta nuestra herramienta de generación de familias morfológicas, conjuntos de palabras ligadas por relaciones derivativas. A continuación

se plantea su uso para el procesamiento de la variación derivativa, evaluando seguidamente su comportamiento en un sistema de Recuperación de Información.

Parte III. Normalización de Términos Complejos

Capítulo 7. En este primer capítulo dedicado al procesamiento de términos complejos realizamos una introducción a la variación lingüística sintáctica y morfosintáctica en este tipo de términos, así como al empleo de dependencias sintácticas en la normalización e indexación de textos. A continuación describimos los dos analizadores sintácticos superficiales empleados en nuestros experimentos para la obtención de dichas dependencias: PATTERNS, basado en patrones, y CASCADE, basado en cascadas de traductores finitos. La validez de nuestra aproximación basada en la indexación de dependencias sintácticas es comprobada empleando los corpus de evaluación, estudiando los resultados obtenidos al emplear tanto la información sintáctica extraída de la consulta como aquélla obtenida a partir de los documentos.

Capítulo 8. Abandonando el paradigma clásico de recuperación basado en documentos, presentamos en este capítulo una nueva aproximación al tratamiento de la variación sintáctica mediante el empleo de un modelo de recuperación basado en la localidad. En primer lugar se realiza un acercamiento al nuevo modelo de recuperación y sus diferencias respecto al modelo clásico, así como las adaptaciones realizadas para su integración en nuestro sistema. Tras presentar los resultados obtenidos con esta propuesta inicial, se presenta una segunda aproximación que utiliza una nueva técnica de fusión de datos mediante intersección, que es seguidamente también evaluada.

Capítulo 9. Finalmente, este último capítulo recoge las conclusiones y aportaciones de este trabajo de tesis, así como las vías de desarrollo futuro del mismo.

Parte IV. Apéndices

Apéndice A. En este primer apéndice se presenta el conjunto de etiquetas morfosintácticas empleadas por el etiquetador-lematizador del sistema, MRTAGOO —capítulo 5—, así como por el preprocesador lingüístico en sus tareas de pre-etiquetación —capítulo 4.

Apéndice B. El segundo de nuestros apéndices recoge las *stopwords* empleadas por el sistema.

Apéndice C. El apéndice recoge los resultados obtenidos durante el proceso de estimación de parámetros para la realimentación por relevancia.

Apéndice D. Se muestran en este apéndice, de forma ordenada, los diferentes sufijos derivativos empleados por el sistema de generación de familias morfológicas mediante morfología derivativa descrito en el capítulo 6.

Apéndice E. Este nuevo apéndice describe los diferentes patrones empleados por el analizador sintáctico superficial PATTERNS para identificar y extraer las dependencias sintácticas del texto —capítulo 7.

Apéndice F. El apéndice recoge los resultados obtenidos durante la puesta a punto de los parámetros de ejecución para los experimentos empleando términos complejos —capítulo 7.

Apéndice G. El último de los apéndices recoge los resultados obtenidos durante la puesta a punto de los parámetros de ejecución de nuestra implementación del modelo basado en localidad —capítulo 8.

1.6. Difusión de Resultados

El trabajo de investigación desarrollado durante la realización de la presente tesis doctoral ha dado lugar a diversos artículos de revista, capítulos de libro y ponencias en congresos. A continuación se detallan dichos trabajos.

Publicaciones de Ámbito Internacional:

1. Jesús Vilares, David Cabrero, y Miguel A. Alonso. **Applying Productive Derivational Morphology to Term Indexing of Spanish Texts**. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volumen 2004 de *Lecture Notes in Computer Science*, páginas 336-348. Springer-Verlag, Berlín-Heidelberg-Nueva York, 2001.

2. Jesús Vilares, Manuel Vilares, y Miguel A. Alonso. **Towards the development of heuristics for automatic query expansion**. In H. C. Mayr, J. Lazansky, G. Quirchmayr y P. Vogel, editores, *Database and Expert Systems Applications*, volumen 2113 de *Lecture Notes in Computer Science*, páginas 887-896. Springer-Verlag, Berlin-Heidelberg-New York, 2001.

3. Jorge Graña, Fco. Mario Barcala, y Jesús Vilares. **Formal Methods of Tokenization for Part-of-Speech Tagging**. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volumen 2276 de *Lecture Notes in Computer Science*, páginas 240–249. Springer-Verlag, Berlín-Heidelberg-Nueva York, 2002.

4. Jesús Vilares, Fco. Mario Barcala, y Miguel A. Alonso. **Using Syntactic Dependency-Pairs Conflation to Improve Retrieval Performance in Spanish**. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volumen 2276 de *Lecture Notes in Computer Science*, páginas 381-390, Springer-Verlag, Berlín-Heidelberg-Nueva York, 2002.

5. Miguel A. Alonso, Jesús Vilares, y Víctor M. Darriba. **On the Usefulness of Extracting Syntactic Dependencies for Text Indexing**. In Michael O'Neill, Richard F. E. Sutcliffe, Conor Ryan, Malachy Eaton y Niall J. L. Griffith, editores, *Artificial Intelligence and Cognitive Science*, volumen 2464 de *Lecture Notes in Artificial Intelligence*, páginas 3–11. Springer-Verlag, Berlín-Heidelberg-Nueva York, 2002.

6. Jesús Vilares, Fco. Mario Barcala, Miguel A. Alonso, Jorge Graña y Manuel Vilares. **Practical NLP-Based Text Indexing**. In Francisco J. Garijo, José C. Riquelme y Miguel Toro, editores, *Advances in Artificial Intelligence - IBERAMIA 2002*, volumen 2527 de *Lecture Notes in Artificial Intelligence*, páginas 635–644. Springer-Verlag, Berlín-Heidelberg-Nueva York, 2002.

7. Jorge Graña, Gloria Andrade y Jesús Vilares. **Compilation of Constraint-based Contextual Rules for Part-of-Speech Tagging into Finite State Transducers**. In J.-M. Champarnaud y D. Maurel, editores, *Implementation and Application of Automata*, volumen 2608 de *Lecture Notes in Computer Science*, páginas 128-137. Springer-Verlag, Berlin-Heidelberg-New York, 2003.

8. Manuel Vilares, Víctor M. Darriba, Jesús Vilares y Leandro Rodríguez-Liñares. **Robust Parsing Using Dynamic Programming**. In O. H. Ibarra y Z. Dang, editores, *Implementation and Application of Automata*, volumen 2759 de *Lecture Notes in Computer Science*, páginas 258-267, Springer-Verlag, Berlin-Heidelberg-New York, 2003.

9. Jesús Vilares, Miguel A. Alonso, Francisco J. Ribadas, y Manuel Vilares. **COLE Experiments at CLEF 2002 Spanish Monolingual Track**. In Carol Peters, Martin Braschler, Julio Gonzalo y Martin Kluck, editores, *Advances in Cross-Language Information Retrieval*, volumen 2785 de *Lecture Notes in Computer Science*, páginas 265–278. Springer-Verlag, Berlín-Heidelberg-Nueva York, 2003.

10. Manuel Vilares, Víctor M. Darriba, y Jesús Vilares. **Parsing Incomplete Sentences Revisited**. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volumen 2945 de *Lecture Notes in Computer Science*, páginas 102-111, Springer-Verlag, Berlin-Heidelberg-New York, 2004.

11. Manuel Vilares, Francisco J. Ribadas y Jesús Vilares. **Phrase Similarity through the Edit Distance**. In Fernando Galindo, Makoto Takizawa y Roland Traunmüller, editores, *Database and Expert Systems Applications*, volumen 3180 de *Lecture Notes in Computer Science*, páginas 306-317. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

12. Jesús Vilares, Miguel A. Alonso y Manuel Vilares. **Morphological and Syntactic Processing for Text Retrieval**. In Fernando Galindo, Makoto Takizawa y Roland Traunmüller, editores, *Database and Expert Systems Applications*, volumen 3180 de *Lecture Notes in Computer Science*, páginas 371–380. Springer-Verlag, Berlín-Heidelberg-Nueva York, 2004.

13. Jesús Vilares, Miguel A. Alonso y Francisco J. Ribadas. **COLE Experiments at CLEF 2003 Spanish Monolingual Track**. In Carol Peters, Julio Gonzalo, Martin Braschler, y Martin Kluck, editores, *Comparative Evaluation of Multilingual Information Access Systems*, volumen 3237 de *Lecture Notes in Computer Science*, páginas 345–357. Springer-Verlag, Berlín-Heidelberg-Nueva York, 2004.

14. Jesús Vilares y Miguel A. Alonso. **Dealing with Syntactic Variation through a Locality-Based Approach**. In Alberto Apostolico y Massimo Melucci, editores, *String Processing and Information Retrieval*, volumen 3246 de *Lecture Notes in Computer Science*, páginas 255–266. Springer-Verlag, Berlín-Heidelberg-Nueva York, 2004.

15. Manuel Vilares, Víctor M. Darriba, Jesús Vilares, y Francisco J. Ribadas. **A Formal Frame for Robust Parsing**. *Theoretical Computer Science*, 328:171-186. Elsevier, 2004.

16. Francisco J. Ribadas, Manuel Vilares y Jesús Vilares. **Semantic Similarity between Sentences through Approximate Tree Matching**. In *Proceedings of the 2nd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2005)*, volumen de *Lecture Notes in Computer Science*. Springer-Verlag, Berlin-Heidelberg-New York, 2005.

17. Enrique Méndez, Jesús Vilares y David Cabrero. **COLE Experiments at QA@CLEF 2004 Spanish Monolingual Track**. A publicar en C. Peters, P.D. Clough, G.J.F. Jones, J. Gonzalo, M.Kluck and B.Magnini, editores, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, volumen de *Lecture Notes in Computer Science*. Springer-Verlag, Berlin-Heidelberg-New York, 2005.

Publicaciones de Ámbito Nacional:

1. Jesús Vilares, Fco. Mario Barcala, y Miguel A. Alonso. **Normalización de Términos Multipalabra mediante Pares de Dependencia Sintáctica**. *Procesamiento del Lenguaje Natural*, 27:123-130, 2001.

2. Jorge Graña, Fco. Mario Barcala, y Jesús Vilares. **Etiquetación Robusta del Lenguaje Natural: Preprocesamiento y Segmentación**. *Procesamiento del Lenguaje Natural*, 27:173-180, 2001.

3. Jesús Vilares, David Cabrero, y Miguel A. Alonso. **Generación Automática de Familias Morfológicas mediante Morfología Derivativa Productiva**. *Procesamiento del Lenguaje Natural*, 27:181-188, 2001.

4. David Cabrero, Jesús Vilares, y Manuel Vilares. **Programación Dinámica y Análisis Parcial**. *Procesamiento del Lenguaje Natural*, 29:129-136, 2002.

5. Jesús Vilares, Fco. Mario Barcala, Santiago Fernández, y Juan Otero. **Manejando la Variación Morfológica y Léxica en la Recuperación de Información Textual**. *Procesamiento del Lenguaje Natural*, 30:99-106, 2003.

6. Manuel Vilares, Víctor M. Darriba, y Jesús Vilares. **Análisis Sintáctico de Sentencias Incompletas**. *Procesamiento del Lenguaje Natural*, 30:107-113, 2003.

7. Jesús Vilares y Miguel A. Alonso. **Un Enfoque Gramatical para la Extracción de Términos Índice**. *Procesamiento del Lenguaje Natural*, 31:243-250, 2003.

8. Miguel A. Alonso, Jesús Vilares, y Francisco J. Ribadas. **Experiencias del Grupo COLE en la Aplicación de Técnicas de Procesamiento del Lenguaje Natural a la Recuperación de Información en Español**. In volumen 22 de *Inteligencia Artificial*, páginas 123–134, 2004.

Contribuciones a Congresos de Ámbito Internacional:

1. Manuel Vilares, Jesús Vilares, y David Cabrero. **Dynamic Programming of Partial Parses**. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nocolov y Nokolai Nikolov, editores, *Proceedings of the EuroConference Recent Advances in Natural Language Processing*, páginas 291-293, Tzigov Chark, Bulgaria, septiembre de 2001.

2. Jorge Graña, Gloria Andrade y Jesús Vilares. **Compilation of Constraint-based Contextual Rules for Part-of-Speech Tagging into Finite State Transducers**. In Jean-Marc Champarnaud y Denis Maurel, editores, *Proceedings of the Seventh International Conference on Implementation and Application of Automata (CIAA 2002)*, páginas 131-140, Tours, Francia, julio de 2002.

3. Fco. Mario Barcala, Jesús Vilares, Miguel A. Alonso, Jorge Graña y Manuel Vilares. **Tokenization and Proper Noun Recognition for Information Retrieval**. In A Min Tjoa y Roland R. Wagner, editores, *Proceedings of the Thirteen International Workshop on Database and Expert Systems Applications. 2-6 September 2002. Aix-en-Provence, France*, páginas 246-250. IEEE Computer Society Press, Los Alamitos, California, USA, 2002.

4. Jesús Vilares, Miguel A. Alonso, Francisco J. Ribadas y Manuel Vilares. **COLE Experiments at CLEF 2002 Spanish Monolingual Track**. In Carol Peters, editora, *Results of the CLEF 2002 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2002 Workshop*, páginas 153-160, Roma, Italia, septiembre de 2002.

5. Jesús Vilares, Miguel A. Alonso, y Francisco J. Ribadas. **COLE Experiments at CLEF 2003 Spanish Monolingual Track**. In Carol Peters y Francesca Borri, editores, *Results of the CLEF 2003 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2003 Workshop*, páginas 197-206, Trondheim, Noruega, agosto de 2003.

6. Jesús Vilares y Miguel A. Alonso. **A Grammatical Approach to the Extraction of Index Terms**. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov y Nikolai Nikolov, editores, *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, páginas 500-504, Borovets, Bulgaria, septiembre de 2003.

7. Enrique Méndez, Jesús Vilares y David Cabrero. **COLE at CLEF 2004: Rapid Prototyping of a QA System for Spanish**. In Carol Peters y Francesca Borri, editores, *Results of the CLEF 2004 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2004 Workshop*, páginas 413-418, Bath, Reino Unido, septiembre de 2004.

8. Francisco J. Ribadas, Jesús Vilares y Miguel A. Alonso. **Integrating Syntactic Information by means of Data Fusion Techniques**. In A. Quesada-Arencibia, R. Moreno-Díaz y J.C. Rodríguez, editores, *Extended abstracts of the 10th Internacional Workshop on Computer Aided Systems Theory (EUROCAST 2005)*, páginas 96-99, Las Palmas de Gran Canaria, España, febrero de 2005.

Contribuciones a Congresos de Ámbito Nacional:

1. David Cabrero, Jesús Vilares, y Manuel Vilares. **Dynamic Programming of Partial Parses**. In *Actas de las Primeras Jornadas sobre Programación y Lenguajes (PROLE)*, páginas 63-76, Almagro (Ciudad Real), España, noviembre de 2001.

2. Fco. Mario Barcala, Eva M. Domínguez, Miguel A. Alonso, David Cabrero, Jorge Graña, Jesús Vilares, Manuel Vilares, Guillermo Rojo, M. Paula Santalla y Susana Sotelo. **Una Aplicación de RI Basada en PLN: el Proyecto ERIAL**. In Emilio Sanchís, Lidia Moreno y Isidoro Gil, editores, *Actas de las I Jornadas de Tratamiento y Recuperación de Información (JOTRI)*, páginas 165-172, Editorial UPV, Valencia, España, julio de 2002.

3. Fco. Mario Barcala, Eva M. Domínguez, Miguel A. Alonso, David Cabrero, Jorge Graña, Jesús Vilares, Manuel Vilares, Guillermo Rojo, M. Paula Santalla y Susana Sotelo. **El Sistema ERIAL: LEIRA, un Entorno para RI Basado en PLN**. In Emilio Sanchís, Lidia Moreno y Isidoro Gil, editores, *Actas de las I Jornadas de Tratamiento y Recuperación de Información (JOTRI)*, páginas 173-174, Editorial UPV, Valencia, España, julio de 2002.

4. Jesús Vilares y Miguel A. Alonso. **Extracción de Términos Índice mediante Cascadas de Expresiones Regulares**. In *Actas de las II Jornadas de Tratamiento y Recuperación de Información (JOTRI 2003)*, páginas 204-211, Leganés (Madrid), España, septiembre de 2003.

1.7. Comunicación con el Autor

Los comentarios y sugerencias acerca de esta memoria y del trabajo en ella reflejado son bienvenidos. Se puede contactar con el autor en

Jesús Vilares Ferro
Departamento de Computación
Facultad de Informática
Campus de Elviña s/n
15071 – La Coruña (España)

jvilares@udc.es
http://www.grupocole.org/~jvilares

En esta página web se encuentra disponible información adicional referente a esta tesis y a trabajos relacionados, así como en la página web del grupo COLE (*COmpiladores y LEnguajes*)

http://www.grupocole.org

Parte I

Conceptos Previos

Capítulo 2

Introducción a la Recuperación de Información

2.1. La Recuperación de Información

La *Recuperación de Información* (IR, *Information Retrieval*) es el área de la ciencia y la tecnología que trata de la adquisición, representación, almacenamiento, organización y acceso a elementos de información [26]. Desde un punto de vista práctico, dada una *necesidad de información* del usuario, un proceso de IR produce como salida un conjunto de documentos cuyo contenido satisface potencialmente dicha necesidad[1]. Esta última puntualización es de suma importancia, ya que la función de un sistema de IR no es la de devolver la información deseada por el usuario, sino únicamente la de indicar qué documentos son potencialmente relevantes para dicha necesidad de información [134]. El ejemplo más popular de un sistema de recuperación de información es el de los motores de búsqueda en Internet tales como *Google*[2], *Altavista*[3] o *Yahoo*[4].

2.1.1. Terminología Básica

Antes de continuar, es necesario introducir algunas definiciones de uso común. En Recuperación de Información el término *documento* hace referencia, de forma genérica, a la unidad de texto almacenado por el sistema y disponible para su recuperación. De este modo, dependiendo de la aplicación o de su ámbito de uso, se tratará de artículos de prensa, páginas web, documentos legales, tesis doctorales, etc., bien completos, bien particionados. Podemos, por ejemplo, procesar por separado cada uno de los capítulos de un libro o cada una de las secciones de un documento si consideramos que su longitud total es excesiva. Por su parte, *colección* denota el repositorio de documentos disponible para resolver las necesidades de información del usuario. Cada una de las unidades léxicas (palabras) que componen un documento —y por extensión, la colección— se denomina *término*. Por su parte, la necesidad de información del usuario, expresada en términos que el sistema pueda comprender, se denomina *consulta* (*query*). Asimismo, los resultados obtenidos son, por lo general, ordenados por grado de similaridad o relevancia respecto a la consulta, introduciendo el concepto de *ordenación* (*ranking*) [203].

El concepto mismo de *relevancia* merece particular atención, ya que si bien se habla de la

[1] Nos estamos restringiendo, pues, a la Recuperación de Información textual, ya que existen actualmente nuevos campos de trabajo como la recuperación multimedia sobre imágenes [149], por ejemplo.

[2] `http://www.google.com`

[3] `http://www.altavista.com`

[4] `http://www.yahoo.com`

relevancia del documento respecto a la consulta, en un sentido estricto tal afirmación no es correcta, ya que el usuario juzgará la relevancia del documento devuelto respecto a su necesidad de información original, no respecto a la consulta en la que ésta ha sido reflejada [110]. Se trata, por tanto, de un concepto con un alto componente de subjetividad.

2.1.2. Recuperación de Información y Sistemas de Bases de Datos

Existen dos grandes tipos de sistemas para el procesamiento de elementos de información [26, 131]: los sistemas de Recuperación de Información y los sistemas de Bases de Datos. Mientras los sistemas de Bases de Datos están optimizados para el manejo de datos estructurados con una semántica bien definida, los sistemas de Recuperación de Información, por el contrario, están diseñados para el procesamiento de texto en lenguaje natural, raramente estructurado y, por lo general, de semántica ambigua. En un sistema de Bases de Datos el usuario introduce una consulta específica expresada en álgebra relacional, obteniendo como salida, en forma tabular, todos los resultados que satisfacen dicho requerimiento sin posibilidad alguna de error —ya que invalidaría por completo el resultado. Sin embargo, en el caso de los sistemas de Recuperación de Información los resultados frecuentemente contienen errores, y no tienen por qué ser completos. De hecho, el objetivo de un sistema de Recuperación de Información es maximizar el número de documentos relevantes devueltos a la vez que se minimiza el número de documentos no relevantes devueltos [131].

2.1.3. Tareas de Recuperación de Información

Los buscadores web, si bien los más populares, no son los únicos sistemas de Recuperación de Información actualmente en funcionamiento, ya que existen diferentes tipos de sistemas dependiendo de la naturaleza de la tarea a realizar. Podemos distinguir los siguientes tipos de tareas de Recuperación de Información:

Recuperación ad hoc. Probablemente la tarea más representativa por ser aquélla en la que se basan los buscadores web. En ella el conjunto de documentos sobre el que se realizan las consultas permanece estable[5] mientras que nuevas consultas procedentes de los usuarios llegan al sistema de forma continua. La colección tiene, pues, un carácter *estático*, mientras que son las consultas las que tienen un carácter variable o *dinámico*.

Categorización o clasificación de documentos.
Consiste en la asignación de un documento a una o más clases de documentos fijadas con anterioridad en función de su contenido. Si bien la recuperación ad hoc puede verse también como una tarea de clasificación de documentos en aquéllos relevantes y no relevantes para una consulta dada, en un sentido estricto es necesario diferenciar ambas tareas. Mientras en el caso de la recuperación denominada ad hoc estamos hablando de necesidades de información puntuales y específicas, en el caso de la *categorización* las necesidades —recogidas en los denominados *perfiles* (*profiles*)— permanecen estables en el tiempo, con escasas modificaciones, y su carácter es menos específico que el de las consultas ad hoc, ya que recoge simultáneamente diversas necesidades de información potenciales del usuario [110]. Por esta razón los perfiles tienden a contener un número muy superior de términos que de las consultas, lo que aumenta su complejidad y disuade a los usuarios de introducir modificaciones en los mismos. Las necesidades de información son, por tanto, de naturaleza *estática* en este caso. Dentro de la tarea general de clasificación podemos

[5] En el caso de la web esta condición se relaja, puesto que con el paso del tiempo se crean, eliminan y modifican múltiples páginas web.

destacar dos tareas concretas de particular interés: el *enrutamiento* (*routing*) [98, 42] y el *filtrado* (*filtering*) [137]. En ambos casos los documentos que van llegando al sistema son comparados con los perfiles preexistentes de los usuarios. La diferencia básica entre las dos tareas reside en que mientras el *enrutamiento* introduce una *ordenación* de los resultados devueltos en función de su similaridad respecto al perfil, el *filtrado* es más estricto, ya que se limita a emitir un juicio respecto a la relevancia del documento, aceptándolo o rechazándolo en función del mismo.

***Clustering* de documentos.** Mientras que en el caso de la clasificación de documentos se asume la preexistencia de una serie de clases o grupos de documentos, el objetivo de la tarea de *clustering* es la de generar una serie de clases o *clústers* a partir de un conjunto dado de documentos. Dichas clases o clústers deben atenerse a los principios de maximización de la similaridad intra-clúster y de minimización de la similaridad inter-clúster [121].

Segmentación de documentos. Consiste en la división automática de un documento en subpartes semánticamente coherentes. Es decir, un documento mayor se particiona en secciones que traten subtemas diferentes [102], bien de cara a su procesamiento por separado como si de documentos diferentes se tratase, bien de cara a mostrar al usuario las partes relevantes del documento devuelto.

Si bien nuestro trabajo podría ser aplicado en cualquiera de estas tareas, nuestros experimentos se han limitado a la recuperación ad hoc. Por esta razón, de aquí en adelante, cuando nos refiramos, en general, a *Recuperación de Información*, nos estaremos refiriendo realmente al concepto, más específico, de *recuperación ad hoc*.

2.2. Modelos de Representación Interna

2.2.1. El Paradigma *Bag-of-Terms*

Obtener una representación adecuada de un documento o consulta es una cuestión clave [230]. Por razones históricas, los documentos han sido generalmente representados como conjuntos de términos. Dichos términos, denominados *términos índice* o *palabras clave*, son generados manualmente por especialistas —el caso de las fichas de una biblioteca, por ejemplo—, o bien automáticamente a partir del contenido del documento, como en el caso de los sistemas de Recuperación de Información. Este tipo de representación interna de los documentos, denominada *bag-of-terms* [26], se basa en una interpretación extrema del denominado *principio de composicionalidad*, según el cual la semántica de un documento reside únicamente en los términos que lo forman [121]. En consecuencia, podemos presumir que, si una palabra determinada aparece en un documento, dicho documento trata dicho tema [130]. De forma similar, si una consulta y un documento comparten uno o más términos índice, el documento debe tratar, de algún modo, el tema sobre el que versa la consulta [24].

Este tipo de representación de documentos (y consultas) mediante conjuntos de términos índice resulta insuficiente para una representación completa y adecuada de la semántica del documento. No tiene en cuenta, por ejemplo, las relaciones entre dichos términos —no es lo mismo *"Juan mató a Pedro"* que *"Pedro mató a Juan"*— o la existencia de diferentes sentidos para una misma palabra. A pesar de estas insuficiencias, dicho paradigma ha venido dominando durante décadas el campo de la Recuperación de Información. Las claves de este éxito residen en su sencillez conceptual y de implementación, y al hecho de que su rendimiento fuera bastante satisfactorio en la práctica, a pesar de la pérdida de semántica inherente a su utilización.

2.2.2. Peso Asociado a un Término

Si bien la semántica de un documento se representa mediante un conjunto de términos índice, resulta patente que no todos los términos tendrán la misma importancia a la hora de representar dicha semántica. Esta importancia se representa asignándole a cada uno de dichos términos un valor numérico que denominaremos *peso* (*weight*), de tal forma que a mayor peso, mayor es la importancia del término [131, 26]. Hablaremos, pues, del peso w_{ij} de un término i en un documento j.

Existen dos factores críticos a la hora de calcular el peso de un término: su frecuencia dentro del documento, y su distribución dentro de la colección. Su importancia radica en dos suposiciones:

1. Los términos que aparecen repetidamente en un documento pueden considerarse como representativos válidos de su semántica, por lo que debería asignárseles un peso mayor. Por ejemplo, si en un documento se hace referencia repetidamente al término *chocolatina*, es lógico pensar que dicho documento habla sobre las chocolatinas. Los estudios de Luhn [140] apoyan este punto, al afirmar que el grado de significatividad de un término dentro de un elemento de información es directamente proporcional a su frecuencia dentro de dicho elemento.

2. A mayor número de documentos en los que aparece un término, menor su poder de discriminación, por lo que deberían recibir, consecuentemente, un peso menor. Por ejemplo, si el término *chocolatina* aparece en gran parte de los documentos de la colección, parece lógico pensar que su utilidad es bastante menor que si sólo apareciese en un pequeño subconjunto de ellos.

La primera de estas suposiciones hace referencia a la frecuencia del término i dentro de un documento j, mientras que la segunda de ellas hace referencia a la frecuencia inversa de documento del término i. Definamos, a continuación, formalmente, dichos factores. Sea N el número total de documentos de la colección, y n_i el número de dichos documentos en los que el término t_i aparece. El factor *frecuencia del término i en el documento j* (tf_{ij}) se define como el número de veces que aparece el término t_i en el documento j. Por otra parte, el factor *frecuencia inversa de documento del término i* (idf_i) se calcula como [117]

$$idf_i \;=\; log\,\frac{N}{n_i} \tag{2.1}$$

donde la aplicación de logaritmos pretende suavizar los valores obtenidos para colecciones de gran tamaño.

Durante el cálculo de pesos es también frecuente la introducción de un tercer factor que tenga en cuenta el tamaño del documento [200], ya que a mayor longitud, mayor la probabilidad de que se produzcan correspondencias, por lo que los documentos de mayor longitud se verían en principio favorecidos respecto a los documentos de menor longitud.

Es frecuente también asumir que los términos índice están incorrelados; es decir, asumir la independencia mutua de los pesos de los términos. De esta forma, conocer el peso w_{ij} de un término i en un documento j no nos permitiría afirmar nada respecto al peso w_{kj} de otro término k en ese mismo documento j. Esto es una simplificación para facilitar el cálculo de dichos pesos, ya que dicha afirmación no es en modo alguno cierta. Por ejemplo, dado un documento en el que aparece la palabra *árbitro*, es mucho más probable que aparezca también la palabra *fútbol* que la palabra *chocolatina*. Sin embargo, experimentos llevados a cabo teniendo en cuenta dicha correlación entre términos no han dado lugar a mejoras significativas en los resultados obtenidos, mientras que sí han aumentado notablemente la complejidad del proceso de cálculo de pesos [26].

2.2.3. Modelos de Recuperación

A la hora de diseñar un sistema de Recuperación de Información es preciso establecer previamente cómo representar los documentos y las necesidades de información del usuario, y cómo comparar ambas representaciones . Es preciso pues, definir el *modelo de recuperación* sobre el que ha de desarrollarse el sistema.

En [26] se define formalmente el concepto de *modelo de recuperación* como una cuadrupla [**D**, **Q**, $\mathcal{F}$, $R(q_i, d_j)$] donde

1. **D** es el conjunto de representaciones de los documentos de la colección.
2. **Q** es el conjunto de representaciones de las necesidades de información del usuario, representaciones denominadas *consultas*.
3. $\mathcal{F}$ es el marco formal dentro del cual modelizar las representaciones de documentos, consultas, y las relaciones entre ambos.
4. $R(q_i, d_j)$ es una función de ordenación que asocia un número real a los diferentes pares consulta $q_i \in \mathbf{Q}$ – representación de documento $d_j \in \mathbf{D}$. Dicha ordenación define una relación de orden entre los documentos de la colección respecto a la consulta q_i.

Seguidamente describiremos los modelos clásicos más representativos: los modelos booleano, vectorial y probabilístico.

El Modelo Booleano

Conceptualmente muy simple, el modelo booleano es el más sencillo de los tres aquí descritos, y se basa en la teoría de conjuntos y el álgebra de Boole [26]. En este modelo inicial el usuario especifica en su consulta una expresión booleana formada por una serie de términos ligados mediante operadores booleanos —comúnmente AND, OR y NOT. Dada la expresión lógica de la consulta, el sistema devolverá aquellos documentos que la satisfacen y que conformarán el conjunto de documentos relevantes. De esta forma, el sistema simplemente particiona los documentos de la colección en dos conjuntos, aquéllos que cumplen la condición especificada (relevantes), y aquéllos que no la cumplen (no relevantes), sin ordenación interna alguna, de forma similar a lo que ocurriría con una base de datos tradicional. Un documento es, por tanto, simplemente relevante o no.

Supongamos, por ejemplo, que queremos tener acceso a aquellos documentos que contengan los términos *modelo* y *booleano*, pero que no contengan el término *vectorial*. La consulta asociada a esta necesidad podría ser:

modelo AND *booleano* AND NOT *vectorial*

mientras que el conjunto de documentos relevantes se correspondería gráficamente con el área rayada de la figura 2.1.

La popularidad del modelo booleano, sobre todo en sus inicios, viene dada por su sencillez tanto a nivel conceptual, por la claridad de sus formalismos, como a nivel de implementación. Además, puesto que las consultas son formuladas a modo de expresiones booleanas, de semántica sumamente precisa, el usuario sabe por qué un documento ha sido devuelto por el sistema, lo que no siempre ocurre en otros modelos más complejos. Por otra parte, dado que los documentos son meros *bag-of-terms*, el proceso de recuperación es extremadamente rápido.

Sin embargo, existen también una serie de desventajas importantes asociadas al modelo booleano. La primera de ellas viene dada por la dificultad que conlleva la formalización de la

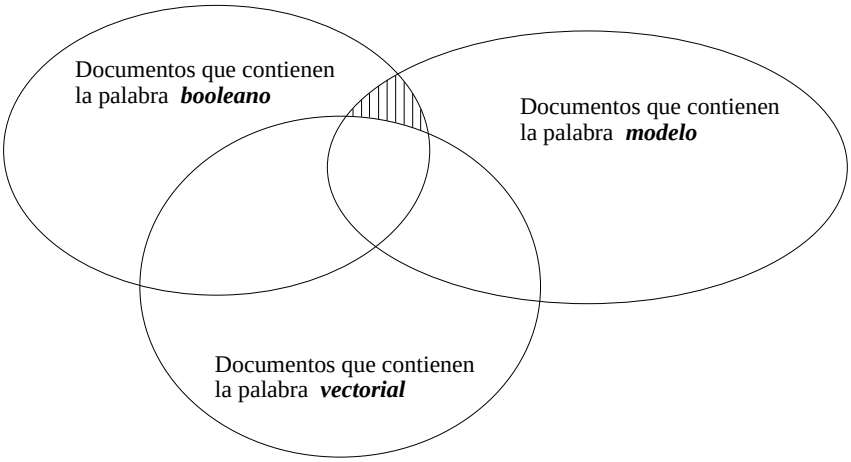


Figura 2.1: Modelo booleano: consulta '*modelo* AND *booleano* AND NOT *vectorial*'

necesidad de información del usuario en forma de expresión booleana, sobre todo cuando se trata de usuarios inexpertos y de necesidades complejas. A ello se suma el hecho de que ligeros cambios en la formulación pueden dar lugar a cambios considerables en el conjunto respuesta. Otro de los grandes inconvenientes del modelo booleano viene dado por su propia naturaleza, de carácter binario. De esta forma, dada una consulta, un documento simplemente es o no relevante dependiendo de si cumple la condición expresada por la consulta. Por lo tanto, no existen ni el concepto de correspondencia parcial ni el concepto de gradación de relevancia. Al no permitir correspondencias parciales, el sistema podría no devolver documentos que, aún siendo relevantes, no verificasen por completo la condición estipulada [266]. Del mismo modo, todos los términos de la consulta tienen la misma importancia, cuando es lógico pensar que la semántica de un texto dado se concentre en mayor grado en ciertos términos —tal y como quedó patente cuando introdujimos el concepto de *peso*. Por otra parte, al no existir ninguna ordenación por relevancia, el usuario se ve obligado a examinar la totalidad del conjunto resultado devuelto.

Si bien bastante popular hace tiempo, por las razones antes comentadas, en la actualidad el modelo booleano se encuentra relegado dentro de los grandes sistemas de Recuperación de Información frente a los restantes modelos a causa de sus desventajas. Sin embargo, continúa empleándose en ciertos ámbitos donde se precisan correspondencias exactas, como en el caso de algunos sistemas de información legislativa [110].

El Modelo Vectorial

Para dar solución a los problemas planteados por el modelo booleano, el modelo vectorial [202, 200] plantea un marco formal diferente en el que se permite tanto la asignación de correspondencias parciales, como la existencia de grados de relevancia en base a los pesos de los términos en consultas y documentos.

En este nuevo modelo, ambos, consultas y documentos, son representados mediante vectores dentro de un espacio multidimensional definido por los propios términos, de tal forma que cada uno de los términos (diferentes) del sistema —es decir, cada uno de los términos del vocabulario— define una dimensión. De este modo, un vocabulario de tamaño t definirá un espacio t-dimensional donde un documento d_j es representado como un vector

$$\overrightarrow{d_j} = (w_{1j}, w_{2j}, \ldots, w_{tj})$$

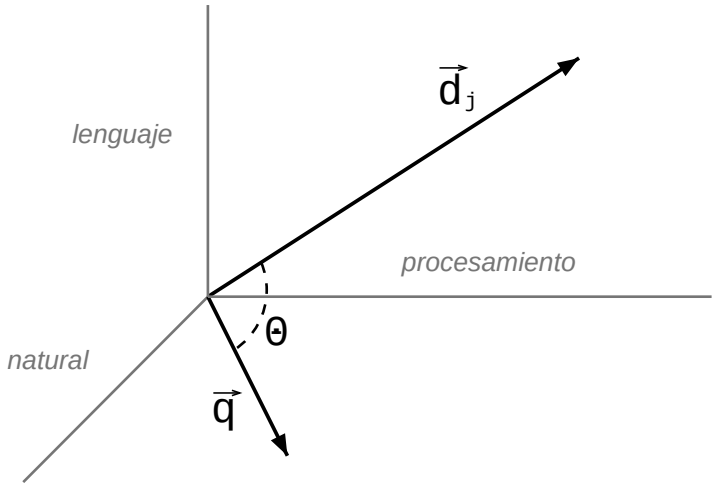


Figura 2.2: Modelo vectorial: espacio tridimensional definido por el vocabulario {*procesamiento*, *lenguaje*, *natural*}

y, paralelamente, una consulta q es representado como un vector

$$\overrightarrow{q} \;=\; (\, w_{1q},\, w_{2q},\, \ldots,\, w_{tq}\,)$$

siendo $w_{ij} \geq 0$ y $w_{iq} \geq 0$ los pesos del término t_i —el i-ésimo término del vocabulario— en el documento d_j y la consulta q, respectivamente.

Desde un punto de vista geométrico, si ambos vectores, consulta y documento, están próximos, es factible asumir que el documento es similar a la consulta —en otras palabras, el documento es posiblemente relevante. Por lo tanto, a mayor proximidad entre ambos vectores, mayor relevancia del documento. En concreto, el modelo vectorial plantea medir la similitud entre un documento d_j y una consulta q en base a la proximidad entre sus vectores correspondientes, en lugar de basarse en criterios de inclusión/exclusión como en el caso del modelo booleano. A su vez, dicha proximidad entre vectores es medida en base al coseno del ángulo Θ que tales vectores forman. Llegados a este punto cabe decir que, al asumir que los términos están incorrelados, esto nos permite suponer que las dimensiones son ortogonales, simplificando notablemente los cálculos. La figura 2.2 muestra un ejemplo gráfico para un vocabulario mínimo de tres términos {*procesamiento*, *lenguaje* y *natural*} que define un espacio tridimensional. De esta forma,

$$sim(d_j, q) \;=\; cos\,(\Theta) \;=\; \frac{\overrightarrow{d_j} \bullet \overrightarrow{q}}{|\overrightarrow{d_j}| \;\times\; |\overrightarrow{q}|} \;=\; \frac{\sum_{i=1}^{t} w_{ij} \;\times\; w_{iq}}{\sqrt{\sum_{i=1}^{t} {w^2}_{ij}} \;\times\; \sqrt{\sum_{i=1}^{t} {w^2}_{iq}}} \tag{2.2}$$

siendo $|\overrightarrow{d_j}|$ y $|\overrightarrow{q}|$ las normas de los vectores documento y consulta, respectivamente.[6]

El modelo vectorial no se limita, pues, a comprobar si los términos especificados en la consulta están o no presentes en el documento, como en el caso del modelo booleano, sino que la similitud entre ambos se calculan en base a los pesos de los términos involucrados, permitiendo de este modo, por un lado, la existencia de correspondencias parciales, y por otro, el cálculo de grados de similaridad o relevancia conforme a los cuales los documentos pueden ser devueltos por orden de mayor a menor relevancia, facilitando notablemente el trabajo del usuario, que puede concentrar

[6]Dado que $|\overrightarrow{q}|$ es constante para una consulta dada, dicho factor puede ser simplificado.

sus esfuerzos en los primeros documentos devueltos —aquellos más relevantes— o incluso definir umbrales de relevancia por debajo de los cuales un documento no es tenido en consideración.

Sin embargo, antes de calcular el grado de similaridad entre los vectores, es necesario calcular los pesos de los términos. Dichos pesos pueden ser calculados de múltiples maneras, si bien el esquema de pesos *tf–idf* y sus derivados [200, 45, 220] se han convertido en los más populares [26]. En el esquema *tf–idf* básico el peso w_{ij} de un término i en un documento j viene dado por la fórmula

$$w_{ij} \;=\; tf_{ij} \;\times\; idf_i \tag{2.3}$$

A la serie de fórmulas para el cómputo de pesos formada por este esquema inicial y las variantes a las que da lugar se las denomina *esquemas tf–idf*.

Los buenos resultados obtenidos con el modelo vectorial, unidos a la simplicidad a nivel de concepto e implementación, su bondad a la hora de aceptar consultas en lenguaje natural, y su capacidad para permitir correspondencias parciales y ordenamiento por relevancia, han hecho de este modelo una de las principales bases sobre la que se han desarrollado gran parte de los experimentos y sistemas en todo el ámbito de la Recuperación de Información [205, 204, 196, 99]. Sus buenas características, unidas al hecho de que sea uno de los modelos de representación más utilizados, le han convertido frecuentemente en el sistema de referencia respecto al cual comparar resultados a la hora de desarrollar nuevos modelos de recuperación [26].

El Modelo Probabilístico

Frente al modelo booleano, basado en teoría de conjuntos, y el modelo vectorial, de carácter algebraico, el modelo probabilístico formaliza el proceso de recuperación en términos de teoría de probabilidades. Las bases del modelo probabilístico fueron establecidas por Robertson y Sparck Jones en [189]. El objetivo perseguido en el modelo es el de calcular la probabilidad de que un documento sea relevante para la consulta dado que dicho documento posee ciertas propiedades [190], propiedades en forma de los términos índice que dicho documento contiene. Según el *principio de orden por probabilidades* [187], el rendimiento óptimo de un sistema se consigue cuando los documentos son ordenados de acuerdo a sus probabilidades de relevancia. En consecuencia, el sistema devolverá los documentos en orden decreciente de las probabilidades de relevancia estimadas mediante el modelo probabilístico.

El modelo parte de las siguientes suposiciones:

1. Todo documento es, bien relevante, bien no relevante para la consulta.
2. El hecho de juzgar un documento dado como relevante o no relevante no aporta información alguna sobre la posible relevancia o no relevancia de otros documentos (suposición de independencia).

En base a ellas, y dada una consulta q, el modelo asigna a cada documento d_j, como medida de similaridad respecto a la consulta, el ratio *P(d_j relevante para q)/P(d_j no relevante para q)*, medida según la cual los documentos son devueltos, ordenadamente, al usuario.

Sea R, pues, el conjunto de documentos que sabemos (o hemos estimado) relevantes, y sea $\overline{R}$ su complementario (es decir, los no relevantes). Sea $P(R|\overrightarrow{d_j})$ la probabilidad de que el documento d_j sea relevante para la consulta q, y $P(\overline{R}|\overrightarrow{d_j})$ la probabilidad de que el documento d_j sea no relevante para la consulta q. La medida de similaridad $sim(d_j, q)$ del documento d_j respecto a la consulta q se define como el ratio

$$sim(d_j, q) \;=\; \frac{P(R|\overrightarrow{d_j})}{P(\overline{R}|\overrightarrow{d_j})} \tag{2.4}$$

que podemos descomponer, aplicando Bayes, en

$$sim(d_j, q) \;=\; \frac{P(\overrightarrow{d_j}|R) \;\times\; P(R)}{P(\overrightarrow{d_j}|\overline{R}) \;\times\; P(\overline{R})} \tag{2.5}$$

donde $P(\overrightarrow{d_j}|R)$ representa la probabilidad de seleccionar aleatoriamente el documento d_j de entre el conjunto de relevantes R y $P(R)$ representa la probabilidad de que un documento de la colección sea relevante. Sus análogos y complementarios vienen dados por $P(\overrightarrow{d_j}|\overline{R})$ y $P(\overline{R})$.

Puesto que $P(R)$ y $P(\overline{R})$ son constantes para todos los documentos de la colección, pueden ser simplificados, obteniendo

$$sim(d_j, q) \;\sim\; \frac{P(\overrightarrow{d_j}|R)}{P(\overrightarrow{d_j}|\overline{R})} \tag{2.6}$$

Asumiendo la independencia entre los términos índice se obtiene

$$sim(d_j, q) \;\sim\; \frac{\left(\prod_{g_i(\overrightarrow{d_j})=1} P(t_i|R)\right) \;\times\; \left(\prod_{g_i(\overrightarrow{d_j})=0} P(\overline{t_i}|R)\right)}{\left(\prod_{g_i(\overrightarrow{d_j})=1} P(t_i|\overline{R})\right) \;\times\; \left(\prod_{g_i(\overrightarrow{d_j})=0} P(\overline{t_i}|\overline{R})\right)} \tag{2.7}$$

donde g_i es una función que devuelve el peso asociado al término k_i dentro de un vector t-dimensional —es decir, $g_i(\overrightarrow{d_j}) = w_{ij}$, con $w_{ij} \in \{0, 1\}$—, $P(t_i|R)$ representa la probabilidad de que un documento seleccionado aleatoriamente de R contenga el término índice t_i, mientras que $P(\overline{t_i}|R)$ representa la probabilidad de que un documento seleccionado aleatoriamente de R no contenga dicho término índice t_i. Las probabilidades asociadas con el conjunto $\overline{R}$ tienen significados análogos.

Tras aplicar logaritmos y eliminar algunos factores constantes dentro de una misma consulta, y sabiendo que $P(t_i|R) + P(\overline{t_i}|R) = 1$, obtenemos

$$sim(d_j, q) \;\sim\; \sum_{i=1}^{t} w_{iq} \;\times\; w_{ij} \;\times\; \left(log\, \frac{P(t_i|R)}{1 - P(t_i|R)} \;+\; log\, \frac{1 - P(t_i|\overline{R})}{P(t_i|\overline{R})} \right) \tag{2.8}$$

donde t es el número de términos que componen el vocabulario del sistema y los pesos de los términos son binarios, $w_{ij} \in \{0, 1\}$ y $w_{iq} \in \{0, 1\}$), indicando meramente la aparición o no del término en el documento o consulta, respectivamente.

Dado que inicialmente el conjunto R no es conocido, se hace necesario estimar las probabilidades $P(t_i|R)$ y $P(t_i|\overline{R})$. De este modo, sea V un subconjunto de los documentos inicialmente devueltos, y que es considerado relevante[7], y sea V_i el subconjunto de V cuyos documentos contienen el término t_i. Aproximaremos $P(t_i|R)$ mediante la distribución del término t_i en V:

$$P(t_i|R) \;=\; \frac{|V_i|}{|V|} \tag{2.9}$$

donde $|V|$ y $|V_i|$ representan el número de elementos en los conjuntos V y V_i, respectivamente. De forma similar, y suponiendo que el resto de los documentos son no relevantes, aproximaremos $P(t_i|\overline{R})$ mediante:

$$P(t_i|\overline{R}) \;=\; \frac{n_i - |V_i|}{N - |V|} \tag{2.10}$$

[7]Bien tras haberlos examinado el usuario, bien seleccionados automáticamente —p.ej., los r primeros documentos devueltos.

donde N es el tamaño de la colección de documentos y n_i el número de documentos de la colección que contienen el término t_i. Para evitar problemas con valores pequeños de $|V|$ y $|V_i|$, comunes en la práctica, se introduce un factor de ajuste, obteniendo finalmente:

$$P(t_i|R) \;=\; \frac{|V_i| + 0{,}5}{|V| + 1} \tag{2.11}$$

$$P(t_i|\overline{R}) \;=\; \frac{n_i - |V_i| + 0{,}5}{N - |V| + 1} \tag{2.12}$$

Substituyendo dichas estimaciones en la expresión 2.8 obtenemos finalmente, tras operar:

$$sim(d_j, q) \;\sim\; \sum_{i=1}^{t} w_{iq} \;\times\; w_{ij} \;\times\; w^{(1)} \tag{2.13}$$

donde $w^{(1)}$ es el denominado *peso Robertson-Sparck Jones* [189], de importancia clave en los esquemas de peso probabilísticos, y que se define como:

$$w^{(1)} \;=\; log\, \frac{(|V_i| + 0{,}5)/(|V| - |V_i| + 0{,}5)}{(n_i - |V_i| + 0{,}5)/(N - n_i - |V| + |V_i| + 0{,}5)} \tag{2.14}$$

Múltiples medidas de similaridad basadas en esta expresión inicial han venido siendo empleadas en diversos sistemas, siendo uno de los más conocidos el sistema *Okapi* [194, 191, 192], cuyo esquema de pesos BM25[193] se encuentra entre los más efectivos y, junto al vectorial *tf–idf*, es punto de referencia para el desarrollo y evaluación de nuevos modelos y nuevos esquemas de pesos. Empleando dicho esquema, el Okapi BM25, la medida de similaridad entre un documento d_j y una consulta q se calcula como:

$$sim(d_j, q) \;=\; \sum_{i=1}^{t} w^{(1)} \;\times\; \frac{(k_1 + 1) \times tf_{ij}}{K + tf_{ij}} \;\times\; \frac{(k_3 + 1) \times tf_{iq}}{k_3 + tf_{iq}} \tag{2.15}$$

donde t es el número de términos que componen el vocabulario
$w^{(1)}$ es el peso Robertson-Sparck Jones definido en la fórmula 2.14
K es calculado como $K = k_1 \times ((1 - b) + b \times dl_j/avdl)$
k_1, b y k_3 son parámetros constantes cuyo valor viene dado en función de la naturaleza de la consulta y de la colección
tf_{ij} es la frecuencia del término t_i en el documento d_j
tf_{iq} es la frecuencia del término t_i en la consulta q
dl_j es la longitud del documento d_j
$avdl$ es la longitud media de los documentos de la colección

2.3. Normalización e Indexación de Documentos

2.3.1. Generación de Términos Índice: Normalización

El proceso de generación de los términos asociados a un documento o consulta se lleva a cabo mediante una serie de transformaciones sucesivas sobre el texto de entrada denominadas genéricamente *operaciones de texto* [26]. Esta serie de transformaciones persigue, además, la reducción del texto a algún tipo de forma canónica que facilite el establecimiento de correspondencias durante el posterior proceso de búsqueda. A este proceso se le denomina *normalización* (*conflation*) [114].

En las aproximaciones clásicas estas operaciones incluyen el análisis léxico del texto, la eliminación de las denominadas *stopwords*, la eliminación de mayúsculas y signos ortográficos, el *stemming* de los términos resultantes y, finalmente, la selección de los componentes de los documentos que serán utilizados por el sistema de recuperación, y que denominaremos *términos índice* para distinguirlos de los términos del documento.

Análisis Léxico del Texto

El proceso de análisis léxico, o *tokenización*, es aquél consistente en la conversión de una secuencia de caracteres —el texto de los documentos o consultas—, en una secuencia de palabras candidatas a ser adoptadas como términos índice por el sistema. Por lo tanto, el objetivo principal de esta primera fase es el de la identificación de las palabras que conforman el texto.

Para ello se considerarán habitualmente tres tipos de caracteres [131]: caracteres de palabra, caracteres interpalabra, y caracteres especiales. Una palabra estará formada por una secuencia de caracteres de palabra delimitada por símbolos interpalabra. Ejemplos de símbolos de palabra serían las letras y los números, mientras que espacios en blanco, comas y puntos son frecuentemente adoptados como símbolos interpalabra. El tercer grupo de símbolos, los símbolos denominados especiales, estarían formados por caracteres que requerirían un procesamiento especial. El guión, por ejemplo, podría ser considerado un carácter especial ya que puede ser empleado de diferentes formas: a final de línea indicaría la continuación de la palabra en la siguiente línea, en otras ocasiones conecta palabras independientes —p.ej., en la expresión inglesa *single-word term* (término unipalabra)—, etc. De esta forma, cuando el sistema detectase un guión —o cualquier otro carácter especial—, debería aplicar una serie de reglas que le permitiesen identificar el caso particular del que se trata y actuar en consecuencia.

Dependiendo del dominio de aplicación y de las características del idioma, la composición de los diferentes tipos de caracteres y su tratamiento variará. De todos modos, el nivel de complejidad de los *tokenizadores* suele ser escaso, siendo herramientas sencillas y rápidas.

Eliminación de *Stopwords*

Denominamos *stopwords* [131, 26] a aquellas palabras de escasa utilidad debido a que su excesiva frecuencia anula su capacidad discriminante —p.ej., formas verbales de *ser* o *estar*— o a que su contenido semántico es escaso —p.ej., artículos y preposiciones. Dada su poca utilidad, dichas palabras son desechadas, lo que permite a su vez un considerable ahorro de recursos, ya que si bien estas palabras representan una parte ínfima del vocabulario —algunas decenas de palabras diferentes—, suponen, en cambio, una cantidad muy importante del número de términos a procesar[8], lo que permite reducir considerablemente el espacio de almacenamiento de las estructuras generadas. En experimentos citados por [266] dicho ahorro supuso en torno a un 25 %, mientras en [26] se habla de un 40 % o incluso más.

Asimismo, en el caso del texto de las consultas es conveniente eliminar también el contenido de su *metanivel* [163, 162], es decir, aquellas expresiones correspondientes a la formulación de la consulta y de las preferencias del usuario acerca de la misma y que no aportan información alguna a la búsqueda, introduciendo únicamente ruido. Tal es el caso, por ejemplo, de *"encuentre*

[8] Lo que concuerda con lo estipulado por la ley de Zipf [272], según la cual dado un corpus suficientemente grande de un idioma dado, si contamos el número de veces que aparece cada palabra, y listamos a continuación dichas palabras por orden de frecuencia, su posición p en dicha lista y su frecuencia f guardan una relación constante

$$f \times r = k$$

de tal forma que, por ejemplo, la palabra en el puesto 50 de dicho ranking es 3 veces más frecuente que aquélla que ocupa el puesto 150.

los documentos que describan ... ". Para ello se debería emplear, en el caso de las consultas, una segunda lista de *stopwords* a mayores de la general, y que podemos denominar de *metastopwords*.

Eliminación de Mayúsculas y Signos Ortográficos

El paso de mayúsculas a minúsculas de los términos procesados es habitual en los sistemas de Recuperación de Información, ya que así se puede evitar la falta de correspondencia con términos que, por ejemplo, estén en mayúsculas únicamente por estar al principio de la oración [131]. Es también frecuente que signos ortográficos tales como tildes o diéresis sean eliminados.

Stemming

Una característica del lenguaje humano es la de que un mismo concepto puede ser formulado de maneras diferentes, que denominaremos *variantes* [111]. Esto supone que a la hora de la comparación de documentos y consultas nos podemos encontrar con que aún refiriéndose a conceptos equivalentes —al menos desde el punto de vista de un sistema de Recuperación de Información [105]—, puedan no producirse correspondencias debido a que ambos estén empleando términos diferentes. Este sería el caso, por ejemplo, de las formas verbales *cocinar* y *cocinaré*, y de las formas nominales *cocina* y *cocinas*. Para minimizar en lo posible el impacto de estos fenómenos, los sistemas de Recuperación de Información recurren a técnicas de *stemming* implementadas mediante herramientas denominadas *stemmers*.

El *stemming* consiste en la reducción de una palabra a su *stem* o supuesta raíz[9] mediante la eliminación de sus terminaciones, siendo ésta la partícula que presumiblemente contiene la semántica básica del concepto [26, 131]. De esta forma, por ejemplo, los términos anteriores se verían reducidos a la cadena *cocin-*, permitiendo de este modo las correspondencias entre los mismos. Si bien el objetivo principal del *stemming* es el de reducir las diferentes formas lingüísticas de una palabra a una forma común o *stem*, y así facilitar el acceso a la información durante el posterior proceso de búsqueda, paralelamente se está reduciendo el número de términos diferentes del sistema, lo que permite a su vez una segunda reducción de los recursos de almacenamiento requeridos.

Entre los algoritmos clásicos de *stemming* destacan el algoritmo de Porter [179] —el algoritmo de *stemming* por excelencia y el más popular—, y el algoritmo de Lovins [139]. En ambos casos podemos diferenciar dos fases: una fase de eliminación de sufijos en base a una lista prefijada de los mismos, y una fase de recodificación de la cadena resultante de acuerdo a una serie de reglas. Asimismo se aplican una serie de restricciones respecto a la longitud del *stem* resultante. Algunos *stemmer* más avanzados, denominados *dictionary look-up stemmers* [131], comprueban además la validez del *stem* obtenido contrastándolo contra un diccionario de *stems* antes de aceptarlo.

Tomemos como ejemplo el caso de la normalización del término inglés *derivational* (derivacional) mediante el algoritmo de Porter. En un primer paso, *derivational* es transformado en *derivate* mediante la aplicación de la regla[10]

$$(m > 0)\ \textit{-ational} \rightarrow \textit{-ate}$$

siendo m la longitud del término en número de secuencias vocal-consonante. De esta forma, *derivational* es primero transformado en *deriv-* en un paso intermedio eliminando la terminación *-ational*, y dado que su longitud ($m = 2$) verifica la condición estipulada ($m > 0$), la transformación es válida, por lo que dicha transformación se completa concatenando al *stem* intermedio el sufijo *-ate*, obteniendo así el término *derivate*. En un segundo paso, *derivate* es

[9] Debemos precisar que el *stem* no tiene por qué coincidir en absoluto con la raíz de la palabra.

[10] Estas reglas se componen frecuentemente de una condición y un par de sufijos, uno a eliminar y otro a concatenar, que son aplicados en caso de que se verifique la condición.

finalmente transformado en *deriv-* mediante la aplicación de la regla

$$(m > 1)\ \text{-}ate \rightarrow \varepsilon$$

donde ε es la cadena vacía. De esta forma, tanto el verbo *to derivate* (derivar) como su adjetivo derivado *derivational* (derivacional), son normalizados ambos a una forma común *deriv-*, lo que permite establecer correspondencias entre ambos.

Sin embargo, la aplicación de mecanismos de *stemming* no está libre de errores [114]. Pueden producirse errores de *sobre-stemming*, donde dos palabras con escasa o nula relación semántica son reducidas a un *stem* común, dando lugar a correspondencias erróneas. Este es el caso, por ejemplo, de *general* (general) y *generous* (generoso), normalizadas ambas como *gener-*. El caso contrario es el de los errores por *infra-stemming*, donde dos términos muy próximos semánticamente son reducidos a *stems* diferentes, impidiendo el establecimiento de correspondencias. Este es el caso, por ejemplo, de *recognize* (reconocer), normalizado como *recogn-*, y *recognition* (reconocimiento), normalizado como *recognit-*.

Selección de Términos Índice

Finalmente, los términos resultantes de las transformaciones de texto previas son adoptados como términos índice, asociándoseles, de ser requerido, el *peso* correspondiente. En la actualidad la mayoría de los sistemas emplean una representación *a texto completo* del texto [26], entendiendo como tal que todos los términos índice generados son empleados para la representación del texto durante el posterior proceso de búsqueda. Cabe citar, sin embargo, que existe también la posibilidad de seleccionar, bien manualmente —por especialistas—, bien automáticamente, un subconjunto de los mismos y que éstos sean los únicos términos índice adoptados para la representación del texto procesado.

2.3.2. Generación de Índices: Indexación

Existen dos alternativas a la hora de realizar una búsqueda dentro de una colección. La primera de ellas es la de realizar una búsqueda secuencial, como es el caso de las herramientas `grep` de Unix o de algunos sistemas de Recuperación de Información como el `seft` [64]. Este tipo de búsquedas secuenciales o *en línea*, resultan apropiadas cuando se trata de colecciones pequeñas, del orden de Megabytes, o cuando ésta sea la única posibilidad debido a la falta de recursos de almacenamiento o a la alta volatilidad de la colección —es decir, que ésta esté siendo modificada amplia y continuamente. La segunda posibilidad, y la más extendida, consiste en la generación de estructuras de datos auxiliares, denominadas *índices*, que permitan acelerar la búsqueda. El sistema de Recuperación estaría formado entonces por dos componentes: las estructuras índice, que contienen las estructuras auxiliares asociadas a la colección, y el *motor de indexación*, el componente software encargado de su generación, manejo, mantenimiento, e interrogación. El empleo de índices es la mejor opción en el caso de colecciones de gran tamaño y de carácter estático o semiestático, entendiendo como tal que dichas estructuras puedan ser actualizadas a intervalos regulares —diario o semanal, por ejemplo— sin perjuicio para el comportamiento del sistema. Este es el caso, por ejemplo, de bases de datos documentales, hemerotecas electrónicas, o la propia web.

Un *índice invertido* o *fichero invertido*, es una estructura de datos de uso frecuente en sistemas de Bases de Datos y Recuperación de Información y cuyo objetivo es el de acelerar los procesos de búsqueda [131, 26]. En dicha estructura podemos diferenciar dos elementos: el vocabulario o diccionario y sus apariciones o *postings*. El *vocabulario* o *diccionario* es, como cabe esperar, el vocabulario del sistema, el conjunto de todos los términos índice diferentes en la colección. Para cada uno de ellos se almacena una lista de sus apariciones (*postings*) dentro de la colección, generalmente a nivel de documento. En otras palabras, el índice consiste en una lista de los

términos índice existentes en la colección y, para cada uno de ellos, una lista de los documentos en los cuales aparece dicho término. Para posibilitar el proceso de cálculo de pesos, se almacenan una serie de datos asociados bien a cada entrada del vocabulario —su frecuencia de documento, por ejemplo—, bien a cada una de las apariciones del término en la colección —su frecuencia dentro de dicho documento, por ejemplo, o directamente su peso.

Asimismo es también frecuente la existencia de un tercer componente, un *fichero de documentos* [131] consistente en una lista de los documentos almacenados por el sistema, y para cada uno de ellos, datos como su longitud, la frecuencia máxima de sus términos, etc.

En lo que respecta a la implementación concreta de la estructura de datos, ésta admite diversas posibilidades: mediante tablas de dispersión, árboles-B, etc.

El proceso de generación del índice se denomina *indexación*[11]. En la figura 2.3 mostramos un ejemplo de dicho proceso. Partimos de tres documentos, denominados DOC_A, DOC_B y DOC_C. El documento DOC_A contiene, entre otros, los términos *management* y *derivational*, que se convertirán en *manag-* y *deriv-*, respectivamente, mediante las correspondientes operaciones de texto. El documento DOC_B contiene los términos, *derivate*, *derivational* y *manages*, que serán normalizados como *deriv-*, *deriv-*, y *manag-*, respectivamente. Finalmente, DOC_C contiene el término *management*, normalizable a *manag-* como se ha indicado.

Una vez obtenidos los términos índice asociados a cada documento, procedemos a su indexación, generándose el índice cuyas estructuras podemos ver en la parte inferior de la figura. El fichero de documentos es, en este caso, bastante simple, ya que únicamente almacena un identificador interno (DID), y la longitud del documento (LONG)—habiendo supuesto longitudes de 200, 414 y 70 términos. Por su parte, el diccionario contiene, para cada término de la colección, su frecuencia de documento (DF), es decir, el número de documentos en los que aparece. Finalmente, para cada uno de dichos documentos existe una entrada o *posting* en el fichero de apariciones con la información asociada a dicho término en dicho documento, en este caso la frecuencia en el documento (TF).

2.4. El Proceso de Búsqueda

Como ya hemos comentado, el objetivo primario de un sistema de Recuperación de Información es el de transformar una *necesidad de información* del usuario en una lista de documentos de la colección cuyo contenido cubra dicha necesidad. Para ello el primer paso consiste en que el usuario plasme su necesidad de información en una *consulta* aceptada por el sistema. Por su parte, el sistema transformará dicha consulta en una representación interna que permita su comparación con los documentos indexados de acuerdo con el modelo de recuperación empleado. Dado que esta comparación se basa por lo general en la aparición de los términos índice de la consulta en el documento, el sistema deberá aplicar sobre la consulta formulada por el usuario las mismas operaciones de texto aplicadas en el caso de los documentos, para así obtener una representación compatible que permita dicha comparación.

La *consulta* supone, pues, un intento por parte del usuario de especificar las condiciones que permitan acotar dentro de la colección aquel subconjunto de documentos que contienen la información que desea. Por lo tanto, el sistema parte de la consulta formulada por el usuario, no de la necesidad de información original, por lo que una formulación incorrecta o insuficiente no podrá guiar adecuadamente al sistema durante el proceso de búsqueda. A este respecto los mayores problemas a los que ha de hacer frente el sistema de IR son, por una parte, la escasa habilidad del usuario a la hora de formular su necesidad en forma de consulta y, por otra, que a

[11] Es también frecuente en la literatura denominar *indexación* al proceso completo formado por los procesos de generación de términos índice y de generación del índice propiamente dicho.

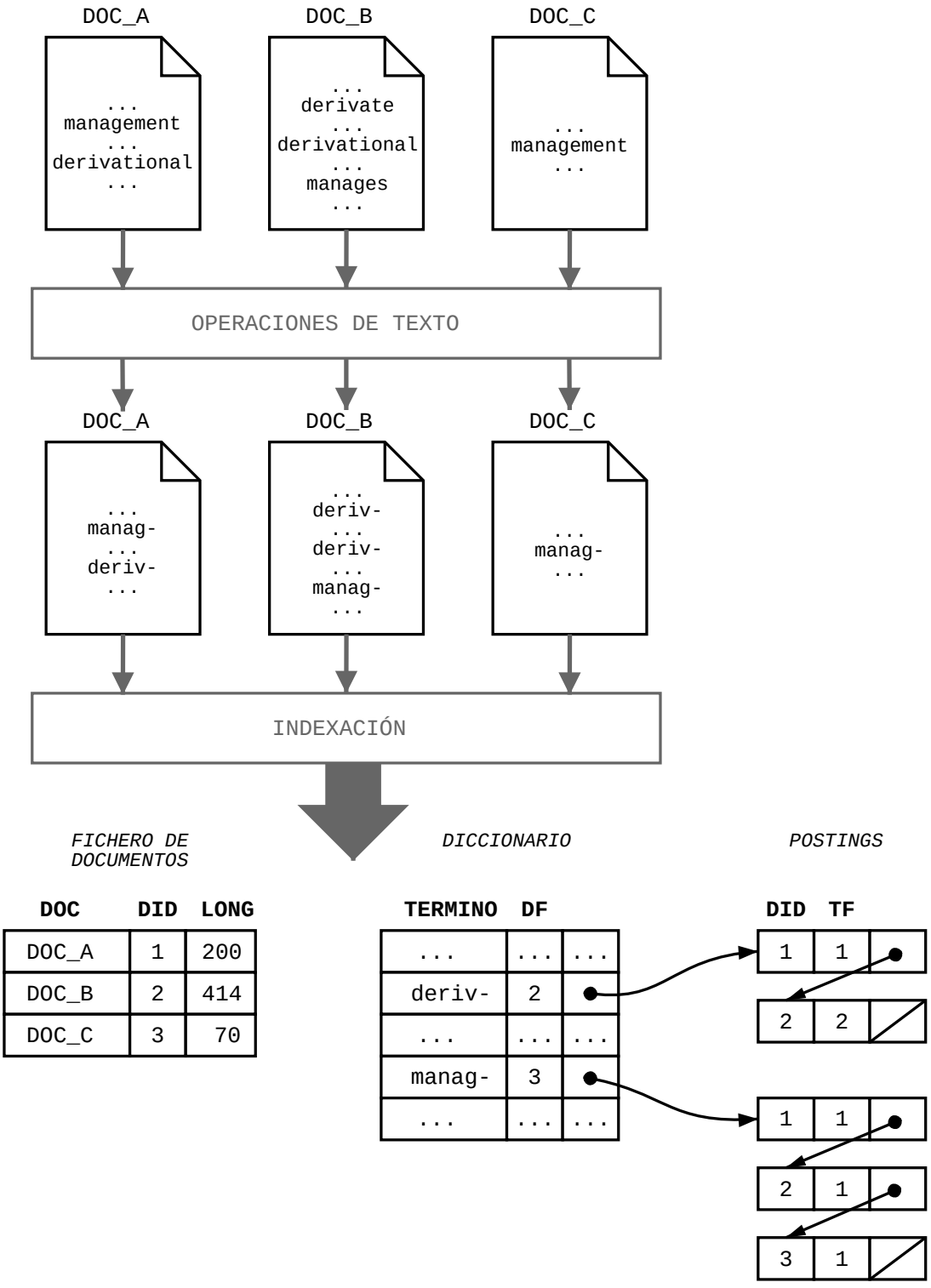


Figura 2.3: Generación de un índice

la hora de describir un mismo concepto los términos empleados por el usuario y los autores de los documentos suelen diferir, impidiendo el establecimiento de correspondencias [268].

Para tratar de paliar esta situación los sistemas de Recuperación de Información suelen incluir mecanismos de expansión de la consulta que permitan la reformulación de la consulta inicial para aumentar su efectividad.

2.4.1. Expansión de Consultas

Bajo la expresión *expansión de consultas* (*query expansion*) se engloban una serie de procesos automáticos o semiautomáticos que permiten la reformulación o refinamiento de la consulta inicial con objeto de aumentar la efectividad del proceso de recuperación, generalmente mediante la adición de nuevos términos, bien relacionados con los términos inicialmente introducidos por el usuario, bien asociados a documentos que se saben o suponen relevantes.

El empleo de técnicas de expansión de consultas permite, por lo general, una mejora de los resultados obtenidos, si bien también se incurre en el riesgo de introducir incorrectamente términos que no guarden relación alguna con el objetivo de la búsqueda y que, por tanto, dañen el rendimiento del sistema [110].

Entre las técnicas de expansión de consultas destacan dos: aquéllas basadas en tesauros y aquéllas basadas en realimentación.

Expansión de Consultas Mediante Tesauros

Un *tesauro* (*thesaurus*) es una base de datos lexicográfica que almacena una representación jerárquica de un lexicón de acuerdo a las relaciones semánticas existentes entre sus palabras [148]. Dependiendo de su ámbito de aplicación, los tesauros pueden ser de carácter general [157, 263] o específico [8]. En lo que respecta a su construcción, ésta puede ser llevada a cabo manualmente por especialistas [157, 263], o bien automáticamente a partir de un corpus empleando técnicas estadísticas [91]. Mención aparte merecen sistemas híbridos como el desarrollado por Fernández Lanza [71], en el cual a partir de un diccionario común impreso de sinónimos y antónimos [35] se genera automáticamente un diccionario electrónico con medidas ponderadas de sinonimia y antonimia.

El empleo de tesauros nos permite reformular la consulta inicial lanzada contra el sistema bien de forma manual [131] —navegando por la estructura jerárquica del tesauro y eligiendo los términos a utilizar—, bien automáticamente [261, 241].

Uno de las herramientas más utilizadas en este tipo de expansión es WordNet [158, 156, 97, 70, 33], una base de datos lexicográfica del inglés en la que sustantivos, verbos, adjetivos y adverbios se organizan en *synsets*, conjuntos de sinónimos de tal modo que cada *synset* está asociado a un sentido determinado. WordNet establece relaciones semánticas de diferente signo:

- **Sinonimia:** dos palabras serán consideradas sinónimas si tienen algún menos un sentido en común; p.ej., *pipe* (tubería) y *tube* (tubería). Como se ha precisado, es la relación básica en torno a la cual se articula WordNet en base a la noción de *synset*.
- **Antonimia:** entre palabras de significados contrarios; p.ej., *wet* (mojado) y *dry* (seco).
- **Hiperonimia e hiponimia:** relación *es-un*. El término más general es el hiperónimo —p.ej., *flower* (flor)—, y el más específico el hipónimo —p.ej., *rose* (rosa).
- **Meronimia y holonimia:** establece una jerarquía *parte-de*. El conjunto es el holónimo —p.ej., *fleet* (flota)—, y la parte es el merónimo —p.ej., *ship* (barco).

Asimismo existe también una base de datos paralela para lenguas europeas denominada EuroWordNet [263].

Por lo general, las aproximaciones actuales basadas en tesauros no logran mejorar los resultados obtenidos. Esto se debe fundamentalmente a los problemas asociados a la introducción de ruido durante el proceso de expansión [261]. Dado que una misma palabra puede tener asociados diferentes sentidos, antes de expandir una palabra sería necesaria una *desambiguación del sentido de las palabras* (WSD, *Word-Sense Disambiguation*) [226, 68] para así expandir únicamente con los términos relacionados semánticamente con ese sentido de la palabra.[12]

Expansión de Consultas Mediante Realimentación

Los métodos de expansión basados en realimentación por relevancia (*relevance feedback*) son, actualmente, el método de reformulación de consultas más extendido debido a su buen comportamiento general [26].

Las técnicas de realimentación se basan en el empleo de la información acerca de la relevancia, o no relevancia, de un subconjunto de los documentos devueltos mediante la consulta inicial para:

- Expandir la consulta en sentido estricto, añadiendo nuevos términos pertenecientes a los documentos considerados relevantes.
- Modificar los pesos de los términos de la consulta buscando optimizar su rendimiento.

La información acerca de la relevancia o no de los documentos devueltos inicialmente se puede obtener mediante interacción con el usuario, o de forma automática tomando como relevantes los n primeros documentos devueltos inicialmente. En este último caso, denominado *expansión por pseudo-relevancia* o *expansión ciega* [131], existe el riesgo de tomar como relevantes documentos que no lo sean, dañando así el rendimiento del sistema. Los resultados obtenidos, sin embargo, son por lo general bastante satisfactorios [110].

La expansión mediante realimentación presenta, además, una serie de ventajas:

- Aisla al usuario de los detalles del proceso de reformulación, debiendo simplemente indicar su criterio de relevancia respecto a los documentos devueltos inicialmente (en el caso de que no se trate de una expansión ciega).
- Permite dividir el proceso completo de búsqueda en una secuencia de pasos más pequeños y fáciles de manejar.

Este tipo de expansión fue puesto en práctica inicialmente en el contexto del modelo vectorial [196], aunque posteriormente su aplicación se extendería al modelo probabilístico [189].

La realimentación por relevancia aplicada al modelo vectorial parte de la suposición de que los documentos relevantes son similares entre sí y de que, por su parte, aquellos documentos no relevantes son disimilares respecto a los sí relevantes. De esta forma, habiendo identificado una serie de documentos relevantes y no relevantes entre los documentos devueltos por la consulta inicial, la idea consiste en reformular el vector consulta de modo que se aproxime al centroide de los documentos relevantes identificados (realimentación *positiva*) y se aleje además del centroide de los no relevantes identificados (realimentación *negativa*). Formalmente, partiendo de un vector consulta

$$\overrightarrow{q} = (w_{1q}, w_{2q}, \ldots, w_{tq})$$

[12] En un sentido estricto, sin embargo, para que el proceso fuese completo sería preciso también un proceso de desambiguación similar en los documentos, lo que conllevaría unos costes inasumibles.

el proceso de realimentación da lugar a un nuevo vector

$$\overrightarrow{q}' = (w'_{1q}, w'_{2q}, \ldots, w'_{tq}, w'_{(t+1)q}, \ldots, w'_{(t+k)q})$$

donde los pesos iniciales w_{ij} han sido actualizados a unos nuevos pesos w'_{ij} y donde se han introducido k nuevos términos. El cálculo del nuevo vector $\overrightarrow{q}'$ se lleva a cabo mediante alguna de las expresiones propuestas por Rocchio [196] e Ide [108], bastante simples y similares entre sí:

$$\textit{Rocchio}: \quad \overrightarrow{q}' = \alpha \overrightarrow{q} + \beta \sum_{j=1}^{n_1} \frac{r_j}{n_1} - \gamma \sum_{j=1}^{n_2} \frac{s_j}{n_2} \tag{2.16}$$

$$\textit{Ide normal}: \quad \overrightarrow{q}' = \alpha \overrightarrow{q} + \beta \sum_{j=1}^{n_1} r_j - \gamma \sum_{j=1}^{n_2} s_j \tag{2.17}$$

$$\textit{Ide dec-hi}: \quad \overrightarrow{q}' = \alpha \overrightarrow{q} + \beta \sum_{j=1}^{n_1} r_j - \gamma\, s_1 \tag{2.18}$$

donde $\overrightarrow{q}'$ es el nuevo vector de la consulta
$\overrightarrow{q}$ es el vector de la consulta inicial
r_j es el vector del j-ésimo documento relevante
s_j es el vector del j-ésimo documento no relevante
n_1 es el número de documentos relevantes examinados
n_2 es el número de documentos no relevantes examinados
y α, β y γ son, respectivamente, los parámetros que controlan las contribuciones relativas de la consulta original, los documentos relevantes, y los documentos no relevantes.

Es frecuente desechar el componente de realimentación negativa fijando el factor γ a cero, ya que si bien dicho componente aleja el vector consulta de los vectores documento no relevantes, eso no quiere decir que lo aproxime más a los vectores documento relevantes, que en última instancia es lo que se pretende [131].

En lo que respecta a la realimentación dentro del modelo probabilístico, ésta presenta unas características diferentes al caso vectorial. Mientras en el modelo vectorial se realizaba simultáneamente una expansión y una modificación de los pesos, en el caso del modelo probabilístico se trata de dos pasos claramente diferenciados.

La modificación de pesos por realimentación es contemplada inherentemente por el propio modelo, ya que el componente $w^{(1)}$ del *peso Robertson-Sparck Jones* (fórmula 2.14) integra la información acerca de la relevancia de los documentos devueltos.

En lo que respecta a la fase de expansión propiamente dicha, ésta fue introducida posteriormente. Se trata de aproximaciones donde los términos procedentes de los documentos relevantes son ordenados en base a algún tipo de función, expandiendo luego la consulta con los términos mejor posicionados. Existen diversas funciones de ordenación al respecto [99], entre la que podemos citar la propuesta de Robertson [188], que emplea el propio peso del término t ponderado por su distribución en el conjunto de documentos relevantes. De esta forma, la puntuación p_i asignada a un término t_i vendría dada por:

$$p_i = |V_i| \times w^{(1)} \tag{2.19}$$

donde $|V_i|$ era el número de documentos relevantes que contenían el término t_i y $w^{(1)}$ es el peso Robertson-Sparck Jones definido en la fórmula 2.14.

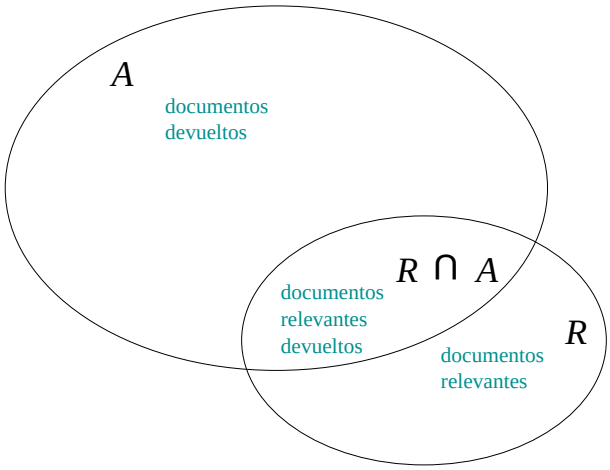


Figura 2.4: Documentos relevantes y documentos devueltos

2.5. Evaluación

A la hora de evaluar un sistema de Recuperación de Información existen múltiples aspectos a tener en cuenta [26]: su *eficiencia* referida a sus costes espacio-temporales asociados; su *efectividad* a la hora de devolver el mayor número de documentos relevantes, minimizando a la vez el número de no relevantes devueltos [235]; el *esfuerzo* realizado por el usuario a la hora de formular o modificar su consulta; y la *amigabilidad* del interfaz de presentación de resultados en relación al esfuerzo requerido por el usuario para su interpretación.

En este trabajo nos centraremos en la evaluación de la efectividad del sistema, diferenciando dos aspectos: por una parte las medidas de evaluación empleadas y, por otra, las colecciones de referencia necesarias para dicha evaluación. Discutiremos estos puntos en los apartados siguientes, así como la metodología de evaluación aplicada en nuestros experimentos y el motor de indexación empleado en los mismos.

2.5.1. Medidas de Evaluación

Las medidas de evaluación aquí descritas pueden calcularse bien como medidas puntuales relativas a una consulta concreta, bien como medidas globales relativas a un conjunto de consultas. En este último caso las medidas son calculadas promediando los valores obtenidos para cada consulta individual respecto al número de consultas empleado, salvo en el caso de la precisión media de documento, que será descrito en detalle más adelante.

Precisión y Cobertura

Las medidas básicas de evaluación en un sistema de Recuperación de Información son la *precisión* (*precision*), que mide la capacidad del sistema para recuperar sólo documentos relevantes, y la *cobertura* (*recall*), que mide la capacidad del sistema para recuperar todos los documentos que son relevantes.

Tal como se aprecia en la figura 2.4, y dada una consulta q, sea R, de tamaño $|R|$, el conjunto de documentos relevantes para dicha consulta; sea A, de tamaño $|A|$, el conjunto de documentos devueltos por el sistema; y sea $R \cap A$, de tamaño $|R \cap A|$, el conjunto de documentos relevantes devueltos por el sistema. Definimos formalmente las medidas de *precisión* y *cobertura* como:

$$Precisión = \frac{|R \cap A|}{|A|} \tag{2.20}$$

$$Cobertura = \frac{|R \cap A|}{|R|} \tag{2.21}$$

Precisión a los 11 Niveles Estándar de Cobertura

Tanto la precisión como la cobertura evalúan la calidad del conjunto de documentos devuelto como tal, como un conjunto, sin tener en cuenta el orden en que han sido devueltos los documentos, y requiriendo además que dicho conjunto respuesta haya sido examinado en su totalidad.

Sin embargo, al usuario se le presentan los documentos devueltos por el sistema de forma ordenada de acuerdo con su relevancia o similaridad respecto a la consulta, de mayor a menor grado de relevancia. Por lo tanto, los valores de precisión y cobertura irán variando conforme el usuario examina, ordenadamente, los documentos devueltos. De esta forma, podemos calcular la precisión para el conjunto de documentos examinados cuando se hayan alcanzado determinados valores de cobertura, es decir, cuando se haya recuperado ya un determinado porcentaje de documentos relevantes. Generalmente se muestran los valores de precisión obtenidos al nivel 0.0 de cobertura (correspondiente al 0 % de cobertura), al nivel 0.1 (correspondiente al 10 %), al 0.2 (al 20 %), y así sucesivamente en incrementos de 0.1 (10 %) hasta alcanzar el nivel 1 (100 %). A dichos niveles se les denomina los *11 niveles estándar de cobertura*.

Dado que no siempre es posible calcular la precisión a un nivel de cobertura concreto, las precisiones son interpoladas. Sea r_j el j-ésimo nivel de cobertura estándard —p.ej., r_5 denota el nivel de 50 % de cobertura—, entonces la precisión interpolada $Pr(r_j)$ a dicho nivel de cobertura se calcula como

$$Pr(r_j) = max_{\,r_j \leq r \leq r_{j+1}}\, Pr(r)$$

es decir, la precisión interpolada al j-ésimo nivel de cobertura estándar es la precisión máxima conocida en cualquier valor de cobertura entre dicho nivel j-ésimo y el nivel $(j+1)$-ésimo.

A modo de ejemplo, supongamos una colección de 20 documentos, donde únicamente 4 son relevantes para una consulta dada. Supongamos que nuestro sistema, ante dicha consulta, ha devuelto dichos documentos en las posiciones 1ª, 2ª, 4ª y 15ª. Al examinar los documentos devueltos, los valores de precisión y cobertura para cada documento relevante encontrado son:

- **1º documento relevante:** precisión 1, cobertura 0.25 (1 documento relevante, 1 documento recuperado)
- **2º documento relevante:** precisión 1, cobertura 0.50 (2 documento relevantes, 2 documentos recuperados)
- **3º documento relevante:** precisión 0.75, cobertura 0.75 (3 documento relevantes, 4 documentos recuperados)
- **4º documento relevante:** precisión 0.27, cobertura 1 (4 documento relevantes, 15 documentos recuperados)

Por lo tanto, de acuerdo a la regla de interpolación establecida, la precisión a los niveles de cobertura del 0 al 5 es 1, la precisión para los niveles 6 y 7 es 0.75, y la precisión para los niveles 8, 9 y 10 es 0.27.

Precisión a los n Documentos Devueltos

Otra posibilidad, similar a la anterior, de cara a comparar dos conjuntos resultado ordenados, es mostrar la precisión obtenida no a determinados valores de cobertura, sino a un número determinado de documentos devueltos.

Precisión Media no Interpolada

Se trata de una medida que refleja el rendimiento del sistema sobre el conjunto de documentos relevantes, pero no sólo considerando el porcentaje de documentos relevantes recuperados, sino también el orden en que éstos han sido devueltos. Este valor es calculado como la media de las precisiones obtenidas después de que cada documento relevante es recuperado respecto al número de documentos relevantes existentes para esa consulta. De esta forma se premia a los sistemas que devuelven los documentos relevantes en posiciones superiores.

Retomando el ejemplo mostrado para la precisión a los 11 niveles estándar de cobertura, su precisión media no interpolada sería

$$\textit{Pr. no int.} \;=\; \frac{1+1+0{,}75+0{,}27}{4} \;=\; 0{,}76$$

Precisión Media de Documento

Similar a la precisión media no interpolada, en cuanto a que también se calculan, como paso intermedio, las precisiones obtenidas después de que cada documento relevante es recuperado.

Como ya hemos comentado, la precisión media no interpolada para una consulta determinada se calcula como la media de dichas precisiones respecto al número de documentos relevantes existentes para la consulta. Posteriormente, la precisión media no interpolada global se calculará, como es usual, como la media de las precisiones medias no interpoladas de cada consulta respecto al número de consultas.

Sin embargo, la precisión media de documento no tiene sentido a nivel de una única consulta determinada, sino que se calcula a nivel global para un conjunto de consultas. Para ello se promedian las sumas de todas las precisiones obtenidas, para todas las consultas empleadas, después de que cada documento relevante es recuperado. Asimismo esta media se hace respecto al número global de documentos relevantes existentes para todas las consultas, en lugar de respecto al número de consultas.

Supongamos que, además de la consulta que venimos usando como ejemplo en la precisión a los 11 niveles estándar de cobertura y en la precisión media no interpolada, tenemos otra consulta que cuenta con 2 documentos relevantes que son devueltos en posiciones 2^{a} y 4^{a}. Sus precisiones para cada documento devuelto serían:

- **1º documento relevante:** precisión 0.5 (1 documento relevante, 2 documentos recuperados)
- **2º documento relevante:** precisión 0.5 (2 documento relevantes, 4 documentos recuperados)

y la precisión media de documento se calcularía, a nivel global del conjunto de las dos consultas, como:

$$\textit{Pr. doc.} \;=\; \frac{\big(1+1+0{,}75+0{,}27\big) \;+\; \big(0{,}5+0{,}5\big)}{4 \;+\; 2} \;=\; 0{,}67$$

Precisión-R (R-*Precision*)

Es la precisión obtenida a los R documentos devueltos, donde R es el número de documentos relevantes para esa consulta.

Retomando de nuevo el ejemplo para la precisión a los 11 niveles estándar de cobertura, y dado que existían 4 documentos relevantes para la consulta, la precisión-R será la precisión a los 4 documentos devueltos, que era de 0.75.

2.5.2. Colecciones de Referencia

Hasta hace poco uno de los mayores problemas en el ámbito de la Recuperación de Información era la falta de colecciones de evaluación con suficiente entidad y de libre acceso, para que de este modo permitiesen una evaluación de los sistemas lo más completa posible y que dichos resultados fuesen comparables.

Una colección de evaluación está compuesta por tres elementos: los documentos, las consultas, y una lista de los documentos de la colección que son relevantes para cada consulta.

La elección de una colección adecuada es de la suma importancia a la hora de evaluar nuestro sistema, ya que únicamente así tendremos la convicción de que los resultados obtenidos son fiables y representativos. La calidad de una colección viene dada por diversos aspectos:

- *Su disponibilidad para la comunidad científica.* El libre acceso a una colección promueve su utilización por otros investigadores, facilitando la comparación de resultados.
- *El tamaño de la colección.* Cuanto mayor sea el repositorio de documentos y el número de consultas a utilizar, más se ajustarán los resultados obtenidos al comportamiento real del sistema [105].
- *La calidad de las consultas.* Dicha calidad depende de su variedad, de la diversidad de construcciones empleadas, y de si dichas consultas se corresponden o no a necesidades de información realísticas.
- *La calidad de los documentos.* Viene dada por la variedad de los mismos y por su realismo en cuanto a que no hayan sido sometidos a ningún tipo de tratamiento especial.

Colecciones como *Cranfield* [118], o CACM e ISI [76], por ejemplo, se mostraban inadecuadas a las necesidades para una evaluación fiable del sistema debido a su pequeño tamaño (número de documentos: 1400, 3204 y 1460, respectivamente; número de consultas: 22, 52, 35+41). En años recientes esta situación ha cambiado gracias al trabajo de la *Text REtrieval Conference (TREC)* y del *Cross-Language Evaluation Forum (CLEF).*

Text REtrieval Conference (TREC)

La situación inicial de falta de colecciones de evaluación estándar dio un giro tras la celebración, en 1992, del primer *Text REtrieval Conference (TREC)* [1], congreso de carácter anual organizado por el *National Institute of Standards and Technology (NIST)* y la *Information Technology Office* de la *Defense Advanced Research Projects Agency (DARPA).* El objetivo perseguido por la organización del TREC es el de apoyar la investigación en el campo de la Recuperación de Información. Para ello:

- Facilita la infraestructura, herramientas y metodologías necesarias para la evaluación a gran escala de sistemas de Recuperación de Información.

- Promueve la comunicación entre industria, gobiernos e investigadores, facilitando de este modo la transferencia tecnológica.

Además de la recuperación *ad hoc*, TREC da cabida a otros campos de aplicación dentro de sus diferentes secciones, denominadas *tracks*, tales como enrutamiento (*routing*), filtrado (*filtering*), recuperación sobre transcripciones de documentos hablados, etc. Algunas de estas secciones varían según la edición.

Las colecciones de documentos empleadas en TREC son del orden de decenas, o incluso centenares, de miles de documentos. En su mayor parte se trata de artículos periodísticos procedentes de periódicos o agencias de noticias. Tal es el caso, por ejemplo, de la colección de *Los Angeles Times*, formada por artículos publicados en dicho periódico durante el año 1994 y que suponen 131896 artículos (475 MB) con una longitud media de 527 palabras por documento.

Junto con las colecciones de documentos, TREC suministra una serie de requerimientos o necesidades de información, denominados *topics*, en torno a 50 por edición. El proceso de convertir dichos *topics* en *consultas* efectivas debe ser llevado a cabo por el propio sistema. Los participantes deben enviar a la organización, dentro de un plazo dado, los resultados devueltos por sus sistemas para dichos *topics*.

El mayor problema surge a la hora de identificar los documentos relevantes para cada *topic* debido al gran número de documentos y consultas existentes. Para esta tarea se emplea una técnica denominada *pooling* [260], consistente en que, para cada *topic*, se toman los K primeros documentos devueltos para ese *topic* por cada uno de los sistemas participantes —generalmente los K=100 primeros. Dichos documentos son luego revisados por especialistas, que son quienes establecen la relevancia o no de cada uno de ellos. En lo que se refiere al concepto de *relevancia*, la organización de TREC optó por el siguiente criterio:[13]

> *"Si estuviera redactando un informe sobre el tema del* topic *en cuestión y pudiese usar para dicho informe la información contenida en el documento examinado, entonces dicho documento será considerado relevante"*

Cross-Language Evaluation Forum (CLEF)

A pesar de su inestimable contribución, las diferentes ediciones de TREC se han centrado en el inglés, salvo contadas excepciones, por lo que la investigación en sistemas de Recuperación de Información en otros idiomas seguía encontrándose con los mismos problemas. Tal era el caso del español, ya que las colecciones empleadas en su sección dedicada al español de las ediciones TREC-4 y TREC-5, y formada por artículos de noticias escritos en español (de México), no están ya disponibles.

Inicialmente en nuestras investigaciones se empleó una colección de evaluación propia [21], creada recopilando artículos periodísticos en español peninsular que cubrían el año 2000. El tamaño final de la colección era de 21.899 documentos, con una longitud media de 447 palabras. Por otra parte, el conjunto de consultas empleado era limitado, 14 consultas, con una longitud media de 7.85 palabras, 4.36 de ellas palabras con contenido.

El conjunto de documentos relevantes para cada consulta fue creado siguiendo la filosofía de *pooling* del TREC. Para cada una de las diferentes técnicas de normalización empleadas, y para cada uno de los diferentes motores de indexación empleados, se tomaron los 100 primeros documentos devueltos. Dichos documentos fueron examinados manualmente para juzgar su relevancia.

[13] Véase `http://trec.nist.gov/data/reljudge_eng.html`

```
<DOC>
<DOCNO>EFE19940101-00002</DOCNO>
<DOCID>EFE19940101-00002</DOCID>
<DATE>19940101</DATE>
<TIME>00.34</TIME>
<SCATE>VAR</SCATE>
<FICHEROS>94F.JPG</FICHEROS>
<DESTINO>ICX MUN EXG</DESTINO>
<CATEGORY>VARIOS</CATEGORY>
<CLAVE>DP2404</CLAVE>
<NUM>100</NUM>
<PRIORIDAD>U</PRIORIDAD>
<TITLE>   IBM-WATSON
           FALLECIO HIJO FUNDADOR EMPRESA DE COMPUTADORAS
</TITLE>
<TEXT>    Nueva York, 31 dic (EFE).- Thomas Watson junior, hijo del fundador
 de International Business Machines Corp. (IBM), falleció hoy,
 viernes, en un hospital del estado de Connecticut a los 79 años de
 edad, informó un portavoz de la empresa.
    Watson falleció en el hospital Greenwich a consecuencia de
 complicaciones tras sufrir un ataque cardíaco, añadió la fuente.
    El difunto heredó de su padre una empresa dedicada principalmente
 a la fabricación de máquinas de escribir y la transformó en una
 compañía líder e innovadora en el mercado de las computadoras. EFE
    PD/FMR
    01/01/00-34/94
</TEXT>
</DOC>
```

Figura 2.5: Documento de ejemplo: documento EFE19940101-00002

Sin embargo, aunque dicha colección servía a los propósitos de evaluación deseados, no era lo suficientemente amplia, sobre todo en lo que respecta al conjunto de consultas empleado. Además, al tratarse de una colección propia, no permitía la comparación de resultados con otros investigadores.

La incorporación en la segunda edición del *Cross-Language Evaluation Forum (CLEF-2001)* [2] de una colección para el español peninsular, cambió esta situación. CLEF es un congreso de corte similar al TREC, pero dedicado a las lenguas europeas, tanto en tareas monolingües como multilingües. Si en su primera edición, en el año 2000, los idiomas disponibles eran inglés, francés, alemán e italiano, las colecciones disponibles se han ido ampliando a español, portugués, finlandés, ruso, búlgaro y húngaro.

2.5.3. Metodología de Evaluación Empleada

En los experimentos recogidos en este trabajo se han empleado las colecciones de evaluación para el español de las ediciones del 2001 [173], 2002 [174] y 2003 [175] del CLEF.

Respecto a las colecciones de documentos, en CLEF 2001 y 2002 se empleó el mismo corpus, denominado EFE 1994 y formado por teletipos de la agencia española de noticias EFE[14] correspondientes al año 1994. Los documentos se encuentran formateados en SGML, tal como se aprecia en la figura 2.5, donde recogemos uno de ellos a modo de ejemplo. Este corpus inicial

[14] http://www.efe.es

colección	*tamaño* (MB)	*#docs.*	*long.*
EFE 1994	509	215738	317.64
EFE 1995	577	238307	325.33
EFE 1994+1995	1086	454045	321.67

Tabla 2.1: Colecciones de evaluación: composición de los corpus de documentos

```
<top>
<num> C044 </num>
<ES-title> Indurain gana el Tour </ES-title>
<ES-desc> Reacciones al cuarto Tour de Francia ganado por Miguel Indurain.
</ES-desc>
<ES-narr> Los documentos relevantes comentan las reacciones a la cuarta
victoria consecutiva de Miguel Indurain en el Tour de Francia. Los
documentos que discuten la relevancia de Indurain en el ciclismo mundial
después de esta victoria también son relevantes. </ES-narr>
</top>
```

Figura 2.6: *Topic* de ejemplo: *topic* número 44

fue ampliado en el CLEF 2003 con una segunda colección, denominada EFE 1995, formada por teletipos de EFE del año 1995. La composición de ambos corpus —tamaño de la colección, número de documentos, y longitud media—, juntos y por separado, se recoge en la tabla 2.1.[15]

En lo que respecta a los *topics* empleados, fueron 50 en 2001 (números 41 a 90), de nuevo 50 en 2002 (91 a 140), y finalmente 60 en 2003 (141 a 200). Los *topics* están formados, tal como se aprecia en la figura 2.6, por tres campos: *título* (*title*), un breve título como su nombre indica; *descripción* (*description*), una somera frase de descripción; y *narrativa* (*narrative*), un pequeño texto especificando los criterios que utilizarán los revisores para establecer la relevancia de un documento respecto a la consulta.

Sin embargo, los experimentos recogidos en este trabajo no fueron realizados directamente con este conjunto inicial de *topics*. En primer lugar se eliminaron aquéllos con un número de documentos relevantes menor de 6. La razón para ello estriba en que, cuando el número de documentos relevantes es muy pequeño, un cambio en la posición de uno o dos documentos devueltos puede acarrear cambios muy marcados en los resultados obtenidos para dicha consulta, distorsionando los resultados globales obtenidos para el conjunto total [106].

Dado el elevado número de consultas disponibles —100 para el corpus de documentos del CLEF 2001 y 2002, y 60 para CLEF 2003—, decidimos crear 3 corpus:

- **CLEF 2001-02·A:** corpus de entrenamiento y estimación de parámetros [42]. Está formado por los *topics* impares de CLEF 2001 y 2002. En caso de no ser necesaria tal fase de entrenamiento y estimación se emplearía igualmente a modo de corpus de evaluación. Se optó por combinar los *topics* de ambas ediciones para homogeneizar en lo posible las colecciones empleadas, en previsión de que hubiese diferencias importantes entre las consultas de ambas ediciones[16]. La elección de las consultas correspondientes al corpus EFE 1994 para crear un corpus de entrenamiento, en lugar del corpus ampliado de

[15] Debemos llamar la atención sobre el hecho de que, al tratarse de teletipos, de estilo poco cuidado y escritos con escasa atención, éstos contienen numerosos errores ortográficos [74].

[16] Debe tenerse en cuenta a este respecto que la edición 2001 era la primera en la que se empleaba el español, por lo que el equipo encargado de dicho idioma gozaba de poca experiencia al respecto.

corpus	*colec. docs.*	*#topics*	*long.*	
			cortas	*largas*
CLEF 2001-02·A	EFE 1994	46	19.28	55.23
CLEF 2001-02·B	EFE 1994	45	20.24	60.22
CLEF 2003	EFE 1994+1995	47	21.31	59.14

Tabla 2.2: Colecciones de evaluación: composición final de los corpus

CLEF 2003, viene dada por el menor número de consultas disponibles para éste último, 60 —47 tras eliminar las de insuficientes documentos relevantes—, lo que hace más difícil la división del mismo en un corpus de entrenamiento y un corpus de evaluación. Dicha afirmación se basa en los resultados obtenidos por Voorhees [262], en base a los cuales podemos considerar que 25 consultas es el número de consultas mínimo necesario para eliminar las posibles perturbaciones debidos a errores en la emisión de juicios de relevancia, lo que no permitiría una división adecuada de las consultas del corpus CLEF 2003.

- **CLEF 2001-02·B:** corpus de evaluación. Formado por los *topics* pares de las ediciones del 2001 y 2002.

- **CLEF 2003:** corpus de evaluación. Formado por los *topics* de la edición del 2003.

Asimismo se emplearon dos tipos de consultas durante la evaluación, las denominadas consultas *cortas*, generadas a partir de los campos *título* y *descripción*, y las denominadas consultas *largas*, que emplean la totalidad de los campos del *topic*. De esta forma podemos comprobar el comportamiento del sistema ante ambos tipos de consultas, siendo las cortas más próximas a las empleadas en sistemas comerciales [105][17]. Por otra parte, en el caso de las consultas largas, se ha primado la información aportada por el campo *título*, doblando su relevancia respecto a los otros dos campos, ya que es el campo que concentra la semántica básica de la consulta.

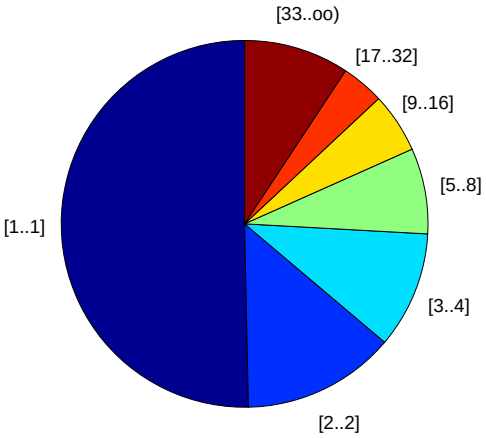
Las estadísticas de los corpus resultantes, eliminadas ya las consultas con menos de 6 documentos relevantes, se recogen en la tabla 2.2. Éstas incluyen: colección de documentos empleada, número de *topics* empleados, y longitud media de consultas, tanto cortas como largas.

En nuestros experimentos emplearemos una aproximación basada en *stemming* como línea base contra la que comparar inicialmente las diferentes aproximaciones propuestas como técnicas de normalización. Para ello emplearemos la versión para español del *stemmer* `Snowball` [3], de amplio uso por la comunidad científica, y desarrollado por el propio Porter empleando su ya clásico algoritmo de *stemming* [179]. Previamente las *stopwords* del texto han sido eliminadas en base a la lista de *stopwords* para español proporcionada con el motor de indexación empleado[18] (ver apartado 2.5.4). A mayores, en el caso de las consultas, se ha empleado una lista de *metastopwords* confeccionada tras examinar el conjunto de consultas empleado. Ambas listas de *stopwords* se recogen en el apéndice B.

Por otra parte, durante el proceso de indexación también fueron desechados aquellos términos de uso marginal en la colección, aquéllos con una frecuencia de documento (df) —número de documentos en los que aparecen— por debajo de un umbral dado. Su eliminación pretende evitar un consumo innecesario de recursos de almacenamiento y procesamiento, y se justifica en dos puntos. En primer lugar, que tras un examen de dichos términos se pudo apreciar que

[17] De hecho, la posición oficial dentro de la competición asociada a la celebración de cada edición de CLEF viene dada por los resultados obtenidos para una ejecución empleando las consultas que nosotros denominamos *cortas*.

[18] `ftp://ftp.cs.cornell.edu/pub/smart/spanish.stop`



df	*#stems*	*%stems*
[1..1]	177305	50.33 %
[2..2]	47530	13.49 %
[3..4]	36275	10.30 %
[5..8]	26593	7.55 %
[9..16]	18621	5.29 %
[17..32]	13276	3.77 %
[33..∞)	32665	9.27 %
Total	352265	100 %

Figura 2.7: Distribución de *stems* de la colección por frecuencia de documento (*df*)

una altísima proporción de los mismos correspondían a errores ortográficos[19]. En segundo lugar, que aún tratándose algunos de términos válidos, su altísimo grado de especifidad hace muy improbable su utilización dentro del vocabulario del usuario, por lo que rara vez serán utilizados en consultas [131], y en consecuencia pueden ser despreciados dada su escasa contribución [42].

Para fijar dicho umbral se estudió la distribución de los términos —en este caso *stems*— en el conjunto de ambas colecciones de documentos, EFE 1994 y EFE 1995. Los resultados obtenidos se recogen en la figura 2.7. En su parte derecha se muestra, para cada rango considerado (columna *df*), el número de términos, tanto absoluto (*#stems*) como relativo (*%stems*), cuya frecuencia de documento se encuentra dentro de dicho rango. Asimismo, en la parte izquierda de la figura se muestra gráficamente dicha distribución mediante un diagrama de sectores. Se puede apreciar que aproximadamente un 75 % de las entradas corresponden a términos con *df* menor que 5[20]. En base a ello, se optó por eliminar de los índices aquellas entradas correspondientes a términos con una frecuencia de documento menor que 5.

Las medidas de evaluación empleadas incluyen: número de documentos devueltos, número de documentos relevantes esperados, número de documentos relevantes devueltos, precisión media no interpolada, precisión media de documentos, precisión-R, precisión a los 11 niveles estándar de cobertura, y precisión a los n documentos devueltos. Los valores correspondientes son obtenidos mediante la herramienta software `trec_eval` [41], de uso ampliamente difundido en este tipo de cometidos.

Finalmente debemos puntualizar que, si bien los experimentos recogidos en este trabajo han sido realizados siguiendo el procedimiento establecido por CLEF, éstos no pueden ser considerados *oficiales*, ya que para ello la evaluación debería haber sido llevada a cabo por la propia organización de CLEF.

2.5.4. El Motor de Indexación Empleado: SMART

El proyecto SMART [199, 204, 40], desarrollado en la universidad de Cornell, es la implementación por excelencia del modelo vectorial, y ha sido pionero, en muchos aspectos, en el campo de la Recuperación de Información. Se trata, de hecho, de un sistema diseñado para la investigación en IR. Nuevos esquemas de cálculo de pesos [200], expansión de consultas y realimentación por relevancia [108, 196, 201, 44], normalización de la longitud de los

[19] De nuevo llamamos la atención sobre el gran número de errores ortográficos presentes en el corpus CLEF, ya apuntado por Figuerola et al en [74]

[20] Lo que concuerda, a su vez, con lo establecido por la ley de Zipf [272].

documentos [219, 218], son algunos de los campos explorados empleando SMART, sin dejar de lado su notable contribución, directa o indirecta, a TREC [43, 42, 44, 45] y a CLEF [211, 212].

La versión del software empleada es la 11 [4], con un esquema de pesos `atn·ntc` [211, 212] —`atn` para los términos de los documentos y `ntc` para las consultas. En SMART, los esquemas de pesos se denotan mediante una tripla de letras [200] donde la primera de ellas indica la componente que refleja la frecuencia del término en el documento, la segunda la componente correspondiente a la frecuencia dentro de la colección, y la tercera la componente de normalización respecto a la longitud del vector/documento. Las componentes utilizadas en nuestro caso son, para un término t_i en un documento d_j:

Componente de la frecuencia del término en el documento

`n`	tf_{ij}	se emplea la frecuencia pura del término t_i en el documento d_j (tf_{ij})
`a`	$0{,}5 + 0{,}5 \frac{tf_{ij}}{\max_j \ tf_{ij}}$	la frecuencia del término (tf_{ij}) normalizada respecto a la frecuencia máxima en dicho documento, y aumentada luego para que los valores finales obtenidos se sitúen entre 0.5 y 1

Componente de la frecuencia del término en la colección

`t`	$log \frac{N}{n_i}$	la componente de frecuencia en el documento se multiplica por la frecuencia inversa de documento del término (idf_i), calculada como el logaritmo del cociente del número de documentos en la colección (N) partido el número de ellos que contienen el término en cuestión (n_i)

Componente de normalización

`n`	1	no hay modificación; el resultado de multiplicar los anteriores factores componente se mantiene (ya que se multiplica por 1)
`c`	$\frac{1}{\sqrt{\sum_{i=1}^{t} {w^2}_{ij}}}$	normalización del coseno; el resultado de multiplicar los anteriores factores componente se mantiene se divide por la norma del vector

Capítulo 3

Introducción al Procesamiento del Lenguaje Natural

3.1. El Procesamiento del Lenguaje Natural

El lenguaje es uno de los aspectos fundamentales no sólo del comportamiento humano, sino de su propia naturaleza. En su forma escrita nos permite guardar un registro del conocimiento que se transmite de generación en generación, y en su forma hablada constituye el principal medio de comunicación en nuestro día a día.

El *Procesamiento del Lenguaje Natural* (NLP, *Natural Language Processing*) es la rama de las ciencias computacionales encargada del diseño e implementación de los elementos software y hardware necesarios para el tratamiento computacional del *lenguaje natural*, entendiendo como tal todo lenguaje humano, en contraposición a los *lenguajes formales* [146] propios del ámbito lógico, matemático, o computacional [110]. El objetivo último que se persigue es el de la comprensión del lenguaje humano por parte de la computadora. La consecución de un objetivo tan ambicioso, del que todavía se está muy lejos, supondría una auténtica revolución. Por una parte, los ordenadores podrían tener por fin acceso al conocimiento humano, y por otra, una nueva generación de interfaces, en lenguaje natural, facilitaría en grado sumo la accesibilidad a sistemas complejos.

3.1.1. Niveles de Análisis

Para cumplir su objetivo, un sistema de NLP necesitará hacer uso de una cantidad considerable de conocimiento acerca de la estructura del lenguaje. Este conocimiento se puede estructurar en niveles:

1. **Conocimiento morfológico**[1]**:** para determinar cómo son las palabras que constituyen el lenguaje y cómo éstas se forman a partir de unidades más pequeñas denominadas *morfemas*.

2. **Conocimiento sintáctico:** para determinar cómo se combinan las palabras para dar lugar a *sintagmas* y *frases*, así como el papel estructural que desempeña cada palabra y cada sintagma en la frase resultante.

3. **Conocimiento semántico:** para determinar el *significado* de cada palabra y cómo se construye el significado de una frase a partir de los significados de las palabras que la constituyen.

[1] También denominado conocimiento *léxico*.

4. **Conocimiento pragmático:** para determinar cómo se relaciona el lenguaje con los contextos en los que se usa.

Paralelamente a estos niveles de conocimiento se establecen cuatro niveles de análisis en los que se incluyen los diversos modelos computacionales y algoritmos para su tratamiento:

1. **Análisis morfológico**[2]**:** mediante el cual se determinan las palabras que integran un texto, así como su etiqueta morfosintáctica, utilizando para ello modelos computacionales de la morfología, basados generalmente en autómatas de estado finito, expresiones regulares, traductores de estado finito, modelos de Markov ocultos y n-gramas.

2. **Análisis sintáctico:** que realiza el agrupamiento de las palabras en sintagmas y frases mediante modelos computacionales como son las gramáticas independientes del contexto, las gramáticas lexicalizadas y las estructuras de rasgos.

3. **Análisis semántico:** mediante el cual se determina el significado de las frases de acuerdo con el significado de los sintagmas, palabras y morfemas que las forman, utilizando para ello modelos computacionales tales como la lógica de predicados de primer orden y las redes semánticas.

4. **Análisis pragmático:** que establece la identidad de las personas y objetos que aparecen en los textos, determina la estructura del discurso y gestiona el diálogo en un entorno conversacional.

En el caso del tratamiento del habla, existiría además un nivel previo de *reconocimiento del habla* y posiblemente un nivel posterior de *síntesis del habla*, los cuales harían uso de conocimiento fonético y fonológico.

3.1.2. Ambigüedad

A la hora de procesar un texto en lenguaje natural, el problema principal con el que nos hemos de enfrentar en los diferentes niveles de análisis es el de la *ambigüedad*.

A nivel morfológico, nos encontramos con que una palabra puede recibir diversas etiquetas. Por ejemplo, la palabra *sobre* puede ser un sustantivo masculino singular, una preposición, o la primera o tercera persona del presente de subjuntivo del verbo *sobrar*. En ciertos contextos la tarea de determinar la etiqueta correcta puede ser relativamente fácil, pero en frases como *"pon lo que sobre sobre el sobre"* la complejidad de este proceso es patente.

A nivel sintáctico, el hecho de que una frase sea ambigua se traduce en que es posible asociar dos o más estructuras sintagmáticas correctas a dicha frase. Tomemos el ejemplo clásico de la frase *"Juan vio a un hombre con un telescopio en una colina"*. Diferentes ubicaciones de las subestructuras correspondientes a los fragmentos *"con un telescopio"* y *"en una colina"* dan lugar a diferentes estructuras sintagmáticas de la frase, todas ellas correctas, y que se corresponden con los significados siguientes:

- Juan vio a un hombre que estaba en una colina y que tenía un telescopio;
- Juan estaba en una colina, desde donde vio a un hombre que tenía un telescopio;
- Juan estaba en una colina, desde donde miraba con un telescopio, a través del cual vio a un hombre.

[2] O análisis *léxico*.

A nivel semántico, nos encontramos con que una palabra puede tener diferentes significados o sentidos. Por ejemplo, la palabra *banda* puede referirse a:

- un grupo de personas;
- una tira de tela;
- los laterales de un barco;
- un conjunto de frecuencias del espectro radioeléctrico.

Como el significado de una frase se construye a partir de las aportaciones semánticas realizadas por las palabras que la componen, es preciso determinar en primer lugar el significado correcto de cada una de ellas. Sin embargo, el significado de una frase puede ser ambiguo incluso aun cuando las palabras que lo componen no lo son. Por ejemplo, la frase *"todos los alumnos de la facultad hablan dos idiomas"* admite dos interpretaciones distintas:

- Existen dos idiomas L y L' tales que todos los alumnos de la facultad los hablan.
- Cada uno de los alumnos de la facultad habla un par de idiomas, pero dos estudiantes distintos pueden hablar idiomas distintos.

A su vez las ambigüedades pueden ser locales o globales. Una *ambigüedad local* es aquella que surge en un momento del análisis pero que es eliminada posteriormente al analizar una porción mayor del texto. Una *ambigüedad global* es aquella que permanece una vez terminado de analizar todo el texto.

Llegados a este punto es interesante destacar que los distinto niveles de análisis no tienen porqué ser totalmente independientes entre sí, ya que, por ejemplo, y tal como hemos visto, el análisis léxico puede ofrecer diferentes etiquetas para una palabra dada, dejando que sean el analizador sintáctico e incluso el semántico los encargados de determinar aquélla más conveniente.

3.1.3. Dos Clases de Aproximaciones: Simbólica y Estadística

Es posible distinguir dos grandes tipos de aproximaciones a la hora de enfrentarse al problema del Procesamiento del Lenguaje Natural: aquéllas de carácter *simbólico*, y aquéllas de tipo empírico o *estadístico*. Hoy en día, sin embargo, parece claro que una aproximación híbrida es la más adecuada.

Aproximaciones Simbólicas

Desde sus inicios en los años 50, el Procesamiento del Lenguaje Natural ha sido abordado mediante diferentes técnicas de carácter simbólico basadas en el empleo de reglas —u otras formas de representación similares— que codifican explícitamente nuestro conocimiento del dominio, y que han sido desarrolladas por expertos humanos en el ámbito de aplicación [58, 110]. Se trata, pues, de aproximaciones *basadas en el conocimiento*, próximas a los modelos tradicionales de Inteligencia Artificial, y que precisan de una fase previa de estudio y análisis del dominio para que, de este modo, los expertos puedan identificar y describir mediante reglas las regularidades del mismo. Desde un punto de vista metodológico, se trata de una aproximación descendente, ya que intentamos imponer sobre los textos los modelos que nosotros hemos desarrollado.

Aproximaciones Estadísticas

Durante la última década, y gracias al incremento de la potencia y velocidad de los ordenadores, han cobrado especial protagonismo las aproximaciones denominadas empíricas o estadísticas, fundamentadas en el análisis y descripción estadística del lenguaje a partir de grandes corpus de texto [141, 110]. Se opta, en este caso, por un punto de vista cuantitativo, donde las diferentes posibilidades fruto de la ambigüedad lingüística son evaluadas en función de sus probabilidades asociadas empleando técnicas estadísticas. Al contrario que antes, nos encontramos ante aproximaciones ascendentes, ya que el modelo es desarrollado partiendo de los propios textos. Para ello se precisa de textos de entrenamiento sobre los que aplicar técnicas de tipo estadístico para la identificación de los patrones y asociaciones presentes en los mismos, siendo capaces incluso de capturar, en ocasiones, aspectos implícitos en el modelo que el experto es incapaz de ver.

3.2. Nivel Morfológico

En este y subsiguientes apartados abordaremos en mayor detalle los diferentes niveles de procesamiento lingüístico.

Todo lenguaje humano, sea hablado o escrito, se compone de palabras. De este modo podemos considerar a las palabras como los "ladrillos" del lenguaje. Es lógico, por tanto, empezar nuestro análisis por el procesamiento de las palabras que forman un texto. De este modo, abordaremos en nuestro primer punto el *nivel morfológico*, también referido en ocasiones como *nivel léxico*.

La *morfología* es la parte de la gramática que se ocupa del estudio de la estructura de las palabras y de sus mecanismos de formación. Las palabras están formadas por unidades mínimas de significado denominadas *morfemas* [135], los cuales podemos clasificar en dos clases: morfemas léxicos y morfemas gramaticales.

Los *morfemas léxicos*, comúnmente denominados *lexemas* o *raíces*, son los elementos que aportan el significado principal a la palabra (p.ej., *hablar*). Por el contrario, los *morfemas gramaticales*, comúnmente denominados *afijos* o, por extensión, simplemente *morfemas*, poseen únicamente significado gramatical, y nos permiten modificar el significado básico del lexema (p.ej., *hablases*).

Conforme a su posición, los afijos se clasifican en *prefijos*, antepuestos al lexema (p.ej., *innecesario*), *sufijos*, postpuestos al lexema (p.ej., *hablador*), e *infijos*, elementos que aparecen intercalados en el interior de la estructura de una palabra (p.ej., *humareda*). Desde el punto de vista de cómo éstos alteran el significado del lexema, los afijos se clasifican en flexivos y derivativos. Los *afijos flexivos* representan conceptos gramaticales tales como género y número (p.ej., *habladoras*), persona, modo, tiempo y aspecto (p.ej., *hablases*). Los *afijos derivativos*, por su parte, producen un cambio semántico respecto al lexema base, y frecuentemente también un cambio de categoría sintáctica (p.ej., *hablador*).

A la hora de estudiar las técnicas y herramientas desarrolladas a nivel morfológico en el área del Procesamiento del Lenguaje Natural nos centraremos en dos aspectos: el análisis morfológico, y la etiquetación.

3.2.1. Análisis Morfológico

El análisis morfológico de una palabra consiste en que, dada una forma de una palabra, obtener los diferentes rasgos morfológicos asociados a la misma [224], tales como su categoría gramatical, género, número, persona, etc. Por ejemplo, dada la palabra *gatos*, un analizador morfológico nos indicaría que se trata de una forma nominal masculina plural.

El análisis morfológico se encuentra íntimamente ligado a la denominada *morfología de dos niveles* [129], que considera las palabras como una correspondencia entre el *nivel léxico*, que representa la concatenación de los morfemas que constituyen una palabra, y el *nivel superficial*, que representa la forma escrita real de una palabra. De esta forma, el análisis morfológico de una palabra se lleva a cabo mediante un conjunto de reglas que hacen corresponder secuencias de letras del nivel superficial a secuencias de morfemas y rasgos morfológicos del nivel léxico. Por ejemplo, la forma superficial *gatos* se convertiría en la forma léxica *gat +Sust +Masc +Sing* mediante la cual se indica que dicha palabra es un sustantivo masculino singular.

Para realizar la correspondencia entre los niveles superficial y léxico se necesita disponer de una información mínima [121]:

1. Un **lexicón** que recoja las raíces y afijos a emplear, junto con la información básica acerca de los mismos. Por ejemplo, si se trata de una raíz nominal, verbal, etc.

2. Un modelo de ordenación para la aplicación de los morfemas, y que se conoce como **morfotácticas**. Por ejemplo, los morfemas flexivos de número se postponen al sustantivo.

3. Una serie de **reglas ortográficas** que modelen los cambios que se producen en la palabra durante la adjunción de los morfemas. Por ejemplo, en inglés, un sustantivo terminado en consonante seguido por *-y* cambia ésta por *-ie* al concatenar el morfema flexivo plural *-s*, como en el caso de *city/cities* (ciudad/ciudades).

A la hora de la implementación de esta correspondencia se utilizan traductores de estado finito [121] que se encargan de traducir un conjunto de símbolos en otro. Para esta tarea de análisis los traductores son utilizados habitualmente en cascada: primero se utiliza un traductor que reconoce el morfema léxico de las palabras y lo convierte en su forma regular, al tiempo que indica su categoría gramatical; posteriormente, se aplican traductores especializados en el reconocimiento de morfemas específicos de género, número, tiempo, persona, etc., que son transformados en rasgos morfológicos. La potencia de los traductores de estado finito viene determinada por el hecho de que la misma cascada, con las mismas secuencias de estados, puede ser utilizada tanto para obtener la forma léxica a partir de la forma superficial como para generar la forma superficial a partir de la forma léxica.

3.2.2. Etiquetación

Los problemas surgen cuando, dado un texto a analizar, nos encontramos con ambigüedades morfológicas en el mismo. Un analizador morfológico únicamente conoce la forma de la palabra, por lo que no cuenta con información suficiente para analizar correctamente cada palabra en caso de ambigüedad, ya que para ello es necesario acceder al contexto de la palabra. En una frase como *"pon lo que sobre sobre el sobre"* únicamente nos podría indicar que existen tres opciones posibles para cada aparición de la palabra *"sobre"*: sustantivo, preposición y verbo.

Al proceso de desambiguación en función del cual a cada palabra del texto le es asignado su análisis morfológico correcto —codificado por medio de una *etiqueta* (*tag*)— se le denomina *etiquetación* (*tagging*) [39], y constituye el primer paso de cara a la realización de análisis más profundos del texto, bien de carácter sintáctico o semántico. Las herramientas que implementan este proceso se denominan *etiquetadores* (*taggers*).

Fuentes de Información Relevantes para la Etiquetación

A la hora de decidir cuál es la etiqueta correcta de una palabra existen, esencialmente, dos fuentes de información [141]:

1. La primera de ellas consiste en examinar su contexto, es decir, las etiquetas de las palabras circundantes. Aunque esas palabras podrían ser también ambiguas, el hecho de observar secuencias de varias etiquetas nos puede dar una idea de cuáles son comunes y cuáles no lo son. Por ejemplo, en inglés, una secuencia como artículo-adjetivo-sustantivo es muy común, mientras que otras secuencias como artículo-adjetivo-verbo resultan muy poco frecuentes o prácticamente imposibles. Por tanto, si hubiera que elegir entre sustantivo o verbo para etiquetar la palabra `play` en la frase `a new play`, obviamente optaríamos por sustantivo.

   Este tipo de estructuras constituyen la fuente de información más directa para el proceso de etiquetación, aunque por sí misma no resulte demasiado exitosa: uno de los primeros etiquetadores basado en reglas deterministas que utilizaba este tipo de patrones sintagmáticos etiquetaba correctamente sólo el 77 % de las palabras [90]. Una de las razones de este rendimiento tan bajo es que en inglés las palabras que pueden tener varias etiquetas son muy numerosas, debido sobre todo a procesos productivos como el que permite a casi todos los sustantivos que podamos tener en el diccionario transformarse y funcionar como verbos, con la consiguiente pérdida de la información restrictiva que es necesaria para el proceso de etiquetación.

2. La segunda fuente de información consiste en el simple conocimiento de la palabra concreta, que puede proporcionarnos datos muy valiosos acerca de la etiqueta correcta. Por ejemplo, existen palabras que, aunque puedan ser usadas como verbos, su aparición es mucho más probable cuando funcionan como sustantivos. La utilidad de esta información fue demostrada de manera concluyente por Charniak, quien puso de manifiesto que un etiquetador que simplemente asigne la etiqueta más común a cada palabra puede alcanzar un índice de acierto del 90 % [52].

   La información léxica de las palabras resulta tan útil porque la distribución de uso de una palabra a lo largo de todas sus posibles etiquetas suele ser rara. Incluso las palabras con un gran número de etiquetas aparecen típicamente con un único uso o etiqueta particular.

Consecuentemente, la distribución de uso de las palabras proporciona una información adicional de gran valor, y es por ello por lo que parece lógico esperar que las aproximaciones estadísticas al proceso de etiquetación den mejores resultados que las aproximaciones basadas en reglas deterministas. En éstas últimas, uno sólo puede decir que una palabra puede o no puede ser un verbo, por ejemplo, existiendo la tentación de desechar la posibilidad de que sea un verbo cuando ésta es muy rara, creyendo que esto aumentará el rendimiento global. Por el contrario, en una aproximación estadística se puede decir *a priori* que una palabra tiene una alta probabilidad de ser un sustantivo, pero también que existe una posibilidad, por remota que sea, de ser un verbo o incluso cualquier otra etiqueta. A día de hoy, los etiquetadores modernos utilizan de alguna manera una combinación de la información sintagmática proporcionada por las secuencias de etiquetas y de la información léxica proporcionada por las palabras.

Rendimiento y Precisión de los Etiquetadores

Las cifras de rendimiento conocidas para los etiquetadores se encuentran casi siempre dentro del rango del 95 al 97 % de acierto[3]. Sin embargo, es importante señalar que estas cifras no son tan buenas como parecen, ya que implica que, en frases largas —caso de artículos periodísticos, por ejemplo—, un rendimiento del 95 % todavía supone que pueden aparecer entre una y dos palabras mal etiquetadas en cada frase. Además, estos errores no siempre se localizan en las categorías

[3]Habiéndose calculado sobre el conjunto de todas las palabras del texto. Algunos autores proporcionan la precisión sólo para los términos ambiguos, en cuyo caso las cifras serán menores.

más pobladas, tales como sustantivos, adjetivos o verbos, donde en principio parece más probable el encontrarse con palabras desconocidas, sino que muchas veces los errores aparecen asociados a las partículas que conectan los sintagmas entre sí, tales como preposiciones, conjunciones o relativos, con lo que pueden hacer que una frase tome un significado muy distinto del original.

Dejando ya de lado estas cuestiones, el rendimiento depende considerablemente de una serie de factores [141]:

- El tamaño del corpus de entrenamiento disponible. En general, a mayor disponibilidad de textos de entrenamiento, mayor y mejor será el conocimiento extraído y mejor será la etiquetación.
- El *juego de etiquetas* (*tag set*). Normalmente, cuanto más grande es el conjunto de etiquetas considerado, mayor será la ambigüedad potencial, con lo que se agrava el problema de la dispersión de datos, y la tarea de etiquetación se vuelve más compleja.
- La diferencia entre, por un lado, el diccionario y el corpus de entrenamiento empleados, y por otro, el corpus de aplicación. Si los textos de entrenamiento y los textos que posteriormente se van a etiquetar proceden de la misma fuente —por ejemplo, textos de la misma época o estilo—, entonces la precisión obtenida será mayor. Sin embargo, si los textos de aplicación pertenecen a un periodo o género distintos —p.ej., textos científicos contra textos periodísticos—, entonces el rendimiento será menor.
- Las palabras desconocidas. Un caso especial del punto anterior es la cobertura del diccionario. La aparición de palabras desconocidas puede degradar el rendimiento, situación común, por ejemplo, al intentar etiquetar material procedente de algún dominio técnico.

Un cambio en cualquiera de estas cuatro condiciones puede producir un fuerte impacto en la precisión alcanzada por el etiquetador. Es importante señalar que estos factores son externos al proceso de etiquetación y al método elegido para realizar dicho proceso, siendo su efecto a menudo mucho mayor que la influencia ejercida por el propio método en sí.

Etiquetación Basada en Reglas

Los primeros etiquetadores abordaban el problema de la desambiguación mediante aproximaciones basadas en reglas empleando una arquitectura en dos etapas [100, 128]. En una primera fase se le asigna a cada palabra una lista de sus etiquetas potenciales en base a un diccionario. Es entonces cuando, en una segunda etapa, se aplican las reglas de desambiguación para identificar la etiqueta correcta.

El primer algoritmo para la asignación de etiquetas que se conoce estaba incorporado en el analizador sintáctico utilizado en el proyecto TDAP, implementado entre 1958 y 1969 en la Universidad de Pennsylvania [100]. Anteriormente, los sistemas de procesamiento del lenguaje natural utilizaban diccionarios con información morfológica de las palabras pero, que se sepa, no realizaban desambiguación de etiquetas. El sistema TDAP realizaba esta desambiguación mediante 14 reglas escritas a mano que eran ejecutadas en un orden basado en la frecuencia relativa de las etiquetas de cada palabra.

Poco después del TDAP surgió el sistema CGC de Klein y Simmons [128], con sus tres componentes: un lexicón, un analizador morfológico y un desambiguador por contexto. El pequeño diccionario de 1.500 palabras incluía aquellas palabras raras que no podían ser tratadas por el analizador morfológico, tales como sustantivos, adjetivos y verbos irregulares. El analizador morfológico utilizaba los sufijos flexivos y derivativos para asignar un conjunto de etiquetas a cada palabra. En ese momento entraban en acción un conjunto de 500 reglas

encargadas de seleccionar la etiqueta correcta, consultando para ello las islas de palabras contiguas no ambiguas. El juego de etiquetas constaba de 30 etiquetas.

Etiquetación Estocástica

Actualmente, uno de los modelos de etiquetación más extendidos, es el de la utilización de procedimientos estadísticos basados en la probabilidad de aparición conjunta de secuencias de n palabras o *n-gramas*. La matemática subyacente a los n-gramas fue propuesta por primera vez por Markov [143], quien utilizó bigramas y trigramas para predecir si la siguiente letra de una palabra rusa sería una vocal o una consonante. Shannon [216] aplicó posteriormente los n-gramas para calcular aproximaciones a las secuencias de palabras en inglés. A partir de los años 50, y gracias al trabajo de Shannon, los modelos de Markov fueron ampliamente utilizados para modelar secuencias de palabras. En décadas posteriores su uso decayó, principalmente debido a la argumentación de muchos lingüistas, entre ellos Chomsky [53], de que los modelos de Markov eran incapaces de modelar completamente el conocimiento gramatical humano. Los modelos de n-gramas resurgen en los años 70 al hacerse públicos los trabajos realizados en el centro de investigación Thomas J. Watson de IBM [115, 27] y en la Universidad de Carnegie Mellon [29], en los que se utilizan con éxito n-gramas para tareas de reconocimiento del habla.

En los años 70 se creó el corpus Lancaster-Oslo/Bergen (LOB) de inglés británico. Para su etiquetación se utilizó el etiquetador CLAWS [145], basado en un algoritmo probabilístico que puede considerarse una aproximación al enfoque actual basado en la utilización de modelos de Markov ocultos. El algoritmo utilizaba la probabilidad de aparición conjunta de dos etiquetas, pero en lugar de almacenar dicha probabilidad directamente, la clasificaba como *rara* ($P(\text{etiqueta} \mid \text{palabra}) < 0,01$), *infrecuente* ($0,01 \leq P(\text{etiqueta} \mid \text{palabra}) < 0,10$) o *normalmante frecuente* ($P(\text{etiqueta} \mid \text{palabra}) \geq 0,10$).

El etiquetador probabilístico de Church [55] seguía una aproximación muy cercana a la de los modelos de Markov ocultos, extendiendo la idea de CLAWS para asignar la probabilidad real a cada combinación palabra/etiqueta, utilizando el algoritmo de Viterbi [259, 75] para encontrar la mejor secuencia de etiquetas. Sin embargo, al igual que CLAWS, almacenaba la probabilidad de una etiqueta dada la palabra para calcular

$$P(\text{etiqueta} \mid \text{palabra}) \times P(\text{etiqueta} \mid n \text{ etiquetas anteriores})$$

en lugar de almacenar la probabilidad de una palabra dada la etiqueta, tal y como actualmente hacen los etiquetadores basados en modelos de Markov ocultos para calcular

$$P(\text{palabra} \mid \text{etiqueta}) \times P(\text{etiqueta} \mid n \text{ etiquetas anteriores})$$

Los etiquetadores posteriores ya introdujeron explícitamente la utilización de modelos de Markov ocultos. Tal es el caso del etiquetador TnT de Brants [37], y MrTagoo de Graña [83] que constituyen claros ejemplos de las herramientas recientes de alto rendimiento que utilizan modelos de Markov ocultos basados en n-gramas.

Antes de describir en qué consiste un *modelo de Markov oculto*, debemos describir en qué consiste un *modelo de Markov observable* [141]. Consideremos un sistema que en cada instante de tiempo se encuentra en un determinado estado. Dicho estado pertenece a un conjunto finito de estados Q. Regularmente, transcurrido un espacio de tiempo discreto, el sistema cambia de estado de acuerdo con un conjunto de probabilidades de transición asociadas a cada uno de los estados del modelo. Los instantes de tiempo asociados a cada cambio de estado se denotan como $t = 1, 2, \ldots, T$, y el estado actual en el instante de tiempo t se denota como q_t. En general, una descripción probabilística completa del sistema requeriría la especificación del estado actual,

así como de todos los estados precedentes. Sin embargo, las cadenas de Markov presentan dos características de suma importancia:

1. La *propiedad del horizonte limitado*, que permite truncar la dependencia probabilística del estado actual y considerar, no todos los estados precedentes, sino únicamente un subconjunto finito de ellos. Una cadena de Markov de orden n es la que utiliza n estados previos para predecir el siguiente estado. Por ejemplo, para el caso de las cadenas de Markov de tiempo discreto de primer orden tenemos que $P(q_t = j|q_{t-1} = i, q_{t-2} = k, \dots) = P(q_t = j|q_{t-1} = i)$, es decir, dependería únicamente del estado anterior; en caso de ser de segundo orden, de los dos estados anteriores, y así sucesivamente.

2. La *propiedad del tiempo estacionario*, que nos permite considerar sólo aquellos procesos en los cuales $P(q_t = j|q_{t-1} = i)$ es independiente del tiempo, lo que a su vez nos lleva a definir una matriz de probabilidades de transición independientes del tiempo $A = \{a_{ij}\}$, donde $\forall i, j; \quad 1 \leq i, j \leq N; \quad a_{ij} = P(q_t = j|q_{t-1} = i) = P(j|i)$ y se cumplen las restricciones estocásticas estándar: $a_{ij} \geq 0$ para todo i y j, y $\sum_{j=1}^{N} a_{ij} = 1$ para todo i. Adicionalmente, es necesario especificar el vector $\pi = \{\pi_i\}$ que almacena la probabilidad $\pi_i \geq 0$ que tiene cada uno de los estados de ser el estado inicial: $\forall i; 1 \leq i \leq N; \pi_i = P(q_1 = i)$.

A un proceso estocástico que satisface estas características se le puede llamar un *modelo de Markov observable*, porque su salida es el conjunto de estados por los que pasa en cada instante de tiempo, y cada uno de estos estados se corresponde con un suceso observable. Esta modelización puede resultar demasiado restrictiva a la hora de ser aplicada a problemas reales. A continuación extenderemos el concepto de modelos de Markov de tal manera que sea posible incluir aquellos casos en los cuales la observación es una función probabilística del estado. El modelo resultante, denominado *modelo de Markov oculto* (HMM, *Hidden Markov Model*), es un modelo doblemente estocástico, ya que uno de los procesos no se puede observar directamente (está oculto), y sólo se puede observar a través de otro conjunto de procesos estocásticos, los cuales producen la secuencia de observaciones. Un HMM se caracteriza por la 5-tupla (Q, V, π, A, B) donde:

1. $Q = \{1, 2, \dots, N\}$ es el conjunto de estados del modelo. Aunque los estados permanecen ocultos, para la mayoría de las aplicaciones prácticas se conocen a priori. Por ejemplo, para el caso de la etiquetación de palabras, cada etiqueta del juego de etiquetas utilizado sería un estado. Generalmente los estados están conectados de tal manera que cualquiera de ellos se puede alcanzar desde cualquier otro en un solo paso, aunque existen muchas otras posibilidades de interconexión. El estado actual en el instante de tiempo t se denota como q_t. El uso de instantes de tiempo es apropiado, por ejemplo, en la aplicación de los HMM al procesamiento de voz. No obstante, para el caso de la etiquetación de palabras, no hablaremos de los instantes de tiempo, sino de las posiciones de cada palabra dentro de la frase.

2. V es el conjunto de los distintos sucesos que se pueden observar en cada uno de los estados. Por tanto, cada uno de los símbolos individuales que un estado puede emitir se denota como $\{v_1, v_2, \dots, v_M\}$. En el caso de la etiquetación de palabras, M es el tamaño del diccionario y cada $v_k, 1 \leq k \leq M$, es una palabra distinta.

3. $\pi = \{\pi_i\}$, es la distribución de probabilidad del estado inicial, cumpliéndose que $\pi_i \geq 0$, $\forall i; \quad 1 \leq i \leq N; \quad \pi_i = P(q_1 = i)$, y $\sum_{i=1}^{N} \pi_i = 1$.

4. $A = \{a_{ij}\}$ es la distribución de probabilidad de las transiciones entre estados, esto es, $\forall i, j, t; \quad 1 \leq i \leq N, 1 \leq i \leq N, 1 \leq t \leq T; \quad a_{ij} = P(q_t = j|q_{t-1} = i) = P(j|i)$, cumpliéndose que $a_i, j \geq 0$ y que $\sum_{j=1}^{N} a_{ij} = 1$ para todo i.

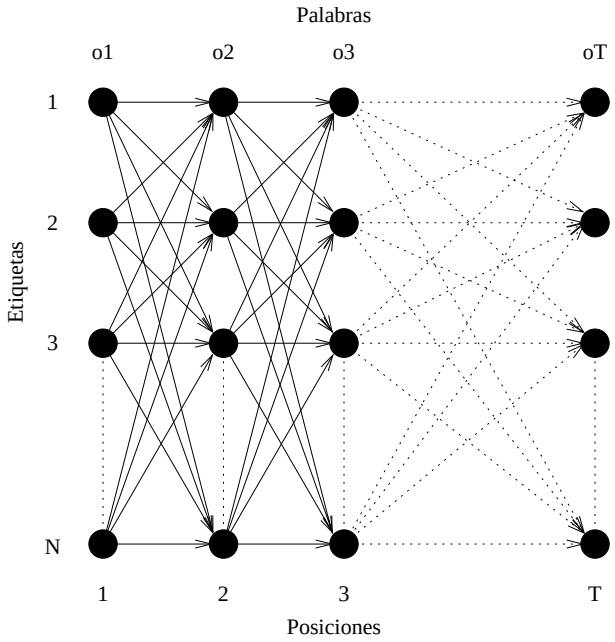


Figura 3.1: Enrejado genérico de T observaciones y N estados

5. $B = \{b_j(v_k)\}$ es la distribución de probabilidad de los sucesos observables, es decir, $\forall j, k, t; \quad 1 \leq j \leq N, 1 \leq k \leq M, 1 \leq t \leq T; \quad b_j(v_k) = P(o_t = v_k | q_t = j) = P(v_k|j)$, cumpliéndose que $\sum_{k=1}^{M} b_j(v_k) = 1$ para todo j. Este conjunto de probabilidades se conoce también con el nombre de *conjunto de probabilidades de emisión*.

Los parámetros del modelo —las probabilidades de transición y las probabilidades de salida de los estados— son estimados mediante un proceso de entrenamiento a partir de un corpus previamente desambiguado manualmente a tal efecto [37].

En base a dicho modelo, y dada una secuencia de observaciones (palabras) $O = (o_1, o_2, \ldots, o_T)$, $o_i \in V$, queremos determinar la secuencia de estados $S = (q_1, q_2, \ldots, q_T)$ óptima, es decir, aquélla que mejor *explica* la secuencia de observaciones. De una forma más sencilla, dada una secuencia de palabras O a etiquetar, queremos determinar la secuencia de etiquetas S más probable. Para ello se genera el *enrejado* o *diagrama de Trellis* correspondiente a dicha secuencia y modelo, tal como se aprecia en la figura 3.1, y que recoge todas las secuencias posibles de etiquetas para dicho secuencia de palabras. Sobre este enrejado se calculará la secuencia de etiquetas más probable empleando el algoritmo de Viterbi [259, 75]. De hecho, en el caso concreto de la etiquetación de palabras, los cálculos involucrados en el algoritmo de Viterbi se realizan frase por frase sobre enrejados simplificados como el de la figura 3.2, donde en cada posición no se consideran todos los estados posibles —o sea, todas la etiquetas del juego de etiquetas utilizado—, sino sólo las etiquetas candidatas que proponga el diccionario para cada palabra.

Etiquetación Basada en Transformaciones

Algunas de las hipótesis de funcionamiento de los modelos de Markov no se adaptan bien a las propiedades sintácticas de los lenguajes naturales, por lo que surge inmediatamente la idea de utilizar modelos más sofisticados que puedan establecer condiciones no sólo sobre las etiquetas precendentes, sino también sobre las palabras precedentes, o que permitan emplear contextos

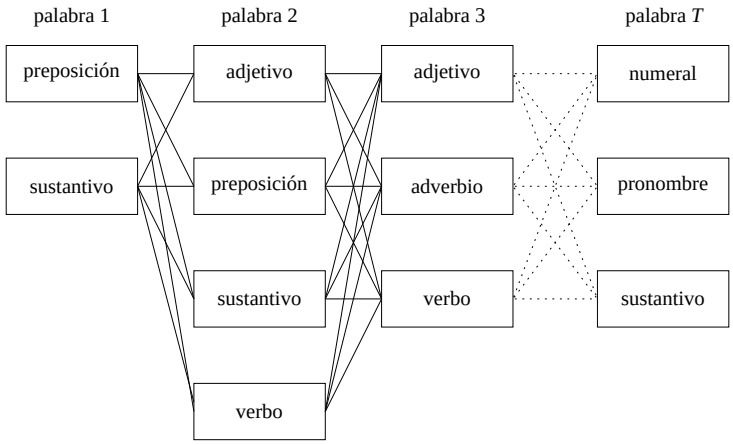


Figura 3.2: Enrejado simplificado para la etiquetación de una frase de T palabras

mayores a los asumibles empleando modelos de Markov[4]. Bajo estas premisas, Brill definió un sistema de etiquetación basado en reglas [38] que a partir de un corpus de entrenamiento infiere automáticamente las reglas de transformación. El así denominado *etiquetador de Brill* alcanza una corrección comparable a la de los etiquetadores estocásticos y, a diferencia de éstos, la información lingüística no se captura de manera indirecta a través de grandes tablas de probabilidades, sino que se codifica directamente bajo la forma de un pequeño conjunto de reglas no estocásticas muy simples, pero capaces de representar interdependencias muy complejas entre palabras y etiquetas.

El proceso de etiquetación consta de tres partes, que se infieren automáticamente a partir de un corpus de entrenamiento: un etiquetador léxico, un etiquetador de palabras desconocidas, y un etiquetador contextual:

1. Un *etiquetador léxico*, que etiqueta inicialmente cada palabra con su etiqueta más probable, sin tener en cuenta el contexto en el que dicha palabra aparece. Esta etiqueta se estima previamente mediante el estudio del corpus de entrenamiento. A las palabras desconocidas se les asigna en un primer momento la etiqueta correspondiente a sustantivo propio si la primera letra es mayúscula, o la correspondiente a sustantivo común en otro caso. Posteriormente, el etiquetador de palabras desconocidas aplica en orden una serie de reglas de transformación léxicas. Si se dispone de un diccionario previamente construido, es posible utilizarlo junto con el que el etiquetador de Brill genera automáticamente.

2. Un *etiquetador de palabras desconocidas*, que opera justo después de que el etiquetador léxico haya etiquetado todas las palabras presentes en el diccionario, y justo antes de que se apliquen las reglas contextuales. Este módulo intenta *adivinar* una etiqueta para una palabra desconocida en función de su sufijo, de su prefijo, y de otras propiedades relevantes similares. Básicamente, cada transformación consta de dos partes: una descripción del contexto de aplicación, y una regla de reescritura que reemplaza una etiqueta por otra.

3. Un *etiquetador contextual*, que actúa justo después del etiquetador de palabras desconocidas, aplicando en orden una secuencia de reglas contextuales que, al igual que las léxicas, también han sido previamente inferidas de manera automática a partir del corpus de entrenamiento.

[4]El orden de los HMM está limitado a valores pequeños debido a la carga computacional que implican y a la gran cantidad de nuevos parámetros que necesitaríamos estimar.

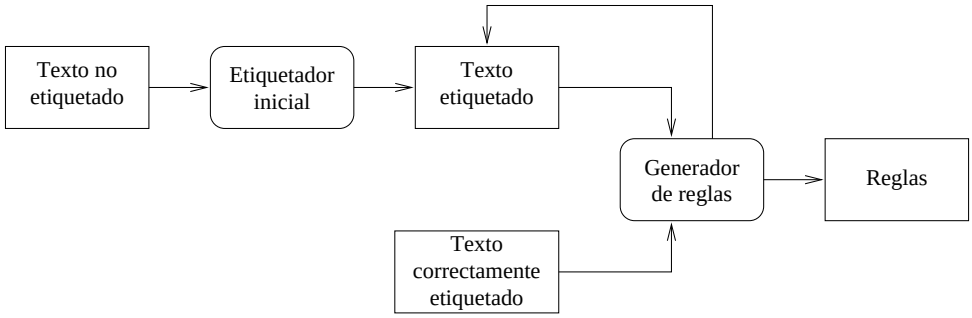


Figura 3.3: Proceso de aprendizaje de reglas en un etiquetador de Brill

El proceso de aprendizaje de las reglas, tanto las léxicas en el caso del etiquetador de palabras desconocidas, como las contextuales en el caso del etiquetador contextual, selecciona el mejor conjunto de transformaciones y determina su orden de aplicación. El algoritmo consta de los pasos que se ilustran en la figura 3.3. En primer lugar, se toma una porción de texto no etiquetado, se pasa a través de la fase o fases de etiquetación anteriores, se compara la salida con el texto correctamente etiquetado, y se genera una lista de errores de etiquetación con sus correspondientes contadores. Entonces, para cada error, se determina qué instancia concreta de la plantilla genérica de reglas produce la mayor reducción de errores. Se aplica la regla, se calcula el nuevo conjunto de errores producidos, y se repite el proceso hasta que la reducción de errores cae por debajo de un umbral dado.

La técnica de etiquetación de Brill resulta considerablemente más lenta que las basadas en modelos probabilísticos. No sólo el proceso de entrenamiento consume una gran cantidad de tiempo, sino que el proceso de etiquetación es también inherentemente lento. La principal razón de esta ineficiencia computacional es la potencial interacción entre las reglas, de manera que el algoritmo puede producir cálculos innecesarios.

Etiquetación Basada en Gramáticas de Restricciones

Las técnicas para la etiquetación de textos vistas hasta ahora son las que podríamos denominar *clásicas*. No obstante, estos métodos difícilmente permiten sobrepasar la cota del 96 % de precisión obtenida. Por otra parte, en el caso de los etiquetadores estocásticos esta cifra se reduce todavía más cuando los corpus de entrenamiento y aplicación son de tipos distintos.

Estas deficiencias abrieron paso a investigaciones sobre nuevos métodos de etiquetación, fruto de las cuales es el paradigma de *etiquetación mediante reglas de restricción*. Dentro de este campo, el sistema de etiquetación por excelencia es el sistema EngCG[5] [264]. En este sistema encontramos un conjunto de reglas escritas a mano que manejan el contexto global o, mayormente, el contexto local. No existe una verdadera noción de gramática formal, sino más bien una serie de restricciones, casi siempre negativas, que van eliminando sucesivamente los análisis imposibles de acuerdo con el contexto [207]. La idea es similar al aprendizaje basado en transformaciones, excepto por el hecho de que es un humano, y no un algoritmo, el que modifica iterativamente el conjunto de reglas de etiquetación para minimizar el número de errores. En cada iteración, el conjunto de reglas se aplica al corpus y posteriormente se intentan modificar dichas reglas de manera que los errores más importantes queden manualmente corregidos.

Podría pensarse que se trata de un retroceso a los métodos tradicionales basados en reglas, sin embargo la idea general en la que se basa este nuevo planteamiento consiste en la utilización de reglas de menor compromiso para evitar así errores en situaciones dudosas. De este modo se

[5] *English Constraint Grammar.*

ha logrado obtener una serie de métodos de alta precisión, con el inconveniente de que en algunas palabras la ambigüedad no ha sido eliminada por completo después del proceso de etiquetación, ya que no utiliza reglas de compromiso máximo. A pesar de esto, la mayoría de las palabras tendrán una única etiqueta tras el proceso de etiquetación.

Por otra parte, existe también la posibilidad de emplear este formalismo en combinación con un etiquetador tradicional como, por ejemplo, un etiquetador estocástico, que sería el encargado de completar el proceso de desambiguación. Esta solución, estudiada por el autor de esta memoria en [85], consiste en podar el enrejado inicial mediante la aplicación de reglas de restricción, eliminando combinaciones de etiquetas imposibles. Sobre el enrejado resultante se aplicaría el algoritmo de Viterbi para proceder a la desambiguación final.

El empleo de este nuevo paradigma basado en restricciones parece ofrecer mejores resultados que los etiquetadores basados en modelos de Markov ocultos —en torno al 99 % en el caso del sistema EngCG—, especialmente cuando los corpus de entrenamiento y de aplicación no provienen de la misma fuente, ya que las reglas son, en principio, universales, al no haber sido extraídas a partir de un corpus de entrenamiento. Sin embargo, la comparación de estos dos modelos es difícil de realizar, ya que cuando el sistema EngCG no es capaz de resolver determinadas ambigüedades, éste devuelve el conjunto de etiquetas obtenido para la palabra. El problema de esta técnica es, al igual que en los modelos tradicionales basados en reglas, la necesidad de participación de expertos lingüistas para la creación de las reglas, lo que supone un problema en comparación con el aprendizaje automático de los HMMs.

La Real Academia Española está desarrollando también un formalismo de reglas de restricciones denominado sistema Rtag [223]. Este sistema aplica gramáticas de reglas de contexto ponderadas sobre textos anotados ambiguamente. De esta forma, cuando un contexto satisface la descripción estructural de una regla, recibe la puntuación que indica la regla. Esta puntuación puede ser positiva, para promover lecturas, o negativa, para penalizarlas. Una vez finalizado el proceso, permanecen las lecturas con mayor puntuación siempre que estén por encima de un umbral definido previamente. El sistema también intenta eliminar lecturas imposibles en función del contexto, sin pérdida de lecturas posibles aunque éstas sean poco probables. Para la poda de lecturas en función del contexto se utiliza información derivada del propio texto (características estructurales, tipográficas o secuenciales), información gramatical (sobre todo concordancia y restricciones de aparición conjunta) e información gramatical estructural (toma de decisiones con ayuda de la información estructural derivable de la secuencia lineal del texto).

Otros Paradigmas de Etiquetación

Existen también otros paradigmas de etiquetación a mayores de los descritos anteriormente, algunos de los cuales presentaremos brevemente.

Ratnaparkhi emplea *modelos de máxima entropía* en su etiquetador JMX [181]. Esta técnica, de naturaleza también probabilística, combina las ventajas de los etiquetadores basados en transformaciones y de los etiquetadores estocásticos basados en modelos de Markov, ya que se trata de una técnica de gran flexibilidad que permite manejar un abanico de propiedades del lenguaje mayor que los modelos de Markov, acercándose al caso de Brill, y que además, al generar las distribuciones de probabilidad de etiquetas para cada palabra, permite su integración dentro de un marco probabilístico.

Los *árboles de decisión* son también empleados en tareas de etiquetación, como en el caso del etiquetador TreeTagger [215]. Un *árbol de decisión* se puede ver como un mecanismo que etiqueta todas las hojas dominadas por un nodo con la etiqueta de la clase mayoritaria de ese nodo. Posteriormente, a medida que descendemos por el árbol, reetiquetamos las hojas de los nodos hijos, si es que difieren de la etiqueta del nodo padre, en función de las respuestas

a las cuestiones o decisiones que aparecen en cada nodo. Esta manera de ver los árboles de decisión guarda ciertas similitudes con el aprendizaje basado en transformaciones, ya que ambos paradigmas realizan series de reetiquetados trabajando con subconjuntos de datos cada vez más pequeños.

Otro de los paradigmas clásicos de computación, las *redes de neuronas artificiales*, es también empleado en tareas de etiquetación. Este es el caso de la propuesta de Marques y Lopes [144] para el portugués.

Queda patente, pues, el amplio abanico de posibilidades a la hora de implementar un etiquetador gracias a la continua investigación sobre el tema. Muestra de ello es, por ejemplo, el reciente desarrollo de aproximaciones basadas en *algoritmos evolutivos* [25] o *support vector machines* [81].

3.3. Nivel Sintáctico

Una vez identificadas y analizadas las palabras individuales que componen un texto, el siguiente paso lógico consiste en estudiar cómo éstas se organizan y relacionan entre sí para formar unidades superiores (sintagmas y frases), y las funciones que representan las unidades inferiores dentro de la unidad superior. Se trata, por lo tanto, de estudiar la estructura sintáctica del texto.

3.3.1. Conceptos Básicos: Lenguajes, Gramáticas y Ambigüedad

La acotación de un lenguaje, la obtención de una representación manejable del mismo, es un paso necesario para posibilitar su procesamiento. La forma más simple de lograr este objetivo es enumerar sus cadenas constituyentes, pero este procedimiento resulta poco práctico cuando el lenguaje consta de más de unas pocas cadenas o pretendemos definir propiedades o clasificaciones entre los lenguajes. De ahí que surja la necesidad de establecer algún mecanismo para generar lenguajes con una notación finita. Estos generadores de lenguajes son las *gramáticas*, sistemas matemáticos adaptados al tratamiento computacional. De este modo definimos una *gramática* como una 4-tupla $\mathcal{G} = (N, \Sigma, P, S)$ donde:

- Σ es el alfabeto finito de la gramática o conjunto finito de *símbolos terminales*, o *palabras*, o *categorías léxicas*,
- N es un conjunto finito de *símbolos no terminales*, o *variables*, o *categorías sintácticas*, $N \cap \Sigma = \emptyset$,
- P es un subconjunto finito de $(N \cup \Sigma)^* N (N \cup \Sigma)^* \times (N \cup \Sigma)^*$ a cuyos elementos denominaremos *producciones*, *reglas*, o *reglas de producción*, y
- $S \in N$ es el *símbolo inicial*, o *axioma* de la gramática.

Con frecuencia se prefiere representar las producciones $(\alpha, \beta) \in P$ como $\alpha \to \beta \in P$. Al primer miembro α de una regla de producción $\alpha \to \beta$ se le suele llamar *parte izquierda* de la regla de producción, mientras que el segundo miembro β recibe el nombre de *parte derecha* de la regla. A las reglas cuya parte derecha es la cadena vacía ε, reglas de la forma $\alpha \longrightarrow \varepsilon$, se les llama *reglas-$\varepsilon$* o *producciones-$\varepsilon$*. Cuando dos producciones $\alpha \to \beta$ y $\alpha \to \gamma$ tienen la misma parte izquierda, se pueden escribir abreviadamente como $\alpha \to \beta \mid \gamma$.

De esta forma, un ejemplo de gramática sería aquélla que genera el lenguaje los numeros binarios pares, es decir, aquéllos terminados en 0:

$$\mathcal{G} = (\{S\}, \{0, 1\}, \{S \to A0, A \to 0A, A \to 1A, A \to \varepsilon\}, S) \tag{3.1}$$

Las cadenas del lenguaje se construyen partiendo del símbolo inicial S, siendo las producciones las encargadas de describir cómo se lleva a cabo esa generación. Empleando las reglas de producción de la gramática, se pueden construir distintas secuencias de símbolos terminales y no terminales a partir del símbolo inicial. Se denominará *formas sentenciales* a dichas secuencias, que podemos definir recursivamente de la siguiente manera. Sea $\mathcal{G} = (N, \Sigma, P, S)$ una gramática, entonces:

- S es una *forma sentencial*.
- Si $\alpha\beta\gamma$ es una forma sentencial y $\beta \to \delta \in P$, entonces $\alpha\delta\gamma$ también es una *forma sentencial*.

Intuitivamente, S es la forma sentencial más simple. A partir de ella se generan las demás formas sentenciales. Dada una forma sentencial y una regla de producción se generará una nueva forma sentencial sustituyendo una aparición de la parte izquierda de la regla en la primera, por la parte derecha de dicha regla. Un tipo especialmente interesante de forma sentencial es aquella que está formada exclusivamente por símbolos terminales. De esta forma, dada una gramática $\mathcal{G} = (N, \Sigma, P, S)$, denominaremos *frase generada por una gramática* a cualquier forma sentencial que únicamente contenga símbolos terminales. Las frases son, por lo tanto, cadenas de símbolos terminales obtenidas a través de la aplicación de reglas de producción de la gramática[6], partiendo del símbolo raíz S. Por lo tanto, son las cadenas que formarán parte del lenguaje generado por la gramática.

A modo de ejemplo, y retomando de nuevo la gramática definida en 3.1 para la generación de binarios pares, tenemos que:

Siendo S forma sentencial,	dado que $S \to A0 \in P$,	$A0$ es forma sentencial.
Siendo $A0$ forma sentencial,	dado que $A \to 0A \in P$,	$0A0$ es forma sentencial.
Siendo $0A0$ forma sentencial,	dado que $A \to 1A \in P$,	$01A0$ es forma sentencial.
Siendo $01A0$ forma sentencial,	dado que $A \to \varepsilon \in P$,	010 es una frase.

La generación de formas sentenciales y frases descrita anteriormente puede formalizarse empleando el concepto de *derivación*. Sea $\mathcal{G} = (N, \Sigma, P, S)$ una gramática, se define una *derivación directa* o *derivación en un solo paso*, $\Rightarrow$, como sigue:

$$\text{Si } \alpha\beta\gamma \in (N \cup \Sigma)^* \text{ y } \beta \to \delta \in P, \text{ entonces } \alpha\beta\gamma \Rightarrow \alpha\delta\gamma.$$

En el caso de una cadena de derivaciones directas, se dirá que $\alpha\beta\gamma$ *deriva indirectamente* $\alpha\delta\gamma$ si y sólo si:

- $\beta \Rightarrow \delta_1 \Rightarrow \delta_2 \ldots \Rightarrow \delta_n \Rightarrow \delta$, que notaremos $\alpha\beta\gamma \overset{+}{\Rightarrow} \alpha\delta\gamma$, o bien
- $\beta = \delta$ ó $\alpha\beta\gamma \overset{+}{\Rightarrow} \alpha\delta\gamma$, que notaremos $\alpha\beta\gamma \overset{*}{\Rightarrow} \alpha\delta\gamma$

En caso de conocer el número exacto k de derivaciones directas, se usará la notación $\alpha\beta\gamma \overset{k}{\Rightarrow} \alpha\delta\gamma$.

Por otra parte, la gramática impone una estructura arborescente sobre la frase o forma sentencial generada, de tal modo que dada una regla $\alpha \to \beta$, ésta conforma en sí misma un árbol donde el nodo raíz es el símbolo de la parte izquierda, siendo sus nodos hijo los símbolos de la parte derecha. Esta estructura arborescente se denomina *arbol sintáctico* o *de derivación* [182]. A modo de ejemplo, y continuando el ejemplo de los números binarios pares, recogemos en la figura 3.4 el árbol sintáctico correspondiente al número 010.

Las formas sentenciales, frases incluídas, serán aquellas que se pueden derivar a partir del símbolo inicial de la gramática. El conjunto de todas las frases generadas por una gramática

[6] Las reglas de producción que hemos usado para generar unas formas sentenciales a partir de otras.

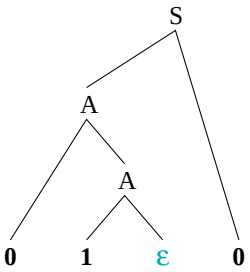


Figura 3.4: Árbol sintáctico del número binario 010

forma un lenguaje sobre el alfabeto Σ de la gramática, que podemos definir formalmente de la siguiente manera. Sea $\mathcal{G} = (N, \Sigma, P, S)$ una gramática, el *lenguaje generado por la gramática* es el conjunto $L(\mathcal{G})$ definido del siguiente modo:

$$L(\mathcal{G}) = \left\{ w | w \in \Sigma^*,\ S \overset{*}{\Rightarrow} w \right\}$$

Finalmente, introduciremos el concepto de *ambigüedad*, que se produce cuando para una misma forma sentencial existe más de un árbol sintáctico válido. En base a ello podemos definir los conceptos de gramática y lenguaje ambiguos, de tal forma que se dice que una gramática $\mathcal{G} = (N, \Sigma, P, S)$ es una *gramática ambigua* si y sólo si $\exists x \in L(\mathcal{G})$, para la cual existen al menos dos árboles sintácticos válidos. Asimismo, diremos que un lenguaje L *no es ambiguo* si y sólo si existe una gramática $\mathcal{G}$ no ambigua tal que $L(\mathcal{G}) = L$. En caso contrario diremos que L es un *lenguaje ambiguo*.

Tomemos como ejemplo una pequeña gramática aproximativa de las oraciones *sujeto-verbo-complemento* con reglas

$$\begin{aligned}
&\text{S} \rightarrow \text{NP}\ \ \text{VP}\\
&\text{S} \rightarrow \text{S}\ \ \text{PP}\\
&\text{NP} \rightarrow \text{Sust}\\
&\text{NP} \rightarrow \text{Det}\ \ \text{Sust}\\
&\text{NP} \rightarrow \text{NP}\ \ \text{PP}\\
&\text{PP} \rightarrow \text{Prep}\ \ \text{NP}\\
&\text{VP} \rightarrow \text{Verbo}\ \ \text{NP}
\end{aligned}$$

Esta gramática resulta ambigua puesto que la frase *"Juan vio un hombre con un telescopio"* puede ser generada de dos formas diferentes, dando lugar a dos árboles sintácticos distintos, tal y como se aprecia, en línea contínua y discontínua, en la figura 3.5.

3.3.2. Jerarquía de Chomsky

Dependiendo de la forma de las reglas de producción, podremos obtener lenguajes más o menos complejos. De este modo, podemos clasificar los lenguajes en función de las gramáticas que los generan y, más concretamente, en función de la forma de dichas reglas de producción. Así, Chomsky [54] propone una jerarquía con cuatro clases. En ella se clasifican, de menor a mayor complejidad, las gramáticas formales y sus lenguajes asociados, de forma que cada nivel de la jerarquía incluye a las gramáticas y lenguajes del nivel anterior, tal como se muestra en la figura 3.6.

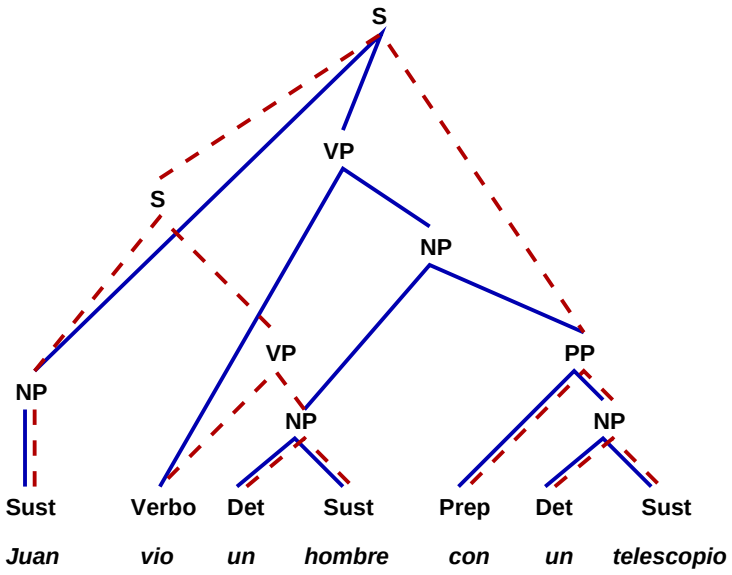


Figura 3.5: Ejemplo de ambigüedad sintáctica

Gramáticas regulares. En este caso, las producciones son de la forma: $A \rightarrow x$ ó $A \rightarrow xB$. Este tipo de producciones nos asegura que todas las formas sentenciales generadas contendrán a lo sumo un único símbolo no terminal. Los lenguajes que pueden ser generados por este tipo de gramáticas se denominan *lenguajes regulares*.

Gramáticas independientes del contexto. Sus producciones tienen un único símbolo no terminal en la parte izquierda: $A \rightarrow \beta$. De esta forma, a la hora de realizar un paso de derivación directo, es posible decidir qué símbolo no terminal queremos reescribir independientemente del contexto que lo rodea. Los lenguajes que pueden ser generados por este tipo de gramáticas se denominan *lenguajes independientes del contexto*.

Gramáticas dependientes del contexto. La parte izquierda de las producciones pueden contener cualquier combinación de símbolos terminales y no terminales, siempre y cuando sea de longitud menor o igual que la parte derecha. De esta forma aseguramos que al aplicar una derivación sobre una forma sentencial obtendremos otra forma sentencial de igual o mayor longitud. Las producciones siguen el patrón $\alpha \rightarrow \beta$, $|\alpha| \leq |\beta|$, siendo $|\alpha|$ la longitud de α, esto es, el número de símbolos en α. Los lenguajes que pueden ser generados por este tipo de gramáticas se denominan *lenguajes sensibles al contexto*.

Gramáticas con estructura de frase. No existe ninguna restricción sobre las producciones. Los lenguajes que pueden ser generados por este tipo de gramáticas se denominan *lenguajes recursivamente enumerables*.

En el caso de los *lenguajes naturales*, no se sabe a ciencia cierta qué lugar ocuparían en esta jerarquía, aunque se cree que estarían situadas entre los lenguajes independientes del contexto y los lenguajes dependientes del contexto, posiblemente más cerca de los primeros que de los segundos, tal y como podemos apreciar en la figura 3.6. Esta suposición se basa en el hecho de que la mayoría de las construcciones sintácticas sólo dependen *suavemente* del contexto en el cual son aplicadas.

Debemos reseñar que la jerarquía de Chomsky no es la única forma de clasificar lenguajes (por ejemplo, las *gramáticas contextuales* [142] son ortogonales a la jerarquía de Chomsky), aunque sí la más común.

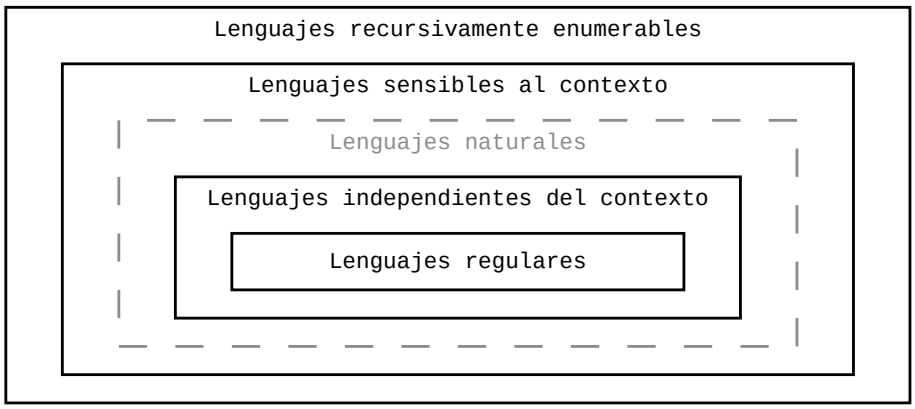


Figura 3.6: Diagrama de Venn correspondiente a la jerarquía de Chomsky

3.3.3. Análisis Sintáctico

Hasta ahora nos hemos centrado en dos conceptos fundamentales, el de lenguaje como un conjunto de cadenas y el de gramática como formalismo descriptivo de un lenguaje. El problema del análisis sintáctico se centra en encontrar un mecanismo que sirva para establecer la gramaticalidad de una cadena, es decir, reconocer si ésta pertenece al lenguaje generado por la gramática, y proponer una representación apropiada de dicho proceso de análisis. Los algoritmos que realizan sólo la primera de las dos acciones se denominan *reconocedores sintácticos*, mientras que a aquéllos capaces de generar además una representación del proceso —es decir, capaces de obtener el árbol sintáctico de la cadena procesada— se les denomina *analizadores sintácticos*. En este punto, podemos introducir una primera clasificación de los algoritmos de análisis sintáctico:

- Los *algoritmos ascendentes* son aquellos que construyen el árbol desde las hojas hasta la raíz.
- Los *algoritmos descendentes* actúan en sentido contrario a los ascendentes, de la raíz a las hojas.
- Las *estrategias mixtas* combinan los dos enfoques anteriores. Aunque existen algoritmos puros, tanto ascendentes como descendentes, lo más habitual es hacer uso de estas estrategias, que de alguna forma aportan lo mejor de cada mundo.

Podemos igualmente establecer clasificaciones de algoritmos de análisis sintáctico basándonos en otros criterios. El primero de éstos es el tratamiento del posible no determinismo en el análisis, factor de especial importancia en el caso de los lenguajes naturales debido a su ambigüedad inherente:

- *Algoritmos basados en retroceso.* En estos algoritmos el no determinismo se simula mediante un mecanismo de retroceso [13]. Cuando varias alternativas son posibles, se escoge sólo una, y, si ésta resulta infructuosa, se retrocede hasta el último punto de no determinismo y se escoge otra. Los cálculos realizados en las alternativas exploradas anteriormente se desechan. Este enfoque es sencillo, pues economiza espacio y recursos, pero presenta varios problemas:
  - Los cálculos realizados en las alternativas exploradas anteriormente se desechan. Por tanto, si éstos vuelven a ser son necesarios en una alterntiva posterior, deberán ser calculados de nuevo.

  - El criterio de selección de las alternativas puede no ser óptimo, llevándonos a una elección incorrecta de alternativas que no conducen a una solución y, por tanto, a cálculos innecesarios.
  - En caso de ambigüedad de la gramática, puede haber varias soluciones diferentes. Si se desea encontrarlas todas, se deberá forzar el retroceso tanto si se encuentran soluciones como si no, agravando los problemas anteriores.

- *Algoritmos basados en programación dinámica.* Mediante técnicas de programación dinámica [46, 65, 67], se almacenan los cálculos ya realizados de forma que no sea necesario repetirlos en caso de que se vuelvan a necesitar. Esto nos permite, incluso, compartir cálculos entre las diversas alternativas de análisis derivadas de una gramática ambigua, solucionando en parte los problemas de los algoritmos basados en retroceso, en particular la multiplicación innecesaria de cálculos y los problemas de no terminación.

Otra posible clasificación de los algoritmos de análisis sintáctico es en función de su dependencia de la estructura gramatical durante el análisis:

- *Guiados por la gramática.* La elección de las alternativas se realiza con la información proporcionada por las reglas de producción.

- *Guiados por control finito.* En estos algoritmos existe una fase de pre-procesamiento antes del análisis. En ella, se utiliza la información de las reglas de la gramática para construir un mecanismo de control que se encargará de la elección de alternativas durante el proceso de análisis.

En el contexto del lenguaje natural, ambiguo, complejo, y propenso a contener errores, cobran protagonismo, frente a las técnicas clásicas de *análisis sintáctico completo* o convencional, ciertos tipos de análisis sintáctico capaces de abordar esta problemática:

- *Análisis sintáctico robusto.* Al contrario que ocurre con los lenguajes formales, en el lenguaje natural no siempre es posible conseguir una cadena de entrada correcta y completa —debido, por ejemplo, al uso incorrecto de la lengua por parte del interlocutor—, ni una gramática exhaustiva que cubra todas las posibles cadenas de entrada —debido a su complejidad. Esta situación nos obliga a realizar el análisis sintáctico en presencia de lagunas gramaticales e, incluso, de errores. A este tipo de análisis se le califica de *robusto* [246, 245]. Debemos precisar que esta clase de análisis está dirigido a obtener la mayor cantidad de información posible a partir de una cadena de entrada con errores. Otra aproximación diferente sería intentar corregir dichos errores para obtener un análisis sintáctico completo [60]. Ambas soluciones no son, sin embargo, excluyentes, pudiendo combinarse [247, 248].

- *Análisis sintáctico parcial.* Emplearemos este término para referirnos a las técnicas de análisis capaces no sólo de obtener, de ser posible, el análisis completo de una entrada, sino también, en su defecto, sus posibles subanálisis [197, 198, 257, 47].

- *Análisis sintáctico superficial.* No siempre es necesario realizar un análisis detallado de la estructura sintáctica del texto. Para algunas tareas basta realizar un análisis *superficial* de la misma [94, 92], identificando únicamente las estructuras de mayor entidad, tales como frases nominales, grupos preposicionales, etc. En este contexto es común la utilización de cascadas de autómatas o traductores finitos [11, 10].

3.3.4. Formalismos Gramaticales

Existen diferentes *formalismos gramaticales* que pueden ser empleados a la hora de abordar el problema del análisis sintáctico en lenguaje natural.

A partir de los años 60, la mayor parte de los modelos computacionales para el procesamiento del lenguaje natural se basaron en gramáticas independientes del contexto debido a la disponibilidad de algoritmos eficientes para realizar el análisis de este tipo de gramáticas, tales como el CYK [271, 123] o el algoritmo de Earley [67].[7]

También es frecuente extender las gramáticas independientes del contexto mediante la decoración de producciones y árboles de análisis con probabilidades para así posibilitar un mejor tratamiento de las ambigüedades [36]. De cara a su análisis se desarrollaron extensiones análogas de los correspondientes algoritmos clásicos de análisis [116, 228].

Sin embargo, las lenguas naturales presentan construcciones que no pueden ser descritas mediante gramáticas independientes del contexto. Surge entonces la necesidad de contar con formalismos más adecuados que permitan llenar el hueco descriptivo existente.

Una de las posibilidades es la del empleo de la operación de unificación en entornos gramaticales [125, 56]. Entre los formalismos con unificación más extendidos se encuentran las *gramáticas de cláusulas definidas*, una generalización de las gramáticas independientes del contexto basada en lógica de primer orden [171]. Sobre la base de una gramática independiente del contexto, se generalizan los símbolos de la misma añadiendo información adicional, *atributos* del símbolo. De este modo los símbolos de la gramática nos permiten representar un conjunto infinito de elementos, extendiendo de este modo su dominio de definición. A continuación se establece una operación que nos permita la manipulación de los símbolos gramaticales con atributos y se adapta convenientemente el mecanismo de derivación de la gramática de forma que tenga en cuenta la información contenida en éstos. La extensión se realiza mediante términos lógicos de primer orden, considerando la unificación [195] como mecanismo de manipulación.

Otros formalismos que utilizan unificación, en este caso unificación de estructuras de rasgos, son las gramáticas léxico-funcionales [122, 169], las gramáticas con estructura de frase dirigidas por el núcleo [178], y las gramáticas categoriales de unificación [234].

Puesto que la estructura sintáctica asociada a las frases es una estructura jerárquica representada normalmente como un árbol o, en el caso de frases ambiguas, como un conjunto de árboles, parece natural pensar que un formalismo que manipule árboles y que presente cierta dependencia suave del contexto resultaría adecuado para la descripción de los fenómenos sintácticos que aparecen en el lenguaje natural. Con este objetivo nacen las *gramáticas de adjunción de árboles* [119], uno de los formalismos gramaticales derivados de las gramáticas independientes del contexto más ampliamente difundidos. En este tipo de gramáticas la estructura fundamental es el árbol, en lugar de la producción. Los árboles se clasifican en iniciales y auxiliares. Los árboles iniciales suelen utilizarse para representar las estructuras de las frases elementales, mientras que los árboles auxiliares se utilizan para representar estructuras recursivas mínimas que se pueden añadir a otros árboles. Los árboles se combinan mediante las operaciones de adjunción y sustitución. Desde el punto de vista lingüístico las grandes ventajas de las gramáticas de adjunción de árboles provienen de su carácter lexicalizado —ya que permiten asociar una palabra con cada árbol— y de su dominio de localidad extendido, posibilitando el establecimiento de relaciones de larga distancia entre los nodos de árboles elementales. También en este caso existen adaptaciones de los algoritmos clásicos de análisis para el caso de las gramáticas de adjunción de árboles [213]. Debemos destacar también la investigación se ha hecho en torno al análisis sintáctico de gramáticas de adjunción de árboles, tanto en análisis

[7] Una visión conjunta de la mayor parte de los algoritmos de análisis sintáctico para gramáticas independientes del contexto puede encontrarse en la obra de Sikkel [217].

bidireccional [20, 16], como mediante autómatas [16, 66].

Existen multitud de formalismos equivalentes a las gramáticas de adjunción de árboles. Entre ellos destacan las gramáticas lineales de índices [18, 19], las gramáticas categoriales combinatorias [225], y las gramáticas de núcleo [186]. Todos estos formalismos se engloban en la clase de los formalismos gramaticales suavemente sensibles al contexto [120].

Existen otros formalismos gramaticales que no se basan en las gramáticas independientes del contexto. Por ejemplo, las *gramáticas de dependencia* [150], que se fundamentan en las relaciones existentes entre palabras y no en las relaciones entre constituyentes.

3.4. Nivel Semántico

La *semántica* es el estudio del significado lingüístico. Consecuentemente, a la hora de realizar un análisis semántico de un texto, nuestro objetivo será el de obtener el significado de las frases que lo componen. En este apartado realizaremos una breve introducción a este campo, menos detallada que en el caso de los niveles anteriores, ya que el nivel semántico, al igual que el nivel pragmático, no es abordado profundamente en nuestro trabajo.

El primer punto a abordar es el de las *representaciones semánticas*, ya que las diferentes aproximaciones al análisis semántico parten de la base de que la semántica de los diferentes elementos lingüísticos —palabras, sintagmas— puede ser capturada mediante estructuras formales. Estas estructuras deberían cumplir una serie de características:

1. **Verificabilidad.** Debemos ser capaces de determinar la verdad o falsedad acerca del enunciado expresado por nuestra representación de acuerdo a nuestra *base de conocimiento*.

2. **No ambigüedad.** Si bien pueden existir ambigüedades lingüísticas a nivel semántico, como en el caso de la frase *"todos los alumnos de la facultad hablan dos idiomas"*, no debemos confundir esta ambigüedad en el enunciado con una ambigüedad en la representación de dicho enunciado. Por lo tanto, independientemente de la existencia de ambigüedades en el texto fuente, el tipo de representación semántica empleada debe admitir una única interpretación no ambigua, interpretación que en su caso sí deberá reflejar la ambigüedad del enunciado.

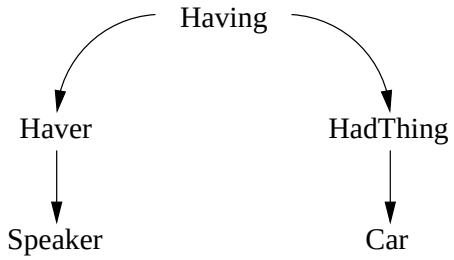
3. **Existencia de una forma canónica.** Debemos ser capaces de asociar una única representación a entradas diferentes con formas diferentes pero igual significado. De este modo evitaremos el riesgo de evaluar de diferente manera la verdad o falsedad de una aserción según la manera en que ésta hubiese sido formulada. Esto supone tratar la *variación lingüística* del lenguaje, es decir, cómo un mismo concepto puede ser expresado de formas diferentes mediante el empleo, por ejemplo, de sinónimos (p.ej., *listo/inteligente*), construcciones gramaticales equivalentes (p.ej., *Juan asesinó a Pedro/Pedro fue asesinado por Juan*), etc.

4. **Disponibilidad de mecanismos de inferencia y uso de variables.** De esta forma el sistema deberá ser capaz de decidir acerca de la verdad o falsedad de proposiciones que no estén explícitamente representadas en su base de conocimiento, pero que sí sean derivables a partir de la misma. Por su parte, el empleo de variables permitirá el manejo de entradas con referencias no totalmente definidas.

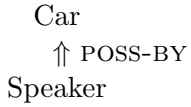
5. **Expresividad.** El tipo de representación empleada debe ser capaz de representar cualquier aserción de interés para la aplicación.

$$\exists\, x, y\ \text{Having}(x) \wedge \text{Haver}(\text{Speaker}, x) \wedge \text{HadThing}(y, x) \wedge \text{Car}(y)$$

(a) Predicado lógico de primer orden



(b) Red semántica



(c) Diagrama de dependencia conceptual

```
Having
  Haver:      Speaker
  HadThing:   Car
```

(d) *Frame*

Figura 3.7: Representaciones semánticas de la oración *"I have a car"* ("Yo tengo un coche")

La figura 3.7 recoge, para el ejemplo *"I have a car"* ("Yo tengo un coche"), algunas de estructuras formales de representación semántica más utilizadas, y comunes al ámbito de la Inteligencia Artificial clásica [185].

La primera de ellas recoge una aproximación basada en el *cálculo de predicados de primer orden*, una de las soluciones más extendidas. Los inicios de su empleo para la captura del significado de textos en lenguaje natural data de la década de los 60, cuando Woods [267] investiga la posibilidad de utilizar representaciones basadas en lógica de predicados para los sistemas de búsqueda de respuestas en lugar de representaciones ad-hoc como venía siendo corriente hasta entonces.

Por esa misma época, aquellos investigadores interesados en el modelado cognitivo del lenguaje y de la memoria trabajaban en varias formas de representación basadas en redes asociativas. Es en este periodo cuando se comienza a investigar con profusión en el ámbito de las *redes semánticas* [147], el segundo caso recogido en la figura 3.7. En una red semántica los objetos son representados como nodos en un grafo, mientras que las relaciones entre los mismos son representadas mediante arcos etiquetados.

La tercera de las estructuras es un diagrama de *dependencia conceptual* [214]. Se trata de una forma de representación de amplio uso en el campo del lenguaje natural, y que emplea una serie de primitivas conceptuales que se pueden combinar entre sí para expresar un significado dado.

El último caso recogido en la figura 3.7 se trata de una representación basada en *frames*, estructuras de conocimiento que constan de una *cabecera*, que identifica el *frame*, y de una serie

de atributos —denominados *slots*—, que pueden contener tanto valores atómicos como nuevos *frames* anidados.

A la hora de realizar el *análisis semántico* propiamente dicho —y contando ya con una estructura de representación adecuada—, nuestro objetivo es el de obtener la representación semántica de la frase componiendo de algún modo las representaciones individuales de sus componentes. Uno de los enfoques más utilizados es el denominado *análisis dirigido por la sintaxis* (*syntax-driven semantic analysis*) [121]. Éste se basa en el llamado *principio de composicionalidad*[8], y según el cual la semántica de una objeto puede ser obtenida a partir de la semántica de sus componentes. Fue Montague [166] quien mostró que el enfoque composicional podía ser aplicado a una parte importante del lenguaje natural, introduciendo la estructura de modelos teóricos en la teoría lingüística, y dando lugar de este modo a una integración mucho más fuerte entre las teorías de la sintaxis formal y un amplio rango de estructuras semánticas.

Sin embargo, si bien el significado de una frase puede obtenerse a partir de los significados de las palabras y sintagmas que la componen, también es cierto que los meros significados aislados de los mismos no son suficientes. De esta forma, si partimos de un conjunto de palabras {*Juan*, *matar*, *Pedro*}, no es en absoluto lo mismo decir *"Juan mató a Pedro"* que *"Pedro mató a Juan"*. Por lo tanto, debemos matizar nuestra afirmación anterior, ya que el significado de una frase no se obtiene únicamente a partir de las palabras que la forman, sino que también viene dado por la forma en que éstas se relacionan entre sí. En otras palabras, el significado de la frase depende parcial pero inexorablemente de su estructura sintáctica. De esta forma, en el *análisis dirigido por la sintaxis* el sistema parte de las representaciones de significado de los componentes para, guiado por la estructura o sintaxis de la frase, obtener la representación resultante de la frase.

En relación a lo anterior, debemos destacar que uno de los entornos aplicativos más representativos en los cuales se trata de capturar la semántica de los textos es el de la propia *Recuperación de Información*, puesto que, como ya se apuntó en el apartado 2.2.1, la mayor parte de los sistemas de recuperación de información actuales están basados en una interpretación extrema del principio de composicionalidad, al considerar que la semántica de los documentos reside únicamente en las palabras que lo forman, sin tener en cuenta el orden de los constituyentes ni su estructura sintáctica. Es lo que se conoce habitualmente como aproximación basada en *bag-of-terms*.

Uno de las herramientas más utilizadas en tareas de procesamiento semántico es la base de datos lexicográfica WordNet [158, 156, 97, 70, 33], en el caso del inglés, o su equivalente EuroWordNet [263], en el caso de otras lenguas europeas —ya abordadas en el apartado 2.4.1.

El hecho de que una misma palabra pueda tener diversos significados según el contexto en el que ésta se utilice constituye uno de los principales problemas del análisis semántico. Las técnicas de *desambiguación del sentido de las palabras* [226, 68] tratan de resolver esta ambigüedad léxica seleccionando el sentido adecuado de cada palabra en una frase. La complejidad de esta tarea viene determinada por la cantidad de palabras homónimas y polisémicas presentes en el vocabulario del idioma. En esencia, se aplican técnicas similares a las utilizadas para realizar la etiquetación de las palabras en el nivel morfológico, pero en lugar de utilizar etiquetas morfosintácticas se utilizan etiquetas semánticas que identifican el sentido de las palabras. Por tanto se tratará de obtener el sentido más probable de una palabra en relación con los sentidos de las palabras vecinas.

[8] Comúnmente conocido como *principio de composicionalidad de Frege*, aún cuando Frege nunca se refirió explícitamente a él.

3.5. Nivel Pragmático

La *pragmática* es el estudio de la relación entre el lenguaje y el contexto en el que se utiliza. El contexto incluye elementos como la identidad de las personas y los objetos participantes, y por tanto la pragmática incluye el estudio de cómo se utiliza el lenguaje para referenciar a personas y cosas. También incluye el contexto del discurso y, por consiguiente, el estudio de cómo se estructura el discurso y de cómo los participantes en una conversación gestionan el diálogo. En consecuencia, para realizar el análisis pragmático se precisa de algoritmos para la resolución de la anáfora, modelos computacionales para recuperar la estructura de monólogos y diálogos, y modelos de gestión del diálogo.

La importancia de la correcta interpretación de la *anáfora* viene dada por su necesidad a la hora de procesar correctamente textos escritos en lenguaje natural [159], especialmente en el caso de tareas como la extracción de información y la creación de resúmenes de textos. Los primeros trabajos sobre resolución de la anáfora trataban de explotar el conocimiento lingüístico y del dominio que se tenía, el cual era difícil tanto de representar como de procesar, requiriendo una notable participación humana. La necesidad de desarrollar soluciones robustas de bajo coste computacional hizo que muchos investigadores optasen por técnicas que hiciesen uso de un conjunto limitado de recursos lingüísticos. Este enfoque vino también motivado por la existencia de herramientas fiables y eficientes para el tratamiento de corpus, tales como etiquetadores-lematizadores y analizadores sintácticos superficiales.

En lo referente al procesamiento de *diálogos*, los primeros sistemas conversacionales, como el ELIZA [265], eran sistemas muy simples, basados fundamentalmente en el emparejamiento de patrones. Se hizo necesaria una mejor comprensión de los mecanismos del diálogo humano para el desarrollo de gestores del diálogo más sofisticados. Se estableció, por ejemplo, el concepto de subdiálogo, y se observó que los diálogos orientados a la realización de una determinada tarea presentaban una estructura cercana a la de la tarea que estaba siendo realizada. En el caso del *monólogo*, su tratamiento es similar al del diálogo, si bien menos complejo, ya que por ejemplo el tratamiento de la anáfora requiere analizar, en el diálogo, tanto el texto del actuante como el de los otros interlocutores.

En la actualidad uno de los principales ámbitos de aplicación del análisis pragmático es el de la *traducción automática* (*machine translation*) [107]. Las primeras investigaciones en este campo se remontan al década de los 50. El optimismo inicial dio paso, al poco tiempo, a una etapa de oscurantismo debido a la falta de recursos software y hardware adecuados para la tarea. Si bien algunos investigadores siguieron trabajando en el campo —caso del sistema SYSTRAN [5]— fue a partir de los 80 cuando cobró nuevo interés. Frente a las primeras aproximaciones de esta década, basadas en el significado y en la utilización de una interlingua, la investigación actual gira en torno a la utilización de métodos estadísticos y basados en la alineación de corpus multilingüe paralelos [184, 109], gracias a la disponibilidad de corpus de gran tamaño y de herramientas computacionales de suficiente potencia. Este nuevo interés radica en el aumento de las relaciones comerciales internacionales, la puesta en práctica de políticas gubernamentales que propician la traducción de documentos oficiales a varias lenguas —caso de la Unión Europea—, y la difusión mediante Internet de una ingente cantidad de información en formato electrónico.

En la misma línea, y por su relación con la temática de esta tesis, llamamos la atención sobre un campo de investigación en continuo desarrollo desde hace algunos años: la *Recuperación de Información Translingüe* (CLIR, *Cross-Lingual Information Retrieval*) [93]. Se trata de uno de los campos dentro de la Recuperación de Información, y en el cual consultas y documentos están en idiomas diferentes.

3.6. Procesamiento del Lenguaje Natural y Recuperación de Información

La comunidad científica que investiga la Recuperación de Información ha mostrado en repetidas ocasiones su interés por el empleo de técnicas de Procesamiento de Lenguaje Natural. La razón para este interés reside en el hecho de que decidir acerca de la relevancia de un documento dado respecto a una consulta consiste, en esencia, en decidir acerca de si el texto del documento satisface la necesidad de información expresada por el usuario, lo que implica que el sistema debe comprender, en cierta medida, el contenido de dicho documento [229].

Tal y como ya hemos indicado anteriormente, los sistemas de IR actuales se basan en una interpretación extrema del *principio de composicionalidad*, que nos dice que la semántica de un documento reside únicamente en los términos que lo forman [121]. De este modo, podemos suponer que cuando una palabra determinada está presente en un documento, dicho documento trata del tema indicado por dicha palabra [130]. De igual modo, cuando una consulta y un documento comparten términos índice, se puede presumir que el documento aborda, de algún modo, el tema sobre el que trata la consulta [24] (véase apartado 2.2.1). En base a ello ambos, consultas y documentos, son representados mediante conjuntos de términos índice o palabras clave —paradigma *bag-of-terms* [26]—, de tal forma que la decisión acerca de la relevancia o no de un documento respecto a una consulta es tomada de acuerdo al grado de correspondencia entre el conjunto de términos índice asociados al documento y el conjunto de términos índice asociados a la consulta. Asimismo, la utilización de pesos a la hora de medir el mayor o menor poder discriminante de un determinado término (véase apartado 2.2.2), así como el empleo de funciones de ordenación (véase apartado 2.2.3), permiten la ordenación de los documentos pertenecientes al conjunto respuesta de acuerdo a su grado de relevancia respecto a la consulta.

En este contexto, una de las principales limitaciones a las que han de hacer frente los sistemas de IR es la *variación lingüística* inherente al lenguaje humano [24], es decir, aquellas alteraciones de carácter lingüístico que un término puede sufrir y que impiden el correcto establecimiento de correspondencias —con el correspondiente detrimento de precisión y cobertura— en situaciones como la existencia de cambios en la flexión de una palabra —p.ej., *gato* vs. *gatas*—, el empleo de sinónimos —p.ej., *matar* vs. *asesinar*—, la presencia de ambigüedades semánticas —p.ej. *banda* (de tela) vs. *banda* (de forajidos)—, etc.

Se hace patente, pues, que el lenguaje no es un mero repositorio de palabras, tal como pretende el paradigma *bag-of-terms*, sino que nos permite comunicar conceptos, entidades, y relaciones, de múltiples maneras diferentes. Del mismo modo, las palabras se combinan a su vez en unidades lingüísticas de mayor complejidad, cuyo significado no siempre viene dado por el significado de sus palabras componente.

La aplicación de técnicas de Procesamiento del Lenguaje Natural al ámbito de la Recuperación de Información surge como respuesta a la necesidad de mejorar el tratamiento de la variación lingüística. El desarrollo de nuevas herramientas de NLP, más eficientes, robustas, y precisas, así como la cada vez mayor potencia de las nuevas generaciones de ordenadores han promovido el desarrollo de dicha aplicación. Sin embargo, debemos precisar a este respecto que el trabajo de investigación llevado a cabo hasta la fecha ha estado primordialmente centrado en el caso del inglés, y si bien otras lenguas como el francés o el alemán han sido también objeto de estudio, el español ha quedado relegado frecuentemente a un segundo plano. Por otra parte, la mayor complejidad lingüística del español frente al inglés en todos sus niveles no permite una extrapolación inmediata al español de los resultados obtenidos para el inglés, requiriendo la realización de experimentos específicos.

A continuación describiremos los diferentes niveles de variación lingüística existentes, así como las diferentes aproximaciones propuestas para abordar estos niveles.

3.6.1. Variación Morfológica

La *morfología* es la parte de la gramática que se ocupa del estudio de la estructura de las palabras y de sus mecanismos de formación en base a unidades mínimas de significado denominadas *morfemas* (ver apartado 3.2). Dentro de la morfología podemos hablar de morfología flexiva y morfología derivativa. La *morfología flexiva* hace referencia a aquellos cambios predecibles fruto de las variaciones de género y número (p.ej., *hablador* vs. *habladoras*), persona, modo, tiempo y aspecto (p.ej., *hablar* vs. *hablases*), etc., los cuales no conllevan una modificación de la categoría gramatical de la palabra, ni tampoco cambios relevantes de significado. Por contra, la *morfología derivativa* estudia la formación de nuevo léxico en base a mecanismos de *derivación*, la unión de morfemas individuales o grupos de morfemas —en este caso morfemas derivativos— para formar términos más complejos. Al contrario que en el caso de la flexión, las modificaciones derivativas sí producen un cambio semántico respecto al término original, y frecuentemente también un cambio de categoría sintáctica (p.ej., *hablar* vs. *hablador*).

La *variación morfológica* conlleva, por tanto, una pérdida de cobertura por parte del sistema, ya que impide establecer correspondencias entre términos próximos debido a las alteraciones morfológicas flexivas o derivativas que ha sufrido. Las soluciones clásicas a la hora de mitigar los efectos de la variación de carácter morfológico pasan por la *expansión de la consulta* mediante las variantes morfológicas de los términos originales [168], o por el empleo de técnicas de *stemming*. Ambas técnicas fueron ya introducidas en los apartados 2.4.1 y 2.3.1, respectivamente, y si bien su efecto es equivalente, la técnica más extendida a la hora de su empleo para la normalización morfológica de un texto es el *stemming*.

Sin embargo, las técnicas tradicionales de *stemming* —el algoritmo de Porter, por ejemplo—, son bastante agresivas, pudiendo dar lugar a normalizaciones erróneas que incidan negativamente en la precisión. Por ejemplo, en inglés, un algoritmo basado en Porter normalizaría las palabras *general* (general) y *generous* (generoso), en una forma común *gener-*. Este problema se agrava en el caso de lenguas de morfología más compleja e irregular que la del inglés [24, 233], como ocurre en el caso del español [74].

A nivel flexivo, Arampatzis et al. [24] proponen una solución más conservadora en la que el proceso de normalización retenga la categoría gramatical de la palabra original. Para ello se propone el empleo de técnicas de *lematización*, en las que los términos que componen el texto sean reducidos a su *lema* o forma canónica —forma masculina singular en nombres y adjetivos e infinitivo en verbos—, eliminando de esta forma la flexión de una palabra. La aproximación al nivel derivativo debe ser, sin embargo, más cauta, debido a los cambios semánticos y de categoría gramatical que conllevan con frecuencia las relaciones derivativas. Algunas relaciones podrían venir indicadas por la propia sintaxis, tales como la nominalización de la acción de un verbo, mientras que otras relaciones más indirectas podrían requerir el empleo de información semántica. No obstante, el potencial de su uso, especialmente en el caso de lenguajes de morfología rica —como el español—, es notable [209, 233, 114].

3.6.2. Variación Semántica

La *variación semántica* viene dada por la *polisemia*, el hecho de que una misma palabra pueda tener diferentes significados o sentidos en función de su contexto. Tal es el caso, por ejemplo, de *banda*: banda de música, banda de delincuentes, banda de tela, etc. Esto incide negativamente en la precisión del sistema, ya que una consulta referente a, por ejemplo, *bandas municipales* podría devolver, equivocadamente, documentos sobre *bandas de delincuentes*.

Para reducir en lo posible la variación semántica de un texto se hace preciso recurrir entonces a técnicas de *desambiguación del sentido de las palabras* [226, 68] para identificar el sentido concreto de cada palabra. Dichas técnicas fueron ya tratadas en el apartado 3.4

3.6.3. Variación Léxica

La *variación léxica* hace referencia a la posibilidad de emplear términos diferentes a la hora de representar un mismo significado, como ocurre en el caso de los *sinónimos*. Este tipo de variación lingüística incide también negativamente en la cobertura del sistema, ya que una consulta que hiciese referencia al término *automóvil* no devolvería documentos que únicamente se refiriesen al término *coche*.

A la hora de tratar estos fenómenos debe tenerse en cuenta el gran impacto que la variación semántica tiene en los procesos de tratamiento de la variación léxica, ya que la elección de uno u otro término semánticamente equivalente a una palabra dada depende del sentido de la misma en su contexto. Es por ello que a la hora de tratar la variación léxica se hace necesario eliminar, en primer lugar, la variación semántica del texto mediante procesos de desambiguación del sentido. Se estima, de hecho, que una desambiguación con una efectividad menor del 90 % puede ser incluso contraproducente [208] en este tipo de procesos, si bien otros trabajos, como el de Stokoe et al. [227] apuntan a que una efectividad del 50 %-60 % es suficiente.

Algunas de las soluciones propuestas para este problema pasan por la expansión de consultas con términos relacionados léxico-semánticamente —sinónimos, hipónimos, etc.—, el empleo de distancias conceptuales a la hora de comparar consultas y documentos, y la indexación mediante *synsets* de WordNet [158, 156, 97, 70, 33]. Asimismo, es precisamente esta base de datos léxica, WordNet, la fuente de información semántica más común.

La expansión de consultas mediante términos relacionados léxico-semánticamente ha sido empleada en repetidas ocasiones, mostrando buenos resultados en el caso de consultas cortas o incompletas, pero escasa o nula incidencia en el caso de consultas suficientemente completas [261].

Por otra parte, experimentos empleando recuperación basada en distancias semánticas [222] han mostrado mejoras en los resultados, si bien dichos experimentos fueron limitados, por lo que no pueden considerarse plenamente representativos.

Finalmente, la indexación mediante *synsets* [82] en lugar de palabras únicamente produce mejoras cuando el sentido de las palabras de las consultas ha sido plenamente desambiguado.

3.6.4. Variación Sintáctica

El tratamiento de la *variación sintáctica*, fruto de las modificaciones en la estructura sintáctica de un discurso manteniendo su significado, han sido tratadas tradicionalmente mediante dos aproximaciones diferentes: aquéllas que operan sobre estructuras sintácticas, y aquéllas que emplean frases a modo de *términos índice complejos*. En ambos casos el objetivo perseguido es aumentar la precisión en el proceso de recuperación, salvando en lo posible las limitaciones del paradigma *bag-of-terms* [233] a la hora de considerar la información sintáctica del texto.

El empleo de representaciones complejas en base a estructuras sintácticas durante el proceso de indexación y/o búsqueda, como podrían ser el caso de árboles [182, 256] o grafos [167], plantea problemas debido a su alto coste, haciéndolas poco adecuadas para su empleo a gran escala en entornos prácticos.

La solución más extendida pasa por el empleo de frases como términos índice dentro de un paradigma de recuperación clásico. La hipótesis sobre la que se sustenta su uso es la de que las frases denotan conceptos o entidades más significativos que en el caso de las palabras individuales, por lo que presumiblemente deberían constituir términos índice más precisos y descriptivos [230, 24]. En lo que respecta a la cobertura del sistema, ésta no se ve inicialmente afectada, ya que los términos simples que componen de una frase hubieran también dado lugar a correspondencias entre documento y consulta de haber empleado únicamente términos simples [161].

Tradicionalmente se han considerado dos tipos de frases en IR: las frases *estadísticas*, obtenidas mediante técnicas estadísticas que buscan secuencias de palabras contiguas que coocurren con una frecuencia significativa [162, 42], y las frases *sintácticas*, formadas por conjuntos de palabras relacionadas sintácticamente, y obtenidas mediante técnicas de NLP [168, 130, 112, 172, 106]. La mayor utilidad de uno u otro tipo de frases en tareas de IR es una cuestión todavía por discernir plenamente, aunque existen resultados que apuntan hacia las frases sintácticas como mejor opción, al menos en un futuro a medio plazo ante la presumible disponibilidad de técnicas de análisis y desambiguación sintáctica adecuadas [24]. Por otra parte, debemos puntualizar que gran parte de las soluciones investigadas hasta ahora en el caso de las soluciones sintácticas suelen emplear como términos índice complejos únicamente sintagmas nominales [132, 161, 106]. Es también común, tanto en el caso de frases estadísticas como sintácticas, que los términos complejos empleados consten nada más que de dos constituyentes, descomponiendo de ser preciso aquellos términos de más de dos constituyentes en compuestos de únicamente dos elementos [24, 172, 69].

Debe tenerse también en cuenta que los términos complejos son utilizados mayormente en combinación con términos simples [168, 161, 106, 230, 42], ya que el empleo único de frases como términos índice permite capturar sólamente una vista parcial e insuficiente del documento, lo que redunda en un empeoramiento de los resultados [161].

Parte II

Normalización de Términos Simples

Capítulo 4

Preprocesamiento y Segmentación

4.1. Introducción

Una de las tareas previas más importantes en el procesamiento automático del lenguaje natural es la correcta segmentación y preprocesamiento de los textos, ya que las palabras y frases identificadas en esta fase constituirán las unidades fundamentales sobre las que deben trabajar las fases posteriores —etiquetadores, analizadores sintácticos, sistemas de Recuperación de Información, etc. No obstante, esta fase es a menudo obviada en muchos de los desarrollos actuales, lo cual, en el caso particular de los sistemas de IR, conduce en ocasiones a normalizaciones erróneas que afectan negativamente al rendimiento del sistema. Sin embargo, la creciente disponibilidad de grandes corpus ha hecho que los procesos de preprocesamiento y segmentación cobren nueva importancia, prestando especial atención a su robustez.

Se trata, pues, de una tarea de enorme importancia práctica, que puede involucrar procesos mucho más complejos que la simple identificación de las diferentes frases del texto y de cada uno de sus componentes individuales. Abordarla con pleno rigor científico, sin caer repetidamente en el análisis de la casuística particular de cada fenómeno detectado, es una labor que resulta especialmente compleja, dado que actualmente ni siquiera existe una estrategia clara y definida respecto al orden en el que se deben abordar los diversos fenómenos lingüísticos involucrados.

Esta dificultad se ve acentuada por el alto grado de dependencia que presenta dicha tarea a diferentes niveles [170]. En primer lugar, su dependencia respecto al idioma a procesar, que presentará una serie de fenómenos lingüísticos particulares a tratar. En segundo lugar, respecto a la aplicación en la que se integrará el preprocesador, y que dictará las necesidades a cubrir, ya que no debemos ver el preprocesamiento como una fase aislada, sino como un proceso estrechamente ligado al diseño e implementación de los otros componentes del sistema en el cual se integra. Finalmente, respecto a su ámbito de aplicación, puesto que, por ejemplo, el preprocesamiento de textos literarios dista mucho del de textos científicos, al presentar ambos fenómenos muy diferentes —debemos, por tanto, incidir de nuevo en la robustez.

En el presente capítulo desarrollamos un esquema de preprocesamiento orientado a la desambiguación y etiquetación robusta del español para su aplicación en el ámbito de la Recuperación de Información. No obstante, se trata de una propuesta de arquitectura general que puede ser aplicada a otros idiomas mediante sencillas modificaciones. Muestra de ello es el hecho de que actualmente se encuentran en explotación dos versiones del mismo, una para el español y otra para el gallego [31, 30].

4.2. El Concepto de *Token* y el Concepto de Palabra

Los diferentes tipos de etiquetadores que existen actualmente asumen que el texto de entrada aparece ya correctamente segmentado, es decir, dividido de manera adecuada en *tokens* o unidades de información de alto nivel de significado, que identifican perfectamente cada uno de los componentes de dicho texto. Esta hipótesis de trabajo no es en absoluto realista debido a la naturaleza heterogénea tanto de los textos de aplicación como de las fuentes donde se originan. Entre otras tareas, el proceso de segmentación se encarga de identificar unidades de información tales como las frases o las propias palabras. Esta operación es más compleja de lo que pueda parecer *a priori*, especialmente en idiomas como el español y el gallego, y dista mucho de la sencillez del preprocesamiento léxico propia de los *tokenizadores* empleados en el caso de los lenguajes de programación [12].

En el caso de la segmentación del texto en palabras, la problemática se centra en el hecho de que el concepto ortográfico de palabra no siempre coincide con el concepto lingüístico. Se nos presentan entonces dos alternativas:

1. Las aproximaciones más sencillas consideran igualmente las palabras ortográficas, ampliando el conjunto de etiquetas para así representar aquellos fenómenos que sean relevantes. Por ejemplo, el término `reconocerse` podría etiquetarse conjuntamente como *verbo infinitivo con un único pronombre enclítico*, cuando en realidad está formado por un verbo y un pronombre enclítico. El ejemplo contrario sería el de las palabras de la locución `en vez de`, que se etiquetarían, respectivamente, como *primer elemento de una conjunción de tres elementos*, *segundo elemento de una conjunción de tres elementos* y *tercer elemento de una conjunción de tres elementos*, aun cuando constituyen un único término. Sin embargo, en idiomas como el español y el gallego, tal planteamiento no es viable, ya que su gran complejidad morfológica produciría un crecimiento excesivo del juego de etiquetas.

2. No ampliar el juego de etiquetas básico. Como ventajas, la complejidad del proceso de etiquetación no se verá afectada por un número excesivo de etiquetas, y la información relativa a cada término lingüístico se puede expresar de manera más precisa. Por ejemplo, a lo que antes era un simple pronombre enclítico se le pueden atribuir ahora valores de persona, número, caso, etc. Como desventaja, la complejidad del preprocesador aumenta, ya que no sólo se limitará a identificar las palabras ortográficas, sino que deberá también identificar las palabras lingüísticas, lo que conlleva la partición o agrupación de las palabras ortográficas para así obtener las palabras lingüísticas asociadas.

   Las mayores dificultades surgen cuando la segmentación es ambigua. Este es el caso, por ejemplo, de la expresión `sin embargo`: puede tratarse de una conjunción —como en la frase *"es atrevido, sin embargo me gusta"*—, en cuyo caso se etiquetará de manera conjunta como una conjunción, o bien de la secuencia formada por la preposición `sin` y el sustantivo `embargo` —como en la frase *"los intercambios comerciales prosiguieron sin embargo económico alguno"*—. De igual forma, la palabra `ténselo` puede ser una forma del verbo `tener` con dos pronombres enclíticos, `se` y `lo`, o bien una forma del verbo `tensar` con un solo pronombre enclítico, `lo`[1].

En nuestra aproximación se ha optado por la segunda opción, es decir, la de identificar las palabras lingüísticas. En cualquier caso, la primera opción, la de trabajar a nivel de

[1] Este fenómeno es especialmente común en el gallego, no sólo en el caso de los pronombres enclíticos, sino también en el caso de algunas contracciones. Por ejemplo, la palabra `polo` puede ser un sustantivo (en español, pollo), o bien la contracción de la preposición `por` (por) y del artículo masculino `o` (el), o incluso la forma verbal `pos` (pones) con el pronombre enclítico de tercera persona `o` (lo).

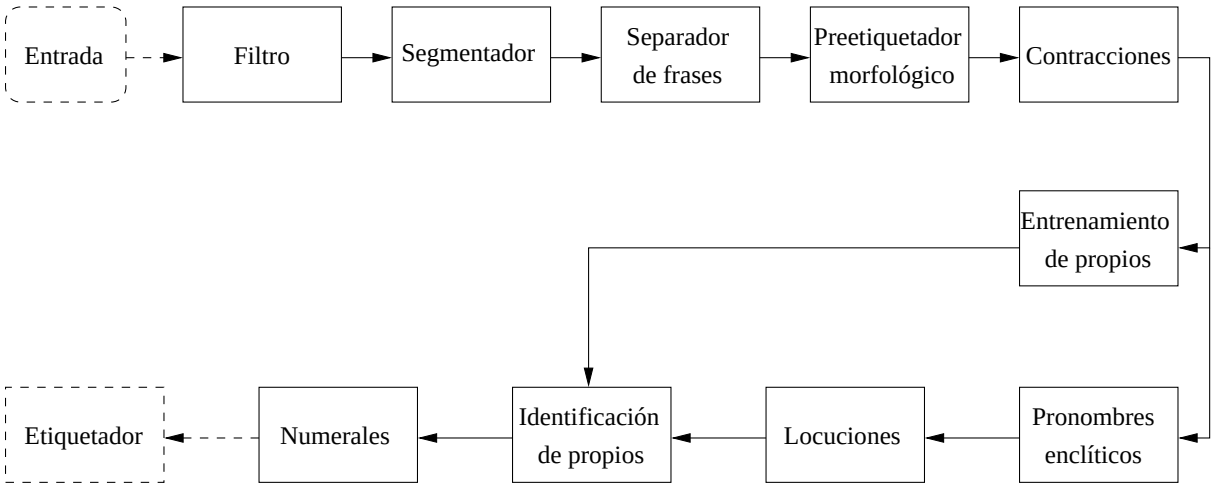


Figura 4.1: Estructura general del preprocesador

la palabra ortográfica, requeriría una fase de postprocesamiento después de la fase de etiquetación para poder identificar los diferentes componentes sintácticos del texto. Dicha fase de postprocesamiento realizaría labores análogas a las que involucra nuestro preprocesador.

4.3. Estructura del Preprocesador

El preprocesador que pasamos a describir ha sido diseñado para ser utilizado como fase previa a la etiquetación, centrándonos en el preprocesado de textos planos, es decir, se tratarán directamente los fenómenos lingüísticos anteriormente mencionados (contracciones, pronombres enclíticos, locuciones, etc.). Si bien es cierto que existen otros formatos de documentos que presentan otras problemáticas a resolver (eliminación de códigos HTML, SGML, XML, etc.), una vez resueltas, dan lugar a los mismos fenómenos que se afrontan aquí, ya que éstos son inherentes a la escritura del idioma.

El objetivo final perseguido es el desarrollo de un preprocesador modular, con algoritmos generales, de manera que pueda ser utilizado para diferentes idiomas, pero obteniendo un comportamiento mucho más fino mediante la introducción en el sistema de información lingüística relativa a una lengua en particular. Por lo tanto, resulta también de especial importancia la definición del tipo de información lingüística que va a resultar útil en dicho proceso, permitiendo su integración en el sistema en los casos en los que ésta esté disponible.

Además de esto, y como segundo objetivo, no hay que olvidar que al estar diseñado como fase previa a la etiquetación de textos, también deberá realizar tareas de preetiquetación. La idea subyacente consiste en que en un proceso de etiquetación-desambiguación completo, el módulo que más información tiene sobre algún fenómeno sea el que desambigüe dicho fenómeno.

A continuación se describen los diferentes módulos que incorpora el preprocesador. Dichos módulos son los que se muestran en la figura 4.1.

4.3.1. Filtro

Mediante el empleo de expresiones regulares, el módulo de filtro es el encargado de compactar los separadores redundantes que existen en el texto, es decir, eliminar múltiples espacios, espacios a inicio de frase, etc. Podrían también incluirse, asimismo, procesos de conversión de otros formatos —p.ej., HTML, SGML, XML, PDF, etc.— a texto plano, si bien dicha conversión podría también llevarse a cabo en una fase previa al preprocesamiento, solución por la que

hemos optado en nuestro sistema a la hora de procesar las etiquetas SGML de los documentos y consultas del corpus de evaluación.

4.3.2. Segmentador

La función del módulo segmentador es la de identificar y separar los *tokens* presentes en el texto, de manera que cada palabra ortográfica individual y cada signo de puntuación constituyan un *token* diferente. El módulo tiene en consideración la existencia de abreviaturas —p.ej., *etc.*—, siglas —p.ej., *CC.OO.*—, números con decimales —p.ej., *12,5*—, y fechas en formato numérico —p.ej., *12/10/1492*—, para no separar de los elementos anteriores y/o posteriores los signos de puntuación que contienen. Para ello se emplea un diccionario de abreviaturas, así como una serie de patrones y reglas heurísticas implementados en forma de expresiones regulares para la detección de siglas, números con decimales, y fechas en formato numérico.

4.3.3. Separador de Frases

Dado que nuestro etiquetador ha sido diseñado y entrenado para optimizar la secuencia de etiquetas de una frase, el preprocesador debe delimitar, por tanto, las frases del texto, siendo ésta la función de este módulo [95, 155, 153, 154]. Se trata de una fase que presenta numerosas dificultades, a pesar de su aparente sencillez. La regla general consiste en separar una frase ante un punto seguido de mayúscula[2], si bien existen otros casos a considerar, entre los que podemos señalar los siguientes:

- Abreviaturas a final de frase no seguidas por punto, como en la frase *"Traje queso, patatas, etc. Mi amigo trajo pollo"*.
- Signos de interrogación y admiración finalizando la frase no seguidos por punto. Por ejemplo: *"¿Cuándo llegaste? No te vi entrar"*.
- Empleo de puntos suspensivos como fin de enunciado. Por ejemplo: *"Dudé... Tenía miedo"*.
- Omisión ocasional del punto fin de frase tras una sigla. Por ejemplo: *"Trabaja en CC.OO. Da cursillos a parados"*.

Por otra parte, existen múltiples excepciones a tales casos que deben ser también tenidas en cuenta, como pueden ser:

- Existencia de abreviaturas especiales como las empleadas en el tratamiento formal, direcciones postales, etc., las cuales suelen ir acompañadas de mayúscula. Por ejemplo: *Sr. Vilares* y *avda. Fernández Latorre*.
- Empleo de abreviaturas en los nombres propios. Por ejemplo: *José L. Freire* y *Miguel A. Alonso*.
- Empleo de puntos suspensivos para introducir matices de intriga o duda. Por ejemplo: *"Me regaló... un coche"*.

Como podemos apreciar, la casuística no es en absoluto sencilla, y es fruto tanto de las propias normas de uso del lenguaje, como de su uso incorrecto por parte del escritor, motivando la necesidad de dotar al módulo de una gran robustez frente a los errores ortográficos.

[2]Seguimos, pues, un criterio estructural más que gramatical, al no separar, por ejemplo, las oraciones coordinadas entre sí o la subordinada de su principal.

Para desarrollar su función, el módulo emplea dos lexicones, uno de abreviaturas y otro de siglas, así como diversos patrones y reglas heurísticas que permiten identificar cada caso y resolverlo adecuadamente. Debemos tener en cuenta que la validez de tales reglas viene dada, en ocasiones, por el estilo y dominio de los documentos sobre los que el preprocesador trabajará, por lo que puede incurrir en errores ante casos particulares para los que no está preparado o que distan del uso más general.

4.3.4. Preetiquetador Morfológico

La función de este módulo es etiquetar aquellos elementos cuya etiqueta se puede deducir a partir de la morfología de la palabra, sin que exista otra manera más fiable de hacerlo. La identificación y etiquetación de dichos términos se realiza en base a diversos patrones y reglas heurísticas. De este modo, los números se etiquetan como `Cifra`, al igual que los porcentajes. De igual manera, se asigna la etiqueta `Fecha` a las fechas en formatos diversos tales como `7/4/82`, `7 de abril de 1982` o `7 de abril`. En estos últimos casos, se utiliza el símbolo `&` para unir los diferentes elementos que integran el *token*. Por ejemplo, `7 de abril de 1982` produciría la siguiente salida:

```
7&de&abril&de&1982      [Data 7&de&abril&de&1982]
```

donde los elementos entre corchetes hacen referencia a la etiqueta y al lema del *token* considerado.

4.3.5. Contracciones

El módulo de contracciones se encarga de desdoblar una contracción en sus diferentes *tokens*, etiquetando además cada uno de ellos. Para ello utiliza un diccionario externo que especifica cómo se deben descomponer dichas contracciones, de manera que sólo es necesario modificar esta información para adaptar este módulo a otro idioma.

Como ejemplo, la salida correspondiente a la contracción `del` es:

```
de      [X    de]
+el     [DAMS el]
```

Es decir, `del` ha sido descompuesta en la preposición `de` y en el artículo `+el`. Nótese que el símbolo `+` indica que se ha producido una segmentación de la palabra.

4.3.6. Pronombres Enclíticos

Este módulo se encarga de analizar los pronombres enclíticos que aparecen en las formas verbales. Este es un problema importante en algunas lenguas, y en particular para el español y el gallego, donde pueden aparecer varios pronombres enclíticos yuxtapuestos al verbo, hasta cuatro en el caso del gallego.

El objetivo que se persigue es el de separar el verbo de sus pronombres, etiquetando correctamente cada una de las partes. Se hace necesario, por tanto, ir más allá del mero análisis léxico, y aplicar técnicas similares a las propias de *lenguajes aglutinantes* tales como el alemán o el turco, que emplean algoritmos basados en el empleo combinado de heurísticas y lexicones [170].

Para realizar su función, nuestro módulo utiliza los siguientes elementos:

- Un diccionario que contiene el máximo número de formas verbales.
- Un diccionario con el máximo número de raíces verbales que pueden llevar pronombres enclíticos.

- Una lista con todas las combinaciones válidas de pronombre enclíticos.
- Una lista con todos los posibles pronombres enclíticos, junto con sus correspondientes etiquetas y lemas.

El proceso consiste en comprobar si una palabra puede ser un verbo con pronombres enclíticos. Para ello se analiza la palabra de izquierda a derecha, carácter a carácter, comprobando si se trata de una posible raíz verbal que pueda tener enclíticos. Si es así, se verifica si los caracteres restantes constituyen una combinación de enclíticos válida para esa raíz. En este caso, se procede a la segmentación y etiquetación de los correspondientes componentes.

Por ejemplo, la descomposición de `cógeselo` es:

```
cóge    [V2SRM coger]
+se     [PY3P  le] [PY3P  se] [PY3S  le] [PY3S  se]
+lo     [PY3S  lo]
```

donde se observa que los componentes son: `coge`, forma verbal de la segunda persona del singular del imperativo de `coger`; `+se`, pronombre personal de la tercera persona, con cuatro pares etiqueta/lema posibles dependiendo de si se trata de una flexión de `le` o de `se`, o de una forma singular o plural; y `+lo`, pronombre personal de la tercera persona del singular. Nótese que el preprocesador devuelve todas las alternativas etiqueta/lema posibles, ya que será el etiquetador quien resuelva la ambigüedad, eligiendo la más adecuada.

4.3.7. Locuciones

El módulo de locuciones se encarga de concatenar los *tokens* que componen una locución y de etiquetarlos como una unidad conjunta [51]. Para ello se emplean dos diccionarios de locuciones: uno con las locuciones que se sabe con seguridad que siempre son locuciones (morfológicamente hablando), y otro donde se encuentran las que pueden serlo o no. Por ejemplo, `en vez de` es una locución segura, mientras que `sin embargo` es insegura, ya que no podemos saber si estamos bien ante la locución o bien ante la secuencia formada por la preposición `sin` y el sustantivo `embargo`.

Una locución insegura conlleva una ambigüedad en la segmentación que no puede resolverse con la información de la que dispone el módulo a este nivel. Surge así la necesidad de un elemento de representación de este tipo de fenómenos con el objeto de que sea el módulo adecuado (en este caso, el etiquetador) el que elija una de las dos alternativas. Nuestro preprocesador representaría el ejemplo mencionado, `sin embargo`, de la siguiente manera:

```
<alternativa>
<alternativa1>
sin
embargo
</alternativa1>
<alternativa2>
sin&embargo
</alternativa2>
</alternativa>
```

Como se puede apreciar, se utiliza nuevamente el símbolo `&` para concatenar los diferentes elementos que forman un mismo *token*.

Este tipo de problema no se presenta únicamente en las locuciones, y profundizaremos algo más dentro de él en el apartado 4.3.10. No obstante, podemos adelantar que el etiquetador debe poder considerar este tipo de representación para proceder a la desambiguación.

Por otra parte, el preprocesador tampoco etiquetará una locución segura si ésta cuenta con más de una etiqueta posible de acuerdo con el diccionario. En estos casos el preprocesador concatenará en un único *token* la secuencia que forma la locución, pero dejará que sea el etiquetador quien la desambigüe y la etiquete. Este es el caso, por ejemplo, de la locución adjetiva `fuera de quicio`, que será un adjetivo calificativo masculino o femenino, singular o plural, dependiendo de a quién esté modificando.

4.3.8. Procesamiento de Nombres Propios

La identificación de nombres propios constituye una de las tareas más complejas del preprocesamiento, desde un punto de vista computacional. Ante la imposibilidad de disponer de un diccionario con todos los posibles nombres de personas, lugares y entidades, se optó por dotar al sistema de la capacidad de aprender los nombres propios que aparecen en los documentos a indexar.

Para ello, el proceso de identificación de nombres propios se divide en dos fases. En una primera *fase de entrenamiento* el sistema aprende los nombres propios no ambiguos contenidos en los documentos para, a continuación, proceder a la *fase de identificación* propiamente dicha. Ambos procesos son llevados a cabo por módulos diferentes.

Entrenamiento de Propios

Partiendo de los trabajos de Andrei Mikheev [152, 155, 153, 154], se lleva a cabo una fase de entrenamiento de nombres propios en la cual se identifican los nuevos nombres propios situados en posiciones no ambiguas del texto, es decir, aquellas palabras ubicadas en posiciones en donde la utilización de mayúsculas indica sin ambigüedad alguna que estamos ante un nombre propio. No se considerarían, por ejemplo, aquellas palabras en mayúscula que apareciesen inmediatamente después de un punto. Con estas palabras detectadas se genera un nuevo diccionario, denominado *de entrenamiento*, que será utilizado por el módulo siguiente.

Con el fin de poder identificar nombres propios complejos, se extraen también todas aquellas secuencias de palabras en mayúsculas que aparezcan interconectadas con alguno de los posibles nexos válidos —ciertas preposiciones y determinantes— previamente establecido, formando así un único *token*, de forma similar al procesamiento realizado en ciertos sistemas de Extracción de Información [103] y Búsqueda de Respuestas [231]. Debe tenerse en cuenta que, ante una secuencia de palabras en mayúsculas, no se puede determinar con seguridad si se trata de un único nombre propio o de una secuencia de nombres propios, por lo que deben considerarse todas las posibilidades. Como ejemplo, ante la secuencia `Consejo Superior de Cámaras de Comercio`, nuestro módulo generaría y almacenaría en el diccionario de entrenamiento los siguientes nombres propios válidos:

```
Consejo&Superior&de&Cámaras&de&Comercio

Consejo&Superior&de&Cámaras

Consejo&Superior

Superior&de&Cámaras&de&Comercio

Superior&de&Cámaras

Cámaras&de&Comercio
```

Frente a este planteamiento, Kwak et al. plantean en [133] la identificación y extracción de nombres propios en base a reglas generadas automáticamente a partir de corpus etiquetados preexistentes. No hemos seguido esta aproximación ya que las reglas de formación de nombres propios complejos pueden ser definidas manualmente de forma sencilla partiendo de que se trata de secuencias de palabras en mayúscula.

Identificación de Nombres Propios

A partir del diccionario de entrenamiento generado en la fase anterior, y empleando también el contenido de un diccionario externo de nombres propios, este segundo módulo etiqueta los nombres propios del texto, tanto simples como compuestos, y tanto en posiciones no ambiguas como ambiguas.

Para ello, en el caso de posiciones no ambiguas, se detecta en primer lugar el posible alcance del nombre propio (secuencias válidas que empiezan y terminan con una palabra en mayúscula). Si el alcance total o una subsecuencia de él se encuentran en el diccionario externo, se etiqueta con la etiqueta correspondiente del diccionario. Si por el contrario no existe una subsecuencia en el diccionario externo, se etiqueta como nombre propio pero sin especificar el género.

En el caso de una posición ambigua, se sigue un proceso similar. Se detecta el posible alcance del nombre propio. Si este alcance o una subsecuencia de él se encuentran en el diccionario externo, se le asigna su correspondiente etiqueta. Si por el contrario no existe ninguna subsecuencia en el diccionario externo, pero sí en el diccionario de entrenamiento de nombres propios, se etiqueta como nombre propio sin especificar el género. Y si finalmente no existe ninguna subsecuencia en ninguno de los diccionarios, este módulo no asigna ninguna etiqueta.

Así, por ejemplo, si aparece el nombre propio `Javier Pérez del Río`, y durante el entrenamiento sólo ha aparecido `Pérez del Río` en una posición no ambigua, pero además tenemos que `Javier` figura en el diccionario externo como nombre propio masculino singular, todo el nombre se etiquetaría como nombre propio masculino singular.

4.3.9. Numerales

Este módulo se encarga de la identificación de numerales compuestos mediante reglas heurísticas. Así, ante la aparición de un numeral compuesto, se concatenan sus componentes de la misma manera que una locución, produciendo un único *token*. Por ejemplo, ante el numeral `mil doscientas`, el preprocesador genera:

```
mil&doscientas      [DCFP mil&doscientas] [PCFP mil&doscientas]
```

En este caso el preprocesador ha identificado el numeral compuesto, proponiendo sus posibles etiquetas, siendo el etiquetador el encargado de desambiguar. Sin embargo, en el caso de las locuciones, como se ha visto anteriormente, el preprocesador no asigna posibles etiquetas, sino que simplemente genera todas las segmentaciones posibles, siendo el etiquetador el que seleccione y etiquete posteriormente la alternativa más probable.

4.3.10. Problemas Combinados

Para dar una idea de la complejidad de los problemas que debe afrontar el preprocesador, planteamos algunos casos especialmente complejos que se han resuelto:

Ejemplo 1

Un ejemplo de conflicto entre dos posibles descomposiciones de pronombres enclíticos aplicable al español sería el caso de `ténselo`, que puede tratarse bien de la forma verbal `tense`

(de `tensar`) más el enclítico `lo`, o bien la forma verbal `ten` (de `tener`) más dos enclíticos, `se` y `lo`. La salida devuelta por el preprocesador es:

```
<alternativa>
<alternativa1>
ténse    [V2SRM   tensar]
+lo      [PY3S    lo]
</alternativa1>
<alternativa2>
tén      [V2SRM   tener]
+se      [PY3P    le] [PY3P    se] [PY3S    le] [PY3S    se]
+lo      [PY3S    lo]
</alternativa2>
</alternativa>
```

Ejemplo 2

Recogemos excepcionalmente un ejemplo perteneciente al gallego, que merece ser incluido dada su especial complejidad. Se trata de la expresión `polo tanto` (*por lo tanto*/*por el tanto*). En este caso, estamos ante una locución insegura, ya que `polo tanto` puede ser una locución (*por lo tanto*), si bien considerados por separado `polo` a su vez puede ser un sustantivo (*pollo*), una contracción (de *por* y *el*) o un verbo con pronombres enclíticos (*ponerlo*), mientras que por su parte `tanto` puede ser sustantivo (*punto de juego*) o adverbio (*tanto*). La salida del preprocesador sería la siguiente:[3]

```
<alternativa>
<alternativa1>
polo     [Scms polo]
tanto
</alternativa1>
<alternativa2>
por      [P por]
+o       [Ddms o]
tanto
</alternativa2>
<alternativa3>
po       [Vpi2s0 pór] [Vpi2s0 poñer]
+o       [Raa3ms o]
tanto
</alternativa3>
<alternativa4>
por&+o&tanto
</alternativa4>
</alternativa>
```

[3]El conjunto de etiquetas empleado para el gallego pertenece al proyecto Corga (COrpus de Referencia del Gallego Actual). Las etiquetas empleadas en este ejemplo son:
`Ddms`: Artículo determinante masculino singular.
`P`: Preposición.
`Raa3ms`: Pronombre átono acusativo masculino, tercera persona del singular.
`Scms`: Sustantivo común masculino singular.
`Vpi2s0`: Verbo presente indicativo, segunda persona del singular.

Un ejemplo del empleo de las diferentes acepciones sería:

- **Sustantivo+Adverbio**: `Coméche-lo polo tanto, que non quedaron nin os osos` (*comiste el pollo tanto, que no quedaron ni los huesos*).
- **Preprosición+Artículo+Sustantivo**: `Gañaron o partido polo tanto da estrela` (*ganaron el partido por el tanto de la estrella*).
- **Verbo+Pronombre+Adverbio**: `Pois agora polo tanto ti coma el` (*pues ahora lo pones tanto tú como él*).
- **Locución**: `Estou enfermo, polo tanto quédome na casa` (*estoy enfermo, por lo tanto me quedo en casa*).

4.4. Adaptaciones para el Empleo en Recuperación de Información

A la hora de integrar el preprocesador en nuestro sistema de Recuperación de Información ha sido necesario realizar algunas modificaciones en el diseño inicial, que detallaremos a continuación. Finalmente, mostraremos un pequeño ejemplo práctico.

4.4.1. Modificaciones al Procesamiento de Locuciones

La primera de las modificaciones vino dada por la no disponibilidad actual de un etiquetador que soporte segmentaciones ambiguas, como es en el caso de las locuciones inseguras, ya tratado en el apartado 4.3.7. Tal y como veremos en el siguiente capítulo, si bien hemos desarrollado un planteamiento teórico, no existe todavía en explotación un etiquetador que lo implemente. Por esta razón hubo de desactivarse la opción de obtención de segmentaciones ambiguas, de tal modo que el preprocesador integrado en nuestro sistema, en vez de devolver las dos alternativas posibles en presencia de una locución insegura, únicamente devuelve la secuencia no concatenada. Debido a esto se modificaron también los diccionarios de locuciones, ampliando el diccionario de locuciones seguras con aquellas locuciones inseguras para las cuales su empleo como locución era marcadamente más frecuente, como es el caso de `sin embargo`.

4.4.2. Modificaciones al Procesamiento de Nombres Propios

La segunda de las modificaciones efectuadas al preprocesador se debió a la necesidad de modificar el comportamiento del módulo de tratamiento de nombres propios, ya que el planteamiento inicial penalizaba el rendimiento del sistema de Recuperación de Información final [32]. Dichas causas comprendían:

1. La presencia de diferentes variantes para designar una misma entidad. Por ejemplo, *George Bush*, *G. W. Bush* y *Bush*, hacen referencia a la misma persona, sin embargo el módulo de identificación de nombres propios generará diferentes términos para cada uno.

2. Entidades que son consideradas nombres comunes o nombres propios dependiendo de si se emplean minúsculas o mayúsculas, respectivamente. Por ejemplo, *secretaría general* y *Secretaría General*. Su consideración como nombre propio o no tiene importantes implicaciones, ya que en función de ello variará su procesamiento posterior (normalización de términos, análisis sintáctico, establecimiento de coocurrencias, etc.).

3. La indexación de entidades como nombres propios complejos no permite correspondencias parciales, como, por ejemplo, en el caso de *Ministerio de Educación y Cultura* frente a *Ministro de Educación* o *Ministro de Cultura*.

Para resolver estos problemas optamos por modificar el algoritmo de procesamiento de nombres propios para contemplar los componentes individuales de los nombres propios compuestos.

Algoritmo 4.1 Algoritmo de etiquetación de nombres propios modificado para su empleo en Recuperación de Información:

for all palabras en mayúscula W del nombre propio compuesto **do**
  if W está en el diccionario externo de nombres propios **then**
    W se etiqueta como *nombre propio*
  else if W está en el lexicón principal con etiqueta T **then**
    W se etiqueta como T
  else if (W está en posición no ambigua)
      or (W está en posición ambigua **and** W está en el diccionario de entrenamiento) **then**
    W se etiqueta como *nombre propio*
  else
    W se etiqueta como *desconocida*
  end if
end for

□

Aplicando este nuevo algoritmo, *George Bush* es ahora etiquetado como una secuencia de dos nombres propios; *Secretaría General* es etiquetado como una secuencia formada por un nombre común y un adjetivo calificativo; y *Ministerio de Educación y Cultura* es etiquetado como una secuencia formada por un nombre común, una preposición, un nombre común, una coordinación, y un nombre común.

Otros autores han investigado también, en el campo de la IR, el impacto del reconocimiento de nombre propios. Pfeifer et al. estudian en [177] el efecto de diferentes métodos de identificación de nombres propios. Thompson y Dozier tratan en [232] el efecto del procesamiento diferenciado de los nombres propios de persona respecto al procesamiento de otros términos. Dichos autores no indexan los nombres propios de persona contenidos en el texto, sino que identifican aquéllos que figuran en la consulta, para luego buscarlos en los documentos mediante operadores de proximidad: deben aparecer en el mismo orden, separados a lo sumo por dos términos que no sean *stopwords*. Ambas propuestas [177, 232] emplean textos parcialmente anotados manualmente, mientras que nuestra aproximación es totalmente automática.

4.5. Un Ejemplo Práctico

Mostramos finalmente un caso práctico en el que se preprocesa un pequeño texto de ejemplo, formateado en SGML de forma similar a como ocurre en los documentos de las colecciones de evaluación que emplearemos:

```
<TEXT>
Juan José pescó un salmón de 10.5 kg. y Pedro uno de 15 kg. Sin embargo, en vez
de vendérselos al pescadero... los liberaron.
</TEXT>
```

Tras aplicar previamente un módulo de conversión de formato SGML a texto plano, el texto es preprocesado, obteniendo a la salida:

```
##TEXT##

Juan    [Sp00    Juan]
José    [Sp00    José]
pescó
un
salmón
de
10.5    [Cifra   10.5]
kg.
y
Pedro   [Sp00    Pedro]
uno
de
15      [Cifra   15]
kg.


sin&embargo     [C       sin&embargo]
,
en&vez&de
vendér  [VRI     vender]
+se     [PY3P    le] [PY3P    se] [PY3S    le] [PY3S    se]
+los    [PY3P    lo]
a       [X       a]
+el     [DAMS    el]
pescadero
...
los
liberaron
.

##/TEXT##
```

Nótese que el formato de salida es el de un término por línea, separando cada frase por una línea en blanco.

4.6. Discusión

En este capítulo hemos presentado realizaciones concretas de mecanismos específicos de preprocesamiento y segmentación.Como se ha podido comprobar, la complejidad del tratamiento de muchos de los fenómenos que aparecen en este nivel es elevada, y por esta razón muchos de ellos suelen ser obviados en las aplicaciones finales.

La presentación ha sido orientada hacia la obtención de mejoras en la etiquetación robusta de los textos. No obstante, la utilización y necesidad de los mecanismos de preprocesamiento y segmentación no se limita únicamente a la desambiguación automática. El lugar del etiquetador

podría ser ocupado por cualquier otra herramienta de análisis (sintáctico, semántico, etc.), o simplemente por un *scanner* que proporcione todas las alternativas de segmentación y sus correspondientes etiquetaciones, permitiendo así realizar de manera más cómoda las labores de desambiguación manual en la creación de corpus de entrenamiento. Actualmente, este último uso está siendo explotado intensamente en el caso del gallego [31, 30], lengua para la cual la carencia de recursos lingüísticos es casi absoluta.

Por otra parte, es necesario el desarrollo de nuevas técnicas que permitan la realización tanto de las tareas de preprocesamiento que ya hemos contemplado aquí, como de otras todavía en estudio, pero en cualquier caso sin la utilización de una gran cantidad de recursos lingüísticos (que pueden no existir y ser muy costosos de construir). En caso de estar disponibles, estos recursos se utilizarían para un refinamiento del comportamiento global, pero no como un requisito.

Capítulo 5

Tratamiento de la Variación Morfológica Flexiva: Etiquetación y Lematización

5.1. Introducción

Dado un documento, las operaciones de texto efectuadas para la obtención de su representación lógica se han limitado, hasta ahora, a su segmentación en las frases y palabras lingüísticas que lo componen. Sin embargo, los términos de indexación deben ser todavía identificados y normalizados para así eliminar la variación lingüística presente en los mismos. Este proceso será el que detallemos en este y subsiguientes capítulos.

En el caso del inglés dicho proceso resulta sencillo, ya que la eliminación de *stopwords* y la normalización de los términos restantes mediante técnicas de *stemming* obtienen de por sí resultados aceptables, no siendo necesario en la mayoría de las veces ningún otro tipo de procesamiento lingüístico o pseudo-lingüístico más complejo. Esto se debe a la relativa sencillez morfológica, sobre todo a nivel flexivo, del inglés [121, 24].

La efectividad del *stemming* viene dada por la morfología de la lengua sobre la que se aplica, mostrando múltiples problemas de rendimiento ante lenguas con una morfología compleja o con muchas irregularidades [24, 233, 74]. En el caso del español, éste presenta fenómenos de una notable complejidad que requieren un mayor procesamiento, incluso a nivel flexivo. Dichos fenómenos incluyen modificaciones flexivas a múltiples niveles (de género y número en el caso de nombres y adjetivos, y de persona, número, tiempo y modo en el caso de los verbos), además de multitud de irregularidades [249]. En base a ello se pueden identificar en el español en torno a 20 grupos de variación para género y 10 para número en el caso de sustantivos y adjetivos, mientras que en el caso de los verbos existen 3 grupos regulares y unos 40 irregulares, con más de 100 formas flexionadas cada uno. Este nivel de complejidad no es abordable únicamente por medio del *stemming*. Por otra parte, la pérdida de información que implica el *stemming* del texto de cara a un procesamiento posterior desaconseja también su utilización [131].

Es por ello que en el caso de la normalización de textos en español, la *lematización*, consistente en la obtención de la forma canónica de una palabra o *lema*[1], se presenta como una alternativa conveniente al *stemming*, ya que permite abordar los complejos fenómenos flexivos que caracterizan esta lengua.

[1]Forma masculina singular en nombres y adjetivos e infinitivo en verbos.

5.2. El Etiquetador-Lematizador: MrTagoo

Una vez que han sido reconocidas las palabras que componen un texto, el primer paso en el proceso de lematización de un texto consiste en determinar su etiqueta correcta. Para ello es preciso desambiguar y etiquetar la salida suministrada por el preprocesador de base lingüística. De entre los diversos etiquetadores de alto rendimiento disponibles (TnT [37], Brill [38], TreeTagger [215], etc.), hemos optado por emplear MrTagoo [83], un etiquetador estocástico desarrollado en el seno de nuestro grupo de investigación y que presenta como características más destacadas una estructura de almacenamiento y búsqueda especialmente eficiente —gracias al empleo de autómatas de estado finito—, la posibilidad de integrar diccionarios externos en su marco probabilístico, el manejo de palabras desconocidas, y su capacidad lematizadora.

5.2.1. Características Generales

MrTagoo [83] es un etiquetador estocástico basado en un Modelo de Markov Oculto de segundo orden, en el cual los estados del modelo representan pares de etiquetas, y sus salidas, palabras [37]. En lo referente a las probabilidades asociadas a las transiciones, éstas dependen de los estados, es decir, de los pares de etiquetas, mientras que las probabilidades asociadas a las salidas dependen únicamente de la categoría más reciente. La secuencia de etiquetas más probable para una frase se obtiene aplicando el algoritmo de Viterbi [259, 75] sobre el enrejado correspondiente al modelo y calculando:

$$\arg \max_{t_1 \ldots t_n} \prod_{i=1}^{n} [P(w_i|t_i) \times P(t_i|t_{i-2}t_{i-1})]$$

donde $w_1 \ldots w_n$ es la frase de longitud n a etiquetar, y donde $t_1 \ldots t_n$ es una secuencia de elementos del conjunto de etiquetas o *tag set*, tal que t_i es la etiqueta seleccionada para la palabra w_i. Las probabilidades de transición y las probabilidades de salida de los estados del modelo son estimadas mediante un proceso de entrenamiento a partir de un corpus previamente desambiguado manualmente a tal efecto [37].[2]

Por otra parte, debe tenerse en cuenta que las probabilidades asociadas a los trigramas generados a partir del corpus de entrenamiento presentan un defecto de *dispersión de datos* o *sparse data*, es decir, existen trigramas posibles en la práctica para los cuales, o bien no hay ejemplos suficiente en el corpus como para que las probabilidades estimadas a partir de ellos sean fiables, o bien no existe ejemplo alguno en el corpus —lo que conllevaría probabilidades de transición nulas

Para eliminar este problema debemos recurrir a *métodos de suavización* o *smoothing*, los cuales permiten que, a partir de muestras pequeñas, se puedan estimar unas probabilidades que sean más representativas del comportamiento real de la población objeto de estudio. En nuestro caso, MrTagoo realiza una *interpolación lineal* de trigramas, bigramas y unigramas, de tal forma que la probabilidad de un trigrama $t_{i-2}t_{i-1}t_i$ se calcula como:

$$P(t_i|t_{i-2}t_{i-1}) = \lambda_3 \, P(t_i|t_{i-2}t_{i-1}) + \lambda_2 \, P(t_i|t_{i-1}) + \lambda_1 \, P(t_i)$$

donde todos los pesos λ_i deben ser no negativos y satisfacer la restricción $\lambda_1 + \lambda_2 + \lambda_3 = 1$, para que de este modo P pueda seguir representando una distribución de probabilidad. Los valores de λ_1, λ_2 y λ_3 son también estimados mediante un proceso de entrenamiento.

Finalmente, las frecuencias léxicas pueden ampliarse mediante palabras no contenidas en el corpus de entrenamiento y que serían aportadas por diccionarios externos, tal y como se comentará en el apartado 5.2.3.

[2] Los recursos lingüísticos empleados a tal efecto serán tratados en profundidad en el apartado 5.3.

5.2.2. Implementación mediante Autómatas de Estado Finito

El empleo de un paradigma de etiquetación estocástico, como ocurre en el caso de nuestro etiquetador, requiere de estructuras auxiliares —diccionarios— para el almacenamiento de las probabilidades asociadas a cada una de las combinaciones posibles palabra-etiqueta. Su implementación mediante gestores de bases de datos no es adecuada, ya que no sólo nos interesa un acceso flexible a los datos, sino también un acceso lo más rápido posible, que permita su utilización en entornos prácticos.

Una de las alternativas que se podría considerar sería la del empleo de un diccionario donde tan sólo se almacenasen las raíces de las palabras, utilizando luego un conjunto de reglas de concatenación de morfemas flexivos para obtener las palabras completas. Sin embargo, este enfoque no es viable si deseamos incorporar información adicional relativa a las palabras y no a las raíces, como es nuestro caso, ya que necesitamos almacenar las probabilidades correspondientes a los pares palabra-etiqueta.

MrTagoo emplea una implementación basada en autómatas finitos que permite obtener la representación más compacta que se puede diseñar para albergar toda la información léxica relativa a las palabras presentes en un diccionario [86]. Se trata, además, de una representación muy flexible que permite incorporar información adicional a los pares palabra-etiqueta de una forma sencilla; en nuestro caso particular, el lema correspondiente a dicho par, de cara a la lematización del texto —véase el apartado 5.2.5.

5.2.3. Integración de Diccionarios Externos

Las frecuencias léxicas empleadas en la estimación del modelo pueden ampliarse mediante palabras no contenidas en el corpus de entrenamiento y que son aportadas por diccionarios externos[3]. Esta información ha de integrarse, pues, dentro del marco probabilístico del etiquetador [88].

La forma más intuitiva de llevar a cabo esta integración es la de emplear el *método Adding-One* [55], que consiste en emplear el diccionario como un corpus etiquetado adicional donde cada par palabra-etiqueta ocurre sólo una vez, asignándosele una frecuencia de 1. No obstante, este método no da lugar a una representación coherente del modelo estimado, pudiendo generar alteraciones importantes en los parámetros de trabajo.

Por esta razón se decidió emplear el *método de Good-Turing* basado en las fórmulas del mismo nombre [9]. Conforme a esta técnica, cada par palabra-etiqueta presente únicamente en el diccionario puede considerarse como un evento de frecuencia nula en el corpus de entrenamiento que, sin embargo, puede ser tratado por las fórmulas Good-Turing de tal forma que le sea asignada una probabilidad mayor que 0, ya que se trata de un suceso que, aunque no ha sido observado en el corpus de entrenamiento, sí es un suceso posible en la práctica. Por otro lado, esta técnica genera un menor grado de distorsión en el modelo, incrementando así el rendimiento del proceso de etiquetado cuando el corpus de entrenamiento es reducido.

5.2.4. Tratamiento de Palabras Desconocidas

Las probabilidades de emisión de las palabras, tal y como hemos visto, se estiman a partir de sus apariciones bien en el corpus de entrenamiento, bien en los diccionarios externos. Sin embargo, a la hora de etiquetar nuevos textos, podemos dar por seguro que nos encontraremos

[3] A este respecto debe destacarse el hecho de que la escasez general de recursos lingüísticos de libre acceso para el español —y de corpus de entrenamiento en particular—, contrasta con la disponibilidad de amplios lexicones, gracias al esfuerzo realizado en fechas recientes en esta dirección.

con palabras que no estaban presentes ni en el corpus ni en los diccionarios, y que nos son, por tanto, desconocidas.

Para hacer frente a este problema, las etiquetas candidatas de una palabra desconocida y sus probabilidades se obtienen a partir de su terminación, de su *sufijo*[4]. La distribución de probabilidad de un sufijo particular se genera a partir de las palabras del corpus de entrenamiento que lo comparten. La probabilidades obtenidas se suavizan mediante una técnica denominada *de abstracción sucesiva* [206]. A partir de las probabilidades obtenidas se puede generar la distribución de probabilidades para cada etiqueta y cada longitud de sufijo, y así poder integrar dicha información dentro del marco probabilístico del modelo.

5.2.5. Capacidad de Lematización

La capacidad de lematización de nuestro etiquetador viene dada por la posibilidad de almacenar, para cada combinación palabra-etiqueta, sus posibles lemas. De este modo, una vez que el etiquetador ha desambiguado la palabra asignándole una etiqueta, basta con devolver el lema correspondiente.

No obstante, su capacidad de lematización no es plena, ya que si bien es capaz de determinar la etiqueta de una palabra desconocida, no puede establecer su lema al no disponer de información sobre dicha palabra en sus diccionarios. Actualmente, el lema de una palabra desconocida es su forma.[5]

5.2.6. Manejo de Segmentaciones Ambiguas

Dado su interés práctico, describiremos brevemente el desarrollo teórico necesario para el manejo de las segmentaciones ambiguas generadas por el preprocesador, aún en fase de implementación.

En primer lugar debe tenerse en cuenta que la existencia de segmentaciones ambiguas en la secuencia de *tokens* a procesar añade un nuevo nivel de complejidad a la etiquetación, ya que el etiquetador no sólo debe asignar la etiqueta apropiada a cada *token*, sino que debe también decidir si algunos de estos *token* forman o no un único término, asignando en cada caso el número de etiquetas apropiado en función de las alternativas propuestas por el preprocesador.

Una primera aproximación intuitiva al problema consiste en la evaluación individual de cada uno de los enrejados a los que da lugar cada una de las diferentes segmentaciones posibles y que integran el enrejado total, tomando como solución final a devolver aquella solución parcial más probable [87].

Tomemos como ejemplo la frase *"Juan fue a pesar de nuevo la fruta"*, un caso especialmente interesante, ya que existen dos ambigüedades de segmentación en *a pesar de* y *de nuevo*, las cuales a su vez se superponen en la palabra *de*. El enrejado completo correspondiente a este ejemplo se muestra en la parte superior de la figura 5.1 —en resaltado, el camino óptimo—, mostrándose a continuación los enrejados que lo integran, uno por cada alternativa posible para la segmentación.

Para llevar a cabo la comparación de las probabilidades obtenidas para cada una de estas segmentaciones posibles, debe fijarse un criterio objetivo de comparación. En el caso de los

[4]El término *sufijo*, tal y como se utiliza aquí, no se corresponde con su significado lingüístico, sino que se refiere simplemente a la *secuencia final de caracteres de la palabra*, lo cual a su vez requiere fijar previamente la longitud máxima de los sufijos a considerar.

[5]Se podrían ampliar las capacidades del módulo de tratamiento de palabras desconocidas mediante un proceso de descomposición de los afijos flexivos, que serían reemplazados por los afijos flexivos correspondientes al lema [121]. Sin embargo, debe ser tenido en cuenta el hecho de que una buena parte de las palabras desconocidas son debidas a errores ortográficos.

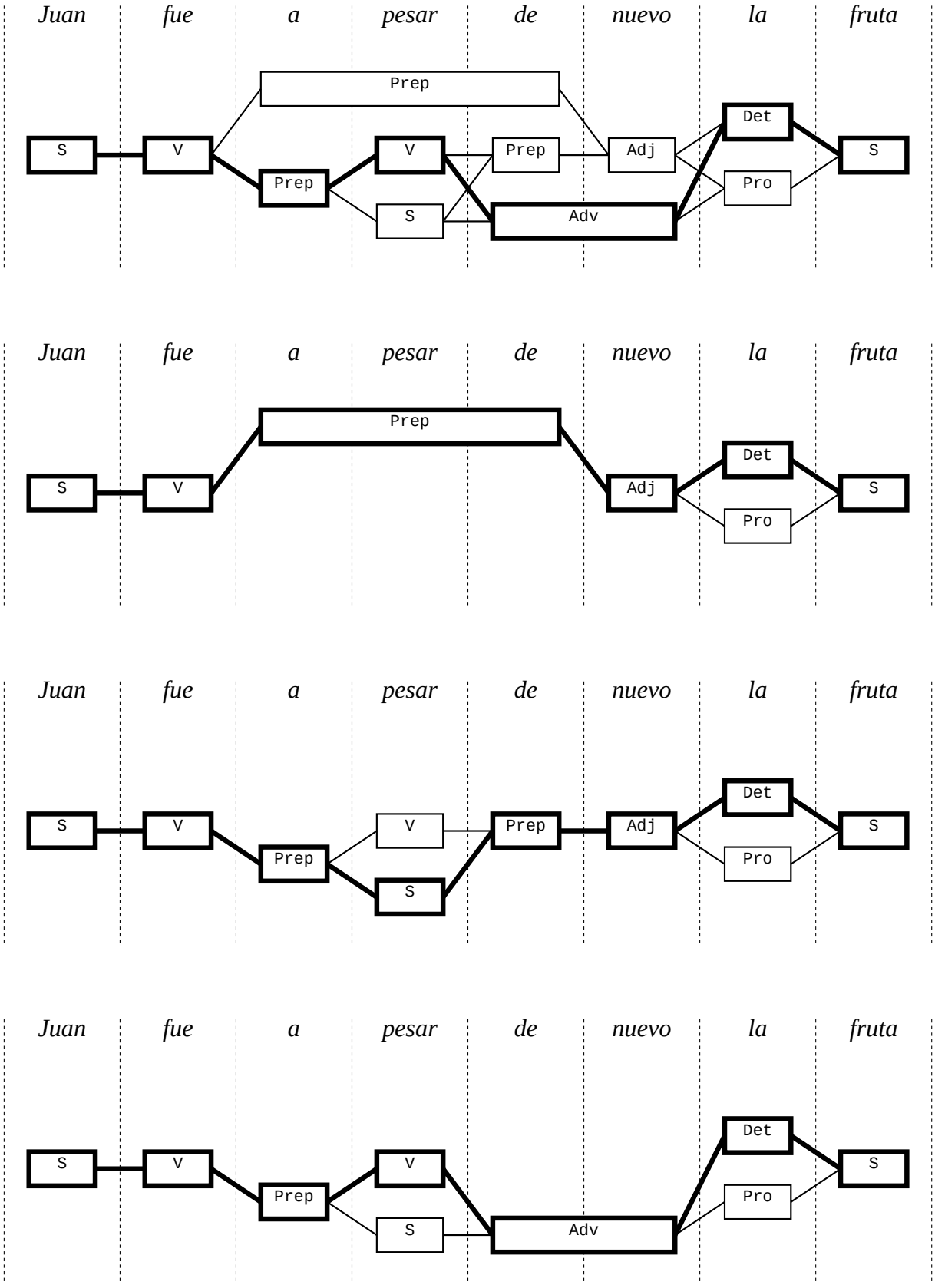


Figura 5.1: Enrejado completo y componentes para el ejemplo *"Juan fue a pesar de nuevo la fruta"*

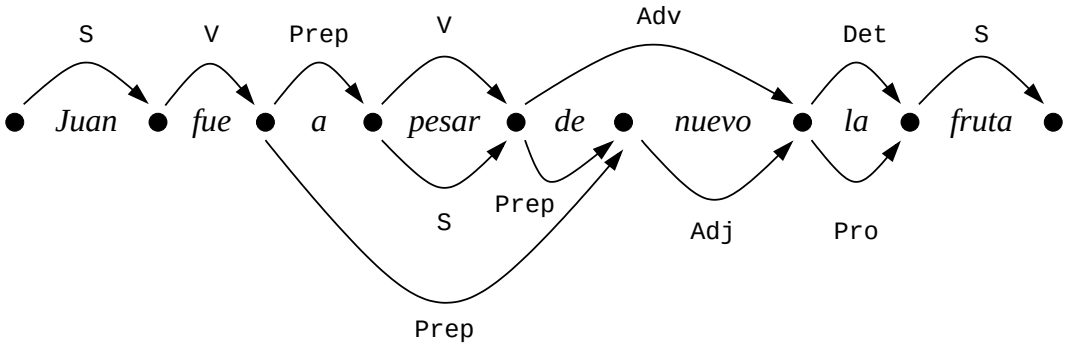


Figura 5.2: Diagrama de arcos del ejemplo *"Juan fue a pesar de nuevo la fruta"*

Modelos de Markov Ocultos, este criterio será el de la comparación de las probabilidades acumuladas una vez normalizadas[6]. Una de las razones que promovieron el empleo de Modelos de Markov Ocultos en nuestro sistema, respecto a otros paradigmas de etiquetación, fue, precisamente, la simplicidad de su criterio de comparación.

Sin embargo, esta solución, si bien válida, involucra un considerable aumento del coste computacional, ya que si, por ejemplo, otra de las palabras de nuestra frase tuviese 2 posibles segmentaciones, nos encontraríamos con $3 \times 2 = 6$ enrejados diferentes a evaluar. Por esta razón se optó por desarrollar una extensión del algoritmo de Viterbi empleado en el cálculo del camino óptimo dentro del enrejado, que fuese capaz de evaluar secuencias de *tokens* de longitudes diferentes dentro del mismo *trellis*, y cuya complejidad fuese comparable a la del algoritmo clásico.

Esta solución, desarrollada en el seno de nuestro grupo por Graña et al en [89], se asienta sobre dos fundamentos:

1. Operar sobre arcos en lugar de sobre enrejados, ya que aquéllos permiten representar de un modo mucho más cómodo e intuitivo las ambigüedades en la segmentación, especialmente en el caso de dependencias superpuestas. Mostramos en la figura 5.2 la estructura correspondiente al enrejado del ejemplo.

2. Efectuar las normalizaciones dinámicamente, es decir, sobre la marcha, y no al final del cálculo tras encontrar el camino óptimo. Dicha normalización se realiza en base a la longitud de los diferentes caminos encontrados durante la aplicación del algoritmo.

Esta extensión dinámica del algoritmo de Viterbi está siendo implementada, y pronto se procederá a su integración final en MrTagoo.

5.3. Recursos Lingüísticos: Proyecto Erial

Se incluyen en este apartado las descripciones de los recursos lingüísticos empleados por el etiquetador MrTagoo, y que contemplan, por una parte, un lexicón empleado como diccionario externo y por otra, un corpus de entrenamiento sobre el que estimar los parámetros del modelo. Estos recursos fueron desarrollados expresamente a tal efecto en el marco del proyecto Erial [31, 30].

[6]Sea p_i la probabilidad acumulada correspondiente al camino óptimo (resaltado) del i-ésimo enrejado componente de la figura 5.1. Dichos valores —p_1, p_2 y p_3— no son comparables directamente. Sin embargo, si empleamos probabilidades logarítmicas podemos normalizar sus valores diviéndolos por la longitud del camino —el número de *tokens* que contiene—. De esta forma, $p_1/6$, $p_2/8$ y $p_3/7$ son ya comparables.

categoría	*#formas*	*#lemas*
abreviatura	249	249
adjetivo	110118	22067
conjunción	171	166
determinante	530	168
pronombre	608	160
nombre común	131169	57002
nombre propio	1172	1115
signo de puntuación	14	14
verbo	361104	5472
adverbio	4054	4047
preposición	171	168
sigla	202	182
interjección	4	4
TOTAL	609566	90814

Tabla 5.1: Distribución por categorías de las formas y lemas del lexicón ERIAL

#etiquetas	*#formas*
1	379467
2	83164
3	15731
4	2151
5	1510
6	21
7	3
9	27
11	2
12	1
TOTAL	609566

Tabla 5.2: Distribución por número de etiquetas de las formas del lexicón ERIAL

5.3.1. El Lexicón ERIAL

A pesar de la amplia cobertura del lexicón original [31, 30], éste fue ampliado mediante la inclusión de nuevos gentilicios, topónimos, antropónimos, abreviaturas y siglas de cara a su utilización en nuestro sistema de Recuperación de Información. Las nuevas entradas fueron obtenidas a partir de fuentes de libre acceso disponible en Internet o bien generadas de forma semiautomática a partir del propio corpus de documentos —como en el caso de siglas y acrónimos.

La composición final del lexicón del sistema consta de 609566 palabras y 90814 lemas, distribuidas tal como se describe en la tabla 5.1. Los niveles de ambigüedad del lexicón —en función del número de etiquetas posibles para cada palabra—, se recogen en la tabla 5.2.

Por su parte, el juego de etiquetas empleado por el etiquetador, con 222 etiquetas posibles, se recoge en el Apéndice A.

#etiquetas	*#formas*	*#palabras*
1	3950	12857
2	158	2386
3	16	132
4	7	167
7	1	491
9	1	71
TOTAL	4133	16104

Tabla 5.3: Distribución de formas únicas y palabras en el corpus de entrenamiento ERIAL

5.3.2. El Corpus de Entrenamiento ERIAL

El corpus de entrenamiento empleado para la estimación de los parámetros del etiquetador fue desarrollado dentro del proyecto ERIAL [31, 30] empleando noticias de *El Correo Gallego*[7], periódico de ámbito regional perteneciente al grupo empresarial Editorial Compostela, una de las entidades privadas participantes en dicho proyecto.

El corpus consta de 16104 palabras, con una entrada por línea de sintaxis

```
forma etiqueta lema
```

y donde el fin de frase se indica mediante una línea en blanco. El corpus contiene 777 frases, lo que supone un número medio de 21 palabras por frase.

El lexicón o diccionario constituido por todas las formas que aparecen en el corpus está formado por 4133 formas correspondientes a 2991 lemas y con 4358 combinaciones forma-etiqueta posibles. La distribución de dichas formas por número de etiquetas posibles se recoge en la tabla 5.3. Si calculamos el porcentaje de formas ambiguas y el número medio de etiquetas por forma, obtenemos:

$$\%\text{ formas ambiguas} = \frac{\#\text{ formas ambiguas}}{\#\text{ formas}} \times 100 = \frac{4133 - 3950}{4133} \times 100 = 4{,}43\,\%$$

$$\#\text{ medio de etiquetas por forma} = \frac{\#\text{ etiquetas}}{\#\text{ formas}} = \frac{4358}{4133} = 1{,}05 \text{ etiquetas por forma.}$$

Si calculamos dichas cifras respecto al número de palabras del corpus en lugar de sobre el número de formas únicas, obtenemos la distribución de palabras por número de etiquetas posibles, recogida en la tabla 5.3, y los siguientes porcentajes de palabras ambiguas y número medio de etiquetas por palabra:

$$\%\text{ palabras ambiguas} = \frac{\#\text{ palabras ambiguas}}{\#\text{ palabras}} \times 100 = \frac{16104 - 12857}{16104} \times 100 = 20{,}16\,\%$$

$$\#\text{ medio de etiquetas por palabra} = \frac{\#\text{ etiquetas}}{\#\text{ palabras}} = \frac{22769}{16104} = 1{,}41 \text{ etiquetas por palabra.}$$

Finalmente, y a modo de ejemplo, mostramos a continuación las entradas correspondientes a una de las frases que componen el corpus.[8]

[7] `http://www.elcorreogallego.es`

[8] El juego de etiquetas empleado por el sistema se recoge, como ya se indicó anteriormente, en el Apéndice A.

```
la DAFS el
selección NCFS selección
femenina AQFS femenino
afronta V3SRI afrontar
esta DDFS este
tarde NCFS tarde
ante X ante
el DAMS el
Letonio Spms Letonio
el DAMS el
primer DRMS primer
partido NCMS partido
de X de
clasificación NCFS clasificación
para X para
el DAMS el
Europeo Spms Europeo
de X de
+el DAMS el
2001 Cifra 2001
. P .
```

5.4. Tratamiento de Frases en Mayúsculas

Una característica de las colecciones de evaluación en Recuperación de Información que tiene un impacto considerable en el rendimiento de las técnicas de indexación con base lingüística, es la presencia de un gran número de errores ortográficos en los documentos, aspecto común al corpus CLEF, tal y como indican Figuerola et al. en [74]. En particular, los títulos de documentos y secciones se escriben por lo general totalmente en mayúscula y sin signos ortográficos, confundiendo a los módulos de preprocesamiento y etiquetación, y produciendo, en consecuencia, etiquetaciones y lematizaciones erróneas. Esto constituye un problema ya que los títulos suelen ser muy indicativos del tema tratado en el documento. Sin embargo, no podemos dar por hecho que los títulos de los documentos del corpus se encuentran en mayúsculas, ya que la solución aplicada entonces no podría considerarse válida, puesto que no sería generalizable a otros textos.

Existen aproximaciones previas sobre recuperación de signos ortográficos [270, 269, 151, 78], capaces incluso de manejar documentos en los que éstos han sido eliminados por completo. Sin embargo, en nuestro caso la práctica totalidad del documento está escrita correctamente, respetando minúsculas y signos, y sólo algunas frases están en mayúscula, con lo que el propio texto se convierte en una fuente fiable de información contextual. Por otra parte, nuestro campo de aplicación es diferente al de las soluciones clásicas, ya que no es necesaria la recuperación de la forma flexionada original, sino sólo la de su etiqueta[9] y su lema, con lo que aspectos como el de la identificación del tiempo verbal correcto, de resolución bastante compleja [270, 269], pueden obviarse.

Es por ello que hemos desarrollado una solución propia, incorporando un módulo para el procesado de secuencias en mayúsculas. Dicho módulo detectará automáticamente este tipo de entradas, pasándolas a minúscula de ser preciso y recuperando los signos ortográficos necesarios.

[9] En concreto sólo su categoría gramatical.

Para ello emplearemos el contexto léxico de la frase en mayúsculas, a partir del cual obtendremos la información necesaria para la recuperación de etiquetas y lemas.

Para evitar, en la medida de lo posible, interferencias de siglas, nombres de entidades, etc., sólo serán procesadas aquellas frases en mayúsculas de longitud mayor o igual que tres y en las cuales al menos tres palabras tengan más de tres caracteres. A cada una de estas secuencias, fácilmente identificables, se aplica el algoritmo de recuperación.

Algoritmo 5.1 Algoritmo de recuperación de frases en mayúsculas:

1. Se obtiene el contexto circundante a la frase de entrada (3 frases).
2. Para cada palabra de la frase de entrada se examina el contexto en busca de entradas con su misma *forma plana*[10], constituyendo éstas su conjunto asociado de candidatas:
   - *a*) Si existe alguna candidata, se toma la que cuenta con más apariciones en el contexto y, en caso de empate, aquélla con una aparición más cercana a la frase.
   - *b*) Si no existen candidatas, se consulta el lexicón buscando aquéllas entradas cuya forma plana sea igual a la de la palabra procesada[11], que se convertirán en candidatas, agrupándolas de acuerdo con su categoría gramatical y lema (la forma no es de interés):
     1) Si existe alguna candidata, se toma de nuevo la más numerosa y, en caso de empate, aquélla con una ocurrencia más cercana a la frase.
     2) Si no existen candidatas, se mantienen la etiqueta y el lema actuales.

   - En caso de que la palabra procesada de la frase en mayúsculas preserve alguno de sus signos ortográficos —por ejemplo la *'˜'* en una *'Ñ'*—, las candidatas deberán observar dichas restricciones.

□

Para mostrar el funcionamiento de este módulo, tomaremos como ejemplo el siguiente documento del corpus:

```
<DOC>
<DOCNO>EFE19940101-00016</DOCNO>
<DOCID>EFE19940101-00016</DOCID>
<DATE>19940101</DATE>
<TIME>08.53</TIME>
<SCATE>POX</SCATE>
<FICHEROS>94F.JPG</FICHEROS>
<DESTINO>MUN EXG</DESTINO>
<CATEGORY>POLITICA</CATEGORY>
<CLAVE>DP2446</CLAVE>
<NUM>196</NUM>
<PRIORIDAD>U</PRIORIDAD>
<TITLE>
           EMPERADORES VIAJARAN A ESPAÑA PROXIMO OTOÑO
</TITLE>
<TEXT>    Tokio, 1 ene (EFE).- Los emperadores de Japón, Akihito y Michiko,
```

[10] Forma en minúscula y sin signos ortográficos; p.ej., *gonzalez* por *González*.

[11] Este proceso ha sido también implementado mediante autómatas de estado finito para así minimizar su impacto en el rendimiento.

```
realizarán una visita oficial a España el próximo otoño en el ámbito
de una gira por Europa, informaron hoy, sábado, fuentes oficiales.
   La pareja imperial viajará también a Francia y a Suiza en las
mismas fechas, aún sin determinar, pero en el otoño.
   Fuentes del Gobierno indicaron que el emperador Akihito y la
emperatriz Michiko modificaron sus planes de visitar Inglaterra en la
misma ocasión debido a las ocupaciones de la Reina Isabel II.
   La gira europea de Akihito y Michiko durará diez días. En junio
visitarán Estados Unidos, en donde permanecerán durante dos semanas.
   Será la primera vez que los emperadores de Japón visiten España,
país que ya han conocido dos de los tres hijos de Akihito y Michiko.
   El príncipe heredero Naruhito ha estado en varias ocasiones en
España, invitado por los Reyes Juan Carlos y Sofía, en la apertura de
los Juegos Olímpicos de Barcelona, y en el día dedicado a Japón en la
Exposición Universal de Sevilla.
   Por su parte, la princesa Nori pasó dos veranos en la residencia
estival de los Reyes de Bélgica en la localidad andaluza de Motril
(sur de España). EFE
   CD/vv/ha
   01/01/08-53/94
</TEXT>
```

Como podemos ver, el título de la noticia, EMPERADORES VIAJARAN A ESPAÑA PROXIMO OTOÑO, está enteramente en mayúsculas y sin tildes, de tal modo que, al ser preprocesado y etiquetado, obtendríamos una salida incorrecta:[12]

```
emperadores     NCMP    emperador
VIAJARAN        Zgms    *
A               NCMP    *
ESPAÑA          Zgms    *
PROXIMO         Zgms    *
OTOÑO           Zgms    *
```

Sin embargo, partiendo del contexto circundante:

```
<TITLE>
          EMPERADORES VIAJARAN A ESPAÑA PROXIMO OTOÑO
</TITLE>
<TEXT>    Tokio, 1 ene (EFE).- Los [emperadores] de Japón, Akihito y Michiko,
 realizarán una visita oficial [a] [España] el [próximo] [otoño] en el ámbito
 de una gira por Europa, informaron hoy, sábado, fuentes oficiales.
    La pareja imperial viajará también [a] Francia y [a] Suiza en las
 mismas fechas, aún sin determinar, pero en el [otoño].
    Fuentes del Gobierno indicaron que el emperador Akihito y la
 emperatriz Michiko modificaron sus planes de visitar Inglaterra en la
 misma ocasión debido a las ocupaciones de la Reina Isabel II.
                              ...
```

y tomando como propias las etiquetas y lemas asignadas a las formas correspondientes provenientes del contexto, obtendríamos una solución bastante aproximada:

```
emperadores     NCMP    emperador
VIAJARAN        Zgms    *
a               X       a
```

[12] El símbolo * indica 'lema desconocido'.

```
España          Sp00    España
próximo         AQMS    próximo
otoño           NCMS    otoño
```

La forma `VIAJARAN` no pudo ser recuperada pues no existe entrada alguna en el contexto con su misma forma plana. Sin embargo, al interrogar al lexicón, encontramos dos entradas cuya forma plana coincide con la deseada, ambas formas verbales de `viajar`:

```
viajarán        V3PFI   viajar
viajaran        VM3PIS  viajar
```

lo que nos permite recuperar también la forma restante. En este caso, si bien cualquiera de las dos soluciones posibles nos hubiera valido (ambas verbos y del mismo lema), el sistema toma por defecto la primera de las formas, que en este caso resulta ser la solución exacta:

```
emperadores     NCMP    emperador
viajarán        V3PFI   viajar
a               X       a
España          Sp00    España
próximo         AQMS    próximo
otoño           NCMS    otoño
```

Hemos recuperado, pues, las categorías gramaticales y los lemas correctos para las entradas de la frase en mayúsculas, respetando los signos ortográficos (p.ej., *próximo*) y el uso de minúsculas y mayúsculas (p.ej., España).

Debemos señalar además que, en el caso particular del corpus CLEF, tal y como muestra este ejemplo, los títulos no sólo están escritos totalmente en mayúsculas, sino que además suelen omitir artículos y preposiciones, formando oraciones no del todo correctas gramaticalmente. Este fenómeno, si bien hubiera perjudicado estrategias de recuperación de signos ortográficos basadas en el empleo etiquetadores [78], no afecta, sin embargo, a nuestra aproximación.

5.5. Identificación y Normalización de Términos

Llegados a este punto, nos encontramos con que el texto ha sido *tokenizado* y segmentado mediante el preprocesador de base lingüística, para ser luego desambiguado y lematizado mediante el etiquetador-lematizador MrTagoo. Sin embargo, resta todavía identificar los términos de indexación.

En esta primera propuesta para la normalización de textos en español mediante técnicas de Procesamiento del Lenguaje Natural se ha optado por una aproximación a nivel léxico, en la que los términos de indexación sean los lemas de las *palabras con contenido* del texto [114]: sustantivos, verbos o adjetivos. El propio concepto de *palabra con contenido* hace referencia a que es en estas palabras en las cuales se concentra el contenido semántico del texto [114, 130], frente a las restantes categorías gramaticales —determinantes, preposiciones, etc.—, que son las encargadas de organizar y estructurar dicho contenido[13]. Al emplear como términos de indexación los lemas de las palabras con contenido del texto se persigue un doble objetivo:

- Capturar el contenido semántico del texto procesado, pues es en estas palabras donde se encuentra.

[13]Es por ello que, realmente, sólo precisamos conocer la categoría gramatical de la palabra, y no su etiqueta completa.

- Eliminar la variación lingüística de carácter morfológico flexivo, ya que pasaremos a trabajar sobre el lema, la forma canónica de la palabra.

Una vez que el término de indexación ha sido identificado, éste es sometido a un segundo proceso de normalización con vistas a incrementar la robustez del sistema frente a los múltiples errores ortográficos presentes todavía en los textos, mayormente el uso incorrecto de tildes y mayúsculas. Para ello aplicaremos soluciones comunes al caso de los *stemmers*, como son la eliminación de signos ortográficos y el paso de mayúsculas a minúsculas.

Si bien desde un punto de vista estrictamente lingüístico la eliminación de los signos ortográficos de un texto —tildes, diéresis, etc.— conlleva una pérdida de información, en el marco de aplicación actual no existe razón alguna para su preservación. En el caso de las tildes, aquéllas correspondientes a los sufijos flexivos de conjugación verbal —p.ej., *cantará* y *cantara*— han desaparecido con la lematización, ya que hemos pasado a trabajar con la forma canónica —*cantar*. Por otra parte, podría argumentarse que la eliminación de tildes incluye también a las *tildes diacríticas*, aquéllas que diferencian palabras con igual forma gráfica y significado diferente —p.ej., *té* (bebida) vs. *te* (pronombre personal)—. Sin embargo, no existe tampoco necesidad alguna de preservarlas, puesto que no se da ningún caso en el que una tilde diacrítica diferencie dos lemas de palabras con contenido. Se puede concluir, por tanto, que las tildes no son ya útiles una vez se ha lematizado el texto. Similar es también el caso de las diéresis y otros signos ortográficos, de uso mucho menos frecuente en el español y, por lo tanto, de escasa incidencia. Por otra parte, sí conservaremos las *'ñ'* del texto, no convirtiéndolas en *'n'*, pues si bien es frecuente cometer errores en el caso de las tildes, una confusión entre *'ñ'* y *'n'* sería ciertamente inusitada, ya que la *'ñ'* es una letra con identidad propia, distinta a la *'n'*, y no una *'n'* con un signo *'˜'* encima. Su eliminación, por tanto, podría introducir ruido en el sistema al normalizar palabras como *cana* y *caña* en un mismo término.

Una vez concluido este proceso de normalización, el texto es indexado previa eliminación de las *stopwords* que contiene. En el caso de la indexación de los lemas de las palabras con contenido del texto, gran parte de las palabras que constituyen una lista de *stopwords* —determinantes, preposiciones, adverbios, etc.— no son siquiera consideradas para su indexación, puesto que sólo nombres, verbos y adjetivos son tenidos en cuenta. Sin embargo, restan por eliminar todavía términos de escasa significatividad como pueden ser verbos auxiliares —p.ej., *ser*, *haber*—, palabras de uso generalizado —p.ej., *decir*, *ejemplo*—, etc. Dichas palabras son eliminadas conforme a una lista de *stopwords* constituida por los lemas de las palabras con contenido —convenientemente normalizados— de la lista de *stopwords* inicial empleada para *stemming*. Dichas listas se recogen en el apéndice B.

5.6. Resultados Experimentales

5.6.1. Resultados sin Realimentación

La tabla 5.4 muestra los resultados obtenidos mediante nuestra propuesta basada en la lematización (*lem*), señalando para cada caso el tanto por ciento de mejora obtenido ($\%\Delta$) frente al *stemming* (*stm*), nuestra línea base —remarcando en negrita los casos en los que se produce un incremento del rendimiento. Como se puede apreciar, el empleo de la lematización como técnica de normalización conlleva, en general, un incremento del rendimiento a todos los niveles, exceptuando en algunos casos la precisión en los niveles más bajos de cobertura y en los primeros documentos devueltos. En ocasiones se produce también una leve disminución de la precisión en los últimos niveles de cobertura y documentos devueltos, pero este fenómeno guarda poco interés ya que estos documentos rara vez son tenidos en cuenta por el usuario.

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*stm*	*lem*	%Δ	*stm*	*lem*	%Δ	*stm*	*lem*	%Δ	*stm*	*lem*	%Δ	*stm*	*lem*	%Δ	*stm*	*lem*	%Δ
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2719	2700	-0.70	2768	2762	-0.22	2345	**2363**	**0.77**	2401	**2410**	**0.37**	2137	**2159**	**1.03**	2181	2175	-0.28
Pr. no int.	.4720	**.4829**	**2.31**	.5166	**.5239**	**1.41**	.4674	**.4678**	**0.09**	.5095	**.5256**	**3.16**	.4304	**.4324**	**0.46**	.4545	**.4659**	**2.51**
Pr. doc.	.5155	**.5327**	**3.34**	.5598	**.5690**	**1.64**	.5541	**.5667**	**2.27**	.5797	**.6089**	**5.04**	.4632	**.4724**	**1.99**	.5131	**.5201**	**1.36**
R-pr.	.4599	**.4848**	**5.41**	.4955	**.5075**	**2.42**	.4584	.4549	-0.76	.4865	**.5030**	**3.39**	.4479	.4362	-2.61	.4457	**.4493**	**0.81**
Pr. a 0 %	.8443	.8293	-1.78	.9206	.8845	-3.92	.8645	.8234	-4.75	.8807	.8763	-0.50	.7881	**.8124**	**3.08**	.8151	**.8366**	**2.64**
Pr. a 10 %	.7361	**.7463**	**1.39**	.7788	**.7914**	**1.62**	.6937	.6845	-1.33	.7618	.7517	-1.33	.7002	**.7057**	**0.79**	.7034	**.7197**	**2.32**
Pr. a 20 %	.6377	**.6771**	**6.18**	.6892	**.7136**	**3.54**	.6280	**.6287**	**0.11**	.6874	**.7095**	**3.22**	.6099	**.6118**	**0.31**	.6301	**.6453**	**2.41**
Pr. a 30 %	.5769	**.6138**	**6.40**	.6470	**.6591**	**1.87**	.5723	**.5829**	**1.85**	.6178	**.6463**	**4.61**	.5610	.5503	-1.91	.5744	**.5788**	**0.77**
Pr. a 40 %	.5351	**.5502**	**2.82**	.5891	**.6085**	**3.29**	.5394	**.5555**	**2.98**	.5810	**.6100**	**4.99**	.4932	**.5059**	**2.58**	.5258	**.5348**	**1.71**
Pr. a 50 %	.4812	**.4931**	**2.47**	.5385	**.5557**	**3.19**	.4979	**.5136**	**3.15**	.5337	**.5658**	**6.01**	.4464	**.4657**	**4.32**	.4720	**.4889**	**3.58**
Pr. a 60 %	.4489	**.4496**	**0.16**	.4910	**.5006**	**1.96**	.4356	**.4413**	**1.31**	.4828	**.4977**	**3.09**	.3961	**.4017**	**1.41**	.4206	**.4339**	**3.16**
Pr. a 70 %	.3776	**.3853**	**2.04**	.4102	**.4161**	**1.44**	.4078	.3943	-3.31	.4429	**.4569**	**3.16**	.3378	**.3422**	**1.30**	.3710	**.3789**	**2.13**
Pr. a 80 %	.3246	**.3277**	**0.96**	.3533	.3509	-0.68	.3126	.3122	-0.13	.3427	**.3679**	**7.35**	.2652	**.2700**	**1.81**	.3138	**.3206**	**2.17**
Pr. a 90 %	.2410	.2356	-2.24	.2522	.2492	-1.19	.2362	.2319	-1.82	.2825	**.2845**	**0.71**	.1909	.1821	-4.61	.2060	**.2067**	**0.34**
Pr. a 100 %	.1169	**.1197**	**2.40**	.1194	**.1289**	**7.96**	.1158	**.1209**	**4.40**	.1422	**.1547**	**8.79**	.1008	.0932	-7.54	.1141	**.1164**	**2.02**
Pr. a 5 docs.	.6391	**.6609**	**3.41**	.6913	**.6957**	**0.64**	.6089	.5956	-2.18	.6933	.6844	-1.28	.5745	**.5915**	**2.96**	.6340	.6213	-2.00
Pr. a 10 docs.	.5935	**.6283**	**5.86**	.6304	**.6543**	**3.79**	.5400	**.5667**	**4.94**	.5844	**.6000**	**2.67**	.5426	.5149	-5.11	.5681	.5596	-1.50
Pr. a 15 docs.	.5551	**.5928**	**6.79**	.6014	**.6188**	**2.89**	.5081	**.5170**	**1.75**	.5452	**.5689**	**4.35**	.4908	.4738	-3.46	.5191	.5106	-1.64
Pr. a 20 docs.	.5174	**.5446**	**5.26**	.5685	**.5880**	**3.43**	.4878	.4878	0.00	.5200	**.5422**	**4.27**	.4468	.4457	-0.25	.4787	**.4819**	**0.67**
Pr. a 30 docs.	.4710	**.4928**	**4.63**	.5138	**.5304**	**3.23**	.4422	**.4452**	**0.68**	.4770	**.4948**	**3.73**	.3986	**.4000**	**0.35**	.4149	**.4255**	**2.55**
Pr. a 100 docs.	.3157	**.3300**	**4.53**	.3396	**.3509**	**3.33**	.2922	**.2993**	**2.43**	.2998	**.3147**	**4.97**	.2477	**.2513**	**1.45**	.2647	**.2702**	**2.08**
Pr. a 200 docs.	.2186	**.2234**	**2.20**	.2310	**.2315**	**0.22**	.1979	**.1997**	**0.91**	.2059	**.2093**	**1.65**	.1611	**.1617**	**0.37**	.1688	**.1694**	**0.36**
Pr. a 500 docs.	.1097	.1090	-0.64	.1132	.1115	-1.50	.0980	**.0983**	**0.31**	.1013	**.1023**	**0.99**	.0813	**.0827**	**1.72**	.0836	**.0842**	**0.72**
Pr. a 1000 docs.	.0591	.0587	-0.68	.0602	.0600	-0.33	.0521	**.0525**	**0.77**	.0534	**.0536**	**0.37**	.0455	**.0459**	**0.88**	.0464	.0463	-0.22

Tabla 5.4: Resultados obtenidos mediante *stemming* (*stm*), caso base, y lematización (*lem*)

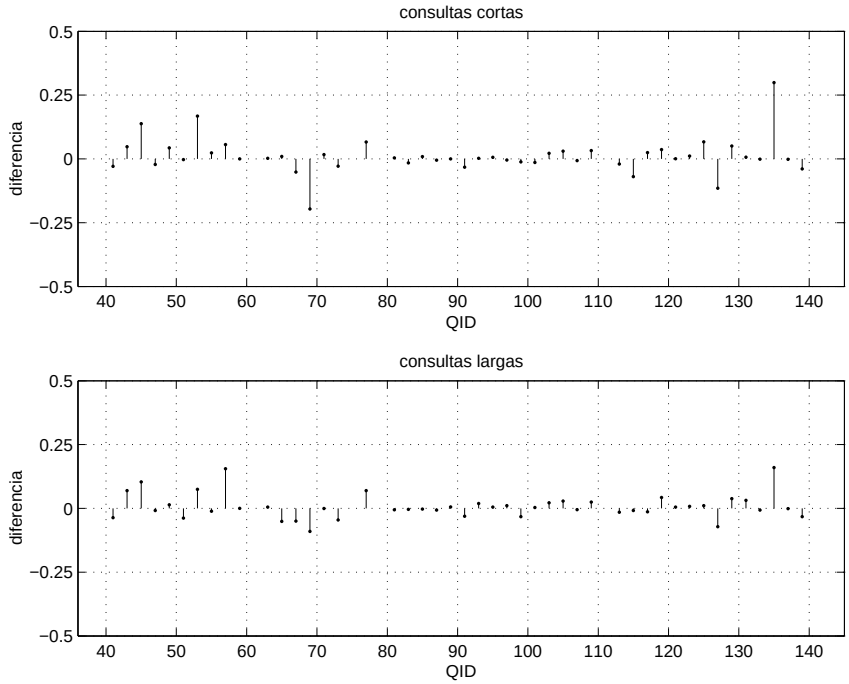


Figura 5.3: Diferencias en las precisiones no interpoladas: *stemming* vs. lematización. Corpus CLEF 2001-02·A

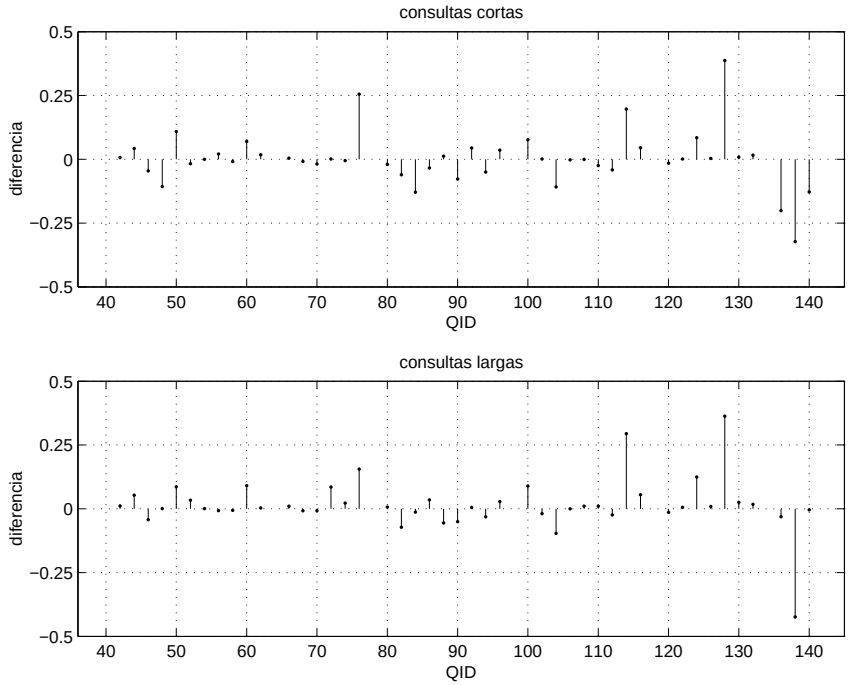


Figura 5.4: Diferencias en las precisiones no interpoladas: *stemming* vs. lematización. Corpus CLEF 2001-02·B

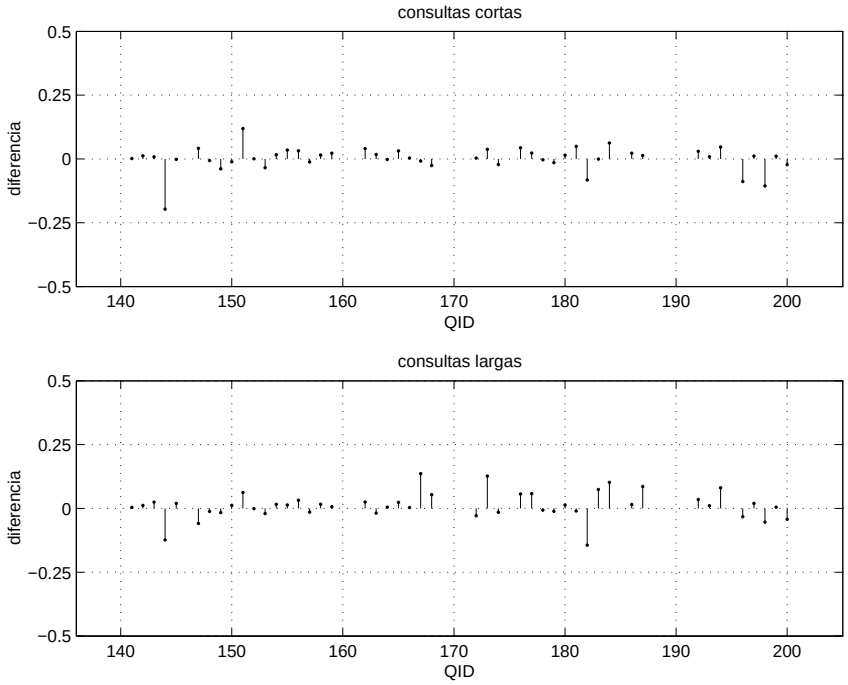


Figura 5.5: Diferencias en las precisiones no interpoladas: *stemming* vs. lematización. Corpus CLEF 2003

En el caso del primer corpus empleado, el CLEF 2001-02·A, la tabla 5.4 muestra una clara mejora —especialmente para el caso de las consultas cortas—, exceptuando un ligero descenso en el número de documentos relevantes devueltos y en la precisión obtenida para el nivel más bajo de cobertura (0 %). Las diferencias obtenidas para cada consulta en sus precisiones no interpoladas se recogen en la figura 5.3. Tal y como se puede apreciar, en el caso de las consultas cortas, existe una tendencia a la mejora de los resultados, especialmente en las primeras y últimas consultas, diferencias que disminuyen en el caso de las consultas largas.

Con respecto a los resultados obtenidos para el corpus CLEF 2001-02·B, la mejora obtenida mediante la sustitución del *stemming* por la lematización es, de nuevo, general, si bien menor en el caso de las consultas cortas, como podemos apreciar en el caso de la precisión no interpolada, donde el impacto de la lematización es escaso. Asimismo, se produce de nuevo un ligero descenso de la precisión en los niveles más bajos de cobertura y en los 5 primeros documentos devueltos, de incidencia algo mayor en el caso de las consultas cortas. Por el contrario, la mejora obtenida para las consultas largas es evidente, y mayor incluso que la obtenida para el corpus anterior. En lo referente a las consultas individuales, la figura 5.4 recoge, para las consultas cortas, varios picos de mejora, que son contrarrestados por algunas caídas en la precisión de las últimas consultas. La mayoría de estos picos negativos son eliminados, junto con otras caídas menores, al introducir información procedente del campo narrativa, siendo en este caso la lematización la técnica más beneficiada.

Para el último corpus de prueba, el CLEF 2003, su comportamiento para el caso de las consultas cortas es uno de los menos beneficiados por el empleo de la lematización, ya que para el caso de la precisión a los primeros documentos devueltos, si bien existe un aumento claro a los 5 documentos, se produce un descenso hasta llegar a los 20 primeros. En el caso de emplear consultas largas la mejora es general y bastante más amplia, si bien existe también una disminución de la precisión hasta los 15 primeros documentos devueltos. La gráfica de las diferencias de precisiones a nivel de consulta, figura 5.5, viene a confirmar estas observaciones,

ya que las diferencias obtenidas para las consultas cortas son mínimas, apreciándose luego una clara mejora con el empleo de consultas largas.

5.6.2. Introducción de la Realimentación: Estimación de Parámetros

Para analizar el comportamiento de la lematización cuando se aplica realimentación por relevancia, se realizó una nueva tanda de experimentos en los que las consultas fueron expandidas automáticamente aplicando la implementación de SMART para el método de Rocchio [196], en base a la expresión:

$$Q_1 = \alpha Q_0 + \beta \sum_{k=1}^{n_1} \frac{R_k}{n_1} - \gamma \sum_{k=1}^{n_2} \frac{S_k}{n_2}$$

donde Q_1 es el nuevo vector de la consulta
Q_0 es el vector de la consulta inicial
R_k es el vector del documento relevante k
S_k es el vector del documento no relevante k
n_1 es el número de documentos relevantes examinados
n_2 es el número de documentos no relevantes examinados
y α, β y γ son, respectivamente, los parámetros que controlan las contribuciones relativas de la consulta original, los documentos relevantes, y los documentos no relevantes.

A estos parámetros debemos añadir un nuevo parámetro t para el caso de esta implementación de SMART, puesto que de los términos de los documentos considerados, sólo los t mejores términos son empleados en la expansión. En nuestro caso concreto, al tratarse de una expansión ciega sin interacción con el usuario, se tomarán como relevantes los n_1 primeros documentos devueltos por la consulta inicial. Fijaremos también γ a 0, para así eliminar la realimentación negativa y considerar únicamente la información proveniente de los documentos tomados como sí relevantes. Por otra parte, dado que los valores concretos que tomen α y β no son en sí importantes, sino su ratio α/β, no es necesario estimar ambos parámetros separadamente, sino que basta con fijar el valor de uno y variar el otro. En nuestro caso se fijó β a 0.10, con lo que únicamente restarán por estimar los restantes tres parámetros —α, n_1, y t— correspondientes a la expresión:

$$Q_1 = \alpha Q_0 + 0{,}10 \sum_{k=1}^{n_1} \frac{R_k}{n_1}$$

Estos parámetros fueron estimados mediante un proceso de entrenamiento llevado a cabo con el corpus destinado a tal efecto, el CLEF 2001-02·A. En procesos similares, como los llevados a cabo en [211, 212], se aceptaba como buena aquélla combinación de parámetros que ofrecía mejores resultados, sin embargo esto puede llevarnos a sobreentrenar el sistema con respecto al corpus de entrenamiento y, de este modo, perder generalidad. Por esta razón nuestro proceso de entrenamiento no se basa en la búsqueda del mejor rendimiento absoluto, sino en el del mayor incremento del mismo.

El criterio de búsqueda es el de la media de las precisiones no interpoladas para la combinación de parámetros considerada, y donde los rangos de parámetros considerados son:

$$\alpha \in \{0.10,\ 0.20,\ 0.40,\ 0.80,\ 1.20,\ 1.80,\ 2.40,\ 3.20\}^{14}$$
$$n_1 \in \{5{,}10{,}15\}$$
$$t \in \{5{,}10{,}15{,}20{,}30{,}40{,}50\}$$

[14] Con ratios α/β 1, 0.5, 0.25, 0.125, 0.0833, 0.0555, 0.0416 y 0.0312, respectivamente.

De este modo, a la hora de determinar el α a utilizar en los experimentos se tomará como criterio la media de las precisiones obtenidas para todas las combinaciones posibles de los valores de n_1 y t con un α dado. Una vez fijado el α calcularemos el número de documentos a emplear (n_1) tomando como criterio la media de las precisiones obtenidas para todos los t con un n_1 dado y el valor de α ya estimado. Finalmente, fijados ya α y n_1, tomaremos como criterio la precisión no interpolada obtenidos para esos valores de α y n_1 y cada t posible. En cada uno de estos pasos el algoritmo de búsqueda es el mismo.

Algoritmo 5.2 Algoritmo de búsqueda para la optimización de los parámetros de realimentación:

1. Se toma como valor por defecto del parámetro a optimizar el primer valor para el cual la media de sus precisiones obtiene una mejora respecto a la precisión obtenida sin realimentación.

2. Partiendo del valor por defecto, se exploran los valores siguientes mientras exista un incremento en la media de las precisiones respecto a la media anterior, tomando como nuevo valor por defecto aquél para el cual dicho incremento sea máximo y de al menos un porcentaje δ respecto a la media anterior (en nuestro caso se ha tomado δ=0.90 %).

3. Partiendo del valor por defecto actual, se exploran los siguientes valores para los cuales sigue habiendo un incremento en la media de precisiones respecto a la media anterior, deteniéndonos en el último valor antes de que dicho incremento caiga por debajo de un porcentaje δ respecto a la media de precisiones anterior.

□

A partir de las precisiones no interpoladas obtenidas para todas las combinaciones posibles de los valores de parámetros considerados —recogidas en el apéndice C—, este proceso se aplicó separadamente para consultas cortas y largas, tal y como se indica en la tabla 5.5, en la cual se muestra para cada paso del proceso: el parámetro a optimizar en esa fase —α, n_1 y t—, la media de las precisiones no interpoladas para las combinaciones de los restantes parámetros por estimar —$\mu_{Pr.}$—, y el porcentaje de incremento respecto a la media anterior —%Δ—. Los valores obtenidos mediante este proceso, y que serán utilizados de aquí en adelante, fueron los siguientes:

Consultas cortas: α=0.80, n_1=5, t=10.

Consultas largas: α=1.20, n_1=5, t=10.

5.6.3. Resultados Aplicando Realimentación

La tabla 5.6 compara los resultados obtenidos mediante lematización (*lem*) y *stemming* (*stm*) cuando la consulta es expandida automáticamente mediante la técnica de Rocchio[15], empleando para ello los parámetros establecidos previamente durante el proceso de entrenamiento:

Consultas cortas: α=0.80, β=0.10, γ=0, n_1=5, t=10.

Consultas largas: α=1.20, β=0.10, γ=0, n_1=5, t=10.

[15]En este caso se toma como línea base el *stemming* con realimentación.

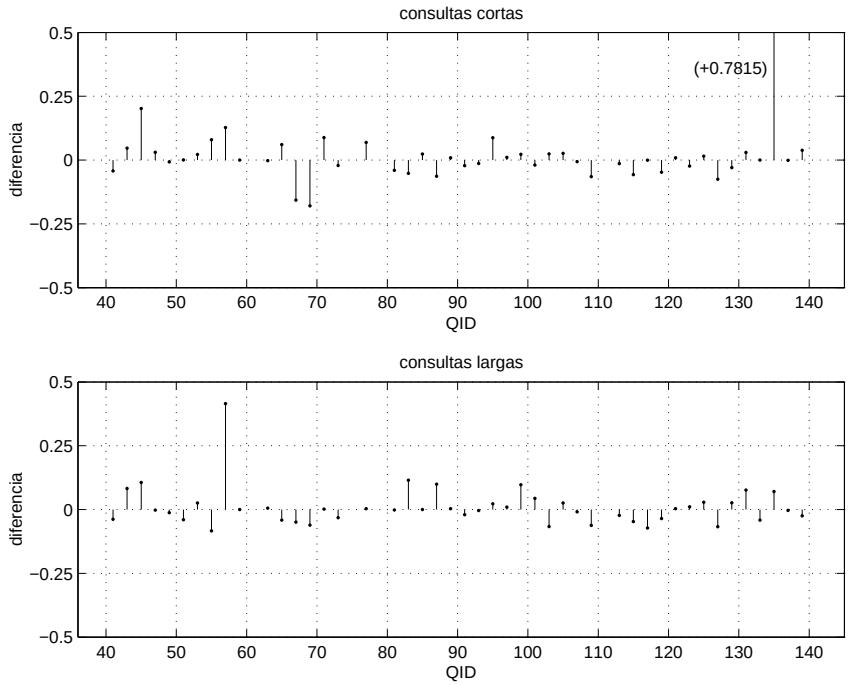


Figura 5.6: Diferencias en las precisiones no interpoladas aplicando realimentación: *stemming* vs. lematización. Corpus CLEF 2001-02·A

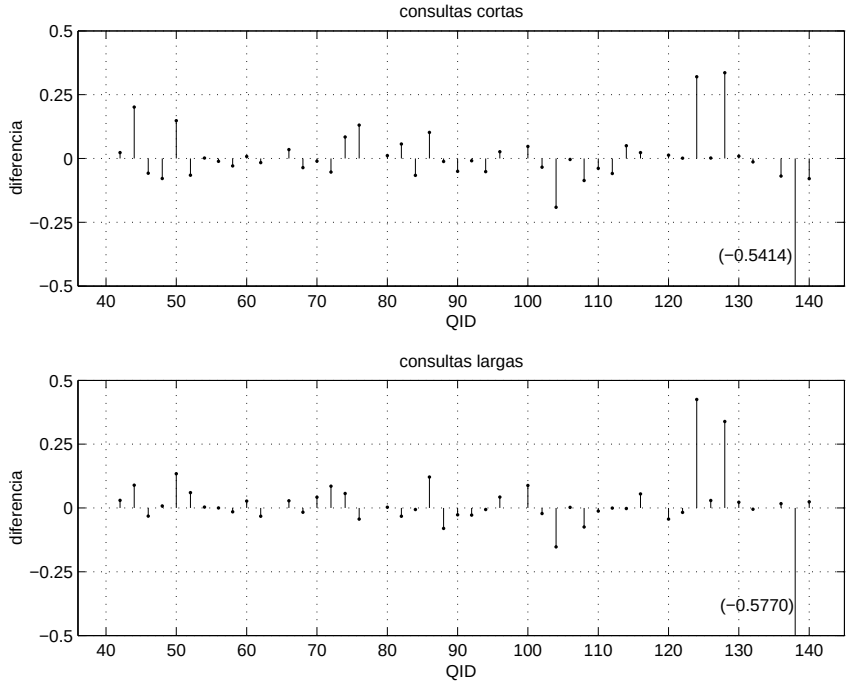


Figura 5.7: Diferencias en las precisiones no interpoladas aplicando realimentación: *stemming* vs. lematización. Corpus CLEF 2001-02·B

Consultas cortas (Pr. no int. sin realimentación: .4829)								
α	$\mu_{Pr.}$	%Δ	n_1	$\mu_{Pr.}$	%Δ	t	$\mu_{Pr.}$	%Δ
.10	.5039	4.35	***5***	***.5222***	***8.14***	5	.5150	6.65
.20	.5087	0.95	10	.5265	.82	***10***	***.5220***	***1.36***
.40	.5158	1.40	15	.5229	-.68	15	.5242	.42
.80	***.5239***	***1.57***				20	.5216	-.50
1.20	.5277	0.74				30	.5253	.71
1.80	.5300	0.42				40	.5239	-.27
2.40	.5301	0.03				50	.5235	-.08
3.20	.5291	-0.20						

Consultas largas (Pr. no int. sin realimentación: .5239)								
α	$\mu_{Pr.}$	%Δ	n_1	$\mu_{Pr.}$	%Δ	t	$\mu_{Pr.}$	%Δ
.10	.5223	-.30	***5***	***.5562***	***6.16***	5	.5554	6.01
.20	.5283	1.15	10	.5550	-.21	***10***	***.5604***	***.90***
.40	.5378	1.80	15	.5561	.19	15	.5624	.36
.80	.5494	2.14				20	.5576	-.85
1.20	***.5558***	***1.17***				30	.5532	-.79
1.80	.5599	.75				40	.5510	-.40
2.40	.5607	.14				50	.5533	.42
3.20	.5602	-.10						

Tabla 5.5: Estimación de parámetros de realimentación

La introducción de la realimentación produce, como era de esperar, un aumento del rendimiento, pudiendo a la vez comprobar de nuevo que el empleo de la lematización como técnica de normalización de términos simples conlleva una mejora de los resultados con respecto al clásico *stemming*, debiendo señalar como única excepción, casos en los que se produce una ligera disminución de la precisión en los niveles más bajos de cobertura y en los primeros documentos devueltos. Sin embargo, esta disminución es menos frecuente y menos extendida que en el caso de no emplear realimentación.

Para el corpus CLEF 2001-02·A, el análisis de los resultados recogidos en la tabla 5.6 es similar al dado para el caso sin realimentación, al producirse una mejora apreciable y general de los resultados cuando se aplica la lematización, a excepción del nivel mínimo de cobertura (0 %), en el que se produce una ligera disminución de la precisión. La mejora obtenida es, de nuevo, mucho más patente en el caso de las consultas cortas, ampliándose incluso, destacando una importante caída en el número de documentos relevantes devueltos —161 documentos— mediante *stemming* frente a la lematización, si bien ésta se centra mayormente en consultas puntuales. Los resultados para las consultas largas muestran también una mejora general respecto al *stemming*, salvo para el nivel mínimo de cobertura y una disminución marginal de la precisión a partir de los 100 documentos, si bien éstos son de menor interés para nosotros. En cuanto a las diferencias de precisión a nivel de consulta, éstas se recogen en la figura 5.6.

En lo que respecta al corpus CLEF 2001-02·B, la mejora obtenida es nuevamente menos apreciable en el caso de las consultas cortas, produciéndose incluso una leve disminución de la precisión no interpolada, al ser algo menor para la lematización, si bien muy similar a la del *stemming*. Debemos señalar, por otra parte, que ya antes de la expansión estas precisiones eran del todo similares, si bien de signo contrario. Sigue también produciéndose para ambos tipos de consultas, y al igual que antes, una disminución de la precisión en los niveles más bajos de

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*stm*	*lem*	*%Δ*	*stm*	*lem*	*%Δ*	*stm*	*lem*	*%Δ*	*stm*	*lem*	*%Δ*	*stm*	*lem*	*%Δ*	*stm*	*lem*	*%Δ*
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2606	**2767**	**6.18**	2760	**2779**	**0.69**	2362	**2376**	**0.59**	2395	**2406**	**0.46**	2210	**2240**	**1.36**	2227	2223	-0.18
Pr. no int.	.5032	**.5220**	**3.74**	.5510	**.5604**	**1.71**	.4780	.4773	-0.15	.5281	**.5392**	**2.10**	.4756	**.5024**	**5.63**	.5135	**.5207**	**1.40**
Pr. doc.	.5213	**.5784**	**10.95**	.5855	**.5912**	**0.97**	.5433	**.5681**	**4.56**	.5754	**.6145**	**6.80**	.5380	**.5530**	**2.79**	.5917	.5802	-1.94
R-pr.	.4796	**.4990**	**4.05**	.5294	**.5366**	**1.36**	.4547	**.4599**	**1.14**	.5002	**.5104**	**2.04**	.4583	**.4912**	**7.18**	.4926	.4871	-1.12
Pr. a 0 %	.8328	.8221	-1.28	.9272	.8895	-4.07	.8538	.8210	-3.84	.8764	.8710	-0.62	.7838	**.8145**	**3.92**	.8072	**.8301**	**2.84**
Pr. a 10 %	.7287	**.7490**	**2.79**	.7996	**.8028**	**0.40**	.6755	**.6861**	**1.57**	.7469	**.7619**	**2.01**	.7145	**.7369**	**3.14**	.7279	**.7421**	**1.95**
Pr. a 20 %	.6518	**.6866**	**5.34**	.7265	**.7352**	**1.20**	.6287	**.6319**	**0.51**	.6903	**.6929**	**0.38**	.6352	**.6632**	**4.41**	.6731	**.6758**	**0.40**
Pr. a 30 %	.6147	**.6573**	**6.93**	.6886	**.6996**	**1.60**	.5567	**.5688**	**2.17**	.6256	**.6497**	**3.85**	.5851	**.6019**	**2.87**	.6242	**.6304**	**0.99**
Pr. a 40 %	.5664	**.5997**	**5.88**	.6394	**.6541**	**2.30**	.5309	.5289	-0.38	.5942	**.6202**	**4.38**	.5229	**.5638**	**7.82**	.5772	**.5975**	**3.52**
Pr. a 50 %	.5177	**.5456**	**5.39**	.5822	**.6005**	**3.14**	.4952	**.5017**	**1.31**	.5633	**.5733**	**1.78**	.4884	**.5410**	**10.77**	.5282	**.5479**	**3.73**
Pr. a 60 %	.4735	**.4994**	**5.47**	.5205	**.5386**	**3.48**	.4620	**.4623**	**0.06**	.5023	**.5194**	**3.40**	.4552	**.4783**	**5.07**	.4945	.4938	-0.14
Pr. a 70 %	.4170	**.4375**	**4.92**	.4594	**.4621**	**0.59**	.4298	.4255	-1.00	.4638	**.4862**	**4.83**	.4081	**.4366**	**6.98**	.4528	**.4587**	**1.30**
Pr. a 80 %	.3661	.3661	0.00	.3839	**.3910**	**1.85**	.3399	.3384	-0.44	.3760	.3753	-0.19	.3452	**.3788**	**9.73**	.3924	.3891	-0.84
Pr. a 90 %	.2854	**.2939**	**2.98**	.3024	**.3130**	**3.51**	.2700	.2651	-1.81	.3207	.3059	-4.61	.2860	**.3102**	**8.46**	.3195	**.3230**	**1.10**
Pr. a 100 %	.1509	**.1547**	**2.52**	.1422	**.1624**	**14.21**	.1612	.1395	-13.46	.1972	.1704	-13.59	.1531	**.1871**	**22.21**	.1804	**.1928**	**6.87**
Pr. a 5 docs.	.6391	**.6609**	**3.41**	.6913	**.6957**	**0.64**	.6089	.5956	-2.18	.6978	.6844	-1.92	.5787	**.5872**	**1.47**	.6340	.6213	-2.00
Pr. a 10 docs.	.6087	**.6457**	**6.08**	.6630	**.6848**	**3.29**	.5578	**.5600**	**0.39**	.6022	**.6178**	**2.59**	.5404	**.5596**	**3.55**	.5936	.5872	-1.08
Pr. a 15 docs.	.5609	**.5884**	**4.90**	.6174	**.6435**	**4.23**	.5274	.5274	0.00	.5600	**.5822**	**3.96**	.5021	**.5305**	**5.66**	.5461	**.5504**	**0.79**
Pr. a 20 docs.	.5478	**.5630**	**2.77**	.5946	**.6043**	**1.63**	.4911	**.5011**	**2.04**	.5256	**.5533**	**5.27**	.4755	**.4883**	**2.69**	.5138	**.5266**	**2.49**
Pr. a 30 docs.	.5000	**.5225**	**4.50**	.5449	**.5580**	**2.40**	.4304	**.4444**	**3.25**	.4793	**.5081**	**6.01**	.4262	**.4433**	**4.01**	.4660	**.4667**	**0.15**
Pr. a 100 docs.	.3274	**.3507**	**7.12**	.3615	.3598	-0.47	.2860	**.2940**	**2.80**	.3027	**.3191**	**5.42**	.2734	**.2770**	**1.32**	.2885	.2853	-1.11
Pr. a 200 docs.	.2157	**.2348**	**8.85**	.2373	.2361	-0.51	.1942	**.1979**	**1.91**	.2014	**.2067**	**2.63**	.1718	**.1753**	**2.04**	.1791	.1791	0.00
Pr. a 500 docs.	.1045	**.1122**	**7.37**	.1126	.1121	-0.44	.0973	**.0980**	**0.72**	.0987	**.1008**	**2.13**	.0855	**.0869**	**1.64**	.0874	.0871	-0.34
Pr. a 1000 docs.	.0567	**.0602**	**6.17**	.0600	**.0604**	**0.67**	.0525	**.0528**	**0.57**	.0532	**.0535**	**0.56**	.0470	**.0477**	**1.49**	.0474	.0473	-0.21

Tabla 5.6: Resultados obtenidos mediante *stemming* (*stm*), caso base, y lematización (*lem*) aplicando en ambos casos realimentación

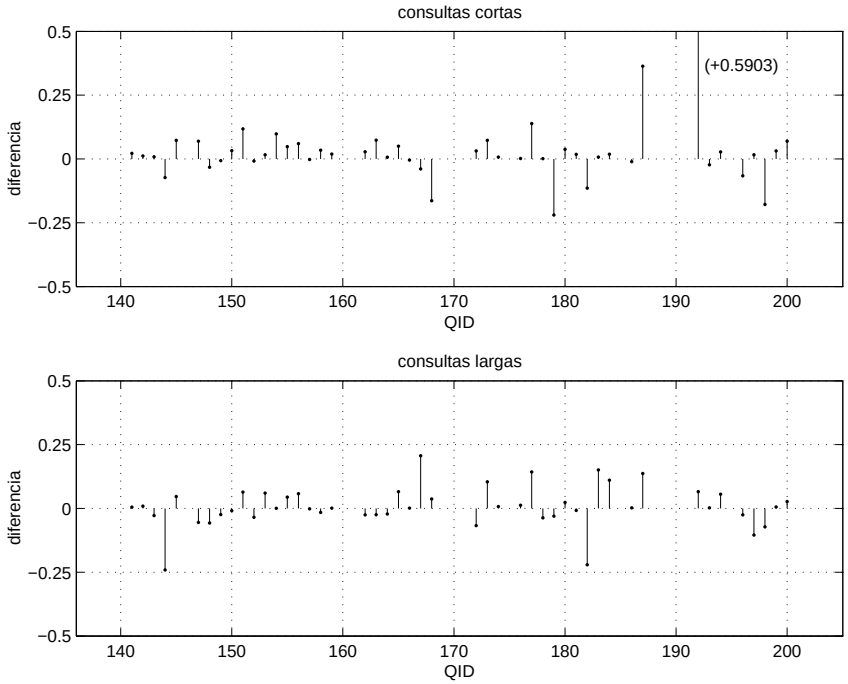


Figura 5.8: Diferencias en las precisiones no interpoladas aplicando realimentación: *stemming* vs. lematización. Corpus CLEF 2003

cobertura —aunque ya sólo en el 0 %— y de número de documentos devueltos. La gráfica 5.7 recoge las diferencias de precisión obtenidas para cada consulta, repitiendo de nuevo algunas consultas picos máximos y mínimos.

Finalmente, para el caso del corpus CLEF 2003, de comportamiento algo más errático en los experimentos iniciales, se produce una mejora clara y general en el rendimiento del sistema cuando se emplean consultas cortas, si bien en el caso de las consultas largas su comportamiento sigue resultando algo errático, especialmente en el caso de la precisión a los N documentos devueltos. Su comportamiento a nivel de cada consulta particular se recoge en la figura 5.8.

5.7. Discusión

Recogemos en este capítulo la primera de nuestras propuestas para la normalización de términos simples en español mediante técnicas de Procesamiento del Lenguaje Natural. En este caso se trata de la lematización del texto y la utilización, como términos de indexación, de los lemas de las palabras con contenido que éste contiene. Los buenos resultados obtenidos señalan a la lematización como una alternativa viable a las técnicas clásicas basadas en *stemming*. Nuestros resultados positivos contrastan con aquéllos desfavorables obtenidos en el *ad hoc Spanish retrieval task* del TREC por [221], con una aproximación similar a la nuestra en la que, tras etiquetarse el texto, se indexaban conjuntamente la forma canónica obtenida junto con su categoría gramatical.

Por otra parte, la etiquetación y lematización del texto a procesar constituyen la base para el desarrollo de métodos de normalización más elaborados que hagan frente a los fenómenos de variación lingüística más complejos, ya que para ello es necesario, en primer lugar, eliminar la ambigüedad léxica del texto a procesar [230, 121, 113].

Asimismo, se ha hecho ya patente que uno de los principales problemas a los que deben hacer frente las técnicas de procesamiento lingüístico en la Recuperación de Información es la

presencia de errores ortográficos y gramaticales, por lo que debe incidirse en la medida de lo posible en la robustez del sistema desarrollado.

Por otra parte, debemos llamar la atención sobre el hecho de que la normalización mediante lematización ha sido también probada con éxito en el caso del gallego [31, 30].

Capítulo 6

Tratamiento de la Variación Morfológica Derivativa: Familias Morfológicas

6.1. Introducción

En el capítulo anterior se planteó el empleo de la lematización para la eliminación de la variación flexiva del español como alternativa a las técnicas clásicas basadas en *stemming*. Sin embargo, restan todavía por tratar los fenómenos derivativos del lenguaje, para así dar cobertura completa a la variación morfológica.

En el caso del inglés, este proceso generalmente también se lleva a cabo mediante la aplicación de técnicas de *stemming*. Sin embargo, tal y como se mostró en el capítulo anterior, su empleo en el caso del español presenta inconvenientes debido a los problemas derivados de la compleja morfología de esta lengua, y debido también a la pérdida de información que conlleva el empleo del *stemming* de cara a fases posteriores. Ello nos ha llevado a desarrollar una herramienta para el procesamiento de los fenómenos morfológico-derivativos del español para su aplicación en el procesamiento automático del lenguaje, dadas sus potenciales aplicaciones [209, 233, 114].

En este capítulo describiremos los fundamentos lingüísticos y las técnicas de implementación de esta herramienta, así como los experimentos realizados en el marco de la normalización de términos simples para comprobar la viabilidad de su aplicación en este marco.

6.2. Aspectos Lingüísticos de la Formación de Palabras en Español

Antes de entrar en detalles de diseño e implementación es preciso describir, desde un punto de vista lingüístico, los mecanismos de formación del léxico en español, para de este modo llegar a una mayor comprensión de la problemática asociada a su procesamiento. A la hora de obtener unos patrones regulares de formación de palabras nos valdremos de las llamadas *reglas de formación*, basadas en teorías tales como la Gramática Generativa Transformacional y el desarrollo de la denominada Fonología Derivativa [135]. Aunque dicho paradigma no es completo, supone un avance considerable puesto que nos permite diseñar un sistema de generación automática de derivados con un grado aceptable de corrección y completud.

6.2.1. La Estructura de la Palabra

Consideraremos que una palabra está formada por morfemas léxicos y gramaticales, siendo el *morfema* la unidad gramatical mínima distintiva que ya no puede ser significativamente subdividida en términos gramaticales [135]. El *morfema léxico*, también denominado *lexema* o *raíz*, es aquel morfema con significado léxico pleno y que contiene, por tanto, la significación de la palabra (p.ej., *sol*, *cantar*). Por el contrario, el morfema gramatical, también denominado *afijo* o, por extensión, simplemente *morfema*, posee únicamente significado gramatical, y sirve para relacionar a los lexemas en la oración (p.ej., *de*, *no*, *el* libro), o para modificar la significación de un lexema (p.ej., casa*s*, cas*ero*). Si una palabra posee un lexema, su significado es léxico; tal es el caso de los nombres, los adjetivos, los verbos, las interjecciones y los adverbios. Sin embargo, si está constituida únicamente por morfemas gramaticales, su significado es gramatical (o relacional), como ocurre en el caso de los artículos, las preposiciones y las conjunciones. Los pronombres tienen un significado léxico accidental, el del nombre al cual sustituyen.

El caso que nos ocupa en este trabajo, y al que nos circunscribiremos de aquí en adelante, es el de las palabras con significado léxico, aquéllas formadas por un lexema y, en la mayoría de los casos, por uno o más afijos que lo modifican.

Conforme a su posición, los afijos se clasifican en *prefijos*, antepuestos al lexema (p.ej., *in*necesario), *sufijos*, pospuestos al lexema (p.ej., *habl*ador), e *infijos*, elementos que aparecen intercalados en el interior de la estructura de una palabra (p.ej., *huma*reda).

De acuerdo con la forma en la que éstos alteran el significado del lexema, los afijos se clasifican en flexivos y derivativos. Los *afijos flexivos* representan conceptos gramaticales tales como género y número (p.ej., *hablador* vs. *habladoras*), persona, modo, tiempo y aspecto (p.ej., *habl*o vs. *habl*ases). Los *afijos derivativos* producen un cambio semántico respecto al lexema base, y frecuentemente también un cambio de categoría sintáctica (p.ej., *habl*ar vs. *habl*ador).

6.2.2. Mecanismos de Formación de Palabras

A la hora de estudiar los mecanismos de generación léxica, dicho estudio puede realizarse bien desde un punto de vista diacrónico, remitiéndose a los orígenes etimológicos de los patrones morfológicos de formación de las palabras y examinando los patrones morfológicos que regían en el pasado, o bien desde un punto de vista sincrónico, analizando los sistemas y reglas de los mecanismos internos del lenguaje independientemente del desarrollo histórico y etnocultural. Puesto que nuestro objetivo es la automatización e implementación de dichos mecanismos, nuestra perspectiva debe ser, por tanto, sincrónica.

Tradicionalmente los mecanismos responsables de la creación y renovación del léxico han sido divididos en composición y derivación, mecanismos morfológicos referidos a la combinación de las palabras y sus subunidades, si bien existen otros procedimientos tales como préstamos de otras lenguas, acrónimos, apócopes, neologismos, etc. Hablamos de *composición* ante la combinación de lexemas independientes, y de *derivación* ante la aplicación de morfemas derivativos. Se trata en ambos casos de procedimientos morfológicos, unión de morfemas individuales o grupos de morfemas en unidades superiores para formar lexemas complejos. Los nuevos términos obtenidos pueden actuar, a su vez, como términos *base* o *primitivos* para la formación de nuevas palabras. La estructura morfémica es, por tanto, fundamental para el análisis de los procesos de formación de palabras. Debe tenerse en cuenta, sin embargo, que en las lenguas romances, y especialmente en el español, la propia palabra antes que cualquiera de sus componentes morfémicos constituye la base de la derivación.

Llegados a este punto debemos recordar que nuestro objetivo es el tratamiento de la variación lingüística de carácter morfológico derivativo, por lo que nos limitaremos únicamente al estudio y tratamiento de los fenómenos derivativos.

6.2.3. Mecanismos de Derivación

Los mecanismos básicos de derivación en español son la prefijación, la sufijación, la parasíntesis y la derivación regresiva.

Prefijación

La *prefijación* consiste en la anteposición de un prefijo a una forma base (p.ej., *necesario* $\xrightarrow{in-}$ *innecesario*[1]). El tratamiento automático de los prefijos es menos complejo que el de los sufijos, ya que su tendencia a la polisemia es mucho menor que la de éstos, su adjunción no implica alteraciones en la acentuación del lexema, y tampoco suelen conllevar cambios en la categoría gramatical de la base. Sin embargo, la profundidad de los cambios semánticos que introducen desaconsejan su procesamiento automático en tareas de Recuperación de Información [230, 105], por lo que la prefijación no es abordada actualmente por nuestra herramienta.

Sufijación

La *sufijación* consiste en la adición de un sufijo a la forma base (p.ej., *hablar* $\xrightarrow{-ador}$ *hablador*), debiendo distinguir entre sufijación apreciativa y no apreciativa.

Entendemos por *sufijación apreciativa* aquélla que altera semánticamente el lexema base de un modo subjetivo emocional pero sin cambiar su categoría gramatical. A su vez, los sufijos apreciativos pueden subdividirse en: diminutivos, que transmiten una idea de pequeñez o afectividad (p.ej., *beso* $\xrightarrow{-ito}$ *besito*); aumentativos, que implican amplia dimensión, fealdad o grandiosidad (p.ej., *coche* $\xrightarrow{-azo}$ *cochazo*); y peyorativos, que implican desagrado o ridiculez (p.ej., *pueblo* $\xrightarrow{-ucho}$ *pueblucho*). El cambio semántico introducido por la sufijación apreciativa es, por tanto, marginal.

Por contra, la *sufijación no apreciativa* conlleva un cambio fundamental más que marginal en el significado del lexema base, y frecuentemente implica además un cambio de categoría sintáctica (p.ej., *hablar* $\xrightarrow{-ador}$ *hablador*).

El repertorio de sufijos no apreciativos del español cuenta con cientos de morfemas derivativos, cuyo inventario no está todavía fijado por completo. Este elevado número de sufijos existentes plantea un problema a la hora de su clasificación. En nuestro caso, el criterio que hemos seguido ha sido doble. En primer lugar, en función de la categoría gramatical de la base, podemos distinguir entre sufijos:

- *Denominales*, que toman como base un sustantivo.
- *Deadjetivales*, que toman como base un adjetivo.
- *Deverbales*, que toman como base un verbo (los más frecuentes).

En segundo lugar, y de acuerdo a la categoría gramatical del derivado resultante, se puede hablar de sufijos

- *Nominalizadores*, que derivan sustantivos (los más comunes).
- *Adjetivizadores*, que derivan adjetivos.
- *Verbalizadores*, que derivan verbos.

[1] Mediante esta notación representamos el hecho de que *innecesario* se deriva de *necesario* mediante el prefijo *in-*.

Desde el punto de vista semántico todos los sufijos no apreciativos son significativos dado que el significado del derivado es siempre diferente del que poseía la base. Sin embargo, debe tenerse en cuenta el hecho de que la mayoría de los sufijos son polisémicos; por ejemplo:

$$muchach\underline{ada}\begin{cases}\text{nombre colectivo } (grupo\ de\ muchachos)\\ \text{acción propia de la base } (chiquillada)\end{cases}$$

El diseño de nuestra herramienta aborda sólamente la sufijación no apreciativa, por ser ésta la imperante en los textos y por su mayor interés de cara a la Recuperación de Información.

Parasíntesis

La *parasíntesis* es un mecanismo derivativo de relativa frecuencia en el español y que consiste en la adición simultánea de un prefijo y un sufijo (p.ej., *rojo* $\overset{en-\ -ecer}{\longrightarrow}$ *enrojecer*). Es necesario precisar que en el caso de la parasíntesis no cabe un análisis diacrónico de dichas formaciones léxicas, dado que supondría la sufijación previa y posterior prefijación (o viceversa) sobre formas intermedias inexistentes como **enrojo* o **rojecer*.

Derivación Regresiva

La *derivación regresiva* es un mecanismo derivativo para la formación de sustantivos a partir de verbos —de gran importancia en el español contemporáneo— en el que, en lugar de incrementar el tamaño del lexema base, se produce una reducción del mismo, al añadir tan solo una vocal fuerte[2] a la raíz verbal. Por ejemplo:

$$\begin{array}{ll} tom\underline{ar} & \rightarrow tom\underline{a}\\ altern\underline{ar} & \rightarrow altern\underline{e}\\ dañ\underline{ar} & \rightarrow dañ\underline{o}\end{array}$$

Alomorfia

Otro fenómeno de notable importancia a tener en cuenta en el proceso de formación del léxico es la *alomorfia*, la posible existencia de diferentes variantes —*alomorfos*— para un mismo morfema derivativo. El alomorfo a utilizar en cada caso puede estar determinado por la fonología o venir impuesto por convención o por la etimología, como podemos observar en los siguientes ejemplos:

des-/dis-	*des*cosido	*dis*culpar
-able/-ible	alcanz*able*	defend*ible*

6.2.4. Condiciones Fonológicas

Dado que toda operación morfológica implica a su vez una alteración fonológica de la base, no debe descuidarse el análisis de las condiciones fonológicas que presiden el proceso de formación de palabras. Dichas alteraciones pueden ser regulares o aparentemente irregulares. Consideremos los siguientes ejemplos:

(a)
$$\begin{array}{ll} responder & \rightarrow respondón\\ vencer & \rightarrow invencible\\ grosero & \rightarrow grosería\end{array}$$

(b)
$$\begin{array}{ll} pan & \rightarrow panadero\\ agua & \rightarrow acuatizar\\ carne & \rightarrow carnicero\end{array}$$

[2]Vocales 'a','e', y 'o'

En (a) los sufijos se adjuntan a sus bases o raíces de manera morfológicamente regular, con resultados previsibles tanto morfológica como fonológicamente, mientras que en (b) este proceso semeja irregular, puesto que el resultado no es el esperado (**pandero*, **agüizar*, **carnero*). Sin embargo, algunos de esos fenómenos son lo suficientemente frecuentes como para que deban ser forzosamente incluidos al estudiar la morfología léxica del español.

El esfuerzo por intentar explicar situaciones como las indicadas condujo a la creación de las llamadas *reglas de reajuste*. Aronoff [135] clasifica dichas reglas en dos grupos: reglas alomórficas, relativas a aspectos de alomorfia, y reglas de truncamiento, aquéllas que suprimen un morfema terminal ante la adjunción de un nuevo sufijo:[3]

pan	→ *panadero*	inserción de /aδ/ entre la raíz y el sufijo
/pan/	→ /panaδero/	
agua	→ *acuatizar*	conversión de la sonora /γ/ en sorda /k/ e inserción
/aγua/	→ /akwatiθar/	de /t/ antes del morfema de infinitivo
carne	→ *carnicero*	inserción de /iθ/ entre la raíz y el sufijo
/karne/	→ /karniθero/	

Las condiciones fonológicas concretas que se han considerado en este trabajo se detallan en el apartado 6.3.4.

6.2.5. Reglas y Restricciones

Incluso los patrones más productivos de formación de palabras en español están sujetos a restricciones, de tal forma que una regla de adición morfológica de un afijo a una base puede ser restringida con el paso del tiempo. En consecuencia, el grado de aceptación de una derivación bien formada desde el punto de vista morfológico resulta imposible de predecir únicamente a partir de la forma.

Sin embargo, al comparar las restricciones aplicables al español y las de otras lenguas, se puede comprobar que el español goza de una gran flexibilidad en su morfología derivativa. Si lo comparásemos directamente, por ejemplo, con el caso del inglés o el francés (lengua romance también), podríamos ver que sus mecanismos de formación de palabras no son tan productivos como los del español, particularmente en lo que atañe a la sufijación [135], como se muestra en las tablas 6.1 y 6.2. Por medio de los ejemplos que ambas contienen podemos percatarnos de cómo el español prefiere la derivación, particularmente mediante sufijación [209], a otros métodos de generación de palabras.

6.3. Implementación

Tomando como base los fundamentos teóricos expuestos en los apartados anteriores, se ha desarrollado una herramienta para la generación automática de familias morfológicas.

6.3.1. Planteamientos Previos

La implementación propuesta en este trabajo se apoya sobre el concepto de *familia morfológica*, que [111] define formalmente de la siguiente forma: sea $\mathfrak{L}$ el conjunto de lemas, y $M \in \mathfrak{L} \leftrightarrow \mathfrak{L}$ una relación binaria de $\mathfrak{L}$ en $\mathfrak{L}$ definida por los lexemas que asocia cada lema con su(s) lexema(s)[4], definimos *familia morfológica* $F_M(l)$ de un lema l como el conjunto de lemas

[3] Para cada palabra mostramos, entre barras, su transcripción fonética.

[4] M no es una función pues los lemas compuestos tienen más de un lexema.

inglés	***español***	***inglés***	***español***
fist	*puño*	orange	*naranja*
dagger	*puñal*	orange tree	*naranjo*
stab	*puñalada*	orange grove	*naranjal*
punch	*puñetazo*	orangeade	*naranjada*
fistful	*puñado*	orange coloured	*anaranjado*
to grasp	*empuñar*	orange seller	*naranjero*
to stab	*apuñalar*	blow with an orange	*naranjazo*
sword hilt	*empuñadura*	small orange	*naranjita*

Tabla 6.1: Formación de palabras en inglés y en español

francés	***español***	***francés***	***español***
coup de pied	*patada*	*tête*	*cabeza*
petit ami	*amiguito*	*coup de tête*	*cabezazo*
mal de mer	*mareo*	*incliner la tête*	*cabecear*
salle à manger	*comedor*	*chef d'èmeute*	*cabecilla*
belle-mère	*suegra*	*tète de lit*	*cabecera*
coup à la porte	*aldabonazo*	*de tète grosse*	*cabezudo*

Tabla 6.2: Formación de palabras en francés y en español

(l incluido) que comparten con l el mismo lexema común:

$$\begin{cases} F_M \in \mathbb{P}(\mathfrak{L}) \\ \forall l \in \mathfrak{L}, F_M(l) = \{l' \in \mathfrak{L} \ / \ \exists r \in \mathfrak{L}, (l,r) \in M \wedge (l',r) \in M\} = M^{-1}(M(l)) \end{cases}$$

siendo $\mathbb{P}(\mathfrak{L})$ el conjunto potencia de $\mathfrak{L}$.

En consecuencia, y restringiéndonos a las relaciones derivativas objeto de nuestro estudio, definimos de un modo informal *familia morfológica* como el conjunto de palabras obtenibles a partir de un mismo lexema mediante procesos derivativos [242, 243].

Además de la restricción a los fenómenos derivativos, deben tenerse en cuenta, a la hora de la implementación de la herramienta, otros aspectos. Por una parte, y con vistas a una automatización plena del proceso, nuestro acercamiento debe ser sincrónico, es decir, independiente del desarrollo histórico y etnocultural del lenguaje, centrándonos únicamente en los sistemas, reglas y mecanismos internos del mismo. Por otra parte, debemos extender el concepto extrictamente lingüístico y etimológico de derivación al concepto más relajado de *relación morfoléxica* [209], para de este modo incluir el emparentamiento de vocablos a través de una terminación por coincidir gráficamente con un sufijo y aportar además su misma semántica y funcionalidad. De esta forma podemos establecer relaciones que, si bien son ajenas al concepto estrictamente lingüístico de derivación, consiguen igualmente el objetivo perseguido, establecer relaciones semánticamente válidas entre palabras relacionadas derivativamente en última instancia. De este modo, podemos relacionar *psiquiatra* con *psiquiatría* simplemente mediante el sufijo *-ía*, sin tener que recurrir a establecer una relación previa en base al lexema *psiqui-*. Del mismo modo, podemos relacionar *multiplicar* y *multipicable* mediante el sufijo *-able*, cuando etimológicamente sería incorrecto, pues *multiplicable* procede del latín *multiplicabilis*, derivado —dentro del latín— de *multiplicare*, palabra latina de la cual procede la española *multiplicar*.

A partir de ahora, si bien seguiremos hablando de *derivación*, nos estaremos refiriendo —salvo que se indique lo contrario— al concepto ampliado de *relación morfoléxica*.

6.3.2. Algoritmo de Generación de Familias Morfológicas

Llegados a este punto, debemos llamar nuevamente la atención sobre la escasa disponibilidad de grandes recursos lingüísticos libremente accesibles en el caso del español (corpora etiquetados, bancos de árboles, diccionarios avanzados, etc.). Ante esta situación se ha optado por diseñar una herramienta que afronte la tarea desde un nivel léxico, restringiéndose a unos recursos mínimos:

1. Un lexicón, el mismo empleado por el etiquetador, donde para cada entrada se incluye su forma junto con la etiqueta y lema correspondientes.
2. Un modelo de ordenación para la aplicación de los morfemas, en este caso implícito en el algoritmo.
3. Una serie de reglas que modelan las alteraciones de la palabra durante la adjunción de los morfemas. Estas reglas se describen en el apartado 6.3.4.

Estos recursos coinciden con la información mínima necesaria para la implementación de un analizador morfológico [121], lo cual no es de extrañar, pues los principios de actuación de ambas herramientas son similares, si bien el analizador *analiza*, y nuestra herramienta *genera*. Por otra parte, los analizadores morfológicos son frecuentamente implementados mediante *traductores finitos* que pueden funcionar tanto a modo de analizadores como de generadores. Nuestra herramienta se limitaría a esta segunda función.

Nuestra aproximación opta por un algoritmo sobregenerativo [233], es decir, que crea todas las derivaciones válidas posibles, pero no necesariamente existentes, de una palabra. Este tipo de solución aporta como ventaja una mayor cobertura, al tener en consideración todos los sufijos posibles y, para cada uno de ellos, todas sus variantes alomórficas. El conjunto de sufijos empleados y sus correspondientes alomorfos se recogen en el apéndice D, habiendo sido recopilados a partir de los trabajos de Lang [135], Bajo [28], y Fernández [73].

Por otra parte, es necesario controlar esta sobregeneración ya que, de no ser así, podría dar lugar a una cierta degradación [233]. A fin de ejercer este control, el algoritmo confronta las formas generadas contra el lexicón del sistema, aceptando únicamente aquellos lemas existentes con la categoría gramatical deseada [233, 24]. Esta solución es similar al sistema empleado por los llamados *dictionary look-up stemmers* [131], y según el cual el *stem* obtenido es contrastado contra un diccionario de *stems* para verificar su validez. Mediante este filtrado en base al lexicón estamos a la vez resolviendo el problema de la decisión acerca de la validez y aceptación del término derivado.

Describimos a continuación el proceso de generación de familias:

Algoritmo 6.1 Algoritmo de generación de familias morfológicas:

Dados dos lemas w_x y w'_y del lexicón, con categorías gramaticales x e y respectivamente, denotamos mediante $w_x \triangleright w'_y$ el hecho de que w'_y sea obtenido a partir de w_x mediante alguno de los mecanismos derivativos considerados. Conforme a esto, calculamos la familia morfológica de w_x como su *cierre transitivo y reflexivo* mediante derivación, denotado por $cierre(w_x)$ y definido recursivamente como:

$w_x \in cierre(w_x)$
if $w_x \triangleright w'_y$ **then** $w'_y \in cierre(w_x)$
if $w'_y \triangleright w_x$ **then** $w'_y \in cierre(w_x)$

El conjunto de familias morfológicas asociadas a un lexicón dado se obtiene mediante la aplicación de $cierre(w_x)$ a cada lema w_x contenido en el lexicón.

□

Se trata, pues, de un algoritmo incremental, donde los nuevos derivados son obtenidos a partir tanto de la forma base inicial como de los derivados intermedios (convertidos asimismo en bases). De esta forma se soluciona el problema de la predicción del orden de los morfemas en los casos de afijación recursiva con acumulación de morfemas derivativos, como se muestra en los siguientes ejemplos:

lematización	lema+*-izar*+*-ción*
gobernabilidad	gobernar+*-able*+*-dad*

En estos casos existe una especie de control sintáctico del orden de los morfemas implicados [135]. Por ejemplo, la nominalización de un verbo requiere un sufijo nominalizador pospuesto al morfema verbal (*-ción* tras *-izar* en *lematización*). Sin embargo, en otros casos, dichos criterios de ordenación de la secuencia no están completamente definidos. La aplicación del concepto relajado de relación morfoléxica frente al estrictamente lingüístico de derivación nos permite solucionar el problema de la sufijación acumulativa mediante un algoritmo incremental, donde cada nueva forma derivada puede actuar a su vez como base, sin tener que preocuparse de criterios etimilógicos ni diacrónicos.

En lo que respecta a la derivación regresiva, ésta es implementada por el algoritmo de modo implícito ya que, en lugar de derivar el sustantivo a partir del verbo, es el sustantivo el que genera el verbo mediante verbalización denominal. Esto es posible, de nuevo, gracias a la aplicación del concepto relajado de relación morfoléxica.

Debemos precisar que, dado que únicamente se han contemplado los mecanismos derivativos de generación de nombres, verbos y adjetivos a partir de dichas mismas categorías, sólo se procesarán las entradas del lexicón correspondientes a tales categorías: nombres[5], verbos, y adjetivos calificativos.

6.3.3. Tratamiento de la Alomorfia

Muchos morfemas derivativos presentan formas variables, alomorfos. Dentro de dichos fenómenos alomórficos existen variantes que son excluyentes entre sí, como pueden ser aquellas cuya elección viene dada por la vocal temática[6] —p.ej., *-amiento* para la vocal *'a'* e *-imiento* para las vocales *'e'* e *'i'*—, mientras que otros no lo son, como es el caso de *-ado/-ato* y *-azgo* (variantes vulgar y culta, respectivamente), dando lugar en ocasiones a resultados diferentes aplicados a la misma base; por ejemplo:

$$líder \rightarrow \begin{cases} liderato \\ liderazgo \end{cases}$$

La variante a aplicar en cada caso dependerá de cada sufijo en particular, pudiendo influir otros factores como la vocal temática y la forma de la base. Es por ello que los diferentes casos para cada sufijo particular han sido considerados e implementados por separado.

[5]Únicamente sustantivos comunes, ya que la inclusión de propios podría introducir a posteriori distorsión en el sistema.

[6]Segmento vocálico, determinado conjugacionalmente, que aparece en el derivado entre la raíz y el sufijo —*'a'* para la primera conjugación, *'e'* para la segunda, e *'i'* para la tercera.

6.3.4. Tratamiento de las Condiciones Fonológicas

Como hemos apuntado anteriormente, toda operación morfológica —en este caso de derivación— implica a su vez una alteración fonológica de la base. No se trata pues, de una mera concatenación de morfemas, sino que deberán aplicarse, en ocasiones, ciertas modificaciones ortográficas fruto de dichas alteraciones [121]. Las reglas de ajuste fonológico [28, 73] implementadas por nuestro sistema son:

- *Supresión de la vocal final átona*: constituye el comportamiento por defecto del sistema. A la hora de concatenar los sufijos el sistema trabaja en principio sobre el término base, tras eliminar la vocal final en el caso de sustantivos y adjetivos; en caso de finalizar en consonante, el término base no se modifica. En todo caso, el término original se mantiene disponible. Ejemplos:

$$arena \rightarrow aren\text{-} \xrightarrow{-oso} arenoso$$
$$amor \rightarrow amor \xrightarrow{-oso} amoroso$$

- *Eliminación de cacofonías*: en ocasiones, al concatenar el sufijo al término base obtenido mediante el proceso anterior, dos vocales iguales quedan adyacentes. Ambas se fusionan para eliminar la cacofonía resultante, como en el caso de:

$$galanteo \rightarrow galante\text{-} \xrightarrow{-ería} galantee ría \rightarrow galantería$$

- *Vocal temática*: en el caso de que el término primitivo sea un verbo, basta con comprobar si acaba en *-ar/-er/-ir/-ír* para conocer la vocal temática y así ser tenida en cuenta, por ejemplo, a la hora de escoger la variante alomórfica a utilizar. Una muestra es el caso de *-miento/-amiento/-imiento/-mento*, donde *-amiento* sólo se emplea con vocal temática *'a'* e *-imiento* con las vocales *'e'* e *'i'*. Por ejemplo:

$$alzar \rightarrow alz\text{-} \xrightarrow{-amiento} alzamiento$$
$$aburrir \rightarrow aburr\text{-} \xrightarrow{-imiento} aburrimiento$$

- *Monoptongación de la raíz diptongada*: se sustituye el diptongo por la forma pertinente. Se considera la monoptongación de *'ie'* en *'e'* y de *'ue'* en *'o'*. Por ejemplo:

$$\text{ie} \rightarrow \text{e} \qquad diente \xrightarrow{-al} diental \rightarrow dental$$
$$\text{ue} \rightarrow \text{o} \qquad fuerza \xrightarrow{-udo} fuerzudo \rightarrow forzudo$$

- *Cambio en la posición del acento*: puesto que los sufijos producen generalmente un cambio en la acentuación, dicha situación debe ser considerada, ya que puede conllevar cambios ortográficos debidos a la aparición o desaparición de tildes. La práctica totalidad de los sufijos son tónicos, con lo que es inmediato saber si se debe introducir o eliminar una tilde aplicando las reglas ortográficas pertinentes. Por ejemplo:

$$europeo \rightarrow europe\text{-} \xrightarrow{-ista} europeísta$$
$$novela \rightarrow novel\text{-} \xrightarrow{-ista} novelista$$

- *Mantenimiento de los fonemas consonánticos finales*: partiendo de la forma base se puede conocer el fonema original —p.ej., la segunda 'c' de *co<u>c</u>er* corresponde al fonema /θ/ y no al /k/. De la misma forma, conociendo el fonema se puede deducir la grafía final. Por ejemplo, la 'z' en *cerveza* corresponde al fonema /θ/ y, por consiguiente:

$$cerve\underline{z}a \rightarrow cervez\text{-} \xrightarrow{-ería} cerve\underline{c}ería$$

  Los fonemas y cambios cubiertos son:

/k/	c → qu
/γ/	g → gü
/γ/	g → gu
/θ/	z → c
/θ/	c → z

- *Reglas ad-hoc*: nos referimos a ajustes varios tales como modificaciones en la consonante final de la raíz en casos como:

$$conce\underline{d}er \rightarrow conce\underline{s}ión$$

  Estos casos se resuelven mediante reglas ad-hoc, es decir, que operan para un sufijo dado. Frecuentemente vienen dados por la presencia de fonemas dentales /δ/ o /t/.

6.3.5. Medidas Adicionales para el Control de la Sobregeneración

Con objeto de llevar a cabo un control más riguroso de la sobregeneración durante el proceso de generación de familias para, de este modo, minimizar en la medida de lo posible el ruido introducido, se han efectuado una serie de modificaciones en el sistema de generación inicial [242, 243]. Tales modificaciones comprenden:

- Establecimiento de una longitud mínima de las raíces intermedias obtenidas durante el proceso de generación de familias, de forma similar a como ocurre en el caso de los algoritmos de *stemming* [179, 139]. Se exige que una raíz intermedia válida deberá tener una longitud mínima de 3 letras y contener al menos una secuencia vocal+consonante. La razón de esta medida estriba en que las raíces intermedias de menor longitud son más proprensas a introducir ruido debido a la generación de relaciones derivativas incorrectas, como por ejemplo:

$$ano \xrightarrow{-al} anal \xleftarrow{-al*} ana$$

- Eliminación de las palabras de uso marginal presentes en el lexicón de entrada. El lexicón empleado por el sistema es extremadamente completo, pues no sólo contiene el vocabulario de uso común del español, sino que contempla además una ingente cantidad de americanismos, tecnicismos, cultismos, y otras palabras de uso poco frecuente. Puesto que nuestro sistema ha sido desarrollado para su empleo en un sistema de recuperación de información general, no especializado, no es necesaria la inclusión de estos términos, ya que su empleo en tal contexto será escaso. Por otra parte, a mayor número de términos en el lexicón a procesar, mayor será la probabilidad de introducir ruido en el sistema mediante el establecimiento de relaciones derivativas incorrectas.

	lex. inicial	*lex. filtrado*
sustantivos	57002	20429
adjetivos	22061	8776
verbos	5472	4287
TOTAL	84535	33492

Tabla 6.3: Distribución de lemas de palabras con contenido en el lexicón de entrada antes y después del filtrado

Dado el tamaño del lexicón —84535 lemas de palabras con contenido—, la eliminación manual de estos términos resulta poco viable. Por esta razón se optó por realizar el filtrado de forma automática, en base a un corpus lo suficientemente amplio como para ser representativo de las palabras de uso más común del lenguaje. El corpus de documentos CLEF 2003 fue elegido como referencia para el filtrado por su gran extensión —454045 documentos. De esta forma, de cara a la generación de familias, fueron eliminadas del lexicón aquellas entradas correspondientes a lemas que no estuviesen presentes en el diccionario de lemas obtenido durante la indexación del corpus CLEF 2003.

- La aplicación de la parasíntesis, contemplada inicialmente [242, 243], ha sido descartada ya que tendía a introducir relaciones derivativas incorrectas. Por ejemplo:

$$\text{bollo} \overset{a- \ -ar*}{\longrightarrow} \text{abollar} \qquad \text{lente} \overset{a- \ -ar*}{\longrightarrow} \text{alentar} \qquad \text{nuncio} \overset{a- \ -ar*}{\longrightarrow} \text{anunciar}$$

- Las familias que contenían un alto número de lemas han sido descartadas, ya que en su práctica totalidad se trataba de familias que integraban dos o más familias que habían sido ligadas mediante el establecimiento de alguna relación derivativa incorrecta, como en el caso de:

$$\ldots \longrightarrow \textit{internar} \overset{-ción}{\longrightarrow} \textit{internación} \overset{-al*}{\longrightarrow} \textit{internacional} \longrightarrow \ldots$$

6.4. Familias Morfológicas y Normalización de Términos

El primer paso en la normalización de términos mediante esta nueva aproximación basada en familias morfológicas es la propia generación de las familias. Partiendo del lexicón inicial del sistema, se procede a su filtrado contra el diccionario del índice de lemas del corpus CLEF 2003, para de este modo obtener el conjunto final de lemas que alimentarán el algoritmo de generación. La composición de dicho lexicón se recoge en la tabla 6.3. Una vez filtrado el lexicón de entrada, se ejecuta el algoritmo de generación. La distribución de las familias obtenidas y sus lemas asociados en función del tamaño de la familia —número de lemas que la integran— se detalla en la tabla 6.4.

Una vez que las familias han sido generadas, el proceso de normalización consiste en una ampliación del proceso llevado a cabo mediante lematización. En una primera fase, se elimina la variación morfológica flexiva. Para ello, el texto es primero *tokenizado* y segmentado mediante el preprocesador de base lingüística, para ser luego desambiguado y lematizado mediante el etiquetador-lematizador MrTagoo, proceso ya descrito en el capítulo 5.

A continuación se identifican los lemas de las palabras con contenido, los cuales son reemplazados por un representante de su familia morfológica fijado previamente. De esta forma

tamaño	*#familias*	*%familias*	*#lemas*	*%lemas*
1	10262	59.77 %	10262	30.64 %
2	3504	20.41 %	7008	20.92 %
3	1438	8.38 %	4314	12.88 %
4	797	4.64 %	3188	9.52 %
5	401	2.34 %	2005	5.99 %
6	243	1.42 %	1458	4.35 %
7	169	0.98 %	1183	3.53 %
8	94	0.55 %	752	2.25 %
9	60	0.35 %	540	1.61 %
10	49	0.29 %	490	1.46 %
+10	153	0.89 %	2292	6.84 %
TOTAL	17170	100.00 %	33492	100.00 %

Tabla 6.4: Distribución de las familias y sus lemas asociados en función del tamaño de la familia

se pretende resolver la variación morfológica derivativa, ya que todas las palabras pertenecientes a la misma familia son normalizadas al mismo término índice. Se comporta, pues, como un *stemmer* avanzado de base lingüística para el español.

En lo referente a los costes respecto a la normalización mediante lematización, éstos son mínimos. El proceso de generación de familias y de asignación de representantes se realiza a priori, por lo que los costes del proceso original no se ven incrementados. En lo que respecta al proceso de sustitución de los lemas por el representante de su familia, una vez elegido el representante —el término escogido es indiferente— existen dos alternativas posibles:

1. Si no es preciso seguir teniendo acceso al lema de la palabra, el proceso de sustitución puede llevarse a cabo mediante el propio etiquetador-lematizador. Para ello basta sustituir el lexicón empleado por el etiquetador-lematizador, con entradas `forma etiqueta lema`, por un lexicón modificado donde para las entradas de las palabras con contenido el lema de la palabra sea substituido por el representante de su familia, dando lugar a una entrada `forma etiqueta representante`. De esta forma, al etiquetar y lematizar el texto se le estará asignando a cada palabra no ya el lema, sino su representante. Esta solución no implica, por tanto, aumento alguno en el coste computacional del sistema.

2. De ser necesario seguir teniendo acceso al lema de la palabra, el proceso de sustitución deberá llevarse a cabo, por lo tanto, posteriormente al proceso de etiquetación-lematización. De este modo se hace necesario contar con una estructura de almacenamiento de los pares `lema representante` lo más eficiente posible, tanto de cara al almacenamiento como de cara a la búsqueda, para así minimizar los costes asociados al proceso de sustitución. La solución a este problema pasa por el empleo de autómatas de estado finito [86], solución ya empleada anteriormente en la implementación del lexicón del etiquetador-lematizador —véase apartado 5.2.2.

Tabla 6.5: Resultados obtenidos mediante lematización (*lem*), caso base, y normalización mediante familias morfológicas (*fam*)

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	*fam*	%Δ	*lem*	*fam*	%Δ	*lem*	*fam*	%Δ	*lem*	*fam*	%Δ	*lem*	*fam*	%Δ	*lem*	*fam*	%Δ
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2700	**2703**	**0.11**	2762	**2776**	**0.51**	2363	**2391**	**1.18**	2410	**2418**	**0.33**	2159	**2161**	**0.09**	2175	**2182**	**0.32**
Pr. no int.	.4829	.4777	-1.08	.5239	.5175	-1.22	.4678	**.4819**	**3.01**	.5256	.5243	-0.25	.4324	.4289	-0.81	.4659	.4584	-1.61
Pr. doc.	.5327	**.5355**	**0.53**	.5690	**.5717**	**0.47**	.5667	.5661	-0.11	.6089	.5939	-2.46	.4724	**.4864**	**2.96**	.5201	.5137	-1.23
R-pr.	.4848	.4609	-4.93	.5075	.4946	-2.54	.4549	**.4778**	**5.03**	.5030	.4997	-0.66	.4362	.4346	-0.37	.4493	.4343	-3.34
Pr. a 0 %	.8293	**.8371**	**0.94**	.8845	.8489	-4.02	.8234	.8195	-0.47	.8763	.8716	-0.54	.8124	.7742	-4.70	.8366	.8220	-1.75
Pr. a 10 %	.7463	.7277	-2.49	.7914	.7796	-1.49	.6845	**.7070**	**3.29**	.7517	.7437	-1.06	.7057	.6829	-3.23	.7197	.7185	-0.17
Pr. a 20 %	.6771	.6307	-6.85	.7136	.6980	-2.19	.6287	**.6627**	**5.41**	.7095	.6901	-2.73	.6118	**.6125**	**0.11**	.6453	.6335	-1.83
Pr. a 30 %	.6138	.5861	-4.51	.6591	.6475	-1.76	.5829	**.6001**	**2.95**	.6463	.6406	-0.88	.5503	**.5631**	**2.33**	.5788	.5673	-1.99
Pr. a 40 %	.5502	.5486	-0.29	.6085	.6039	-0.76	.5555	**.5593**	**0.68**	.6100	.5890	-3.44	.5059	.4970	-1.76	.5348	.5197	-2.82
Pr. a 50 %	.4931	**.4977**	**0.93**	.5557	.5508	-0.88	.5136	.5105	-0.60	.5658	.5590	-1.20	.4657	.4512	-3.11	.4889	.4725	-3.35
Pr. a 60 %	.4496	**.4521**	**0.56**	.5006	.4972	-0.68	.4413	**.4624**	**4.78**	.4977	.4947	-0.60	.4017	.3899	-2.94	.4339	.4100	-5.51
Pr. a 70 %	.3853	**.3973**	**3.11**	.4161	**.4265**	**2.50**	.3943	**.4174**	**5.86**	.4569	.4557	-0.26	.3422	.3304	-3.45	.3789	.3670	-3.14
Pr. a 80 %	.3277	**.3427**	**4.58**	.3509	**.3575**	**1.88**	.3122	**.3215**	**2.98**	.3679	.3633	-1.25	.2700	**.2766**	**2.44**	.3206	.3158	-1.50
Pr. a 90 %	.2356	**.2368**	**0.51**	.2492	**.2544**	**2.09**	.2319	**.2401**	**3.54**	.2845	**.2996**	**5.31**	.1821	**.2024**	**11.15**	.2067	**.2217**	**7.26**
Pr. a 100 %	.1197	.1018	-14.95	.1289	.1282	-0.54	.1209	**.1254**	**3.72**	.1547	**.1625**	**5.04**	.0932	**.0968**	**3.86**	.1164	**.1186**	**1.89**
Pr. a 5 docs.	.6609	.6304	-4.61	.6957	.6913	-0.63	.5956	**.6356**	**6.72**	.6844	.6844	0.00	.5915	.5702	-3.60	.6213	**.6298**	**1.37**
Pr. a 10 docs.	.6283	.5870	-6.57	.6543	.6370	-2.64	.5667	**.5800**	**2.35**	.6000	**.6044**	**0.73**	.5149	**.5298**	**2.89**	.5596	.5532	-1.14
Pr. a 15 docs.	.5928	.5667	-4.40	.6188	.6043	-2.34	.5170	**.5274**	**2.01**	.5689	.5674	-0.26	.4738	**.4894**	**3.29**	.5106	.5007	-1.94
Pr. a 20 docs.	.5446	.5359	-1.60	.5880	.5717	-2.77	.4878	**.4933**	**1.13**	.5422	.5311	-2.05	.4457	**.4468**	**0.25**	.4819	.4660	-3.30
Pr. a 30 docs.	.4928	.4862	-1.34	.5304	.5261	-0.81	.4452	**.4481**	**0.65**	.4948	.4852	-1.94	.4000	.3915	-2.13	.4255	.4121	-3.15
Pr. a 100 docs.	.3300	.3259	-1.24	.3509	.3437	-2.05	.2993	**.3011**	**0.60**	.3147	.3100	-1.49	.2513	**.2547**	**1.35**	.2702	.2668	-1.26
Pr. a 200 docs.	.2234	.2227	-0.31	.2315	.2302	-0.56	.1997	**.2032**	**1.75**	.2093	.2060	-1.58	.1617	**.1656**	**2.41**	.1694	**.1720**	**1.53**
Pr. a 500 docs.	.1090	**.1096**	**0.55**	.1115	**.1133**	**1.61**	.0983	**.0992**	**0.92**	.1023	.1018	-0.49	.0827	**.0834**	**0.85**	.0842	**.0846**	**0.48**
Pr. a 1000 docs.	.0587	**.0588**	**0.17**	.0600	**.0603**	**0.50**	.0525	**.0531**	**1.14**	.0536	**.0537**	**0.19**	.0459	**.0460**	**0.22**	.0463	**.0464**	**0.22**

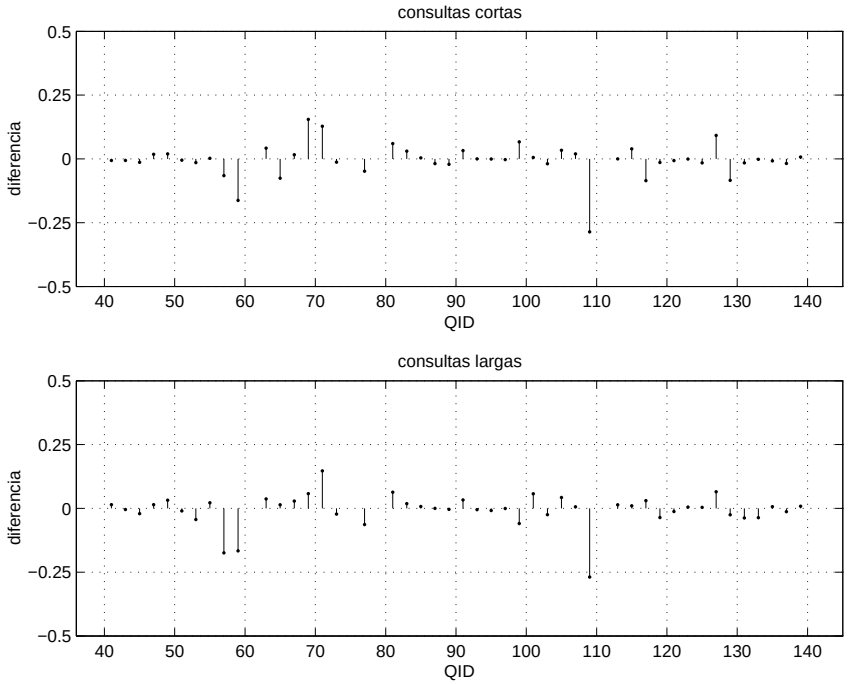


Figura 6.1: Diferencias en las precisiones no interpoladas: lematización vs. familias. Corpus CLEF 2001-02·A

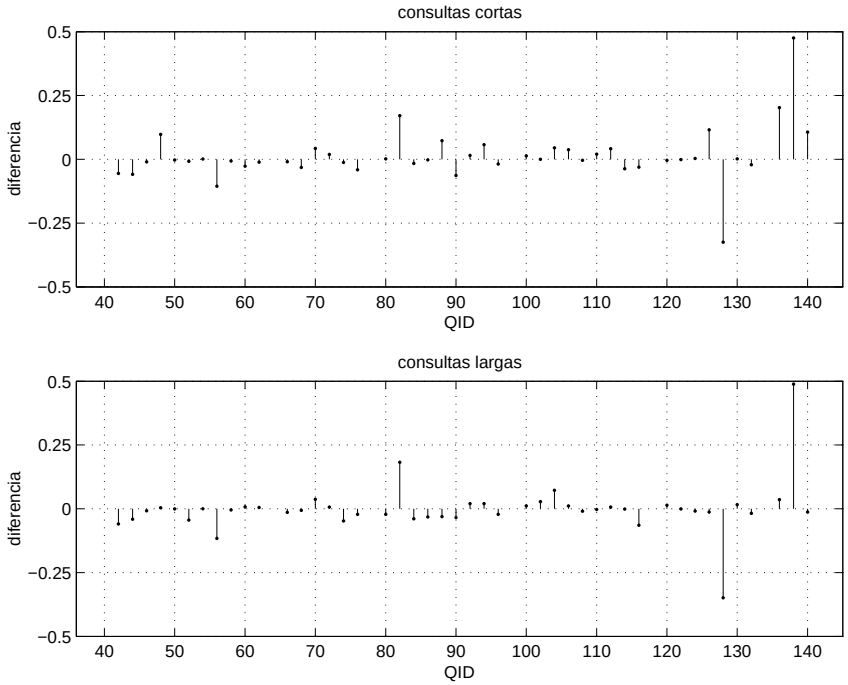


Figura 6.2: Diferencias en las precisiones no interpoladas: lematización vs. familias. Corpus CLEF 2001-02·B

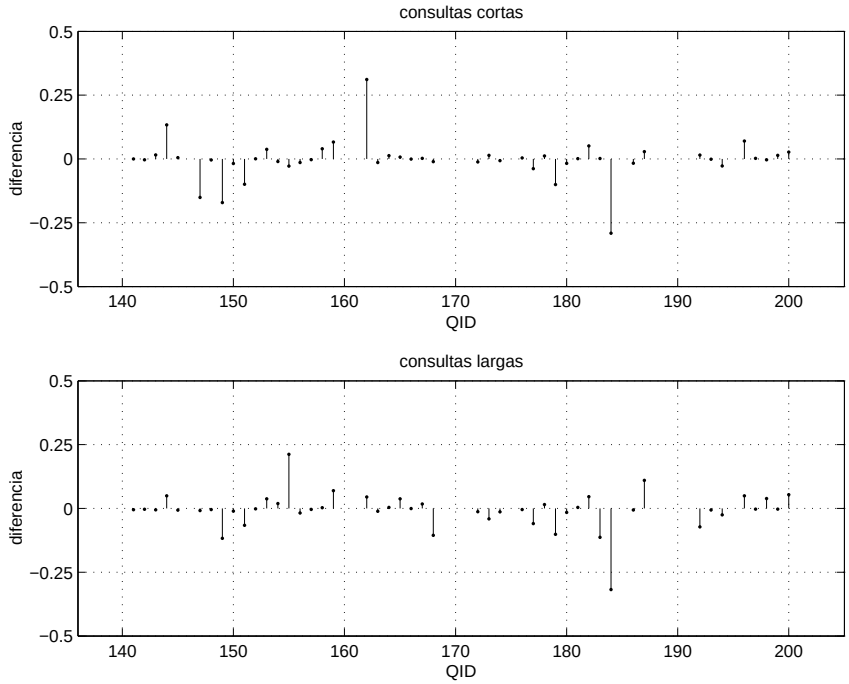


Figura 6.3: Diferencias en las precisiones no interpoladas: lematización vs. familias. Corpus CLEF 2003

6.5. Resultados Experimentales

6.5.1. Resultados sin Realimentación

La tabla 6.5 muestra los resultados obtenidos mediante nuestra propuesta para la resolución de la variación derivativa basada en familias morfológicas (*fam*). Para cada caso se ha señalado el porcentaje de mejora obtenido ($\%\Delta$) frente a la línea base, que en este caso será la normalización mediante lematización (*lem*), nuestra propuesta para la resolución de la variación flexiva. Los casos en los que se produce un incremento del rendimiento han sido remarcados en negrita.

Como se puede apreciar, el comportamiento del sistema al aplicar familias morfológicas presenta ciertas irregularidades. Las mejoras obtenidas mediante el empleo de familias se restringen en su mayor parte a las consultas cortas —de forma similar a como ocurre en las aproximaciones clásicas basadas en *stemming* [105]—, siendo menores en el caso del corpus CLEF 2001-02·A. Esta diferencia entre consultas cortas y largas se debe al menor margen de mejora que las consultas largas tiene respecto a las cortas, ya que la información introducida a mayores por los términos de la narrativa permite identificar de un modo bastante aproximado el conjunto de documentos relevantes sin tener que recurrir necesariamente a técnicas más complejas.

A nivel de consulta particular, las gráficas de diferencia de precisiones no intepoladas —figuras 6.1, 6.2 y 6.3— muestran, en general, una escasa incidencia del empleo de familias respecto a la lematización. Existen también excepciones donde la variación en el rendimiento es patente, como ocurre en la consulta 138 del corpus CLEF2001-02·B[7], acerca de los extranjerismos en la lengua francesa. En dicha consulta el empleo de familias morfológicas permite correspondencias entre los términos *extranjerismo* y *extranjero*, lo cual redunda en un incremento del rendimiento

[7] **Consulta C138:** <ES-title> Extranjerismos en el francés </ES-title> <ES-desc> Encontrar documentos que traten del papel de los extranjerismos en la lengua francesa. </ES-desc> <ES-narr> El tribunal constitucional ha revisado, para hacerla menos estricta, la ley de Toubon sobre el uso del idioma francés. Los documentos relevantes discuten el papel y el uso de extranjerismos en el francés. </ES-narr>

del sistema, ya que los artículos relevantes apenas hacían referencia al término *extranjerismo*, sino que empleaban expresiones como *palabras extranjeras* o similares. Existen también ejemplos de lo contrario, consultas en las que el empleo de familias influye negativamente en el rendimiento. Este es el caso de la consulta 109 también del corpus CLEF2001-02·B[8], sobre la seguridad en las redes informáticas. En ella aparece, con un alto poder discriminante, el término *informático*, a cuya familia pertenece también, de modo incorrecto, el término *información*. Debido al ruido que esto introduce durante el proceso de normalización, nos encontramos que al lanzar la consulta contra el sistema, éste nos devuelve gran cantidad de documentos no relevantes referentes a *redes de información*, tanto refiriéndose con ello a *redes de comunaciones*, como a *servicios de inteligencia*. Dichos documentos no eran devueltos, sin embargo, en el caso de la lematización.

El análisis de los resultados obtenidos nos permite afirmar que el problema de la sobregeneración no ha podido ser resuelto en su totalidad mediante el filtrado contra el lexicón de las formas generadas. Esto se traduce en ruido introducido en el sistema durante el proceso de normalización de términos, lo cual puede afectar negativamente a su rendimiento. Las fuentes principales de ruido pueden clasificarse en:

- *Variación semántica.* Los fenómenos de sinonimia asociados a un término se propagan a toda su familia, aumentando su impacto. Tómese como ejemplo *atracar*, ya que tanto su término derivado *atraco*, correspondiente a su acepción *"asaltar con propósito de robo"*, como su también derivado *atraque*, esta vez correspondiente a su acepción *"arrimar un buque a tierra o a otro"*, son ambos normalizados al mismo término.

- *Distancia semántica excesiva entre términos de la misma familia.* Este es el caso de *cargar* y *carguero*, ya que si bien pertenecen a la misma familia y acepción, una consulta sobre *(buques) cargueros* se vería completamente distorsionada si el término *carguero*, muy específico, es normalizado al mismo término índice que *cargar*, excesivamente genérico y con escasa relación con el término general en sus contextos de uso.

- *Especialización de significados.* Como ocurre con el término *embellecedor (de un coche)* y su base *embellecer*.

- *Derivaciones incorrectas.* Como por ejemplo en *internar* $\overset{-ción}{\longrightarrow}$ *internación* $\overset{-al*}{\longrightarrow}$ *internacional*.

Se trata, en todos los casos, de una cuestión semántica, donde la solución a dichos problemas pasa, por una parte, por la aplicación de mecanismos de desambiguación del sentido de las palabras, y por otra parte, por que no se establezca una relación entre dos términos de no existir una relación semántica suficiente entre ambos, independientemente de si existe una relación puramente derivativa entre los mismos. Sin embargo, la falta de recursos apropiados de libre acceso para el español —tales como corpus anotados semánticamente—, constituye un inconveniente de cara a su aplicación, además del incremento de costes que tal solución conllevaría.

6.5.2. Resultados Aplicando Realimentación

La tabla 6.6 compara nuevamente los resultados obtenidos mediante la normalización basada en familias morfológicas (*fam*) y la basada en lematización[9] (*lem*), si bien en esta nueva

[8] **Consulta C109:** `<ES-title>` Seguridad Informática `</ES-title>` `<ES-desc>` ¿Cuál es la situación de la seguridad informática en relación con el acceso a través de una red? `</ES-desc>` `<ES-narr>` Los documentos deben referirse a la seguridad informática con respecto al uso de redes. Los informes que hacen referencia sólo a la seguridad física de los sistemas informáticos no son relevantes. `</ES-narr>`

[9] Se tomará como línea base la lematización con realimentación.

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	*fam*	%Δ	*lem*	*fam*	%Δ	*lem*	*fam*	%Δ	*lem*	*fam*	%Δ	*lem*	*fam*	%Δ	*lem*	*fam*	%Δ
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2767	**2777**	**0.36**	2779	2776	-0.11	2376	**2401**	**1.05**	2406	**2413**	**0.29**	2240	2232	-0.36	2223	**2242**	**0.85**
Pr. no int.	.5220	.5140	-1.53	.5604	.5508	-1.71	.4773	**.5007**	**4.90**	.5392	.5372	-0.37	.5024	.4733	-5.79	.5207	.5131	-1.46
Pr. doc.	.5784	**.5819**	**0.61**	.5912	**.5915**	**0.05**	.5681	.5614	-1.18	.6145	.5950	-3.17	.5530	.5266	-4.77	.5802	.5618	-3.17
R-pr.	.4990	.4863	-2.55	.5366	.5254	-2.09	.4599	**.4727**	**2.78**	.5104	.5098	-0.12	.4912	.4673	-4.87	.4871	**.4910**	**0.80**
Pr. a 0 %	.8221	**.8332**	**1.35**	.8895	.8503	-4.41	.8210	.8177	-0.40	.8710	.8699	-0.13	.8145	.7740	-4.97	.8301	.8200	-1.22
Pr. a 10 %	.7490	.7352	-1.84	.8028	.7816	-2.64	.6861	**.7003**	**2.07**	.7619	.7484	-1.77	.7369	.6951	-5.67	.7421	**.7489**	**0.92**
Pr. a 20 %	.6866	.6686	-2.62	.7352	.7177	-2.38	.6319	**.6572**	**4.00**	.6929	**.7045**	**1.67**	.6632	.6266	-5.52	.6758	**.6769**	**0.16**
Pr. a 30 %	.6573	.6298	-4.18	.6996	.6764	-3.32	.5688	**.6001**	**5.50**	.6497	.6481	-0.25	.6019	.5708	-5.17	.6304	.6146	-2.51
Pr. a 40 %	.5997	.5859	-2.30	.6541	.6400	-2.16	.5289	**.5618**	**6.22**	.6202	.6019	-2.95	.5638	.5345	-5.20	.5975	.5679	-4.95
Pr. a 50 %	.5456	.5447	-0.16	.6005	.5955	-0.83	.5017	**.5329**	**6.22**	.5733	.5659	-1.29	.5410	.4946	-8.58	.5479	.5198	-5.13
Pr. a 60 %	.4994	**.5019**	**0.50**	.5386	.5381	-0.09	.4623	**.4912**	**6.25**	.5194	.5146	-0.92	.4783	.4572	-4.41	.4938	.4817	-2.45
Pr. a 70 %	.4375	.4354	-0.48	.4621	**.4643**	**0.48**	.4255	**.4528**	**6.42**	.4862	.4717	-2.98	.4366	.4180	-4.26	.4587	.4518	-1.50
Pr. a 80 %	.3661	**.3696**	**0.96**	.3910	**.4032**	**3.12**	.3384	**.3725**	**10.08**	.3753	**.3837**	**2.24**	.3788	.3538	-6.60	.3891	.3748	-3.68
Pr. a 90 %	.2939	.2793	-4.97	.3130	.3050	-2.56	.2651	**.2842**	**7.20**	.3059	**.3196**	**4.48**	.3102	.2838	-8.51	.3230	.3146	-2.60
Pr. a 100 %	.1547	.1345	-13.06	.1624	.1544	-4.93	.1395	**.1730**	**24.01**	.1704	**.1890**	**10.92**	.1871	.1612	-13.84	.1928	**.2048**	**6.22**
Pr. a 5 docs.	.6609	.6304	-4.61	.6957	.6913	-0.63	.5956	**.6311**	**5.96**	.6844	.6844	0.00	.5872	.5702	-2.90	.6213	**.6340**	**2.04**
Pr. a 10 docs.	.6457	.6087	-5.73	.6848	.6652	-2.86	.5600	**.5822**	**3.96**	.6178	.6111	-1.08	.5596	.5489	-1.91	.5872	.5851	-0.36
Pr. a 15 docs.	.5884	**.5899**	**0.25**	.6435	.6319	-1.80	.5274	**.5481**	**3.92**	.5822	.5822	0.00	.5305	.5050	-4.81	.5504	.5390	-2.07
Pr. a 20 docs.	.5630	.5630	0.00	.6043	**.6065**	**0.36**	.5011	**.5089**	**1.56**	.5533	.5489	-0.80	.4883	.4691	-3.93	.5266	.5128	-2.62
Pr. a 30 docs.	.5225	.5188	-0.71	.5580	.5420	-2.87	.4444	**.4593**	**3.35**	.5081	.5000	-1.59	.4433	.4206	-5.12	.4667	.4525	-3.04
Pr. a 100 docs.	.3507	.3459	-1.37	.3598	.3572	-0.72	.2940	**.2960**	**0.68**	.3191	.3151	-1.25	.2770	.2726	-1.59	.2853	.2834	-0.67
Pr. a 200 docs.	.2348	.2336	-0.51	.2361	**.2375**	**0.59**	.1979	**.1994**	**0.76**	.2067	.2021	-2.23	.1753	.1736	-0.97	.1791	.1757	-1.90
Pr. a 500 docs.	.1122	.1122	0.00	.1121	**.1125**	**0.36**	.0980	**.0995**	**1.53**	.1008	.0997	-1.09	.0869	.0869	0.00	.0871	**.0877**	**0.69**
Pr. a 1000 docs.	.0602	**.0604**	**0.33**	.0604	.0603	-0.17	.0528	**.0534**	**1.14**	.0535	**.0536**	**0.19**	.0477	.0475	-0.42	.0473	**.0477**	**0.85**

Tabla 6.6: Resultados obtenidos mediante lematización (*lem*), caso base, y normalización mediante familias morfológicas (*fam*) aplicando en ambos casos realimentación

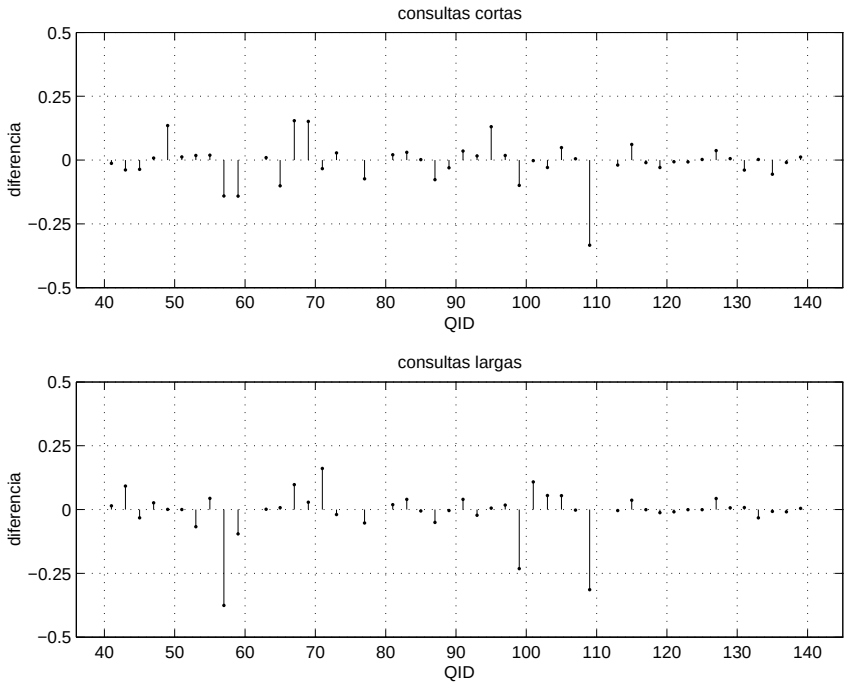


Figura 6.4: Diferencias en las precisiones no interpoladas aplicando realimentación: lematización vs. familias. Corpus CLEF 2001-02·A

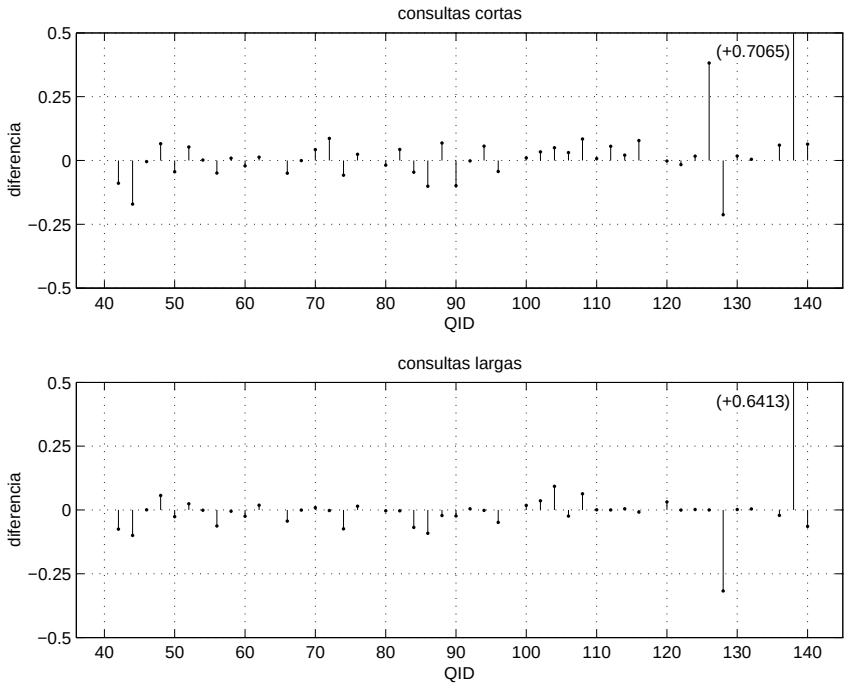


Figura 6.5: Diferencias en las precisiones no interpoladas aplicando realimentación: lematización vs. familias. Corpus CLEF 2001-02·B

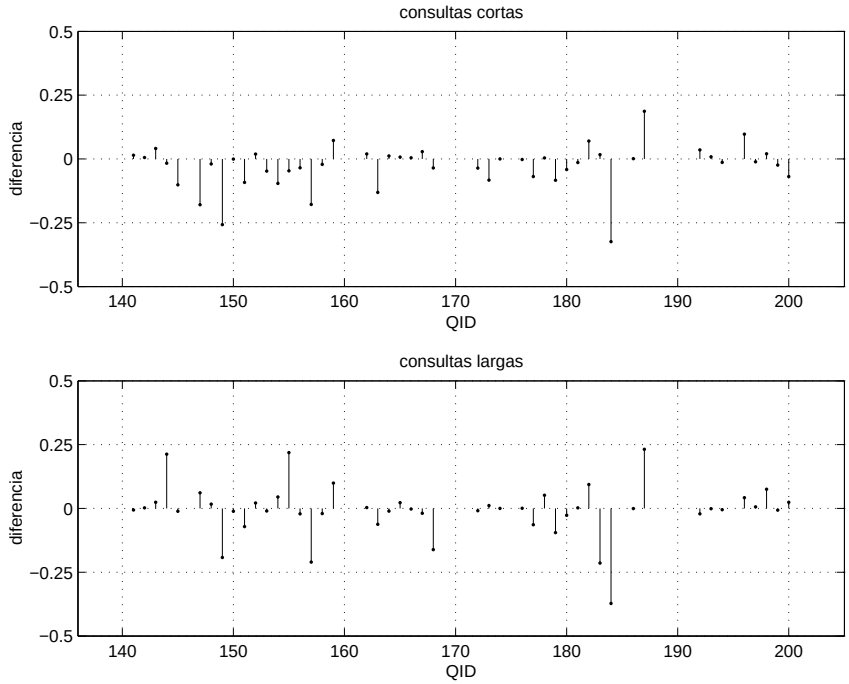


Figura 6.6: Diferencias en las precisiones no interpoladas aplicando realimentación: lematización vs. familias. Corpus CLEF 2003

serie de pruebas las consultas son expandidas automáticament mediante la técnica de Rocchio, empleando los parámetros establecidos anteriormente:

Consultas cortas: α=0.80, β=0.10, γ=0, n_1=5, t=10.

Consultas largas: α=1.20, β=0.10, γ=0, n_1=5, t=10.

Se puede apreciar en los resultados que la introducción de la realimentación conlleva nuevamente un aumento del rendimiento, si bien el comportamiento relativo de las familias respecto a la lematización varía de acuerdo con la calidad de los documentos en base a los cuales se realizó la expansión —los 5 primeros, n_1=5. De esta forma, en el caso de las consultas cortas del corpus CLEF 2001-02·B o de las consultas largas del CLEF 2003, los resultados siguen siendo similares o incluso mejores, ya que los documentos empleados en la expansión eran mejores, tal como indicaban las precisiones a los 5 documentos de la tabla 6.5. Sin embargo, en el caso de las consultas cortas del corpus CLEF 2003, los 5 primeros documentos devueltos por las familias eran peores que aquéllos devueltos mediante lematización —obsérvese la caída en la precisión a los 5 documentos en la tabla 6.5—, lo que provoca una caída del rendimiento relativo respecto a la lematización a la hora de expandir la consulta. Por su parte, las gráficas de diferencias a nivel de consulta en la precisión no interpolada —figuras 6.4, 6.5 y 6.6—, corroboran los resultados generales.

6.6. Discusión

En este capítulo se ha mostrado cómo los mecanismos de derivación para la formación de palabras en español pueden ser formalizados y utilizados en la construcción de una herramienta para la generación automática de familias morfológicas, conjuntos de palabras que comparten

una misma raíz. Las familias obtenidas pueden ser empleadas, a modo de *stemmer* avanzado de base lingüística, en tareas de normalización de términos simples para el tratamiento de la variación morfológica derivativa en la recuperación de información de textos en español.

Dado que únicamente se precisa de un lexicón para la generación de familias, esto posibilita la aplicación de esta aproximación en lenguas con escasos recursos lingüísticos disponibles, como es el caso del español. Muestra también de la validez de esta afirmación es la existancia de una versión en funcionamiento para el gallego [31, 30].

Sin embargo, nuestra propuesta no ha resultado ser por completo inmune al principal problema de las soluciones clásicas basadas en *stemming*, y que es el ruido introducido durante el proceso de normalización cuando dos términos de escasa o nula relación semántica son normalizados al mismo término índice [105, 74, 24]. En el caso concreto de las familias la razón para esto reside en la sobregeneración [233, 127] de la herramienta de creación de familias.

Parte III

Normalización de Términos Complejos

Capítulo 7

Tratamiento de la Variación Sintáctica mediante Análisis Sintáctico Superficial

7.1. Introducción

A lo largo de los capítulos anteriores hemos demostrado la viabilidad de las técnicas de Procesamiento del Lenguaje Natural para el tratamiento de la variación lingüística a nivel de palabra. El siguiente paso consistirá en aplicar técnicas de análisis a nivel de frase con un doble objetivo [24, 114, 230]: en primer lugar, obtener términos índice más precisos y descriptivos que los términos empleados hasta el momento; en segundo lugar, reducir la *variación sintáctica* del texto, que supone el mayor inconveniente a la hora de la utilización de términos complejos.

Para ello se hace necesario procesar el texto para identificar su estructura sintáctica. Esta tarea podría realizarse de un modo bastante preciso mediante cualquiera de los potentes algoritmos de análisis sintáctico descritos en la literatura [217] empleando una gramática de amplia cobertura del español. Desafortunadamente, el alto coste computacional de dichos algoritmos de análisis completo desaconseja su utilización en el procesamiento masivo de documentos, puesto que los analizadores sintácticos para gramáticas independientes del contexto generales tienen un coste temporal que crece cúbicamente con respecto al tamaño de los textos a analizar. Esta complejidad es incluso mayor en el caso de formalismos gramaticales más adecuados a la descripción lingüística de los fenómenos sintácticos [17]. Además de su alto coste, la falta de robustez de dichos algoritmos supone un inconveniente notable al pretender realizar un análisis completo de la oración, ya que reduce considerablemente la cobertura del sistema [24]. Este problema se agrava todavía más en el caso del español, dado que se ha de hacer frente, a mayores, a la falta de una gramática de amplia cobertura libremente accesible, o de un banco de árboles a partir del cual generarla. Debemos buscar, por lo tanto, un compromiso entre la calidad de la información sintáctica extraída y la facilidad de su obtención [24].

En este marco, el empleo de técnicas de *análisis sintáctico superficial* permite, por una parte, reducir la complejidad del sistema a una complejidad lineal y, por otra, incrementar la robustez del mismo. Conforme a ello, hemos optado por una aproximación basada en análisis sintáctico superficial frente a otras aproximaciones basadas en análisis completo aplicadas en otros sistemas [172]. El análisis superficial [11] ha demostrado su utilidad en diversos campos del NLP, en particular en el caso de los sistemas de Extracción de Información [23, 96, 103]. Su aplicación en Recuperación de Información no ha sido todavía estudiada en profundidad, centrándose prácticamente en idiomas diferentes al español, y a menudo limitándose al análisis

y obtención de frases nominales simples [132, 161, 106]. Por el contrario, nuestro sistema ha sido desarrollado para el español, cubriendo no sólo los sintagmas nominales sino también sus variantes sintácticas y morfosintácticas [114], incluso aquéllas que implican el análisis de sintagmas verbales.

En este capítulo se abordará la utilización de términos índice complejos como complemento al empleo de términos simples, así como los dos analizadores sintácticos superficiales —PATTERNS y CASCADE— desarrollados a tal efecto. Finalmente se evaluará comportamiento del sistema resultante.

7.2. Los Términos Complejos como Términos Índice

Las representaciones basadas en la indexación de *términos simples* permiten capturar de forma parcial la semántica de un documento. Para paliar esta carencia es frecuente recurrir al empleo de sintagmas para crear *términos índice complejos* o *multipalabra* [114] —términos que contienen dos o más palabras con contenido— y así complementar la información recogida por los términos simples.

Pensemos, por ejemplo, en los *sintagmas nominales*, los cuales hacen referencia a conceptos complejos y estáticos, siendo su papel fundamental en todo lenguaje humano al ser quienes actúan como sujetos y objetos, amén de su participación en sintagmas preposicionales. Por otra parte, los *sintagmas verbales* describen hechos, eventos y procesos de carácter dinámico al establecer relaciones entre el verbo principal de la oración y un número dado de sintagmas nominales u otros sintagmas [130, 24]. En ambos casos, el grado de especifidad del sintagma es mucho mayor que el del conjunto de sus términos componente aislados, dando lugar a términos índice más precisos y descriptivos que dichos componentes [230, 24]. Por su significatividad, los sintagmas a emplear como términos índice deberían incluir, al menos, a los sintagmas nominales y sus modificadores, así como las relaciones en las que están involucrados sintagmas verbales, tales como sujetos, objetos, y otros complementos.

Dado que el número posible de sintagmas de una lengua es prácticamente ilimitado, el espacio de términos multipalabra es mucho más disperso que el espacio de términos de los términos simples [130, 24], por lo que la necesidad de mecanismos de normalización apropiados se hace mucho más patente, de forma que las diversas formas semánticamente equivalentes de un sintagma puedan ser representadas mediante un mismo término índice.

7.2.1. Variantes Sintácticas

Las variantes sintácticas de un sintagma surgen como resultado de la flexión de sus componentes y de la modificación de su estructura sintáctica en base a los siguientes mecanismos:

Coordinación: empleo de construcciones coordinantes —copulativas o disyuntivas— bien en el núcleo del sintagma, bien en alguno de sus modificadores. Por ejemplo, los términos *desarrollo agrícola* y *desarrollo rural* se combinan en *desarrollo agrícola y rural*, que podemos considerar tanto una variante de *desarrollo agrícola* como de *desarrollo rural*.

Sustitución: adición de modificadores para así aumentar la especifidad del término considerado. Por ejemplo, *desarrollo del arte* puede transformarse en *desarrollo tardío del arte* sustituyendo *desarrollo* por el más específico *desarrollo tardío*.

Permutación: hace referencia a la permutación de palabras alrededor de un elemento pivote central. Por ejemplo, *saco viejo* y *viejo saco* son permutaciones del mismo término multipalabra.

Sinapsis: si bien las construcciones anteriores eran binarias, ésta es una construcción unaria donde la preposición empleada es sustituida por otra o bien un determinante es añadido o eliminado. Por ejemplo, al obtener la variante *abono para plantas* a partir de *abono para las plantas*.

7.2.2. Variantes Morfosintácticas

Las variantes morfosintácticas de un sintagma difieren de las sintácticas en que al menos una de las palabras con contenido del término original se transforma en otra palabra con la que guarda una relación derivativa. Dichas variantes se clasifican en función de la naturaleza de las transformaciones morfológicas que sufren sus términos:

Iso-categóricas: aquéllas para las que la derivación morfológica no altera la categoría gramatical de la palabra como, por ejemplo, en *producción artesanal* y *producto artesanal* (relación semántica proceso-resultado) o en *compuesto ionizador* y *compuesto ionizado* (relación semántica agente-resultado).

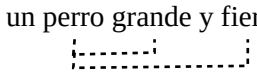
Hetero-categóricas: son aquéllas para las que sí se produce un cambio en la categoría gramatical de la palabra derivada. Esto nos permite englobar, por ejemplo, el caso de construcciones adjetivas y preposicionales equivalentes como *cambio del clima* y *cambio climático*, o las relaciones entre sintagmas nominales y verbales, como en *recortar los gastos* y *recorte de gastos* (relación semántica proceso-resultado).

7.2.3. Normalización en forma de Pares de Dependencia Sintáctica

El objetivo último a la hora de normalizar un sintagma es el de obtener una forma base común para todas las formas semánticamente equivalentes de una frase.

En nuestro caso, a la hora de indexar los términos multipalabra presentes en el texto, se ha optado por una representación de nivel intermedio, a medio camino entre una representación plana y fácilmente computable, y una representación sintáctica más completa pero compleja, como podrían ser árboles [182, 256] o grafos [167].

La forma que se ha elegido a la hora de representar el contenido de un sintagma es por medio de las dependencias que se establecen entre las distintas palabras que lo conforman. Si restringimos el número de palabras participantes en dichas dependencias a dos, se obtienen *pares de dependencia sintáctica*, tal y como se ilustra en el siguiente ejemplo:



donde, como se puede apreciar, el sintagma contiene dos dependencias:

1. Entre el sustantivo *perro* y el adjetivo *grande* que lo está modificando, dando lugar al par *(perro, grande)*.
2. Entre el sustantivo *perro* y el adjetivo *fiero* que también lo está modificando, dando lugar al par *(perro,fiero)*.

De esta forma, se extraerán del texto, mediante un proceso de análisis sintáctico superficial, aquellos pares de palabras que se encuentran ligadas por medio de alguno de las dependencias sintácticas más significativas [50]:

- Nombre-Adjetivo, relacionando el núcleo del sintagma nominal con su adjetivo modificador.
- Nombre-Complemento Nominal, relacionando el núcleo del sintagma nominal con el núcleo de su complemento.
- Sujeto-Verbo, relacionando el núcleo del sujeto con el verbo principal.
- Verbo-Complemento, relacionando el verbo principal con el núcleo de su complemento.

Dichos pares de dependencia serán empleados como términos índice complejos. Se trata, pues, de pares *núcleo-modificador*, una solución adoptada frecuentemente en aproximaciones similares [230, 130, 24]. Llegados a este punto debemos puntualizar que si bien dicha relación núcleo-modificador captura de forma indirecta la relación semántica pretendida, lo que estamos obteniendo es una relación de carácter estrictamente sintáctico [163, 172, 24].

Una vez que los pares han sido identificados y extraídos mediante análisis sintáctico superficial, los términos simples que lo componen son a su vez normalizados, primero mediante lematización —capítulo 5— y, seguidamente, mediante familias morfológicas —capítulo 6. De este modo estamos eliminando, por una parte, la flexión asociada tanto a variantes sintácticas como morfosintácticas y, por otra, los cambios derivativos propios de las variantes morfosintácticas.

Tomemos, por ejemplo, el caso de *aumento del gasto* y su variante morfosintáctica *los gastos aumentan*. Las dependencias extraídas son aquéllas entre los sustantivos *aumento* y *gastos* en el caso del término original, y entre la forma verbal *aumentan* y el sustantivo *gastos* en el caso de su variante. De eliminar únicamente la variación flexiva mediante lematización se generarían términos índice diferentes ya que, si bien en ambos pares está presente el lema *gasto*, diferirían en los componentes *aumento* y *aumentar*. Sin embargo, al existir una relación de carácter derivativo entre *aumentar* y *aumento*, podemos obtener el mismo par de dependencia para ambos términos en caso de emplear familias morfológicas para la normalización de los términos simples componentes[1]. De esta forma estamos normalizando el término original y la variante a un término común que permite su correspondencia. Debemos precisar, llegado este punto, que al trabajar sobre términos multipalabra el problema de la ambigüedad por sobregeneración en las familias se reduce, pues el propio par constituye en sí un pequeño contexto lingüístico que permite la desambigüación —al menos parcial— de sus componentes, reduciéndose así el ruido producido por familias mal construidas [233, 111].

En la literatura se describen otras aproximaciones que aplican también mecanismos de normalización sobre los componentes de los términos multipalabra: en [130] dichos componentes son lematizados para eliminar su flexión; [230] y [161] aplican técnicas de *stemming* para el mismo fin, lo que permite además dar cobertura también a los fenómenos derivativos; y en [24] se propone la nominalización de verbos y la verbalización de sustantivos con fines similares.

7.3. Análisis Sintáctico Superficial Basado en Patrones: el Analizador PATTERNS

En esta primera aproximación a la extracción de dependencias mediante análisis sintáctico superficial, éste es llevado a cabo por medio de patrones obtenidos aplanando los árboles sintácticos de las estructuras de interés.

[1] Lo que equivale a substituir cada lema por el representante de su familia morfológica —ver capítulo 6.

Para el desarrollo de dicho analizador sintáctico superficial, al que hemos denominado PATTERNS, se ha utilizado la siguiente gramática simple y aproximada del español :[2]

$$\begin{array}{lcll} S & \rightarrow & SN \;\; V \;\; W? \;\; (SN|SP)* & (1) \\ SN & \rightarrow & D? \;\; SA* \;\; N \;\; (SA|SP)* & (2) \\ SA & \rightarrow & W? \;\; A & (3) \\ SP & \rightarrow & P \;\; SN & (4) \end{array}$$

donde D, A, N, W, V y P son las etiquetas que representan respectivamente a determinantes, adjetivos, sustantivos, adverbios, verbos y preposiciones[3], y para la cual la motivación de las reglas de la gramática era la siguiente:

(1) representa una oración de estructura *Sujeto-Verbo-Predicado*.

(2) define el sintagma nominal como un sustantivo modificado por adjetivos y sintagmas preposicionales.

(3) permite que los adjetivos sean modificados por adverbios.

(4) representa un sintagma preposicional formado por una preposición y un sintagma nominal.

Dicha gramática es empleada a modo de guía, persiguiendo así un compromiso entre exhaustividad y exactitud. La metodología aplicada a la hora de establecer los mecanismos de identificación y extracción de las dependencias sintácticas del texto ha sido la siguiente [21, 258, 22]:

- En una primera etapa se estudiaron las construcciones sintácticas a considerar, tomando como base la estructura de los sintagmas nominales del español. Para ello se levantaron los árboles correspondientes a cada una de las posibles formas de construir un sintagma nominal, teniendo en cuenta sus posibles complementos, y tratando de generar un árbol lo más bajo posible. Una vez creados los árboles, se procedió a aplicar sobre ellos las transformaciones sintácticas y morfosintácticas más frecuentes en español [114, 112], dando lugar a nuevos árboles, cuyo alcance puede llegar a abarcar la oración completa en el caso de entrar en juego verbos derivados de los términos del árbol original. El conjunto de árboles obtenidos para los términos multipalabra (sintagmas nominales y sus variantes) se puede dividir en cuatro grandes grupos: *sustantivo modificado por adjetivo*, *sustantivo modificado por un sintagma preposicional*, *verbo-objeto* y *sujeto-verbo*.

- En una segunda etapa los árboles resultantes fueron aplanados con el fin de obtener una expresión regular en base a las categorías gramaticales de los términos involucrados, y que representase de forma aproximada los sintagmas y frases generados por los árboles obtenidos en la etapa precedente, limitando la complejidad de los patrones imponiendo restricciones respecto al uso de los operadores de repetición $^+$ y *. Mediante esta infraespecificación de la estructura se busca obtener una mayor generalidad del patrón y así poder aumentar la cobertura [111]. Finalmente, se identifican las dependencias sintácticas entre pares de palabras con contenido dentro del árbol sintáctico del término multipalabra (*pares de dependencia sintáctica*); dichos pares quedan asociados al patrón correspondiente a ese árbol. Las dependencias contempladas se corresponden con aquéllas descritas en el apartado 7.2.3:

[2] A lo largo de este capítulo haremos uso frecuente de los operadores clásicos de definición de expresiones regulares: **?** expresará opcionalidad, + indicará repetición (1 o más veces), ***** indicará repetición con opcionalidad (0 o más veces), y | separará alternativas.

[3] A lo largo del capítulo —por ejemplo, a la hora de construir las variantes— se emplearán también conjunciones coordinantes (representadas por Cc), subordinantes (Cs) y signos de puntuación (Q).

1. *Sustantivo-modificador* (adjetivo o complemento nominal): asumiendo la *distributividad* de dicha relación de dependencia [130, 24, 168] —consideramos que cada modificador modifica de forma independientemente al término modificado—, se obtendrá un par de dependencia por cada uno de los núcleos de los modificadores y cada uno de los núcleos de sus modificados. Por ejemplo, de *coches y camiones rojos* obtendremos los pares (*coche, rojo*) y (*camión, rojo*).
2. *Sujeto-verbo*: el par principal será el formado por el núcleo del sujeto y el verbo. Por ejemplo, de *los perros comen carne* obtendremos el par (*perro, comer*).
3. *Verbo-objeto*: el par principal será en este caso el formado por el verbo y el sustantivo núcleo del objeto o complemento verbal. Por ejemplo, de *recortar gastos* obtendremos el par (*recortar, gasto*).

El conjunto de patrones resultantes empleados por el sistema, así como sus dependencias asociadas, están recogidos en el apéndice E.

- Finalmente, nuestro analizador sintáctico superficial emplea dichas expresiones regulares para generar una lista de los pares de palabras ligadas por dependencias sintácticas presentes en el texto a analizar, para su posterior normalización e indexación. De este modo, estamos identificando los pares de dependencia mediante una simple correspondencia de patrones sobre la salida del etiquetador-lematizador, abordando el problema desde una aproximación de procesamiento superficial a nivel léxico, lo que conlleva una considerable reducción del coste de ejecución. Por otra parte, el empleo de técnicas de estado finito para la implementación del analizador nos permite analizar el texto en un tiempo lineal respecto a su longitud.

7.3.1. Ejemplo de Extracción mediante Patrones

Para ilustrar el funcionamiento del sistema durante el proceso de extracción, incluimos en este apartado un pequeño ejemplo. Partiremos de los términos recogidos en las figuras 7.1 y 7.2, que comprenden, respectivamente, una serie de variantes sintácticas y morfosintácticas del grupo nominal *una caída de las ventas*. En el caso de las variantes sintácticas, éstas han sido obtenidas aplicando sucesivamente los diferentes mecanismos de variación posibles. Para cada variante se indican los mecanismos sintácticos y morfosintácticos aplicados para su obtención, además de mostrar su estructura básica, las dependencias sintácticas que contiene (indicadas mediante líneas punteadas), y los pares de dependencia a los que éstas dan lugar.

En concreto examinaremos el caso del término original —*una caída de las ventas*—, la variante sintáctica final obtenida —*unas inusuales caídas de ventas y de beneficios*— y su variante morfosintáctica —*las ventas han caído*. La información común a los tres términos corresponde a la relación establecida entre los conceptos *caída/caer* y *venta*. El objetivo es el de extraer y normalizar adecuadamente dicha información para cada uno de los diferentes términos de forma que se posibilite el establecimiento de correspondencias entre los mismos.

En la figura 7.3 mostramos las correspondencias de los diferentes términos, tanto el original como sus variantes, con los patrones que capturarán las dependencias de nuestro interés —descritos en el apéndice E. Concretamente, los patrones empleados son el patrón `CN01` (sustantivo modificado por un complemento nominal) para el término original, el `CN02` (sustantivo modificado por una coordinación de complementos nominales) para su variante sintáctica, y el `SV01` (dependencia sujeto–verbo) para su variante morfosintáctica.

Para cada uno de los términos se muestran, además, los pares extraídos de acuerdo con las dependencias asociadas a los patrones empleados (líneas punteadas). Estos pares son:

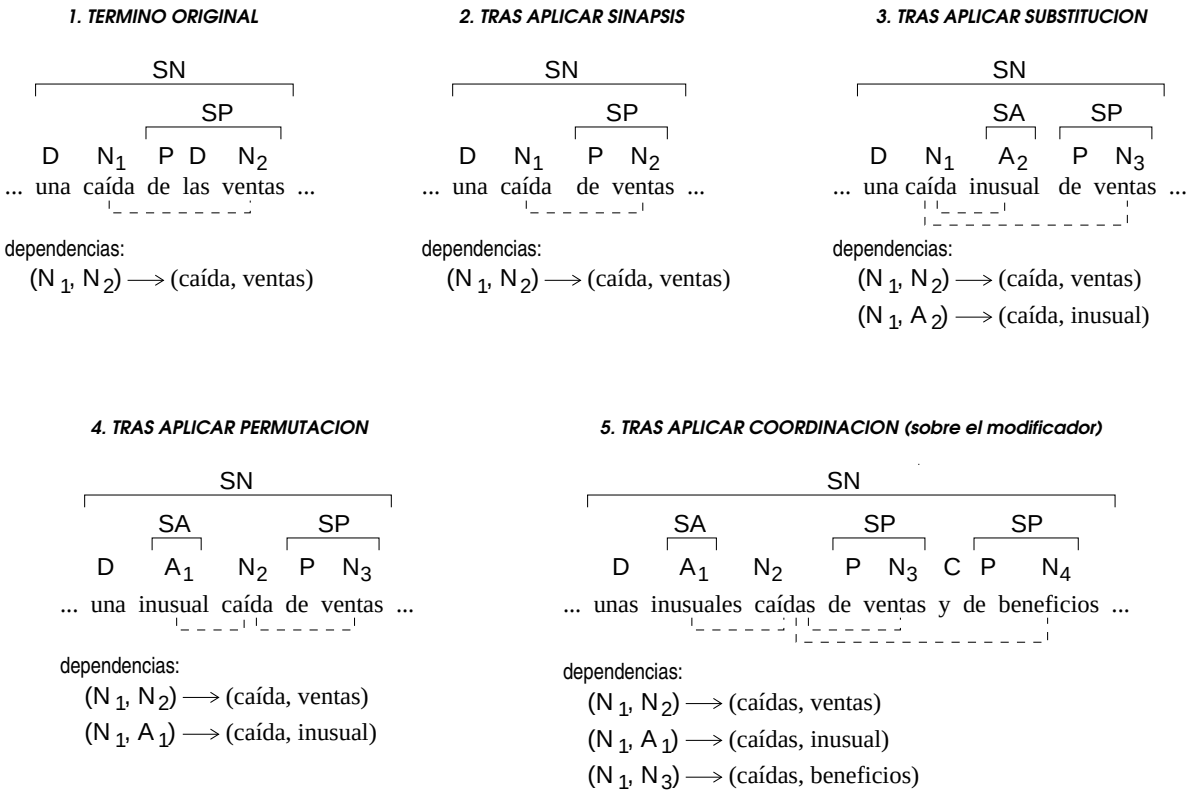


Figura 7.1: Variantes sintácticas de *una caída de las ventas*

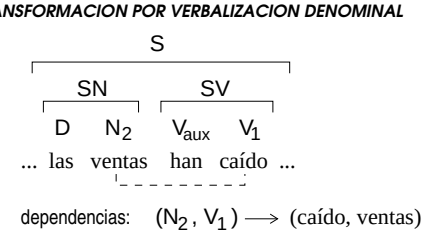


Figura 7.2: Variante morfosintáctica de *una caída de las ventas*

- En el término original: *(caída,ventas)*.
- En la variante sintáctica: *(caídas, ventas)* y *(caídas, beneficios)*.
- En la variante morfosintáctica: *(caído, ventas)*.

Los pares de interés para este ejemplo son aquéllos que hacen referencia al mismo concepto, es decir, *(caída, ventas)*, *(caídas, ventas)* y *(caído, ventas)*, que deben ser todavía normalizados.

Una vez extraídos, los términos componentes del par son primero lematizados para eliminar la flexión asociada a variantes sintácticas y morfosintácticas. De este modo, obtenemos ya un mismo término *(caída, venta)* para el término original y su variante sintáctica. En el caso de la variante morfosintáctica el término obtenido es *(caer, venta)*, por lo que todavía no sería posible establecer una correspondencia entre ambos.

En un segundo paso, los lemas obtenidos para los términos componentes del par son normalizados aplicando familias morfológicas. Sea $r_f(l)$ la función que, dado un lema l, devuelve el representante correspondiente a su familia morfológica. Dado que *caer* y *caída*

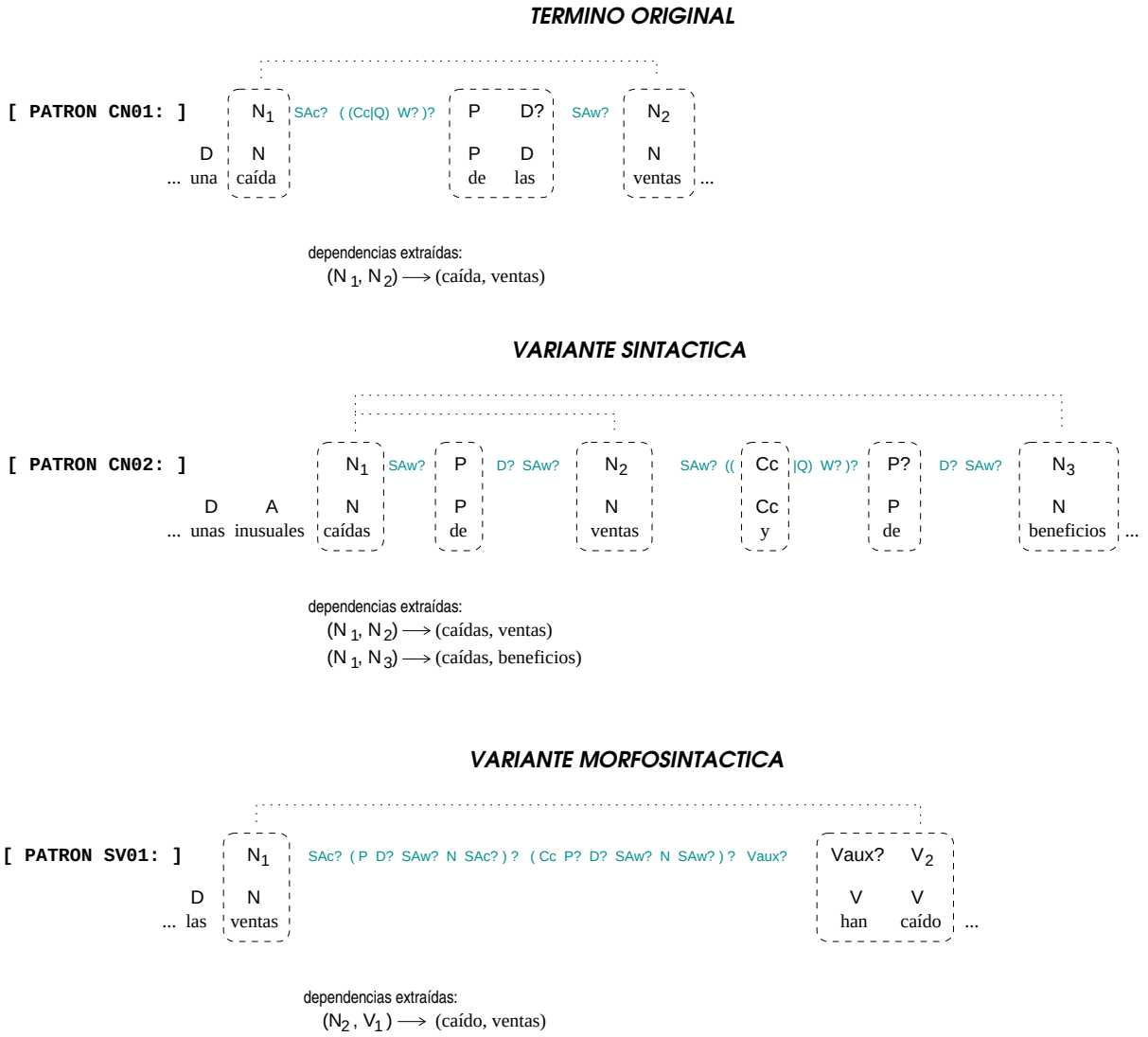


Figura 7.3: Ejemplo de extracción de dependencias mediante patrones

pertenecen a la misma familia morfológica, tienen el mismo representante, al que denotaremos como $repr_{\{caer,\,caída\}}$ siendo $repr_{\{caer,\,caída\}} = r_f(caer) = r_f(caída)$. De esta forma los tres pares iniciales diferentes obtenidos son normalizados a un mismo término, el par $(repr_{\{caer,\,caída\}}, r_f(venta))$, con lo que habríamos conseguido el objetivo deseado.

7.4. Análisis Sintáctico Superficial mediante Cascadas de Traductores Finitos: el Analizador CASCADE

Dadas las limitaciones del analizador PATTERNS en cuanto a su rigidez, difícil mantenimiento e incapacidad para establecer dependencias lejanas correspondientes a dependencias verbales (sujetos y objetos/complementos), se decidió desarrollar un analizador más avanzado, modular, robusto, y de aplicación más general. Dicho analizador, denominado CASCADE, fue desarrollado con un doble propósito, su integración, por una parte, dentro de un sistema de Recuperación de Información —el caso que nos ocupa— y, por otra, para su empleo en tareas de Extracción de Información.

De acuerdo con estas necesidades se optó por desarrollar un analizador sintáctico superficial basado en cascadas de traductores finitos. La base de este nuevo modelo de analizador parte de la Teoría de Lenguajes Formales [104] según la cual, dada una gramática independiente del contexto y una cadena de entrada, los subárboles sintácticos de altura k generados por un analizador sintáctico pueden ser recreados mediante k capas de traductores finitos: la primera capa obtiene los nodos etiquetados por no terminales correspondiente a la parte izquierda de producciones que sólo contienen terminales en su parte derecha; la segunda capa obtiene aquellos nodos que involucran símbolos terminales y aquellos símbolos no-terminales generados en la primera capa; y así sucesivamente. Evidentemente, la acotación en la altura de los árboles limita el tipo de construcciones sintácticas que se puede reconocer. Sin embargo, otras aproximaciones similares para el análisis [11, 10] han sido empleadas con éxito en otros campos próximos del Procesamiento automático del Lenguaje Natural, como es el caso de la Extracción de Información [23, 96, 103]. Su aplicación en el ámbito de la Recuperación de Información no ha sido todavía explorada en profundidad, si bien ha sido ya ensayada por XEROX en el TREC para el caso del inglés [106], demostrando su superioridad respecto a aproximaciones clásicas basadas en pares de palabras contiguas.

7.4.1. Arquitectura del Sistema

De acuerdo con lo anteriormente expuesto, hemos desarrollado un analizador sintáctico superficial basado en una arquitectura de cinco capas. donde la entrada al analizador es proporcionada por el etiquetador-lematizador descrito en el capítulo 5.

A continuación, describiremos el funcionamiento de cada una de las capas. Para ello utilizaremos como notación reglas independientes del contexto en las cuales permitimos el empleo de los operadores clásicos de definición de expresiones regulares. Por otra parte, los identificadores con mayúscula referenciarán conjuntos de términos, tanto preterminales (las etiquetas resultantes del proceso de etiquetación léxica) como elementos de una categoría gramatical dada. Cuando se requiera la presencia de un lema en concreto, éste se indicará empleando la fuente `typewriter`.

Capa 0: Ampliación del Preprocesado

El objetivo perseguido en el procesamiento llevado a cabo en esta capa es doble: por una parte, mejorar las capacidades del preprocesador y etiquetador-lematizador originales y, por otra, eliminar el ruido generado por ciertas construcciones sintácticas en las etapas de análisis posteriores. Para ello, la salida obtenida tras la acción conjunta del preprocesador lingüístico y el etiquetador-lematizador es procesada de diversas formas:

- *Tratamiento de cifras en formato no numérico*. Para evitar interferencias en el reconocimiento de estructuras sintácticas en las capas superiores, es preciso identificar aquellas secuencias de palabras que se corresponden con cifras escritas parcial o totalmente en formato no numérico, como por ejemplo *cinco millones* y *5 millones*. Esta tarea no resulta tan sencilla como aparenta ya que surgen ambigüedades ligadas a la aparición de la conjunción coordinada *y*. Muchos de estos casos se pueden resolver diseñando una gramática para los numerales, permiténdonos determinar que formas como *cuatro y cinco* o *veintisiete y ocho* no forman una sola cifra. No obstante, todavía quedarían algunos casos, más complejos, sin resolver. Este sería el caso de la frase *vendí cuarenta y cinco quedaron sin vender*, para la cual sería preciso realizar un análisis sintáctico completo de la oración para determinar que la conjunción no forma parte de la cifra *cuarenta y cinco*, sino que está coordinando dos cláusulas.

Afortunadamente, la extracción de cifras en los sistemas de Recuperación de Información no es tan importante como en el caso de los sistemas de Extracción de Información, ya que las consultas expresadas en lenguaje natural raramente involucran términos numéricos[4], siendo suficiente aplicar heurísticas que resuelvan los casos más frecuentes. Este sería el caso, por ejemplo, de las secuencias *entre* ⟨*cifra*⟩ *y* ⟨*cifra*⟩, para las cuales podemos asegurar con bastante seguridad que estamos ante la coordinación de dos cifras diferentes, aunque en este caso también se podrían plantean problemas con excepciones del tipo *entre cuatro y cinco mil*, pues las cifras involucradas no son realmente *cuatro* y *cinco mil* sino *cuatro mil* y *cinco mil*.

- *Tratamiento de expresiones de cantidad*. Nos referimos a expresiones del tipo *algo más de dos millones* o *unas dos docenas*, que si bien se refieren a una cifra, establecen una cierta vaguedad en relación al valor de la misma. Para tratar tales expresiones ha sido necesario definir una serie de reglas. El primer grupo de dichas reglas define el conjunto de modificadores y *numerales colectivos*[5] permitidos:

$$\begin{aligned} PreModifCant &\rightarrow \texttt{casi} \\ PreModifCant &\rightarrow \texttt{cerca de} \\ PreModifCant &\rightarrow (\texttt{poco} \mid \texttt{algo})?\ (\texttt{más} \mid \texttt{menos})\ \texttt{de} \end{aligned}$$

$$NumeralColectivo \rightarrow \texttt{decena} \mid \texttt{docena} \mid \texttt{centenar} \mid \texttt{millar} \mid \ldots$$

Partiendo de estas reglas auxiliares, podemos ya definir expresiones de cantidad del tipo *poco más de cinco millones*, formadas por un modificador opcional, una cifra, un numeral colectivo y una preposición *de* que sólo será necesario si la expresión tiene función de determinante, pues si funciona como pronombre no acompañará a ningún sustantivo. Para ello emplearemos la regla:

$$SNum \rightarrow PreModifCant?\ Cifra\ NumeralColectivo\ \texttt{de}?$$

La siguiente regla nos permite identificar expresiones que denotan una cantidad aproximada de números colectivos, como por ejemplo *cientos de miles de*. Aunque en teoría es posible iterar indefinidamente esta construcción, dando lugar a expresiones del tipo *cientos de miles de millones de billones*, tales formas raramente son empleadas precisamente por su complejidad. Hemos optado, pues, por acotarlas:

$$\begin{aligned} SNum \rightarrow\ & PreModifCant?\ (\texttt{medio} \mid \texttt{un})?\ (NumeralColectivo\ \texttt{de})? \\ & NumeralColectivo\ (\texttt{y medio})?\ \texttt{de}? \end{aligned}$$

La última regla nos permite identificar aquellas expresiones que no contienen numerales colectivos, como por ejemplo *poco más de trecientos treinta*:

$$SNum \rightarrow PreModifCant\ Cifra$$

- *Simplificación de expresiones verbales*. Ciertas expresiones verbales de difícil normalización [130] son consideradas como una unidad para así simplificar el trabajo de las capas superiores. De este modo, la expresión *tener en cuenta*, por ejemplo, debe considerarse una unidad, sinónima del verbo *considerar*, para de este modo evitar que capas posteriores identifiquen *en cuenta* como complemento del verbo, lo que a su vez podría

[4] Las consultas del tipo "folletos de cámaras que cuesten menos de sesenta y cinco euros" son más adecuadas para sistemas de gestión de bases de datos estructuradas que para sistemas de RI, hasta el momento incapaces de resolver este tipo de consultas sobre textos generales.

[5] Sustantivos que representan como unidad un número determinado de elementos; p.ej., *docena = 12 elementos*.

impedir la correcta identificación de otros complementos verbales de interés. En la versión actual del sistema, el criterio que hemos seguido a la hora de identificar estas expresiones es el de considerar sólo aquellas secuencias Verbo-Preposición-Sustantivo especialmente frecuentes, que sean fácilmente sustituibles por un sinónimo verbal o perifrástico y en las que no se pierda información al realizar dicha sustitución. En concreto, hemos considerado las siguientes expresiones:

poner en práctica	=	*aplicar*	*dejar en libertad*	=	*liberar*
tener en cuenta	=	*considerar*	*poner en libertad*	=	*liberar*
entrar en contacto	=	*contactar*	*poner de manifiesto*	=	*mostrar*
estar en contacto	=	*contactar*	*poner de relieve*	=	*mostrar*
poner en contacto	=	*contactar*	*estar en condición*	=	*poder*
poner en marcha	=	*iniciar*	*llevar a cabo*	=	*realizar*
estar a punto	=	*ir a*			

Capa 1: Sintagmas Adverbiales y Grupos Verbales de Primer Nivel

Esta primera capa de análisis propiamente dicho consta de reglas que contienen únicamente etiquetas y/o lemas en su parte derecha. Con el fin de que las capas siguientes puedan extraer pares de dependencias, nos interesa poder asociar al no terminal del lado izquierdo de cada regla el lema correspondiente al núcleo del sintagma que se esté reconociendo, así como una etiqueta con los rasgos morfosintácticos pertinentes. La notación que utilizaremos para este mecanismo de herencia está inspirada en la empleada para la especificación del conjunto de restricciones en las gramáticas basadas en estructuras de rasgos [49].

La primera de las reglas empleadas por el analizador nos permite identificar *sintagmas adverbiales* ($SAdv$) formados por secuencias de adverbios (W). Aunque no van a participar en la formación de los pares de dependencia, es necesario identificar este tipo de sintagmas para que las etapas posteriores puedan funcionar correctamente. Consideraremos que el último adverbio constituye el núcleo del sintagma, por lo que su lema y su etiqueta constituirán el lema y la etiqueta del no terminal $SAdv$:

$$SAdv \rightarrow W^* \; W_1 \quad \begin{cases} SAdv.lem \doteq W_1.lem \\ SAdv.etiq \doteq W_1.etiq \end{cases}$$

Existen también expresiones tales como *de forma rápida*, con un adjetivo (A) como núcleo, que sin embargo equivalen a un adverbio, en este caso *rápidamente*. Estas construcciones son procesadas por la regla:

$$SAdv \rightarrow \texttt{de} \; (\texttt{forma} \mid \texttt{manera} \mid \texttt{modo}) \; A \quad \begin{cases} SAdv.lem \doteq A.lem \\ SAdv.etiq \doteq A.etiq \end{cases}$$

El siguiente conjunto de reglas nos permite identificar *grupos verbales de primer nivel* o no perifrásticos ($GV1$) correspondientes a formas pasivas, tanto simples —p.ej., *soy observado*— como compuestas —p.ej., *he sido observado*. La primera de estas reglas se encarga de procesar las formas compuestas: la etiqueta se toma del verbo (V) auxiliar `haber`, mientras que el lema se toma del verbo principal, que debe ser un participio, al igual que el verbo auxiliar `ser`. La segunda regla trata las formas simples: la etiqueta se obtiene de la forma del verbo auxiliar `ser`, mientras que el lema se toma del verbo principal, de nuevo un participio.

$$GV1 \rightarrow V_1 \; V_2 \; V_3 \quad \begin{cases} GV1.lem \doteq V_3.lem \\ GV1.etiq \doteq V_1.etiq \\ GV1.voz \doteq \text{PAS} \\ V_1.lem \doteq \texttt{haber} \\ V_2.lem \doteq \texttt{ser} \\ V_2.tiempo \doteq \text{PART} \\ V_3.tiempo \doteq \text{PART} \end{cases}$$

$$GV1 \rightarrow V_1 \ V_2 \begin{cases} GV1.lem \doteq V_2.lem \\ GV1.etiq \doteq V_1.etiq \\ GV1.voz \doteq \text{PAS} \\ V_1.lem \doteq \texttt{ser} \\ V_2.tiempo \doteq \text{PART} \end{cases}$$

Las formas activas, compuestas y simples respectivamente, son identificadas de forma similar mediante las siguientes reglas:

$$GV1 \rightarrow V_1 \ V_2 \begin{cases} GV1.lem \doteq V_2.lem \\ GV1.etiq \doteq V_1.etiq \\ GV1.voz \doteq \text{ACT} \\ V_1.lem \doteq \texttt{haber} \\ V_2.tiempo \doteq \text{PART} \end{cases}$$

$$GV1 \rightarrow V_1 \begin{cases} GV1.lem \doteq V_1.lem \\ GV1.etiq \doteq V_1.etiq \\ GV1.voz \doteq \text{ACT} \end{cases}$$

Es interesante observar que, para garantizar la correcta identificación de los tiempos verbales en caso de que puedan aplicarse varias reglas, tendrá preferencia la que presente una parte derecha más larga (*longest matching*).

Capa 2: Sintagmas Adjetivales y Grupos Verbales de Segundo Nivel

Los sintagmas adjetivales ($SAdj$) como *azul* o *muy alto* son tratados en esta capa, al igual que los *grupos verbales de segundo nivel* ($GV2$) como *tengo que ir*.

Definimos un *sintagma adjetival* ($SAdj$) como aquél cuyo núcleo es un adjetivo (A), pudiendo ser modificado por un sintagma adverbial previo:

$$SAdj \rightarrow SAdv? \ A \begin{cases} SAdj.lem \doteq A.lem \\ SAdj.etiq \doteq A.etiq \end{cases}$$

En lo que respecta a los *grupos verbales de segundo nivel* ($GV2$), éstos incluyen los grupos verbales perifrásticos. Las *perífrasis verbales* son uniones de dos o más formas verbales que funcionan como una unidad, dotando a la semántica del verbo principal de matices de significado tales como obligación, grado de desarrollo de la acción, etc., que no pueden ser expresados mediante las formas simples o compuestas del verbo. En lo concerniente a su estructura, las perífrasis están formadas generalmente por un verbo auxiliar conjugado que aporta la flexión, un verbo en forma no personal (infinitivo, gerundio o participio) que aporta el significado principal, y un elemento opcional (preposición o conjunción) de enlace entre ambos.

Las denominadas *perífrasis de infinitivo* se identifican mediante la siguiente regla, en la que también se contempla la posibilidad de que los verbos vayan seguidos de un pronombre enclítico (separado previamente de la forma verbal por el etiquetador-lematizador) en el caso de que el verbo auxiliar sea reflexivo. Por cuestiones prácticas, sólo se han considerado los nexos más comunes. La etiqueta se hereda del verbo auxiliar, mientras que el lema y la voz se heredan del verbo principal:

$$GV2 \rightarrow GV1_1 \ (\texttt{me} \mid \texttt{te} \mid \texttt{se})? \ (\texttt{que} \mid \texttt{de} \mid \texttt{a})? \ GV1_2 \begin{cases} GV2.lem \doteq GV1_2.lem \\ GV2.etiq \doteq GV1_1.etiq \\ GV2.voz \doteq GV1_2.voz \\ GV1_1.voz \doteq \text{ACT} \\ GV1_2.tiempo \doteq \text{INF} \end{cases}$$

Las *perífrasis de gerundio* se tratan de modo similar, si bien en este caso no se permite la presencia de nexos:

$$GV2 \rightarrow GV1_1 \ (\texttt{me} \mid \texttt{te} \mid \texttt{se})? \ GV1_2 \quad \begin{cases} GV2.lem \doteq GV1_2.lem \\ GV2.etiq \doteq GV1_1.etiq \\ GV2.voz \doteq GV1_2.voz \\ GV1_1.voz \doteq \text{ACT} \\ GV1_2.tiempo \doteq \text{GER} \end{cases}$$

De forma análoga se procede con las *perífrasis de participio*:

$$GV2 \rightarrow GV1_1 \ (\texttt{me} \mid \texttt{te} \mid \texttt{se})? \ GV1_2 \quad \begin{cases} GV2.lem \doteq GV1_2.lem \\ GV2.etiq \doteq GV1_1.etiq \\ GV2.voz \doteq GV1_2.voz \\ GV1_1.voz \doteq \text{ACT} \\ GV1_2.tiempo \doteq \text{PART} \end{cases}$$

Por último, los grupos verbales de primer nivel que no forman parte de grupos perifrásticos, son promocionados a grupos verbales de segundo nivel:

$$GV2 \rightarrow GV1 \quad \begin{cases} GV2.lem \doteq GV1.lem \\ GV2.etiq \doteq GV1.etiq \\ GV2.voz \doteq GV1.voz \end{cases}$$

Capa 3: Sintagmas Nominales

En esta capa son procesados los *sintagmas nominales* (SN), aquéllos cuyo núcleo es un elemento de función sustantiva. En la definición de las reglas para su identificación y procesado se ha contemplado la posibilidad de que vengan precedidos de un *complemento partitivo*[6]. Una vez más, por cuestiones prácticas sólo hemos contemplado los determinantes y pronombres indefinidos más frecuentes, dando lugar a la siguiente regla para la identificación de dichos complementos. Nos interesará guardar el número del partitivo correspondiente, tal y como se explica a continuación:

$$CPartitivo \rightarrow (\quad \begin{array}{l} \texttt{algún} \mid \texttt{alguno} \mid \\ \texttt{cualquier} \mid \texttt{cualquiera} \mid \\ \texttt{ningún} \mid \texttt{ninguno} \mid \\ \texttt{mucho} \mid \texttt{uno} \ )_1 \ \texttt{de} \end{array} \quad \{ \ CPartitivo.num \doteq ()_1.num$$

Tras el núcleo del sintagma nominal puede aparecer un modificador en forma de dos sintagmas adjetivales unidos por una conjunción coordinada (representada por Cc), o bien una secuencia de uno, dos y hasta tres sintagmas adjetivales:

$$\begin{aligned} PostModifSAdj &\rightarrow SAdj \ \ Cc \ \ Sadj \\ PostModifSAdj &\rightarrow SAdj \\ PostModifSAdj &\rightarrow SAdj \ \ SAdj \\ PostModifSAdj &\rightarrow SAdj \ \ SAdj \ \ SAdj \end{aligned}$$

El núcleo del sintagma nominal estará formado por un nombre común (representado por N), una sigla o un nombre propio; su etiqueta y lema determinarán la etiqueta y lema del sintagma completo. En el caso de la aparición sucesiva de varios candidatos a núcleo, consideraremos

[6] Expresiones de cuantificación del tipo *alguno de*, *ninguno de*, etc.

que el último es el que realiza esta función. Por otra parte, la etiqueta establecida para el sintagma puede verse modificada en presencia de un complemento partitivo, ya que en este caso, y de cara a establecer concordancias con otros sintagmas, el número del sintagma nominal se corresponderá con el aportado por dicho partitivo. Por ejemplo, se debe decir *"Cualquiera de ellos lo sabe"*, y no **"Cualquiera de ellos lo saben"*.

Opcionalmente, antes del núcleo pueden aparecer uno o más determinantes (D) y un sintagma adjetival. La aparición de modificadores adjetivales tras el núcleo es también opcional, dando lugar, finalmente, a la regla:

$$\begin{array}{lll} SN & \rightarrow & CPartitivo? \\ & & D^* \ \ (SAdj \mid Cifra \mid SNum)? \\ & & (N \mid Sigla \mid Propio\ SNNucleo_1)^* \\ & & (N \mid Sigla \mid Propio\ SNNucleo_1)_1 \\ & & PostModifSAdj? \end{array} \quad \left\{ \begin{array}{l} SN.lem \doteq ()_1.lem \\ SN.etiq \doteq ()_1.etiq \\ SN.num \doteq CPartitivo.num \end{array} \right.$$

Capa 4: Sintagmas Preposicionales

Por último, la capa 4 se encarga de la identificación de los *sintagmas preposicionales* (SP, $SPde$, $SPpor$), aquéllos formados por un sintagma nominal (SN) precedido de una preposición (P). Con objeto de facilitar la extracción de dependencias, nos interesará distinguir del resto los sintagmas que comienzan por las preposiciones *de* y *por*, dando lugar a las siguientes reglas:

$$SPde \rightarrow P \ \ SN \quad \left\{ \begin{array}{l} P.lem \doteq \texttt{de} \\ SP.lem \doteq SN.lem \\ SP.etiq \doteq SN.etiq \end{array} \right.$$

$$SPpor \rightarrow P \ \ SN \quad \left\{ \begin{array}{l} P.lem \doteq \texttt{por} \\ SP.lem \doteq SN.lem \\ SP.etiq \doteq SN.etiq \end{array} \right.$$

$$SP \rightarrow P \ \ SN \quad \left\{ \begin{array}{l} SP.lem \doteq SN.lem \\ SP.etiq \doteq SN.etiq \end{array} \right.$$

7.4.2. Extracción de Dependencias Sintácticas

Como hemos apuntado anteriormente, el objetivo final del análisis sintáctico del texto es la extracción de pares de palabras ligadas por relaciones de dependencia sintáctica. En el caso de CASCADE, dicho proceso se desarrolla en dos fases: una primera fase de *identificación de funciones sintácticas* de los sintagmas identificados durante el proceso de análisis, y una segunda fase de *extracción de dependencias* propiamente dicha.

Fase de Identificación de Funciones Sintácticas

En esta primera etapa, y debido a la superficialidad del análisis realizado, nos tenemos que enfrentar a diversas limitaciones, entre las que destaca el problema de establecer los límites de cada oración. Tomando como hipótesis de trabajo que para cada núcleo verbal existe una oración asociada, y en base a consideraciones prácticas, consideraremos que existe un fin de oración cuando se alcance uno de los siguientes elementos:

1. *Signos de puntuación*: sin limitarse únicamente al punto de cierre de las oraciones, pues tampoco podremos retomar con garantías el análisis de una frase cuando éste sea interrumpido, por ejemplo, por un inciso entre comas o una cita entre comillas.

2. *Relativos*: ya que separan una cláusula subordinada de la oración principal.

3. *Conjunciones*: debido a las ambigüedades sintácticas que introducen respecto a cuáles son las partes de la oración que conectan.

4. *Grupos verbales de segundo nivel (GV2) cuyo núcleo es una forma personal* cuando no existe ningún otro límite de oración entre dicho grupo verbal y el anterior, ya que por norma general la aparición de un verbo implica que estamos ante una nueva oración.

Esta limitación del ámbito en el cual trabaja el extractor de dependencias no representa un gran obstáculo en el contexto de la extracción de términos índice para un sistema de RI, ya que lo que se persigue no es tanto la exhaustividad como la fiabilidad de las dependencias obtenidas.

En última instancia, lo que se persigue es identificar oraciones que sigan alguno de los siguientes criterios de formación:

- Sujeto activo + grupo verbal predicativo activo + complemento directo.
- Sujeto activo + grupo verbal copulativo + atributo.
- Sujeto pasivo + grupo verbal predicativo pasivo + complemento agente.

Evidentemente, tales estructuras ideales raramente aparecen en estado puro, ya que lo habitual es encontrarse con sintagmas diversos entre el sujeto y el grupo verbal y entre el grupo verbal y sus complementos.

Las funciones sintácticas identificadas, junto con los criterios empleados para ello, son las siguientes:

Complemento nominal. Debido a la ambigüedad en la adjunción de sintagmas preposicionales en cuanto a si nos encontramos realmente ante un complemento nominal o bien ante un complemento verbal, sólo hemos considerado el caso de los sintagmas preposicionales introducidos por *de*, los $SPde$, por ser altamente fiables. En consecuencia, cuando nos encontremos con un $SPde$ que siga inmediatamente a un sintagma nominal o preposicional, aquél será etiquetado como complemento nominal.

Sujeto. El sintagma nominal (SN) más próximo que antecede a un grupo verbal ($GV2$) se toma como su sujeto. Adicionalmente, consideraremos que carecen de sujeto aquellos grupos verbales cuyo núcleo es una forma no personal (infinitivo, gerundio o participio).

Atributo. En presencia de un verbo copulativo, identificaremos como atributo aquel $SAdj$ no ligado, núcleo del SN o $SPde$ más próximo que sigue al grupo verbal.

Objeto directo. El SN más próximo que aparece después de un $GV2$ predicativo activo se considera su complemento directo.

Agente. El $SPpor$ más próximo que sigue a un $GV2$ predicativo pasivo es considerado su complemento agente.

Complemento circunstancial. Debido al citado problema de adjunción de sintagmas preposicionales, se ha optado por seguir un criterio conservador a la hora de identificar los complementos circunstanciales del verbo. El objetivo es minimizar el ruido introducido por complementos incorrectamente identificados. Por ello, consideraremos como complemento circunstancial sólo a aquel sintagma preposicional posterior al verbo, más próximo a él, y anterior a todo complemento verbal o atributo previamente identificado.

Fase de Extracción de Dependencias

Una vez identificadas las funciones sintácticas de los sintagmas, la siguiente fase consiste en la *extracción de las dependencias sintácticas* existentes entre éstos. Para ello, se crean los pares de dependencia formados por:

- Un sustantivo y cada uno de los adjetivos que lo modifican. Debemos llamar la atención sobre el hecho de que si bien el resto de las dependencias son extraídas tras finalizar el proceso de análisis, estas dependencias se extraen realmente en la fase de identificación de sintagmas nominales de la capa 3. Esto se debe a que son dependencias internas al sintagma nominal y, de no extraerlas entonces, dicha información se perdería una vez el sintagma es reducido a su núcleo.
- Un sustantivo y el núcleo de su complemento nominal.
- El núcleo del sujeto y el verbo predicativo.
- El núcleo del sujeto y el del atributo. Los verbos atributivos se consideran meros elementos copulativos, por lo que la dependencia se establece directamente entre el sujeto y el atributo.
- Un verbo activo y el núcleo de su objeto directo.
- Un verbo pasivo y el núcleo de su complemento agente.
- Un verbo predicativo y el núcleo de su complemento circunstancial.
- El núcleo del sujeto y el del complemento circunstancial de su verbo, únicamente en caso de que el verbo sea atributivo, dado su especial comportamiento.

Para cada dependencia extraída se almacena su tipo, así como los lemas y etiquetas de sus componentes, sin normalización alguna salvo la propia lematización. De este modo se mantiene la generalidad, ya que a partir de esta información inicial se pueden aplicar fácilmente diferentes esquemas de normalización, obteniendo así diferentes tipos de términos índice según nuestras necesidades. En este caso se ha aplicado de nuevo el esquema de normalización basado en familias morfológicas que hemos venido aplicando hasta ahora.

7.4.3. Implementación del Analizador Sintáctico

La base del analizador consiste en construir un traductor de estado finito a partir de cada una de las reglas involucradas en las diferentes etapas del proceso de análisis. A diferencia de lo que ocurre en otras aplicaciones como son la Extracción de Información [103] o la extracción de patrones léxicos [94], nuestro objetivo no es el de obtener a la salida de todo el proceso una versión parentizada del texto de entrada, con los paréntesis identificando los sintagmas, sino una lista, en forma de pares, de las dependencias del texto. Puesto que en la construcción de dichos pares sólo intervienen los núcleos de los sintagmas, para cada sintagma identificado sólo nos interesa preservar el lema de dicho núcleo, junto con los rasgos morfosintácticos que le corresponden. Nuestro analizador se comporta, pues, como un *filtro de estado finito* y no como un *marcador de estado finito* [92]. Al generalizar esta forma de proceder a todas las etapas, un texto estará formado en todo momento por ternas *lema-etiqueta-no terminal*. Para mantener la uniformidad a la hora de la inicialización del sistema, consideraremos que toda categoría gramatical constituye un no terminal válido, lo que nos permite considerar el texto de entrada a la primera capa como formado por ternas *lema-etiqueta-categoría gramatical*. Para ilustrarlo,

tomemos el siguiente ejemplo:

Docenas de niños muy alegres han tenido que aprender hoy en el colegio una lección de historia.

La salida inicial del etiquetador son ternas *forma*-etiqueta-lema:[7]

```
[docenas NCFP docena] [de X de] [niños NCMP niño] [muy WQ muy]
[alegres AQFP alegre] [han V3PRI haber] [tenido VPMS tener]
[que C que] [aprender VRI aprender] [hoy WI hoy] [en X en]
[el DAMS el] [colegio NCMS colegio] [una DAFS un]
[lección NCFS lección] [de X de] [historia NCFS historia]
```

la cual es transformada al formato de entrada requerido por el analizador, *lema*-etiqueta-no terminal. Recordemos que en esta fase inicial del análisis, la categoría gramatical del término es la que constituye dicho no terminal:

```
[docena NCFP N] [de P P] [niño NCMP N] [muy WQ W]
[alegre AQFP A] [haber V3PRI V] [tener VPMS V] [que Cs Cs]
[aprender VRI V] [hoy WI W] [en P P] [el DAMS DA]
[colegio NCMS N] [un DAFS DA] [lección NCFS N] [de P P]
[historia NCFS N]
```

Una ventaja de utilizar esta representación es que nos permite referenciar en la parte derecha de las reglas a cualesquiera de los componentes: lemas, etiquetas y no terminales.

El primer paso a la hora de definir el traductor de estado finito asociado a una regla determinada consiste en adaptar la parte derecha de las producciones al sistema elegido para la representación de los textos. Tomemos como ejemplo una regla sencilla, como es el caso de la identificación de sintagmas adverbiales ($SAdv$):

$$SAdv \;\rightarrow\; W^* \;\; W_1 \;\; \begin{cases} SAdv.lem \doteq W_1.lem \\ SAdv.etiq \doteq W_1.etiq \end{cases}$$

Dicha regla es transformada en[8]

$$\langle\rangle_1\square\langle\rangle_2\square SAdv\blacksquare \;\rightarrow\; (\alpha^+\square\alpha^+\square W\blacksquare)^* \;\; (\langle\alpha^+\rangle_1\square\langle\alpha^+\rangle_2\square W\blacksquare)$$

Con esta regla le estamos indicando al analizador que un sintagma adverbial está formado por una secuencia de 0 o más entradas correspondientes a adverbios, seguidas por una única entrada perteneciente al adverbio que actúa como núcleo sintagmático y del cual heredará lema y etiqueta. Cada una de las ternas de entrada está separada de las demás por un tabulador (■), y a su vez sus campos están separados entre sí por espacios en blanco (□). Los dos primeros campos, ambos formados por cadenas alfanuméricas (α^+), representan, recordemos, el lema y la etiqueta morfosintáctica asociados al núcleo del no terminal indicado por el tercer campo, inicialmente constituido por la categoría gramatical de la palabra. Puesto que exigimos que las entradas pertenezcan a adverbios, obligamos a que sus no terminales sean, pues, adverbios (denotados

[7] Adviértase que tanto en éste como en los siguientes ejemplos, los corchetes son un añadido del autor durante la redacción de la memoria para facilitar la lectura de los mismos.

[8] Extendemos la notación empleada hasta el momento con α, que representa un carácter alfanumérico, □, que denota el carácter de espaciado, y ■, que indica el carácter de tabulación.

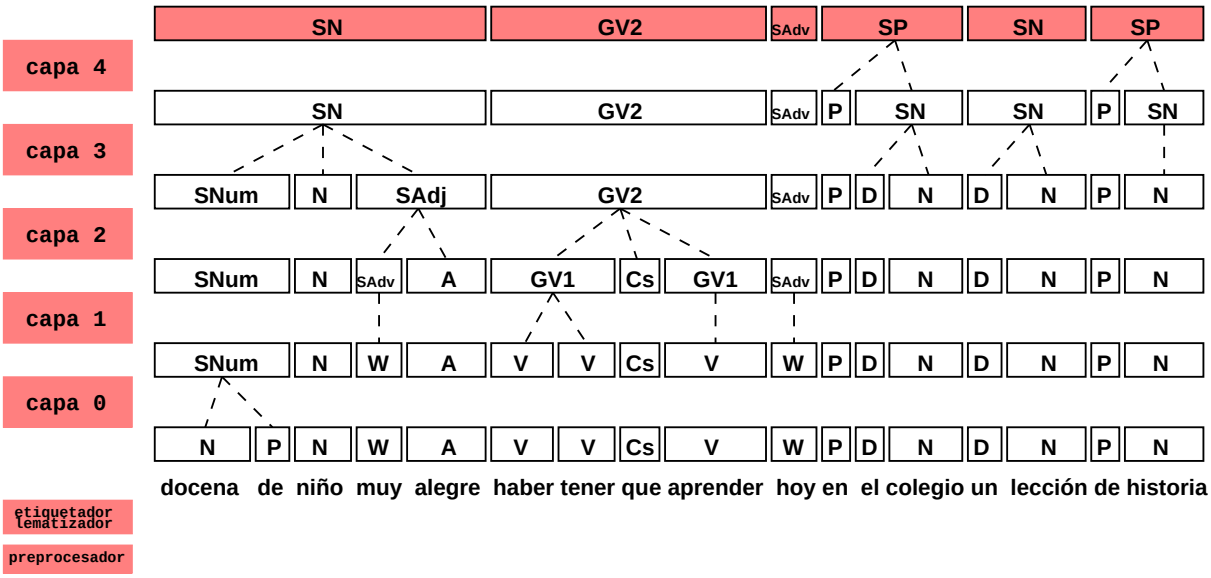


Figura 7.4: Resumen del proceso de análisis para el ejemplo de ejecución

mediante W). Una vez el sistema encuentra una correspondencia de la parte derecha de la regla sobre las ternas de entrada, la regla es reducida, sustituyendo las ternas de la correspondencia por la terna indicada en la parte izquierda de la regla. Dado que, para esta regla, el sintagma hereda el lema y la etiqueta de su núcleo, le indicamos al analizador que los campos correspondientes a las correspondencias $\langle\rangle_1$ y $\langle\rangle_2$ pasarán a ser, respectivamente, el lema y la etiqueta asociados a la terna del no terminal $SAdv$. Por lo tanto, si aplicamos dicha regla sobre el fragmento del ejemplo:

$$\cdots \texttt{\textit{hoy}}\ \texttt{WI}\ W\ \cdots$$

se obtendría a la salida:

$$\cdots \texttt{\textit{hoy}}\ \texttt{WI}\ SAdv\ \cdots$$

donde, como se puede apreciar, el sistema ha identificado una secuencia de adverbios, en este caso de longitud 1, reduciéndola a un $SAdv$ con núcleo y etiqueta los núcleo y etiqueta del último —y único— adverbio de la secuencia, *hoy* y WI, respectivamente.

7.4.4. Ejemplo Detallado de Ejecución

Para ilustrar mejor el funcionamiento del analizador y de sus reglas, retomaremos nuestro ejemplo y mostraremos la evolución de su análisis a través de las diferentes capas. Dicho análisis se muestra, de forma resumida, en la figura 7.4.

Capa 0. La aplicación de las reglas descritas en el apartado 7.4.1 nos permite completar el preprocesado de la entrada de cara al análisis. De este modo, la secuencia de ternas `[`*docena* `NCFP` N`]` `[`*de* `P` P`]`, correspondiente a la expresión *docenas de*, es reducida al sintagma numeral `[`*docena*`&`*de*[9] `Cifra` $SNum$`]`:

[9]El carácter `&` es empleado a la hora de concatenar palabras.

[*docena&de* Cifra $SNum$] [*niño* NCMP N] [*muy* WQ W]
[*alegre* AQFP A] [*haber* V3PRI V] [*tener* VPMS V] [*que* Cs Cs]
[*aprender* VRI V] [*hoy* WI W] [*en* P P] [*el* DAMS DA]
[*colegio* NCMS N] [*un* DAFS DA] [*lección* NCFS N] [*de* P P]
[*historia* NCFS N]

Capa 1: De acuerdo a las reglas del apartado 7.4.1, identificamos los sintagmas adverbiales ($SAdv$) *muy*, definido por la terna de entrada [*muy* WQ W], y *hoy*, definido por la terna de entrada [*hoy* WI W]. Una vez reducidos, se obtienen las ternas de salida [*muy* WQ $SAdv$] y [*hoy* WI $SAdv$], respectivamente. Al mismo tiempo son identificados dos grupos verbales de primer nivel ($GV1$). El primero de ellos, *han tenido*, dado por [*haber* V3PRI V] [*tener* VPMS V], es reducido a [*tener* V3PRI $GV1$], mientras que el segundo, *aprender*, dado por la entrada [*aprender* VRI V], es transformado en [*aprender* VRI $GV1$]:

[*docena&de* Cifra $SNum$] [*niño* NCMP N] [*muy* WQ $SAdv$]
[*alegre* AQFP A] [*tener* V3PRI $GV1$] [*que* Cs Cs]
[*aprender* VRI $GV1$] [*hoy* WI $SAdv$] [*en* P P] [*el* DAMS DA]
[*colegio* NCMS N] [*un* DAFS DA] [*lección* NCFS N] [*de* P P]
[*historia* NCFS N]

Capa 2: Los sintagmas adjetivales ($SAdj$) y grupos verbales de segundo nivel ($GV2$) del texto son procesados mediante las reglas contempladas en el apartado 7.4.1. El único sintagma adjetival de nuestro ejemplo es *muy alegres*, dado por [*muy* WQ $SAdv$] [*alegre* AQFP A] —recordemos que *muy* fue reducido a un $SAdv$ en la capa anterior—, y que a su vez es reducido a [*alegre* AQFP $SAdj$]. En cuanto a los grupos verbales $GV2$, el texto únicamente contiene el grupo perifrástico *han tenido que aprender*, el cual, tras haber sido procesados en la capa anterior los $GV1$ que lo integran, ha quedado reducido a la secuencia de ternas [*tener* V3PRI $GV1$] [*que* Cs Cs] [*aprender* VRI $GV1$]. Una vez procesado, se obtiene a la salida la terna [*aprender* V3PRI $GV2$]:

[*docena&de* Cifra $SNum$] [*niño* NCMP N] [*alegre* AQFP $SAdj$]
[*aprender* V3PRI $GV2$] [*hoy* WI $SAdv$] [*en* P P] [*el* DAMS DA]
[*colegio* NCMS N] [*un* DAFS DA] [*lección* NCFS N] [*de* P P]
[*historia* NCFS N]

Capa 3: En esta fase son procesados los sintagmas nominales del texto (SN), de acuerdo con las reglas presentadas en el apartado 7.4.1; paralelamente, se extraen las dependencias sintácticas internas a los mismos. El primero de los sintagmas encontrados es *docenas de niños muy alegres*, cuyos componentes han sido repetidamente reducidos hasta [*niño* NCMP N] [*alegre* AQFP $SAdj$], que a su vez son ahora transformados en [*niño* NCMP SN]. Los otros tres sintagmas identificados son más sencillos: *el colegio* —[*el* DAMS DA] [*colegio* NCMS N]—, *una lección* —[*un* DAFS DA] [*lección* NCFS N]— e *historia* —[*historia* NCFS N]. Su reducción a sintagmas produce: [*colegio* NCMS SN], [*lección* NCFS SN] e [*historia* NCFS SN]:

[*niño* NCMP SN] [*aprender* V3PRI $GV2$] [*hoy* WI $SAdv$] [*en* P P]
[*colegio* NCMS SN] [*lección* NCFS SN] [*de* P P]
[*historia* NCFS SN]

De los cuatro sintagmas nominales identificados, únicamente *docenas de niños muy alegres*

contiene dependencias *sustantivo-adjetivo* (SA) a extraer:[10]

SA (*niño* NCMP , *alegre* AQFP)

Capa 4: El último paso consiste en identificar los sintagmas preposicionales (SP, $SPde$ y $SPpor$) del texto, en este caso dos, en base a las reglas del apartado 7.4.1. El primero de ellos es el *SP en el colegio*, representado en este punto por la secuencia [*en* P P] [*colegio* NCMS SN], y transformado a continuación en [*colegio* NCMS SP]. El segundo sintagma identificado es el *SPde de historia*, [*de* P P] [*historia* NCFS SN], que es reducido a [*historia* NCFS $SPde$]:

[*niño* NCMP SN] [*aprender* V3PRI $GV2$] [*hoy* WI $SAdv$]
[*colegio* NCMS SP] [*lección* NCFS SN] [*historia* NCFS $SPde$]

Una vez finalizada la fase de análisis sintáctico propiamente dicha, se procede a la extracción de las dependencias contenidas en el texto. Para ello primero identificaremos las funciones sintácticas de los sintagmas obtenidos en el análisis y a continuación extraeremos sus dependencias asociadas. Debemos recordar, de nuevo, que las dependencias internas a los sintagmas nominales, aquéllas entre su núcleo nominal y sus adjetivos modificadores, constituyen una excepción a este proceso pues son ya extraídas en la capa 3, en el momento de identificar los sintagmas nominales.

Por lo tanto, el primer paso será identificar las funciones sintácticas de los sintagmas resultantes del análisis, de acuerdo con los criterios establecidos en la Sección 7.4.2. Mostramos a continuación, sobre la salida del analizador, las funciones sintácticas de los sintagmas identificados, pudiendo comprobar que hemos detectado un sujeto activo ($SUJact$), un grupo verbal predicativo activo ($Vact$), su complemento circunstacial (CC), su objeto directo (OD), y el complemento nominal de éste último (CN):

[*niño*	NCMP	SN]	--	⟨ $SUJact$ ⟩
[*aprender*	V3PRI	$GV2$]	--	⟨ $Vact$ ⟩
[*hoy*	WI	$SAdv$]	--	⟨ ⟩
[*colegio*	NCMS	SP]	--	⟨ CC ⟩
[*lección*	NCFS	SN]	--	⟨ OD ⟩
[*historia*	NCFS	$SPde$]	--	⟨ CN ⟩

Una vez identificadas las funciones sintácticas de cada sintagma, extraemos sus dependencias asociadas, obteniendo finalmente como resultado:

SA	(*niño*	NCMP , *alegre*	AQFP)
SUJ_A	(*aprender*	V3PRI , *niño*	NCMP)
OD	(*aprender*	V3PRI , *lección*	NCFS)
CN	(*lección*	NCFS , *historia*	NCFS)
CC	(*aprender*	V3PRI , *colegio*	NCMS)

Las pruebas realizadas con el analizador mostraron unos buenos resultados a la hora de identificar la existencia de posibles dependencias sintácticas entre dos términos, si bien no siempre dichas dependencias eran correctamente clasificadas. Este es el caso, por ejemplo, de sujetos postpuestos al verbo, los cuales eran identificados por las heurísticas como objetos o

[10] El tipo de la dependencia junto con los lemas y etiquetas de sus componentes, principal y modificador, se muestran mediante la notación:

tipo (*lema-núcleo* etiqueta-núcleo , *lema-modificador* etiqueta-modificador)

df	*#lemas*	*%lemas*	*#pares*	*%pares*
[1..1]	216403	51.76 %	3922172	56.95 %
[2..2]	56524	13.52 %	1036809	15.06 %
[3..4]	42330	10.12 %	747078	10.85 %
[5..8]	30345	7.26 %	495083	7.19 %
[9..16]	21026	5.03 %	307203	4.46 %
[17..32]	14652	3.50 %	182384	2.65 %
[33..∞)	36805	8.81 %	195746	2.84 %
Total	418085	100.00 %	6886475	100.00 %

Tabla 7.1: Distribución por frecuencia de documento (*df*) de lemas y pares de dependencias —obtenidos empleando el analizador PATTERNS— en la colección CLEF 2003

complementos del verbo en lugar de como sujetos. Sin embargo, el analizador sí identificaba la existencia de una dependencia entre ambos términos, dependencia que era extraída y normalizada. Este tipo de errores en la identificación de la función sintáctica del sintagma no suponen un problema para el proceso de IR, ya que a la hora de normalizar el par para su posterior indexación, la información correspondiente al tipo de dependencia no es necesaria y se desecha, por lo que sólo resulta de interés el hecho de que la dependencia existente haya sido identificada.

7.4.5. Indexación de los Términos Extraídos

El empleo de términos complejos debe verse como complemento a los términos simples, ya que el empleo exclusivo de términos complejos a modo de términos índice —al igual que en el caso del empleo exclusivo de términos simples—, permite capturar únicamente una vista parcial e insuficiente de la semántica del texto [161].

Por otra parte, de emplear únicamente términos complejos, la cobertura del sistema se vería reducida considerablemente debido al alto grado de dispersión de su espacio de términos. Esto quiere decir que el número de pares de dependencia existentes en una colección es mucho mayor que el número palabras que ésta contiene, ya que dado un conjunto de palabras, el número de frases que se pueden formar a partir de ellas es mucho mayor que el número de palabras que forman el conjunto. A modo de ejemplo la tabla 7.4.5 recoge la distribución de lemas y pares de dependencia en la colección CLEF 2003. También por la misma razón, es mucho menos frecuente la repetición de una frase en dos documentos que la repetición de las palabras que la conforman, por lo que la probabilidad de que se produzca una correspondencia durante el proceso de recuperación es mucho menor. Por ejemplo, si en un documento aparece la frase *"merendé chocolate"*, y la consulta se refiere a *"comer chocolate"*, el empleo de términos simples permitiría una correspondencia parcial entre ambas vía el término *"chocolate"*. Sin embargo, si se emplean términos multipalabra —en este caso pares de dependencia—, no habría correspondencia, ya que *(merendar, chocolate)* y *(comer, chocolate)* son pares diferentes.

Por estas razones los términos complejos son utilizados en combinación con los términos simples [168, 161, 106, 230, 42]. En nuestro caso, los pares de dependencia sintáctica serán empleados como términos índice en combinación con los lemas de las palabras con contenido del texto —capítulo 5. Sin embargo, dicho empleo combinado de términos simples y complejos plantea a su vez una serie de problemas.

El primero de ellos es que el empleo conjunto de ambos tipos de términos supone una violación de la suposición de independencia, puesto que las palabras que integran el par de dependencia

sintáctica han sido también indexadas como términos simples [168].

Por otra parte, existe una sobreponderación de los pesos de los términos complejos, mucho menos frecuentes que los términos simples, debido a su alto grado de dispersión, y por tanto, con un peso asignado mucho mayor [230].

Esto se traduce en una creciente inestabilidad del sistema, en tanto que al producirse correspondencias no deseadas de términos complejos con documentos no relevantes, su puntuación asignada aumenta considerablemente, disparando su nivel de relevancia. Al mismo tiempo, cuando se producen correspondencias de términos complejos con documentos sí relevantes se produce una clara mejora de los resultados respecto al empleo de términos simples. De acuerdo con esto podría argumentarse que deberían esperarse unos resultados similares a los obtenidos para los términos simples. Sin embargo, esto no es así, ya que las correspondencias de términos complejos son mucho menos frecuentes que las de términos simples —debido de nuevo a su alto grado de dispersión—, por lo que las correspondencias casuales de términos complejos, de producirse, son muchísimo más perjudiciales que para el caso de términos simples, cuyo efecto tiende a compensarse debido a la influencia de las restantes correspondencias. Podemos afirmar, pues, que el ruido introducido en el sistema por falsas correspondencias se ve amplificado. Debemos, por tanto, tratar de corregir esa sobrevaloración de los términos índice complejos, y así minimizar el efecto negativo de las correspondencias no deseadas.

La solución a ambos problemas pasa por disminuir el peso relativo de los términos complejos respecto a los simples mediante un factor de ponderación [168, 106].

7.5. Resultados Experimentales con Información Sintáctica Extraída de las Consultas

A continuación se recogen los resultados obtenidos para la evaluación experimental del empleo de pares de dependencia sintáctica a modo de términos índice complejos como complemento de los términos índice simples. En este apartado mostramos una primera serie de experimentos que emplea la información sintáctica extraída de las consultas, mientras que en el apartado 7.6 se muestra una segunda serie de experimentos que utiliza la información sintáctica extraída de los documentos.

7.5.1. Resultados para Sintagmas Nominales y Verbales

En este primer conjunto de resultados emplearemos de forma combinada términos simples lematizados y términos complejos obtenidos a partir de las consultas. Ambos, documentos y consultas, son normalizados simultáneamente mediante lematización por una parte, y mediante análisis sintáctico superficial por otra. Los términos obtenidos en el caso de los documentos, tanto simples como complejos, son indexados de forma combinada de acuerdo con lo expuesto en el apartado 7.4.5. En el caso de las consultas se procede de forma similar, combinando los lemas y pares de dependencia extraídos.

El peso de los términos simples es multiplicado por un factor de ponderación ω para de esta forma disminuir la contribución relativa de los términos complejos en un *ratio* $1/\omega$ dado. El factor de ponderación ω a emplear se estima previamente a partir del corpus de entrenamiento CLEF 2001-02·A. Los resultados obtenidos durante esta fase de puesta a punto se encuentran recogidos en el apéndice F. Los factores de ponderación ω considerados —con sus *ratios* correspondientes indicados entre paréntesis— son:

$$\omega \in \{1\ \textit{(1)},\ 2\ \textit{(0.500)},\ 3\ \textit{(0.333)},\ 4\ \textit{(0.250)},\ 5\ \textit{(0.200)},\ 8\ \textit{(0.125)},\ 10\ \textit{(0.100)},\\ 12\ \textit{(0.083)},\ 14\ \textit{(0.071)},\ 16\ \textit{(0.062)},\ 18\ \textit{(0.055)},\ 20\ \textit{(0.050)}\}$$

Se han realizado dos series de experimentos en este apartado: la primera de ellas para los pares de dependencia obtenidos mediante nuestro analizador basado en patrones PATTERNS —apartado 7.3—, y la segunda para los pares obtenidos mediante nuestro analizador en cascada CASCADE —apartado 7.4. Comentaremos en primer lugar los resultados obtenidos mediante el análisis basado en patrones empleando el analizador PATTERNS.

Tras la fase de puesta a punto del factor de ponderación ω —tablas F.1 a F.4 del apéndice F— se estimó, tanto para consultas cortas como para largas, un valor ω=8 (*ratio* $1/\omega$=0.125) como factor óptimo. Los resultados generales conseguidos empleando los pares generados por PATTERNS (*FNF*), así como la mejora obtenida respecto a la normalización mediante lematización de términos simples (*lem*), se recogen en la tabla 7.2.

Los resultados conseguidos son ciertamente positivos, ya que existe una clara mejora, a todos los niveles, respecto a los resultados obtenidos empleando únicamente términos simples. Por otra parte, nuestros resultados son cualitativamente mejores que aquéllos obtenidos por [163] para el inglés o por [132] para el holandés, ya que no sólo se produce una mejora de la precisión para los primeros documentos devueltos, como ocurría también en el caso de los trabajos citados, sino que se trata de una mejora general, que afecta incluso a las precisiones globales.

Los resultados obtenidos a nivel de consulta se recogen en las gráficas 7.5, 7.6 y 7.7. En esta ocasión, dado que nuestra meta es la de mejorar las precisiones de los primeros documentos devueltos, las gráficas de comparativa de rendimiento no se han generado, como hasta ahora, para la precisión no interpolada, sino para la precisión a los 10 documentos devueltos[11]. Como se puede ver, los casos de mejora superan tanto en cuantía como en magnitud a aquéllos casos para los cuales se produce un descenso de la precisión, confirmando los resultados globales anteriormente comentados.

En lo que respecta a los resultados obtenidos empleando la información sintáctica extraída por el analizador CASCADE, en primer lugar se procedió a calcular su correspondiente factor de ponderación ω —tablas F.5 a F.8 del apéndice F. Si bien en términos estrictos el factor óptimo era ω=4 (*ratio* $1/\omega$=0.250), se decidió seguir empleando un factor ω=8 (*ratio* $1/\omega$=0.125) para facilitar la comparación con la aproximación basada en patrones, dado que la diferencias de resultados eran escasas.

Los resultados obtenidos para el analizador CASCADE se recogen en la tabla 7.3 y en las gráficas 7.8, 7.9 y 7.10. Estos resultados siguen siendo satisfactorios, si bien el comportamiento del corpus CLEF 2003 es bastante irregular, particularmente en el caso de las consultas cortas.

Además de las comparativas de resultados, también se recogen en las figuras 7.11 y 7.12 las distribuciones por frecuencia de documento (*df*) de los términos complejos obtenidos para la colección CLEF 2003 mediante ambos analizadores. En ambas se incluyen el número y porcentaje de pares totales para cada rango de frecuencia, así como el total de pares indexados —aquéllos con una frecuencia de documento mayor o igual que 5. Las cifras muestran que las diferencias entre ambas distribuciones son ínfimas, y que sólo una pequeña parte de los pares —un 17 %— son indexados, fruto del alto grado de dispersión del espacio de términos. Sin embargo, el número de pares distintos generados por CASCADE es notablemente inferior, un 43 % menor, lo que repercute notablemente en el tamaño del índice generado. Esta amplia diferencia cuantitativa se debe a la permisividad de PATTERNS en el caso de las dependencias *sustantivo-complemento nominal*. CASCADE es sumamente conservador en este aspecto, permitiendo la adjunción de un sintagma preposicional al sustantivo de su izquierda únicamente en el caso de la preposición *de*. La razón para ello viene dada por la imposibilidad de nuestro sistema de desambiguar con un margen de confianza suficiente la estructura de dependencias en el caso de sintagmas preposicionales introducidos por otra preposición diferente a la *de* —ver apartado 7.4.2. Por su

[11] Se tomaron 10 documentos pues éste suele ser el número de documentos devuelto por un buscador en su primera página de resultados, aquélla a la que suele restringirse el usuario.

corpus	*CLEF 2001-02 A*						*CLEF 2001-02 B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2700	**2738**	**1.41**	2762	**2784**	**0.80**	2363	**2367**	**0.17**	2410	**2416**	**0.25**	2159	**2173**	**0.65**	2175	**2199**	**1.10**
Pr. no int.	.4829	**.4949**	**2.48**	.5239	**.5356**	**2.23**	.4678	**.4861**	**3.91**	.5256	**.5454**	**3.77**	.4324	**.4464**	**3.24**	.4659	**.4763**	**2.23**
Pr. doc.	.5327	**.5439**	**2.10**	.5690	**.5807**	**2.06**	.5667	**.5777**	**1.94**	.6089	**.6265**	**2.89**	.4724	**.4861**	**2.90**	.5201	**.5307**	**2.04**
R-pr.	.4848	**.4851**	**0.06**	.5075	**.5118**	**0.85**	.4549	**.4675**	**2.77**	.5030	**.5206**	**3.50**	.4362	**.4608**	**5.64**	.4493	**.4715**	**4.94**
Pr. a 0 %	.8293	**.8406**	**1.36**	.8845	**.8870**	**0.28**	.8234	.8155	-0.96	.8763	**.8768**	**0.06**	.8124	**.8597**	**5.82**	.8366	**.8499**	**1.59**
Pr. a 10 %	.7463	**.7612**	**2.00**	.7914	**.8102**	**2.38**	.6845	**.7201**	**5.20**	.7517	**.7836**	**4.24**	.7057	**.7249**	**2.72**	.7197	**.7455**	**3.58**
Pr. a 20 %	.6771	**.6885**	**1.68**	.7136	**.7496**	**5.04**	.6287	**.6643**	**5.66**	.7095	**.7260**	**2.33**	.6118	**.6368**	**4.09**	.6453	**.6717**	**4.09**
Pr. a 30 %	.6138	**.6386**	**4.04**	.6591	**.6683**	**1.40**	.5829	**.6145**	**5.42**	.6463	**.6599**	**2.10**	.5503	**.5585**	**1.49**	.5788	**.6011**	**3.85**
Pr. a 40 %	.5502	**.5590**	**1.60**	.6085	**.6122**	**0.61**	.5555	**.5589**	**0.61**	.6100	**.6265**	**2.70**	.5059	**.5067**	**0.16**	.5348	**.5465**	**2.19**
Pr. a 50 %	.4931	**.5040**	**2.21**	.5557	**.5605**	**0.86**	.5136	**.5247**	**2.16**	.5658	**.5841**	**3.23**	.4657	**.4668**	**0.24**	.4889	**.5053**	**3.35**
Pr. a 60 %	.4496	**.4638**	**3.16**	.5006	**.5116**	**2.20**	.4413	**.4675**	**5.94**	.4977	**.5219**	**4.86**	.4017	**.4163**	**3.63**	.4339	**.4391**	**1.20**
Pr. a 70 %	.3853	**.3856**	**0.08**	.4161	**.4300**	**3.34**	.3943	**.4043**	**2.54**	.4569	**.4638**	**1.51**	.3422	.3380	-1.23	.3789	**.3959**	**4.49**
Pr. a 80 %	.3277	**.3333**	**1.71**	.3509	**.3670**	**4.59**	.3122	**.3287**	**5.29**	.3679	**.3832**	**4.16**	.2700	**.2783**	**3.07**	.3206	**.3241**	**1.09**
Pr. a 90 %	.2356	**.2484**	**5.43**	.2492	**.2606**	**4.57**	.2319	**.2338**	**0.82**	.2845	**.2993**	**5.20**	.1821	**.1997**	**9.67**	.2067	**.2137**	**3.39**
Pr. a 100 %	.1197	**.1213**	**1.34**	.1289	**.1349**	**4.65**	.1209	**.1251**	**3.47**	.1547	**.1617**	**4.52**	.0932	**.1013**	**8.69**	.1164	**.1179**	**1.29**
Pr. a 5 docs.	.6609	**.6783**	**2.63**	.6957	**.7217**	**3.74**	.5956	**.6356**	**6.72**	.6844	**.7111**	**3.90**	.5915	.5872	-0.73	.6213	.6213	0.00
Pr. a 10 docs.	.6283	**.6391**	**1.72**	.6543	**.6674**	**2.00**	.5667	**.5889**	**3.92**	.6000	**.6311**	**5.18**	.5149	**.5404**	**4.95**	.5596	**.5745**	**2.66**
Pr. a 15 docs.	.5928	.5913	-0.25	.6188	**.6246**	**0.94**	.5170	**.5481**	**6.02**	.5689	**.5926**	**4.17**	.4738	**.4865**	**2.68**	.5106	**.5234**	**2.51**
Pr. a 20 docs.	.5446	**.5587**	**2.59**	.5880	**.5967**	**1.48**	.4878	**.5100**	**4.55**	.5422	**.5544**	**2.25**	.4457	**.4553**	**2.15**	.4819	**.4851**	**0.66**
Pr. a 30 docs.	.4928	**.5051**	**2.50**	.5304	**.5391**	**1.64**	.4452	**.4578**	**2.83**	.4948	**.5000**	**1.05**	.4000	**.4113**	**2.82**	.4255	**.4333**	**1.83**
Pr. a 100 docs.	.3300	**.3317**	**0.52**	.3509	.3489	-0.57	.2993	**.3056**	**2.10**	.3147	**.3216**	**2.19**	.2513	**.2560**	**1.87**	.2702	**.2736**	**1.26**
Pr. a 200 docs.	.2234	**.2260**	**1.16**	.2315	**.2343**	**1.21**	.1997	**.2014**	**0.85**	.2093	**.2127**	**1.62**	.1617	**.1637**	**1.24**	.1694	**.1707**	**0.77**
Pr. a 500 docs.	.1090	**.1106**	**1.47**	.1115	**.1132**	**1.52**	.0983	**.0993**	**1.02**	.1023	**.1032**	**0.88**	.0827	**.0840**	**1.57**	.0842	**.0851**	**1.07**
Pr. a 1000 docs.	.0587	**.0595**	**1.36**	.0600	**.0605**	**0.83**	.0525	**.0526**	**0.19**	.0536	**.0537**	**0.19**	.0459	**.0462**	**0.65**	.0463	**.0468**	**1.08**

Tabla 7.2: Resultados obtenidos mediante lematización (*lem*), caso base, y pares de dependencia sintáctica obtenidos mediante el analizador Patterns a partir de la consulta (*FNF*)

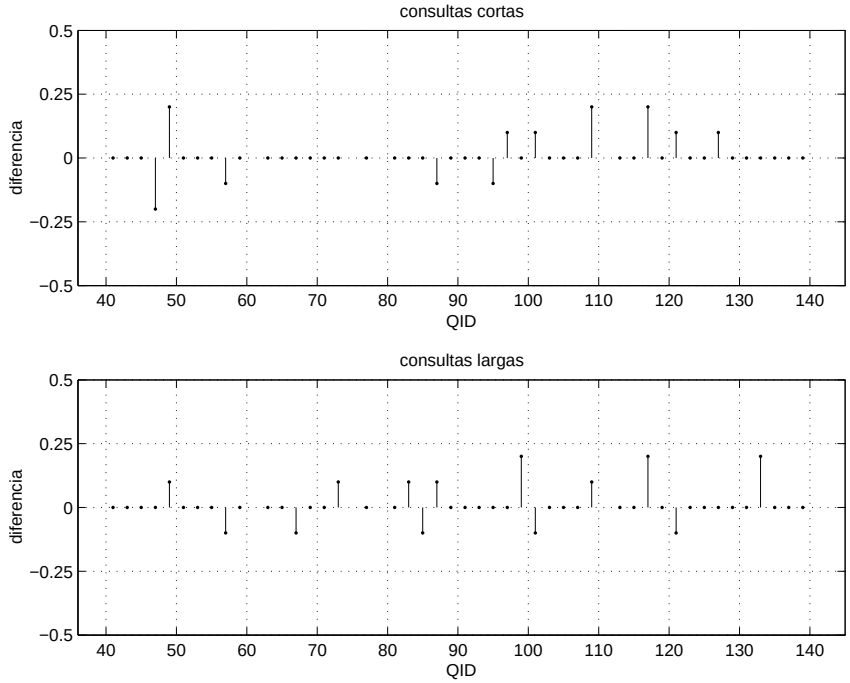


Figura 7.5: Diferencias en las precisiones a los 10 documentos: lematización vs. pares de dependencia sintáctica obtenidos mediante el analizador PATTERNS a partir de la consulta. Corpus CLEF 2001-02·A

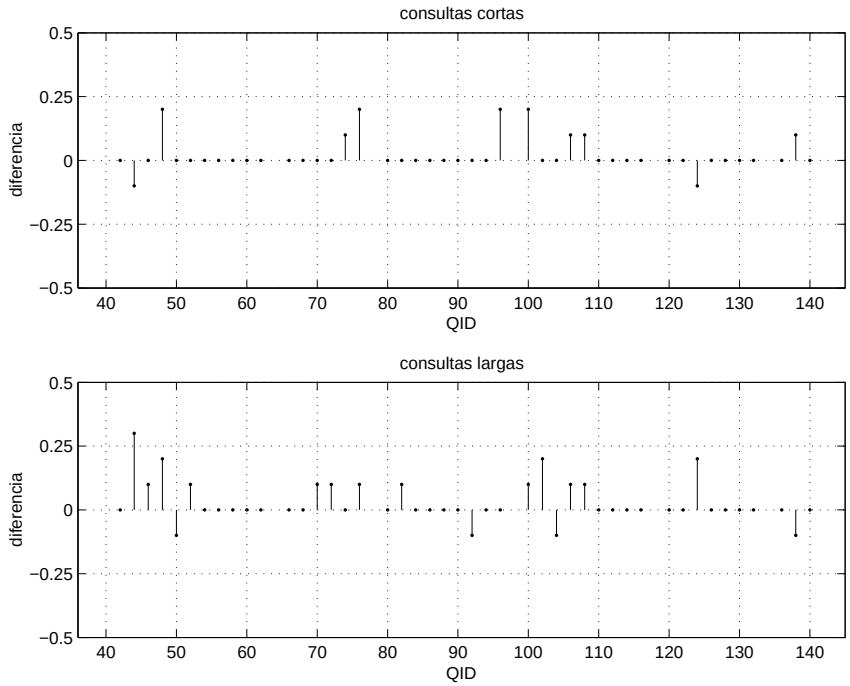


Figura 7.6: Diferencias en las precisiones a los 10 documentos: lematización vs. pares de dependencia sintáctica obtenidos mediante el analizador PATTERNS a partir de la consulta. Corpus CLEF 2001-02·B

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46000	46000	–	46000	46000	–	45000	45000	–	45000	45000	–	47000	47000	–	47000	47000	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2700	**2728**	**1.04**	2762	**2786**	**0.87**	2363	**2371**	**0.34**	2410	**2415**	**0.21**	2159	**2169**	**0.46**	2175	**2198**	**1.06**
Pr. no int.	.4829	**.4965**	**2.82**	.5239	**.5339**	**1.91**	.4678	**.4810**	**2.82**	.5256	**.5401**	**2.76**	.4324	**.4377**	**1.23**	.4659	**.4700**	**0.88**
Pr. doc.	.5327	**.5491**	**3.08**	.5690	**.5826**	**2.39**	.5667	**.5728**	**1.08**	.6089	**.6213**	**2.04**	.4724	**.4738**	**0.30**	.5201	**.5253**	**1.00**
R-pr.	.4848	**.4895**	**0.97**	.5075	**.5185**	**2.17**	.4549	**.4672**	**2.70**	.5030	**.5133**	**2.05**	.4362	**.4490**	**2.93**	.4493	**.4550**	**1.27**
Pr. a 0 %	.8293	**.8406**	**1.36**	.8845	.8685	-1.81	.8234	.8137	-1.18	.8763	.8674	-1.02	.8124	**.8409**	**3.51**	.8366	.8342	-0.29
Pr. a 10 %	.7463	**.7640**	**2.37**	.7914	**.8067**	**1.93**	.6845	**.7128**	**4.13**	.7517	**.7789**	**3.62**	.7057	.6993	-0.91	.7197	**.7354**	**2.18**
Pr. a 20 %	.6771	**.6925**	**2.27**	.7136	**.7331**	**2.73**	.6287	**.6608**	**5.11**	.7095	**.7262**	**2.35**	.6118	**.6364**	**4.02**	.6453	**.6702**	**3.86**
Pr. a 30 %	.6138	**.6338**	**3.26**	.6591	**.6728**	**2.08**	.5829	**.6137**	**5.28**	.6463	**.6586**	**1.90**	.5503	**.5566**	**1.14**	.5788	**.5923**	**2.33**
Pr. a 40 %	.5502	**.5622**	**2.18**	.6085	**.6138**	**0.87**	.5555	**.5575**	**0.36**	.6100	**.6204**	**1.70**	.5059	.4992	-1.32	.5348	.5243	-1.96
Pr. a 50 %	.4931	**.5076**	**2.94**	.5557	**.5655**	**1.76**	.5136	**.5232**	**1.87**	.5658	**.5787**	**2.28**	.4657	.4625	-0.69	.4889	.4874	-0.31
Pr. a 60 %	.4496	**.4648**	**3.38**	.5006	**.5162**	**3.12**	.4413	**.4552**	**3.15**	.4977	**.5124**	**2.95**	.4017	**.4103**	**2.14**	.4339	**.4353**	**0.32**
Pr. a 70 %	.3853	**.3909**	**1.45**	.4161	**.4292**	**3.15**	.3943	**.4002**	**1.50**	.4569	**.4641**	**1.58**	.3422	.3385	-1.08	.3789	**.3927**	**3.64**
Pr. a 80 %	.3277	**.3364**	**2.65**	.3509	**.3632**	**3.51**	.3122	**.3280**	**5.06**	.3679	**.3845**	**4.51**	.2700	**.2758**	**2.15**	.3206	**.3214**	**0.25**
Pr. a 90 %	.2356	**.2483**	**5.39**	.2492	**.2581**	**3.57**	.2319	**.2352**	**1.42**	.2845	**.2981**	**4.78**	.1821	**.1913**	**5.05**	.2067	**.2096**	**1.40**
Pr. a 100 %	.1197	.1197	0.00	.1289	**.1318**	**2.25**	.1209	**.1242**	**2.73**	.1547	**.1572**	**1.62**	.0932	**.0992**	**6.44**	.1164	**.1182**	**1.55**
Pr. a 5 docs.	.6609	**.6913**	**4.60**	.6957	**.7043**	**1.24**	.5956	**.6178**	**3.73**	.6844	**.6933**	**1.30**	.5915	.5745	-2.87	.6213	**.6340**	**2.04**
Pr. a 10 docs.	.6283	**.6500**	**3.45**	.6543	**.6783**	**3.67**	.5667	**.5822**	**2.74**	.6000	**.6178**	**2.97**	.5149	.5128	-0.41	.5596	.5574	-0.39
Pr. a 15 docs.	.5928	**.6029**	**1.70**	.6188	**.6319**	**2.12**	.5170	**.5333**	**3.15**	.5689	**.5822**	**2.34**	.4738	**.4794**	**1.18**	.5106	**.5163**	**1.12**
Pr. a 20 docs.	.5446	**.5620**	**3.20**	.5880	**.5967**	**1.48**	.4878	**.5078**	**4.10**	.5422	**.5500**	**1.44**	.4457	.4436	-0.47	.4819	.4745	-1.54
Pr. a 30 docs.	.4928	**.5036**	**2.19**	.5304	**.5442**	**2.60**	.4452	**.4526**	**1.66**	.4948	**.5059**	**2.24**	.4000	**.4007**	**0.17**	.4255	**.4305**	**1.18**
Pr. a 100 docs.	.3300	**.3348**	**1.45**	.3509	.3509	0.00	.2993	**.3024**	**1.04**	.3147	**.3196**	**1.56**	.2513	.2506	-0.28	.2702	**.2734**	**1.18**
Pr. a 200 docs.	.2234	**.2263**	**1.30**	.2315	**.2346**	**1.34**	.1997	**.2004**	**0.35**	.2093	**.2119**	**1.24**	.1617	**.1633**	**0.99**	.1694	**.1703**	**0.53**
Pr. a 500 docs.	.1090	**.1103**	**1.19**	.1115	**.1127**	**1.08**	.0983	**.0992**	**0.92**	.1023	**.1029**	**0.59**	.0827	**.0833**	**0.73**	.0842	**.0846**	**0.48**
Pr. a 1000 docs.	.0587	**.0593**	**1.02**	.0600	**.0606**	**1.00**	.0525	**.0527**	**0.38**	.0536	**.0537**	**0.19**	.0459	**.0461**	**0.44**	.0463	**.0468**	**1.08**

Tabla 7.3: Resultados obtenidos mediante lematización (*lem*), caso base, y pares de dependencia sintáctica obtenidos mediante el analizador CASCADE a partir de la consulta (*FNF*)

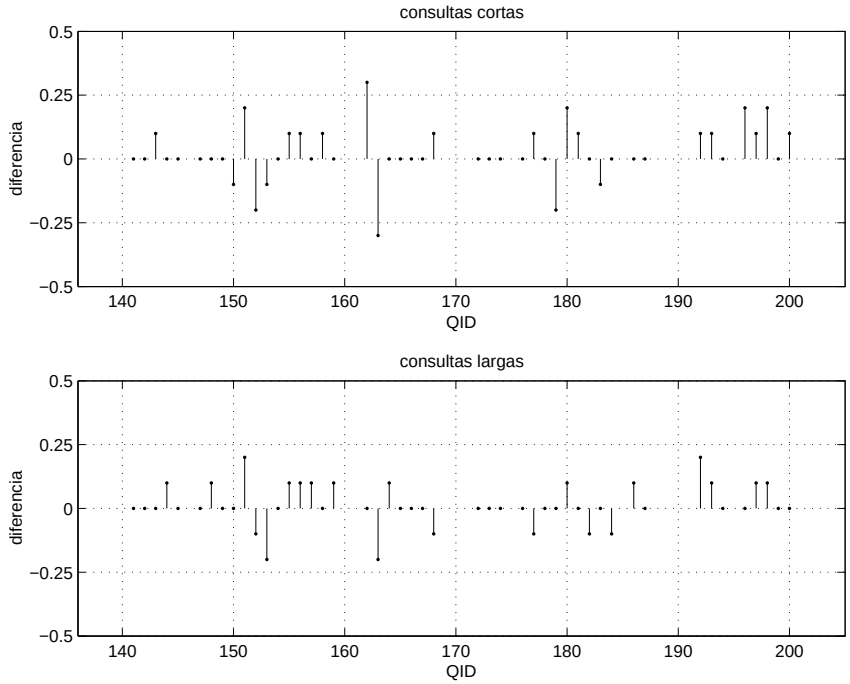


Figura 7.7: Diferencias en las precisiones a los 10 documentos: lematización vs. pares de dependencia sintáctica obtenidos mediante el analizador PATTERNS a partir de la consulta. Corpus CLEF 2003

parte, PATTERNS permite la adjunción de cualquier sintagma preposicional al sustantivo situado a su izquierda, disparando de esta forma el número de dependencias generadas, si bien con una alta proporción de dependencias incorrectas que introducen ruido en el sistema.

Sin embargo, puesto que el comportamiento para los corpus restantes es similar, y dadas las mejores características de robustez, modularidad, mantenibilidad y capacidad para establecer dependencias a mayor distancia, se decidió emplear el analizador CASCADE como analizador base para desarrollos futuros. De este modo, los experimentos recogidos en los apartados siguientes han sido realizados empleando exclusivamente el analizador CASCADE.

7.5.2. Resultados para Sintagmas Nominales

Trabajando ya únicamente sobre nuestro analizador basado en cascadas de traductores finitos, y para facilitar la comparación con otras aproximaciones clásicas basadas en la indexación de sintagmas nominales únicamente [101], se realizó una nueva serie de experimentos, restringiéndose esta vez a las dependencias correspondientes a sintagmas nominales, aquéllas entre un núcleo nominal y sus adjetivos (SA), y entre un núcleo nominal y sus complementos nominales (CN).

Tras una fase inicial de estimación de los factores de ponderación, tablas F.9 a F.12 del apéndice F, se emplearon unos valores ω=5 (*ratio* $1/\omega$=0.200) para el caso de las consultas cortas, y ω=10 (*ratio* $1/\omega$=0.100) para el caso de las consultas largas.

Los resultados obtenidos con este nuevo método se recogen en la tabla 7.4, y si bien existe una clara tendencia a la baja debido a la pérdida de la información aportada por las dependencias correspondientes a sintagmas verbales, los resultados son muy similares. Este punto queda patente en la tabla 7.5, en la cual comparamos su rendimiento con el de la propuesta inicial.

Por su parte, la gráfica 7.13 recoge la distribución por frecuencia de documento de los

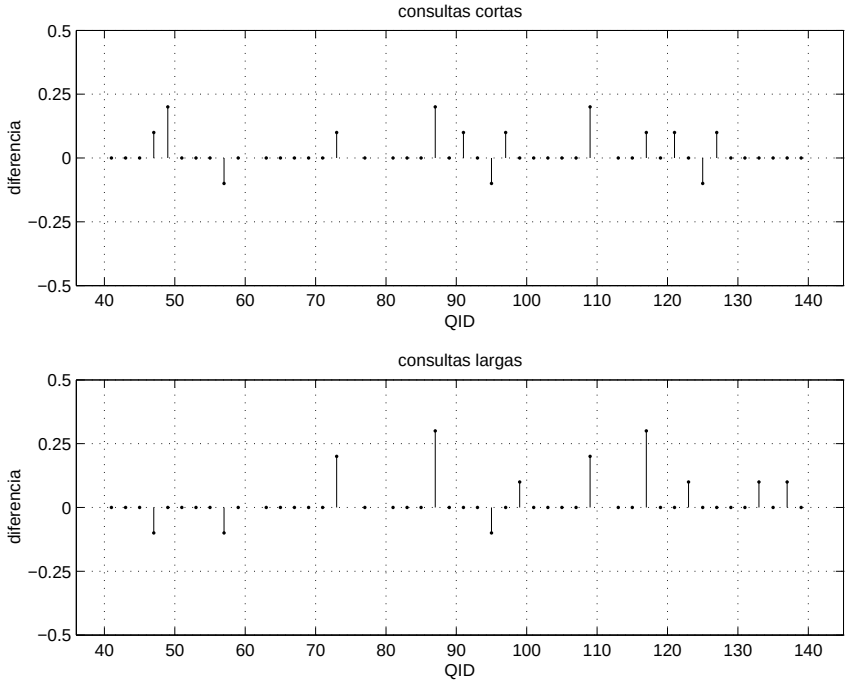


Figura 7.8: Diferencias en las precisiones a los 10 documentos: lematización vs. pares de dependencia sintáctica obtenidos mediante el analizador CASCADE a partir de la consulta. Corpus CLEF 2001-02·A

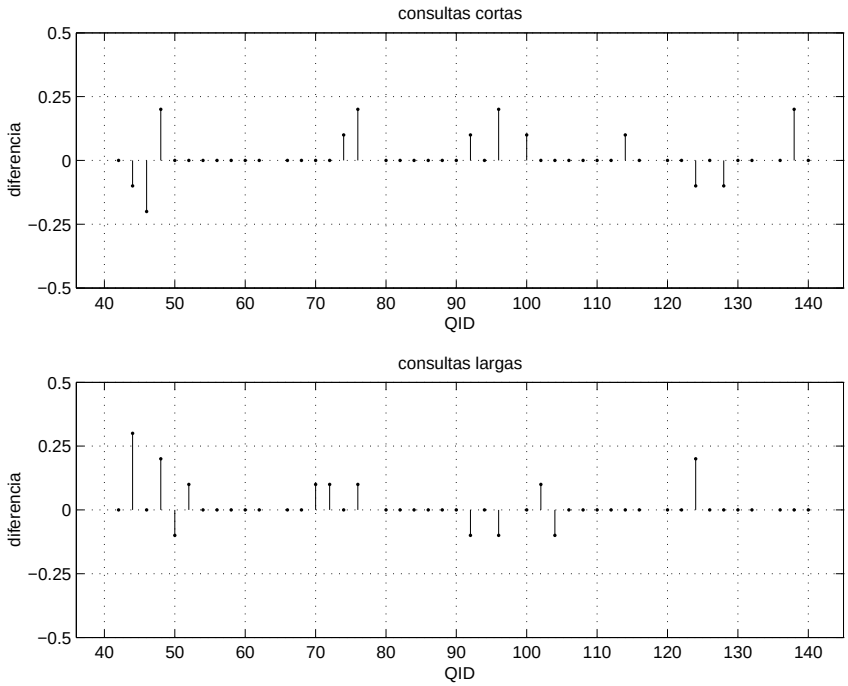


Figura 7.9: Diferencias en las precisiones a los 10 documentos: lematización vs. pares de dependencia sintáctica obtenidos mediante el analizador CASCADE a partir de la consulta. Corpus CLEF 2001-02·B

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	***SNF***	*%Δ*	*lem*	***SNF***	*%Δ*	*lem*	***SNF***	*%Δ*	*lem*	***SNF***	*%Δ*	*lem*	***SNF***	*%Δ*	*lem*	***SNF***	*%Δ*
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2700	**2726**	**0.96**	2762	**2781**	**0.69**	2363	**2368**	**0.21**	2410	**2413**	**0.12**	2159	**2172**	**0.60**	2175	**2197**	**1.01**
Pr. no int.	.4829	**.4940**	**2.30**	.5239	**.5303**	**1.22**	.4678	**.4752**	**1.58**	.5256	**.5371**	**2.19**	.4324	.4324	0.00	.4659	**.4699**	**0.86**
Pr. doc.	.5327	**.5457**	**2.44**	.5690	**.5791**	**1.78**	.5667	**.5683**	**0.28**	.6089	**.6191**	**1.68**	.4724	.4675	-1.04	.5201	**.5254**	**1.02**
R-pr.	.4848	**.4858**	**0.21**	.5075	**.5128**	**1.04**	.4549	**.4635**	**1.89**	.5030	**.5126**	**1.91**	.4362	**.4529**	**3.83**	.4493	**.4575**	**1.83**
Pr. a 0 %	.8293	**.8498**	**2.47**	.8845	.8617	-2.58	.8234	.8041	-2.34	.8763	.8759	-0.05	.8124	.8085	-0.48	.8366	**.8446**	**0.96**
Pr. a 10 %	.7463	**.7628**	**2.21**	.7914	**.8020**	**1.34**	.6845	**.7076**	**3.37**	.7517	**.7806**	**3.84**	.7057	.6775	-4.00	.7197	**.7355**	**2.20**
Pr. a 20 %	.6771	**.6878**	**1.58**	.7136	**.7277**	**1.98**	.6287	**.6547**	**4.14**	.7095	**.7185**	**1.27**	.6118	**.6259**	**2.30**	.6453	**.6511**	**0.90**
Pr. a 30 %	.6138	**.6273**	**2.20**	.6591	**.6678**	**1.32**	.5829	**.5844**	**0.26**	.6463	**.6524**	**0.94**	.5503	**.5556**	**0.96**	.5788	**.5909**	**2.09**
Pr. a 40 %	.5502	**.5614**	**2.04**	.6085	**.6124**	**0.64**	.5555	.5530	-0.45	.6100	**.6172**	**1.18**	.5059	.4991	-1.34	.5348	.5324	-0.45
Pr. a 50 %	.4931	**.5036**	**2.13**	.5557	**.5619**	**1.12**	.5136	**.5192**	**1.09**	.5658	**.5758**	**1.77**	.4657	.4599	-1.25	.4889	**.4942**	**1.08**
Pr. a 60 %	.4496	**.4618**	**2.71**	.5006	**.5128**	**2.44**	.4413	**.4526**	**2.56**	.4977	**.5097**	**2.41**	.4017	**.4049**	**0.80**	.4339	**.4400**	**1.41**
Pr. a 70 %	.3853	**.3881**	**0.73**	.4161	**.4248**	**2.09**	.3943	**.4002**	**1.50**	.4569	**.4631**	**1.36**	.3422	.3380	-1.23	.3789	**.3917**	**3.38**
Pr. a 80 %	.3277	**.3358**	**2.47**	.3509	**.3587**	**2.22**	.3122	**.3219**	**3.11**	.3679	**.3833**	**4.19**	.2700	**.2745**	**1.67**	.3206	**.3234**	**0.87**
Pr. a 90 %	.2356	**.2477**	**5.14**	.2492	**.2568**	**3.05**	.2319	**.2381**	**2.67**	.2845	**.2965**	**4.22**	.1821	**.1889**	**3.73**	.2067	**.2084**	**0.82**
Pr. a 100 %	.1197	**.1199**	**0.17**	.1289	**.1319**	**2.33**	.1209	**.1280**	**5.87**	.1547	**.1581**	**2.20**	.0932	**.0994**	**6.65**	.1164	**.1175**	**0.95**
Pr. a 5 docs.	.6609	**.6870**	**3.95**	.6957	.6957	0.00	.5956	**.6222**	**4.47**	.6844	**.6889**	**0.66**	.5915	.5915	0.00	.6213	**.6255**	**0.68**
Pr. a 10 docs.	.6283	**.6522**	**3.80**	.6543	**.6761**	**3.33**	.5667	**.5778**	**1.96**	.6000	**.6200**	**3.33**	.5149	**.5277**	**2.49**	.5596	**.5660**	**1.14**
Pr. a 15 docs.	.5928	**.5971**	**0.73**	.6188	**.6304**	**1.87**	.5170	**.5333**	**3.15**	.5689	**.5793**	**1.83**	.4738	.4738	0.00	.5106	.5092	-0.27
Pr. a 20 docs.	.5446	**.5587**	**2.59**	.5880	**.5935**	**0.94**	.4878	**.4989**	**2.28**	.5422	**.5467**	**0.83**	.4457	.4362	-2.13	.4819	.4723	-1.99
Pr. a 30 docs.	.4928	**.5036**	**2.19**	.5304	**.5391**	**1.64**	.4452	**.4533**	**1.82**	.4948	**.5044**	**1.94**	.4000	.3922	-1.95	.4255	**.4326**	**1.67**
Pr. a 100 docs.	.3300	**.3361**	**1.85**	.3509	**.3517**	**0.23**	.2993	.2987	-0.20	.3147	**.3184**	**1.18**	.2513	.2513	0.00	.2702	**.2730**	**1.04**
Pr. a 200 docs.	.2234	**.2257**	**1.03**	.2315	**.2341**	**1.12**	.1997	**.2000**	**0.15**	.2093	**.2116**	**1.10**	.1617	**.1622**	**0.31**	.1694	**.1703**	**0.53**
Pr. a 500 docs.	.1090	**.1101**	**1.01**	.1115	**.1123**	**0.72**	.0983	**.0989**	**0.61**	.1023	**.1029**	**0.59**	.0827	**.0828**	**0.12**	.0842	**.0845**	**0.36**
Pr. a 1000 docs.	.0587	**.0593**	**1.02**	.0600	**.0605**	**0.83**	.0525	**.0526**	**0.19**	.0536	.0536	0.00	.0459	**.0462**	**0.65**	.0463	**.0467**	**0.86**

Tabla 7.4: Resultados obtenidos mediante lematización (*lem*), caso base, y pares de dependencia sintáctica correspondientes a sintagmas nominales obtenidos a partir de la consulta (*SNF*)

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	***FNF***	***SNF***	***%Δ***	***FNF***	***SNF***	***%Δ***	***FNF***	***SNF***	***%Δ***	***FNF***	***SNF***	***%Δ***	***FNF***	***SNF***	***%Δ***	***FNF***	***SNF***	***%Δ***
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2728	2726	-0.07	2786	2781	-0.18	2371	2368	-0.13	2415	2413	-0.08	2169	**2172**	**0.14**	2198	2197	-0.05
Pr. no int.	.4965	.4940	-0.50	.5339	.5303	-0.67	.4810	.4752	-1.21	.5401	.5371	-0.56	.4377	.4324	-1.21	.4700	.4699	-0.02
Pr. doc.	.5491	.5457	-0.62	.5826	.5791	-0.60	.5728	.5683	-0.79	.6213	.6191	-0.35	.4738	.4675	-1.33	.5253	**.5254**	**0.02**
R-pr.	.4895	.4858	-0.76	.5185	.5128	-1.10	.4672	.4635	-0.79	.5133	.5126	-0.14	.4490	**.4529**	**0.87**	.4550	**.4575**	**0.55**
Pr. a 0 %	.8406	**.8498**	**1.09**	.8685	.8617	-0.78	.8137	.8041	-1.18	.8674	**.8759**	**0.98**	.8409	.8085	-3.85	.8342	**.8446**	**1.25**
Pr. a 10 %	.7640	.7628	-0.16	.8067	.8020	-0.58	.7128	.7076	-0.73	.7789	**.7806**	**0.22**	.6993	.6775	-3.12	.7354	**.7355**	**0.01**
Pr. a 20 %	.6925	.6878	-0.68	.7331	.7277	-0.74	.6608	.6547	-0.92	.7262	.7185	-1.06	.6364	.6259	-1.65	.6702	.6511	-2.85
Pr. a 30 %	.6338	.6273	-1.03	.6728	.6678	-0.74	.6137	.5844	-4.77	.6586	.6524	-0.94	.5566	.5556	-0.18	.5923	.5909	-0.24
Pr. a 40 %	.5622	.5614	-0.14	.6138	.6124	-0.23	.5575	.5530	-0.81	.6204	.6172	-0.52	.4992	.4991	-0.02	.5243	**.5324**	**1.54**
Pr. a 50 %	.5076	.5036	-0.79	.5655	.5619	-0.64	.5232	.5192	-0.76	.5787	.5758	-0.50	.4625	.4599	-0.56	.4874	**.4942**	**1.40**
Pr. a 60 %	.4648	.4618	-0.65	.5162	.5128	-0.66	.4552	.4526	-0.57	.5124	.5097	-0.53	.4103	.4049	-1.32	.4353	**.4400**	**1.08**
Pr. a 70 %	.3909	.3881	-0.72	.4292	.4248	-1.03	.4002	.4002	0.00	.4641	.4631	-0.22	.3385	.3380	-0.15	.3927	.3917	-0.25
Pr. a 80 %	.3364	.3358	-0.18	.3632	.3587	-1.24	.3280	.3219	-1.86	.3845	.3833	-0.31	.2758	.2745	-0.47	.3214	**.3234**	**0.62**
Pr. a 90 %	.2483	.2477	-0.24	.2581	.2568	-0.50	.2352	**.2381**	**1.23**	.2981	.2965	-0.54	.1913	.1889	-1.25	.2096	.2084	-0.57
Pr. a 100 %	.1197	**.1199**	**0.17**	.1318	**.1319**	**0.08**	.1242	**.1280**	**3.06**	.1572	**.1581**	**0.57**	.0992	**.0994**	**0.20**	.1182	.1175	-0.59
Pr. a 5 docs.	.6913	.6870	-0.62	.7043	.6957	-1.22	.6178	**.6222**	**0.71**	.6933	.6889	-0.63	.5745	**.5915**	**2.96**	.6340	.6255	-1.34
Pr. a 10 docs.	.6500	**.6522**	**0.34**	.6783	.6761	-0.32	.5822	.5778	-0.76	.6178	**.6200**	**0.36**	.5128	**.5277**	**2.91**	.5574	**.5660**	**1.54**
Pr. a 15 docs.	.6029	.5971	-0.96	.6319	.6304	-0.24	.5333	.5333	0.00	.5822	.5793	-0.50	.4794	.4738	-1.17	.5163	.5092	-1.38
Pr. a 20 docs.	.5620	.5587	-0.59	.5967	.5935	-0.54	.5078	.4989	-1.75	.5500	.5467	-0.60	.4436	.4362	-1.67	.4745	.4723	-0.46
Pr. a 30 docs.	.5036	.5036	0.00	.5442	.5391	-0.94	.4526	**.4533**	**0.15**	.5059	.5044	-0.30	.4007	.3922	-2.12	.4305	**.4326**	**0.49**
Pr. a 100 docs.	.3348	**.3361**	**0.39**	.3509	**.3517**	**0.23**	.3024	.2987	-1.22	.3196	.3184	-0.38	.2506	**.2513**	**0.28**	.2734	.2730	-0.15
Pr. a 200 docs.	.2263	.2257	-0.27	.2346	.2341	-0.21	.2004	.2000	-0.20	.2119	.2116	-0.14	.1633	.1622	-0.67	.1703	.1703	0.00
Pr. a 500 docs.	.1103	.1101	-0.18	.1127	.1123	-0.35	.0992	.0989	-0.30	.1029	.1029	0.00	.0833	.0828	-0.60	.0846	.0845	-0.12
Pr. a 1000 docs.	.0593	.0593	0.00	.0606	.0605	-0.17	.0527	.0526	-0.19	.0537	.0536	-0.19	.0461	**.0462**	**0.22**	.0468	.0467	-0.21

Tabla 7.5: Resultados obtenidos mediante pares de dependencia sintáctica obtenidos a partir de la consulta empleando la totalidad de las dependencias (*FNF*) o sólo aquéllas correspondientes a sintagmas nominales (*SNF*)

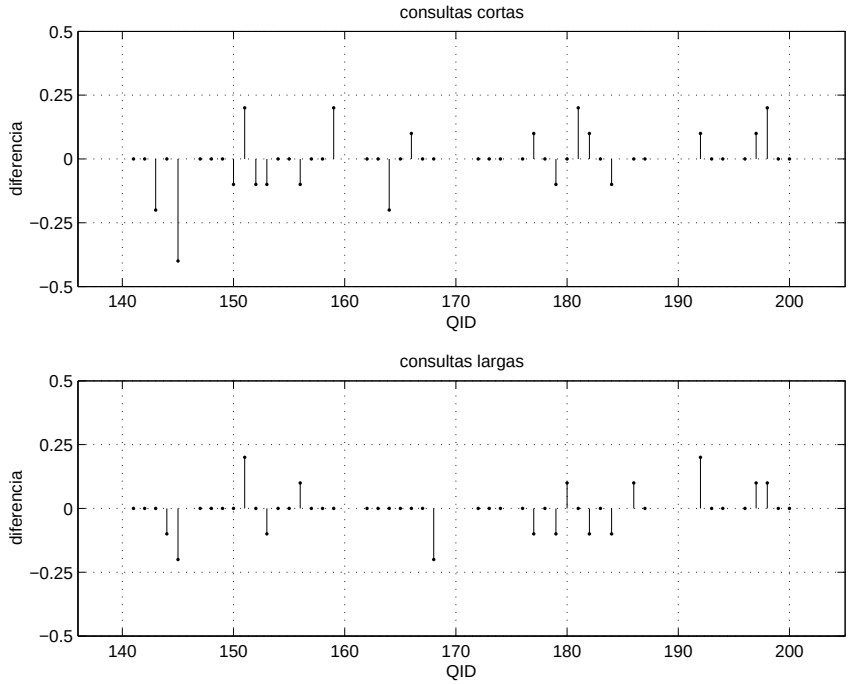


Figura 7.10: Diferencias en las precisiones a los 10 documentos: lematización vs. pares de dependencia sintáctica obtenidos mediante el analizador CASCADE a partir de la consulta. Corpus CLEF 2003

términos generados. Podemos observar que si bien la distribución sigue siendo prácticamente la misma que para el conjunto completo de dependencias —gráfica 7.12—, el número de términos distintos a indexar se ve decrementado en un 48 %, con la consiguiente reducción en el tamaño del índice obtenido. Sin embargo, llegados a este punto, debemos llamar la atención sobre el hecho de que si bien se produce una reducción de los costes asociados al almacenamiento y manejo del índice, el coste de generación de los términos es el mismo, ya que las dependencias *sustantivo–adjetivo* y *sustantivo–complemento nominal* precisan del análisis de sintagmas nominales y preposicionales, correspondientes a las capas finales del analizador.

7.5.3. Resultados Aplicando Realimentación

Siguiendo en la línea establecida en capítulos anteriores, estudiaremos las repercusiones en el rendimiento de la introducción de la realimentación en el sistema. Los parámetros de expansión automática de consultas son los usuales:

Consultas cortas: α=0.80, β=0.10, γ=0, n_1=5, t=10.

Consultas largas: α=1.20, β=0.10, γ=0, n_1=5, t=10.

La tabla 7.6 recoge los resultados generales obtenidos empleando expansión automática de consultas sobre nuestra propuesta basada en la combinación de términos simples y complejos (de todo tipo de dependencias) extraídos de las consultas. Podemos constatar que la introducción de la realimentación produce, como era de esperar, un aumento general del rendimiento, manteniéndose además las mejoras relativas respecto a la utilización de términos simples (habiéndoseles aplicado también realimentación). Las comparativas de precisiones a los 10 documentos —gráficas 7.14, 7.15 y 7.16— apoyan dicha información.

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*	*lem*	***FNF***	*%Δ*
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2767	**2811**	**1.59**	2779	2775	-0.14	2376	**2383**	**0.29**	2406	**2419**	**0.54**	2240	**2251**	**0.49**	2223	2222	-0.04
Pr. no int.	.5220	**.5434**	**4.10**	.5604	**.5770**	**2.96**	.4773	**.4993**	**4.61**	.5392	**.5629**	**4.40**	.5024	**.5037**	**0.26**	.5207	**.5241**	**0.65**
Pr. doc.	.5784	**.6045**	**4.51**	.5912	**.6059**	**2.49**	.5681	**.5754**	**1.28**	.6145	**.6321**	**2.86**	.5530	**.5542**	**0.22**	.5802	**.5852**	**0.86**
R-pr.	.4990	**.5210**	**4.41**	.5366	**.5693**	**6.09**	.4599	**.4880**	**6.11**	.5104	**.5207**	**2.02**	.4912	.4773	-2.83	.4871	**.4975**	**2.14**
Pr. a 0 %	.8221	**.8432**	**2.57**	.8895	.8755	-1.57	.8210	.8138	-0.88	.8710	.8657	-0.61	.8145	**.8425**	**3.44**	.8301	.8301	0.00
Pr. a 10 %	.7490	**.7786**	**3.95**	.8028	**.8127**	**1.23**	.6861	**.7164**	**4.42**	.7619	**.7724**	**1.38**	.7369	.7168	-2.73	.7421	**.7529**	**1.46**
Pr. a 20 %	.6866	**.7112**	**3.58**	.7352	**.7523**	**2.33**	.6319	**.6675**	**5.63**	.6929	**.7297**	**5.31**	.6632	.6526	-1.60	.6758	**.6955**	**2.92**
Pr. a 30 %	.6573	**.6723**	**2.28**	.6996	**.7107**	**1.59**	.5688	**.6219**	**9.34**	.6497	**.6725**	**3.51**	.6019	**.6107**	**1.46**	.6304	**.6345**	**0.65**
Pr. a 40 %	.5997	**.6257**	**4.34**	.6541	**.6699**	**2.42**	.5289	**.5545**	**4.84**	.6202	**.6380**	**2.87**	.5638	.5579	-1.05	.5975	.5780	-3.26
Pr. a 50 %	.5456	**.5864**	**7.48**	.6005	**.6316**	**5.18**	.5017	**.5216**	**3.97**	.5733	**.5995**	**4.57**	.5410	.5378	-0.59	.5479	.5345	-2.45
Pr. a 60 %	.4994	**.5360**	**7.33**	.5386	**.5720**	**6.20**	.4623	**.4704**	**1.75**	.5194	**.5475**	**5.41**	.4783	**.4877**	**1.97**	.4938	**.5016**	**1.58**
Pr. a 70 %	.4375	**.4540**	**3.77**	.4621	**.4999**	**8.18**	.4255	**.4409**	**3.62**	.4862	**.5109**	**5.08**	.4366	**.4440**	**1.69**	.4587	.4545	-0.92
Pr. a 80 %	.3661	**.3811**	**4.10**	.3910	**.4121**	**5.40**	.3384	**.3653**	**7.95**	.3753	**.4137**	**10.23**	.3788	**.3908**	**3.17**	.3891	**.4053**	**4.16**
Pr. a 90 %	.2939	**.2962**	**0.78**	.3130	**.3237**	**3.42**	.2651	**.2780**	**4.87**	.3059	**.3212**	**5.00**	.3102	**.3212**	**3.55**	.3230	**.3258**	**0.87**
Pr. a 100 %	.1547	**.1664**	**7.56**	.1624	**.1735**	**6.83**	.1395	**.1444**	**3.51**	.1704	**.1790**	**5.05**	.1871	**.1900**	**1.55**	.1928	**.2053**	**6.48**
Pr. a 5 docs.	.6609	**.6913**	**4.60**	.6957	**.7043**	**1.24**	.5956	**.6178**	**3.73**	.6844	**.6933**	**1.30**	.5872	.5745	-2.16	.6213	**.6383**	**2.74**
Pr. a 10 docs.	.6457	**.6739**	**4.37**	.6848	**.7043**	**2.85**	.5600	**.5933**	**5.95**	.6178	**.6422**	**3.95**	.5596	.5468	-2.29	.5872	**.6064**	**3.27**
Pr. a 15 docs.	.5884	**.6246**	**6.15**	.6435	**.6522**	**1.35**	.5274	**.5437**	**3.09**	.5822	**.6015**	**3.32**	.5305	.5135	-3.20	.5504	.5489	-0.27
Pr. a 20 docs.	.5630	**.5946**	**5.61**	.6043	**.6315**	**4.50**	.5011	**.5156**	**2.89**	.5533	**.5689**	**2.82**	.4883	.4872	-0.23	.5266	.5202	-1.22
Pr. a 30 docs.	.5225	**.5543**	**6.09**	.5580	**.5739**	**2.85**	.4444	**.4681**	**5.33**	.5081	**.5207**	**2.48**	.4433	**.4461**	**0.63**	.4667	**.4716**	**1.05**
Pr. a 100 docs.	.3507	**.3550**	**1.23**	.3598	**.3707**	**3.03**	.2940	**.2987**	**1.60**	.3191	**.3249**	**1.82**	.2770	.2768	-0.07	.2853	.2847	-0.21
Pr. a 200 docs.	.2348	**.2357**	**0.38**	.2361	**.2404**	**1.82**	.1979	**.2011**	**1.62**	.2067	**.2104**	**1.79**	.1753	**.1783**	**1.71**	.1791	**.1821**	**1.68**
Pr. a 500 docs.	.1122	**.1135**	**1.16**	.1121	.1120	-0.09	.0980	**.0996**	**1.63**	.1008	**.1018**	**0.99**	.0869	**.0888**	**2.19**	.0871	**.0874**	**0.34**
Pr. a 1000 docs.	.0602	**.0611**	**1.50**	.0604	.0603	-0.17	.0528	**.0530**	**0.38**	.0535	**.0538**	**0.56**	.0477	**.0479**	**0.42**	.0473	.0473	0.00

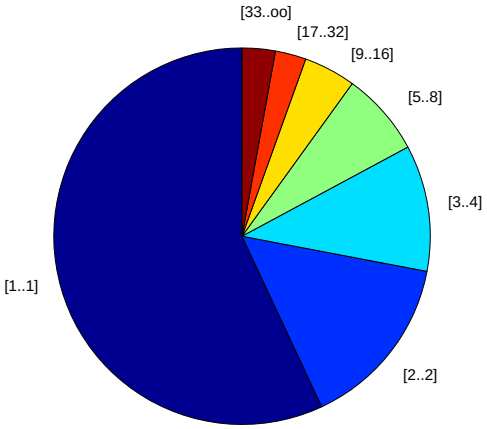
Tabla 7.6: Resultados obtenidos mediante lematización (*lem*), caso base, y pares de dependencia sintáctica obtenidos a partir de la consulta (*FNF*) aplicando en ambos casos realimentación

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	***SNF***	**%Δ**	*lem*	***SNF***	**%Δ**	*lem*	***SNF***	**%Δ**	*lem*	***SNF***	**%Δ**	*lem*	***SNF***	**%Δ**	*lem*	***SNF***	**%Δ**
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2767	**2807**	**1.45**	2779	2770	-0.32	2376	**2390**	**0.59**	2406	**2414**	**0.33**	2240	**2259**	**0.85**	2223	**2229**	**0.27**
Pr. no int.	.5220	**.5372**	**2.91**	.5604	**.5697**	**1.66**	.4773	**.4838**	**1.36**	.5392	**.5577**	**3.43**	.5024	**.5066**	**0.84**	.5207	**.5240**	**0.63**
Pr. doc.	.5784	**.5974**	**3.28**	.5912	**.6006**	**1.59**	.5681	.5625	-0.99	.6145	**.6276**	**2.13**	.5530	**.5596**	**1.19**	.5802	**.5841**	**0.67**
R-pr.	.4990	**.5167**	**3.55**	.5366	**.5597**	**4.30**	.4599	**.4671**	**1.57**	.5104	**.5213**	**2.14**	.4912	**.4956**	**0.90**	.4871	**.5006**	**2.77**
Pr. a 0 %	.8221	**.8541**	**3.89**	.8895	.8671	-2.52	.8210	.8047	-1.99	.8710	**.8724**	**0.16**	.8145	**.8156**	**0.14**	.8301	**.8409**	**1.30**
Pr. a 10 %	.7490	**.7805**	**4.21**	.8028	**.8047**	**0.24**	.6861	**.7110**	**3.63**	.7619	**.7789**	**2.23**	.7369	.7149	-2.99	.7421	**.7508**	**1.17**
Pr. a 20 %	.6866	**.7112**	**3.58**	.7352	**.7547**	**2.65**	.6319	**.6627**	**4.87**	.6929	**.7229**	**4.33**	.6632	**.6721**	**1.34**	.6758	**.6860**	**1.51**
Pr. a 30 %	.6573	**.6686**	**1.72**	.6996	**.7049**	**0.76**	.5688	**.6080**	**6.89**	.6497	**.6688**	**2.94**	.6019	**.6211**	**3.19**	.6304	**.6346**	**0.67**
Pr. a 40 %	.5997	**.6153**	**2.60**	.6541	**.6626**	**1.30**	.5289	**.5447**	**2.99**	.6202	**.6356**	**2.48**	.5638	**.5678**	**0.71**	.5975	.5825	-2.51
Pr. a 50 %	.5456	**.5720**	**4.84**	.6005	**.6239**	**3.90**	.5017	**.5141**	**2.47**	.5733	**.5939**	**3.59**	.5410	**.5440**	**0.55**	.5479	.5397	-1.50
Pr. a 60 %	.4994	**.5236**	**4.85**	.5386	**.5620**	**4.34**	.4623	.4588	-0.76	.5194	**.5355**	**3.10**	.4783	**.4933**	**3.14**	.4938	**.5000**	**1.26**
Pr. a 70 %	.4375	**.4437**	**1.42**	.4621	**.4848**	**4.91**	.4255	.4201	-1.27	.4862	**.5022**	**3.29**	.4366	**.4438**	**1.65**	.4587	.4570	-0.37
Pr. a 80 %	.3661	**.3690**	**0.79**	.3910	**.4034**	**3.17**	.3384	.3334	-1.48	.3753	**.4015**	**6.98**	.3788	**.3908**	**3.17**	.3891	**.4080**	**4.86**
Pr. a 90 %	.2939	**.2963**	**0.82**	.3130	**.3206**	**2.43**	.2651	.2576	-2.83	.3059	**.3176**	**3.82**	.3102	**.3176**	**2.39**	.3230	**.3295**	**2.01**
Pr. a 100 %	.1547	**.1618**	**4.59**	.1624	**.1734**	**6.77**	.1395	.1375	-1.43	.1704	**.1753**	**2.88**	.1871	.1834	-1.98	.1928	**.1987**	**3.06**
Pr. a 5 docs.	.6609	**.6826**	**3.28**	.6957	.6957	0.00	.5956	**.6222**	**4.47**	.6844	**.6889**	**0.66**	.5872	**.5915**	**0.73**	.6213	**.6255**	**0.68**
Pr. a 10 docs.	.6457	**.6609**	**2.35**	.6848	**.6870**	**0.32**	.5600	**.5844**	**4.36**	.6178	**.6400**	**3.59**	.5596	**.5681**	**1.52**	.5872	**.6021**	**2.54**
Pr. a 15 docs.	.5884	**.6145**	**4.44**	.6435	**.6464**	**0.45**	.5274	**.5363**	**1.69**	.5822	**.6000**	**3.06**	.5305	.5277	-0.53	.5504	**.5518**	**0.25**
Pr. a 20 docs.	.5630	**.5913**	**5.03**	.6043	**.6196**	**2.53**	.5011	**.5056**	**0.90**	.5533	**.5644**	**2.01**	.4883	**.5011**	**2.62**	.5266	.5202	-1.22
Pr. a 30 docs.	.5225	**.5507**	**5.40**	.5580	**.5696**	**2.08**	.4444	**.4615**	**3.85**	.5081	**.5185**	**2.05**	.4433	**.4539**	**2.39**	.4667	**.4681**	**0.30**
Pr. a 100 docs.	.3507	**.3530**	**0.66**	.3598	**.3676**	**2.17**	.2940	.2931	-0.31	.3191	**.3247**	**1.75**	.2770	**.2813**	**1.55**	.2853	.2853	0.00
Pr. a 200 docs.	.2348	.2337	-0.47	.2361	**.2397**	**1.52**	.1979	**.1990**	**0.56**	.2067	**.2087**	**0.97**	.1753	**.1789**	**2.05**	.1791	**.1803**	**0.67**
Pr. a 500 docs.	.1122	**.1136**	**1.25**	.1121	**.1123**	**0.18**	.0980	**.0999**	**1.94**	.1008	**.1013**	**0.50**	.0869	**.0886**	**1.96**	.0871	**.0876**	**0.57**
Pr. a 1000 docs.	.0602	**.0610**	**1.33**	.0604	.0602	-0.33	.0528	**.0531**	**0.57**	.0535	**.0536**	**0.19**	.0477	**.0481**	**0.84**	.0473	**.0474**	**0.21**

Tabla 7.7: Resultados obtenidos mediante lematización (*lem*), caso base, y pares de dependencia sintáctica correspondientes a sintagmas nominales obtenidos a partir de la consulta (*SNF*) aplicando en ambos casos realimentación

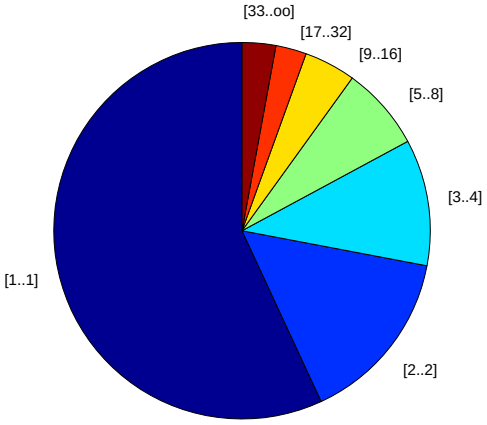
corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	***FNF***	***SNF***	**%Δ**	***FNF***	***SNF***	**%Δ**	***FNF***	***SNF***	**%Δ**	***FNF***	***SNF***	**%Δ**	***FNF***	***SNF***	**%Δ**	***FNF***	***SNF***	**%Δ**
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2811	2807	-0.14	2775	2770	-0.18	2383	**2390**	**0.29**	2419	2414	-0.21	2251	**2259**	**0.36**	2222	**2229**	**0.32**
Pr. no int.	.5434	.5372	-1.14	.5770	.5697	-1.27	.4993	.4838	-3.10	.5629	.5577	-0.92	.5037	**.5066**	**0.58**	.5241	.5240	-0.02
Pr. doc.	.6045	.5974	-1.17	.6059	.6006	-0.87	.5754	.5625	-2.24	.6321	.6276	-0.71	.5542	**.5596**	**0.97**	.5852	.5841	-0.19
R-pr.	.5210	.5167	-0.83	.5693	.5597	-1.69	.4880	.4671	-4.28	.5207	**.5213**	**0.12**	.4773	**.4956**	**3.83**	.4975	**.5006**	**0.62**
Pr. a 0 %	.8432	**.8541**	**1.29**	.8755	.8671	-0.96	.8138	.8047	-1.12	.8657	**.8724**	**0.77**	.8425	.8156	-3.19	.8301	**.8409**	**1.30**
Pr. a 10 %	.7786	**.7805**	**0.24**	.8127	.8047	-0.98	.7164	.7110	-0.75	.7724	**.7789**	**0.84**	.7168	.7149	-0.27	.7529	.7508	-0.28
Pr. a 20 %	.7112	.7112	0.00	.7523	**.7547**	**0.32**	.6675	.6627	-0.72	.7297	.7229	-0.93	.6526	**.6721**	**2.99**	.6955	.6860	-1.37
Pr. a 30 %	.6723	.6686	-0.55	.7107	.7049	-0.82	.6219	.6080	-2.24	.6725	.6688	-0.55	.6107	**.6211**	**1.70**	.6345	**.6346**	**0.02**
Pr. a 40 %	.6257	.6153	-1.66	.6699	.6626	-1.09	.5545	.5447	-1.77	.6380	.6356	-0.38	.5579	**.5678**	**1.77**	.5780	**.5825**	**0.78**
Pr. a 50 %	.5864	.5720	-2.46	.6316	.6239	-1.22	.5216	.5141	-1.44	.5995	.5939	-0.93	.5378	**.5440**	**1.15**	.5345	**.5397**	**0.97**
Pr. a 60 %	.5360	.5236	-2.31	.5720	.5620	-1.75	.4704	.4588	-2.47	.5475	.5355	-2.19	.4877	**.4933**	**1.15**	.5016	.5000	-0.32
Pr. a 70 %	.4540	.4437	-2.27	.4999	.4848	-3.02	.4409	.4201	-4.72	.5109	.5022	-1.70	.4440	.4438	-0.05	.4545	**.4570**	**0.55**
Pr. a 80 %	.3811	.3690	-3.18	.4121	.4034	-2.11	.3653	.3334	-8.73	.4137	.4015	-2.95	.3908	.3908	0.00	.4053	**.4080**	**0.67**
Pr. a 90 %	.2962	**.2963**	**0.03**	.3237	.3206	-0.96	.2780	.2576	-7.34	.3212	.3176	-1.12	.3212	.3176	-1.12	.3258	**.3295**	**1.14**
Pr. a 100 %	.1664	.1618	-2.76	.1735	.1734	-0.06	.1444	.1375	-4.78	.1790	.1753	-2.07	.1900	.1834	-3.47	.2053	.1987	-3.21
Pr. a 5 docs.	.6913	.6826	-1.26	.7043	.6957	-1.22	.6178	**.6222**	**0.71**	.6933	.6889	-0.63	.5745	**.5915**	**2.96**	.6383	.6255	-2.01
Pr. a 10 docs.	.6739	.6609	-1.93	.7043	.6870	-2.46	.5933	.5844	-1.50	.6422	.6400	-0.34	.5468	**.5681**	**3.90**	.6064	.6021	-0.71
Pr. a 15 docs.	.6246	.6145	-1.62	.6522	.6464	-0.89	.5437	.5363	-1.36	.6015	.6000	-0.25	.5135	**.5277**	**2.77**	.5489	**.5518**	**0.53**
Pr. a 20 docs.	.5946	.5913	-0.55	.6315	.6196	-1.88	.5156	.5056	-1.94	.5689	.5644	-0.79	.4872	**.5011**	**2.85**	.5202	.5202	0.00
Pr. a 30 docs.	.5543	.5507	-0.65	.5739	.5696	-0.75	.4681	.4615	-1.41	.5207	.5185	-0.42	.4461	**.4539**	**1.75**	.4716	.4681	-0.74
Pr. a 100 docs.	.3550	.3530	-0.56	.3707	.3676	-0.84	.2987	.2931	-1.87	.3249	.3247	-0.06	.2768	**.2813**	**1.63**	.2847	**.2853**	**0.21**
Pr. a 200 docs.	.2357	.2337	-0.85	.2404	.2397	-0.29	.2011	.1990	-1.04	.2104	.2087	-0.81	.1783	**.1789**	**0.34**	.1821	.1803	-0.99
Pr. a 500 docs.	.1135	**.1136**	**0.09**	.1120	**.1123**	**0.27**	.0996	**.0999**	**0.30**	.1018	.1013	-0.49	.0888	.0886	-0.23	.0874	**.0876**	**0.23**
Pr. a 1000 docs.	.0611	.0610	-0.16	.0603	.0602	-0.17	.0530	**.0531**	**0.19**	.0538	.0536	-0.37	.0479	**.0481**	**0.42**	.0473	**.0474**	**0.21**

Tabla 7.8: Resultados obtenidos mediante pares de dependencia sintáctica obtenidos a partir de la consulta empleando la totalidad de las dependencias (*FNF*) o sólo aquéllas correspondientes a sintagmas nominales (*SNF*) aplicando en ambos casos realimentación



df	*#pares*	*%pares*
[1..1]	3922172	56.95 %
[2..2]	1036809	15.06 %
[3..4]	747078	10.85 %
[5..8]	495083	7.19 %
[9..16]	307203	4.46 %
[17..32]	182384	2.65 %
[33..∞)	195746	2.84 %
Total	6886475	100.00 %
df≥5	1180416	17.14 %

Figura 7.11: Distribución por frecuencia de documento (*df*) de los *términos complejos* de la colección CLEF 2003 obtenidos empleando el analizador PATTERNS



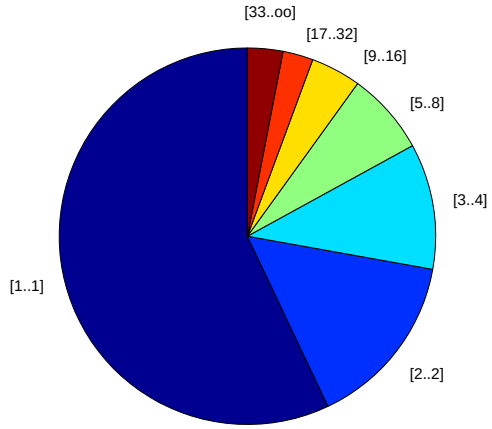
df	*#pares*	*%pares*
[1..1]	2227774	56.91 %
[2..2]	591121	15.10 %
[3..4]	425179	10.86 %
[5..8]	280921	7.18 %
[9..16]	173262	4.43 %
[17..32]	103181	2.64 %
[33..∞)	112995	2.89 %
Total	3914433	100.00 %
df≥5	670359	17.13 %

Figura 7.12: Distribución por frecuencia de documento (*df*) de los *términos complejos* de la colección CLEF 2003 obtenidos empleando el analizador CASCADE

En lo que respecta al comportamiento de los términos complejos basados en dependencias nominales, la tabla 7.7 muestra los resultados obtenidos al expandir automáticamente las consultas. Al igual que para el caso general, se produce también, como era de esperar, un incremento del rendimiento del sistema. Sin embargo, las diferencias respecto al empleo del conjunto completo de dependencias se hacen ahora palpables, quedando patentes en tabla 7.8, en la cual comparamos de nuevo las propuestas basadas en ambos tipos de pares y donde el empleo exclusivo de dependencias internas a frases nominales demuestra un peor comportamiento, especialmente en el caso de las consultas cortas —salvo para el caso del corpus CLEF 2003, más irregular.

7.6. Resultados Experimentales con Información Sintáctica Extraída de los Documentos

Los resultados obtenidos hasta el momento empleando la información sintáctica extraída a partir de las consultas han sido satisfactorios, obteniendo una mejora consistente respecto al empleo exclusivo de términos simples. Sin embargo, consideramos que podría existir todavía un margen para la mejora. Por ello decidimos desviar nuestra atención desde la información



df	*#pares*	*%pares*
[1..1]	1171381	57.04 %
[2..2]	311022	15.14 %
[3..4]	221998	10.81 %
[5..8]	145041	7.06 %
[9..16]	88544	4.31 %
[17..32]	53491	2.60 %
[33..∞)	62253	3.03 %
Total	2053730	100.00 %
df≥5	349329	17.01 %

Figura 7.13: Distribución por frecuencia de documento (*df*) de los *términos complejos* correspondientes a sintagmas nominales de la colección CLEF 2003 obtenidos empleando el analizador CASCADE

sintáctica de las consultas a la información sintáctica de los documentos.

A la hora de acceder a esa información se optó por una solución basada en la realimentación, de tal forma que si bien el proceso de indexación combinada de términos simples y complejos sigue siendo el mismo, el proceso de interrogación al sistema se desarrolla ahora en tres fases:

1. La consulta inicial, formada por términos simples —los lemas de las palabras con contenido de la consulta—, es lanzada contra el sistema.

2. Se obtienen las dependencias más informativas de los documentos devueltos por esta consulta inicial, las cuales son empleadas para expandir la consulta inicial de términos simples lematizados. Dichas dependencias son seleccionadas mediante el algoritmo de realimentación de Rocchio [196], seleccionando los pares situados entre los t' mejores términos de los n'_1 primeros documentos devueltos.

3. El sistema es interrogado con la consulta expandida resultante, obteniendo el conjunto final de documentos a devolver.

Buckley et al. [42] también expanden las consultas con las X mejores frases de los primeros documentos devueltos (además de con los Y mejores términos simples); no obstante, existen diferencias manifiestas entre su aproximación y la nuestra. En primer lugar, Buckley et al. expanden con las X mejores frases, mientras que en nuestra aproximación expandimos con los mejores pares comprendidos entre los X mejores términos. En segundo lugar, se trata de frases no lingüísticas basadas en meros pares de *no-stopwords* adyacentes. En tercer lugar, los autores aplican su propuesta a un sistema de *enrutamiento* en lugar de a un sistema de recuperación *ad-hoc* como ocurre en nuestro caso, por lo que los criterios de relevancia acerca de los documentos a partir de los cuales los términos son extraídos para la realimentación han sido establecidos, en su caso, manualmente —en lugar de suponer relevantes los n_1 documentos devueltos, como es nuestro caso—, por lo que dichos documentos resultan fuentes de información mucho más fiables. Por último, si bien ambas aproximaciones emplean SMART, el equipo de Buckley emplea un esquema de pesos más simple y de peor rendimiento, por lo que su margen de mejora es mayor.

7.6.1. Resultados para Sintagmas Nominales y Verbales

Antes de evaluar nuestra nueva aproximación sintáctica, se hace necesaria una fase previa de estimación de los parámetros del modelo. En esta ocasión no sólo debe fijarse el valor

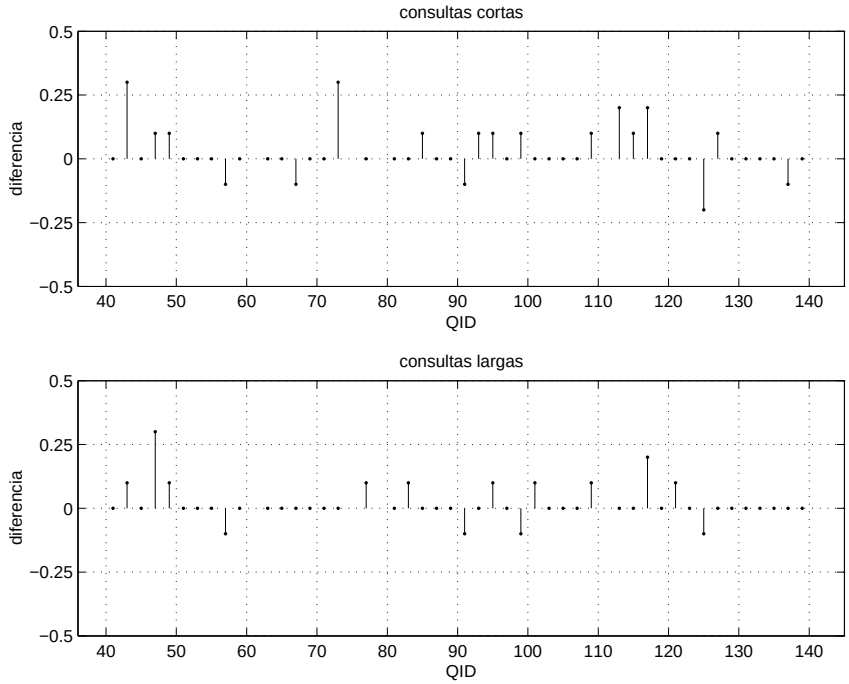


Figura 7.14: Diferencias en las precisiones a los 10 documentos aplicando realimentación: lematización vs. pares de dependencia sintáctica obtenidos a partir de la consulta. Corpus CLEF 2001-02·A

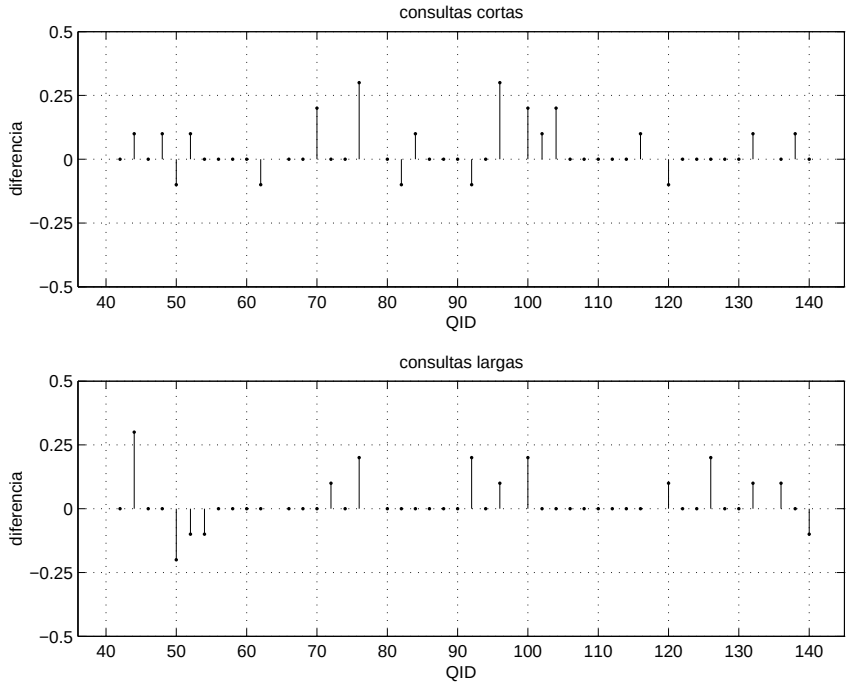


Figura 7.15: Diferencias en las precisiones a los 10 documentos aplicando realimentación: lematización vs. pares de dependencia sintáctica obtenidos a partir de la consulta. Corpus CLEF 2001-02·B

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	*DSD*	%Δ	*lem*	*DSD*	%Δ	*lem*	*DSD*	%Δ	*lem*	*DSD*	%Δ	*lem*	*DSD*	%Δ	*lem*	*DSD*	%Δ
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2700	**2758**	**2.15**	2762	**2765**	**0.11**	2363	**2394**	**1.31**	2410	**2417**	**0.29**	2159	**2246**	**4.03**	2175	**2240**	**2.99**
Pr. no int.	.4829	**.5286**	**9.46**	.5239	**.5577**	**6.45**	.4678	**.4990**	**6.67**	.5256	**.5475**	**4.17**	.4324	**.4690**	**8.46**	.4659	**.5015**	**7.64**
Pr. doc.	.5327	**.5768**	**8.28**	.5690	**.5902**	**3.73**	.5667	**.5892**	**3.97**	.6089	**.6286**	**3.24**	.4724	**.5432**	**14.99**	.5201	**.5666**	**8.94**
R-pr.	.4848	**.5119**	**5.59**	.5075	**.5274**	**3.92**	.4549	**.4765**	**4.75**	.5030	**.5267**	**4.71**	.4362	**.4582**	**5.04**	.4493	**.4794**	**6.70**
Pr. a 0 %	.8293	**.8389**	**1.16**	.8845	.8759	-0.97	.8234	.8033	-2.44	.8763	.8298	-5.31	.8124	.7964	-1.97	.8366	.8035	-3.96
Pr. a 10 %	.7463	**.7794**	**4.44**	.7914	**.8233**	**4.03**	.6845	**.6990**	**2.12**	.7517	**.7622**	**1.40**	.7057	**.7300**	**3.44**	.7197	**.7467**	**3.75**
Pr. a 20 %	.6771	**.7244**	**6.99**	.7136	**.7651**	**7.22**	.6287	**.6681**	**6.27**	.7095	**.7209**	**1.61**	.6118	**.6564**	**7.29**	.6453	**.6782**	**5.10**
Pr. a 30 %	.6138	**.6497**	**5.85**	.6591	**.7016**	**6.45**	.5829	**.6224**	**6.78**	.6463	**.6648**	**2.86**	.5503	**.5818**	**5.72**	.5788	**.6209**	**7.27**
Pr. a 40 %	.5502	**.5926**	**7.71**	.6085	**.6431**	**5.69**	.5555	**.5860**	**5.49**	.6100	**.6400**	**4.92**	.5059	**.5499**	**8.70**	.5348	**.5848**	**9.35**
Pr. a 50 %	.4931	**.5534**	**12.23**	.5557	**.5988**	**7.76**	.5136	**.5535**	**7.77**	.5658	**.6048**	**6.89**	.4657	**.5046**	**8.35**	.4889	**.5355**	**9.53**
Pr. a 60 %	.4496	**.5005**	**11.32**	.5006	**.5326**	**6.39**	.4413	**.4779**	**8.29**	.4977	**.5448**	**9.46**	.4017	**.4439**	**10.51**	.4339	**.4798**	**10.58**
Pr. a 70 %	.3853	**.4343**	**12.72**	.4161	**.4468**	**7.38**	.3943	**.4371**	**10.85**	.4569	**.5004**	**9.52**	.3422	**.3959**	**15.69**	.3789	**.4308**	**13.70**
Pr. a 80 %	.3277	**.3678**	**12.24**	.3509	**.3710**	**5.73**	.3122	**.3406**	**9.10**	.3679	**.3901**	**6.03**	.2700	**.3166**	**17.26**	.3206	**.3704**	**15.53**
Pr. a 90 %	.2356	**.2775**	**17.78**	.2492	**.2807**	**12.64**	.2319	**.2581**	**11.30**	.2845	**.3058**	**7.49**	.1821	**.2303**	**26.47**	.2067	**.2559**	**23.80**
Pr. a 100 %	.1197	**.1522**	**27.15**	.1289	**.1483**	**15.05**	.1209	**.1434**	**18.61**	.1547	**.1750**	**13.12**	.0932	**.1100**	**18.03**	.1164	**.1278**	**9.79**
Pr. a 5 docs.	.6609	**.7000**	**5.92**	.6957	**.7304**	**4.99**	.5956	**.6533**	**9.69**	.6844	**.7022**	**2.60**	.5915	**.6128**	**3.60**	.6213	**.6468**	**4.10**
Pr. a 10 docs.	.6283	**.6717**	**6.91**	.6543	**.6913**	**5.65**	.5667	**.5800**	**2.35**	.6000	**.6356**	**5.93**	.5149	**.5745**	**11.58**	.5596	**.6149**	**9.88**
Pr. a 15 docs.	.5928	**.6203**	**4.64**	.6188	**.6420**	**3.75**	.5170	**.5363**	**3.73**	.5689	**.5852**	**2.87**	.4738	**.5390**	**13.76**	.5106	**.5504**	**7.79**
Pr. a 20 docs.	.5446	**.5935**	**8.98**	.5880	**.6109**	**3.89**	.4878	**.5056**	**3.65**	.5422	**.5533**	**2.05**	.4457	**.4957**	**11.22**	.4819	**.5064**	**5.08**
Pr. a 30 docs.	.4928	**.5348**	**8.52**	.5304	**.5529**	**4.24**	.4452	**.4607**	**3.48**	.4948	**.5059**	**2.24**	.4000	**.4468**	**11.70**	.4255	**.4560**	**7.17**
Pr. a 100 docs.	.3300	**.3474**	**5.27**	.3509	**.3578**	**1.97**	.2993	**.3078**	**2.84**	.3147	**.3267**	**3.81**	.2513	**.2719**	**8.20**	.2702	**.2764**	**2.29**
Pr. a 200 docs.	.2234	**.2336**	**4.57**	.2315	**.2373**	**2.51**	.1997	**.2049**	**2.60**	.2093	**.2111**	**0.86**	.1617	**.1739**	**7.54**	.1694	**.1789**	**5.61**
Pr. a 500 docs.	.1090	**.1117**	**2.48**	.1115	**.1130**	**1.35**	.0983	**.1002**	**1.93**	.1023	.1017	-0.59	.0827	**.0875**	**5.80**	.0842	**.0876**	**4.04**
Pr. a 1000 docs.	.0587	**.0600**	**2.21**	.0600	**.0601**	**0.17**	.0525	**.0532**	**1.33**	.0536	**.0537**	**0.19**	.0459	**.0478**	**4.14**	.0463	**.0477**	**3.02**

Tabla 7.9: Resultados obtenidos mediante lematización (*lem*), caso base, y pares de dependencia sintáctica obtenidos a partir de los documentos (*DSD*)

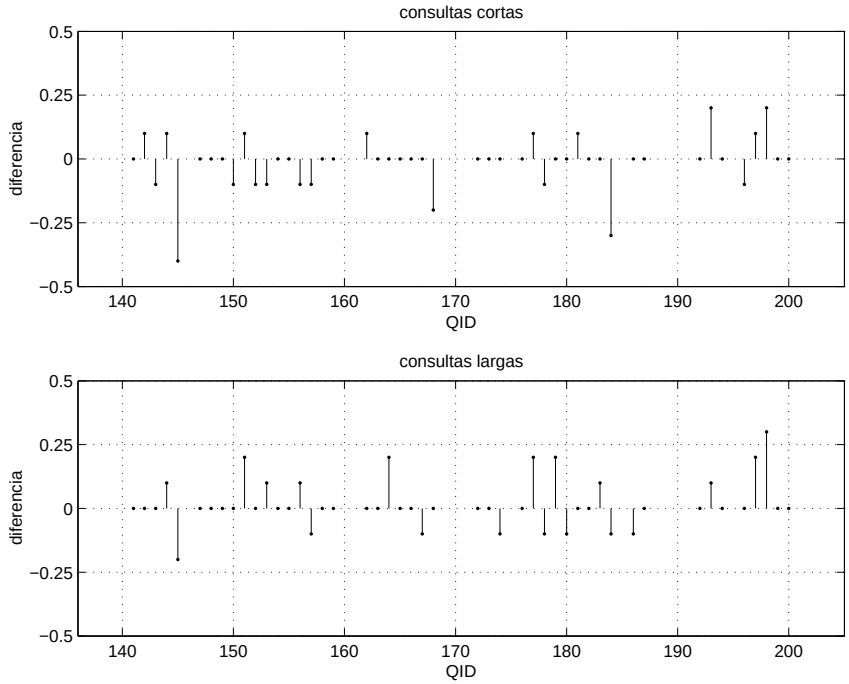


Figura 7.16: Diferencias en las precisiones a los 10 documentos aplicando realimentación: lematización vs. pares de dependencia sintáctica obtenidos a partir de la consulta. Corpus CLEF 2003

del factor de ponderación ω a utilizar, sino también los parámetros de selección de pares: el número de términos t' y el número de documentos n'_1 a partir de los cuales extraer los pares de expansión. Las tablas F.13 a F.16 del apéndice F recogen los resultados obtenidos con el corpus de entrenamiento CLEF 2001-02·A para los siguientes posibles valores de los parámetros n'_1, t' y ω (entre paréntesis el *ratio* $1/\omega$ correspondiente):

$$n'_1 \in 5, 10, 15, 20$$
$$t' \in \{5, 10, 15, 20, 30, 40, 50\}$$
$$\omega \in \{1\ \textit{(1)}, 2\ \textit{(0.500)}, 3\ \textit{(0.333)}, 4\ \textit{(0.250)}, 5\ \textit{(0.200)}, 8\ \textit{(0.125)}, 10\ \textit{(0.100)}, 20\ \textit{(0.050)}\}$$

La estimación se ha hecho tomando como criterio la precisión a los n documentos devueltos, ya que el objetivo perseguido sigue siendo el de aumentar la precisión de los primeros documentos devueltos.

Al analizar el conjunto de resultados obtenidos podemos apreciar que existe una tendencia según la cual, según incrementamos el número t' de términos examinados, también aumenta el rendimiento. Por contra, no existe una tendencia clara en el caso del comportamiento del sistema ante la variación en el número de documentos empleado, siendo especialmente irregular en el caso de las consultas cortas, ya que conforme varía el número de términos t' empleado, también varía, de forma irregular e impredecible, el número óptimo de documentos n'_1. Sí se puede afirmar, en cambio, que los resultados se estabilizan a partir de n'_1=15 documentos, ya que aunque se aumente el número n'_1 de documentos a emplear, éstos no introducen nuevos pares de interés. Finalmente, los parámetros escogidos fueron:

Consultas cortas: n'_1=10, t'=50, ω=3 *(ratio* $1/\omega$*=0.333)*

Consultas largas: n'_1=5, t'=50, ω=2 *(ratio* $1/\omega$*=0.500)*

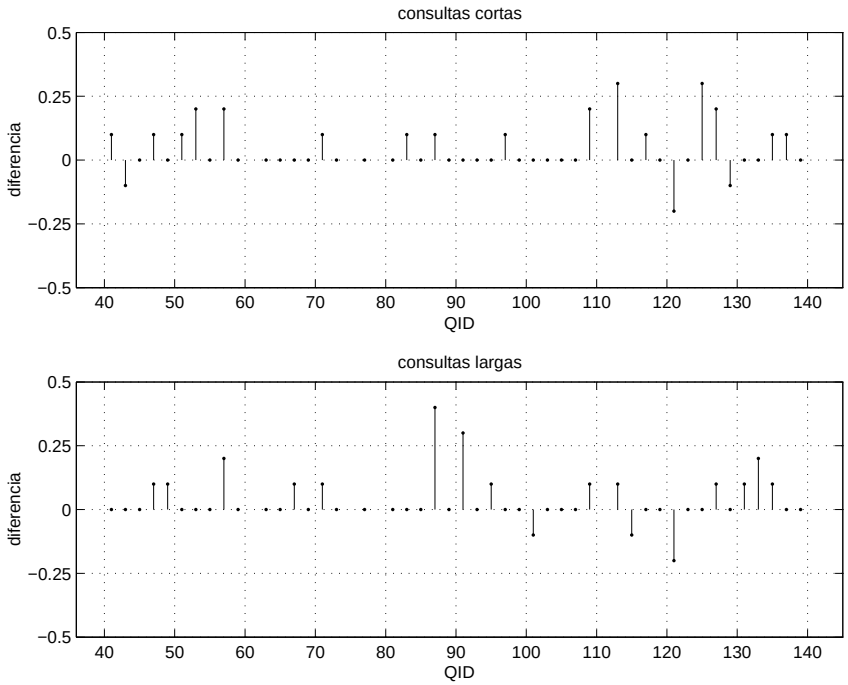


Figura 7.17: Diferencias en las precisiones a los 10 documentos: lematización vs. pares de dependencia sintáctica obtenidos a partir de los documentos. Corpus CLEF 2001-02·A

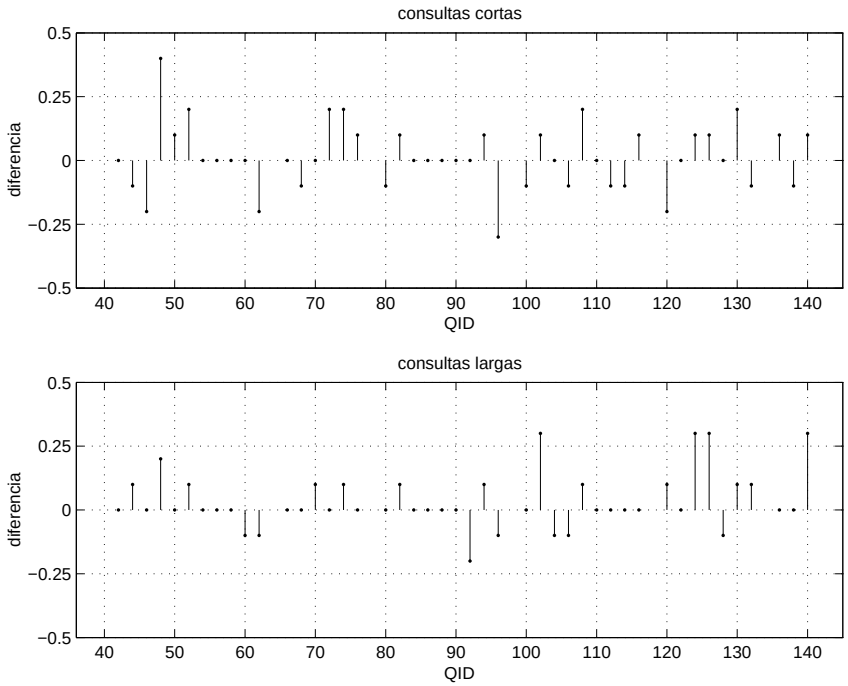


Figura 7.18: Diferencias en las precisiones a los 10 documentos: lematización vs. pares de dependencia sintáctica obtenidos a partir de los documentos. Corpus CLEF 2001-02·B

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	***FNF***	***DSD***	***%Δ***	***FNF***	***DSD***	***%Δ***	***FNF***	***DSD***	***%Δ***	***FNF***	***DSD***	***%Δ***	***FNF***	***DSD***	***%Δ***	***FNF***	***DSD***	***%Δ***
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2728	**2758**	**1.10**	2786	2765	-0.75	2371	**2394**	**0.97**	2415	**2417**	**0.08**	2169	**2246**	**3.55**	2198	**2240**	**1.91**
Pr. no int.	.4965	**.5286**	**6.47**	.5339	**.5577**	**4.46**	.4810	**.4990**	**3.74**	.5401	**.5475**	**1.37**	.4377	**.4690**	**7.15**	.4700	**.5015**	**6.70**
Pr. doc.	.5491	**.5768**	**5.04**	.5826	**.5902**	**1.30**	.5728	**.5892**	**2.86**	.6213	**.6286**	**1.17**	.4738	**.5432**	**14.65**	.5253	**.5666**	**7.86**
R-pr.	.4895	**.5119**	**4.58**	.5185	**.5274**	**1.72**	.4672	**.4765**	**1.99**	.5133	**.5267**	**2.61**	.4490	**.4582**	**2.05**	.4550	**.4794**	**5.36**
Pr. a 0 %	.8406	.8389	-0.20	.8685	**.8759**	**0.85**	.8137	.8033	-1.28	.8674	.8298	-4.33	.8409	.7964	-5.29	.8342	.8035	-3.68
Pr. a 10 %	.7640	**.7794**	**2.02**	.8067	**.8233**	**2.06**	.7128	.6990	-1.94	.7789	.7622	-2.14	.6993	**.7300**	**4.39**	.7354	**.7467**	**1.54**
Pr. a 20 %	.6925	**.7244**	**4.61**	.7331	**.7651**	**4.37**	.6608	**.6681**	**1.10**	.7262	.7209	-0.73	.6364	**.6564**	**3.14**	.6702	**.6782**	**1.19**
Pr. a 30 %	.6338	**.6497**	**2.51**	.6728	**.7016**	**4.28**	.6137	**.6224**	**1.42**	.6586	**.6648**	**0.94**	.5566	**.5818**	**4.53**	.5923	**.6209**	**4.83**
Pr. a 40 %	.5622	**.5926**	**5.41**	.6138	**.6431**	**4.77**	.5575	**.5860**	**5.11**	.6204	**.6400**	**3.16**	.4992	**.5499**	**10.16**	.5243	**.5848**	**11.54**
Pr. a 50 %	.5076	**.5534**	**9.02**	.5655	**.5988**	**5.89**	.5232	**.5535**	**5.79**	.5787	**.6048**	**4.51**	.4625	**.5046**	**9.10**	.4874	**.5355**	**9.87**
Pr. a 60 %	.4648	**.5005**	**7.68**	.5162	**.5326**	**3.18**	.4552	**.4779**	**4.99**	.5124	**.5448**	**6.32**	.4103	**.4439**	**8.19**	.4353	**.4798**	**10.22**
Pr. a 70 %	.3909	**.4343**	**11.10**	.4292	**.4468**	**4.10**	.4002	**.4371**	**9.22**	.4641	**.5004**	**7.82**	.3385	**.3959**	**16.96**	.3927	**.4308**	**9.70**
Pr. a 80 %	.3364	**.3678**	**9.33**	.3632	**.3710**	**2.15**	.3280	**.3406**	**3.84**	.3845	**.3901**	**1.46**	.2758	**.3166**	**14.79**	.3214	**.3704**	**15.25**
Pr. a 90 %	.2483	**.2775**	**11.76**	.2581	**.2807**	**8.76**	.2352	**.2581**	**9.74**	.2981	**.3058**	**2.58**	.1913	**.2303**	**20.39**	.2096	**.2559**	**22.09**
Pr. a 100 %	.1197	**.1522**	**27.15**	.1318	**.1483**	**12.52**	.1242	**.1434**	**15.46**	.1572	**.1750**	**11.32**	.0992	**.1100**	**10.89**	.1182	**.1278**	**8.12**
Pr. a 5 docs.	.6913	**.7000**	**1.26**	.7043	**.7304**	**3.71**	.6178	**.6533**	**5.75**	.6933	**.7022**	**1.28**	.5745	**.6128**	**6.67**	.6340	**.6468**	**2.02**
Pr. a 10 docs.	.6500	**.6717**	**3.34**	.6783	**.6913**	**1.92**	.5822	.5800	-0.38	.6178	**.6356**	**2.88**	.5128	**.5745**	**12.03**	.5574	**.6149**	**10.32**
Pr. a 15 docs.	.6029	**.6203**	**2.89**	.6319	**.6420**	**1.60**	.5333	**.5363**	**0.56**	.5822	**.5852**	**0.52**	.4794	**.5390**	**12.43**	.5163	**.5504**	**6.60**
Pr. a 20 docs.	.5620	**.5935**	**5.60**	.5967	**.6109**	**2.38**	.5078	.5056	-0.43	.5500	**.5533**	**0.60**	.4436	**.4957**	**11.74**	.4745	**.5064**	**6.72**
Pr. a 30 docs.	.5036	**.5348**	**6.20**	.5442	**.5529**	**1.60**	.4526	**.4607**	**1.79**	.5059	.5059	0.00	.4007	**.4468**	**11.50**	.4305	**.4560**	**5.92**
Pr. a 100 docs.	.3348	**.3474**	**3.76**	.3509	**.3578**	**1.97**	.3024	**.3078**	**1.79**	.3196	**.3267**	**2.22**	.2506	**.2719**	**8.50**	.2734	**.2764**	**1.10**
Pr. a 200 docs.	.2263	**.2336**	**3.23**	.2346	**.2373**	**1.15**	.2004	**.2049**	**2.25**	.2119	.2111	-0.38	.1633	**.1739**	**6.49**	.1703	**.1789**	**5.05**
Pr. a 500 docs.	.1103	**.1117**	**1.27**	.1127	**.1130**	**0.27**	.0992	**.1002**	**1.01**	.1029	.1017	-1.17	.0833	**.0875**	**5.04**	.0846	**.0876**	**3.55**
Pr. a 1000 docs.	.0593	**.0600**	**1.18**	.0606	.0601	-0.83	.0527	**.0532**	**0.95**	.0537	.0537	0.00	.0461	**.0478**	**3.69**	.0468	**.0477**	**1.92**

Tabla 7.10: Resultados obtenidos empleando los pares de dependencia sintáctica obtenidos a partir de la consulta (*FNF*) y a partir de los documentos (*DSD*)

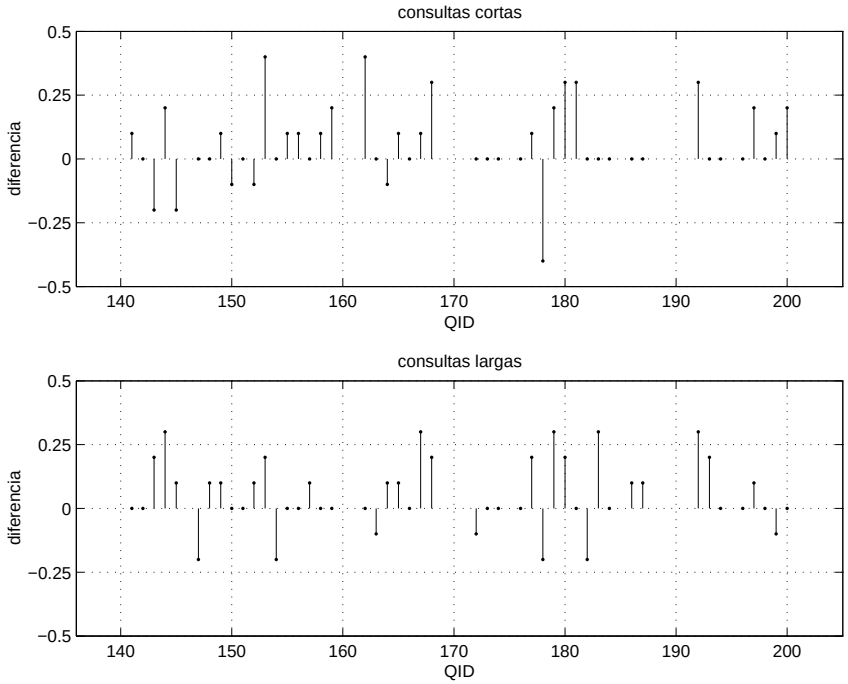


Figura 7.19: Diferencias en las precisiones a los 10 documentos: lematización vs. pares de dependencia sintáctica obtenidos a partir de los documentos. Corpus CLEF 2003

	consultas cortas			***consultas largas***		
corpus	$FNF \setminus DSD$	$DSD \setminus FNF$	$DSD \cap FNF$	$FNF \setminus DSD$	$DSD \setminus FNF$	$DSD \cap FNF$
CLEF 2001-02·A	106	267	38 (26.38 %)	293	461	46 (13.56 %)
CLEF 2001-02·B	121	291	42 (25.76 %)	306	499	53 (14.76 %)
CLEF 2001-03	126	286	48 (27.58 %)	351	436	48 (12.03 %)

Tabla 7.11: Comparación del número de términos complejos introducidos por la consulta (FNF) y por los documentos (DSD)

Los resultados obtenidos para los diferentes corpus de evaluación, así como las mejoras obtenidas respecto al empleo exclusivo de términos simples lematizados, se encuentran recogidos en la tabla 7.9 y en las gráficas 7.17, 7.18 y 7.19. En ellas se constata un incremento generalizado del rendimiento, no sólo respecto al empleo único de términos simples, sino también respecto a los resultados obtenidos con los pares procedentes de las consultas, lo que parece indicar que la información sintáctica procedente de los documentos resulta de mayor utilidad que aquélla proveniente de las consultas a la hora de incrementar la precisión de los documentos devueltos. En la tabla 7.10 se incluye una comparativa del rendimiento de ambas fuentes de información sintáctica: los pares extraídos de la consulta (FNF), y aquéllos obtenidos a partir de los documentos (DSD). La mejora mostrada en estos resultados es consistente y general.

7.6.2. Resultados para Sintagmas Nominales

Ante el interés despertado por los nuevos resultados, decidimos estudiar la posibilidad de que existiese algún tipo de relación constatable entre los términos aportados por cada una de las aproximaciones sintácticas. Para ello se estudió el conjunto de términos complejos introducidos por ambas soluciones, es decir, los pares aportados por la consulta y aquéllos otros aportados por

corpus	***CLEF 2001-02·A***				***CLEF 2001-02·B***				***CLEF 2003***			
consulta	***cortas***		***largas***		***cortas***		***largas***		***cortas***		***largas***	
tipo	FNF	$DSD \cap FNF$	FNF	$DSD \cap FNF$	FNF	$DSD \cap FNF$	FNF	$DSD \cap FNF$	FNF	$DSD \cap FNF$	FNF	$DSD \cap FNF$
SA	37.50	57.89	33.03	52.17	34.35	50.00	33.42	49.05	27.01	31.25	27.06	31.25
CN	45.83	34.21	45.72	36.95	44.78	45.23	44.28	41.50	47.70	64.58	46.11	56.25
SUJ_A	1.38	0.00	2.35	0.00	4.90	2.38	3.06	0.00	6.89	2.08	5.76	0.00
OD	6.25	2.63	11.50	10.86	9.20	7.14	9.19	11.32	7.47	0.00	10.52	6.25
SUJ_P	0.69	0.00	0.29	0.00	0.00	0.00	0.27	0.00	0.00	0.00	0.25	0.00
CA	0.00	0.00	0.00	0.00	0.00	0.00	0.27	1.88	0.00	0.00	0.25	0.00
CC	9.02	5.26	8.25	4.34	7.36	2.38	10.02	1.88	10.91	4.16	10.27	8.33
Atr	0.00	0.00	0.29	0.00	1.22	0.00	0.83	0.00	0.57	0.00	0.75	0.00

Tabla 7.12: Distribución porcentual, por tipos de dependencia asociada, de los términos complejos obtenidos a partir de la consulta (FNF) y aquéllos comunes con los obtenidos a partir de los documentos ($DSD \cap FNF$)

los documentos. La tabla 7.11 recoge el número de términos complejos eliminados al emplear la información sintáctica de los documentos ($FNF \setminus DSD$), el número de nuevos términos complejos introducidos por los documentos ($DSD \setminus FNF$), y el número de términos comunes a ambas soluciones ($DSD \cap FNF$), tanto en número absoluto como el porcentaje de los términos originales —de la consulta— que representa. Se puede apreciar que el porcentaje de pares de dependencia comunes a ambas soluciones se mantiene en unos márgenes constantes, lo que se podría considerar indicativo de la posible existencia de algún tipo de relación subyacente según la cual sólo un determinado tipo de pares de dependencia de la consulta original serían útiles desde el punto de vista del extractor automático.

Para analizar ese punto, se estudió la distribución de los diferentes tipos de dependencias sintácticas a los que correspondían aquellos pares originales obtenidos de la consulta (FNF) y aquellos pares comunes a ambas aproximaciones ($DSD \cap FNF$), para así comprobar la existencia de posibles sesgos o preferencias. Estos datos se recogen en la tabla 7.12, en la que se ha calculado la distribución porcentual de los diferentes tipos de dependencia sintáctica a partir de los cuales se obtiene un par dado[12]: sustantivo–adjetivo (SA), sustantivo–complemento nominal (CN), sujeto agente–verbo (SUJ_A), verbo–objeto (OD), sujeto pasivo–verbo (SUJ_P), verbo–agente (CA), verbo–complemento circunstancial (CC), y sujeto–atributo (Atr). Como se puede apreciar, las dependencias asociadas a sintagmas nominales —SA y CN— son las únicas que se ven favorecidas por el proceso de selección automática, al ser las únicas que ven aumentada su representatividad.

Dada la existencia de una preferencia a la hora de la realimentación mediante pares, la cual favorece a aquéllos pares procedentes de frases nominales, sería conveniente estudiar el comportamiento del sistema cuando únicamente se emplean aquellos términos complejos procedentes de dependencias nominales, de forma similar a como se hizo en el caso de los términos obtenidos a partir de la consulta.

De nuevo, el primer paso consiste en calcular los parámetros del sistema. Tras una primera fase de puesta a punto, —tablas F.17 a F.20 del apéndice F—, se tomaron los siguientes valores:

Consultas cortas: n'_1=10, t'=40, ω=3 *(ratio* $1/\omega$*=0.333)*

Consultas largas: n'_1=5, t'=50, ω=2 *(ratio* $1/\omega$*=0.500)*

[12]Téngase en cuenta que un mismo par puede ser generado simultáneamente por varios tipos de dependencia, siendo computado para cada uno de ellos.

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	*DND*	*%Δ*	*lem*	*DND*	*%Δ*	*lem*	*DND*	*%Δ*	*lem*	*DND*	*%Δ*	*lem*	*DND*	*%Δ*	*lem*	*DND*	*%Δ*
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45000	45000	–	45000	45000	–	47000	47000	–	47000	47000	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2700	**2756**	**2.07**	2762	2760	-0.07	2363	**2387**	**1.02**	2410	**2415**	**0.21**	2159	**2239**	**3.71**	2175	**2236**	**2.80**
Pr. no int.	.4829	**.5190**	**7.48**	.5239	**.5501**	**5.00**	.4678	**.4912**	**5.00**	.5256	**.5451**	**3.71**	.4324	**.4616**	**6.75**	.4659	**.5032**	**8.01**
Pr. doc.	.5327	**.5761**	**8.15**	.5690	**.5866**	**3.09**	.5667	**.5840**	**3.05**	.6089	**.6255**	**2.73**	.4724	**.5279**	**11.75**	.5201	**.5680**	**9.21**
R-pr.	.4848	**.5001**	**3.16**	.5075	**.5239**	**3.23**	.4549	**.4621**	**1.58**	.5030	**.5214**	**3.66**	.4362	**.4608**	**5.64**	.4493	**.4738**	**5.45**
Pr. a 0 %	.8293	**.8584**	**3.51**	.8845	.8718	-1.44	.8234	.8032	-2.45	.8763	.8424	-3.87	.8124	.7943	-2.23	.8366	.8080	-3.42
Pr. a 10 %	.7463	**.7981**	**6.94**	.7914	**.8209**	**3.73**	.6845	**.6872**	**0.39**	.7517	**.7602**	**1.13**	.7057	**.7184**	**1.80**	.7197	**.7586**	**5.41**
Pr. a 20 %	.6771	**.7184**	**6.10**	.7136	**.7589**	**6.35**	.6287	**.6540**	**4.02**	.7095	**.7230**	**1.90**	.6118	**.6421**	**4.95**	.6453	**.6870**	**6.46**
Pr. a 30 %	.6138	**.6377**	**3.89**	.6591	**.6867**	**4.19**	.5829	**.6090**	**4.48**	.6463	**.6623**	**2.48**	.5503	**.5744**	**4.38**	.5788	**.6137**	**6.03**
Pr. a 40 %	.5502	**.5811**	**5.62**	.6085	**.6356**	**4.45**	.5555	**.5746**	**3.44**	.6100	**.6315**	**3.52**	.5059	**.5418**	**7.10**	.5348	**.5791**	**8.28**
Pr. a 50 %	.4931	**.5361**	**8.72**	.5557	**.5978**	**7.58**	.5136	**.5328**	**3.74**	.5658	**.5983**	**5.74**	.4657	**.5034**	**8.10**	.4889	**.5368**	**9.80**
Pr. a 60 %	.4496	**.4836**	**7.56**	.5006	**.5218**	**4.23**	.4413	**.4774**	**8.18**	.4977	**.5327**	**7.03**	.4017	**.4440**	**10.53**	.4339	**.4814**	**10.95**
Pr. a 70 %	.3853	**.4173**	**8.31**	.4161	**.4367**	**4.95**	.3943	**.4341**	**10.09**	.4569	**.4961**	**8.58**	.3422	**.3786**	**10.64**	.3789	**.4328**	**14.23**
Pr. a 80 %	.3277	**.3548**	**8.27**	.3509	**.3652**	**4.08**	.3122	**.3406**	**9.10**	.3679	**.3937**	**7.01**	.2700	**.3118**	**15.48**	.3206	**.3588**	**11.92**
Pr. a 90 %	.2356	**.2721**	**15.49**	.2492	**.2775**	**11.36**	.2319	**.2563**	**10.52**	.2845	**.3034**	**6.64**	.1821	**.2225**	**22.19**	.2067	**.2490**	**20.46**
Pr. a 100 %	.1197	**.1419**	**18.55**	.1289	**.1468**	**13.89**	.1209	**.1384**	**14.47**	.1547	**.1711**	**10.60**	.0932	**.1021**	**9.55**	.1164	**.1287**	**10.57**
Pr. a 5 docs.	.6609	**.7043**	**6.57**	.6957	**.7304**	**4.99**	.5956	**.6400**	**7.45**	.6844	.6800	-0.64	.5915	**.6085**	**2.87**	.6213	**.6596**	**6.16**
Pr. a 10 docs.	.6283	**.6717**	**6.91**	.6543	**.6783**	**3.67**	.5667	.5667	0.00	.6000	**.6311**	**5.18**	.5149	**.5681**	**10.33**	.5596	**.6000**	**7.22**
Pr. a 15 docs.	.5928	**.6203**	**4.64**	.6188	**.6348**	**2.59**	.5170	**.5244**	**1.43**	.5689	**.5867**	**3.13**	.4738	**.5220**	**10.17**	.5106	**.5589**	**9.46**
Pr. a 20 docs.	.5446	**.5837**	**7.18**	.5880	**.6000**	**2.04**	.4878	**.4956**	**1.60**	.5422	**.5478**	**1.03**	.4457	**.4787**	**7.40**	.4819	**.5138**	**6.62**
Pr. a 30 docs.	.4928	**.5326**	**8.08**	.5304	**.5457**	**2.88**	.4452	**.4504**	**1.17**	.4948	**.5074**	**2.55**	.4000	**.4333**	**8.32**	.4255	**.4532**	**6.51**
Pr. a 100 docs.	.3300	**.3470**	**5.15**	.3509	**.3567**	**1.65**	.2993	**.3038**	**1.50**	.3147	**.3213**	**2.10**	.2513	**.2672**	**6.33**	.2702	**.2760**	**2.15**
Pr. a 200 docs.	.2234	**.2315**	**3.63**	.2315	**.2380**	**2.81**	.1997	**.2037**	**2.00**	.2093	**.2106**	**0.62**	.1617	**.1713**	**5.94**	.1694	**.1788**	**5.55**
Pr. a 500 docs.	.1090	**.1123**	**3.03**	.1115	**.1131**	**1.43**	.0983	**.1000**	**1.73**	.1023	.1023	0.00	.0827	**.0865**	**4.59**	.0842	**.0874**	**3.80**
Pr. a 1000 docs.	.0587	**.0599**	**2.04**	.0600	.0600	0.00	.0525	**.0530**	**0.95**	.0536	**.0537**	**0.19**	.0459	**.0476**	**3.70**	.0463	**.0476**	**2.81**

Tabla 7.13: Resultados obtenidos mediante lematización (*lem*), caso base, y pares de dependencia sintáctica correspondientes a sintagmas nominales obtenidos a partir de los documentos (*DND*)

Los resultados obtenidos, si bien positivos, al superar tanto a los términos simples lematizados —tabla 7.13—, como a su contrapartida basada en los pares de la consulta —tabla 7.15—, no son tan buenos como aquéllos obtenidos empleando la totalidad de las dependencias. Esta diferencia de resultados se muestra claramente al comparar explícitamente ambos casos en la tabla 7.14. Esta diferencia es mayor que la obtenida en el caso de emplear la información sintáctica de la consultas, donde las diferencias obtenidas al emplear únicamente dependencias procedentes de sintagmas nominales eran menores.

En lo referente a otros trabajos, la realimentación también es empleada por Khan y Khor [126] para la selección de frases nominales. Sin embargo, nuestros estudios son más completos, al haber empleado en nuestros experimentos un número mucho mayor de consultas y no un conjunto ad-hoc como en su caso.

7.6.3. Resultados Aplicando Realimentación

Dado su superior rendimiento, a lo largo de este apartado emplearemos los pares de dependencia de todos los tipos de dependencias, y no sólo aquéllas asociados a frases nominales.

Los parámetros empleados para la expansión automática son, de nuevo, los usuales. En contra de lo esperado, la introducción de realimentación conlleva en este caso una caída del rendimiento, tal y como se refleja en la tabla 7.16. Este descenso se debe, probablemente, a un ajuste inadecuado de los pesos correspondientes a los términos complejos, ya que durante el proceso de realimentación su peso es reajustado de igual modo que con los términos simples, cuando su tratamiento debería ser diferenciado por los problemas de sobreponderación y violación de la suposición de independencia descritos en el apartado 7.4.5.

Por tanto, seguiremos tomando como referencia la propuesta inicial sin realimentación. Sin embargo, resultaría de interés comprobar si su comportamiento sigue siendo superior al caso del empleo de términos simples con realimentación y al caso de empleo de información sintáctica de las consultas con realimentación. Para ello se han incluido las tablas comparativas correspondientes.

En primer lugar, en la tabla 7.17 se compara el comportamiento de la información sintáctica procendente de los documentos —sin realimentación— con el caso del empleo exclusivo de términos simples con realimentación. Los resultados demuestran que el empleo de términos complejos, aún sin realimentación, continúa superando los resultados obtenidos empleando únicamente términos simples, mejorando la precisión de los primeros documentos devueltos y, en general, ampliándose esta mejora al resto de parámetros de evalución para el caso de las consultas cortas.

Por su parte, la tabla 7.18 recoge los resultados para el segundo supuesto, el empleo de información sintáctica de las consultas con realimentación. En este caso los resultados para los pares de los documentos —sin realimentación— no son, en general, tan buenos como en el caso de los pares de las consultas con realimentación, si bien los documentos devueltos en primer lugar siguen siendo más precisos, y el rendimiento para la precisión a los n documentos es particularmente bueno en el caso del CORPUS 2003.

7.7. Discusión

A lo largo de este capítulo hemos planteado la utilización de dependencias sintácticas para la extracción de términos índice complejos como complemento a los términos índice simples en la Recuperación de Información. El objetivo perseguido era el de tratar los problemas derivados de la variación lingüística de carácter sintáctico y morfosintáctico del español y, de este modo, obtener resultados más precisos.

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*DSD*	*DND*	*%Δ*	*DSD*	*DND*	*%Δ*	*DSD*	*DND*	*%Δ*	*DSD*	*DND*	*%Δ*	*DSD*	*DND*	*%Δ*	*DSD*	*DND*	*%Δ*
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2758	2756	-0.07	2765	2760	-0.18	2394	2387	-0.29	2417	2415	-0.08	2246	2239	-0.31	2240	2236	-0.18
Pr. no int.	.5286	.5190	-1.82	.5577	.5501	-1.36	.4990	.4912	-1.56	.5475	.5451	-0.44	.4690	.4616	-1.58	.5015	**.5032**	**0.34**
Pr. doc.	.5768	.5761	-0.12	.5902	.5866	-0.61	.5892	.5840	-0.88	.6286	.6255	-0.49	.5432	.5279	-2.82	.5666	**.5680**	**0.25**
R-pr.	.5119	.5001	-2.31	.5274	.5239	-0.66	.4765	.4621	-3.02	.5267	.5214	-1.01	.4582	**.4608**	**0.57**	.4794	.4738	-1.17
Pr. a 0 %	.8389	**.8584**	**2.32**	.8759	.8718	-0.47	.8033	.8032	-0.01	.8298	**.8424**	**1.52**	.7964	.7943	-0.26	.8035	**.8080**	**0.56**
Pr. a 10 %	.7794	**.7981**	**2.40**	.8233	.8209	-0.29	.6990	.6872	-1.69	.7622	.7602	-0.26	.7300	.7184	-1.59	.7467	**.7586**	**1.59**
Pr. a 20 %	.7244	.7184	-0.83	.7651	.7589	-0.81	.6681	.6540	-2.11	.7209	**.7230**	**0.29**	.6564	.6421	-2.18	.6782	**.6870**	**1.30**
Pr. a 30 %	.6497	.6377	-1.85	.7016	.6867	-2.12	.6224	.6090	-2.15	.6648	.6623	-0.38	.5818	.5744	-1.27	.6209	.6137	-1.16
Pr. a 40 %	.5926	.5811	-1.94	.6431	.6356	-1.17	.5860	.5746	-1.95	.6400	.6315	-1.33	.5499	.5418	-1.47	.5848	.5791	-0.97
Pr. a 50 %	.5534	.5361	-3.13	.5988	.5978	-0.17	.5535	.5328	-3.74	.6048	.5983	-1.07	.5046	.5034	-0.24	.5355	**.5368**	**0.24**
Pr. a 60 %	.5005	.4836	-3.38	.5326	.5218	-2.03	.4779	.4774	-0.10	.5448	.5327	-2.22	.4439	**.4440**	**0.02**	.4798	**.4814**	**0.33**
Pr. a 70 %	.4343	.4173	-3.91	.4468	.4367	-2.26	.4371	.4341	-0.69	.5004	.4961	-0.86	.3959	.3786	-4.37	.4308	**.4328**	**0.46**
Pr. a 80 %	.3678	.3548	-3.53	.3710	.3652	-1.56	.3406	.3406	0.00	.3901	**.3937**	**0.92**	.3166	.3118	-1.52	.3704	.3588	-3.13
Pr. a 90 %	.2775	.2721	-1.95	.2807	.2775	-1.14	.2581	.2563	-0.70	.3058	.3034	-0.78	.2303	.2225	-3.39	.2559	.2490	-2.70
Pr. a 100 %	.1522	.1419	-6.77	.1483	.1468	-1.01	.1434	.1384	-3.49	.1750	.1711	-2.23	.1100	.1021	-7.18	.1278	**.1287**	**0.70**
Pr. a 5 docs.	.7000	**.7043**	**0.61**	.7304	.7304	0.00	.6533	.6400	-2.04	.7022	.6800	-3.16	.6128	.6085	-0.70	.6468	**.6596**	**1.98**
Pr. a 10 docs.	.6717	.6717	0.00	.6913	.6783	-1.88	.5800	.5667	-2.29	.6356	.6311	-0.71	.5745	.5681	-1.11	.6149	.6000	-2.42
Pr. a 15 docs.	.6203	.6203	0.00	.6420	.6348	-1.12	.5363	.5244	-2.22	.5852	**.5867**	**0.26**	.5390	.5220	-3.15	.5504	**.5589**	**1.54**
Pr. a 20 docs.	.5935	.5837	-1.65	.6109	.6000	-1.78	.5056	.4956	-1.98	.5533	.5478	-0.99	.4957	.4787	-3.43	.5064	**.5138**	**1.46**
Pr. a 30 docs.	.5348	.5326	-0.41	.5529	.5457	-1.30	.4607	.4504	-2.24	.5059	**.5074**	**0.30**	.4468	.4333	-3.02	.4560	.4532	-0.61
Pr. a 100 docs.	.3474	.3470	-0.12	.3578	.3567	-0.31	.3078	.3038	-1.30	.3267	.3213	-1.65	.2719	.2672	-1.73	.2764	.2760	-0.14
Pr. a 200 docs.	.2336	.2315	-0.90	.2373	**.2380**	**0.29**	.2049	.2037	-0.59	.2111	.2106	-0.24	.1739	.1713	-1.50	.1789	.1788	-0.06
Pr. a 500 docs.	.1117	**.1123**	**0.54**	.1130	**.1131**	**0.09**	.1002	.1000	-0.20	.1017	**.1023**	**0.59**	.0875	.0865	-1.14	.0876	.0874	-0.23
Pr. a 1000 docs.	.0600	.0599	-0.17	.0601	.0600	-0.17	.0532	.0530	-0.38	.0537	.0537	0.00	.0478	.0476	-0.42	.0477	.0476	-0.21

Tabla 7.14: Resultados obtenidos mediante pares de dependencia sintáctica obtenidos a partir de los documentos empleando la totalidad de las dependencias (*DSD*) o sólo aquéllas correspondientes a sintagmas nominales (*DND*)

Tabla 7.15: Resultados obtenidos mediante pares de dependencia sintáctica correspondientes a sintagmas nominales obtenidos a partir de la consulta (*SNF*) y a partir de los documentos (*DND*)

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*SND*	*DSD*	*%Δ*	*SND*	*DSD*	*%Δ*	*SND*	*DSD*	*%Δ*	*SND*	*DSD*	*%Δ*	*SND*	*DSD*	*%Δ*	*SND*	*DSD*	*%Δ*
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2726	**2756**	**1.10**	2781	2760	-0.76	2368	**2387**	**0.80**	2413	**2415**	**0.08**	2172	**2239**	**3.08**	2197	**2236**	**1.78**
Pr. no int.	.4940	**.5190**	**5.06**	.5303	**.5501**	**3.73**	.4752	**.4912**	**3.37**	.5371	**.5451**	**1.49**	.4324	**.4616**	**6.75**	.4699	**.5032**	**7.09**
Pr. doc.	.5457	**.5761**	**5.57**	.5791	**.5866**	**1.30**	.5683	**.5840**	**2.76**	.6191	**.6255**	**1.03**	.4675	**.5279**	**12.92**	.5254	**.5680**	**8.11**
R-pr.	.4858	**.5001**	**2.94**	.5128	**.5239**	**2.16**	.4635	.4621	-0.30	.5126	**.5214**	**1.72**	.4529	**.4608**	**1.74**	.4575	**.4738**	**3.56**
Pr. a 0 %	.8498	**.8584**	**1.01**	.8617	**.8718**	**1.17**	.8041	.8032	-0.11	.8759	.8424	-3.82	.8085	.7943	-1.76	.8446	.8080	-4.33
Pr. a 10 %	.7628	**.7981**	**4.63**	.8020	**.8209**	**2.36**	.7076	.6872	-2.88	.7806	.7602	-2.61	.6775	**.7184**	**6.04**	.7355	**.7586**	**3.14**
Pr. a 20 %	.6878	**.7184**	**4.45**	.7277	**.7589**	**4.29**	.6547	.6540	-0.11	.7185	**.7230**	**0.63**	.6259	**.6421**	**2.59**	.6511	**.6870**	**5.51**
Pr. a 30 %	.6273	**.6377**	**1.66**	.6678	**.6867**	**2.83**	.5844	**.6090**	**4.21**	.6524	**.6623**	**1.52**	.5556	**.5744**	**3.38**	.5909	**.6137**	**3.86**
Pr. a 40 %	.5614	**.5811**	**3.51**	.6124	**.6356**	**3.79**	.5530	**.5746**	**3.91**	.6172	**.6315**	**2.32**	.4991	**.5418**	**8.56**	.5324	**.5791**	**8.77**
Pr. a 50 %	.5036	**.5361**	**6.45**	.5619	**.5978**	**6.39**	.5192	**.5328**	**2.62**	.5758	**.5983**	**3.91**	.4599	**.5034**	**9.46**	.4942	**.5368**	**8.62**
Pr. a 60 %	.4618	**.4836**	**4.72**	.5128	**.5218**	**1.76**	.4526	**.4774**	**5.48**	.5097	**.5327**	**4.51**	.4049	**.4440**	**9.66**	.4400	**.4814**	**9.41**
Pr. a 70 %	.3881	**.4173**	**7.52**	.4248	**.4367**	**2.80**	.4002	**.4341**	**8.47**	.4631	**.4961**	**7.13**	.3380	**.3786**	**12.01**	.3917	**.4328**	**10.49**
Pr. a 80 %	.3358	**.3548**	**5.66**	.3587	**.3652**	**1.81**	.3219	**.3406**	**5.81**	.3833	**.3937**	**2.71**	.2745	**.3118**	**13.59**	.3234	**.3588**	**10.95**
Pr. a 90 %	.2477	**.2721**	**9.85**	.2568	**.2775**	**8.06**	.2381	**.2563**	**7.64**	.2965	**.3034**	**2.33**	.1889	**.2225**	**17.79**	.2084	**.2490**	**19.48**
Pr. a 100 %	.1199	**.1419**	**18.35**	.1319	**.1468**	**11.30**	.1280	**.1384**	**8.12**	.1581	**.1711**	**8.22**	.0994	**.1021**	**2.72**	.1175	**.1287**	**9.53**
Pr. a 5 docs.	.6870	**.7043**	**2.52**	.6957	**.7304**	**4.99**	.6222	**.6400**	**2.86**	.6889	.6800	-1.29	.5915	**.6085**	**2.87**	.6255	**.6596**	**5.45**
Pr. a 10 docs.	.6522	**.6717**	**2.99**	.6761	**.6783**	**0.33**	.5778	.5667	-1.92	.6200	**.6311**	**1.79**	.5277	**.5681**	**7.66**	.5660	**.6000**	**6.01**
Pr. a 15 docs.	.5971	**.6203**	**3.89**	.6304	**.6348**	**0.70**	.5333	.5244	-1.67	.5793	**.5867**	**1.28**	.4738	**.5220**	**10.17**	.5092	**.5589**	**9.76**
Pr. a 20 docs.	.5587	**.5837**	**4.47**	.5935	**.6000**	**1.10**	.4989	.4956	-0.66	.5467	**.5478**	**0.20**	.4362	**.4787**	**9.74**	.4723	**.5138**	**8.79**
Pr. a 30 docs.	.5036	**.5326**	**5.76**	.5391	**.5457**	**1.22**	.4533	.4504	-0.64	.5044	**.5074**	**0.59**	.3922	**.4333**	**10.48**	.4326	**.4532**	**4.76**
Pr. a 100 docs.	.3361	**.3470**	**3.24**	.3517	**.3567**	**1.42**	.2987	**.3038**	**1.71**	.3184	**.3213**	**0.91**	.2513	**.2672**	**6.33**	.2730	**.2760**	**1.10**
Pr. a 200 docs.	.2257	**.2315**	**2.57**	.2341	**.2380**	**1.67**	.2000	**.2037**	**1.85**	.2116	.2106	-0.47	.1622	**.1713**	**5.61**	.1703	**.1788**	**4.99**
Pr. a 500 docs.	.1101	**.1123**	**2.00**	.1123	**.1131**	**0.71**	.0989	**.1000**	**1.11**	.1029	.1023	-0.58	.0828	**.0865**	**4.47**	.0845	**.0874**	**3.43**
Pr. a 1000 docs.	.0593	**.0599**	**1.01**	.0605	.0600	-0.83	.0526	**.0530**	**0.76**	.0536	**.0537**	**0.19**	.0462	**.0476**	**3.03**	.0467	**.0476**	**1.93**

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*nof*	*sif*	%Δ	*nof*	*sif*	%Δ	*nof*	*sif*	%Δ	*nof*	*sif*	%Δ	*nof*	*sif*	%Δ	*nof*	*sif*	%Δ
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2758	**2780**	**0.80**	2765	2748	-0.61	2394	2386	-0.33	2417	2410	-0.29	2246	2243	-0.13	2240	2221	-0.85
Pr. no int.	.5286	**.5382**	**1.82**	.5577	.5550	-0.48	.4990	.4929	-1.22	.5475	.5374	-1.84	.4690	**.4996**	**6.52**	.5015	**.5198**	**3.65**
Pr. doc.	.5768	.5600	-2.91	.5902	.5610	-4.95	.5892	.5644	-4.21	.6286	.5992	-4.68	.5432	**.5750**	**5.85**	.5666	**.5870**	**3.60**
R-pr.	.5119	.5097	-0.43	.5274	**.5275**	**0.02**	.4765	.4599	-3.48	.5267	.5040	-4.31	.4582	**.4720**	**3.01**	.4794	**.4806**	**0.25**
Pr. a 0 %	.8389	.8371	-0.21	.8759	.8747	-0.14	.8033	.7948	-1.06	.8298	.8246	-0.63	.7964	.7957	-0.09	.8035	.7921	-1.42
Pr. a 10 %	.7794	.7639	-1.99	.8233	.8187	-0.56	.6990	.6915	-1.07	.7622	.7407	-2.82	.7300	.7264	-0.49	.7467	.7416	-0.68
Pr. a 20 %	.7244	.7143	-1.39	.7651	.7393	-3.37	.6681	.6554	-1.90	.7209	.7001	-2.89	.6564	.6472	-1.40	.6782	.6675	-1.58
Pr. a 30 %	.6497	**.6667**	**2.62**	.7016	.6857	-2.27	.6224	.5933	-4.68	.6648	.6424	-3.37	.5818	**.5852**	**0.58**	.6209	**.6255**	**0.74**
Pr. a 40 %	.5926	**.6089**	**2.75**	.6431	.6221	-3.27	.5860	.5633	-3.87	.6400	.6223	-2.77	.5499	**.5540**	**0.75**	.5848	.5832	-0.27
Pr. a 50 %	.5534	**.5759**	**4.07**	.5988	.5784	-3.41	.5535	.5371	-2.96	.6048	.5942	-1.75	.5046	**.5291**	**4.86**	.5355	**.5421**	**1.23**
Pr. a 60 %	.5005	**.5220**	**4.30**	.5326	.5296	-0.56	.4779	**.4868**	**1.86**	.5448	.5352	-1.76	.4439	**.4798**	**8.09**	.4798	**.4863**	**1.35**
Pr. a 70 %	.4343	**.4560**	**5.00**	.4468	**.4693**	**5.04**	.4371	.4309	-1.42	.5004	.4855	-2.98	.3959	**.4418**	**11.59**	.4308	**.4531**	**5.18**
Pr. a 80 %	.3678	**.3802**	**3.37**	.3710	**.3894**	**4.96**	.3406	**.3522**	**3.41**	.3901	**.3912**	**0.28**	.3166	**.3817**	**20.56**	.3704	**.4070**	**9.88**
Pr. a 90 %	.2775	**.3066**	**10.49**	.2807	**.3140**	**11.86**	.2581	.2481	-3.87	.3058	.2863	-6.38	.2303	**.3174**	**37.82**	.2559	**.3436**	**34.27**
Pr. a 100 %	.1522	**.1804**	**18.53**	.1483	**.1789**	**20.63**	.1434	.1417	-1.19	.1750	.1622	-7.31	.1100	**.1960**	**78.18**	.1278	**.2126**	**66.35**
Pr. a 5 docs.	.7000	**.7043**	**0.61**	.7304	.7304	0.00	.6533	.6533	0.00	.7022	.7022	0.00	.6128	.6128	0.00	.6468	.6468	0.00
Pr. a 10 docs.	.6717	**.6804**	**1.30**	.6913	.6717	-2.84	.5800	.5756	-0.76	.6356	.6333	-0.36	.5745	.5702	-0.75	.6149	.6106	-0.70
Pr. a 15 docs.	.6203	**.6362**	**2.56**	.6420	**.6449**	**0.45**	.5363	**.5422**	**1.10**	.5852	**.5881**	**0.50**	.5390	**.5489**	**1.84**	.5504	**.5688**	**3.34**
Pr. a 20 docs.	.5935	**.6000**	**1.10**	.6109	.6098	-0.18	.5056	**.5067**	**0.22**	.5533	.5444	-1.61	.4957	**.5043**	**1.73**	.5064	**.5138**	**1.46**
Pr. a 30 docs.	.5348	.5319	-0.54	.5529	.5500	-0.52	.4607	.4548	-1.28	.5059	.4970	-1.76	.4468	**.4596**	**2.86**	.4560	**.4582**	**0.48**
Pr. a 100 docs.	.3474	.3372	-2.94	.3578	.3426	-4.25	.3078	.2964	-3.70	.3267	.3142	-3.83	.2719	**.2838**	**4.38**	.2764	**.2802**	**1.37**
Pr. a 200 docs.	.2336	.2258	-3.34	.2373	.2265	-4.55	.2049	.1986	-3.07	.2111	.2033	-3.69	.1739	**.1791**	**2.99**	.1789	**.1795**	**0.34**
Pr. a 500 docs.	.1117	.1114	-0.27	.1130	.1108	-1.95	.1002	.0982	-2.00	.1017	.1004	-1.28	.0875	**.0879**	**0.46**	.0876	.0876	0.00
Pr. a 1000 docs.	.0600	**.0604**	**0.67**	.0601	.0597	-0.67	.0532	.0530	-0.38	.0537	.0536	-0.19	.0478	.0477	-0.21	.0477	.0473	-0.84

Tabla 7.16: Resultados obtenidos mediante pares de dependencia sintáctica obtenidos a partir de los documentos empleando (*sif*) o no (*nof*) realimentación

Tabla 7.17: Resultados obtenidos mediante lematización aplicando realimentación (*lem*), caso base, y pares de dependencia sintáctica obtenidos a partir de los documentos sin realimentación (*DSD*)

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	*DSD*	*%Δ*	*lem*	*DSD*	*%Δ*	*lem*	*DSD*	*%Δ*	*lem*	*DSD*	*%Δ*	*lem*	*DSD*	*%Δ*	*lem*	*DSD*	*%Δ*
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2767	2758	-0.33	2779	2765	-0.50	2376	**2394**	**0.76**	2406	**2417**	**0.46**	2240	**2246**	**0.27**	2223	**2240**	**0.76**
Pr. no int.	.5220	**.5286**	**1.26**	.5604	.5577	-0.48	.4773	**.4990**	**4.55**	.5392	**.5475**	**1.54**	.5024	.4690	-6.65	.5207	.5015	-3.69
Pr. doc.	.5784	.5768	-0.28	.5912	.5902	-0.17	.5681	**.5892**	**3.71**	.6145	**.6286**	**2.29**	.5530	.5432	-1.77	.5802	.5666	-2.34
R-pr.	.4990	**.5119**	**2.59**	.5366	.5274	-1.71	.4599	**.4765**	**3.61**	.5104	**.5267**	**3.19**	.4912	.4582	-6.72	.4871	.4794	-1.58
Pr. a 0 %	.8221	**.8389**	**2.04**	.8895	.8759	-1.53	.8210	.8033	-2.16	.8710	.8298	-4.73	.8145	.7964	-2.22	.8301	.8035	-3.20
Pr. a 10 %	.7490	**.7794**	**4.06**	.8028	**.8233**	**2.55**	.6861	**.6990**	**1.88**	.7619	**.7622**	**0.04**	.7369	.7300	-0.94	.7421	**.7467**	**0.62**
Pr. a 20 %	.6866	**.7244**	**5.51**	.7352	**.7651**	**4.07**	.6319	**.6681**	**5.73**	.6929	**.7209**	**4.04**	.6632	.6564	-1.03	.6758	**.6782**	**0.36**
Pr. a 30 %	.6573	.6497	-1.16	.6996	**.7016**	**0.29**	.5688	**.6224**	**9.42**	.6497	**.6648**	**2.32**	.6019	.5818	-3.34	.6304	.6209	-1.51
Pr. a 40 %	.5997	.5926	-1.18	.6541	.6431	-1.68	.5289	**.5860**	**10.80**	.6202	**.6400**	**3.19**	.5638	.5499	-2.47	.5975	.5848	-2.13
Pr. a 50 %	.5456	**.5534**	**1.43**	.6005	.5988	-0.28	.5017	**.5535**	**10.32**	.5733	**.6048**	**5.49**	.5410	.5046	-6.73	.5479	.5355	-2.26
Pr. a 60 %	.4994	**.5005**	**0.22**	.5386	.5326	-1.11	.4623	**.4779**	**3.37**	.5194	**.5448**	**4.89**	.4783	.4439	-7.19	.4938	.4798	-2.84
Pr. a 70 %	.4375	.4343	-0.73	.4621	.4468	-3.31	.4255	**.4371**	**2.73**	.4862	**.5004**	**2.92**	.4366	.3959	-9.32	.4587	.4308	-6.08
Pr. a 80 %	.3661	**.3678**	**0.46**	.3910	.3710	-5.12	.3384	**.3406**	**0.65**	.3753	**.3901**	**3.94**	.3788	.3166	-16.42	.3891	.3704	-4.81
Pr. a 90 %	.2939	.2775	-5.58	.3130	.2807	-10.32	.2651	.2581	-2.64	.3059	.3058	-0.03	.3102	.2303	-25.76	.3230	.2559	-20.77
Pr. a 100 %	.1547	.1522	-1.62	.1624	.1483	-8.68	.1395	**.1434**	**2.80**	.1704	**.1750**	**2.70**	.1871	.1100	-41.21	.1928	.1278	-33.71
Pr. a 5 docs.	.6609	**.7000**	**5.92**	.6957	**.7304**	**4.99**	.5956	**.6533**	**9.69**	.6844	**.7022**	**2.60**	.5872	**.6128**	**4.36**	.6213	**.6468**	**4.10**
Pr. a 10 docs.	.6457	**.6717**	**4.03**	.6848	**.6913**	**0.95**	.5600	**.5800**	**3.57**	.6178	**.6356**	**2.88**	.5596	**.5745**	**2.66**	.5872	**.6149**	**4.72**
Pr. a 15 docs.	.5884	**.6203**	**5.42**	.6435	.6420	-0.23	.5274	**.5363**	**1.69**	.5822	**.5852**	**0.52**	.5305	**.5390**	**1.60**	.5504	.5504	0.00
Pr. a 20 docs.	.5630	**.5935**	**5.42**	.6043	**.6109**	**1.09**	.5011	**.5056**	**0.90**	.5533	.5533	0.00	.4883	**.4957**	**1.52**	.5266	.5064	-3.84
Pr. a 30 docs.	.5225	**.5348**	**2.35**	.5580	.5529	-0.91	.4444	**.4607**	**3.67**	.5081	.5059	-0.43	.4433	**.4468**	**0.79**	.4667	.4560	-2.29
Pr. a 100 docs.	.3507	.3474	-0.94	.3598	.3578	-0.56	.2940	**.3078**	**4.69**	.3191	**.3267**	**2.38**	.2770	.2719	-1.84	.2853	.2764	-3.12
Pr. a 200 docs.	.2348	.2336	-0.51	.2361	**.2373**	**0.51**	.1979	**.2049**	**3.54**	.2067	**.2111**	**2.13**	.1753	.1739	-0.80	.1791	.1789	-0.11
Pr. a 500 docs.	.1122	.1117	-0.45	.1121	**.1130**	**0.80**	.0980	**.1002**	**2.24**	.1008	**.1017**	**0.89**	.0869	**.0875**	**0.69**	.0871	**.0876**	**0.57**
Pr. a 1000 docs.	.0602	.0600	-0.33	.0604	.0601	-0.50	.0528	**.0532**	**0.76**	.0535	**.0537**	**0.37**	.0477	**.0478**	**0.21**	.0473	**.0477**	**0.85**

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	***FNF***	***DSD***	**%Δ**	***FNF***	***DSD***	**%Δ**	***FNF***	***DSD***	**%Δ**	***FNF***	***DSD***	**%Δ**	***FNF***	***DSD***	**%Δ**	***FNF***	***DSD***	**%Δ**
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2811	2758	-1.89	2775	2765	-0.36	2383	**2394**	**0.46**	2419	2417	-0.08	2251	2246	-0.22	2222	**2240**	**0.81**
Pr. no int.	.5434	.5286	-2.72	.5770	.5577	-3.34	.4993	.4990	-0.06	.5629	.5475	-2.74	.5037	.4690	-6.89	.5241	.5015	-4.31
Pr. doc.	.6045	.5768	-4.58	.6059	.5902	-2.59	.5754	**.5892**	**2.40**	.6321	.6286	-0.55	.5542	.5432	-1.98	.5852	.5666	-3.18
R-pr.	.5210	.5119	-1.75	.5693	.5274	-7.36	.4880	.4765	-2.36	.5207	**.5267**	**1.15**	.4773	.4582	-4.00	.4975	.4794	-3.64
Pr. a 0 %	.8432	.8389	-0.51	.8755	**.8759**	**0.05**	.8138	.8033	-1.29	.8657	.8298	-4.15	.8425	.7964	-5.47	.8301	.8035	-3.20
Pr. a 10 %	.7786	**.7794**	**0.10**	.8127	**.8233**	**1.30**	.7164	.6990	-2.43	.7724	.7622	-1.32	.7168	**.7300**	**1.84**	.7529	.7467	-0.82
Pr. a 20 %	.7112	**.7244**	**1.86**	.7523	**.7651**	**1.70**	.6675	**.6681**	**0.09**	.7297	.7209	-1.21	.6526	**.6564**	**0.58**	.6955	.6782	-2.49
Pr. a 30 %	.6723	.6497	-3.36	.7107	.7016	-1.28	.6219	**.6224**	**0.08**	.6725	.6648	-1.14	.6107	.5818	-4.73	.6345	.6209	-2.14
Pr. a 40 %	.6257	.5926	-5.29	.6699	.6431	-4.00	.5545	**.5860**	**5.68**	.6380	**.6400**	**0.31**	.5579	.5499	-1.43	.5780	**.5848**	**1.18**
Pr. a 50 %	.5864	.5534	-5.63	.6316	.5988	-5.19	.5216	**.5535**	**6.12**	.5995	**.6048**	**0.88**	.5378	.5046	-6.17	.5345	**.5355**	**0.19**
Pr. a 60 %	.5360	.5005	-6.62	.5720	.5326	-6.89	.4704	**.4779**	**1.59**	.5475	.5448	-0.49	.4877	.4439	-8.98	.5016	.4798	-4.35
Pr. a 70 %	.4540	.4343	-4.34	.4999	.4468	-10.62	.4409	.4371	-0.86	.5109	.5004	-2.06	.4440	.3959	-10.83	.4545	.4308	-5.21
Pr. a 80 %	.3811	.3678	-3.49	.4121	.3710	-9.97	.3653	.3406	-6.76	.4137	.3901	-5.70	.3908	.3166	-18.99	.4053	.3704	-8.61
Pr. a 90 %	.2962	.2775	-6.31	.3237	.2807	-13.28	.2780	.2581	-7.16	.3212	.3058	-4.79	.3212	.2303	-28.30	.3258	.2559	-21.45
Pr. a 100 %	.1664	.1522	-8.53	.1735	.1483	-14.52	.1444	.1434	-0.69	.1790	.1750	-2.23	.1900	.1100	-42.11	.2053	.1278	-37.75
Pr. a 5 docs.	.6913	**.7000**	**1.26**	.7043	**.7304**	**3.71**	.6178	**.6533**	**5.75**	.6933	**.7022**	**1.28**	.5745	**.6128**	**6.67**	.6383	**.6468**	**1.33**
Pr. a 10 docs.	.6739	.6717	-0.33	.7043	.6913	-1.85	.5933	.5800	-2.24	.6422	.6356	-1.03	.5468	**.5745**	**5.07**	.6064	**.6149**	**1.40**
Pr. a 15 docs.	.6246	.6203	-0.69	.6522	.6420	-1.56	.5437	.5363	-1.36	.6015	.5852	-2.71	.5135	**.5390**	**4.97**	.5489	**.5504**	**0.27**
Pr. a 20 docs.	.5946	.5935	-0.18	.6315	.6109	-3.26	.5156	.5056	-1.94	.5689	.5533	-2.74	.4872	**.4957**	**1.74**	.5202	.5064	-2.65
Pr. a 30 docs.	.5543	.5348	-3.52	.5739	.5529	-3.66	.4681	.4607	-1.58	.5207	.5059	-2.84	.4461	**.4468**	**0.16**	.4716	.4560	-3.31
Pr. a 100 docs.	.3550	.3474	-2.14	.3707	.3578	-3.48	.2987	**.3078**	**3.05**	.3249	**.3267**	**0.55**	.2768	.2719	-1.77	.2847	.2764	-2.92
Pr. a 200 docs.	.2357	.2336	-0.89	.2404	.2373	-1.29	.2011	**.2049**	**1.89**	.2104	**.2111**	**0.33**	.1783	.1739	-2.47	.1821	.1789	-1.76
Pr. a 500 docs.	.1135	.1117	-1.59	.1120	**.1130**	**0.89**	.0996	**.1002**	**0.60**	.1018	.1017	-0.10	.0888	.0875	-1.46	.0874	**.0876**	**0.23**
Pr. a 1000 docs.	.0611	.0600	-1.80	.0603	.0601	-0.33	.0530	**.0532**	**0.38**	.0538	.0537	-0.19	.0479	.0478	-0.21	.0473	**.0477**	**0.85**

Tabla 7.18: Resultados obtenidos mediante pares de dependencia sintáctica obtenidos a partir de la consulta y aplicando realimentación (*FNF*), caso base, y aquéllos obtenidos a partir de los documentos sin realimentación (*DSD*)

Para extraer dichas dependencias hemos desarrollado dos analizadores sintácticos superficiales del español de naturaleza bastante diferente: PATTERNS, basado en el reconocimiento de patrones, y CASCADE, basado en la emulación del análisis sintáctico completo mediante cascadas de traductores finitos. Si bien ambos nos permiten abordar el procesamiento de grandes colecciones de forma ágil, pues su complejidad es lineal respecto al tamaño del texto de entrada, las mejores características de CASCADE —mayor robustez, modularidad y mantenibilidad— lo capacitan mejor para nuestros propósitos.

Se han ensayado dos aproximaciones diferentes de acuerdo con el origen de la información sintáctica utilizada: la primera, empleando los pares obtenidos a partir de las consultas; la segunda, que ha resultado superior en su conjunto, empleando los pares obtenidos a partir de los documentos.

Por otra parte, se estudió también el impacto de los tipos de dependencias a utilizar, comparando el rendimiento del sistema al emplear únicamente dependencias correspondientes a frases nominales o bien empleando la totalidad de éstas. A este respecto podríamos afirmar que, basándonos estrictamente en la precisión obtenida, la mejor opción pasa por emplear todas las dependencias, si bien se pueden reducir notablemente los costes de indexación y recuperación empleando únicamente dependencias nominales, obteniendo en compensación unos resultados a menudo inferiores.

Los resultados conseguidos nos permiten ser optimistas respecto a nuestro planteamiento, consistiendo el mayor desafío en determinar el mejor modo de utilizar la información sintáctica extraída por el analizador. Nuestras primeras aproximaciones, si bien simples, han sido positivas, por lo que es lógico suponer que técnicas más refinadas nos permitirían mejorar aún más estos resultados.

Apuntar también que existe una versión disponible del analizador PATTERNS para el gallego [31, 30], y que si bien no existe una versión tal del analizador CASCADE, su desarrollo no conllevaría excesivas complicaciones, bastando con adecuar la gramática al nuevo lenguaje.

Capítulo 8

Tratamiento de la Variación Sintáctica mediante un Modelo Basado en Localidad

8.1. Introducción

En el capítulo anterior abordamos la utilización del procesamiento sintáctico para hacer frente a la variación lingüística sintáctica presente en los textos, quedando patente que el empleo de este tipo de técnicas conlleva la necesidad de disponer de algún tipo de analizador sintáctico, para lo cual es necesario contar, a su vez, con una gramática apropiada, por sencilla que ésta sea. Sin embargo, aún en el caso de que dicha información sintáctica pueda ser convenientemente extraída de los textos, persiste todavía el problema de cómo incorporar dicha información al sistema. Como hemos podido comprobar, la aproximación más común, consistente en una combinación ponderada de términos simples y términos multipalabra —constituidos por términos simples relacionados sintácticamente entre sí—, no logra siempre resolver adecuadamente los problemas producidos por la sobrevaloración que el sistema tiende a dar a los términos complejos en detrimento de los términos simples [161].

En este contexto, el empleo de técnicas pseudo-sintácticas basadas en distancias entre términos se presenta como una alternativa práctica que evita dichos problemas, al no ser necesario el desarrollo de gramática o analizador alguno, y al integrar de modo consistente la información obtenida, tanto a nivel de la aparición de los términos en sí, como de su proximidad, frecuentemente ligada a la existencia de una relación sintáctica entre los mismos.

En este capítulo proponemos la utilización de un *modelo basado en localidad*, sustentado sobre similaridades basadas en distancias, como complemento a las técnicas clásicas de Recuperación de Información basadas en la indexación de términos simples, con el fin de incrementar la precisión de los documentos devueltos por el sistema.

8.2. Recuperación de Información Basada en Localidad

8.2.1. Modelo de Recuperación

En el modelo de recuperación imperante entre los sistemas de RI, el denominado *basado en documentos*, el usuario solicita del sistema los documentos relevantes a su necesidad de información. Frente a ello, el modelo *basado en localidad* propuesto por de Kretser y Moffat [63, 62, 61] va un paso más allá y busca las *posiciones* concretas del texto que pueden resultar relevantes al usuario.

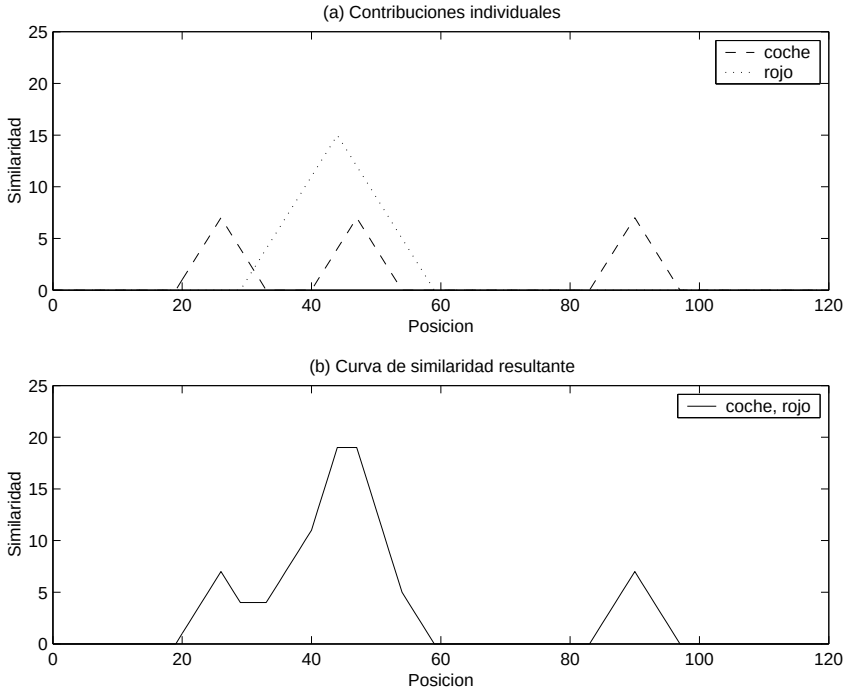


Figura 8.1: Ejemplo de similitudes basadas en localidad: (a) posiciones del texto con aparición de términos de la consulta y sus áreas de influencia; y (b) curva de similaridad resultante

La *Recuperación de Pasajes* [124] puede verse como una aproximación intermedia, ya que su objetivo consiste en identificar aquellas secciones del documento —denominadas *pasajes*— relevantes para la necesidad de información del usuario. Sin embargo, la Recuperación de Pasajes está más próxima al modelo basado en documentos que al modelo basado en localidad, ya que una vez que el documento original ha sido dividido en pasajes, éstos son procesados y ordenados empleando técnicas tradicionales. En este caso, los problemas para la correcta aplicación del modelo vienen dados por qué entendemos por *pasaje*, cómo identificarlo, qué tamaño debe tener, qué grado de superposición es el adecuado, etc. [138].

Por el contrario, el modelo basado en localidad considera la colección a indexar no como un conjunto de documentos, sino como una secuencia de palabras donde cada aparición de un término de la consulta ejerce una influencia sobre los términos circundantes. Dichas influencias son aditivas, de forma que la contribución de diferentes apariciones de términos de la consulta pueden sumarse, dando lugar a una medida de similaridad, tal y como se muestra en la figura 8.1. Aquellas áreas del texto con una mayor densidad de términos de la consulta, o con términos de mayor peso, darán lugar a picos en la curva de influencia resultante, señalando posiciones del texto potencialmente relevantes. Es de destacar que dichas posiciones son identificadas sin necesidad de particionar artificialmente el documento, cosa que sí ocurre en el caso de la Recuperación de Pasajes.

8.2.2. Cálculo de Similaridades

A continuación, describiremos el modelo basado en localidad propuesto originalmente por de Kretser y Moffat [63, 62, 61]. En este modelo únicamente es preciso calcular la medida de similaridad o relevancia sobre aquellas posiciones donde aparecen términos de la consulta, reduciendo de este modo el coste computacional asociado y posibilitando su aplicación en entornos prácticos.

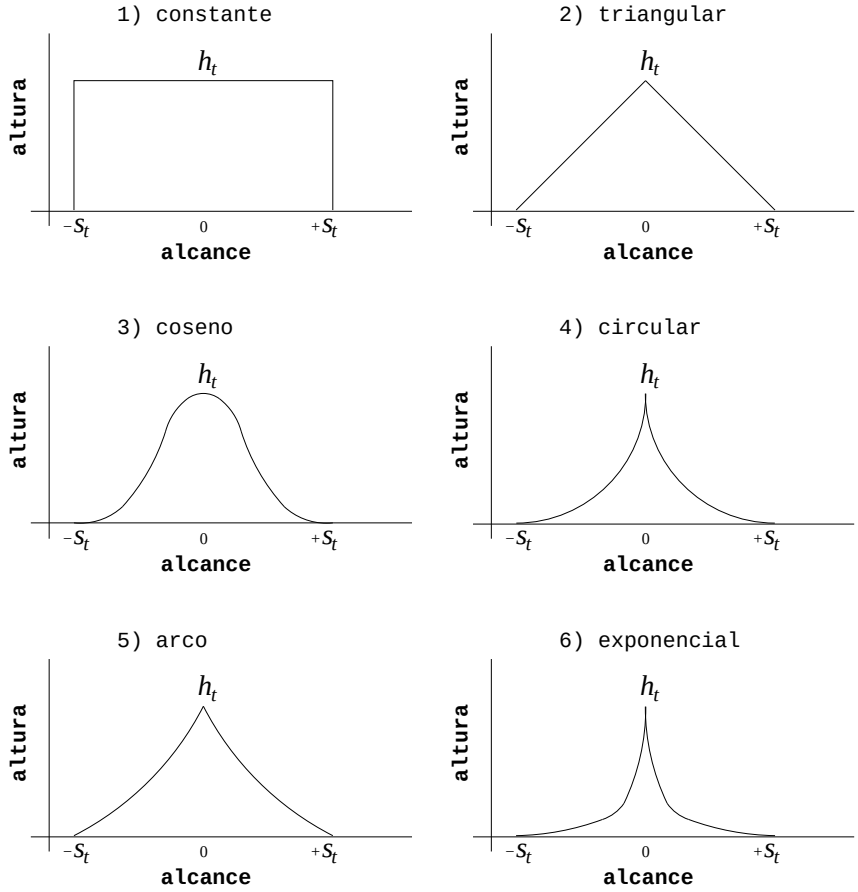


Figura 8.2: Formas de la función de contribución de similaridad c_t

La contribución a dicha similaridad por parte de un término de la consulta viene dada por una *función de contribución de similaridad* c_t definida en base a los siguientes parámetros [61]:

- La *forma* de la función, siendo la misma para todos los términos.
- La *altura máxima* h_t de la función, que se da en la posición del término que ejerce la influencia.
- El *alcance* s_t de la función, es decir, su radio de influencia.
- La distancia en palabras entre los dos términos considerados, $d = |x - l|$, donde l es la posición del término de la consulta que ejerce la influencia y x la posición sobre la que se desea calcular la medida de similaridad.

Seis son las formas de función descritas por los autores, y que recogemos en la figura 8.2:

1. La primera de las formas propuestas, y la más simple, es la función constante (cte), definida como

$$c_t(x, l) = \begin{cases} h_t & \text{si } d \leq s_t \\ 0 & \text{en otro caso} \end{cases} \tag{8.1}$$

y donde el término ejerce una influencia constante de amplitud h_t en toda su área de influencia.

2. La segunda forma estudiada es la triangular (*tri*), dada por

$$c_t(x,l) = \begin{cases} h_t \cdot \left(1 - \frac{d}{s_t}\right) & \text{si } d \leq s_t \\ 0 & \text{en otro caso} \end{cases} \tag{8.2}$$

y en la que se produce una disminución lineal del grado de influencia conforme aumenta la distancia respecto al término que ejerce dicha influencia.

3. La tercera forma propuesta es el coseno (*cos*), cuya fórmula es

$$c_t(x,l) = \begin{cases} \frac{h_t}{2} \cdot \left(1 + cos\left(\pi \cdot \frac{d}{s_t}\right)\right) & \text{si } d \leq s_t \\ 0 & \text{en otro caso} \end{cases} \tag{8.3}$$

y donde, al igual que antes, la magnitud de la influencia disminuye conforme nos alejamos del término, si bien el decaimiento es sinusoidal y no lineal.

4. La siguiente forma propuesta es la circular (*cir*), con una caída más abrupta al comenzar a distanciarse del término, y que equivale a los cuadrantes de dos círculos normalizados con centros en $(h_t, -s_t)$ y (h_t, s_t), siendo su expresión:

$$c_t(x,l) = \begin{cases} h_t \cdot \sqrt{1 - \left(\frac{d}{s_t}\right)^2} & \text{si } d \leq s_t \\ 0 & \text{en otro caso} \end{cases} \tag{8.4}$$

5. La quinta forma descrita por los autores es el arco (*arc*), un compromiso entre la función triangular (*tri*) y la función circular (*cir*), y que se define como la media de ambas:

$$c_t(x,l) = \begin{cases} \frac{h_t}{2} \cdot \left(1 - \frac{d}{s_t} + \sqrt{1 - \left(\frac{d}{s_t}\right)^2}\right) & \text{si } d \leq s_t \\ 0 & \text{en otro caso} \end{cases} \tag{8.5}$$

6. La sexta y última de las formas propuestas es la exponencial (*exp*), con un decaimiento exponencial del grado de influencia conforme la distancia respecto al término se incrementa:

$$c_t(x,l) = \begin{cases} h_t \cdot e^{-\frac{log_e 100}{s_t} \cdot d} & \text{si } d \leq s_t \\ 0 & \text{en otro caso} \end{cases} \tag{8.6}$$

El segundo parámetro conforme al cual se define la función de contribución de similaridad c_t es la altura máxima h_t asociada a un término t de la consulta. Dicha altura es función inversa de su frecuencia en la colección, admitiendo dos formulaciones posibles:

$$h_t = f_{Q,t} \cdot \frac{N}{f_t} \tag{8.7}$$

y

$$h_t = f_{Q,t} \cdot log_e\left(\frac{N}{f_t}\right) \tag{8.8}$$

donde N es el número total de términos en la colección, f_t el número de apariciones del término t en la colección y $f_{Q,t}$ la frecuencia del término t en la consulta Q.

En lo que respecta al siguiente parámetro, el alcance s_t de la influencia de un término t, éste viene dado también por el inverso de su frecuencia en la colección, pero normalizado en base a la frecuencia media:

$$s_t = \frac{n}{N} \cdot \frac{N}{f_t} = \frac{n}{f_t} \tag{8.9}$$

donde n es el número de términos diferentes en la colección, es decir, el tamaño del vocabulario.

Una vez determinados los parámetros a emplear, la medida de similaridad asignada a la posición x del documento en la cual aparece un término de la consulta Q se calcula como:

$$C_Q(x) = \sum_{t \in Q} \sum_{\substack{l \in I_t \\ |l-x| \leq s_t \\ term(x) \neq term(l)}} c_t(x, l) \tag{8.10}$$

donde I_t es el conjunto de posiciones donde ocurre un término t de la consulta Q, y donde $term(w)$ denota el termino asociado a la posición w. En otras palabras, la medida de similaridad o relevancia asociada a una posición es la suma de las influencias ejercidas por los otros términos de la consulta presentes en el documento y dentro de cuyo alcance se encuentra, exceptuando otras apariciones del término existente en la posición considerada [62].

Finalmente, la medida de relevancia $sim(D, Q)$ asignada a un documento D respecto a la consulta Q vendrá dada en función de las similaridades asignadas a las apariciones de términos de la consulta que dicho documento contenga. Para calcular esta medida de Kretser y Moffat [61] emplean un algoritmo iterativo. En dicho algoritmo se seleccionan, repetidamente, las posiciones de mayor puntuación de la colección, puntuaciones que se van sumando en un acumulador asociado al documento D al que corresponden y que almacena la medida de relevancia del mismo respecto a la consulta Q. Este proceso de selección y suma de puntuaciones termina cuando se hayan procesado posiciones correspondientes a r documentos diferentes[1]. Es decir, el proceso se detiene cuando hayamos conseguido calcular medidas de relevancia para r documentos diferentes.

8.2.3. Adaptaciones del Modelo

Dado que el modelo basado en localidad permite trabajar a un nivel más detallado que las técnicas clásicas de RI, al identificar no sólo los documentos relevantes sino también concretar las posiciones de interés dentro de los mismos, hemos optado por integrar dicho modelo en nuestro sistema. Para ello ha sido necesario realizar ciertas adaptaciones de acuerdo con nuestras necesidades, las cuales nos diferencian del planteamiento original del modelo.

El planteamiento elegido a la hora de integrar la similaridad basada en distancias dentro de nuestro sistema de RI, ha sido el del postprocesado de los documentos previamente obtenidos mediante un sistema de recuperación clásico basado en documentos con intención de mejorar la precisión de los primeros documentos devueltos. Este conjunto inicial de documentos es obtenido empleando nuestra propuesta basada en la indexación de los lemas de las palabras con contenido (lem), ya descrito en el capítulo 5. Este primer conjunto de documentos devuelto por el sistema es a continuación procesado empleando el modelo basado en localidad, tomando la ordenación final obtenida en base a distancias como aquélla a devolver como salida al usuario.

Para mejorar en lo posible el rendimiento final del sistema resultante, intentaremos partir de un conjunto inicial de documentos tan bueno como sea posible; es por ello que este conjunto inicial a reordenar es obtenido aplicando realimentación[2]. Sin embargo, el proceso

[1] Siendo r el número de documentos a devolver.

[2] Parámetros empleados: consultas cortas: α=0.8, β=0.1, γ=0, n_1=5, t=10; consultas largas: α=1.2, β=0.1, γ=0, n_1=5, t=10.

de reordenación se realizará en base a los términos de la consulta inicial, sin tener en cuenta los términos añadidos durante la realimentación. Ello se debe a que no existe garantía de que dichos términos guarden relación alguna de naturaleza sintáctica con los términos de la consulta inicial sin ir más allá de la mera coocurrencia a nivel de documento.

Por otra parte, debemos señalar que los parámetros de altura máxima h_t y alcance s_t utilizados durante la reordenación se calculan en base a los parámetros globales de la colección, y no en base a los parámetros locales al subconjunto de documentos devueltos, para así evitar los problemas derivados de la correlación que esto conllevaría.[3]

Otra de las diferencias respecto al modelo original es el del empleo de la lematización frente al *stemming* a la hora de la normalización de consultas y documentos, dado su mejor comportamiento, tal como se mostró en el capítulo 5.

El tercer punto de diferencia viene dado por el algoritmo de cálculo de relevancia de un documento. Dicho cálculo se realiza en base a las medidas de similaridad asignadas a las apariciones de términos de la consulta que dicho documento contiene. Frente al algoritmo iterativo inicial propuesto por de Kretser y Moffat, nuestra solución calcula la medida de relevancia $sim(D, Q)$ de un documento D respecto a una consulta Q como la suma de las medidas de similaridad asignadas a las apariciones de términos de la consulta que en él aparecen:

$$sim(D, Q) = \sum_{\substack{x \in D \\ term(x) \in Q}} C_Q(x) \tag{8.11}$$

8.3. Resultados Experimentales con Distancias

Antes de evaluar el comportamiento de nuestra propuesta basada en la reordenación mediante distancias, se realizó una primera serie de experimentos con el corpus de entrenamiento CLEF 2001-02·A como fase de puesta a punto para establecer la combinación de forma de función y altura máxima h_t que ofrecía mejores resultados. Dicha combinación sería aquélla a emplear de cara a la evaluación final.

Los resultados obtenidos durante esta fase previa de puesta a punto, recogidos en el Apéndice G en sus tablas G.1 a G.4, apuntan, por una parte, a la conveniencia de emplear logaritmos en el cálculo de la altura máxima h_t —fórmula 8.8—, al obtener unos resultados claramente superiores. Por otra parte, la forma de función circular (cir) —expresión 8.4— muestra un mejor comportamiento general frente al resto de formas, tanto en consultas cortas como largas. Por lo tanto, la evaluación de nuestra propuesta se realizará empleando la expresión de altura máxima $h_t = f_{Q,t} \cdot log_e(N/f_t)$ y la forma de función circular (cir).

Los resultados obtenidos se muestran en la tabla 8.1. Siguiendo el esquema habitual, la primera columna de cada grupo recoge los resultados de la línea base, en este caso la indexación de lemas con realimentación (lem), mientras que las siguientes dos columnas muestran, respectivamente, los resultados obtenidos tras la ordenación de lem mediante distancias (cir) y el porcentaje de mejora obtenido ($\%\Delta$).

Como muestran los resultados, la reordenación por distancias ha producido una disminución general del rendimiento del sistema, salvo para los primeros niveles de cobertura y primeros documentos devueltos, donde en algunos casos los resultados son similares o, con mayor frecuencia, incluso mejores.

En lo que respecta al comportamiento del sistema a nivel de consulta particular, las gráficas 8.3, 8.4 y 8.5 muestran de nuevo las diferencias en las precisiones a los 10 primeros documentos

[3]Por ejemplo, el parámetro f_t de número de apariciones de un término t es el número de apariciones de t en toda la colección, no el número de apariciones de t en el conjunto de documentos a reordenar.

Tabla 8.1: Resultados obtenidos mediante reordenación por distancias (*cir*) de la lematización con realimentación (*lem*)

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	*cir*	%Δ	*lem*	*cir*	%Δ	*lem*	*cir*	%Δ	*lem*	*cir*	%Δ	*lem*	*cir*	%Δ	*lem*	*cir*	%Δ
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2767	2767	–	2779	2779	–	2376	2376	–	2406	2406	–	2240	2240	–	2223	2223	–
Pr. no int.	.5220	.4668	-10.57	.5604	.4714	-15.88	.4773	.4278	-10.37	.5392	.4831	-10.40	.5024	.3924	-21.89	.5207	.4005	-23.08
Pr. doc.	.5784	.5039	-12.88	.5912	.5064	-14.34	.5681	.5379	-5.32	.6145	.5789	-5.79	.5530	.4761	-13.91	.5802	.4796	-17.34
R-pr.	.4990	.4651	-6.79	.5366	.4652	-13.31	.4599	.4205	-8.57	.5104	.4592	-10.03	.4912	.3921	-20.18	.4871	.3911	-19.71
Pr. a 0 %	.8221	**.8835**	**7.47**	.8895	**.8979**	**0.94**	.8210	**.8233**	**0.28**	.8710	.8678	-0.37	.8145	**.8230**	**1.04**	.8301	**.8415**	**1.37**
Pr. a 10 %	.7490	**.7870**	**5.07**	.8028	**.8143**	**1.43**	.6861	**.7197**	**4.90**	.7619	**.8084**	**6.10**	.7369	.6626	-10.08	.7421	.6518	-12.17
Pr. a 20 %	.6866	**.6883**	**0.25**	.7352	.7017	-4.56	.6319	**.6378**	**0.93**	.6929	.6808	-1.75	.6632	.5663	-14.61	.6758	.5737	-15.11
Pr. a 30 %	.6573	.6148	-6.47	.6996	.6066	-13.29	.5688	.5464	-3.94	.6497	.6000	-7.65	.6019	.5098	-15.30	.6304	.5030	-20.21
Pr. a 40 %	.5997	.5267	-12.17	.6541	.5372	-17.87	.5289	.4827	-8.74	.6202	.5438	-12.32	.5638	.4439	-21.27	.5975	.4400	-26.36
Pr. a 50 %	.5456	.4656	-14.66	.6005	.4728	-21.27	.5017	.4322	-13.85	.5733	.4987	-13.01	.5410	.4077	-24.64	.5479	.3956	-27.80
Pr. a 60 %	.4994	.4011	-19.68	.5386	.3968	-26.33	.4623	.3742	-19.06	.5194	.4276	-17.67	.4783	.3534	-26.11	.4938	.3437	-30.40
Pr. a 70 %	.4375	.3362	-23.15	.4621	.3402	-26.38	.4255	.3301	-22.42	.4862	.3809	-21.66	.4366	.2713	-37.86	.4587	.3043	-33.66
Pr. a 80 %	.3661	.2817	-23.05	.3910	.2836	-27.47	.3384	.2486	-26.54	.3753	.2984	-20.49	.3788	.2172	-42.66	.3891	.2532	-34.93
Pr. a 90 %	.2939	.1968	-33.04	.3130	.1876	-40.06	.2651	.1892	-28.63	.3059	.2212	-27.69	.3102	.1637	-47.23	.3230	.1936	-40.06
Pr. a 100 %	.1547	.0962	-37.82	.1624	.0865	-46.74	.1395	.0838	-39.93	.1704	.1095	-35.74	.1871	.0683	-63.50	.1928	.0973	-49.53
Pr. a 5 docs.	.6609	**.6913**	**4.60**	.6957	**.7261**	**4.37**	.5956	**.6000**	**0.74**	.6844	.6533	-4.54	.5872	.5532	-5.79	.6213	.5574	-10.28
Pr. a 10 docs.	.6457	.6391	-1.02	.6848	.6522	-4.76	.5600	.5444	-2.79	.6178	.6089	-1.44	.5596	.5064	-9.51	.5872	.4979	-15.21
Pr. a 15 docs.	.5884	**.5899**	**0.25**	.6435	.5971	-7.21	.5274	.5111	-3.09	.5822	.5556	-4.57	.5305	.4624	-12.84	.5504	.4652	-15.48
Pr. a 20 docs.	.5630	.5446	-3.27	.6043	.5674	-6.11	.5011	.4822	-3.77	.5533	.5189	-6.22	.4883	.4181	-14.38	.5266	.4277	-18.78
Pr. a 30 docs.	.5225	.4848	-7.22	.5580	.4971	-10.91	.4444	.4215	-5.15	.5081	.4733	-6.85	.4433	.3702	-16.49	.4667	.3780	-19.01
Pr. a 100 docs.	.3507	.3052	-12.97	.3598	.3048	-15.29	.2940	.2780	-5.44	.3191	.3022	-5.30	.2770	.2400	-13.36	.2853	.2404	-15.74
Pr. a 200 docs.	.2348	.2145	-8.65	.2361	.2124	-10.04	.1979	.1926	-2.68	.2067	.2043	-1.16	.1753	.1594	-9.07	.1791	.1626	-9.21
Pr. a 500 docs.	.1122	.1100	-1.96	.1121	.1107	-1.25	.0980	**.0982**	**0.20**	.1008	**.1013**	**0.50**	.0869	.0857	-1.38	.0871	.0855	-1.84
Pr. a 1000 docs.	.0602	.0602	0.00	.0604	.0604	0.00	.0528	.0528	0.00	.0535	.0535	0.00	.0477	.0477	0.00	.0473	.0473	0.00

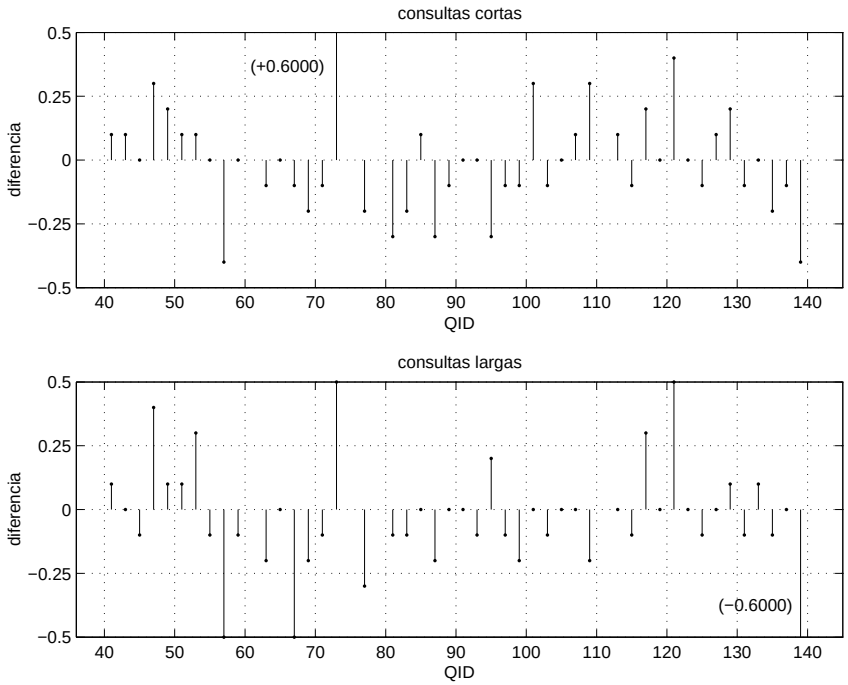


Figura 8.3: Diferencias en las precisiones a los 10 documentos: lematización con realimentación vs. reordenación por distancias. Corpus CLEF 2001-02·A

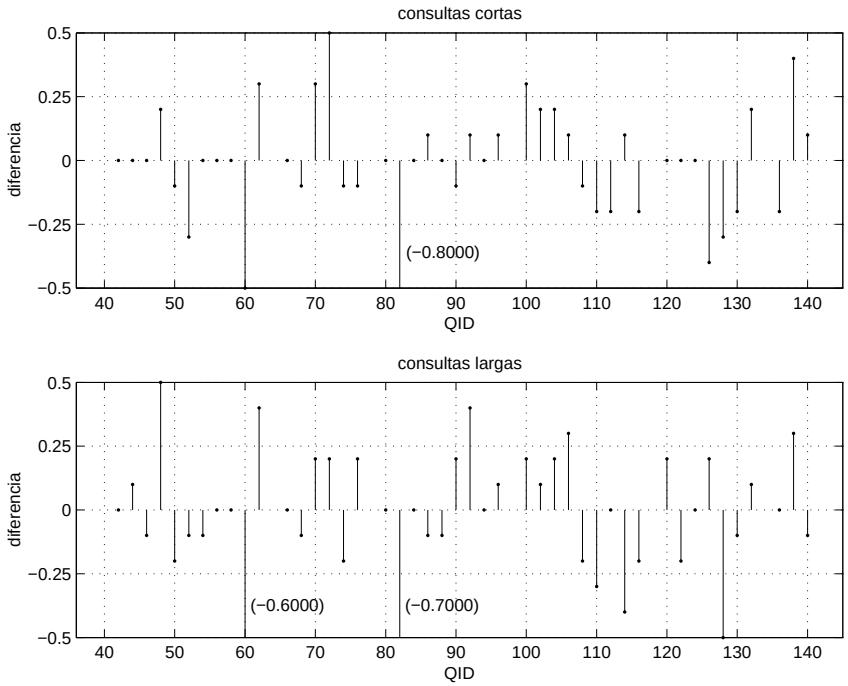


Figura 8.4: Diferencias en las precisiones a los 10 documentos: lematización con realimentación vs. reordenación por distancias. Corpus CLEF 2001-02·B

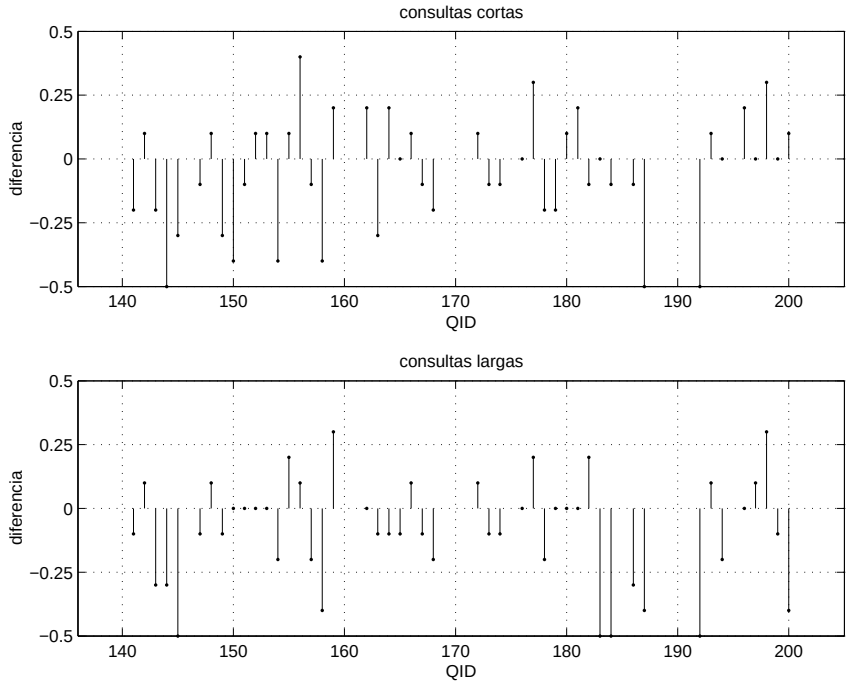


Figura 8.5: Diferencias en las precisiones a los 10 documentos: lematización con realimentación vs. reordenación por distancias. Corpus CLEF 2003

devueltos, ya que lo que pretendemos con esta aproximación no es tanto aumentar el rendimiento general del sistema sino mejorar la calidad de los primeros documentos obtenidos. Como se puede apreciar, las gráficas muestran un comportamiento del sistema algo errático, si bien, salvo en el caso del corpus CLEF 2001-02·B, los descensos en la precisión son más frecuentes que los incrementos.

Podemos concluir que esta primera aproximación no ha demostrado ser de demasiado interés práctico.

	Docs. relevantes							Docs. no relevantes			
K	$D \setminus L$	$L \setminus D$	$L \cap D$	R_{sup}	$Pr(L)$	$Pr(D)$	$Pr(L \cap D)$	$D \setminus L$	$L \setminus D$	$L \cap D$	N_{sup}
5	1.93	1.78	1.52	0.45	**0.66**	**0.69**	***0.80***	1.15	1.30	0.39	0.24
10	3.24	3.30	3.15	0.49	**0.65**	**0.64**	***0.76***	2.61	2.54	1.00	0.28
15	4.17	4.15	4.67	0.53	**0.59**	**0.59**	***0.72***	4.35	4.37	1.80	0.29
20	4.59	4.96	6.30	0.57	**0.56**	**0.54**	***0.72***	6.65	6.28	2.46	0.28
30	5.61	6.74	8.93	0.59	**0.52**	**0.48**	***0.68***	11.22	10.09	4.24	0.28
100	7.00	11.48	23.52	0.72	**0.35**	**0.31**	***0.49***	45.43	40.96	24.04	0.36
200	5.54	9.63	37.35	0.83	**0.23**	**0.21**	***0.36***	90.50	86.41	66.61	0.43
500	2.35	3.39	52.67	0.95	**0.11**	**0.11**	***0.16***	167.43	166.39	277.54	0.62

Tabla 8.2: Distribución de documentos relevantes y no relevantes tras la reordenación mediante distancias (forma circular, altura $h_t = f_{Q,t} \cdot log_e(N/f_t)$). Corpus CLEF 2001-02·A, consultas cortas

8.4. Fusión de Datos mediante Intersección

8.4.1. Justificación

Dado que el número de documentos relevantes devueltos es el mismo, la caída en el rendimiento del sistema en ésta primera aproximación sólo puede ser debida a una peor ordenación de los resultados fruto de la aplicación del modelo basado en distancias. Por esta razón decidimos estudiar la variación en la distribución de documentos relevantes y no relevantes en los K primeros documentos devueltos.

Los resultados obtenidos en dicho estudio son similares para los tres corpus de documentos y para ambos tipos de consultas, y se recogen en su totalidad en la tabla G.5 del Apéndice G, por lo que en estas páginas únicamente comentaremos los resultados obtenidos para el corpus CLEF 2001-02·A empleando consultas cortas, y que se muestran en la tabla 8.2.

Cada fila muestra los resultados obtenidos al comparar los K primeros documentos devueltos inicialmente por el sistema mediante indexación de lemas con realimentación (*lem*) —conjunto de resultados L— con aquéllos devueltos tras su reordenación mediante distancias (*cir*) —conjunto de resultados D. Las columnas muestran los resultados obtenidos para cada uno de los parámetros considerados: número medio de nuevos relevantes obtenidos mediante distancias ($D \setminus L$), número medio de relevantes perdidos con distancias ($L \setminus D$), número medio de relevantes que se mantienen ($L \cap D$), coeficiente de superposición de relevantes (R_{sup}), precisión de *lem* a los K primeros documentos ($Pr(L)$), precisión a los K documentos tras la reordenación por distancias ($Pr(D)$), y precisión en los documentos comunes a ambas aproximaciones dentro de sus K primeros documentos ($Pr(L \cap D)$). En la parte derecha de la tabla se muestran sus equivalentes para el caso de los documentos no relevantes: número medio de no relevantes añadidos, perdidos y comunes, y grado de superposición de no relevantes.

A partir de estos resultados se pueden extraer diversas conclusiones de importancia. En primer lugar, observamos que el número de documentos relevantes obtenidos por ambas aproximaciones dentro de sus K primeros documentos es muy similar —si bien algo menor para las distancias—, tal como se puede apreciar en las cifras absolutas de documentos relevantes entrantes y salientes y en las precisiones a los K documentos de ambas aproximaciones. Esto nos permite confirmar que se trata de un problema de mala ordenación de los resultados.

En segundo lugar debemos referirnos a los coeficientes de superposición de documentos relevantes (R_{sup}) y no relevantes (N_{sup}). Estos coeficientes, definidos por Lee en [136], indican el grado de superposición entre el conjunto de documentos relevantes o no relevantes de dos conjuntos de documentos devueltos. Para dos ejecuciones run_1 y run_2, dichos coeficientes se definen como:

$$R_{sup} = \frac{2\,|Rel(run_1) \cap Rel(run_2)|}{|Rel(run_1)| + |Rel(run_2)|} \tag{8.12}$$

$$N_{sup} = \frac{2\,|Nonrel(run_1) \cap Nonrel(run_2)|}{|Nonrel(run_1)| + |Nonrel(run_2)|} \tag{8.13}$$

donde $Rel(X)$ y $Nonrel(X)$ representan, respectivamente, el conjunto de documentos relevantes y no relevantes devueltos en la ejecución X.

Puede apreciarse en la tabla 8.2 que los factores de superposición de los documentos relevantes es considerablemente mayor que el obtenido para los documentos no relevantes. Por lo tanto ambas aproximaciones devuelven un conjunto similar de documentos relevantes, pero un conjunto diferente de documentos irrelevantes. Se cumple, pues, la denominada *propiedad de la superposición desigual* [136], que dice que diferentes ejecuciones deben devolver conjuntos similares de documentos relevantes a la vez que devolver conjuntos disimilares de no relevantes como primer indicador de la efectividad que tendría la fusión de datos de ambas.

En tercer lugar, y en relación al punto anterior, puede verse que la precisión en los documentos comunes a ambas aproximaciones dentro de sus K primeros documentos ($Pr(L \cap D)$) es mayor que las precisiones alcanzadas tanto por lemas ($Pr(L)$) como por distancias ($Pr(D)$); o lo que es lo mismo, la probabilidad de que un documento sea relevante es mayor cuando es devuelto por ambas aproximaciones. Esto último concuerda con lo afirmado por Saracevic y Kantor [210], según los cuales cuando un documento es devuelto repetidamente para una misma necesidad de información —bien mediante diferentes sistemas de recuperación, bien mediante diferentes ejecuciones—, más indicios hay acerca de su relevancia, y mayor debería ser la relevancia asignada al mismo.

Conforme a estas observaciones, se decidió abordar una nueva aproximación para la reordenación, esta vez basada en la fusión de datos, combinando los resultados obtenidos inicialmente mediante la indexación de lemas con los resultados obtenidos durante su reordenación con distancias.

8.4.2. Descripción del Algoritmo

La *fusión de datos* es una técnica de combinación de evidencias consistente en la combinación de los resultados devueltos empleando diferentes representaciones de consultas o documentos, o empleando múltiples técnicas de recuperación [77, 136, 57].

En nuestro caso hemos optado por una aproximación basada no en la combinación de puntuaciones en base a similaridades [77, 136] o rango [136], sino en un criterio booleano para el cual, una vez fijado un valor K, los documentos son devueltos en el siguiente orden:

1. En primer lugar, los documentos pertenecientes a la intersección de los K primeros documentos de ambas aproximaciones: $L_K \cap D_K$. El objetivo perseguido es el de incrementar la precisión en los primeros documentos devueltos.

2. A continuación, los documentos pertenecientes a los K primeros documentos de ambas aproximaciones que no estén en la intersección: $(L_K \cup D_K) \setminus (L_K \cap D_K)$. El objetivo es añadir a los primeros documentos devueltos aquellos documentos relevantes devueltos únicamente medienta la aproximación basada en distancias, sin perjudicar la ordenación de aquéllos devueltos únicamente por la indexación de lemas.

3. Finalmente, los restantes documentos devueltos por los lemas: $L \setminus (L_K \cup D_K)$.

donde L es el conjunto de resultados devuelto por *lem*, L_K el conjunto de K primeros resultados devuelto con *lem*, y D_K el conjunto de K primeros resultados devuelto mediante la reordenación por distancias.

Con respecto a la ordenación interna de los resultados, se tomará como referencia, por sus mejores resultados, la ordenación obtenida mediante la indexación de lemas (*lem*). De esta forma cuando se devuelva un subconjunto S de resultados, los documentos que lo conforman se devolverán en el mismo orden relativo que existía entre ellos cuando eran devueltos por *lem*.[4]

8.5. Resultados Experimentales con Fusión de Datos

Tras experimentos previos de puesta a punto del valor de K en la que se ensayaron diferentes valores de dicho parámetro —véanse tablas G.6 a G.9 del Apéndice G—, se optó finalmente por emplear, como mejor compromiso, un valor $K = 30$ en el caso de consultas cortas y un valor

[4]Es decir, si la secuencia original en *lem* era *d2-d3-d1* y se toma un subconjunto {*d1,d2*} a devolver, los documentos se obtendrían en el mismo orden relativo original: *d2-d1*.

Tabla 8.3: Resultados obtenidos mediante reordenación por fusión con intersección (*cir*) de la lematización con realimentación (*lem*)

corpus	*CLEF 2001-02·A*						*CLEF 2001-02·B*						*CLEF 2003*					
consulta	*cortas*			*largas*			*cortas*			*largas*			*cortas*			*largas*		
técnica	*lem*	*cir*	*%Δ*	*lem*	*cir*	*%Δ*	*lem*	*cir*	*%Δ*	*lem*	*cir*	*%Δ*	*lem*	*cir*	*%Δ*	*lem*	*cir*	*%Δ*
#consultas	46	46	–	46	46	–	45	45	–	45	45	–	47	47	–	47	47	–
#docs. dev.	46k	46k	–	46k	46k	–	45k	45k	–	45k	45k	–	47k	47k	–	47k	47k	–
#rlvs. esp.	3007	3007	–	3007	3007	–	2513	2513	–	2513	2513	–	2335	2335	–	2335	2335	–
#rlvs. dev.	2767	2767	–	2779	2779	–	2376	2376	–	2406	2406	–	2240	2240	–	2223	2223	–
Pr. no int.	.5220	**.5327**	**2.05**	.5604	.5589	-0.27	.4773	.4768	-0.10	.5392	**.5497**	**1.95**	.5024	.4977	-0.94	.5207	.5167	-0.77
Pr. doc.	.5784	**.5807**	**0.40**	.5912	**.5921**	**0.15**	.5681	**.5764**	**1.46**	.6145	**.6305**	**2.60**	.5530	**.5554**	**0.43**	.5802	.5793	-0.16
R-pr.	.4990	**.5126**	**2.73**	.5366	**.5433**	**1.25**	.4599	.4551	-1.04	.5104	**.5188**	**1.65**	.4912	.4737	-3.56	.4871	.4865	-0.12
Pr. a 0 %	.8221	**.8386**	**2.01**	.8895	**.9091**	**2.20**	.8210	**.8248**	**0.46**	.8710	**.8751**	**0.47**	.8145	**.8163**	**0.22**	.8301	.8257	-0.53
Pr. a 10 %	.7490	**.7758**	**3.58**	.8028	**.8256**	**2.84**	.6861	**.7191**	**4.81**	.7619	**.7740**	**1.59**	.7369	.7283	-1.17	.7421	**.7540**	**1.60**
Pr. a 20 %	.6866	**.7193**	**4.76**	.7352	**.7528**	**2.39**	.6319	**.6426**	**1.69**	.6929	**.7188**	**3.74**	.6632	**.6737**	**1.58**	.6758	**.6834**	**1.12**
Pr. a 30 %	.6573	**.6844**	**4.12**	.6996	.6922	-1.06	.5688	**.5818**	**2.29**	.6497	**.6784**	**4.42**	.6019	.6015	-0.07	.6304	**.6391**	**1.38**
Pr. a 40 %	.5997	**.6164**	**2.78**	.6541	**.6610**	**1.05**	.5289	**.5470**	**3.42**	.6202	**.6460**	**4.16**	.5638	**.5672**	**0.60**	.5975	.5876	-1.66
Pr. a 50 %	.5456	**.5644**	**3.45**	.6005	**.6026**	**0.35**	.5017	.4909	-2.15	.5733	**.5996**	**4.59**	.5410	.5359	-0.94	.5479	.5327	-2.77
Pr. a 60 %	.4994	**.5098**	**2.08**	.5386	.5233	-2.84	.4623	.4531	-1.99	.5194	**.5381**	**3.60**	.4783	.4705	-1.63	.4938	.4865	-1.48
Pr. a 70 %	.4375	**.4391**	**0.37**	.4621	.4554	-1.45	.4255	.4003	-5.92	.4862	.4795	-1.38	.4366	.4191	-4.01	.4587	.4447	-3.05
Pr. a 80 %	.3661	**.3694**	**0.90**	.3910	.3815	-2.43	.3384	.3328	-1.65	.3753	**.3769**	**0.43**	.3788	.3590	-5.23	.3891	.3768	-3.16
Pr. a 90 %	.2939	.2919	-0.68	.3130	.2953	-5.65	.2651	.2560	-3.43	.3059	.3014	-1.47	.3102	.2932	-5.48	.3230	.3077	-4.74
Pr. a 100 %	.1547	.1545	-0.13	.1624	.1537	-5.36	.1395	.1370	-1.79	.1704	.1643	-3.58	.1871	.1698	-9.25	.1928	.1845	-4.30
Pr. a 5 docs.	.6609	**.6739**	**1.97**	.6957	**.7217**	**3.74**	.5956	**.6178**	**3.73**	.6844	**.6933**	**1.30**	.5872	**.6298**	**7.25**	.6213	**.6553**	**5.47**
Pr. a 10 docs.	.6457	**.6761**	**4.71**	.6848	**.7065**	**3.17**	.5600	**.5756**	**2.79**	.6178	**.6400**	**3.59**	.5596	**.5745**	**2.66**	.5872	**.5979**	**1.82**
Pr. a 15 docs.	.5884	**.6188**	**5.17**	.6435	**.6449**	**0.22**	.5274	**.5393**	**2.26**	.5822	**.6000**	**3.06**	.5305	**.5390**	**1.60**	.5504	**.5560**	**1.02**
Pr. a 20 docs.	.5630	**.5826**	**3.48**	.6043	**.6185**	**2.35**	.5011	**.5089**	**1.56**	.5533	**.5722**	**3.42**	.4883	**.5074**	**3.91**	.5266	.5170	-1.82
Pr. a 30 docs.	.5225	.5225	0.00	.5580	**.5652**	**1.29**	.4444	.4444	0.00	.5081	**.5148**	**1.32**	.4433	.4433	0.00	.4667	**.4716**	**1.05**
Pr. a 100 docs.	.3507	.3502	-0.14	.3598	.3539	-1.64	.2940	**.3011**	**2.41**	.3191	**.3304**	**3.54**	.2770	**.2789**	**0.69**	.2853	.2809	-1.54
Pr. a 200 docs.	.2348	**.2349**	**0.04**	.2361	**.2380**	**0.80**	.1979	**.1999**	**1.01**	.2067	**.2100**	**1.60**	.1753	**.1761**	**0.46**	.1791	**.1800**	**0.50**
Pr. a 500 docs.	.1122	.1122	0.00	.1121	**.1126**	**0.45**	.0980	**.0983**	**0.31**	.1008	**.1012**	**0.40**	.0869	**.0871**	**0.23**	.0871	**.0874**	**0.34**
Pr. a 1000 docs.	.0602	.0602	0.00	.0604	.0604	0.00	.0528	.0528	0.00	.0535	.0535	0.00	.0477	.0477	0.00	.0473	.0473	0.00

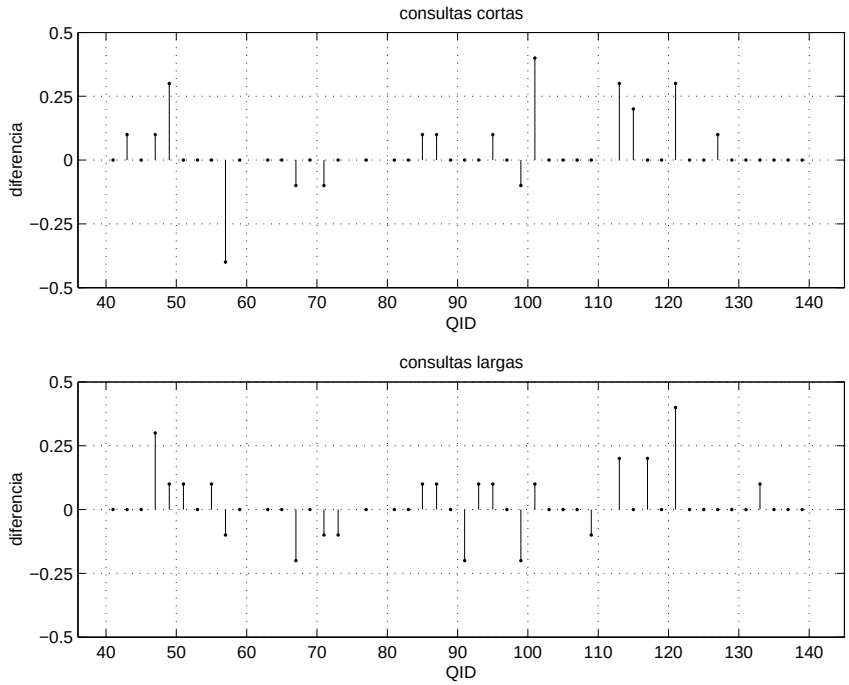


Figura 8.6: Diferencias en las precisiones a los 10 documentos: lematización con realimentación vs. reordenación por fusión. Corpus CLEF 2001-02·A

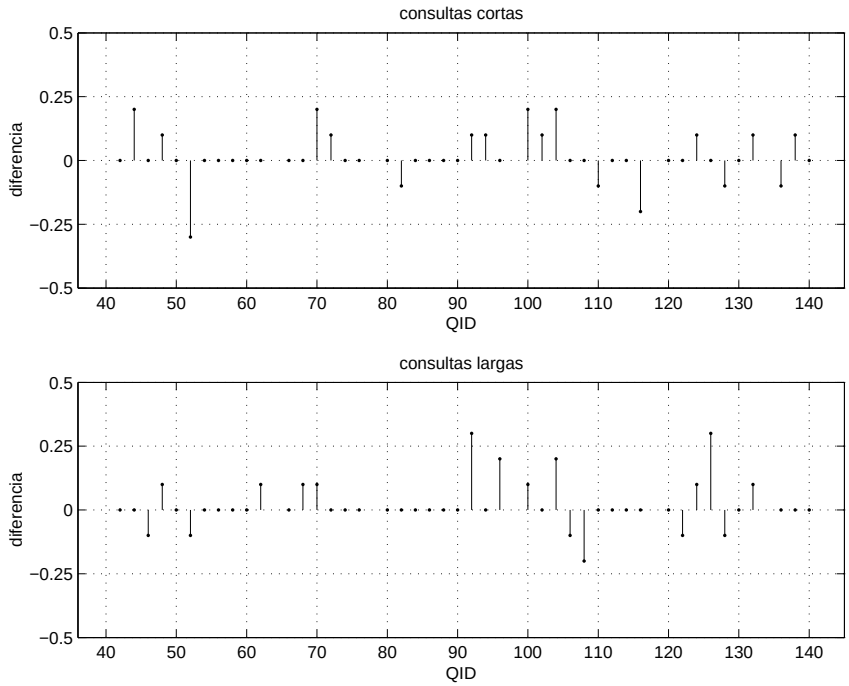


Figura 8.7: Diferencias en las precisiones a los 10 documentos: lematización con realimentación vs. reordenación por fusión. Corpus CLEF 2001-02·B

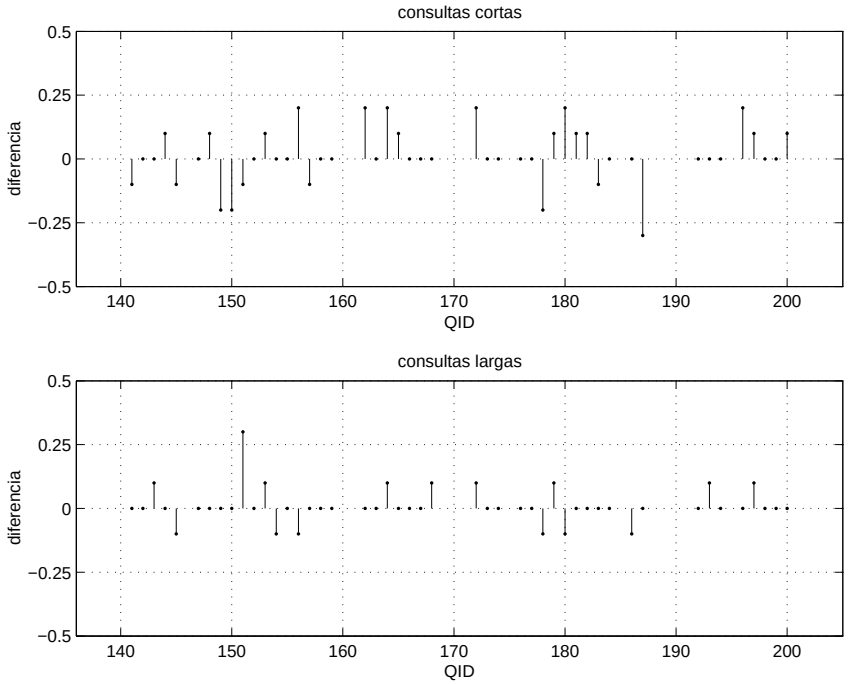


Figura 8.8: Diferencias en las precisiones a los 10 documentos: lematización con realimentación vs. reordenación por fusión. Corpus CLEF 2003

$K = 50$ en el caso de las largas. Los restantes parámetros del modelo se mantienen: forma circular y altura $h_t = f_{Q,t} \cdot log_e(N/f_t)$.

En la tabla 8.3 se muestran los resultados obtenidos para la nueva aproximación. Al igual que antes, la primera columna de cada grupo muestra los resultados de la indexación de lemas con realimentación (*lem*) antes de la reordenación, que actuará como línea base, mientras que las siguientes dos columnas muestran los resultados obtenidos tras la ordenación mediante fusión (*cir*), y el porcentaje de mejora obtenido ($\%\Delta$).

Podemos apreciar que las mejoras obtenidas con la reordenación mediante fusión son consistentes, especialmente en el caso de la precisión a los n documentos devueltos —tal como se pretendía—, si bien dichas mejoras también se extienden al resto de parámetros estudiados, siendo algo menores en el caso del corpus CLEF 2003.

Asimismo, los resultados obtenidos a nivel de consulta particular —gráficas 8.6, 8.7 y 8.8— muestran un predominio de los incrementos en la precisión frente a sus disminuciones.

8.6. Discusión

A lo largo de este capítulo se ha planteado la utilización de un modelo de recuperación basado en distancias entre palabras, o basado en localidad, que permite una aproximación pseudo-sintáctica al problema de la variación lingüística de carácter sintáctico presente en los textos.

Dos han sido las aproximaciones propuestas para su aplicación, ambas enfocadas a la reordenación de resultados, en este caso obtenidos mediante indexación de lemas de palabras con contenido. La primera aproximación, donde el orden obtenido mediante la aplicación del modelo basado en localidad se tomaba como el orden final a devolver, no obtuvo, en general, buenos resultados. Tras analizar el comportamiento del sistema se optó por emplear una aproximación basada en la fusión de datos, y que emplea la intersección de los conjuntos de documentos devueltos por ambos sistemas como guía para la reordenación. Esta segunda aproximación

resultó fructífera, obteniendo una mejora consistente y general en la ordenación de los resultados, sin perjudicar otros aspectos.

Capítulo 9

Conclusiones y Trabajo Futuro

El objetivo perseguido en este trabajo de tesis ha sido el desarrollo de tecnología de base para el Procesamiento del Lenguaje Natural y su aplicación en sistemas de procesamiento automático de la información —y de Recuperación de Información en particular— para el caso del español. A este respecto debemos llamar la atención sobre el hecho de que el español ha sido relegado con frecuencia a un segundo plano por la comunidad científica, que ha centrado sus esfuerzos en el inglés, lengua dominante, y en menor medida en otras lenguas europeas como el francés o el alemán.

Al escaso interés acerca de la investigación en el español se le une, además, una importante carencia de recursos lingüísticos libremente accesibles para los investigadores, dificultando todavía más la tarea. Este obstáculo ha debido ser salvado restringiendo la necesidad de recursos de las soluciones propuestas, centrándose mayormente en el empleo de la información léxica.

9.1. Aportaciones de la Tesis

Continuando con el esquema de organización que se ha seguido en la memoria, recogemos en este apartado las principales contribuciones de nuestro trabajo de tesis.

En primer lugar, se ha desarrollado un preprocesador-segmentador avanzado de base lingüística para la *tokenización* y segmentación de textos en español. Esta tarea suele ser frecuentemente ignorada a pesar de su enorme importancia, ya que las palabras y frases identificadas en esta fase constituirán las unidades fundamentales sobre las que deben trabajar las fases posteriores. Por otra parte, los procesos involucrados en esta fase van más allá de la mera identificación de las diferentes frases del texto y de cada uno de sus componentes individuales. A este efecto hemos desarrolado un esquema de preprocesamiento lingüístico avanzado orientado a la desambiguación y etiquetación robusta del español. No obstante, se trata de una propuesta de arquitectura general que puede ser aplicada a otras tareas de análisis (sintáctico, semántico, etc.). Llegados a este punto debemos indicar que debieron llevarse a cabo algunas modificaciones en el módulo de identificación de nombres propios para mejorar su efectividad en tareas de Recuperación de Información.

En el caso del tratamiento de la variación lingüística de carácter morfológico flexivo, se ha estudiado la utilización de técnicas de desambiguación-lematización para la normalización de términos simples en tareas de Recuperación de Información, empleando como términos de indexación los lemas de las palabras con contenido del texto —nombres, adjetivos y verbos. Asimismo se desarrolló un módulo de tratamiento, en función del contexto, de los errores de etiquetación-lematización producidos por la eliminación de signos ortográficos en las frases en mayúscula. Los buenos resultados obtenidos señalan a la lematización como una alternativa viable a las técnicas clásicas basadas en *stemming*. Por otra parte, la eliminación de la

ambigüedad léxica del texto a procesar mediante su etiquetación y lematización constituye el primer paso para el desarrollo de métodos de normalización más elaborados que hagan frente a fenómenos de variación lingüística más complejos.

El tratamiento del segundo tipo de variación lingüística morfológica, la variación derivativa, se ha llevado a cabo mediante el desarrollo de mecanismos automáticos de generación de familias morfológicas —conjuntos de palabras que comparten la misma raíz y que están ligadas por relaciones derivativas— en base a la morfología derivativa productiva del español. Esta herramienta para el procesamiento de los fenómenos morfológico-derivativos puede ser empleada a modo de *stemmer* avanzado de base lingüística en tareas de normalización de términos simples. Sin embargo, nuestra propuesta no es todavía del todo inmune a los problemas generados por la introducción de ruido durante el proceso de normalización cuando dos términos de escasa o nula relación semántica son normalizados al mismo término índice. Este fenómeno es debido a la sobregeneración durante el proceso creación de familias.

Una vez demostrada la viabilidad de las técnicas de Procesamiento del Lenguaje Natural para el tratamiento de la variación lingüística a nivel de palabra, el siguiente paso lo constituyó la aplicación de técnicas de análisis a nivel de frase para, en primer lugar, obtener términos índice más precisos y descriptivos y, en segundo lugar, para tratar la variación lingüística de carácter sintáctico. Para ello se ha ensayado una aproximación basada en la utilización de dependencias sintácticas a modo de términos índice complejos como complemento a los términos índice simples. Dado que el primer paso en este proceso es el de la obtención de la estructura sintáctica del texto, se desarrollaron dos analizadores sintácticos superficiales de dependencias para el español: el primero, basado en patrones —Patterns—, y el segundo, basado en cascadas de traductores finitos —Cascade. Estas soluciones permitieron reducir la complejidad computacional del sistema a una complejidad lineal, incrementando además la robustez del mismo. Por otra parte, la introducción de mecanismos de tratamiento de la morfología derivativa basados en familias morfológicas permitieron extender el tratamiento de la variación estrictamente sintáctica a la variación morfosintáctica. Asimismo, se ensayaron dos aproximaciones diferentes en base al origen de la información sintáctica empleada: la primera, utilizando las dependencias obtenidas a partir de las consultas; la segunda, que ha resultado superior en su conjunto, empleando las dependencias obtenidas a partir de los documentos. También se estudió el impacto de los tipos de dependencias a utilizar, comparando el rendimiento del sistema al emplear únicamente dependencias correspondientes a frases nominales —con la consiguiente reducción de costes—, o bien empleando la totalidad de éstas —lo que permitió incrementar ligeramente la precisión obtenida.

Finalmente, se evaluó una nueva aproximación pseudo-sintáctica al tratamiento de la variación lingüística de carácter sintáctico, la cual prescinde de la necesidad de contar con gramática o analizador sintáctico alguno. Esta nueva aproximación se basa en la utilización de un modelo basado en localidad, sustentado sobre similaridades basadas en distancias, como complemento a las técnicas clásicas de Recuperación de Información basadas en la indexación de términos simples. El modelo original hubo de ser adaptado primero para su integración en nuestro sistema de Recuperación, proponiéndose seguidamente dos acercamientos diferentes para su aplicación, ambos enfocados a la reordenación de los resultados obtenidos mediante la indexación de lemas de palabras con contenido mediante un modelo clásico basado en documentos . El primero de ellos, que empleaba la ordenación obtenida mediante el modelo basado en localidad como aquélla a devolver, no obtuvo, en general, resultados satisfactorios. Tras analizar el comportamiento del sistema se optó por emplear una aproximación basada en la fusión de datos, y que emplea la intersección de los conjuntos de documentos devueltos por ambos sistemas como guía para la reordenación. Esta segunda aproximación resultó mucho más fructífera.

A la hora de minimizar el coste computacional de nuestras propuestas de cara a su aplicación en entornos prácticos, se ha hecho amplio uso de tecnología de estado finito. Si bien esta tecnología es a menudo tildada como *ad hoc*, proponemos en este trabajo una secuencia de procesos basados en estado finito donde cada etapa se corresponde con elementos intuitivos y universales del lenguaje:

- La existencia de palabras individuales y expresiones que conforman frases y oraciones.
- La existencia de diferentes categorías de palabras que contienen la semántica principal del lenguaje: nombres, adjetivos y verbos.
- La existencia de relaciones semánticas entre palabras pertenecientes a categorías diferentes (p.ej., el nombre correspondiente a la acción de un verbo).
- La existencia de estructuras sintácticas básicas que relacionan las palabras dentro de la frase u oración, tales como las relaciones nombre-modificador, sujeto-verbo o verbo-objeto.

Por otra parte, si bien el trabajo aquí presentado estaba orientado principalmente al tratamiento de textos en español, las técnicas y mecanismos descritos son fácilmente adaptables a otros idiomas de características y comportamiento similar, constituyendo una arquitectura general aplicable a otros idiomas mediante la introducción de las modificaciones oportunas. Prueba de ello es la existencia de versiones en explotación para el gallego de gran parte de las herramientas descritas [31, 30], idioma donde la falta de recursos lingüísticos es aún mayor. Del mismo modo, esta tecnología podría ser fácilmente adaptada a otras lenguas de características similares, como podrían ser el portugués o el catalán.

9.2. Campos de Investigación Relacionados

Si bien esta memoria se centra en el caso de la aplicación del Procesamiento del Lenguaje Natural a la Recuperación de Información, las herramientas desarrolladas han sido también aplicadas con éxito en otros campos como la Extracción de Información —dentro de los proyectos *"Aplicación de Inteligencia Artificial para la Extracción de Información Cognitiva y Cualitativa en Mercados Financieros"* y *"Extracción de Información de Noticias Bursátiles para la Evaluación del Sentimiento del Mercado"*— y la Búsqueda de Respuestas —publicaciones [164, 165] y proyecto *"Recuperación de Información para la Búsqueda de Respuestas en Textos Económicos"*.

A continuación describimos brevemente otras líneas de investigación en las que participa el autor y que guardan relación directa con el tema de tesis:

- Desarrollo de mecanismos eficientes basados en autómatas para la aplicación de reglas de restricción en procesos de etiquetación-desambigüación [85, 84].
- Estudio de las aplicaciones del *post-stemming* combinado con la lematización para el tratamiento de la variación morfológica derivativa [241].
- Estudio de las aplicaciones de los mecanismos de gradación de sinonimia para la expansión de consultas en entornos de Recuperación de Información [240, 239, 241].
- Desarrollo de nuevas técnicas de análisis sintáctico parcial [257, 47, 48].
- Investigación sobre nuevas técnicas de análisis sintáctico robusto [246, 245].

- Desarrollo de nuevas aproximaciones para la corrección automática de errores sintácticos [247, 248].
- Estudio de nuevas técnicas de comparación de árboles sintácticos en base a distancias sintácticas y semánticas [256].
- Ensayo de nuevas aproximaciones para la fusión de datos en Recuperación de Información [183].

9.3. Trabajo Futuro

Los resultados del trabajo presentado en esta memoria marcan un punto de partida para el desarrollo de nuevas aproximaciones en el campo del Procesamiento del Lenguaje Natural y su aplicación al procesamiento automático de la información en español, particularmente para el caso de la Recuperación de Información. Sin embargo, es preciso seguir profundizando en nuestras investigaciones sobre el español para reducir en lo posible la diferencia que todavía nos separa respecto a otros idiomas.

Uno de los primeros temas que pretendemos abordar es el de la extensión de nuestro trabajo a otros modelos de recuperación diferentes al vectorial, como sería el caso de los modelos probabilísticos. A este respecto se han llevado a cabo ya algunas pruebas preliminares empleando el motor de indexazión ZPrise [6] con un esquema de pesos Okapi BM25 [193]. En dichos experimentos [21, 238, 237] la lematización ha seguido mostrado su bondad respecto al *stemming* como técnica de normalización de términos simples.

Del mismo modo, pretendemos extender nuestros experimentos en Recuperación de Información monolingüe al caso translingüe, empleando como lengua origen o destino el español.

De un modo más específico, pretendemos también ampliar nuestro trabajo en los diferentes niveles de procesamiento lingüístico:

- En lo referente a nuestro preprocesador lingüístico, podría resultar de interés la generación automática de un tesauro de nombres propios y acrónimos durante la fase de entrenamiento de nombres propios que nos permitiese posteriormente expandir las consultas con las variantes de los propios o siglas que éstas contuviesen.
- Ya a nivel morfológico, se está considerando la posibilidad de incrementar la efectividad del proceso de normalización de términos simples mediante la integración de nuevos mecanismos de corrección ortográfica automática, campo en el que nuestro grupo cuenta ya con amplia experiencia [252, 253, 244, 251, 250]. Además, pretendemos incrementar también la efectividad en la etiquetación mediante la aplicación de reglas de restricción. Los mecanismos de implementación de las mismas, empleando traductores finitos para incrementar su eficiencia, han sido ya estudiados por el autor en [85, 84], siendo ahora el turno de los lingüistas para desarrollar las reglas pertinentes.
- También a nivel morfológico, pero ya en su aspecto derivativo, pretendemos aplicar mecanismos adicionales de filtrado basados en la aplicación de *restricciones de selección* generadas automáticamente a partir de los propios textos [80, 79]. La idea consistiría en que una posible derivación sería únicamente aceptada si ambos términos, original y derivado, poseen contextos compatibles. Esto nos permitiría reducir todavía más el ruido introducido debido a la sobregeneración.
- Las restricciones de selección señaladas en el punto anterior permitirían también incrementar la capacidad de desambiguación sintáctica del sistema, especialmente

respecto al clásico problema de la adjunción de sintagmas peposicionales, donde también pretendemos aplicarlas. El objetivo, aumentar en lo posible la cobertura del sistema sin perjudicar la precisión.

- Un segundo campo de mejora a nivel sintáctico sería el del modo de emplear la información sintáctica disponible de un modo más eficaz. El almacenamiento de términos simples y complejos en índices separados para su tratamiento independiente y el empleo de técnicas de *fusión de datos* para la combinación de los resultados obtenidos podrían facilitar dicha tarea. Algunos de nuestros primeros ensayos se recogen en [183]. Pretendemos también realizar un tratamiento diferenciado de los pares a la hora del reajuste de pesos durante la expansión de consultas, por los problemas de sobreponderación y violación del principio de independencia que puede suponer su empleo conjunto con términos simples.

- Estamos también estudiando la posibilidad de enriquecer los términos índice con la inclusión de los pares de palabras involucradas en las *colocaciones*[1] [14, 15] que aparecen en textos y consultas. El problema que surge con el tratamiento de las colocaciones es que no existen unas reglas fijas que nos indiquen qué palabras coocurren con qué otras, sino que vienen determinadas por el uso de la lengua. Por ello, para extraer las colocaciones primero se debe disponer de un conjunto de textos en los cuales se hayan identificado manualmente las colocaciones existentes, para a continuación tratar de extraer ciertos patrones estadísticos que nos ayuden a identificar colocaciones en textos que no hayan sido previamente tratados.

- Dentro todavía del nivel sintáctico, pero en lo referente a nuestra segunda aproximación, empleando un modelo basado en localidad, pretendemos dar capacidad al sistema para tratar las variantes no sólo sintácticas sino también morfosintácticas de una expresión mediante el empleo de familias morfológicas para la normalización de términos simples. Sería también de interés realizar experiencias acerca de la posibilidad de restringir la influencia de un término a únicamente aquellos otros términos que se encuentran ligados por él mediante relaciones sintácticas, relaciones que serían identificadas mediante un análisis sintáctico superficial.

- Del mismo modo, se está considerando la posibilidad de introducir técnicas de comparación sintáctica basadas en árboles [182, 256, 255, 254], de mayor complejidad y potencia. Para reducir en lo posible los costes de este proceso se aplicarían dichas técnicas de comparación avanzada únicamente para los N primeros devueltos inicialmente por un sistema de Recuperación de Información convencional. Este tipo de solución permitiría también la aplicación de algoritmos de análisis sintáctico más potentes [46, 257, 47, 48, 246, 245].

- Ya a nivel semántico, se han realizado algunos experimentos iniciales acerca del empleo de relaciones de sinonimia ponderada [241, 240, 239] en base al trabajo de Fernández Lanza [71, 72]. La existencia de una medida del grado de sinonimia entre dos términos permite el establecimiento de umbrales a la hora de expandir consultas, o la ponderación de pesos al producirse correspondencias entre sinónimos.

- Finalmente, continuaremos el trabajo sobre la aplicación de nuestras soluciones para el español en la Búsqueda de Respuestas [164, 165] —especialmente en el caso del modelo basado en localidad— y en la Extracción de Información.

[1] Aunque la definición de *colocación* es ciertamente controvertida, podemos decir que dos palabras forman una colocación si coocurren regularmente en una lengua y el significado de la aparición conjunta es diferente a la simple suma de significados de las palabras individuales. Por ejemplo, el verbo *dar* suele coocurrir con el sustantivo *respiro*, y *darse un respiro* no es lo mismo que respirar ni quiere decir que nos entreguemos un respiro.

Parte IV

Apéndices

Apéndice A

Juego de Etiquetas de Erial

La siguiente lista constituye el juego de etiquetas correspondiente al proyecto Erial [31, 30] y empleado por el etiquetador MrTagoo. La lista consta de tres columnas: la primera columna es la etiqueta, la segunda columna contiene la descripción de la etiqueta, y la tercera una palabra de ejemplo. El cardinal de este conjunto es de 222 etiquetas.

Abreviatura	abreviatura	*Excmo.*
AQFP	adjetivo calificativo femenino plural	*rojas*
AQFS	adjetivo calificativo femenino singular	*roja*
AQMP	adjetivo calificativo masculino plural	*rojos*
AQMS	adjetivo calificativo masculino singular	*rojo*
AQNS	adjetivo calificativo neutro singular	*rojo*
ATFS	adjetivo de tratamiento femenino singular	*doña*
ATMS	adjetivo de tratamiento masculino singular	*don*
C	conjunción	*y*
Cifra	cifra	*125*
DAFP	determinante artículo femenino plural	*las*
DAFS	determinante artículo femenino singular	*la*
DAMP	determinante artículo masculino plural	*los*
DAMS	determinante artículo masculino singular	*el*
DANS	determinante artículo neutro singular	*lo*
DCFP	determinante numeral cardinal femenino plural	*doscientas*
DCFS	determinante numeral cardinal femenino singular	*una*
DCMP	determinante numeral cardinal masculino plural	*doscientos*
DCMS	determinante numeral cardinal masculino singular	*un*
DDFP	determinante demostrativo femenino plural	*estas*
DDFS	determinante demostrativo femenino singular	*esta*
DDMP	determinante demostrativo masculino plural	*estos*
DDMS	determinante demostrativo masculino singular	*este*
DDNS	determinante demostrativo neutro singular	*esto*
DEFP	determinante reflexivo femenino plural	*mismas*
DEFS	determinante reflexivo femenino singular	*misma*
DEMP	determinante reflexivo masculino plural	*mismos*
DEMS	determinante reflexivo masculino singular	*mismo*
DGFP	determinante globalizador femenino plural	*todas*
DGFS	determinante globalizador femenino singular	*toda*
DGMP	determinante globalizador masculino plural	*todos*

DGMS	determinante globalizador masculino singular	*todo*
DGNS	determinante globalizador neutro singular	*todo*
DIFP	determinante distributivo indefinido femenino plural	*cada*
DIFS	determinante distributivo indefinido femenino singular	*cada*
DIMP	determinante distributivo indefinido masculino plural	*cada*
DIMS	determinante distributivo indefinido masculino singular	*cada*
DLFP	determinante relativo femenino plural	*cuyas*
DLFS	determinante relativo femenino singular	*cuya*
DLMP	determinante relativo masculino plural	*cuyos*
DLMS	determinante relativo masculino singular	*cuyo*
DNFP	determinante indefinido femenino plural	*algunas*
DNFS	determinante indefinido femenino singular	*alguna*
DNMP	determinante indefinido masculino plural	*algunos*
DNMS	determinante indefinido masculino singular	*algún*
DOFP	determinante comparativo femenino plural	*más*
DOFS	determinante comparativo femenino singular	*más*
DOMP	determinante comparativo masculino plural	*más*
DOMS	determinante comparativo masculino singular	*más*
DPFP	determinante numeral partitivo femenino plural	*medias*
DPFS	determinante numeral partitivo femenino singular	*media*
DPMP	determinante numeral partitivo masculino plural	*medios*
DPMS	determinante numeral partitivo masculino singular	*medio*
DQFP	determinante cuantitativo femenino plural	*muchas*
DQFS	determinante cuantitativo femenino singular	*mucha*
DQMP	determinante cuantitativo masculino plural	*muchos*
DQMS	determinante cuantitativo masculino singular	*mucho*
DRFP	determinante numeral ordinal femenino plural	*primeras*
DRFS	determinante numeral ordinal femenino singular	*primera*
DRMP	determinante numeral ordinal masculino plural	*primeros*
DRMS	determinante numeral ordinal masculino singular	*primer*
DSFP	determinante posesivo femenino plural	*vuestras*
DSFS	determinante posesivo femenino singular	*vuestra*
DSMP	determinante posesivo masculino plural	*vuestros*
DSMS	determinante posesivo masculino singular	*vuestro*
DTFP	determinante interrogativo/exclamativo femenino plural	*cuántas*
DTFS	determinante interrogativo/exclamativo femenino singular	*cuánta*
DTMP	determinante interrogativo/exclamativo masculino plural	*cuántos*
DTMS	determinante interrogativo/exclamativo masculino singular	*cuánto*
Fecha	fecha	*12/10/1492*
I	interjección	*hola*
NCFP	sustantivo común femenino plural	*niñas*
NCFS	sustantivo común femenino singular	*niña*
NCMP	sustantivo común masculino plural	*niños*
NCMS	sustantivo común masculino singular	*niño*
P	signo de puntuación	*.*
PCFP	pronombre numeral cardinal femenino plural	*doscientas*
PCFS	pronombre numeral cardinal femenino singular	*una*
PCMP	pronombre numeral cardinal masculino plural	*doscientos*
PCMS	pronombre numeral cardinal masculino singular	*uno*

PDFP	pronombre demostrativo femenino plural	*éstas*
PDFS	pronombre demostrativo femenino singular	*ésta*
PDMP	pronombre demostrativo masculino plural	*éstos*
PDMS	pronombre demostrativo masculino singular	*éste*
PDNS	pronombre demostrativo neutro singular	*esto*
PEFP	pronombre reflexivo femenino plural	*mismas*
PEFS	pronombre reflexivo femenino singular	*misma*
PEMP	pronombre reflexivo masculino plural	*mismos*
PEMS	pronombre reflexivo masculino singular	*mismo*
PENS	pronombre reflexivo neutro singular	*mismo*
PGFP	pronombre globalizador femenino plural	*todas*
PGFS	pronombre globalizador femenino singular	*toda*
PGMP	pronombre globalizador masculino plural	*todos*
PGMS	pronombre globalizador masculino singular	*todo*
PGNS	pronombre globalizador neutro singular	*todo*
PIMS	pronombre distributivo indefinido	*cada cual*
PLFP	pronombre relativo femenino plural	*cuales*
PLFS	pronombre relativo femenino singular	*cual*
PLMP	pronombre relativo masculino plural	*cuales*
PLMS	pronombre relativo masculino singular	*cual*
PLNS	pronombre relativo neutro singular	*cual*
PNFP	pronombre indefinido femenino plural	*algunas*
PNFS	pronombre indefinido femenino singular	*alguna*
PNMP	pronombre indefinido masculino plural	*algunos*
PNMS	pronombre indefinido masculino singular	*alguno*
PNNS	pronombre indefinido neutro singular	*algo*
POFP	pronombre comparativo femenino plural	*más*
POFS	pronombre comparativo femenino singular	*más*
POMP	pronombre comparativo masculino plural	*más*
POMS	pronombre comparativo masculino singular	*más*
PONS	pronombre comparativo neutro singular	*más*
PPFP	pronombre numeral partitivo femenino plural	*medias*
PPFS	pronombre numeral partitivo femenino singular	*media*
PPMP	pronombre numeral partitivo masculino plural	*medios*
PPMS	pronombre numeral partitivo masculino singular	*medio*
PQFP	pronombre cuantitativo femenino plural	*muchas*
PQFS	pronombre cuantitativo femenino singular	*mucha*
PQMP	pronombre cuantitativo masculino plural	*muchos*
PQMS	pronombre cuantitativo masculino singular	*mucho*
PQNS	pronombre cuantitativo neutro singular	*mucho*
PRFP	pronombre numeral ordinal femenino plural	*primeras*
PRFS	pronombre numeral ordinal femenino singular	*primera*
PRMP	pronombre numeral ordinal masculino plural	*primeros*
PRMS	pronombre numeral ordinal masculino singular	*primero*
PRNS	pronombre numeral ordinal neutro singular	*primero*
PSFP	pronombre posesivo femenino plural	*vuestras*
PSFS	pronombre posesivo femenino singular	*vuestra*
PSMP	pronombre posesivo masculino plural	*vuestros*
PSMS	pronombre posesivo masculino singular	*vuestro*

PSNS	pronombre posesivo neutro singular	*vuestro*
PTFP	pronombre interrogativo/exclamativo femenino plural	*cuántas*
PTFS	pronombre interrogativo/exclamativo femenino singular	*cuánta*
PTMP	pronombre interrogativo/exclamativo masculino plural	*cuántos*
PTMS	pronombre interrogativo/exclamativo masculino singular	*cuánto*
PTNS	pronombre interrogativo/exclamativo neutro singular	*cuánto*
PY1P	pronombre personal 1ª persona plural	*nosotros*
PY1S	pronombre personal 1ª persona singular	*yo*
PY2P	pronombre personal 2ª persona plural	*vosotros*
PY2S	pronombre personal 2ª persona singular	*tú*
PY3P	pronombre personal 3ª persona plural	*ellos*
PY3S	pronombre personal 3ª persona singular	*él*
Sp00	nombre propio	*Álvarez*
Spfp	nombre propio femenino plural	*Olimpíadas*
Spfs	nombre propio femenino singular	*Córdoba*
Spmp	nombre propio masculino plural	*Celtics*
Spms	nombre propio masculino singular	*Ángel*
V1PAI	verbo, 1ª persona plural del perfecto de indicativo	*comimos*
V1PCI	verbo, 1ª persona plural del condicional de indicativo	*comeríamos*
V1PFI	verbo, 1ª persona plural del futuro de indicativo	*comeremos*
V1PFS	verbo, 1ª persona plural del futuro de subjuntivo	*comiéremos*
V1PII	verbo, 1ª persona plural del imperfecto de indicativo	*comíamos*
V1PIS	verbo, 1ª persona plural del imperfecto de subjuntivo	*comiéramos*
V1PRI	verbo, 1ª persona plural del presente de indicativo	*comemos*
V1PRM	verbo, 1ª persona plural del presente de imperativo	*comamos*
V1PRS	verbo, 1ª persona plural del presente de subjuntivo	*comamos*
V1SAI	verbo, 1ª persona singular del perfecto de indicativo	*comí*
V1SCI	verbo, 1ª persona singular del condicional de indicativo	*comería*
V1SFI	verbo, 1ª persona singular del futuro de indicativo	*comeré*
V1SFS	verbo, 1ª persona singular del futuro de subjuntivo	*comiere*
V1SII	verbo, 1ª persona singular del imperfecto de indicativo	*comía*
V1SIS	verbo, 1ª persona singular del imperfecto de subjuntivo	*comiera*
V1SRI	verbo, 1ª persona singular del presente de indicativo	*como*
V1SRS	verbo, 1ª persona singular del presente de subjuntivo	*coma*
V2PAI	verbo, 2ª persona plural del perfecto de indicativo	*comisteis*
V2PCI	verbo, 2ª persona plural del condicional de indicativo	*comeríais*
V2PFI	verbo, 2ª persona plural del futuro de indicativo	*comeréis*
V2PFS	verbo, 2ª persona plural del futuro de subjuntivo	*comiereis*
V2PII	verbo, 2ª persona plural del imperfecto de indicativo	*comíais*
V2PIS	verbo, 2ª persona plural del imperfecto de subjuntivo	*comierais*
V2PRI	verbo, 2ª persona plural del presente de indicativo	*com'eis*
V2PRM	verbo, 2ª persona plural del presente de imperativo	*comed*
V2PRS	verbo, 2ª persona plural del presente de subjuntivo	*com'ais*
V2SAI	verbo, 2ª persona singular del perfecto de indicativo	*comiste*
V2SCI	verbo, 2ª persona singular del condicional de indicativo	*comerías*
V2SFI	verbo, 2ª persona singular del futuro de indicativo	*comerás*
V2SFS	verbo, 2ª persona singular del futuro de subjuntivo	*comieres*
V2SII	verbo, 2ª persona singular del imperfecto de indicativo	*comías*
V2SIS	verbo, 2ª persona singular del imperfecto de subjuntivo	*comieras*

V2SRI	verbo, 2ª persona singular del presente de indicativo	*comes*
V2SRM	verbo, 2ª persona singular del presente de imperativo	*come*
V2SRS	verbo, 2ª persona singular del presente de subjuntivo	*comas*
V3PAI	verbo, 3ª persona plural del perfecto de indicativo	*comieron*
V3PCI	verbo, 3ª persona plural del condicional de indicativo	*comerían*
V3PFI	verbo, 3ª persona plural del futuro de indicativo	*comerán*
V3PFS	verbo, 3ª persona plural del futuro de subjuntivo	*comieren*
V3PII	verbo, 3ª persona plural del imperfecto de indicativo	*comían*
V3PIS	verbo, 3ª persona plural del imperfecto de subjuntivo	*comieran*
V3PRI	verbo, 3ª persona plural del presente de indicativo	*comen*
V3PRM	verbo, 3ª persona plural del presente de imperativo	*coman*
V3PRS	verbo, 3ª persona plural del presente de subjuntivo	*coman*
V3SAI	verbo, 3ª persona singular del perfecto de indicativo	*comió*
V3SCI	verbo, 3ª persona singular del condicional de indicativo	*comería*
V3SFI	verbo, 3ª persona singular del futuro de indicativo	*comerá*
V3SFS	verbo, 3ª persona singular del futuro de subjuntivo	*comiere*
V3SII	verbo, 3ª persona singular del imperfecto de indicativo	*comía*
V3SIS	verbo, 3ª persona singular del imperfecto de subjuntivo	*comiera*
V3SRI	verbo, 3ª persona singular del presente de indicativo	*come*
V3SRM	verbo, 3ª persona singular del presente de imperativo	*coma*
V3SRS	verbo, 3ª persona singular del presente de subjuntivo	*coma*
VPFP	participio verbal femenino plural	*comidas*
VPFS	participio verbal femenino singular	*comida*
VPMP	participio verbal masculino plural	*comidos*
VPMS	participio verbal masculino singular	*comido*
VPNS	participio verbal neutro singular	*comido*
VRG	gerundio verbal	*comiendo*
VRI	infinitivo verbal	*comer*
W	adverbio general	*rápidamente*
WA	adverbio de afirmación/negación/duda	*sí*
WI	adverbio de tiempo	*siempre*
WL	adverbio relativo	*adonde*
WM	adverbio de modo	*bien*
WO	adverbio comparativo	*más*
WP	adverbio de lugar	*aquí*
WQ	adverbio de cantidad	*mucho*
WT	adverbio interrogativo	*cuándo*
WX	adverbio exclamativo	*qué*
X	preposición	*de*
Zg00	sigla	*I+D*
Zgfp	sigla femenino plural	*APAS*
Zgfs	sigla femenino singular	*ONU*
Zgmp	sigla masculino plural	*GRAPO*
Zgms	sigla masculino singular	*PIB*

Apéndice B

Listas de *Stopwords*

B.1. Normalización mediante *Stemming*

B.1.1. Lista de *Stopwords* Generales

a	actualmente	adelante	además
afirmó	agregó	ahí	ahora
al	algo	algún	alguna
algunas	alguno	algunos	alrededor
ambos	ante	anterior	antes
añadió	apenas	aproximadamente	aquí
aseguró	así	aún	aunque
ayer	bajo	bien	buen
buena	buenas	bueno	buenos
cada	casi	cerca	cierto
cinco	comentó	como	cómo
con	conocer	considera	consideró
contra	cosas	creo	cual
cuales	cualquier	cuando	cuanto
cuatro	cuenta	da	dado
dan	dar	de	debe
deben	debido	decir	dejó
del	demás	dentro	desde
después	dice	dicen	dicho
dieron	diferente	diferentes	dijeron
dijo	dio	donde	dos
durante	e	ejemplo	el
él	ella	ellas	ello
ellos	embargo	en	encuentra
entonces	entre	era	eran
es	esa	esas	ese
eso	esos	esta	ésta
está	estaba	estaban	estamos
están	estar	estará	estas
éstas	este	éste	esto
estos	éstos	estoy	estuvo
ex	existe	existen	explicó

expresó
fin
fue
fuera
fueron
gran
grandes
ha
haber
había
habían
habrá
hace
hacen
hacer
hacerlo
hacia
haciendo
han
hasta
hay
haya
he
hecho
hemos
hicieron
hizo
hoy
hubo
igual
incluso
indicó
informó
junto
la
lado
las
le
les
llegó
lleva
llevar
lo
los
luego
lugar
manera
manifestó
más
mayor
me
mediante
mejor
mencionó
menos
mi
mientras
misma
mismas
mismo
mismos
momento
mucha
muchas
mucho
muchos
muy
nada
nadie
ni
ningún
ninguna
ningunas
ninguno
ningunos
no
nos
nosotras
nosotros
nuestra
nuestras
nuestro
nuestros
nueva
nuevas
nuevo
nuevos
nunca
o
ocho
otra
otras
otro
otros
para
parece
parte
partir
pasada
pasado
pero
pesar
poca
pocas
poco
pocos
podemos
podrá
podrán
podría
podrían
poner
por
porque
posible
primer
primera
primero
primeros
principalmente
propia
propias
propio
propios
próximo
próximos
pudo
pueda
puede
pueden
pues
que
qué
quedó
queremos
quien
quién
quienes
quiere
realizado
realizar
realizó
respecto
se
sea
sean
según
segunda
segundo
seis
señaló
ser
será
serán
sería
si
sí
sido
siempre
siendo
siete
sigue
siguiente
sin
sino
sobre
sola
solamente
solas
solo
sólo
solos
son
su
sus
tal
también
tampoco
tan
tanto
tendrá
tendrán
tenemos
tener
tenga
tengo
tenía
tenido
tercera
tiene
tienen
toda
todas
todavía
todo
todos

total	tras	trata	través
tres	tuvo	última	últimas
último	últimos	un	una
unas	uno	unos	usted
va	vamos	van	varias
varios	veces	ver	vez
y	ya	yo	

B.1.2. Lista de *Metastopwords*

comenta	comentan	comentar	comentará
comentarán	comente	comenten	considera
consideran	considerar	considerará	considerarán
considere	consideren	contendrá	contendrán
contener	contenga	contengan	contiene
contienen	describa	describan	describe
describen	describir	describirá	descripción
descripciones	detalla	detallada	detalladas
detallado	detallados	detallan	detallar
detallará	detallarán	detalle	detalle
detallen	detalles	dexcribirán	discuta
discutan	discute	discuten	discutir
discutirá	discutirán	documento	documentos
encontrar	encontrará	encontrarán	encuentra
encuentran	encuentre	encuentren	habla
hablan	hablar	hablará	hablarán
hable	hablen	incluir	incluirá
incluirán	incluya	incluyan	incluye
incluyen	informa	información	informaciones
informan	informar	informará	informarán
informe	informen	interesante	interesantes
irrelevante	irrelevantes	menciona	mencionan
mencionar	mencionará	mencionarán	mencione
mencionen	narra	narran	narrar
narrará	narrarán	narre	narren
noticia	noticias	proporciona	proporcionan
proporcionar	proporcionará	proporcionarán	proporcione
proporcionen	referencia	referencia	referencian
referenciar	referenciará	referenciarán	referencias
referencie	referencien	referir	referirá
referirán	refiera	refieran	refiere
refieren	relevante	relevantes	

B.2. Normalización mediante Lematización

B.2.1. Lista de *Stopwords* Generales

afirmar	agregar	añadir	anterior

asegurar	bueno	comentar	conocer
considerar	contar	cosa	creer
dar	deber	decir	dejar
diferente	ejemplo	estar	existir
explicar	expresar	fin	grande
haber	hacer	igual	indicar
informar	ir	llegar	llevar
lugar	manera	manifestar	mayor
mejor	mencionar	momento	nuevo
parecer	partir	poder	poner
posible	propio	quedar	querer
realizar	seguir	señalar	ser
siguiente	solo	tener	tratar
ultimo	venir	ver	vez

B.2.2. Lista de *Metastopwords*

comentar	considerar	contener	describir
descripcion	detallado	detallar	detalle
discutir	documento	encontrar	hablar
incluir	informacion	informar	interesante
irrelevante	mencionar	narrar	noticia
proporcionar	referencia	referenciar	referir
relevante			

Apéndice C

Estimación de Parámetros para la Realimentación

C.1. Consultas Cortas

α *(ratio α/β)*	.10 *(1)*			.20 *(.5)*			.40 *(.25)*			.80 *(.125)*		
t \ n_1	5	10	15	5	10	15	5	10	15	5	10	15
5	.4956	.4963	.4936	.5008	.5009	.4980	.5074	.5071	.5046	.5150	.5129	.5104
10	.5037	.5055	.4975	.5083	.5099	.5019	.5142	.5154	.5087	.5220	.5220	.5166
15	.5051	.5038	.5041	.5098	.5084	.5079	.5166	.5146	.5137	.5242	.5221	.5210
20	.4996	.5076	.5091	.5051	.5120	.5133	.5131	.5184	.5197	.5216	.5258	.5268
30	.4992	.5132	.5104	.5056	.5179	.5145	.5148	.5248	.5207	.5253	.5328	.5279
40	.4963	.5143	.5105	.5021	.5190	.5146	.5121	.5269	.5210	.5239	.5348	.5294
50	.4951	.5139	.5074	.5013	.5184	.5123	.5115	.5264	.5196	.5235	.5349	.5283
α *(ratio α/β)*	1.20 *(.0833)*			1.80 *(.0555)*			2.40 *(.0416)*			3.20 *(.0312)*		
t \ n_1	5	10	15	5	10	15	5	10	15	5	10	15
5	.5168	.5150	.5124	.5170	.5154	.5129	.5168	.5143	.5121	.5154	.5123	.5107
10	.5254	.5255	.5201	.5266	.5267	.5218	.5267	.5260	.5213	.5247	.5244	.5195
15	.5278	.5262	.5240	.5299	.5289	.5265	.5301	.5284	.5262	.5294	.5276	.5244
20	.5269	.5296	.5300	.5306	.5322	.5321	.5307	.5322	.5317	.5302	.5309	.5295
30	.5303	.5360	.5315	.5340	.5381	.5330	.5367	.5377	.5329	.5369	.5363	.5317
40	.5306	.5384	.5338	.5344	.5408	.5359	.5365	.5411	.5358	.5367	.5401	.5350
50	.5298	.5391	.5333	.5346	.5415	.5364	.5360	.5420	.5370	.5371	.5413	.5362

Tabla C.1: Precisiones obtenidas para consultas cortas durante el proceso de estimación de parámetros para la realimentación: α, número de documentos n_1 y número de términos t. (β=0.1, γ=0; precisión sin realimentación: .4829)

C.2. Consultas Largas

α *(ratio α/β)*	.10 *(1)*			.20 *(.5)*			.40 *(.25)*			.80 *(.125)*		
t \ n_1	5	10	15	5	10	15	5	10	15	5	10	15
5	.5252	.5154	.5198	.5324	.5209	.5247	.5427	.5294	.5319	.5517	.5378	.5397
10	.5208	.5210	.5301	.5283	.5275	.5348	.5400	.5362	.5421	.5540	.5461	.5492
15	.5188	.5198	.5349	.5263	.5264	.5395	.5386	.5353	.5463	.5537	.5462	.5544
20	.5143	.5271	.5312	.5212	.5332	.5364	.5329	.5421	.5440	.5477	.5532	.5533
30	.5076	.5248	.5319	.5148	.5308	.5368	.5268	.5410	.5446	.5423	.5532	.5544
40	.5050	.5264	.5317	.5118	.5326	.5365	.5235	.5433	.5438	.5393	.5560	.5541
50	.5086	.5220	.5324	.5151	.5282	.5370	.5265	.5385	.5450	.5428	.5519	.5554
α *(ratio α/β)*	1.20 *(.0833)*			1.80 *(.0555)*			2.40 *(.0416)*			3.20 *(.0312)*		
t \ n_1	5	10	15	5	10	15	5	10	15	5	10	15
5	.5554	.5414	.5433	.5573	.5440	.5448	.5546	.5442	.5444	.5527	.5427	.5432
10	.5604	.5509	.5522	.5627	.5531	.5536	.5619	.5530	.5530	.5597	.5518	.5515
15	.5624	.5507	.5576	.5660	.5535	.5590	.5658	.5541	.5587	.5641	.5536	.5566
20	.5576	.5584	.5575	.5647	.5615	.5599	.5668	.5616	.5598	.5663	.5598	.5586
30	.5532	.5601	.5596	.5630	.5645	.5630	.5672	.5647	.5631	.5695	.5637	.5621
40	.5510	.5633	.5602	.5601	.5677	.5645	.5652	.5685	.5653	.5683	.5680	.5647
50	.5533	.5605	.5621	.5629	.5663	.5666	.5673	.5686	.5679	.5708	.5685	.5675

Tabla C.2: Precisiones obtenidas para consultas largas durante el proceso de estimación de parámetros para la realimentación: α, número de documentos n_1 y número de términos t. (β=0.1, γ=0; precisión sin realimentación: .5239)

Apéndice D

Sufijos Considerados

D.1. Sufijos Nominalizadores

D.1.1. Nominalizadores Denominales

-ada

Acentuación: llana.
Significado:

- Movimiento realizado con o golpe producido por la base.
  Ejemplo: *piedra* → *pedrada*.
- Acto propio de (base animada, significado peyorativo).
  Ejemplo: *novato* → *novatada*.
- Medida o magnitud acotada por el objeto.
  Ejemplo: *cuchara* → *cucharada*.

-ado/-ato/-azgo

Acentuación: llana.
Significado:

- Grupo de animados implícitos en la base.
  Ejemplo: *campesino* → *campesinado*.
- Denominaciones de cargos y oficios, oficina o dependencia.
  Ejemplo: *rector* → *rectorado*.

-aje

Acentuación: llana.
Significado:

- Colectivo, en ocasiones con matices despectivos.
  Ejemplo: *ropa* → *ropaje*.
- Similar al *-ado* correspondiente a oficios.
  Ejemplo: *bandido* → *bandidaje*.
- Acción propia de la base.
  Ejemplo: *piloto* → *pilotaje*.
- Medida o proporción.
  Ejemplo: *octano* → *octanaje*.

-al/-ar

Acentuación: aguda.
Significado:

- Lugar de cultivo o conjunto de plantas designadas por la base.
  Ejemplo: *patata* → *patatal*.
- Árbol cuyo fruto es la base.
  Ejemplo: *pera* → *peral*.
- Colectivo de árboles o plantas designadas por la base.
  Ejemplo: *pino* → *pinar*.

-ería

Acentuación: llana.
Significado:

- Lugar de venta o manufactura del primitivo (el comerciante o propietario asociado suele derivarse mediante *-ero*).
  Ejemplo: *fruta* → *frutería*.

- Nombre de cualidad, abstracción del primitivo (matices colectivos o peyorativos).
  Ejemplo: *brujo → brujería.*
- Colectivo.
  Ejemplo: *chiquillo → chiquillería.*
- Acción o dicho propio de la base.
  Ejemplo: *tonto → tontería.*

-ero(a)

Acentuación: llana.
Significado:

- Receptáculo de objetos concretos.
  Ejemplo: *ensalada → ensaladera.*
- Uso agentivo, donde la base es el material u objeto utilizado; profesión.
  Ejemplo: *pistola → pistolero.*
- Árboles.
  Ejemplo: *limón → limonero.*
- Tipo de lugar.
  Ejemplo: *perro → perrera.*
- Colectivo.
  Ejemplo: *abeja → abejera.*

-ismo

Acentuación: llana.
Significado:

- Ideología, creencia, partido, sistema, doctrina.
  Ejemplo: *Marx → marxismo.*
- Derivación paralela a los nombres y adjetivos en *-ista.*
  Ejemplo: *fascista → fascismo.*

-ista

Acentuación: llana.
Significado:

- Derivación paralela a los sustantivos en *-ismo.*
  Ejemplo: *fascismo → fascista.*
- Miembro de un movimiento.
  Ejemplo: *Castro → castrista.*
- Profesión, ocupación artística, deportista.
  Ejemplo: *guión → guionista.*
- Persona que tiene la costumbre de hacer lo indicado por la base.
  Ejemplo: *camorra → camorrista.*

-ario

Acentuación: llana.
Significado:

- Colectivo; serie, índice, registro, repertorio.
  Ejemplo: *poema → poemario.*
- Lugar.
  Ejemplo: *campana → campanario.*
- Profesión.
  Ejemplo: *empresa → empresario.*

-ía

Acentuación: llana.
Significado:

- Nombre colectivo.
  Ejemplo: *ciudadano → ciudadanía.*
- Establecimiento comercial.
  Ejemplo: *carpintero → carpintería.*
- Acción.
  Ejemplo: *grosero → grosería.*
- Dignidad, cargo.
  Ejemplo: *alcalde → alcaldía.*
- Ideología, creencia, partido, sistema, doctrina.
  Ejemplo: *monarca → monarquía.*

-azo/-etazo/-otazo/-onazo

Acentuación: llana.
Significado: golpe, movimiento súbito.
Ejemplo: *zarpa → zarpazo.*

D.1.2. Nominalizadores Deadjetivales

-ancia/-encia/-iencia

Acentuación: llana.
Significado: significación variada, pero siempre derivados a partir de términos en *-ante/-ente/-iente/-ento*.
Ejemplo: *tolerante → tolerancia*.

-ería

Acentuación: llana.
Significado:

- Nombre de acción (hecho o dicho de connotación negativa).
  Ejemplo: *grosero → grosería*.
- Nombre abstracto de cualidad o estado (de base frecuentemente peyorativa).
  Ejemplo: *tacaño → tacañería*.

-ez(a)

Acentuación: aguda (llana).
Significado:

- Nombre de acción (hecho o dicho de connotación negativa).
  Ejemplo: *estúpido → estupidez*.
- Nombre abstracto de cualidad o estado (de base frecuentemente peyorativa).
  Ejemplo: *bello → belleza*.

-ismo

Acentuación: llana.
Significado:

- Ideología, creencia, partido, sistema, doctrina.
  Ejemplo: *católico → catolicismo*.
- Nombre abstracto de cualidad o estado.
  Ejemplo: *pasota → pasotismo*.

-ura

Acentuación: llana.
Significado:

- Nombre de acción (hecho o dicho de connotación negativa).
  Ejemplo: *chalado → chaladura*.
- Nombre abstracto de cualidad o estado.
  Ejemplo: *cuerdo → cordura*.

-ía

Acentuación: llana.
Significado:

- Nombre de acción (hecho o dicho de connotación negativa).
  Ejemplo: *majadero → majadería*.
- Nombre abstracto de cualidad o estado.
  Ejemplo: *lejano → lejanía*.
- Ideología, creencia, partido, sistema, doctrina.
  Ejemplo: *monarca → monarquía*.

-era

Acentuación: llana.
Significado: nombre abstracto de cualidad o estado.
Ejemplo: *ciego → ceguera*.

-or

Acentuación: aguda.
Significado: nombre abstracto de cualidad o estado.
Ejemplo: *verde → verdor*.

-tud

Acentuación: aguda.
Significado: nombre abstracto de cualidad o estado.
Ejemplo: *amplio → amplitud*.

-dad/-edad/-idad/-tad

Acentuación: aguda.
Significado:

- Nombre de acción (hecho o dicho de connotación negativa).
  Ejemplo: *ruin → ruindad*.

- Nombre abstracto de cualidad o estado.
  Ejemplo: *serio → seriedad.*

-ada

Acentuación: llana.
Significado: acto propio de la base (base animada; significado peyorativo).
Ejemplo: *brabucón → brabuconada.*

D.1.3. Nominalizadores Deverbales

-ado/-ido/-ato

Acentuación: llana.
Significado: acción y efecto.
Ejemplo: *encender → encendido.*

-ción/-ación/-ición/-sión

Acentuación: aguda.
Significado: acción y/o efecto.
Ejemplo: *germinar → germinación.*

-miento/-mento/-amiento/-imiento

Acentuación: llana.
Significado: acción y/o efecto.
Ejemplo: *seguir → seguimiento.*

-aje

Acentuación: llana.
Significado: acción y efecto, materialización (similar *-miento/-ción*).
Ejemplo: *embalar → embalaje.*

-ante/-iente/-iente

Acentuación: llana.
Significado: nombre de agente, profesión.
Ejemplo: *delinear → delineante.*

-dero/-edero/-idero

Acentuación: llana.
Significado:

- Lugar donde se produce la acción del verbo.
  Ejemplo: *embarcar → embarcadero.*
- (*-dera/-edera/-idera*) Que hace o realiza, que sirve para.
  Ejemplo: *tapar → tapadera.*

-or/-tor/-ador/-edor/-idor/(a)

Acentuación: aguda (llana).
Significado:

- Nombre de agente, profesión.
  Ejemplo: *sabotear → saboteador.*
- Equipo instrumental.
  Ejemplo: *destornillar → destornillador.*
- Lugar donde se lleva a cabo la acción señalada.
  Ejemplo: *comer → comedor.*

-dura

Acentuación: llana.
Significado: acción y/o efecto, con tendencia a la materialización.
Ejemplo: *torcer → torcedura.*

-ido

Acentuación: llana.
Significado: sonido emitido, prolongado y repetido, o efecto de dicha emisión.
Ejemplo: *rugir → rugido.*

-ada/-ata

Acentuación: llana.
Significado: acción repentina o enérgica; acción y/o efecto.
Ejemplo: *caminar → caminata.*

-azón

Acentuación: aguda.
Significado: acción y/o efecto.
Ejemplo: *picar → picazón.*

-ón

Acentuación: aguda.
Significado: acción y/o efecto, generalmente denotando algo repentino o violento.
Ejemplo: *empujar → empujón.*

D.2. Sufijos Adjetivizadores

D.2.1. Adjetivizadores Denominales

-al

Acentuación: aguda.
Significado: equivalente a un complemento preposicional.
Ejemplo: *primavera → primaveral.*

-ario

Acentuación: llana.
Significado: referente a la base (léxico político o económico).
Ejemplo: *parlamento → parlamentario.*

-ero

Acentuación: llana.
Significado:

- Equivalente a un complemento preposicional.
  Ejemplo: *pesca → pesquero.*
- Formaciones ad hoc con tono peyorativo.
  Ejemplo: *zarzuela → zarzuelero.*
- Designan el carácter, principalmente de personas.
  Ejemplo: *traición → traicionero.*

-ista

Acentuación: llana.
Significado:

- Doblete con dimensión política o deportiva sobre base gentilícea.
  Ejemplo: *europeo → europeísta.*
- Acciones provocadas por el personaje de la base.
  Ejemplo: *Castro → castrista.*

-ístico

Acentuación: esdrújula.
Significado: relativo a. Pertenece al grupo de *-ista*/*-ismo*, pudiendo existir o no una forma intermedia en *-ista.*
Ejemplo: *humor → humorista → humorística*, *boxeo → boxístico.*

-oso

Acentuación: llana.
Significado:

- Abundancia respecto la base (abundante en, dotado de).
  Ejemplo: *arcilla → arcilloso.*
- Valor atributivo (ocasionalmente peyorativo).
  Ejemplo: *verde → verdoso.*

-ico

Acentuación: átono.
Significado: relativo a, propio de.
Ejemplo: *álgebra → algebraico.*

-udo

Acentuación: llana.
Significado:

- Abundancia, preferentemente relativo a partes del cuerpo (tendencia a significado peyorativo si es un atributo físico negativo, o afectivo si es positivo).
  Ejemplo: *panza → panzudo.*
- Con forma de.
  Ejemplo: *gancho → ganchudo.*

-iano

Acentuación: llana.
Significado:

- Con base referente a un lugar habitado que designa demarcación administrativa o eclesiástica, institución, accidente geográfico, etc.
  Ejemplo: *diócesis → diocesano.*
- Relativo a una persona dada (nombre propio)[1].
  Ejemplo: *(Virgen) María → mariano.*

[1] Con este significado presenta un alomorfo, *-iano*; p.ej., *Cristo → cristiano*

-ino

Acentuación: llana.
Significado:

- Relativo a una persona dada (nombre propio).
  Ejemplo: *Isabel → isabelino.*

- Con base nombre de materia (sobre todo gemas) o color.
  Ejemplo: *diamante → diamantino.*

-eño

Acentuación: llana.
Significado: hecho de, relativo a, propio de.
Ejemplo: *marfil → marfileño.*

D.2.2. Adjetivizadores Deadjetivales

-ísimo

Acentuación: esdrújula.
Significado: intensificador.
Ejemplo: *feo → feísimo.*

D.2.3. Adjetivizadores Deverbales

-able/-ible

Acentuación: llana.
Significado: capaz de.
Ejemplo: *defender → defendible.*

-dor/-ador/-edor/-idor/(a)

Acentuación: aguda (llana).
Significado: alternativa a una oración de relativo.
Ejemplo: *acoger → acogedor.*

-ante/-(i)ente

Acentuación: llana.
Significado: que realiza la acción del verbo (semejante a *-dor*).
Ejemplo: *crecer → creciente.*

-dizo/-adizo/-edizo/-idizo

Acentuación: llana.
Significado: indican propensión o aptitud para recibir la acción del verbo base.
Ejemplo: *quebrar → quebradizo.*

-ivo/-tivo/-sivo

Acentuación: llana.
Significado: relativo a.
Ejemplo: *defender → defensivo.*

D.3. Sufijos Verbalizadores

D.3.1. Verbalizadores Denominales

-ar/a--ar/en--ar

Acentuación: aguda.
Ejemplo: *orquesta → orquestar.*

-ear

Acentuación: aguda.
Ejemplo: *agujero → agujerear.*

-ificar

Acentuación: aguda.
Ejemplo: *gas → gasificar.*

-izar/a--izar/en--izar

Acentuación: aguda.
Ejemplo: *vapor → vaporizar.*

D.3.2. Verbalizadores Deadjetivales

-ar/a--ar/en--ar

Acentuación: aguda.
Ejemplo: *gordo → engordar.*

-ear

Acentuación: aguda.
Ejemplo: *amarillo → amarillear.*

-ificar

Acentuación: aguda.
Ejemplo: *puro → purificar.*

-izar/a--izar/en--izar

Acentuación: aguda.
Ejemplo: *visible* → *visibilizar*.

-ecer/a--ecer/en--ecer

Acentuación: aguda.
Ejemplo: *húmedo* → *humadecer*.

Apéndice E

Patrones Empleados por el Analizador Sintáctico Superficial
PATTERNS

E.1. Subpatrones Componente

Antes de describir los patrones empleados, describiremos ciertos *subpatrones* de frecuente utilización que nos permitirán simplificar las expresiones de los patrones. De esta forma, dado un subpatrón S dentro de un patrón P, la expresión regular correspondiente al subpatrón se representará de forma abreviada mediante el identificador del subpatrón. Por otra parte, en lo que respecta a las dependencias sintácticas asociadas al patrón, el establecimiento de una dependencia entre un término T_i del patrón y el subpatrón S implica que se establecerá una dependencia entre el término T_i y cada una de las palabras con contenido del subpatrón S.

SUBPATRÓN I: **Sintagma adjetivo con modificador adverbial opcional**

$$\overbrace{\underbrace{\text{W? A}_1}}^{\text{SA}_\text{w}}$$

El primero de nuestros subpatrones componente corresponde a un sintagma adjetival simple, identificado mediante SAw, formado por un núcleo adjetival que puede verse modificado opcionalmente por un adverbio. El adjetivo núcleo será el elemento con el cual se cree la dependencia en caso de que se establezca una dependencia sintáctica entre el sintagma adjetival y un segundo término. Por ejemplo: $muy\ \overbrace{fuerte}^{A_1}$.

$$\overbrace{\quad\quad}^{SAw}$$

SUBPATRÓN II: **Coordinación opcional de sintagmas adjetivos con modificadores adverbiales**

$$\overbrace{\underbrace{\text{SA}_\text{w}_1\text{(Cc SA}_\text{w}_2\text{)?}}}^{\text{SA}_\text{c}}$$

El subpatrón SAc permite englobar tanto las apariciones de sintagmas adjetivo simples SAw, como de coordinaciones de los mismos. Las dependencias establecidas con el subpatrón darán lugar a sendas dependencias con los respectivos núcleos adjetivales de los sintagmas adjetivales más simples. Por ejemplo: Por ejemplo: $\dfrac{\dfrac{\underline{alto}}{A_1}}{SAw_1}\ y\ \dfrac{muy\ \dfrac{\underline{fuerte}}{A_2}}{SAw_2}$ $\Big/\ SAc$

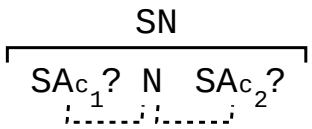
E.2. Patrones de Análisis

Para cada patrón mostraremos la estructura sintáctica que representa junto con las dependencias sintácticas a extraer (en línea discontínua). También se hará una pequeña descripción del mismo para, a continuación, mostrar un ejemplo de su aplicación.

E.2.1. Sintagmas Nominales: Dependencias Sustantivo–Adjetivo

En este primer bloque de patrones de análisis se recogen aquellos patrones encargados del reconocimiento y análisis de sintagmas nominales con dependencias entre sustantivos y adjetivos modificadores.

PATRÓN SA01: **Sustantivo modificado por un adjetivo o coordinación de adjetivos en posición(es) anterior y/o posterior**

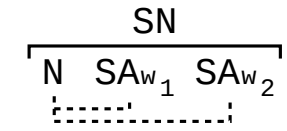


Este patrón captura las dependencias existentes entre un sustantivo y los sintagmas adjetivos simples y/o coordinados que lo anteceden y/o siguen para modificarlo. De esta forma se extraen las dependencias existentes entre el sustantivo núcleo del sintagma y cada uno de los adjetivos que lo modifican. Obsérvese que en la representación del patrón hemos hecho uso del subpatrón componente SAc, estableciendo dependencias entre el subpatrón y el sustantivo, pero a la hora de extraer las dependencias se extraen los pares correspondientes al sustantivo y cada uno de los adjetivos del sintagma adjetival, tal y como ilustra el ejemplo que se incluye a continuación.

Ejemplo: *una* $\underset{A_1}{\underline{hermosa}}$ $\underset{N_1}{\underline{doncella}}$ $\underset{A_2}{\underline{joven}}$ *y* $\underset{A_3}{\underline{dichosa}}$

$$(A_1\,,\ N_1) \longrightarrow (hermosa\,,\ doncella)$$
$$(N_1\,,\ A_2) \longrightarrow (doncella\,,\ joven\ \ )$$
$$(N_1\,,\ A_3) \longrightarrow (doncella\,,\ dichosa\ )$$

PATRÓN SA02: **Sustantivo modificado por sintagmas adjetivos consecutivos**



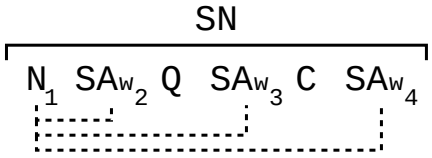
Este patrón captura las dependencias correspondientes a un sintagma nominal formado por un sustantivo modificado por dos sintagmas adjetivos simples. Las dependencias extraídas son

aquéllas existentes entre el sustantivo y cada uno de los adjetivos modificadores.

Ejemplo: *una* $\underset{N_1}{\underline{camioneta}}$ $\underset{A_1}{\underline{blanca}}$ $\underset{A_2}{\underline{destartalada}}$

$$(N_1,\ A_1) \longrightarrow (camioneta,\ blanca\ \ \ \ \ \ )$$
$$(N_1,\ A_2) \longrightarrow (camioneta,\ destartalada)$$

PATRÓN SA03: **Sustantivo modificado por una enumeración**

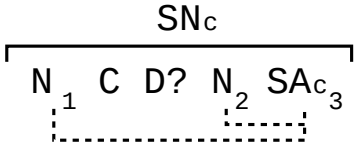


Este patrón recoge el supuesto de un sustantivo modificado por una enumeración de sintagmas adjetivos simples, formando todos ellos un sintagma nominal. Las dependencias extraídas son, al igual que en los casos anteriores, las existentes entre el sustantivo y cada uno de los adjetivos modificadores.

Ejemplo: *un* $\underset{N_1}{\underline{jersey}}$ $\underset{A_1}{\underline{bueno}}$, $\underset{A_2}{\underline{bonito}}$ *y* $\underset{A_3}{\underline{barato}}$

$$(N_1,\ A_1) \longrightarrow (jersey,\ bueno)$$
$$(N_1,\ A_2) \longrightarrow (jersey,\ bonito)$$
$$(N_1,\ A_3) \longrightarrow (jersey,\ barato)$$

PATRÓN SA04: **Sintagma nominal coordinado modificado por un adjetivo o coordinación de adjetivos**

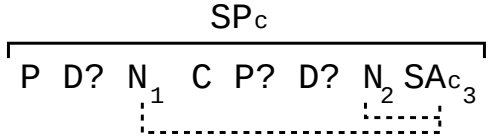


Este nuevo patrón cubre el caso de la existencia de sintagmas nominales coordinados donde dos sustantivos coordinados son modificados por un sintagma adjetivo. Las dependencias estraídas son aquéllas existentes entre los sustantivos y cada uno de sus adjetivos modificadores.

Ejemplo: $\underset{N_1}{\underline{hombres}}$ *y* $\underset{N_2}{\underline{mujeres}}$ $\underset{A_1}{\underline{felices}}$ *y* $\underset{A_2}{\underline{despreocupados}}$

$$(N_1,\ A_1) \longrightarrow (hombres,\ felices\ \ \ \ \ \ )$$
$$(N_1,\ A_2) \longrightarrow (hombres,\ despreocupados)$$
$$(N_2,\ A_1) \longrightarrow (mujeres\ ,\ felices\ \ \ \ \ \ )$$
$$(N_2,\ A_2) \longrightarrow (mujeres\ ,\ despreocupados)$$

PATRÓN SA05: **Sintagma preposicional coordinado modificado por un adjetivo o coordinación de adjetivos**

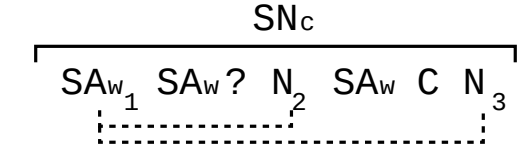


Se trata del patrón equivalente al anterior para el caso de dos sintagmas preposicionales coordinados y modificados conjuntamente por un sintagma adjetivo. Las dependencias estraídas son aquéllas existentes entre los sustantivos y cada uno de sus adjetivos modificadores. Adviértase que la preposición del segundo complemento nominal no es obligatoria, ya que en caso de ser la misma que la del primero complemento nominal es frecuente obviarla.

Ejemplo: *para* $\underset{N_1}{\underline{hombres}}$ *y* $\underset{N_2}{\underline{mujeres}}$ $\underset{A_1}{\underline{felices}}$ *y* $\underset{A_2}{\underline{despreocupados}}$

$$(N_1, A_1) \longrightarrow (hombres, felices)$$
$$(N_1, A_2) \longrightarrow (hombres, despreocupados)$$
$$(N_2, A_1) \longrightarrow (mujeres, felices)$$
$$(N_2, A_2) \longrightarrow (mujeres, despreocupados)$$

PATRÓN SA06: **Sintagma adjetivo en posición anterior modificando una coordinación**

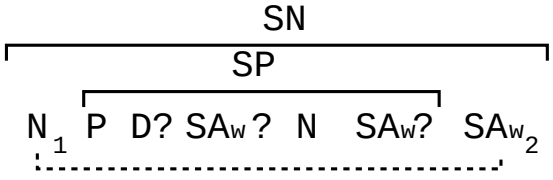


Similar a los dos anteriores, pero con el adjetivo modificador en posición anterior.

Ejemplo: *lleno de* $\underset{A_1}{\underline{agradables}}$ $\underset{N_1}{\underline{olores}}$ *y* $\underset{N_2}{\underline{sabores}}$

$$(A_1, N_1) \longrightarrow (agradables, olores)$$
$$(A_1, N_2) \longrightarrow (agradables, sabores)$$

PATRÓN SA07: **Sintagma adjetivo modificando un sustantivo y su complemento nominal**



En este caso el patrón cubre el caso de la existencia de un sintagma adjetivo simple que modifica un sustantivo que también es modificado previamente por un complemento nominal.

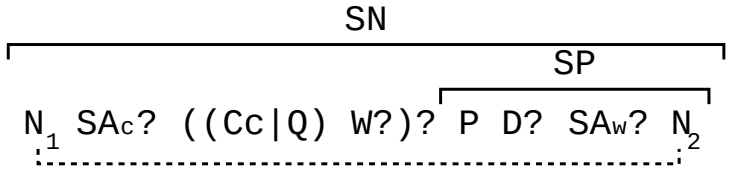
Ejemplo: *los* $\underset{N_1}{\underline{corpus}}$ *de entrenamiento* $\underset{A_1}{\underline{disponibles}}$

$$(N_1\,,\ A_1) \longrightarrow (corpus\,,\ disponibles)$$

E.2.2. Sintagmas Nominales: Dependencias Sustantivo–Complemento Nominal

Una vez descritos los patrones correspondientes a las dependencias sustantivo–adjetivo, corresponde turno a las otras dependencias propias de los sintagmas nominales, las existentes entre el sustantivo núcleo y los complementos nominales que lo modifican.

PATRÓN CN01: **Sustantivo modificado por un complemento nominal**

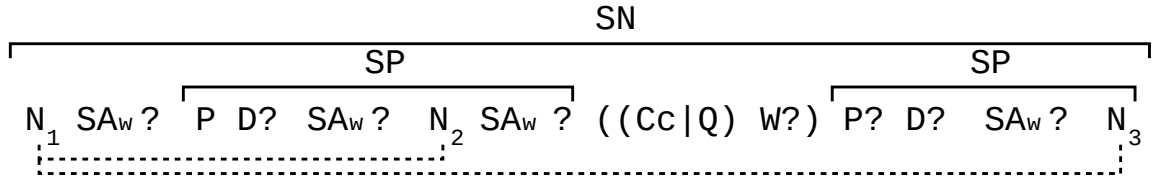


Este primer patrón recoge el caso más simple, el de un sustantivo modificado por un único complemento nominal. El sustantivo puede estar también modificado por sintagmas adjetivales, si bien el patrón extrae únicamente la dependencia sustantivo–complemento nominal, dejando que sean los patrones corespondientes quienes extraigan las posibles dependencias sustantivo–adjetivo. Se contempla también la posibilidad de la existencia de una pausa entre el sustantivo y el complemento o bien que el complemento esté coordinado con el sintagma adjetival.

Ejemplo: *un* $\underset{N_1}{\underline{personaje}}$ *muy inteligente, y con* $\underset{N_1}{\underline{carisma}}$

$$(N_1\,,\ N_1) \longrightarrow (personaje\,,\ carisma)$$

PATRÓN CN02: **Sustantivo modificado por una coordinación de complementos nominales**

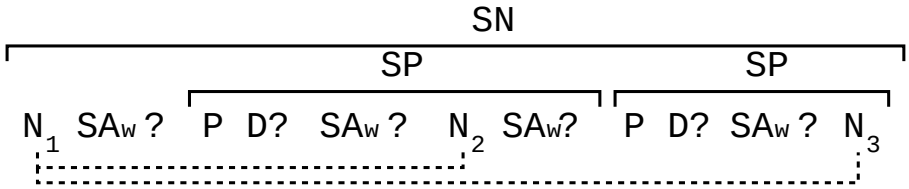


Esta nueva expresión recoge el caso de la existencia de un sintagma nominal en el que el núcleo sea modificado por una coordinación de complementos nominales. Las dependencias se establecen entre el núcleo del sintagma nominal y cada uno de los núcleos de los sintagmas preposicionales que actúan como complementos nominales suyos. Obsérvese que la preposición del segundo complemento nominal no es obligatoria, ya que en caso de ser la misma que la del primero complemento nominal es frecuente obviarla.

Ejemplo: $\underset{N_1}{\underline{ropajes}}$ *de fina* $\underset{N_2}{\underline{seda}}$ *y lujoso* $\underset{N_3}{\underline{oro}}$

$$(N_1\,,\ N_2) \longrightarrow (ropajes\,,\ seda)$$
$$(N_1\,,\ N_3) \longrightarrow (ropajes\,,\ oro\ )$$

PATRÓN CN03: **Sustantivo modificado por dos complementos no coordinados**

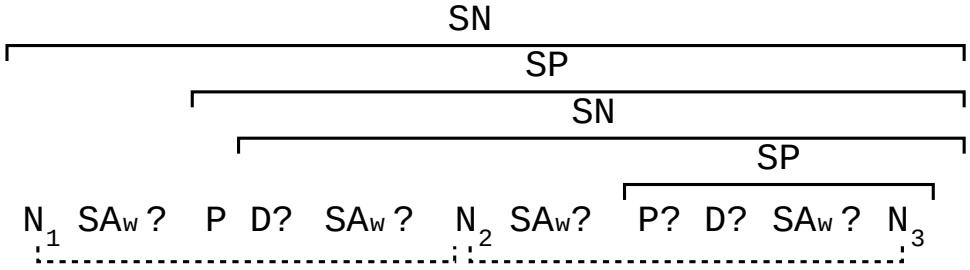


El patrón cubre el supuesto de la existancia de sintagmas nominales donde el núcleo está siendo modificado de forma simultánea por dos complementos nominales no coordinados. De nuevo, las dependencias se establecen entre el núcleo del sintagma nominal y cada uno de los núcleos de los complementos nominales.

Ejemplo: *su* $\underset{N_1}{\underline{precio}}$ *inicial de* $\underset{N_2}{\underline{salida}}$ *en* $\underset{N_3}{\underline{dólares}}$

$$(N_1, N_2) \longrightarrow (precio, salida\ )$$
$$(N_1, N_3) \longrightarrow (precio, dólares)$$

PATRÓN CN04: **Sustantivo modificado por un complemento modificado a su vez por un segundo complemento**

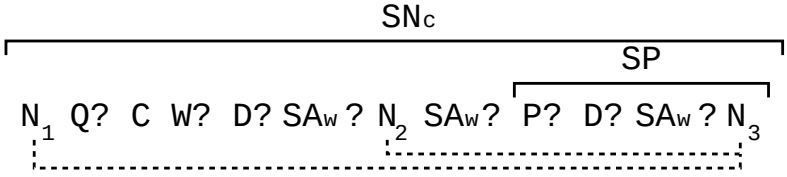


Se decribe el caso en el que un sustantivo se ve modificado por un complemento preposicional que es a su vez modificado por otro complemento. Las dependencias se establecen entre los correspondientes términos modificados y sus modificadores, entre el sustantivo núcleo del sintagma nominal total y el núcleo del primer complemento nominal, y entre éste último y el núcleo del segundo complemento nominal.

Ejemplo: *una* $\underset{N_1}{\underline{empanada}}$ *de* $\underset{N_2}{\underline{berberechos}}$ *de la* $\underset{N_3}{\underline{ría}}$

$$(N_1, N_2) \longrightarrow (empanada\ , berberechos)$$
$$(N_1, N_3) \longrightarrow (berberechos, ría\ \ \ \ \ \ )$$

PATRÓN CN05: **Sintagma nominal coordinado modificado por un complemento nominal**

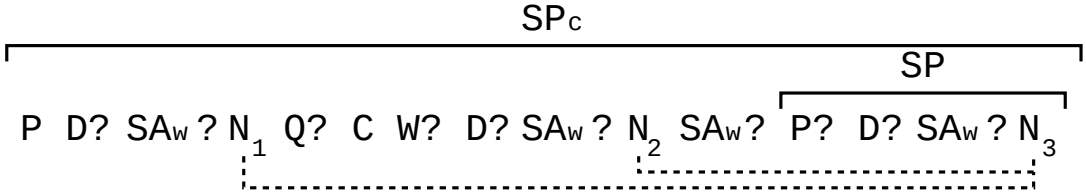


Este patrón captura las dependencias existentes entre un sintagma nominal coordinado y el único complemento preposicional por el que está siendo modificado. Las dependencias extraídas son aquéllas entre cada uno de los sustantivos coordinados y el sustantivo núcleo del complemento que está modificando a ambos.

Ejemplo: *quiero* $\underset{N_1}{\underline{caramelos}}$ *y* $\underset{N_2}{\underline{gominolas}}$ *de muchos* $\underset{N_3}{\underline{sabores}}$

$$(N_1\,,\ N_3) \longrightarrow (caramelos\,,\ sabores)$$
$$(N_2\,,\ N_3) \longrightarrow (gominolas\,,\ sabores)$$

PATRÓN CN06: **Sintagma preposicional coordinado modificado por un complemento nominal**

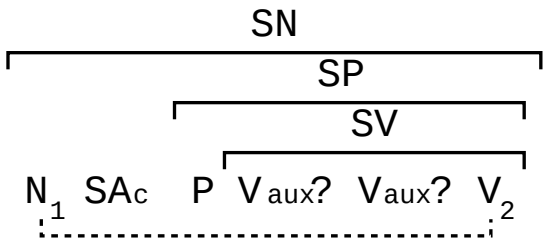


Patrón equivalente al anterior para el caso de sintagmas preposicionales coordinados modificados por un único complemento. Las dependencias extraídas son aquéllas existentes entre cada uno de los núcleos de los sintagmas preposicionales coordinados y el núcleo del complemento modificador de ambos. Obsérvese que la preposición del segundo sintagma coordinado no es obligatoria, ya que en caso de ser la misma que la del primer complemento nominal es frecuente su eliminación.

Ejemplo: *de* $\underset{N_1}{\underline{caramelos}}$ *y* $\underset{N_2}{\underline{gominolas}}$ *de muchos* $\underset{N_3}{\underline{sabores}}$

$$(N_1\,,\ N_3) \longrightarrow (caramelos\,,\ sabores)$$
$$(N_2\,,\ N_3) \longrightarrow (gominolas\,,\ sabores)$$

PATRÓN CN07: **Sustantivo modificado por un complemento preposicional de carácter verbal**



En ocasiones el complemento nominal tiene un núcleo verbal en lugar de nominal. Este patrón recoge dicha situación. La dependencia extraída es la existente entre entre el sustantivo y el núcleo verbal del complemento. Nótese que el verbo puede ser una forma pasiva y/o compuesta —de ahí la existencia de auxiliares.

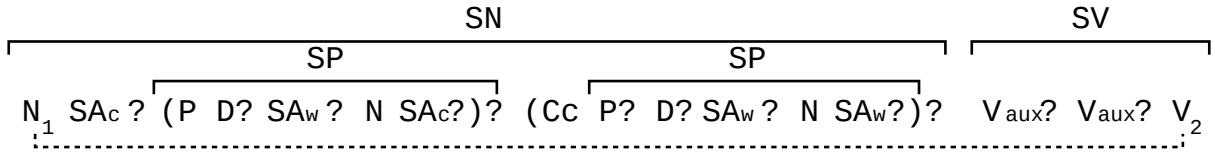
Ejemplo: *una* $\underset{N_1}{\underline{probabilidad}}$ *muy alta de haber* $\underset{V_1}{\underline{sucedido}}$

$$(N_1, V_1) \longrightarrow (probabilidad, sucedido)$$

E.2.3. Estructuras Sujeto–Verbo–Objeto

Una vez procesadas las dependencias internas a los sintagmas nominales, restan por analizar las dependencias correspondientes a relaciones verbales, cuyos patrones son recogidos en este tercer bloque.

PATRÓN SV01: **dependencias sujeto–verbo**

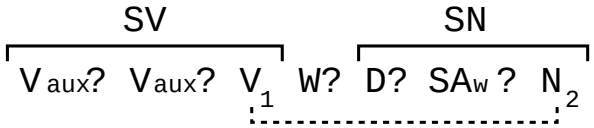


En este primer patrón correspondiente a dependencias verbales se captura la relación entre un verbo y su sujeto. Obsérvese que el sujeto es un sintagma nominal que puede estar siendo modificado por adjetivos o complementos, y que el verbo puede ser de nuevo una forma pasiva y/o compuesta —de ahí la existencia de auxiliares—. La dependencia se establece entre el núcleo del sujeto y el núcleo verbal.

Ejemplo: *varios* $\underset{N_1}{\underline{coches}}$ *de lujo han sido* $\underset{V_1}{\underline{robados}}$

$$(N_1, V_1) \longrightarrow (coches, robados)$$

PATRÓN SV02: **dependencias verbo–objeto**

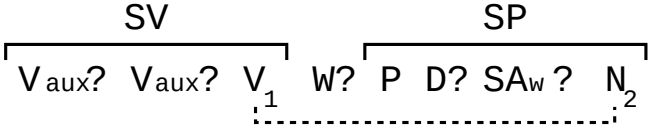


Este segundo patrón recoge el segundo tipo de dependencias verbales, aquéllas entre un verbo y su objeto. La dependencia se establece entre el núcleo verbal y el núcleo de dicho objeto.

Ejemplo: $\underset{V_1}{\underline{compré}}$ $\underset{N_1}{\underline{pollo}}$

$$(V_1, N_1) \longrightarrow (compré, pollo)$$

PATRÓN SV03: **dependencias verbo–complemento verbal**



Este último patrón cubre el caso de la existencia de complementos verbales de tipo preposicional. La dependencia se establece entre el núcleo verbal y el núcleo del complemento.

Ejemplo: $\underset{V_1}{\underline{cenamos}}$ *en un* $\underset{N_1}{\underline{restaurante}}$

$$(V_1\,,\ N_1) \longrightarrow (cenamos\,,\ restaurante)$$

Apéndice F

Puesta a Punto de los Términos Multipalabra

Tabla F.1: Puesta a punto del factor de ponderación ω. Resultados para el corpus CLEF 2001-02.A con el analizador PATTERNS: consultas cortas, $\omega \in \{1,2,3,4,5,8\}$

	lem	PATTERNS											
		$\omega = 1$	%Δ	$\omega = 2$	%Δ	$\omega = 3$	%Δ	$\omega = 4$	%Δ	$\omega = 5$	%Δ	$\omega = 8$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2700	2601	-3.67	2694	-0.22	**2717**	**0.63**	**2727**	**1.00**	**2730**	**1.11**	**2738**	**1.41**
Pr. no int.	.4829	.4325	-10.44	.4713	-2.40	**.4837**	**0.17**	**.4908**	**1.64**	**.4933**	**2.15**	**.4949**	**2.48**
Pr. doc.	.5327	.4431	-16.82	.4965	-6.80	.5183	-2.70	.5299	-0.53	**.5364**	**0.69**	**.5439**	**2.10**
R-pr.	.4848	.4279	-11.74	.4621	-4.68	.4756	-1.90	.4828	-0.41	.4811	-0.76	**.4851**	**0.06**
Pr. a 0 %	.8293	.7913	-4.58	**.8296**	**0.04**	**.8357**	**0.77**	**.8341**	**0.58**	**.8453**	**1.93**	**.8406**	**1.36**
Pr. a 10 %	.7463	.7237	-3.03	.7394	-0.92	.7457	-0.08	**.7509**	**0.62**	**.7552**	**1.19**	**.7612**	**2.00**
Pr. a 20 %	.6771	.6222	-8.11	.6718	-0.78	**.6863**	**1.36**	**.6915**	**2.13**	**.6911**	**2.07**	**.6885**	**1.68**
Pr. a 30 %	.6138	.5532	-9.87	.6040	-1.60	.6087	-0.83	**.6243**	**1.71**	**.6283**	**2.36**	**.6386**	**4.04**
Pr. a 40 %	.5502	.4653	-15.43	.5302	-3.64	**.5555**	**0.96**	**.5663**	**2.93**	**.5614**	**2.04**	**.5590**	**1.60**
Pr. a 50 %	.4931	.4286	-13.08	.4821	-2.23	.4931	0.00	**.5000**	**1.40**	**.5040**	**2.21**	**.5040**	**2.21**
Pr. a 60 %	.4496	.3913	-12.97	.4293	-4.52	.4439	-1.27	**.4583**	**1.94**	**.4596**	**2.22**	**.4638**	**3.16**
Pr. a 70 %	.3853	.3261	-15.36	.3646	-5.37	.3764	-2.31	.3799	-1.40	.3832	-0.55	**.3856**	**0.08**
Pr. a 80 %	.3277	.2781	-15.14	.3136	-4.30	.3246	-0.95	**.3283**	**0.18**	**.3323**	**1.40**	**.3333**	**1.71**
Pr. a 90 %	.2356	.2125	-9.80	**.2375**	**0.81**	**.2448**	**3.90**	**.2480**	**5.26**	**.2503**	**6.24**	**.2484**	**5.43**
Pr. a 100 %	.1197	.1087	-9.19	**.1199**	**0.17**	**.1225**	**2.34**	**.1229**	**2.67**	**.1229**	**2.67**	**.1213**	**1.34**
Pr. a 5 docs.	.6609	.6391	-3.30	.6435	-2.63	.6609	0.00	**.6696**	**1.32**	.6609	0.00	**.6783**	**2.63**
Pr. a 10 docs.	.6283	.5587	-11.08	.6130	-2.44	**.6304**	**0.33**	.6261	-0.35	**.6304**	**0.33**	**.6391**	**1.72**
Pr. a 15 docs.	.5928	.5348	-9.78	.5681	-4.17	.5739	-3.19	.5812	-1.96	.5797	-2.21	.5913	-0.25
Pr. a 20 docs.	.5446	.5022	-7.79	.5370	-1.40	.5446	0.00	**.5489**	**0.79**	**.5500**	**0.99**	**.5587**	**2.59**
Pr. a 30 docs.	.4928	.4449	-9.72	.4797	-2.66	.4920	-0.16	**.4957**	**0.59**	.4920	-0.16	**.5051**	**2.50**
Pr. a 100 docs.	.3300	.2859	-13.36	.3120	-5.45	.3217	-2.52	.3237	-1.91	.3274	-0.79	**.3317**	**0.52**
Pr. a 200 docs.	.2234	.1977	-11.50	.2154	-3.58	.2210	-1.07	**.2251**	**0.76**	**.2257**	**1.03**	**.2260**	**1.16**
Pr. a 500 docs.	.1090	.1040	-4.59	.1074	-1.47	**.1093**	**0.28**	**.1100**	**0.92**	**.1102**	**1.10**	**.1106**	**1.47**
Pr. a 1000 docs.	.0587	.0565	-3.75	.0586	-0.17	**.0591**	**0.68**	**.0593**	**1.02**	**.0593**	**1.02**	**.0595**	**1.36**

Tabla F.2: Puesta a punto del factor de ponderación ω. Resultados para el corpus CLEF 2001-02·A con el analizador PATTERNS: consultas cortas, $\omega \in \{10, 12, 14, 16, 18, 20\}$

	lem	PATTERNS											
		$\omega = 10$	%Δ	$\omega = 12$	%Δ	$\omega = 14$	%Δ	$\omega = 16$	%Δ	$\omega = 18$	%Δ	$\omega = 20$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2700	**2734**	**1.26**	**2730**	**1.11**	**2729**	**1.07**	**2729**	**1.07**	**2729**	**1.07**	**2730**	**1.11**
Pr. no int.	.4829	**.4945**	**2.40**	**.4940**	**2.30**	**.4927**	**2.03**	**.4915**	**1.78**	**.4915**	**1.78**	**.4906**	**1.59**
Pr. doc.	.5327	**.5448**	**2.27**	**.5450**	**2.31**	**.5447**	**2.25**	**.5442**	**2.16**	**.5439**	**2.10**	**.5433**	**1.99**
R-pr.	.4848	**.4851**	**0.06**	**.4872**	**0.50**	.4845	-0.06	**.4853**	**0.10**	**.4861**	**0.27**	**.4861**	**0.27**
Pr. a 0 %	.8293	**.8404**	**1.34**	**.8404**	**1.34**	**.8394**	**1.22**	**.8351**	**0.70**	**.8333**	**0.48**	**.8318**	**0.30**
Pr. a 10 %	.7463	**.7631**	**2.25**	**.7614**	**2.02**	**.7656**	**2.59**	**.7598**	**1.81**	**.7653**	**2.55**	**.7591**	**1.72**
Pr. a 20 %	.6771	**.6892**	**1.79**	**.6835**	**0.95**	**.6775**	**0.06**	.6759	-0.18	**.6781**	**0.15**	**.6787**	**0.24**
Pr. a 30 %	.6138	**.6387**	**4.06**	**.6419**	**4.58**	**.6340**	**3.29**	**.6303**	**2.69**	**.6286**	**2.41**	**.6280**	**2.31**
Pr. a 40 %	.5502	**.5581**	**1.44**	**.5554**	**0.95**	**.5586**	**1.53**	**.5549**	**0.85**	**.5554**	**0.95**	**.5543**	**0.75**
Pr. a 50 %	.4931	**.5009**	**1.58**	**.5024**	**1.89**	**.5032**	**2.05**	**.5033**	**2.07**	**.5040**	**2.21**	**.5032**	**2.05**
Pr. a 60 %	.4496	**.4601**	**2.34**	**.4611**	**2.56**	**.4594**	**2.18**	**.4581**	**1.89**	**.4576**	**1.78**	**.4572**	**1.69**
Pr. a 70 %	.3853	**.3858**	**0.13**	**.3854**	**0.03**	.3846	-0.18	**.3861**	**0.21**	**.3867**	**0.36**	**.3863**	**0.26**
Pr. a 80 %	.3277	**.3344**	**2.04**	**.3363**	**2.62**	**.3361**	**2.56**	**.3356**	**2.41**	**.3354**	**2.35**	**.3353**	**2.32**
Pr. a 90 %	.2356	**.2463**	**4.54**	**.2446**	**3.82**	**.2451**	**4.03**	**.2437**	**3.44**	**.2433**	**3.27**	**.2415**	**2.50**
Pr. a 100 %	.1197	**.1209**	**1.00**	.1188	-0.75	.1189	-0.67	.1190	-0.58	.1191	-0.50	.1192	-0.42
Pr. a 5 docs.	.6609	**.6739**	**1.97**	**.6696**	**1.32**	.6609	0.00	**.6652**	**0.65**	**.6652**	**0.65**	**.6652**	**0.65**
Pr. a 10 docs.	.6283	**.6348**	**1.03**	**.6370**	**1.38**	**.6348**	**1.03**	**.6326**	**0.68**	**.6304**	**0.33**	**.6304**	**0.33**
Pr. a 15 docs.	.5928	**.5942**	**0.24**	.5899	-0.49	.5899	-0.49	.5913	-0.25	.5928	0.00	**.5942**	**0.24**
Pr. a 20 docs.	.5446	**.5543**	**1.78**	**.5587**	**2.59**	**.5565**	**2.19**	**.5543**	**1.78**	**.5533**	**1.60**	**.5543**	**1.78**
Pr. a 30 docs.	.4928	**.5051**	**2.50**	**.5043**	**2.33**	**.5022**	**1.91**	**.5029**	**2.05**	**.5029**	**2.05**	**.5000**	**1.46**
Pr. a 100 docs.	.3300	**.3337**	**1.12**	**.3339**	**1.18**	**.3346**	**1.39**	**.3350**	**1.52**	**.3346**	**1.39**	**.3352**	**1.58**
Pr. a 200 docs.	.2234	**.2261**	**1.21**	**.2261**	**1.21**	**.2263**	**1.30**	**.2259**	**1.12**	**.2258**	**1.07**	**.2255**	**0.94**
Pr. a 500 docs.	.1090	**.1104**	**1.28**	**.1102**	**1.10**	**.1101**	**1.01**	**.1100**	**0.92**	**.1098**	**0.73**	**.1098**	**0.73**
Pr. a 1000 docs.	.0587	**.0594**	**1.19**	**.0593**	**1.02**	**.0593**	**1.02**	**.0593**	**1.02**	**.0593**	**1.02**	**.0593**	**1.02**

Tabla F.3: Puesta a punto del factor de ponderación ω. Resultados para el corpus CLEF 2001-02.A con el analizador PATTERNS: consultas largas, $\omega \in \{1,2,3,4,5,8\}$

	lem	PATTERNS											
		$\omega = 1$	%Δ	$\omega = 2$	%Δ	$\omega = 3$	%Δ	$\omega = 4$	%Δ	$\omega = 5$	%Δ	$\omega = 8$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2762	2681	-2.93	2738	-0.87	2760	-0.07	**2767**	**0.18**	**2772**	**0.36**	**2784**	**0.80**
Pr. no int.	.5239	.4839	-7.64	.5142	-1.85	**.5252**	**0.25**	**.5298**	**1.13**	**.5327**	**1.68**	**.5356**	**2.23**
Pr. doc.	.5690	.4951	-12.99	.5405	-5.01	.5595	-1.67	.5689	-0.02	**.5745**	**0.97**	**.5807**	**2.06**
R-pr.	.5075	.4720	-7.00	.4962	-2.23	.4979	-1.89	**.5097**	**0.43**	**.5100**	**0.49**	**.5118**	**0.85**
Pr. a 0 %	.8845	.8728	-1.32	.8696	-1.68	.8724	-1.37	.8666	-2.02	.8696	-1.68	**.8870**	**0.28**
Pr. a 10 %	.7914	.7813	-1.28	.7820	-1.19	.7887	-0.34	**.7988**	**0.94**	**.8063**	**1.88**	**.8102**	**2.38**
Pr. a 20 %	.7136	.6779	-5.00	**.7166**	**0.42**	**.7314**	**2.49**	**.7348**	**2.97**	**.7418**	**3.95**	**.7496**	**5.04**
Pr. a 30 %	.6591	.6203	-5.89	**.6636**	**0.68**	**.6676**	**1.29**	**.6676**	**1.29**	**.6684**	**1.41**	**.6683**	**1.40**
Pr. a 40 %	.6085	.5323	-12.52	.5772	-5.14	.5922	-2.68	.5993	-1.51	.6035	-0.82	**.6122**	**0.61**
Pr. a 50 %	.5557	.4829	-13.10	.5270	-5.16	.5441	-2.09	.5487	-1.26	.5517	-0.72	**.5605**	**0.86**
Pr. a 60 %	.5006	.4525	-9.61	.4916	-1.80	**.5051**	**0.90**	**.5102**	**1.92**	**.5123**	**2.34**	**.5116**	**2.20**
Pr. a 70 %	.4161	.3852	-7.43	.4107	-1.30	**.4174**	**0.31**	**.4181**	**0.48**	**.4228**	**1.61**	**.4300**	**3.34**
Pr. a 80 %	.3509	.3205	-8.66	**.3535**	**0.74**	**.3624**	**3.28**	**.3642**	**3.79**	**.3666**	**4.47**	**.3670**	**4.59**
Pr. a 90 %	.2492	.2305	-7.50	.2486	-0.24	**.2594**	**4.09**	**.2598**	**4.25**	**.2606**	**4.57**	**.2606**	**4.57**
Pr. a 100 %	.1289	.1186	-7.99	**.1292**	**0.23**	**.1317**	**2.17**	**.1330**	**3.18**	**.1339**	**3.88**	**.1349**	**4.65**
Pr. a 5 docs.	.6957	.6826	-1.88	**.7217**	**3.74**	**.7174**	**3.12**	**.7087**	**1.87**	**.7261**	**4.37**	**.7217**	**3.74**
Pr. a 10 docs.	.6543	.6239	-4.65	.6543	0.00	**.6652**	**1.67**	**.6761**	**3.33**	**.6804**	**3.99**	**.6674**	**2.00**
Pr. a 15 docs.	.6188	.5855	-5.38	.6087	-1.63	.6116	-1.16	.6145	-0.69	**.6203**	**0.24**	**.6246**	**0.94**
Pr. a 20 docs.	.5880	.5641	-4.06	.5761	-2.02	.5815	-1.11	.5880	0.00	.5880	0.00	**.5967**	**1.48**
Pr. a 30 docs.	.5304	.4957	-6.54	.5188	-2.19	.5225	-1.49	.5246	-1.09	**.5319**	**0.28**	**.5391**	**1.64**
Pr. a 100 docs.	.3509	.3070	-12.51	.3302	-5.90	.3374	-3.85	.3426	-2.37	.3441	-1.94	.3489	-0.57
Pr. a 200 docs.	.2315	.2117	-8.55	.2272	-1.86	.2309	-0.26	**.2335**	**0.86**	**.2342**	**1.17**	**.2343**	**1.21**
Pr. a 500 docs.	.1115	.1093	-1.97	**.1120**	**0.45**	**.1130**	**1.35**	**.1136**	**1.88**	**.1137**	**1.97**	**.1132**	**1.52**
Pr. a 1000 docs.	.0600	.0583	-2.83	.0595	-0.83	.0600	0.00	**.0602**	**0.33**	**.0603**	**0.50**	**.0605**	**0.83**

Tabla F.4: Puesta a punto del factor de ponderación ω. Resultados para el corpus CLEF 2001-02·A con el analizador PATTERNS: consultas largas, $\omega \in \{10, 12, 14, 16, 18, 20\}$

	lem	PATTERNS											
		$\omega = 10$	%Δ	$\omega = 12$	%Δ	$\omega = 14$	%Δ	$\omega = 16$	%Δ	$\omega = 18$	%Δ	$\omega = 20$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2762	**2788**	**0.94**	**2787**	**0.91**	**2790**	**1.01**	**2785**	**0.83**	**2786**	**0.87**	**2783**	**0.76**
Pr. no int.	.5239	**.5355**	**2.21**	**.5348**	**2.08**	**.5342**	**1.97**	**.5326**	**1.66**	**.5319**	**1.53**	**.5315**	**1.45**
Pr. doc.	.5690	**.5818**	**2.25**	**.5818**	**2.25**	**.5814**	**2.18**	**.5805**	**2.02**	**.5802**	**1.97**	**.5796**	**1.86**
R-pr.	.5075	**.5145**	**1.38**	**.5159**	**1.66**	**.5160**	**1.67**	**.5162**	**1.71**	**.5164**	**1.75**	**.5164**	**1.75**
Pr. a 0 %	.8845	**.8873**	**0.32**	.8823	-0.25	.8804	-0.46	.8807	-0.43	.8804	-0.46	.8804	-0.46
Pr. a 10 %	.7914	**.8117**	**2.57**	**.8115**	**2.54**	**.8053**	**1.76**	**.8039**	**1.58**	**.8046**	**1.67**	**.8042**	**1.62**
Pr. a 20 %	.7136	**.7476**	**4.76**	**.7437**	**4.22**	**.7390**	**3.56**	**.7320**	**2.58**	**.7313**	**2.48**	**.7311**	**2.45**
Pr. a 30 %	.6591	**.6689**	**1.49**	**.6681**	**1.37**	**.6676**	**1.29**	**.6675**	**1.27**	**.6674**	**1.26**	**.6659**	**1.03**
Pr. a 40 %	.6085	**.6107**	**0.36**	**.6101**	**0.26**	**.6130**	**0.74**	**.6138**	**0.87**	**.6132**	**0.77**	**.6134**	**0.81**
Pr. a 50 %	.5557	**.5598**	**0.74**	**.5601**	**0.79**	**.5617**	**1.08**	**.5630**	**1.31**	**.5623**	**1.19**	**.5607**	**0.90**
Pr. a 60 %	.5006	**.5146**	**2.80**	**.5113**	**2.14**	**.5103**	**1.94**	**.5080**	**1.48**	**.5065**	**1.18**	**.5055**	**0.98**
Pr. a 70 %	.4161	**.4289**	**3.08**	**.4286**	**3.00**	**.4282**	**2.91**	**.4259**	**2.36**	**.4251**	**2.16**	**.4240**	**1.90**
Pr. a 80 %	.3509	**.3666**	**4.47**	**.3670**	**4.59**	**.3660**	**4.30**	**.3647**	**3.93**	**.3631**	**3.48**	**.3621**	**3.19**
Pr. a 90 %	.2492	**.2597**	**4.21**	**.2580**	**3.53**	**.2576**	**3.37**	**.2568**	**3.05**	**.2568**	**3.05**	**.2567**	**3.01**
Pr. a 100 %	.1289	**.1345**	**4.34**	**.1341**	**4.03**	**.1342**	**4.11**	**.1341**	**4.03**	**.1339**	**3.88**	**.1317**	**2.17**
Pr. a 5 docs.	.6957	**.7130**	**2.49**	**.7174**	**3.12**	**.7043**	**1.24**	**.7000**	**0.62**	.6957	0.00	**.7000**	**0.62**
Pr. a 10 docs.	.6543	**.6674**	**2.00**	**.6652**	**1.67**	**.6674**	**2.00**	**.6674**	**2.00**	**.6652**	**1.67**	**.6674**	**2.00**
Pr. a 15 docs.	.6188	**.6319**	**2.12**	**.6319**	**2.12**	**.6275**	**1.41**	**.6290**	**1.65**	**.6261**	**1.18**	**.6246**	**0.94**
Pr. a 20 docs.	.5880	**.5957**	**1.31**	**.5935**	**0.94**	**.5946**	**1.12**	**.5946**	**1.12**	**.5957**	**1.31**	**.5967**	**1.48**
Pr. a 30 docs.	.5304	**.5420**	**2.19**	**.5435**	**2.47**	**.5413**	**2.06**	**.5413**	**2.06**	**.5413**	**2.06**	**.5406**	**1.92**
Pr. a 100 docs.	.3509	.3493	-0.46	.3502	-0.20	.3502	-0.20	.3498	-0.31	.3500	-0.26	.3500	-0.26
Pr. a 200 docs.	.2315	**.2350**	**1.51**	**.2346**	**1.34**	**.2347**	**1.38**	**.2342**	**1.17**	**.2345**	**1.30**	**.2339**	**1.04**
Pr. a 500 docs.	.1115	**.1128**	**1.17**	**.1128**	**1.17**	**.1128**	**1.17**	**.1127**	**1.08**	**.1125**	**0.90**	**.1125**	**0.90**
Pr. a 1000 docs.	.0600	**.0606**	**1.00**	**.0606**	**1.00**	**.0607**	**1.17**	**.0605**	**0.83**	**.0606**	**1.00**	**.0605**	**0.83**

	lem	CASCADE											
		$\omega = 1$	%Δ	$\omega = 2$	%Δ	$\omega = 3$	%Δ	$\omega = 4$	%Δ	$\omega = 5$	%Δ	$\omega = 8$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2700	2625	-2.78	**2701**	**0.04**	**2721**	**0.78**	**2724**	**0.89**	**2728**	**1.04**	**2728**	**1.04**
Pr. no int.	.4829	.4547	-5.84	**.4857**	**0.58**	**.4940**	**2.30**	**.4989**	**3.31**	**.4989**	**3.31**	**.4965**	**2.82**
Pr. doc.	.5327	.4806	-9.78	.5256	-1.33	**.5410**	**1.56**	**.5476**	**2.80**	**.5496**	**3.17**	**.5491**	**3.08**
R-pr.	.4848	.4513	-6.91	.4775	-1.51	**.4869**	**0.43**	**.4883**	**0.72**	**.4894**	**0.95**	**.4895**	**0.97**
Pr. a 0 %	.8293	**.8331**	**0.46**	**.8431**	**1.66**	**.8602**	**3.73**	**.8616**	**3.89**	**.8656**	**4.38**	**.8406**	**1.36**
Pr. a 10 %	.7463	.7408	-0.74	**.7569**	**1.42**	**.7665**	**2.71**	**.7710**	**3.31**	**.7664**	**2.69**	**.7640**	**2.37**
Pr. a 20 %	.6771	.6345	-6.29	.6724	-0.69	**.6791**	**0.30**	**.6879**	**1.60**	**.6879**	**1.60**	**.6925**	**2.27**
Pr. a 30 %	.6138	.5679	-7.48	**.6169**	**0.51**	**.6278**	**2.28**	**.6373**	**3.83**	**.6325**	**3.05**	**.6338**	**3.26**
Pr. a 40 %	.5502	.5068	-7.89	**.5606**	**1.89**	**.5671**	**3.07**	**.5685**	**3.33**	**.5692**	**3.45**	**.5622**	**2.18**
Pr. a 50 %	.4931	.4593	-6.85	**.4994**	**1.28**	**.5093**	**3.29**	**.5111**	**3.65**	**.5094**	**3.31**	**.5076**	**2.94**
Pr. a 60 %	.4496	.4174	-7.16	.4434	-1.38	**.4581**	**1.89**	**.4662**	**3.69**	**.4662**	**3.69**	**.4648**	**3.38**
Pr. a 70 %	.3853	.3548	-7.92	.3765	-2.28	.3844	-0.23	**.3880**	**0.70**	**.3900**	**1.22**	**.3909**	**1.45**
Pr. a 80 %	.3277	.2999	-8.48	.3239	-1.16	**.3329**	**1.59**	**.3365**	**2.69**	**.3368**	**2.78**	**.3364**	**2.65**
Pr. a 90 %	.2356	.2231	-5.31	**.2411**	**2.33**	**.2454**	**4.16**	**.2475**	**5.05**	**.2477**	**5.14**	**.2483**	**5.39**
Pr. a 100 %	.1197	.1146	-4.26	.1171	-2.17	**.1198**	**0.08**	.1197	0.00	.1195	-0.17	.1197	0.00
Pr. a 5 docs.	.6609	.6391	-3.30	**.6696**	**1.32**	**.6739**	**1.97**	**.6957**	**5.27**	**.7000**	**5.92**	**.6913**	**4.60**
Pr. a 10 docs.	.6283	.5804	-7.62	**.6326**	**0.68**	**.6457**	**2.77**	**.6478**	**3.10**	**.6478**	**3.10**	**.6500**	**3.45**
Pr. a 15 docs.	.5928	.5391	-9.06	**.5942**	**0.24**	**.6043**	**1.94**	**.6029**	**1.70**	**.6029**	**1.70**	**.6029**	**1.70**
Pr. a 20 docs.	.5446	.5065	-7.00	**.5511**	**1.19**	**.5609**	**2.99**	**.5630**	**3.38**	**.5641**	**3.58**	**.5620**	**3.20**
Pr. a 30 docs.	.4928	.4638	-5.88	**.4949**	**0.43**	**.5058**	**2.64**	**.5087**	**3.23**	**.5065**	**2.78**	**.5036**	**2.19**
Pr. a 100 docs.	.3300	.3030	-8.18	.3239	-1.85	**.3304**	**0.12**	**.3333**	**1.00**	**.3365**	**1.97**	**.3348**	**1.45**
Pr. a 200 docs.	.2234	.2109	-5.60	.2223	-0.49	**.2254**	**0.90**	**.2265**	**1.39**	**.2264**	**1.34**	**.2263**	**1.30**
Pr. a 500 docs.	.1090	.1044	-4.22	.1086	-0.37	**.1099**	**0.83**	**.1103**	**1.19**	**.1104**	**1.28**	**.1103**	**1.19**
Pr. a 1000 docs.	.0587	.0571	-2.73	.0587	0.00	**.0592**	**0.85**	**.0592**	**0.85**	**.0593**	**1.02**	**.0593**	**1.02**

Tabla F.5: Puesta a punto del factor de ponderación ω. Resultados para el corpus CLEF 2001-02.A con el analizador CASCADE: consultas cortas, $\omega \in \{1, 2, 3, 4, 5, 8\}$

Tabla F.6: Puesta a punto del factor de ponderación ω. Resultados para el corpus CLEF 2001-02·A con el analizador CASCADE: consultas cortas, $\omega \in \{10, 12, 14, 16, 18, 20\}$

	lem	CASCADE											
		$\omega = 10$	%Δ	$\omega = 12$	%Δ	$\omega = 14$	%Δ	$\omega = 16$	%Δ	$\omega = 18$	%Δ	$\omega = 20$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2700	**2726**	**0.96**	**2724**	**0.89**	**2722**	**0.81**	**2722**	**0.81**	**2723**	**0.85**	**2723**	**0.85**
Pr. no int.	.4829	**.4944**	**2.38**	**.4938**	**2.26**	**.4918**	**1.84**	**.4906**	**1.59**	**.4900**	**1.47**	**.4894**	**1.35**
Pr. doc.	.5327	**.5475**	**2.78**	**.5464**	**2.57**	**.5449**	**2.29**	**.5437**	**2.06**	**.5431**	**1.95**	**.5423**	**1.80**
R-pr.	.4848	**.4904**	**1.16**	**.4902**	**1.11**	**.4881**	**0.68**	**.4895**	**0.97**	**.4885**	**0.76**	**.4873**	**0.52**
Pr. a 0 %	.8293	.8287	-0.07	.8283	-0.12	.8268	-0.30	.8272	-0.25	.8274	-0.23	.8275	-0.22
Pr. a 10 %	.7463	**.7608**	**1.94**	**.7644**	**2.43**	**.7579**	**1.55**	**.7563**	**1.34**	**.7550**	**1.17**	**.7548**	**1.14**
Pr. a 20 %	.6771	**.6875**	**1.54**	**.6888**	**1.73**	**.6837**	**0.97**	**.6829**	**0.86**	**.6823**	**0.77**	**.6819**	**0.71**
Pr. a 30 %	.6138	**.6296**	**2.57**	**.6299**	**2.62**	**.6288**	**2.44**	**.6277**	**2.26**	**.6260**	**1.99**	**.6272**	**2.18**
Pr. a 40 %	.5502	**.5577**	**1.36**	**.5581**	**1.44**	**.5547**	**0.82**	**.5542**	**0.73**	**.5531**	**0.53**	**.5522**	**0.36**
Pr. a 50 %	.4931	**.5016**	**1.72**	**.5016**	**1.72**	**.4990**	**1.20**	**.4989**	**1.18**	**.4989**	**1.18**	**.4993**	**1.26**
Pr. a 60 %	.4496	**.4651**	**3.45**	**.4639**	**3.18**	**.4603**	**2.38**	**.4595**	**2.20**	**.4583**	**1.94**	**.4584**	**1.96**
Pr. a 70 %	.3853	**.3909**	**1.45**	**.3903**	**1.30**	**.3888**	**0.91**	**.3890**	**0.96**	**.3890**	**0.96**	**.3893**	**1.04**
Pr. a 80 %	.3277	**.3349**	**2.20**	**.3355**	**2.38**	**.3346**	**2.11**	**.3342**	**1.98**	**.3335**	**1.77**	**.3332**	**1.68**
Pr. a 90 %	.2356	**.2468**	**4.75**	**.2460**	**4.41**	**.2443**	**3.69**	**.2432**	**3.23**	**.2432**	**3.23**	**.2416**	**2.55**
Pr. a 100 %	.1197	.1196	-0.08	.1196	-0.08	.1196	-0.08	.1195	-0.17	.1196	-0.08	.1196	-0.08
Pr. a 5 docs.	.6609	**.6870**	**3.95**	**.6696**	**1.32**	**.6696**	**1.32**	.6609	0.00	**.6652**	**0.65**	.6609	0.00
Pr. a 10 docs.	.6283	**.6500**	**3.45**	**.6478**	**3.10**	**.6435**	**2.42**	**.6413**	**2.07**	**.6391**	**1.72**	**.6370**	**1.38**
Pr. a 15 docs.	.5928	**.5957**	**0.49**	.5913	-0.25	.5928	0.00	.5913	-0.25	.5899	-0.49	.5913	-0.25
Pr. a 20 docs.	.5446	**.5587**	**2.59**	**.5576**	**2.39**	**.5554**	**1.98**	**.5543**	**1.78**	**.5543**	**1.78**	**.5543**	**1.78**
Pr. a 30 docs.	.4928	**.5036**	**2.19**	**.5065**	**2.78**	**.5022**	**1.91**	**.5014**	**1.75**	**.5014**	**1.75**	**.4993**	**1.32**
Pr. a 100 docs.	.3300	**.3348**	**1.45**	**.3352**	**1.58**	**.3339**	**1.18**	**.3337**	**1.12**	**.3335**	**1.06**	**.3330**	**0.91**
Pr. a 200 docs.	.2234	**.2260**	**1.16**	**.2260**	**1.16**	**.2251**	**0.76**	**.2252**	**0.81**	**.2251**	**0.76**	**.2249**	**0.67**
Pr. a 500 docs.	.1090	**.1103**	**1.19**	**.1101**	**1.01**	**.1100**	**0.92**	**.1098**	**0.73**	**.1096**	**0.55**	**.1096**	**0.55**
Pr. a 1000 docs.	.0587	**.0593**	**1.02**	**.0592**	**0.85**	**.0592**	**0.85**	**.0592**	**0.85**	**.0592**	**0.85**	**.0592**	**0.85**

Tabla F.7: Puesta a punto del factor de ponderación ω. Resultados para el corpus CLEF 2001-02-A con el analizador CASCADE: consultas largas, $\omega \in \{1, 2, 3, 4, 5, 8\}$

	lem	CASCADE											
		$\omega = 1$	%Δ	$\omega = 2$	%Δ	$\omega = 3$	%Δ	$\omega = 4$	%Δ	$\omega = 5$	%Δ	$\omega = 8$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2762	2697	-2.35	2747	-0.54	**2765**	**0.11**	**2776**	**0.51**	**2778**	**0.58**	**2786**	**0.87**
Pr. no int.	.5239	.5050	-3.61	**.5278**	**0.74**	**.5330**	**1.74**	**.5342**	**1.97**	**.5343**	**1.99**	**.5339**	**1.91**
Pr. doc.	.5690	.5300	-6.85	.5656	-0.60	**.5767**	**1.35**	**.5812**	**2.14**	**.5825**	**2.37**	**.5826**	**2.39**
R-pr.	.5075	.4959	-2.29	**.5109**	**0.67**	**.5151**	**1.50**	**.5166**	**1.79**	**.5167**	**1.81**	**.5185**	**2.17**
Pr. a 0 %	.8845	.8280	-6.39	.8494	-3.97	.8469	-4.25	.8507	-3.82	.8506	-3.83	.8685	-1.81
Pr. a 10 %	.7914	.7896	-0.23	**.8054**	**1.77**	**.8106**	**2.43**	**.8136**	**2.81**	**.8133**	**2.77**	**.8067**	**1.93**
Pr. a 20 %	.7136	.6989	-2.06	**.7264**	**1.79**	**.7377**	**3.38**	**.7380**	**3.42**	**.7358**	**3.11**	**.7331**	**2.73**
Pr. a 30 %	.6591	.6288	-4.60	.6573	-0.27	**.6688**	**1.47**	**.6772**	**2.75**	**.6783**	**2.91**	**.6728**	**2.08**
Pr. a 40 %	.6085	.5745	-5.59	.5996	-1.46	**.6091**	**0.10**	**.6101**	**0.26**	**.6115**	**0.49**	**.6138**	**0.87**
Pr. a 50 %	.5557	.5314	-4.37	**.5560**	**0.05**	**.5671**	**2.05**	**.5650**	**1.67**	**.5621**	**1.15**	**.5655**	**1.76**
Pr. a 60 %	.5006	.4749	-5.13	**.5063**	**1.14**	**.5121**	**2.30**	**.5154**	**2.96**	**.5179**	**3.46**	**.5162**	**3.12**
Pr. a 70 %	.4161	.4074	-2.09	**.4214**	**1.27**	**.4249**	**2.11**	**.4218**	**1.37**	**.4217**	**1.35**	**.4292**	**3.15**
Pr. a 80 %	.3509	.3501	-0.23	**.3641**	**3.76**	**.3666**	**4.47**	**.3648**	**3.96**	**.3641**	**3.76**	**.3632**	**3.51**
Pr. a 90 %	.2492	.2428	-2.57	**.2586**	**3.77**	**.2632**	**5.62**	**.2621**	**5.18**	**.2603**	**4.45**	**.2581**	**3.57**
Pr. a 100 %	.1289	.1248	-3.18	**.1314**	**1.94**	**.1303**	**1.09**	**.1308**	**1.47**	**.1311**	**1.71**	**.1318**	**2.25**
Pr. a 5 docs.	.6957	.6652	-4.38	**.7217**	**3.74**	**.7261**	**4.37**	**.7087**	**1.87**	**.7087**	**1.87**	**.7043**	**1.24**
Pr. a 10 docs.	.6543	.6391	-2.32	**.6739**	**3.00**	**.6717**	**2.66**	**.6783**	**3.67**	**.6783**	**3.67**	**.6783**	**3.67**
Pr. a 15 docs.	.6188	.6014	-2.81	**.6275**	**1.41**	**.6246**	**0.94**	**.6319**	**2.12**	**.6362**	**2.81**	**.6319**	**2.12**
Pr. a 20 docs.	.5880	.5717	-2.77	**.5913**	**0.56**	**.5967**	**1.48**	**.5967**	**1.48**	**.5978**	**1.67**	**.5967**	**1.48**
Pr. a 30 docs.	.5304	.5087	-4.09	**.5370**	**1.24**	**.5420**	**2.19**	**.5478**	**3.28**	**.5464**	**3.02**	**.5442**	**2.60**
Pr. a 100 docs.	.3509	.3320	-5.39	.3463	-1.31	.3493	-0.46	.3507	-0.06	**.3517**	**0.23**	.3509	0.00
Pr. a 200 docs.	.2315	.2257	-2.51	**.2347**	**1.38**	**.2348**	**1.43**	**.2349**	**1.47**	**.2348**	**1.43**	**.2346**	**1.34**
Pr. a 500 docs.	.1115	.1098	-1.52	**.1121**	**0.54**	**.1129**	**1.26**	**.1129**	**1.26**	**.1130**	**1.35**	**.1127**	**1.08**
Pr. a 1000 docs.	.0600	.0586	-2.33	.0597	-0.50	**.0601**	**0.17**	**.0603**	**0.50**	**.0604**	**0.67**	**.0606**	**1.00**

Tabla F.8: Puesta a punto del factor de ponderación ω. Resultados para el corpus CLEF 2001-02·A con el analizador CASCADE: consultas largas, $\omega \in \{10, 12, 14, 16, 18, 20\}$

	lem	CASCADE											
		$\omega = 10$	%Δ	$\omega = 12$	%Δ	$\omega = 14$	%Δ	$\omega = 16$	%Δ	$\omega = 18$	%Δ	$\omega = 20$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2762	**2784**	**0.80**	**2785**	**0.83**	**2783**	**0.76**	**2783**	**0.76**	**2783**	**0.76**	**2781**	**0.69**
Pr. no int.	.5239	**.5333**	**1.79**	**.5320**	**1.55**	**.5315**	**1.45**	**.5309**	**1.34**	**.5303**	**1.22**	**.5295**	**1.07**
Pr. doc.	.5690	**.5815**	**2.20**	**.5806**	**2.04**	**.5796**	**1.86**	**.5788**	**1.72**	**.5780**	**1.58**	**.5773**	**1.46**
R-pr.	.5075	**.5185**	**2.17**	**.5152**	**1.52**	**.5166**	**1.79**	**.5160**	**1.67**	**.5144**	**1.36**	**.5129**	**1.06**
Pr. a 0 %	.8845	.8760	-0.96	.8752	-1.05	.8753	-1.04	.8824	-0.24	.8815	-0.34	.8814	-0.35
Pr. a 10 %	.7914	**.8100**	**2.35**	**.8062**	**1.87**	**.8014**	**1.26**	**.8007**	**1.18**	**.7997**	**1.05**	**.7998**	**1.06**
Pr. a 20 %	.7136	**.7304**	**2.35**	**.7286**	**2.10**	**.7253**	**1.64**	**.7277**	**1.98**	**.7259**	**1.72**	**.7259**	**1.72**
Pr. a 30 %	.6591	**.6708**	**1.78**	**.6675**	**1.27**	**.6663**	**1.09**	**.6637**	**0.70**	**.6634**	**0.65**	**.6628**	**0.56**
Pr. a 40 %	.6085	**.6130**	**0.74**	**.6102**	**0.28**	**.6141**	**0.92**	**.6128**	**0.71**	**.6113**	**0.46**	**.6113**	**0.46**
Pr. a 50 %	.5557	**.5647**	**1.62**	**.5650**	**1.67**	**.5641**	**1.51**	**.5636**	**1.42**	**.5629**	**1.30**	**.5622**	**1.17**
Pr. a 60 %	.5006	**.5140**	**2.68**	**.5130**	**2.48**	**.5131**	**2.50**	**.5111**	**2.10**	**.5094**	**1.76**	**.5084**	**1.56**
Pr. a 70 %	.4161	**.4277**	**2.79**	**.4251**	**2.16**	**.4258**	**2.33**	**.4245**	**2.02**	**.4228**	**1.61**	**.4221**	**1.44**
Pr. a 80 %	.3509	**.3617**	**3.08**	**.3612**	**2.94**	**.3594**	**2.42**	**.3586**	**2.19**	**.3582**	**2.08**	**.3577**	**1.94**
Pr. a 90 %	.2492	**.2578**	**3.45**	**.2566**	**2.97**	**.2561**	**2.77**	**.2556**	**2.57**	**.2553**	**2.45**	**.2544**	**2.09**
Pr. a 100 %	.1289	**.1318**	**2.25**	**.1321**	**2.48**	**.1323**	**2.64**	**.1323**	**2.64**	**.1320**	**2.40**	**.1319**	**2.33**
Pr. a 5 docs.	.6957	**.7000**	**0.62**	.6957	0.00	**.7000**	**0.62**	.6957	0.00	.6957	0.00	.6957	0.00
Pr. a 10 docs.	.6543	**.6739**	**3.00**	**.6674**	**2.00**	**.6652**	**1.67**	**.6630**	**1.33**	**.6587**	**0.67**	**.6565**	**0.34**
Pr. a 15 docs.	.6188	**.6319**	**2.12**	**.6333**	**2.34**	**.6333**	**2.34**	**.6333**	**2.34**	**.6319**	**2.12**	**.6275**	**1.41**
Pr. a 20 docs.	.5880	**.5924**	**0.75**	**.5902**	**0.37**	**.5891**	**0.19**	**.5913**	**0.56**	**.5924**	**0.75**	**.5924**	**0.75**
Pr. a 30 docs.	.5304	**.5413**	**2.06**	**.5435**	**2.47**	**.5413**	**2.06**	**.5413**	**2.06**	**.5391**	**1.64**	**.5399**	**1.79**
Pr. a 100 docs.	.3509	**.3524**	**0.43**	**.3526**	**0.48**	**.3535**	**0.74**	**.3535**	**0.74**	**.3524**	**0.43**	**.3522**	**0.37**
Pr. a 200 docs.	.2315	**.2347**	**1.38**	**.2347**	**1.38**	**.2346**	**1.34**	**.2340**	**1.08**	**.2337**	**0.95**	**.2336**	**0.91**
Pr. a 500 docs.	.1115	**.1125**	**0.90**	**.1124**	**0.81**	**.1124**	**0.81**	**.1123**	**0.72**	**.1122**	**0.63**	**.1122**	**0.63**
Pr. a 1000 docs.	.0600	**.0605**	**0.83**	**.0605**	**0.83**	**.0605**	**0.83**	**.0605**	**0.83**	**.0605**	**0.83**	**.0605**	**0.83**

Tabla F.9: Puesta a punto del factor de ponderación ω. Resultados para *sintagmas nominales* con el corpus CLEF 2001-02·A con el analizador CASCADE: consultas cortas, $\omega \in \{1, 2, 3, 4, 5, 8\}$

	lem	CASCADE											
		$\omega = 1$	%Δ	$\omega = 2$	%Δ	$\omega = 3$	%Δ	$\omega = 4$	%Δ	$\omega = 5$	%Δ	$\omega = 8$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46k	46k	–	46k	–	46k	–	46k	–	46k	–	46k	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2700	2626	-2.74	2700	0.00	**2719**	**0.70**	**2724**	**0.89**	**2726**	**0.96**	**2725**	**0.93**
Pr. no int.	.4829	.4538	-6.03	.4812	-0.35	**.4906**	**1.59**	**.4937**	**2.24**	**.4940**	**2.30**	**.4925**	**1.99**
Pr. doc.	.5327	.4763	-10.59	.5214	-2.12	**.5368**	**0.77**	**.5432**	**1.97**	**.5457**	**2.44**	**.5457**	**2.44**
R-pr.	.4848	.4490	-7.38	.4727	-2.50	.4837	-0.23	**.4878**	**0.62**	**.4858**	**0.21**	.4844	-0.08
Pr. a 0 %	.8293	.8128	-1.99	**.8321**	**0.34**	**.8498**	**2.47**	**.8500**	**2.50**	**.8498**	**2.47**	**.8401**	**1.30**
Pr. a 10 %	.7463	.7335	-1.72	**.7604**	**1.89**	**.7727**	**3.54**	**.7677**	**2.87**	**.7628**	**2.21**	**.7588**	**1.67**
Pr. a 20 %	.6771	.6230	-7.99	.6671	-1.48	.6759	-0.18	**.6839**	**1.00**	**.6878**	**1.58**	**.6861**	**1.33**
Pr. a 30 %	.6138	.5740	-6.48	.6074	-1.04	**.6268**	**2.12**	**.6307**	**2.75**	**.6273**	**2.20**	**.6270**	**2.15**
Pr. a 40 %	.5502	.5193	-5.62	.5498	-0.07	**.5574**	**1.31**	**.5603**	**1.84**	**.5614**	**2.04**	**.5564**	**1.13**
Pr. a 50 %	.4931	.4683	-5.03	**.4932**	**0.02**	**.4997**	**1.34**	**.5065**	**2.72**	**.5036**	**2.13**	**.4997**	**1.34**
Pr. a 60 %	.4496	.4237	-5.76	.4436	-1.33	**.4545**	**1.09**	**.4610**	**2.54**	**.4618**	**2.71**	**.4617**	**2.69**
Pr. a 70 %	.3853	.3563	-7.53	.3751	-2.65	.3814	-1.01	.3851	-0.05	**.3881**	**0.73**	**.3901**	**1.25**
Pr. a 80 %	.3277	.3048	-6.99	.3270	-0.21	**.3332**	**1.68**	**.3352**	**2.29**	**.3358**	**2.47**	**.3362**	**2.59**
Pr. a 90 %	.2356	.2273	-3.52	**.2435**	**3.35**	**.2459**	**4.37**	**.2474**	**5.01**	**.2477**	**5.14**	**.2482**	**5.35**
Pr. a 100 %	.1197	.1160	-3.09	.1185	-1.00	**.1205**	**0.67**	**.1201**	**0.33**	**.1199**	**0.17**	**.1198**	**0.08**
Pr. a 5 docs.	.6609	.6261	-5.27	.6565	-0.67	**.6696**	**1.32**	**.6783**	**2.63**	**.6870**	**3.95**	**.6826**	**3.28**
Pr. a 10 docs.	.6283	.5891	-6.24	.6239	-0.70	**.6413**	**2.07**	**.6457**	**2.77**	**.6522**	**3.80**	**.6413**	**2.07**
Pr. a 15 docs.	.5928	.5449	-8.08	.5899	-0.49	**.6014**	**1.45**	**.5957**	**0.49**	**.5971**	**0.73**	**.5971**	**0.73**
Pr. a 20 docs.	.5446	.5011	-7.99	**.5489**	**0.79**	**.5565**	**2.19**	**.5598**	**2.79**	**.5587**	**2.59**	**.5576**	**2.39**
Pr. a 30 docs.	.4928	.4630	-6.05	.4928	0.00	**.5036**	**2.19**	**.5058**	**2.64**	**.5036**	**2.19**	**.5007**	**1.60**
Pr. a 100 docs.	.3300	.3004	-8.97	.3237	-1.91	.3296	-0.12	**.3337**	**1.12**	**.3361**	**1.85**	**.3348**	**1.45**
Pr. a 200 docs.	.2234	.2113	-5.42	.2215	-0.85	**.2246**	**0.54**	**.2258**	**1.07**	**.2257**	**1.03**	**.2257**	**1.03**
Pr. a 500 docs.	.1090	.1043	-4.31	.1083	-0.64	**.1097**	**0.64**	**.1100**	**0.92**	**.1101**	**1.01**	**.1103**	**1.19**
Pr. a 1000 docs.	.0587	.0571	-2.73	.0587	0.00	**.0591**	**0.68**	**.0592**	**0.85**	**.0593**	**1.02**	**.0592**	**0.85**

Tabla F.10: Puesta a punto del factor de ponderación ω. Resultados para *sintagmas nominales* con el corpus CLEF 2001-02.A con el analizador CASCADE: consultas cortas, $\omega \in \{10, 12, 14, 16, 18, 20\}$

	lem	CASCADE											
		$\omega = 10$	%Δ	$\omega = 12$	%Δ	$\omega = 14$	%Δ	$\omega = 16$	%Δ	$\omega = 18$	%Δ	$\omega = 20$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46k	46k	–	46k	–	46k	–	46k	–	46k	–	46k	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2700	**2723**	**0.85**	**2721**	**0.78**	**2720**	**0.74**	**2720**	**0.74**	**2721**	**0.78**	**2721**	**0.78**
Pr. no int.	.4829	**.4915**	**1.78**	**.4913**	**1.74**	**.4900**	**1.47**	**.4890**	**1.26**	**.4884**	**1.14**	**.4881**	**1.08**
Pr. doc.	.5327	**.5447**	**2.25**	**.5440**	**2.12**	**.5429**	**1.91**	**.5419**	**1.73**	**.5414**	**1.63**	**.5408**	**1.52**
R-pr.	.4848	.4844	-0.08	.4842	-0.12	**.4866**	**0.37**	**.4869**	**0.43**	**.4874**	**0.54**	**.4872**	**0.50**
Pr. a 0 %	.8293	.8281	-0.14	.8281	-0.14	.8271	-0.27	.8275	-0.22	.8272	-0.25	.8276	-0.20
Pr. a 10 %	.7463	**.7595**	**1.77**	**.7644**	**2.43**	**.7560**	**1.30**	**.7555**	**1.23**	**.7535**	**0.96**	**.7535**	**0.96**
Pr. a 20 %	.6771	**.6863**	**1.36**	**.6880**	**1.61**	**.6828**	**0.84**	**.6804**	**0.49**	**.6795**	**0.35**	**.6791**	**0.30**
Pr. a 30 %	.6138	**.6218**	**1.30**	**.6229**	**1.48**	**.6211**	**1.19**	**.6208**	**1.14**	**.6201**	**1.03**	**.6209**	**1.16**
Pr. a 40 %	.5502	**.5547**	**0.82**	**.5563**	**1.11**	**.5538**	**0.65**	**.5535**	**0.60**	**.5525**	**0.42**	**.5524**	**0.40**
Pr. a 50 %	.4931	**.4982**	**1.03**	**.4981**	**1.01**	**.4974**	**0.87**	**.4975**	**0.89**	**.4974**	**0.87**	**.4985**	**1.10**
Pr. a 60 %	.4496	**.4624**	**2.85**	**.4617**	**2.69**	**.4602**	**2.36**	**.4596**	**2.22**	**.4598**	**2.27**	**.4584**	**1.96**
Pr. a 70 %	.3853	**.3899**	**1.19**	**.3893**	**1.04**	**.3878**	**0.65**	**.3880**	**0.70**	**.3881**	**0.73**	**.3881**	**0.73**
Pr. a 80 %	.3277	**.3350**	**2.23**	**.3350**	**2.23**	**.3343**	**2.01**	**.3334**	**1.74**	**.3328**	**1.56**	**.3327**	**1.53**
Pr. a 90 %	.2356	**.2466**	**4.67**	**.2458**	**4.33**	**.2439**	**3.52**	**.2431**	**3.18**	**.2430**	**3.14**	**.2413**	**2.42**
Pr. a 100 %	.1197	.1197	0.00	.1196	-0.08	.1197	0.00	.1196	-0.08	.1196	-0.08	.1196	-0.08
Pr. a 5 docs.	.6609	**.6783**	**2.63**	**.6696**	**1.32**	**.6696**	**1.32**	**.6652**	**0.65**	**.6652**	**0.65**	**.6652**	**0.65**
Pr. a 10 docs.	.6283	**.6435**	**2.42**	**.6435**	**2.42**	**.6391**	**1.72**	**.6391**	**1.72**	**.6391**	**1.72**	**.6370**	**1.38**
Pr. a 15 docs.	.5928	.5928	0.00	.5899	-0.49	.5884	-0.74	.5899	-0.49	.5884	-0.74	.5884	-0.74
Pr. a 20 docs.	.5446	**.5576**	**2.39**	**.5554**	**1.98**	**.5543**	**1.78**	**.5543**	**1.78**	**.5543**	**1.78**	**.5543**	**1.78**
Pr. a 30 docs.	.4928	**.4993**	**1.32**	**.5007**	**1.60**	**.4993**	**1.32**	**.4986**	**1.18**	**.4986**	**1.18**	**.4971**	**0.87**
Pr. a 100 docs.	.3300	**.3348**	**1.45**	**.3350**	**1.52**	**.3337**	**1.12**	**.3335**	**1.06**	**.3330**	**0.91**	**.3330**	**0.91**
Pr. a 200 docs.	.2234	**.2258**	**1.07**	**.2257**	**1.03**	**.2252**	**0.81**	**.2251**	**0.76**	**.2249**	**0.67**	**.2247**	**0.58**
Pr. a 500 docs.	.1090	**.1100**	**0.92**	**.1098**	**0.73**	**.1098**	**0.73**	**.1095**	**0.46**	**.1093**	**0.28**	**.1093**	**0.28**
Pr. a 1000 docs.	.0587	**.0592**	**0.85**	**.0592**	**0.85**	**.0591**	**0.68**	**.0591**	**0.68**	**.0592**	**0.85**	**.0592**	**0.85**

Tabla F.11: Puesta a punto del factor de ponderación ω. Resultados para *sintagmas nominales* con el corpus CLEF 2001-02·A con el analizador CASCADE: consultas largas, $\omega \in \{1, 2, 3, 4, 5, 8\}$

	lem	CASCADE											
		$\omega = 1$	%Δ	$\omega = 2$	%Δ	$\omega = 3$	%Δ	$\omega = 4$	%Δ	$\omega = 5$	%Δ	$\omega = 8$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46k	46k	–	46k	–	46k	–	46k	–	46k	–	46k	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2762	2696	-2.39	2749	-0.47	**2764**	**0.07**	**2775**	**0.47**	**2776**	**0.51**	**2785**	**0.83**
Pr. no int.	.5239	.5005	-4.47	.5213	-0.50	**.5253**	**0.27**	**.5281**	**0.80**	**.5287**	**0.92**	**.5307**	**1.30**
Pr. doc.	.5690	.5270	-7.38	.5607	-1.46	**.5713**	**0.40**	**.5767**	**1.35**	**.5784**	**1.65**	**.5797**	**1.88**
R-pr.	.5075	.4916	-3.13	.5072	-0.06	**.5128**	**1.04**	**.5099**	**0.47**	**.5107**	**0.63**	**.5112**	**0.73**
Pr. a 0 %	.8845	.8227	-6.99	.8392	-5.12	.8321	-5.92	.8465	-4.30	.8434	-4.65	.8622	-2.52
Pr. a 10 %	.7914	.7794	-1.52	**.8003**	**1.12**	**.7972**	**0.73**	**.8055**	**1.78**	**.7999**	**1.07**	**.7986**	**0.91**
Pr. a 20 %	.7136	.6943	-2.70	**.7208**	**1.01**	**.7310**	**2.44**	**.7270**	**1.88**	**.7251**	**1.61**	**.7289**	**2.14**
Pr. a 30 %	.6591	.6284	-4.66	.6573	-0.27	**.6682**	**1.38**	**.6753**	**2.46**	**.6750**	**2.41**	**.6694**	**1.56**
Pr. a 40 %	.6085	.5759	-5.36	.5925	-2.63	.6016	-1.13	.6037	-0.79	**.6107**	**0.36**	**.6135**	**0.82**
Pr. a 50 %	.5557	.5298	-4.66	.5480	-1.39	.5521	-0.65	**.5576**	**0.34**	**.5571**	**0.25**	**.5622**	**1.17**
Pr. a 60 %	.5006	.4709	-5.93	.4998	-0.16	**.5056**	**1.00**	**.5107**	**2.02**	**.5129**	**2.46**	**.5150**	**2.88**
Pr. a 70 %	.4161	.4014	-3.53	.4129	-0.77	.4155	-0.14	.4150	-0.26	**.4164**	**0.07**	**.4254**	**2.24**
Pr. a 80 %	.3509	.3492	-0.48	**.3592**	**2.37**	**.3623**	**3.25**	**.3621**	**3.19**	**.3617**	**3.08**	**.3607**	**2.79**
Pr. a 90 %	.2492	.2439	-2.13	**.2584**	**3.69**	**.2618**	**5.06**	**.2613**	**4.86**	**.2596**	**4.17**	**.2575**	**3.33**
Pr. a 100 %	.1289	.1264	-1.94	**.1315**	**2.02**	**.1300**	**0.85**	**.1305**	**1.24**	**.1308**	**1.47**	**.1316**	**2.09**
Pr. a 5 docs.	.6957	.6652	-4.38	**.7130**	**2.49**	.6957	0.00	.6913	-0.63	.6913	-0.63	.6957	0.00
Pr. a 10 docs.	.6543	.6239	-4.65	**.6587**	**0.67**	**.6630**	**1.33**	**.6717**	**2.66**	**.6674**	**2.00**	**.6717**	**2.66**
Pr. a 15 docs.	.6188	.5942	-3.98	.6159	-0.47	**.6232**	**0.71**	**.6290**	**1.65**	**.6304**	**1.87**	**.6290**	**1.65**
Pr. a 20 docs.	.5880	.5587	-4.98	.5804	-1.29	**.5902**	**0.37**	**.5891**	**0.19**	**.5935**	**0.94**	**.5935**	**0.94**
Pr. a 30 docs.	.5304	.5051	-4.77	**.5326**	**0.41**	**.5391**	**1.64**	**.5420**	**2.19**	**.5428**	**2.34**	**.5391**	**1.64**
Pr. a 100 docs.	.3509	.3300	-5.96	.3430	-2.25	.3485	-0.68	.3502	-0.20	**.3522**	**0.37**	**.3517**	**0.23**
Pr. a 200 docs.	.2315	.2249	-2.85	**.2328**	**0.56**	**.2336**	**0.91**	**.2342**	**1.17**	**.2345**	**1.30**	**.2340**	**1.08**
Pr. a 500 docs.	.1115	.1098	-1.52	**.1118**	**0.27**	**.1126**	**0.99**	**.1126**	**0.99**	**.1127**	**1.08**	**.1125**	**0.90**
Pr. a 1000 docs.	.0600	.0586	-2.33	.0598	-0.33	**.0601**	**0.17**	**.0603**	**0.50**	**.0603**	**0.50**	**.0605**	**0.83**

Tabla F.12: Puesta a punto del factor de ponderación ω. Resultados para *sintagmas nominales* con el corpus CLEF 2001-02-A con el analizador CASCADE: consultas largas, $\omega \in \{10, 12, 14, 16, 18, 20\}$

	lem	CASCADE											
		$\omega = 10$	%Δ	$\omega = 12$	%Δ	$\omega = 14$	%Δ	$\omega = 16$	%Δ	$\omega = 18$	%Δ	$\omega = 20$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46k	46k	–	46k	–	46k	–	46k	–	46k	–	46k	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2762	**2781**	**0.69**	**2782**	**0.72**	**2779**	**0.62**	**2781**	**0.69**	**2781**	**0.69**	**2779**	**0.62**
Pr. no int.	.5239	**.5303**	**1.22**	**.5298**	**1.13**	**.5291**	**0.99**	**.5292**	**1.01**	**.5286**	**0.90**	**.5279**	**0.76**
Pr. doc.	.5690	**.5791**	**1.78**	**.5785**	**1.67**	**.5776**	**1.51**	**.5774**	**1.48**	**.5768**	**1.37**	**.5760**	**1.23**
R-pr.	.5075	**.5128**	**1.04**	**.5130**	**1.08**	**.5139**	**1.26**	**.5134**	**1.16**	**.5125**	**0.99**	**.5111**	**0.71**
Pr. a 0 %	.8845	.8617	-2.58	.8650	-2.20	.8672	-1.96	.8753	-1.04	.8752	-1.05	.8755	-1.02
Pr. a 10 %	.7914	**.8020**	**1.34**	**.8030**	**1.47**	**.7969**	**0.69**	**.7987**	**0.92**	**.7978**	**0.81**	**.7974**	**0.76**
Pr. a 20 %	.7136	**.7277**	**1.98**	**.7262**	**1.77**	**.7240**	**1.46**	**.7273**	**1.92**	**.7260**	**1.74**	**.7246**	**1.54**
Pr. a 30 %	.6591	**.6678**	**1.32**	**.6652**	**0.93**	**.6634**	**0.65**	**.6620**	**0.44**	**.6612**	**0.32**	**.6610**	**0.29**
Pr. a 40 %	.6085	**.6124**	**0.64**	**.6112**	**0.44**	**.6110**	**0.41**	**.6108**	**0.38**	**.6104**	**0.31**	**.6106**	**0.35**
Pr. a 50 %	.5557	**.5619**	**1.12**	**.5615**	**1.04**	**.5612**	**0.99**	**.5610**	**0.95**	**.5609**	**0.94**	**.5606**	**0.88**
Pr. a 60 %	.5006	**.5128**	**2.44**	**.5126**	**2.40**	**.5125**	**2.38**	**.5104**	**1.96**	**.5091**	**1.70**	**.5081**	**1.50**
Pr. a 70 %	.4161	**.4248**	**2.09**	**.4230**	**1.66**	**.4239**	**1.87**	**.4235**	**1.78**	**.4220**	**1.42**	**.4215**	**1.30**
Pr. a 80 %	.3509	**.3587**	**2.22**	**.3574**	**1.85**	**.3560**	**1.45**	**.3557**	**1.37**	**.3553**	**1.25**	**.3545**	**1.03**
Pr. a 90 %	.2492	**.2568**	**3.05**	**.2560**	**2.73**	**.2557**	**2.61**	**.2553**	**2.45**	**.2553**	**2.45**	**.2545**	**2.13**
Pr. a 100 %	.1289	**.1319**	**2.33**	**.1320**	**2.40**	**.1322**	**2.56**	**.1321**	**2.48**	**.1320**	**2.40**	**.1319**	**2.33**
Pr. a 5 docs.	.6957	.6957	0.00	.6957	0.00	**.7000**	**0.62**	**.7000**	**0.62**	.6913	-0.63	.6913	-0.63
Pr. a 10 docs.	.6543	**.6761**	**3.33**	**.6696**	**2.34**	**.6674**	**2.00**	**.6630**	**1.33**	**.6609**	**1.01**	**.6587**	**0.67**
Pr. a 15 docs.	.6188	**.6304**	**1.87**	**.6290**	**1.65**	**.6304**	**1.87**	**.6290**	**1.65**	**.6275**	**1.41**	**.6261**	**1.18**
Pr. a 20 docs.	.5880	**.5935**	**0.94**	**.5902**	**0.37**	**.5902**	**0.37**	**.5913**	**0.56**	**.5913**	**0.56**	**.5913**	**0.56**
Pr. a 30 docs.	.5304	**.5391**	**1.64**	**.5399**	**1.79**	**.5370**	**1.24**	**.5377**	**1.38**	**.5362**	**1.09**	**.5362**	**1.09**
Pr. a 100 docs.	.3509	**.3517**	**0.23**	**.3524**	**0.43**	**.3526**	**0.48**	**.3533**	**0.68**	**.3528**	**0.54**	**.3524**	**0.43**
Pr. a 200 docs.	.2315	**.2341**	**1.12**	**.2339**	**1.04**	**.2339**	**1.04**	**.2335**	**0.86**	**.2333**	**0.78**	**.2334**	**0.82**
Pr. a 500 docs.	.1115	**.1123**	**0.72**	**.1123**	**0.72**	**.1122**	**0.63**	**.1122**	**0.63**	**.1121**	**0.54**	**.1121**	**0.54**
Pr. a 1000 docs.	.0600	**.0605**	**0.83**	**.0605**	**0.83**	**.0604**	**0.67**	**.0605**	**0.83**	**.0605**	**0.83**	**.0604**	**0.67**

	cortas								*largas*							
ω	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*
Pr. a 5	**3.28**	**3.28**	**2.63**	**1.97**	**1.97**	**1.97**	**1.97**	**0.65**	**2.49**	**1.24**	**1.24**	**1.24**	**1.24**	0.00	0.00	0.00
Pr. a 10	**2.42**	**3.80**	**3.45**	**3.45**	**3.45**	**2.07**	**1.72**	**0.33**	0.00	-0.99	0.00	-0.32	-0.99	0.00	0.00	0.00
Pr. a 15	**2.19**	**2.19**	**1.94**	**1.45**	**1.70**	**2.19**	**1.70**	**0.98**	0.00	**0.94**	**0.94**	**0.71**	**0.47**	0.00	0.00	-0.23
Pr. a 20	**2.79**	**3.20**	**2.59**	**2.79**	**2.99**	**2.59**	**2.19**	**1.78**	0.00	**0.75**	**0.94**	**0.19**	**0.19**	**0.37**	**0.19**	0.00
Pr. a 30	**3.23**	**1.91**	**2.33**	**2.19**	**1.75**	**1.46**	**1.18**	**0.59**	**0.28**	**0.96**	**0.41**	**0.83**	**0.96**	**0.28**	-0.13	**0.15**
Pr. a 100	**0.79**	**2.18**	**2.58**	**2.36**	**2.42**	**1.64**	**1.64**	**1.12**	**0.06**	**0.91**	**1.11**	**1.05**	**0.85**	**0.54**	**0.31**	**0.17**
Pr. a 200	**1.21**	**2.51**	**1.88**	**1.84**	**1.66**	**1.48**	**1.34**	**0.67**	**1.47**	**1.51**	**1.12**	**1.04**	**0.91**	**0.52**	**0.52**	**0.30**
Pr. a 5	**1.97**	**2.63**	**3.28**	**3.28**	**3.28**	**3.28**	**3.28**	**2.63**	**1.87**	**1.24**	**1.24**	**1.24**	**1.24**	0.00	0.00	0.00
Pr. a 10	-2.44	-1.05	-0.35	-0.35	**1.03**	**1.38**	**1.03**	**1.03**	**0.34**	-0.66	0.00	0.00	-0.66	**1.01**	**0.34**	0.00
Pr. a 15	-1.23	-0.25	**0.49**	**0.49**	0.00	**0.49**	**0.98**	**0.49**	**0.47**	**1.18**	**1.87**	**1.41**	**1.18**	**0.47**	0.00	0.00
Pr. a 20	**0.59**	**0.79**	**2.59**	**2.79**	**2.99**	**3.20**	**2.79**	**2.59**	**0.75**	**1.85**	**2.04**	**1.48**	**1.12**	**1.12**	**0.75**	**0.37**
Pr. a 30	**1.60**	**2.50**	**3.51**	**3.37**	**3.51**	**3.08**	**2.78**	**1.32**	**0.15**	**0.83**	**0.83**	**0.55**	**0.96**	**0.55**	**0.15**	**0.28**
Pr. a 100	**0.91**	**3.30**	**3.30**	**3.03**	**2.52**	**2.12**	**1.97**	**1.18**	**0.74**	**1.97**	**1.54**	**1.48**	**1.17**	**0.60**	**0.31**	**0.31**
Pr. a 200	**0.13**	**2.82**	**3.04**	**2.73**	**2.37**	**2.01**	**1.92**	**1.12**	**1.90**	**1.90**	**1.47**	**1.34**	**1.04**	**0.60**	**0.60**	**0.35**
Pr. a 5	**0.65**	**2.63**	**3.28**	**2.63**	**2.63**	**2.63**	**2.63**	**2.63**	**3.12**	**2.49**	**1.24**	**0.62**	**0.62**	-0.63	-0.63	0.00
Pr. a 10	-3.47	-0.70	-0.35	**1.03**	**2.07**	**2.42**	**1.38**	**1.38**	**0.34**	0.00	**2.34**	**2.66**	**2.34**	**2.34**	**0.67**	0.00
Pr. a 15	-2.45	-0.25	**0.73**	**0.73**	**0.24**	**0.24**	**0.98**	**0.73**	**0.94**	**1.41**	**2.81**	**2.81**	**2.59**	**1.18**	**1.18**	**0.47**
Pr. a 20	**0.39**	**1.40**	**2.39**	**3.38**	**3.38**	**2.99**	**2.79**	**2.79**	**0.75**	**3.33**	**2.23**	**1.12**	**1.31**	**1.48**	**0.75**	**0.75**
Pr. a 30	0.00	**1.60**	**3.37**	**3.23**	**3.37**	**3.08**	**2.78**	**1.46**	**2.19**	**3.43**	**3.56**	**2.47**	**2.47**	**2.06**	**1.64**	**1.24**
Pr. a 100	-2.36	**1.45**	**2.36**	**2.52**	**2.12**	**2.03**	**1.73**	**1.12**	-0.06	**1.60**	**1.37**	**1.37**	**1.37**	**0.54**	**0.37**	**0.23**
Pr. a 200	-1.97	**2.24**	**3.04**	**2.91**	**2.51**	**2.06**	**2.06**	**1.16**	**2.55**	**2.76**	**2.03**	**1.56**	**1.34**	**0.56**	**0.43**	**0.26**
Pr. a 5	**0.65**	**3.28**	**3.28**	**2.63**	**2.63**	**2.63**	**2.63**	**2.63**	**2.49**	**1.87**	**0.62**	**0.62**	**0.62**	**0.62**	**0.62**	0.00
Pr. a 10	-3.47	-0.70	-0.35	**1.38**	**2.77**	**2.07**	**2.07**	**1.72**	-0.32	-0.32	**1.01**	**2.00**	**2.66**	**2.34**	**1.01**	0.00
Pr. a 15	-2.45	0.00	**0.73**	**0.73**	**0.73**	**0.24**	**0.98**	**0.73**	**1.87**	**1.87**	**3.05**	**3.52**	**2.81**	**1.65**	**1.65**	**0.94**
Pr. a 20	**0.59**	**1.40**	**2.19**	**2.99**	**2.79**	**3.58**	**3.98**	**3.78**	**0.94**	**3.15**	**2.96**	**2.04**	**2.60**	**2.04**	**1.31**	**1.12**
Pr. a 30	-0.30	**1.18**	**2.92**	**2.92**	**3.51**	**2.33**	**2.33**	**1.91**	**1.64**	**4.37**	**3.96**	**3.28**	**2.34**	**2.06**	**1.64**	**1.24**
Pr. a 100	-4.03	**0.73**	**2.58**	**2.91**	**2.58**	**2.24**	**1.58**	**1.06**	-0.46	**1.85**	**1.85**	**1.74**	**1.80**	**0.85**	**0.48**	**0.23**
Pr. a 200	-2.69	**1.97**	**2.82**	**2.95**	**2.91**	**2.19**	**2.19**	**1.34**	**1.73**	**2.72**	**1.68**	**1.47**	**1.30**	**0.60**	**0.43**	**0.26**
Pr. a 5	-0.67	**1.32**	**3.28**	**5.27**	**4.60**	**3.95**	**4.60**	**3.95**	**4.99**	**3.12**	**2.49**	**0.62**	**0.62**	**2.49**	**2.49**	**0.62**
Pr. a 10	-3.47	**2.07**	**3.45**	**3.45**	**3.80**	**2.77**	**1.72**	**1.38**	**2.00**	**2.00**	**1.33**	**1.67**	**3.00**	**2.66**	**1.67**	-0.32
Pr. a 15	-2.70	**0.24**	**1.21**	**1.21**	**0.98**	**0.73**	**1.45**	**1.70**	**1.41**	**1.87**	**3.05**	**2.81**	**2.59**	**1.41**	**1.18**	**0.94**
Pr. a 20	-0.61	**1.19**	**2.19**	**2.79**	**2.99**	**4.39**	**4.59**	**3.20**	0.00	**1.85**	**2.96**	**2.04**	**1.85**	**2.60**	**2.04**	**1.12**
Pr. a 30	-2.94	**2.33**	**4.55**	**4.69**	**5.28**	**4.55**	**3.51**	**3.08**	**2.47**	**4.51**	**3.28**	**2.88**	**2.88**	**2.60**	**2.47**	**2.06**
Pr. a 100	-7.64	-0.52	**2.24**	**3.48**	**3.48**	**3.36**	**3.30**	**2.30**	-1.25	**2.22**	**2.28**	**1.85**	**1.85**	**1.28**	**0.80**	**0.23**
Pr. a 200	-6.31	**0.09**	**2.06**	**3.13**	**3.40**	**2.73**	**2.60**	**1.70**	**0.26**	**2.16**	**2.38**	**1.68**	**1.21**	**0.78**	**0.56**	**0.26**
Pr. a 5	**3.95**	**5.92**	**7.88**	**8.55**	**7.23**	**6.57**	**5.27**	**3.95**	**4.99**	**4.99**	**4.37**	**3.12**	**3.12**	**3.74**	**3.12**	**0.62**
Pr. a 10	-1.05	**3.10**	**4.84**	**4.49**	**4.84**	**2.42**	**3.10**	**1.72**	**4.66**	**5.65**	**3.33**	**2.66**	**4.33**	**2.66**	**2.00**	**1.01**
Pr. a 15	-2.21	**0.98**	**0.98**	**1.94**	**2.19**	**1.94**	**2.43**	**1.70**	**3.99**	**3.05**	**4.22**	**3.05**	**2.81**	**2.12**	**2.59**	**1.18**
Pr. a 20	-0.40	**3.38**	**3.98**	**3.58**	**4.19**	**5.58**	**5.38**	**4.19**	**2.41**	**2.77**	**3.33**	**2.77**	**2.77**	**3.52**	**3.33**	**1.31**
Pr. a 30	-2.21	**2.50**	**5.28**	**6.45**	**6.76**	**6.17**	**6.03**	**4.40**	**3.96**	**4.24**	**3.43**	**3.70**	**4.11**	**3.43**	**2.73**	**2.34**
Pr. a 100	-7.30	-0.79	**2.24**	**3.64**	**3.82**	**4.27**	**4.15**	**3.24**	-2.31	**2.02**	**3.16**	**2.96**	**3.33**	**2.22**	**1.60**	**0.43**
Pr. a 200	-7.48	-0.09	**2.01**	**2.91**	**3.27**	**3.36**	**3.13**	**2.42**	-1.04	**2.42**	**2.94**	**2.89**	**2.51**	**1.34**	**1.12**	**0.35**
Pr. a 5	**2.63**	**5.27**	**6.57**	**7.23**	**6.57**	**5.92**	**4.60**	**3.28**	**6.24**	**4.99**	**4.37**	**3.74**	**2.49**	**2.49**	**2.49**	**0.62**
Pr. a 10	-1.05	**2.42**	**4.14**	**4.84**	**3.80**	**2.42**	**2.77**	**2.42**	**5.32**	**5.65**	**5.00**	**3.67**	**4.66**	**3.33**	**2.34**	**0.67**
Pr. a 15	-2.70	0.00	**0.73**	**0.73**	**0.98**	**1.21**	**2.43**	**1.45**	**3.28**	**3.75**	**4.22**	**3.52**	**3.52**	**2.34**	**3.05**	**1.18**
Pr. a 20	-0.61	**2.19**	**4.19**	**4.79**	**5.18**	**5.78**	**5.38**	**4.79**	**2.60**	**3.89**	**2.96**	**2.77**	**2.41**	**3.33**	**3.15**	**1.48**
Pr. a 30	-1.93	**4.26**	**6.03**	**6.76**	**7.79**	**6.45**	**5.72**	**4.85**	**3.15**	**4.24**	**3.83**	**4.11**	**3.96**	**3.56**	**2.60**	**2.19**
Pr. a 100	-6.67	0.00	**2.30**	**4.15**	**4.48**	**4.94**	**5.00**	**3.70**	-2.79	**1.97**	**3.02**	**3.16**	**3.45**	**2.48**	**1.91**	**0.74**
Pr. a 200	-6.76	-0.18	**2.55**	**3.72**	**4.39**	**4.07**	**3.67**	**2.64**	-1.25	**2.51**	**3.59**	**2.98**	**2.68**	**1.64**	**1.43**	**0.48**

Tabla F.13: Puesta a punto de los parámetros para la selección automática de dependencias mediante realimentación. Resultados para el Corpus CLEF 2001-02·A (número de documentos n'_1=5; de arriba a abajo, número de términos $t' \in \{5, 10, 15, 20, 30, 40, 50\}$)

	cortas								*largas*							
ω	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*
Pr. a 5	**1.97**	**1.97**	0.00	**0.65**	**0.65**	**1.32**	**1.97**	**1.32**	**1.24**	**1.24**	**1.24**	**1.24**	0.00	-0.63	-0.63	0.00
Pr. a 10	**2.42**	**2.77**	**2.42**	**2.07**	**2.07**	**1.72**	**1.03**	**0.68**	**1.67**	**0.34**	**1.01**	**0.67**	**0.34**	**0.34**	0.00	**0.34**
Pr. a 15	**3.17**	**2.19**	**1.94**	**1.70**	**0.98**	**0.73**	**0.24**	**0.24**	**0.94**	**1.18**	**0.94**	**0.47**	**0.24**	0.00	-0.47	-0.23
Pr. a 20	**4.19**	**3.78**	**2.59**	**2.59**	**2.19**	**1.98**	**1.60**	**1.40**	**0.94**	**0.94**	**0.75**	**0.19**	**0.19**	**0.75**	**0.37**	0.00
Pr. a 30	**3.08**	**2.05**	**1.75**	**1.91**	**1.18**	**0.59**	**0.43**	**0.43**	**1.51**	**1.38**	**1.24**	**1.38**	**1.38**	**0.70**	**0.70**	**0.70**
Pr. a 100	**1.30**	**1.97**	**1.91**	**1.30**	**1.18**	**1.00**	**1.18**	**0.73**	**0.37**	**0.68**	**0.74**	**0.74**	**0.54**	**0.37**	**0.17**	**0.11**
Pr. a 200	**1.84**	**2.73**	**2.28**	**2.19**	**1.84**	**1.61**	**1.61**	**0.72**	**1.51**	**1.21**	**1.21**	**1.08**	**0.99**	**0.56**	**0.43**	**0.30**
Pr. a 5	0.00	0.00	**0.65**	**1.97**	**1.97**	**1.32**	**1.32**	**0.65**	**3.74**	**3.12**	**3.12**	**2.49**	**1.24**	**0.62**	**0.62**	**0.62**
Pr. a 10	**2.42**	**2.07**	**1.38**	**0.33**	**1.03**	**1.38**	**1.38**	**0.33**	**2.34**	**0.34**	**0.67**	**0.67**	**0.34**	**0.67**	0.00	**0.34**
Pr. a 15	-0.25	**0.98**	**0.98**	**0.49**	**0.98**	**0.98**	**0.98**	**0.24**	**2.34**	**2.12**	**2.34**	**1.65**	**1.18**	**0.24**	-0.23	0.00
Pr. a 20	**1.19**	**1.60**	**1.78**	**2.99**	**2.99**	**2.59**	**1.98**	**2.59**	**2.96**	**2.96**	**2.41**	**1.67**	**1.31**	**0.94**	**0.37**	**0.37**
Pr. a 30	**2.33**	**3.37**	**3.37**	**4.10**	**3.37**	**2.33**	**1.60**	**0.87**	**3.70**	**2.73**	**2.60**	**2.19**	**1.92**	**1.38**	**0.83**	**0.83**
Pr. a 100	**0.52**	**2.36**	**3.36**	**2.97**	**2.91**	**2.36**	**2.18**	**1.24**	**0.43**	**0.60**	**0.60**	**0.68**	**0.54**	**0.31**	**0.17**	**0.11**
Pr. a 200	**0.85**	**2.91**	**2.60**	**2.37**	**2.01**	**2.15**	**2.01**	**1.03**	**2.07**	**1.64**	**1.34**	**1.08**	**0.95**	**0.43**	**0.39**	**0.30**
Pr. a 5	**0.65**	**1.32**	**1.32**	**2.63**	**2.63**	**1.32**	**0.65**	**1.97**	**3.74**	**1.24**	**1.24**	**1.24**	**0.62**	**0.62**	**0.62**	**0.62**
Pr. a 10	-1.05	**1.03**	**0.33**	0.00	**1.03**	**2.07**	**2.07**	**0.33**	**1.01**	0.00	**0.67**	**1.01**	**0.34**	**0.67**	0.00	0.00
Pr. a 15	-1.47	**0.98**	**0.73**	**0.24**	**0.73**	**1.21**	**1.21**	**0.98**	**1.41**	**2.34**	**2.59**	**1.87**	**1.18**	**0.47**	-0.23	0.00
Pr. a 20	**0.59**	**2.79**	**2.59**	**3.98**	**3.20**	**2.39**	**2.19**	**2.59**	**1.85**	**3.33**	**2.23**	**1.48**	**1.31**	**0.56**	**0.19**	**0.37**
Pr. a 30	**4.10**	**4.99**	**4.69**	**4.99**	**4.55**	**3.23**	**1.75**	**1.18**	**3.56**	**2.88**	**2.73**	**2.19**	**1.79**	**1.38**	**0.96**	**0.83**
Pr. a 100	**0.52**	**3.15**	**4.03**	**3.70**	**3.03**	**2.97**	**2.91**	**1.45**	**1.42**	**1.74**	**1.65**	**1.65**	**1.42**	**0.37**	**0.23**	**0.31**
Pr. a 200	**0.22**	**2.28**	**1.92**	**2.15**	**1.79**	**1.92**	**2.01**	**1.30**	**2.63**	**2.12**	**1.51**	**1.38**	**1.12**	**0.60**	**0.52**	**0.26**
Pr. a 5	**3.95**	**3.28**	**3.95**	**5.27**	**4.60**	**2.63**	**1.32**	**1.97**	**6.87**	**3.74**	**3.12**	**2.49**	**1.87**	**1.24**	**0.62**	0.00
Pr. a 10	**1.03**	**1.38**	**1.72**	**1.03**	**1.72**	**2.42**	**2.07**	**0.68**	**2.66**	**0.67**	**1.01**	**1.33**	**0.34**	**1.33**	**0.67**	**0.34**
Pr. a 15	**0.49**	**2.19**	**1.21**	**0.73**	**0.73**	**0.73**	**0.73**	**1.21**	**2.34**	**2.34**	**2.59**	**2.34**	**1.87**	**0.71**	**0.24**	**0.24**
Pr. a 20	**3.38**	**3.20**	**3.20**	**3.98**	**3.78**	**3.58**	**3.38**	**3.20**	**1.48**	**3.15**	**2.96**	**2.23**	**2.04**	**1.31**	**0.94**	**0.94**
Pr. a 30	**3.96**	**6.31**	**5.58**	**5.28**	**5.28**	**3.51**	**2.19**	**1.60**	**3.15**	**3.15**	**3.15**	**2.34**	**2.19**	**1.64**	**1.09**	**1.24**
Pr. a 100	**0.27**	**3.70**	**4.27**	**3.88**	**3.48**	**3.03**	**3.03**	**1.97**	**1.17**	**1.91**	**1.80**	**1.74**	**1.42**	**0.54**	**0.23**	**0.31**
Pr. a 200	-0.40	**1.92**	**1.92**	**2.42**	**1.97**	**1.88**	**2.19**	**1.61**	**2.38**	**2.72**	**2.16**	**1.86**	**1.47**	**0.78**	**0.60**	**0.39**
Pr. a 5	**1.32**	**4.60**	**7.23**	**5.27**	**5.92**	**3.28**	**2.63**	**1.97**	**4.99**	**4.99**	**6.24**	**3.12**	**2.49**	**3.12**	**2.49**	0.00
Pr. a 10	-0.35	**2.77**	**2.77**	**3.45**	**3.10**	**2.77**	**2.42**	**1.72**	-0.66	**0.67**	**0.34**	**1.01**	0.00	**0.67**	**0.34**	-0.32
Pr. a 15	0.00	**1.94**	**0.49**	**0.98**	**1.45**	**0.98**	**0.73**	**0.98**	0.00	**0.47**	**1.41**	**1.87**	**0.94**	**0.24**	**0.24**	**0.24**
Pr. a 20	**2.59**	**3.98**	**3.98**	**3.98**	**3.78**	**3.38**	**3.38**	**3.38**	-0.17	**2.04**	**2.23**	**1.67**	**1.85**	**1.12**	**0.94**	**0.56**
Pr. a 30	**1.91**	**4.85**	**5.58**	**5.13**	**5.58**	**4.69**	**3.67**	**2.33**	**1.51**	**2.34**	**2.73**	**1.92**	**1.51**	**2.06**	**1.79**	**1.51**
Pr. a 100	-1.91	**2.52**	**4.21**	**4.82**	**4.33**	**3.36**	**3.36**	**2.03**	-1.42	**1.74**	**1.37**	**1.74**	**1.65**	**0.68**	**0.54**	**0.11**
Pr. a 200	-2.69	**2.46**	**3.09**	**3.58**	**3.09**	**2.51**	**2.37**	**1.75**	**0.48**	**2.03**	**2.51**	**2.03**	**1.60**	**0.73**	**0.60**	**0.30**
Pr. a 5	**3.95**	**3.95**	**7.88**	**9.20**	**8.55**	**5.92**	**4.60**	**3.95**	**8.74**	**8.12**	**7.49**	**6.24**	**5.62**	**4.37**	**3.74**	**0.62**
Pr. a 10	**5.19**	**7.26**	**5.52**	**5.19**	**4.84**	**3.45**	**2.42**	**2.42**	**1.01**	**3.00**	**2.34**	**1.67**	**1.01**	**0.34**	**0.34**	**0.34**
Pr. a 15	**2.92**	**6.60**	**4.64**	**4.39**	**3.90**	**3.41**	**2.19**	**1.45**	**2.34**	**0.47**	**1.41**	**1.41**	**1.18**	**1.41**	**0.71**	**0.24**
Pr. a 20	**5.99**	**8.58**	**8.98**	**8.78**	**7.97**	**5.99**	**5.38**	**3.58**	**1.31**	**2.60**	**1.31**	**0.94**	**1.12**	**1.31**	**0.75**	**1.12**
Pr. a 30	**4.26**	**8.52**	**9.25**	**8.22**	**7.63**	**7.04**	**5.44**	**3.81**	**3.15**	**3.02**	**2.60**	**2.19**	**1.92**	**2.19**	**1.79**	**1.24**
Pr. a 100	-1.39	**3.36**	**5.33**	**6.06**	**6.12**	**5.67**	**5.55**	**3.36**	-2.74	**1.42**	**2.28**	**2.22**	**2.28**	**2.02**	**1.37**	**0.54**
Pr. a 200	-3.54	**2.15**	**3.85**	**4.61**	**4.39**	**3.85**	**3.76**	**2.42**	-0.26	**1.73**	**2.25**	**1.86**	**1.34**	**0.86**	**0.78**	**0.35**
Pr. a 5	**3.95**	**4.60**	**5.92**	**7.23**	**7.88**	**7.23**	**5.92**	**4.60**	**8.12**	**10.62**	**9.37**	**6.24**	**4.99**	**3.74**	**3.12**	**1.24**
Pr. a 10	**5.52**	**7.61**	**6.91**	**4.84**	**4.84**	**4.14**	**3.45**	**2.42**	**3.00**	**3.33**	**1.67**	**2.34**	**2.34**	**1.33**	**0.67**	**0.34**
Pr. a 15	**3.66**	**6.34**	**4.64**	**4.64**	**4.88**	**3.90**	**1.70**	**1.21**	**3.28**	**2.34**	**2.59**	**1.87**	**1.41**	**0.94**	**0.94**	**0.71**
Pr. a 20	**4.79**	**7.58**	**8.98**	**8.17**	**7.58**	**6.37**	**6.98**	**3.78**	**2.41**	**2.96**	**2.77**	**1.85**	**1.48**	**2.41**	**1.12**	**1.48**
Pr. a 30	**2.50**	**7.79**	**8.52**	**8.08**	**7.35**	**6.90**	**5.44**	**4.55**	**4.11**	**4.24**	**3.70**	**3.15**	**2.60**	**1.92**	**2.19**	**1.38**
Pr. a 100	-1.52	**3.09**	**5.27**	**6.33**	**6.79**	**6.00**	**5.67**	**3.64**	-1.88	**1.91**	**2.91**	**3.28**	**3.33**	**2.54**	**1.97**	**0.60**
Pr. a 200	-3.85	**2.42**	**4.57**	**5.15**	**4.83**	**4.43**	**3.94**	**2.91**	-0.35	**2.46**	**2.81**	**2.68**	**2.16**	**1.17**	**0.86**	**0.35**

Tabla F.14: Puesta a punto de los parámetros para la selección automática de dependencias mediante realimentación. Resultados para el Corpus CLEF 2001-02·A (número de documentos n'_1=10; de arriba a abajo, número de términos $t' \in \{5, 10, 15, 20, 30, 40, 50\}$)

	cortas								*largas*							
ω	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*
Pr. a 5	**2.63**	**3.28**	**1.32**	**1.97**	**1.97**	**1.97**	**2.63**	**1.32**	-0.63	**0.62**	**0.62**	**0.62**	-0.63	-0.63	-0.63	0.00
Pr. a 10	**2.42**	**3.45**	**1.72**	**1.38**	**1.38**	**1.72**	**1.03**	**0.68**	**1.33**	**0.34**	**1.01**	**1.01**	**0.67**	**0.67**	**0.34**	**0.34**
Pr. a 15	**2.92**	**1.45**	**0.98**	**0.98**	**0.73**	**0.49**	**0.24**	**0.24**	0.00	**1.18**	**1.41**	**1.18**	**0.71**	**0.71**	0.00	**0.24**
Pr. a 20	**3.98**	**2.79**	**2.19**	**2.39**	**1.98**	**1.98**	**1.40**	**0.99**	0.00	0.00	**0.37**	**0.19**	**0.19**	**0.56**	**0.19**	0.00
Pr. a 30	**3.23**	**2.19**	**1.91**	**1.60**	**1.32**	**0.59**	**0.73**	**0.43**	**1.09**	**1.38**	**0.96**	**0.96**	**0.96**	**0.70**	**0.55**	**0.70**
Pr. a 100	**0.61**	**1.12**	**1.45**	**1.00**	**1.12**	**1.12**	**1.18**	**0.91**	**0.85**	**0.85**	**0.60**	**0.60**	**0.54**	**0.43**	**0.23**	**0.11**
Pr. a 200	**1.70**	**2.19**	**1.88**	**1.79**	**1.61**	**1.34**	**1.30**	**0.54**	**1.94**	**1.64**	**1.38**	**1.12**	**0.99**	**0.60**	**0.48**	**0.30**
Pr. a 5	**1.32**	**1.32**	**1.97**	**1.97**	**1.97**	**1.97**	**1.97**	**1.32**	-0.63	**0.62**	**1.24**	**1.24**	**1.24**	**0.62**	**0.62**	**0.62**
Pr. a 10	**2.42**	**4.14**	**2.42**	**2.07**	**1.72**	**2.07**	**1.72**	**1.03**	**1.67**	**0.34**	**1.67**	**1.67**	**1.01**	**2.00**	**1.01**	**0.67**
Pr. a 15	**3.17**	**1.94**	**1.94**	**1.45**	**1.45**	**0.98**	**0.49**	**0.24**	**0.47**	**1.65**	**1.18**	**0.94**	**0.47**	**0.47**	-0.23	0.00
Pr. a 20	**3.98**	**2.99**	**1.98**	**2.19**	**1.78**	**1.98**	**1.40**	**1.40**	**2.04**	**1.48**	**1.67**	**1.12**	**1.31**	**0.75**	**0.37**	**0.37**
Pr. a 30	**3.81**	**3.23**	**2.92**	**2.78**	**2.50**	**1.01**	**1.18**	**0.43**	**2.73**	**2.34**	**2.19**	**1.79**	**1.51**	**0.96**	**0.55**	**0.70**
Pr. a 100	**1.24**	**1.85**	**2.36**	**1.79**	**1.97**	**1.39**	**1.39**	**1.06**	**0.43**	**0.80**	**0.85**	**0.97**	**0.80**	**0.54**	**0.43**	**0.37**
Pr. a 200	**1.34**	**2.60**	**2.19**	**1.92**	**1.75**	**1.52**	**1.43**	**0.76**	**1.68**	**1.56**	**1.30**	**1.12**	**1.12**	**0.82**	**0.56**	**0.30**
Pr. a 5	**0.65**	**1.97**	**2.63**	**2.63**	**1.97**	**2.63**	**2.63**	**1.97**	**0.62**	**1.24**	**1.24**	**1.87**	**1.87**	**0.62**	**0.62**	**0.62**
Pr. a 10	**0.33**	**2.77**	**1.38**	**1.03**	**2.07**	**3.10**	**3.10**	**1.03**	-0.99	-0.32	**0.67**	**0.67**	**0.67**	**1.33**	**0.34**	**0.34**
Pr. a 15	**1.70**	**1.45**	**1.94**	**2.19**	**1.94**	**1.94**	**1.70**	**0.98**	**0.47**	**1.18**	**1.18**	**1.18**	**0.47**	**0.24**	-0.23	**0.24**
Pr. a 20	**2.59**	**2.99**	**2.79**	**3.20**	**2.39**	**1.98**	**1.19**	**1.78**	**0.94**	**1.12**	**1.31**	**1.12**	**1.12**	**0.75**	**0.56**	**0.37**
Pr. a 30	**3.08**	**3.37**	**3.37**	**2.92**	**3.08**	**2.05**	**1.75**	**0.73**	**2.73**	**2.47**	**2.19**	**1.79**	**1.38**	**1.09**	**0.83**	**0.70**
Pr. a 100	-0.45	**2.58**	**3.64**	**3.24**	**3.42**	**2.82**	**2.58**	**1.30**	**0.74**	**1.37**	**1.54**	**1.48**	**1.28**	**0.74**	**0.54**	**0.31**
Pr. a 200	**0.81**	**3.27**	**2.86**	**2.46**	**2.33**	**1.84**	**1.66**	**1.07**	**2.46**	**2.20**	**1.73**	**1.56**	**1.34**	**0.95**	**0.73**	**0.30**
Pr. a 5	**5.27**	**5.92**	**4.60**	**3.95**	**3.95**	**5.27**	**5.27**	**3.95**	**1.87**	**1.24**	**1.24**	**1.87**	**1.87**	**1.24**	**1.24**	**0.62**
Pr. a 10	**0.68**	**2.42**	**2.42**	**2.07**	**2.77**	**3.45**	**3.10**	**1.72**	**1.01**	**0.67**	**1.67**	**1.67**	**1.67**	**1.01**	**1.01**	**1.01**
Pr. a 15	**1.70**	**2.43**	**2.92**	**3.41**	**2.43**	**1.45**	**1.21**	**0.73**	**1.18**	**1.65**	**1.65**	**1.41**	**0.94**	0.00	0.00	**0.47**
Pr. a 20	**3.78**	**4.59**	**3.98**	**3.38**	**2.79**	**3.20**	**3.20**	**3.20**	**0.75**	**1.67**	**2.04**	**1.85**	**1.67**	**0.94**	**0.94**	**0.56**
Pr. a 30	**4.85**	**5.13**	**4.40**	**4.10**	**4.26**	**2.92**	**2.64**	**1.91**	**2.47**	**2.73**	**2.60**	**2.34**	**2.19**	**1.51**	**1.51**	**0.96**
Pr. a 100	**0.12**	**4.67**	**4.82**	**4.33**	**4.09**	**3.03**	**2.82**	**1.79**	**0.54**	**1.60**	**1.91**	**1.54**	**1.42**	**0.68**	**0.43**	**0.31**
Pr. a 200	**1.30**	**4.03**	**3.94**	**3.72**	**3.45**	**2.46**	**2.28**	**1.75**	**2.46**	**2.25**	**2.03**	**1.73**	**1.34**	**0.82**	**0.73**	**0.26**
Pr. a 5	**5.27**	**7.23**	**7.88**	**7.23**	**6.57**	**4.60**	**4.60**	**3.95**	**1.87**	**3.74**	**3.12**	**1.87**	**1.87**	**1.24**	**1.24**	0.00
Pr. a 10	**2.77**	**5.19**	**4.49**	**4.49**	**4.84**	**3.45**	**3.10**	**2.42**	**0.34**	**2.00**	**1.33**	**1.01**	**1.01**	**1.33**	**1.33**	**0.67**
Pr. a 15	**1.45**	**3.41**	**2.92**	**1.94**	**2.43**	**1.45**	**1.70**	**0.98**	-0.47	0.00	**0.71**	**0.71**	**0.24**	0.00	**0.24**	**0.71**
Pr. a 20	**4.19**	**4.98**	**4.59**	**4.59**	**4.19**	**4.19**	**4.39**	**4.39**	**0.19**	**1.85**	**2.23**	**1.85**	**1.67**	**1.31**	**1.31**	**0.75**
Pr. a 30	**5.13**	**5.72**	**5.72**	**5.58**	**6.17**	**4.26**	**3.81**	**2.78**	**4.11**	**3.83**	**3.43**	**2.88**	**2.73**	**2.06**	**1.38**	**0.96**
Pr. a 100	-0.79	**3.70**	**5.27**	**4.61**	**4.94**	**4.42**	**4.09**	**2.52**	-0.51	**2.02**	**2.28**	**1.74**	**1.65**	**0.97**	**1.05**	**0.54**
Pr. a 200	-1.03	**3.49**	**4.07**	**4.30**	**3.98**	**3.18**	**3.04**	**2.15**	**0.78**	**1.73**	**1.90**	**1.86**	**1.38**	**0.82**	**0.73**	**0.26**
Pr. a 5	**7.23**	**9.20**	**11.18**	**6.57**	**7.88**	**5.27**	**5.92**	**4.60**	**3.12**	**4.37**	**3.74**	**1.87**	**1.87**	**0.62**	**0.62**	0.00
Pr. a 10	**3.10**	**4.84**	**5.19**	**5.87**	**5.87**	**3.10**	**2.42**	**1.72**	**3.00**	**2.66**	**2.34**	**2.00**	**2.34**	**1.67**	**1.67**	**1.33**
Pr. a 15	0.00	**3.17**	**3.41**	**3.17**	**3.41**	**2.19**	**2.19**	**1.70**	**0.47**	**1.18**	**2.12**	**1.65**	**1.18**	**0.47**	**0.71**	**0.71**
Pr. a 20	**1.60**	**4.79**	**5.18**	**5.58**	**6.19**	**5.99**	**5.18**	**4.19**	-0.54	**1.48**	**1.85**	**1.85**	**1.31**	**2.04**	**1.67**	**1.12**
Pr. a 30	**3.81**	**7.35**	**8.22**	**7.63**	**6.76**	**6.31**	**6.17**	**4.10**	**2.73**	**3.96**	**3.83**	**3.15**	**2.88**	**2.34**	**1.79**	**1.09**
Pr. a 100	-0.27	**3.76**	**5.55**	**5.94**	**6.00**	**5.67**	**5.39**	**3.36**	-1.94	**1.11**	**2.39**	**2.22**	**2.28**	**1.23**	**0.97**	**0.48**
Pr. a 200	-2.06	**2.73**	**3.49**	**3.63**	**4.07**	**3.94**	**3.54**	**2.15**	**0.22**	**2.03**	**2.25**	**2.42**	**1.81**	**0.86**	**0.73**	**0.22**
Pr. a 5	**7.88**	**9.20**	**9.87**	**6.57**	**7.23**	**3.95**	**4.60**	**5.27**	**6.24**	**8.12**	**6.24**	**4.37**	**3.74**	**1.87**	**1.24**	**0.62**
Pr. a 10	**3.45**	**5.52**	**5.87**	**5.52**	**6.22**	**3.80**	**2.77**	**2.07**	**4.66**	**5.32**	**3.33**	**2.66**	**2.34**	**2.66**	**3.00**	**1.67**
Pr. a 15	-0.25	**3.17**	**3.41**	**3.41**	**3.66**	**1.94**	**1.70**	**1.94**	**2.59**	**2.34**	**3.52**	**2.81**	**2.59**	**1.65**	**1.65**	**0.71**
Pr. a 20	**0.99**	**5.99**	**6.19**	**6.37**	**6.98**	**5.99**	**5.58**	**4.59**	**0.56**	**2.04**	**1.67**	**0.94**	**2.04**	**2.04**	**1.85**	**1.67**
Pr. a 30	**2.92**	**6.17**	**6.62**	**6.76**	**6.17**	**6.17**	**5.72**	**4.55**	**2.34**	**3.96**	**4.37**	**3.83**	**4.11**	**3.15**	**2.06**	**1.64**
Pr. a 100	-0.45	**3.03**	**4.42**	**5.15**	**5.39**	**5.73**	**5.33**	**3.64**	-1.68	**0.80**	**1.37**	**2.54**	**2.02**	**1.60**	**1.54**	**0.54**
Pr. a 200	-2.82	**2.01**	**3.49**	**3.98**	**4.12**	**3.85**	**3.58**	**2.24**	**0.26**	**3.20**	**2.81**	**2.63**	**2.16**	**0.91**	**0.86**	**0.30**

Tabla F.15: Puesta a punto de los parámetros para la selección automática de dependencias mediante realimentación. Resultados para el Corpus CLEF 2001-02·A (número de documentos n'_1=15; de arriba a abajo, número de términos $t' \in \{5, 10, 15, 20, 30, 40, 50\}$)

	cortas								*largas*							
ω	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*
Pr. a 5	**2.63**	**3.28**	**1.32**	**1.97**	**1.97**	**1.97**	**2.63**	**1.32**	-0.63	**0.62**	**0.62**	**0.62**	-0.63	-0.63	-0.63	0.00
Pr. a 10	**2.42**	**3.45**	**1.72**	**1.38**	**1.38**	**1.72**	**1.03**	**0.68**	**1.33**	**0.34**	**1.01**	**1.01**	**0.67**	**0.67**	**0.34**	**0.34**
Pr. a 15	**2.92**	**1.45**	**0.98**	**0.98**	**0.73**	**0.49**	**0.24**	**0.24**	0.00	**1.18**	**1.41**	**1.18**	**0.71**	**0.71**	0.00	**0.24**
Pr. a 20	**3.98**	**2.79**	**2.19**	**2.39**	**1.98**	**1.98**	**1.40**	**0.99**	0.00	0.00	**0.37**	**0.19**	**0.19**	**0.56**	**0.19**	0.00
Pr. a 30	**3.23**	**2.19**	**1.91**	**1.60**	**1.32**	**0.59**	**0.73**	**0.43**	**1.09**	**1.38**	**0.96**	**0.96**	**0.96**	**0.70**	**0.55**	**0.70**
Pr. a 100	**0.61**	**1.12**	**1.45**	**1.00**	**1.12**	**1.12**	**1.18**	**0.91**	**0.85**	**0.85**	**0.60**	**0.60**	**0.54**	**0.43**	**0.23**	**0.11**
Pr. a 200	**1.70**	**2.19**	**1.88**	**1.79**	**1.61**	**1.34**	**1.30**	**0.54**	**1.94**	**1.64**	**1.38**	**1.12**	**0.99**	**0.60**	**0.48**	**0.30**
Pr. a 5	**1.32**	**1.32**	**1.97**	**1.97**	**1.97**	**1.97**	**1.97**	**1.32**	-0.63	**0.62**	**1.24**	**1.24**	**1.24**	**0.62**	**0.62**	**0.62**
Pr. a 10	**2.42**	**4.14**	**2.42**	**2.07**	**1.72**	**2.07**	**1.72**	**1.03**	**1.67**	**0.34**	**1.67**	**1.67**	**1.01**	**2.00**	**1.01**	**0.67**
Pr. a 15	**3.17**	**1.94**	**1.94**	**1.45**	**1.45**	**0.98**	**0.49**	**0.24**	**0.47**	**1.65**	**1.18**	**0.94**	**0.47**	**0.47**	-0.23	0.00
Pr. a 20	**3.98**	**2.99**	**1.98**	**2.19**	**1.78**	**1.98**	**1.40**	**1.40**	**2.04**	**1.48**	**1.67**	**1.12**	**1.31**	**0.75**	**0.37**	**0.37**
Pr. a 30	**3.81**	**3.23**	**2.92**	**2.78**	**2.50**	**1.01**	**1.18**	**0.43**	**2.73**	**2.34**	**2.19**	**1.79**	**1.51**	**0.96**	**0.55**	**0.70**
Pr. a 100	**1.24**	**1.85**	**2.36**	**1.79**	**1.97**	**1.39**	**1.39**	**1.06**	**0.43**	**0.80**	**0.85**	**0.97**	**0.80**	**0.54**	**0.43**	**0.37**
Pr. a 200	**1.34**	**2.60**	**2.19**	**1.92**	**1.75**	**1.52**	**1.43**	**0.76**	**1.68**	**1.56**	**1.30**	**1.12**	**1.12**	**0.82**	**0.56**	**0.30**
Pr. a 5	**0.65**	**1.97**	**2.63**	**2.63**	**1.97**	**2.63**	**2.63**	**1.97**	**0.62**	**1.24**	**1.24**	**1.87**	**1.87**	**0.62**	**0.62**	**0.62**
Pr. a 10	**0.33**	**2.77**	**1.38**	**1.03**	**2.07**	**3.10**	**3.10**	**1.03**	-0.99	-0.32	**0.67**	**0.67**	**0.67**	**1.33**	**0.34**	**0.34**
Pr. a 15	**1.70**	**1.45**	**1.94**	**2.19**	**1.94**	**1.94**	**1.70**	**0.98**	**0.47**	**1.18**	**1.18**	**1.18**	**0.47**	**0.24**	-0.23	**0.24**
Pr. a 20	**2.59**	**2.99**	**2.79**	**3.20**	**2.39**	**1.98**	**1.19**	**1.78**	**0.94**	**1.12**	**1.31**	**1.12**	**1.12**	**0.75**	**0.56**	**0.37**
Pr. a 30	**3.08**	**3.37**	**3.37**	**2.92**	**3.08**	**2.05**	**1.75**	**0.73**	**2.73**	**2.47**	**2.19**	**1.79**	**1.38**	**1.09**	**0.83**	**0.70**
Pr. a 100	-0.45	**2.58**	**3.64**	**3.24**	**3.42**	**2.82**	**2.58**	**1.30**	**0.74**	**1.37**	**1.54**	**1.48**	**1.28**	**0.74**	**0.54**	**0.31**
Pr. a 200	**0.81**	**3.27**	**2.86**	**2.46**	**2.33**	**1.84**	**1.66**	**1.07**	**2.46**	**2.20**	**1.73**	**1.56**	**1.34**	**0.95**	**0.73**	**0.30**
Pr. a 5	**5.27**	**5.92**	**4.60**	**3.95**	**3.95**	**5.27**	**5.27**	**3.95**	**1.87**	**1.24**	**1.24**	**1.87**	**1.87**	**1.24**	**1.24**	**0.62**
Pr. a 10	**0.68**	**2.42**	**2.42**	**2.07**	**2.77**	**3.45**	**3.10**	**1.72**	**1.01**	**0.67**	**1.67**	**1.67**	**1.67**	**1.01**	**1.01**	**1.01**
Pr. a 15	**1.70**	**2.43**	**2.92**	**3.41**	**2.43**	**1.45**	**1.21**	**0.73**	**1.18**	**1.65**	**1.65**	**1.41**	**0.94**	0.00	0.00	**0.47**
Pr. a 20	**3.78**	**4.59**	**3.98**	**3.38**	**2.79**	**3.20**	**3.20**	**3.20**	**0.75**	**1.67**	**2.04**	**1.85**	**1.67**	**0.94**	**0.94**	**0.56**
Pr. a 30	**4.85**	**5.13**	**4.40**	**4.10**	**4.26**	**2.92**	**2.64**	**1.91**	**2.47**	**2.73**	**2.60**	**2.34**	**2.19**	**1.51**	**1.51**	**0.96**
Pr. a 100	**0.12**	**4.67**	**4.82**	**4.33**	**4.09**	**3.03**	**2.82**	**1.79**	**0.54**	**1.60**	**1.91**	**1.54**	**1.42**	**0.68**	**0.43**	**0.31**
Pr. a 200	**1.30**	**4.03**	**3.94**	**3.72**	**3.45**	**2.46**	**2.28**	**1.75**	**2.46**	**2.25**	**2.03**	**1.73**	**1.34**	**0.82**	**0.73**	**0.26**
Pr. a 5	**5.27**	**7.23**	**7.88**	**7.23**	**6.57**	**4.60**	**4.60**	**3.95**	**1.87**	**3.74**	**3.12**	**1.87**	**1.87**	**1.24**	**1.24**	0.00
Pr. a 10	**2.77**	**5.19**	**4.49**	**4.49**	**4.84**	**3.45**	**3.10**	**2.42**	**0.34**	**2.00**	**1.33**	**1.01**	**1.01**	**1.33**	**1.33**	**0.67**
Pr. a 15	**1.45**	**3.41**	**2.92**	**1.94**	**2.43**	**1.45**	**1.70**	**0.98**	-0.47	0.00	**0.71**	**0.71**	**0.24**	0.00	**0.24**	**0.71**
Pr. a 20	**4.19**	**4.98**	**4.59**	**4.59**	**4.19**	**4.19**	**4.39**	**4.39**	**0.19**	**1.85**	**2.23**	**1.85**	**1.67**	**1.31**	**1.31**	**0.75**
Pr. a 30	**5.13**	**5.72**	**5.72**	**5.58**	**6.17**	**4.26**	**3.81**	**2.78**	**4.11**	**3.83**	**3.43**	**2.88**	**2.73**	**2.06**	**1.38**	**0.96**
Pr. a 100	-0.79	**3.70**	**5.27**	**4.61**	**4.94**	**4.42**	**4.09**	**2.52**	-0.51	**2.02**	**2.28**	**1.74**	**1.65**	**0.97**	**1.05**	**0.54**
Pr. a 200	-1.03	**3.49**	**4.07**	**4.30**	**3.98**	**3.18**	**3.04**	**2.15**	**0.78**	**1.73**	**1.90**	**1.86**	**1.38**	**0.82**	**0.73**	**0.26**
Pr. a 5	**7.23**	**9.20**	**11.18**	**6.57**	**7.88**	**5.27**	**5.92**	**4.60**	**3.12**	**4.37**	**3.74**	**1.87**	**1.87**	**0.62**	**0.62**	0.00
Pr. a 10	**3.10**	**4.84**	**5.19**	**5.87**	**5.87**	**3.10**	**2.42**	**1.72**	**3.00**	**2.66**	**2.34**	**2.00**	**2.34**	**1.67**	**1.67**	**1.33**
Pr. a 15	0.00	**3.17**	**3.41**	**3.17**	**3.41**	**2.19**	**2.19**	**1.70**	**0.47**	**1.18**	**2.12**	**1.65**	**1.18**	**0.47**	**0.71**	**0.71**
Pr. a 20	**1.60**	**4.79**	**5.18**	**5.58**	**6.19**	**5.99**	**5.18**	**4.19**	-0.54	**1.48**	**1.85**	**1.85**	**1.31**	**2.04**	**1.67**	**1.12**
Pr. a 30	**3.81**	**7.35**	**8.22**	**7.63**	**6.76**	**6.31**	**6.17**	**4.10**	**2.73**	**3.96**	**3.83**	**3.15**	**2.88**	**2.34**	**1.79**	**1.09**
Pr. a 100	-0.27	**3.76**	**5.55**	**5.94**	**6.00**	**5.67**	**5.39**	**3.36**	-1.94	**1.11**	**2.39**	**2.22**	**2.28**	**1.23**	**0.97**	**0.48**
Pr. a 200	-2.06	**2.73**	**3.49**	**3.63**	**4.07**	**3.94**	**3.54**	**2.15**	**0.22**	**2.03**	**2.25**	**2.42**	**1.81**	**0.86**	**0.73**	**0.22**
Pr. a 5	**7.88**	**9.20**	**9.87**	**6.57**	**7.23**	**3.95**	**4.60**	**5.27**	**6.24**	**8.12**	**6.24**	**4.37**	**3.74**	**1.87**	**1.24**	**0.62**
Pr. a 10	**3.45**	**5.52**	**5.87**	**5.52**	**6.22**	**3.80**	**2.77**	**2.07**	**4.66**	**5.32**	**3.33**	**2.66**	**2.34**	**2.66**	**3.00**	**1.67**
Pr. a 15	-0.25	**3.17**	**3.41**	**3.41**	**3.66**	**1.94**	**1.70**	**1.94**	**2.59**	**2.34**	**3.52**	**2.81**	**2.59**	**1.65**	**1.65**	**0.71**
Pr. a 20	**0.99**	**5.99**	**6.19**	**6.37**	**6.98**	**5.99**	**5.58**	**4.59**	**0.56**	**2.04**	**1.67**	**0.94**	**2.04**	**2.04**	**1.85**	**1.67**
Pr. a 30	**2.92**	**6.17**	**6.62**	**6.76**	**6.17**	**6.17**	**5.72**	**4.55**	**2.34**	**3.96**	**4.37**	**3.83**	**4.11**	**3.15**	**2.06**	**1.64**
Pr. a 100	-0.45	**3.03**	**4.42**	**5.15**	**5.39**	**5.73**	**5.33**	**3.64**	-1.68	**0.80**	**1.37**	**2.54**	**2.02**	**1.60**	**1.54**	**0.54**
Pr. a 200	-2.82	**2.01**	**3.49**	**3.98**	**4.12**	**3.85**	**3.58**	**2.24**	**0.26**	**3.20**	**2.81**	**2.63**	**2.16**	**0.91**	**0.86**	**0.30**

Tabla F.16: Puesta a punto de los parámetros para la selección automática de dependencias mediante realimentación. Resultados para el Corpus CLEF 2001-02·A (número de documentos n'_1=20; de arriba a abajo, número de términos $t' \in \{5, 10, 15, 20, 30, 40, 50\}$)

	cortas								*largas*							
ω	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*
Pr. a 5	**2.63**	**1.97**	**1.32**	**0.65**	**1.32**	**1.32**	**1.32**	**0.65**	**1.24**	0.00	**0.62**	**0.62**	**0.62**	0.00	0.00	0.00
Pr. a 10	**1.72**	**3.10**	**3.10**	**3.10**	**3.10**	**2.07**	**1.72**	**0.33**	-0.32	-1.31	0.00	-0.32	-0.99	0.00	0.00	0.00
Pr. a 15	**0.73**	**1.94**	**1.45**	**0.98**	**1.21**	**1.94**	**1.45**	**0.73**	0.00	**0.71**	**0.94**	**0.71**	**0.47**	0.00	0.00	-0.23
Pr. a 20	**2.59**	**2.99**	**2.39**	**2.39**	**2.59**	**2.59**	**2.39**	**1.78**	**0.19**	**0.56**	**0.56**	0.00	0.00	**0.37**	**0.19**	0.00
Pr. a 30	**3.37**	**2.78**	**2.78**	**2.33**	**2.05**	**1.46**	**1.18**	**0.59**	**0.83**	**0.96**	**0.55**	**0.70**	**0.70**	0.00	-0.13	**0.15**
Pr. a 100	**0.67**	**1.91**	**2.30**	**2.52**	**2.30**	**1.64**	**1.64**	**1.12**	**0.06**	**0.97**	**1.28**	**1.17**	**0.85**	**0.60**	**0.43**	**0.23**
Pr. a 200	**0.58**	**2.33**	**1.52**	**1.52**	**1.30**	**1.30**	**1.34**	**0.72**	**1.30**	**1.21**	**0.95**	**0.86**	**0.86**	**0.56**	**0.56**	**0.30**
Pr. a 5	**1.32**	**1.32**	**1.97**	**1.97**	**2.63**	**2.63**	**2.63**	**1.97**	**0.62**	0.00	**0.62**	**0.62**	**0.62**	0.00	0.00	0.00
Pr. a 10	-1.73	-1.05	0.00	-0.35	**0.33**	**1.03**	**1.03**	**0.68**	-0.66	-0.99	0.00	0.00	-0.66	**0.67**	**0.34**	0.00
Pr. a 15	-2.21	-0.74	**0.49**	**0.49**	**0.24**	**0.73**	**0.73**	0.00	0.00	**1.18**	**1.41**	**1.18**	**0.71**	-0.23	-0.23	0.00
Pr. a 20	**0.20**	**0.39**	**2.59**	**2.79**	**3.20**	**3.78**	**3.58**	**2.79**	**1.48**	**1.67**	**1.85**	**1.31**	**0.94**	**1.12**	**0.75**	0.00
Pr. a 30	**2.05**	**3.08**	**3.67**	**3.51**	**3.51**	**2.78**	**2.64**	**1.32**	**0.41**	**0.96**	**0.70**	**0.55**	**0.41**	**0.15**	0.00	**0.15**
Pr. a 100	**0.45**	**2.82**	**3.03**	**3.09**	**2.76**	**2.03**	**2.03**	**1.18**	**0.43**	**1.65**	**1.48**	**1.28**	**0.97**	**0.54**	**0.43**	**0.48**
Pr. a 200	**0.04**	**2.64**	**2.55**	**2.33**	**2.15**	**1.88**	**1.92**	**1.03**	**1.68**	**1.60**	**1.38**	**1.12**	**1.04**	**0.65**	**0.65**	**0.35**
Pr. a 5	**1.32**	**1.97**	**3.28**	**1.97**	**2.63**	**2.63**	**2.63**	**1.97**	**3.12**	**1.24**	**1.24**	**0.62**	**0.62**	0.00	0.00	0.00
Pr. a 10	-2.44	-1.05	-0.35	-0.70	0.00	**0.68**	**0.68**	**0.68**	-1.65	-1.31	**0.67**	**1.01**	**0.34**	**1.33**	**0.67**	-0.32
Pr. a 15	-3.91	-1.47	-0.25	**0.49**	**0.73**	**0.49**	**0.73**	**0.49**	-1.41	**0.71**	**0.94**	**1.18**	**0.71**	**0.47**	**0.24**	0.00
Pr. a 20	-1.21	**0.20**	**2.19**	**3.38**	**3.58**	**3.38**	**3.38**	**2.59**	-0.73	**0.75**	**1.31**	**0.75**	**0.94**	**1.12**	**0.37**	0.00
Pr. a 30	-1.34	**1.91**	**3.51**	**3.37**	**3.23**	**2.64**	**2.64**	**1.18**	**0.28**	**2.34**	**2.19**	**1.92**	**1.64**	**1.64**	**1.64**	**0.96**
Pr. a 100	-2.36	**1.06**	**2.18**	**2.70**	**2.36**	**2.24**	**2.03**	**1.24**	-1.11	**0.68**	**1.17**	**1.28**	**0.91**	**0.68**	**0.48**	**0.43**
Pr. a 200	-1.97	**1.84**	**2.51**	**2.42**	**2.19**	**1.92**	**2.01**	**1.03**	**1.51**	**2.03**	**1.56**	**1.43**	**1.30**	**0.73**	**0.48**	**0.26**
Pr. a 5	**1.32**	**1.97**	**1.97**	0.00	**1.32**	**1.97**	**2.63**	**1.97**	**2.49**	**2.49**	**2.49**	**1.24**	**1.24**	0.00	0.00	0.00
Pr. a 10	-3.47	-3.12	-0.70	**0.33**	**0.33**	**0.68**	**1.38**	**1.03**	-1.99	-0.99	**0.67**	**1.33**	**1.33**	**2.34**	**1.33**	0.00
Pr. a 15	-5.15	-2.70	-0.74	**0.24**	**0.73**	**0.73**	**0.98**	**0.49**	**0.71**	**1.87**	**1.87**	**2.34**	**1.41**	**0.94**	**0.71**	**0.24**
Pr. a 20	-1.40	-0.40	**0.99**	**2.99**	**2.99**	**3.78**	**3.78**	**3.58**	**0.75**	**1.48**	**1.85**	**1.31**	**1.85**	**1.48**	**0.75**	**0.75**
Pr. a 30	-1.77	**1.60**	**4.10**	**3.23**	**3.51**	**2.64**	**2.64**	**1.60**	**1.51**	**3.28**	**2.60**	**2.19**	**1.79**	**1.92**	**1.64**	**0.83**
Pr. a 100	-4.48	**0.45**	**2.30**	**2.70**	**2.30**	**2.42**	**1.91**	**1.24**	-0.88	**0.85**	**1.28**	**1.54**	**1.05**	**0.54**	**0.60**	**0.54**
Pr. a 200	-2.73	**1.66**	**2.33**	**2.37**	**2.33**	**1.88**	**2.01**	**1.16**	**1.38**	**2.12**	**1.81**	**1.73**	**1.21**	**0.78**	**0.48**	**0.22**
Pr. a 5	0.00	**2.63**	**2.63**	**3.28**	**3.95**	**2.63**	**3.28**	**3.28**	**3.74**	**3.12**	**1.87**	0.00	0.00	**1.24**	**1.24**	0.00
Pr. a 10	-2.44	**0.68**	**2.77**	**2.77**	**2.07**	**1.03**	**1.72**	**1.03**	**0.67**	**2.00**	**1.33**	**1.33**	**2.00**	**2.34**	**1.67**	0.00
Pr. a 15	-4.17	-0.25	**1.21**	**0.98**	**1.45**	**0.49**	**1.21**	**1.45**	-1.16	**1.18**	**1.65**	**1.87**	**1.18**	**0.24**	**0.71**	**0.94**
Pr. a 20	-2.20	**1.60**	**3.20**	**4.19**	**4.19**	**4.79**	**4.59**	**2.79**	-0.36	**0.19**	**1.31**	**0.75**	**1.12**	**2.23**	**1.48**	**0.94**
Pr. a 30	-4.57	**2.19**	**5.13**	**4.85**	**5.28**	**4.85**	**4.26**	**3.96**	**0.96**	**2.88**	**2.60**	**2.19**	**2.06**	**2.19**	**2.06**	**1.51**
Pr. a 100	-7.85	-0.73	**2.30**	**3.70**	**3.48**	**3.94**	**3.76**	**2.64**	-2.48	**2.17**	**2.48**	**1.97**	**1.74**	**1.23**	**0.97**	**0.43**
Pr. a 200	-5.95	**0.40**	**1.97**	**3.04**	**3.18**	**2.60**	**2.60**	**1.66**	**0.65**	**2.55**	**2.63**	**2.07**	**1.68**	**0.99**	**0.65**	**0.39**
Pr. a 5	**1.32**	**2.63**	**2.63**	**3.95**	**4.60**	**4.60**	**4.60**	**3.28**	**3.74**	**3.74**	**3.12**	**1.87**	**1.24**	**0.62**	**1.87**	**0.62**
Pr. a 10	-3.47	**0.33**	**2.77**	**2.42**	**1.03**	**1.72**	**2.42**	**1.03**	**1.67**	**2.34**	**2.66**	**2.66**	**3.33**	**2.34**	**2.00**	**0.67**
Pr. a 15	-4.40	-0.98	**0.24**	**0.98**	**0.98**	**1.45**	**1.45**	**0.98**	**1.87**	**1.87**	**1.41**	**1.18**	**1.65**	**1.41**	**1.65**	**1.18**
Pr. a 20	-2.61	**1.19**	**2.59**	**3.38**	**3.58**	**4.98**	**4.59**	**3.78**	**1.31**	**1.48**	**2.41**	**1.31**	**2.04**	**3.15**	**2.23**	**1.31**
Pr. a 30	-4.42	**0.87**	**3.67**	**4.40**	**6.31**	**5.28**	**5.44**	**4.26**	**2.19**	**2.73**	**3.28**	**3.56**	**3.43**	**2.73**	**2.47**	**2.06**
Pr. a 100	-8.09	-1.64	**1.58**	**2.97**	**3.42**	**3.70**	**3.82**	**3.09**	-3.16	**1.65**	**2.71**	**2.85**	**2.59**	**1.97**	**1.80**	**0.48**
Pr. a 200	-7.25	-0.09	**1.79**	**2.78**	**3.13**	**2.91**	**2.73**	**2.01**	-0.48	**2.51**	**3.02**	**2.76**	**2.51**	**1.34**	**1.12**	**0.60**
Pr. a 5	**1.32**	**2.63**	**1.32**	**2.63**	**4.60**	**4.60**	**3.95**	**2.63**	**4.99**	**4.99**	**4.37**	**4.37**	**3.12**	**1.87**	**1.87**	**0.62**
Pr. a 10	-3.47	**0.33**	**2.42**	**2.07**	**1.72**	**1.03**	**2.07**	**1.03**	**2.34**	**3.67**	**3.99**	**3.33**	**3.99**	**3.33**	**2.66**	**0.34**
Pr. a 15	-4.66	-1.96	**0.49**	**0.73**	**0.98**	**1.21**	**1.70**	**0.73**	**0.71**	**2.59**	**1.65**	**1.41**	**1.87**	**1.65**	**1.87**	**1.41**
Pr. a 20	-2.20	**1.19**	**3.20**	**4.19**	**4.39**	**5.58**	**5.18**	**3.78**	**1.48**	**2.04**	**2.04**	**1.67**	**2.41**	**2.60**	**2.23**	**1.48**
Pr. a 30	-3.39	**2.78**	**4.69**	**6.03**	**7.20**	**6.45**	**6.17**	**4.99**	**2.19**	**2.88**	**3.43**	**3.43**	**3.15**	**2.88**	**2.47**	**2.06**
Pr. a 100	-7.36	-0.85	**2.24**	**3.82**	**4.42**	**4.48**	**4.82**	**3.64**	-3.53	**1.65**	**2.59**	**2.79**	**2.54**	**2.11**	**2.02**	**0.74**
Pr. a 200	-6.04	**0.58**	**2.24**	**3.94**	**4.61**	**4.07**	**3.49**	**2.55**	-0.43	**2.81**	**3.37**	**3.24**	**2.81**	**1.68**	**1.51**	**0.60**

Tabla F.17: Puesta a punto de los parámetros para la selección automática de dependencias, correspondientes a sintagmas nominales, mediante realimentación. Resultados para el Corpus CLEF 2001-02·A (número de documentos n'_1=5; de arriba a abajo, número de términos $t' \in \{5, 10, 15, 20, 30, 40, 50\}$)

	cortas								*largas*							
ω	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*
Pr. a 5	**2.63**	**2.63**	**0.65**	**1.32**	**1.32**	**1.97**	**1.97**	**1.32**	**1.87**	**1.87**	**1.87**	**1.87**	**0.62**	0.00	0.00	0.00
Pr. a 10	**2.07**	**2.42**	**2.42**	**1.72**	**1.72**	**1.38**	**1.03**	**0.68**	**1.01**	0.00	**0.67**	**0.34**	0.00	**0.34**	0.00	**0.34**
Pr. a 15	**2.68**	**1.94**	**1.70**	**1.21**	**0.49**	**0.73**	**0.24**	**0.24**	**0.94**	**1.18**	**0.71**	**0.24**	**0.24**	0.00	-0.47	-0.23
Pr. a 20	**3.20**	**3.38**	**2.39**	**1.98**	**1.60**	**1.60**	**1.19**	**1.19**	**0.56**	**0.94**	**0.56**	**0.19**	**0.19**	**0.75**	**0.37**	**0.19**
Pr. a 30	**3.08**	**2.19**	**2.05**	**2.05**	**1.32**	**0.73**	**0.59**	**0.43**	**1.64**	**1.24**	**1.24**	**1.09**	**0.96**	**0.55**	**0.70**	**0.70**
Pr. a 100	**1.97**	**2.12**	**2.18**	**1.64**	**1.45**	**1.12**	**1.30**	**0.85**	**0.54**	**0.68**	**0.80**	**0.80**	**0.54**	**0.37**	**0.31**	**0.11**
Pr. a 200	**1.84**	**2.55**	**2.06**	**1.97**	**1.52**	**1.48**	**1.39**	**0.67**	**1.51**	**1.08**	**1.12**	**1.04**	**0.95**	**0.56**	**0.43**	**0.30**
Pr. a 5	**1.32**	**1.97**	**2.63**	**2.63**	**3.95**	**2.63**	**2.63**	**1.32**	**4.99**	**3.74**	**3.12**	**2.49**	**1.24**	**1.24**	**1.24**	**0.62**
Pr. a 10	**2.42**	**2.42**	**2.42**	**1.38**	**1.38**	**1.03**	**1.03**	**0.33**	**1.67**	**0.67**	**0.67**	**0.67**	**0.34**	**0.67**	**0.34**	**0.34**
Pr. a 15	-0.74	**0.98**	**0.98**	**0.24**	**0.73**	**1.21**	**0.98**	**0.24**	**1.87**	**1.65**	**1.65**	**1.18**	**0.47**	0.00	-0.47	0.00
Pr. a 20	**1.40**	**0.99**	**1.98**	**2.79**	**2.79**	**2.79**	**2.19**	**2.19**	**2.96**	**2.04**	**1.85**	**1.48**	**1.12**	**0.94**	**0.37**	**0.19**
Pr. a 30	**3.37**	**3.37**	**3.23**	**3.96**	**3.08**	**2.05**	**1.46**	**0.73**	**3.02**	**2.19**	**2.19**	**1.79**	**1.64**	**1.24**	**0.83**	**0.70**
Pr. a 100	**1.39**	**2.70**	**3.36**	**3.15**	**3.09**	**2.42**	**2.30**	**1.45**	**0.37**	**0.60**	**0.80**	**0.74**	**0.48**	**0.31**	**0.17**	**0.17**
Pr. a 200	**1.16**	**2.95**	**2.55**	**2.37**	**2.06**	**2.19**	**1.97**	**1.03**	**1.94**	**1.47**	**1.17**	**0.99**	**0.91**	**0.48**	**0.43**	**0.30**
Pr. a 5	0.00	**1.97**	**2.63**	**1.97**	**3.28**	**1.97**	**1.97**	**2.63**	**3.12**	**1.87**	**1.24**	**1.24**	**0.62**	**0.62**	**0.62**	**0.62**
Pr. a 10	-1.73	**0.33**	**0.68**	**0.68**	**1.38**	**2.07**	**2.42**	**0.68**	-0.32	-0.66	**0.34**	**1.01**	**0.34**	**0.67**	**0.34**	0.00
Pr. a 15	-2.21	**0.73**	**1.21**	**0.49**	**1.21**	**1.45**	**1.21**	**0.73**	**1.18**	**1.87**	**2.12**	**1.87**	**0.94**	**0.24**	-0.23	**0.24**
Pr. a 20	**0.79**	**2.59**	**3.20**	**3.58**	**3.38**	**2.99**	**2.79**	**2.79**	**2.04**	**2.41**	**2.04**	**1.48**	**1.31**	**0.56**	0.00	**0.19**
Pr. a 30	**4.26**	**4.69**	**4.55**	**4.40**	**3.96**	**2.50**	**1.32**	**0.87**	**2.88**	**2.47**	**2.47**	**1.79**	**1.51**	**1.09**	**0.83**	**0.55**
Pr. a 100	**0.21**	**2.97**	**3.76**	**3.48**	**3.15**	**2.97**	**2.97**	**1.52**	**0.80**	**0.97**	**1.37**	**1.37**	**0.97**	**0.31**	**0.17**	**0.31**
Pr. a 200	**0.18**	**2.28**	**1.84**	**1.92**	**1.70**	**1.84**	**1.97**	**1.21**	**1.99**	**1.34**	**1.17**	**1.17**	**0.99**	**0.52**	**0.52**	**0.22**
Pr. a 5	**0.65**	**1.32**	**3.28**	**3.28**	**4.60**	**2.63**	**1.97**	**1.97**	**5.62**	**3.74**	**3.12**	**2.49**	**1.87**	**1.24**	**1.24**	0.00
Pr. a 10	0.00	0.00	**1.03**	**1.03**	**1.03**	**1.72**	**2.07**	**0.68**	**0.67**	-0.66	**0.34**	**1.33**	**0.34**	**0.67**	**0.34**	0.00
Pr. a 15	-0.98	**1.21**	**1.21**	**0.49**	**0.98**	**0.98**	**0.73**	**0.73**	**0.94**	**2.12**	**2.12**	**2.59**	**1.87**	**0.71**	**0.24**	**0.47**
Pr. a 20	**2.39**	**2.19**	**2.99**	**3.20**	**3.38**	**3.78**	**3.20**	**2.99**	**1.12**	**2.41**	**2.60**	**2.04**	**1.85**	**0.94**	**0.37**	**0.56**
Pr. a 30	**3.67**	**5.44**	**4.69**	**4.10**	**3.96**	**2.64**	**1.46**	**0.87**	**2.06**	**2.73**	**2.34**	**1.92**	**1.79**	**1.64**	**0.96**	**0.83**
Pr. a 100	-0.39	**3.15**	**3.82**	**3.48**	**3.36**	**2.97**	**2.97**	**1.91**	**0.17**	**0.85**	**1.17**	**1.17**	**1.05**	**0.48**	**0.17**	**0.17**
Pr. a 200	-0.54	**1.79**	**1.84**	**2.24**	**1.97**	**1.79**	**2.06**	**1.34**	**1.56**	**1.86**	**1.64**	**1.51**	**1.34**	**0.86**	**0.52**	**0.30**
Pr. a 5	**0.65**	**3.28**	**5.27**	**4.60**	**5.27**	**2.63**	**2.63**	**1.97**	**5.62**	**5.62**	**4.99**	**1.87**	**1.87**	**1.24**	**1.24**	-0.63
Pr. a 10	-1.73	**0.68**	**2.07**	**3.10**	**2.77**	**2.42**	**2.42**	**1.72**	-2.32	-0.66	-0.32	**1.33**	**0.67**	**0.34**	**0.34**	0.00
Pr. a 15	-1.96	**0.73**	**0.49**	**0.49**	**0.98**	**0.49**	**0.49**	**0.49**	-1.41	**0.94**	**0.94**	**1.41**	**0.71**	-0.23	0.00	**0.24**
Pr. a 20	**0.59**	**1.98**	**2.99**	**3.20**	**3.20**	**3.78**	**3.78**	**3.38**	-0.54	**0.94**	**1.48**	**1.12**	**1.48**	**1.12**	**0.75**	**0.56**
Pr. a 30	**0.43**	**4.85**	**4.85**	**4.69**	**5.13**	**3.96**	**3.08**	**1.91**	**0.55**	**2.60**	**2.06**	**1.51**	**1.38**	**1.92**	**1.51**	**1.09**
Pr. a 100	-3.09	**1.12**	**2.97**	**3.94**	**4.03**	**3.09**	**3.09**	**1.97**	-2.48	**0.85**	**0.97**	**1.28**	**1.60**	**0.80**	**0.48**	**0.23**
Pr. a 200	-2.78	**1.97**	**2.46**	**2.51**	**2.51**	**2.33**	**2.33**	**1.70**	**0.43**	**1.73**	**2.16**	**1.94**	**1.47**	**0.91**	**0.73**	**0.26**
Pr. a 5	**3.28**	**3.95**	**6.57**	**7.88**	**9.20**	**6.57**	**5.27**	**5.27**	**9.37**	**8.12**	**7.49**	**5.62**	**4.99**	**3.12**	**3.12**	**0.62**
Pr. a 10	**3.80**	**7.26**	**6.91**	**6.22**	**5.87**	**3.80**	**2.42**	**1.72**	-0.32	**2.34**	**2.34**	**1.67**	**2.00**	**0.67**	**0.67**	**0.34**
Pr. a 15	**2.68**	**5.36**	**4.64**	**4.15**	**4.15**	**2.68**	**2.68**	**1.21**	**1.65**	**0.71**	**0.47**	**0.94**	**0.71**	**1.18**	**0.24**	**0.24**
Pr. a 20	**4.19**	**6.78**	**7.18**	**6.78**	**6.19**	**5.38**	**5.38**	**4.19**	**0.94**	**1.48**	**0.94**	**0.37**	**1.12**	**1.12**	**0.56**	**1.12**
Pr. a 30	**3.96**	**7.93**	**8.08**	**6.90**	**6.45**	**5.44**	**4.99**	**3.23**	**2.19**	**3.15**	**2.34**	**1.92**	**1.64**	**1.79**	**1.64**	**0.96**
Pr. a 100	-1.24	**3.64**	**5.15**	**6.06**	**6.12**	**5.00**	**4.82**	**2.97**	-2.74	**0.74**	**2.11**	**2.17**	**2.17**	**2.02**	**1.60**	**0.60**
Pr. a 200	-3.09	**1.97**	**3.63**	**4.30**	**4.43**	**3.85**	**3.67**	**2.01**	-0.30	**2.16**	**2.46**	**1.94**	**1.47**	**1.04**	**0.78**	**0.39**
Pr. a 5	**3.28**	**3.95**	**5.92**	**7.23**	**8.55**	**5.27**	**4.60**	**4.60**	**7.49**	**9.37**	**6.87**	**4.99**	**4.37**	**2.49**	**2.49**	**1.24**
Pr. a 10	**2.42**	**5.19**	**5.87**	**4.49**	**3.45**	**3.45**	**2.42**	**1.38**	-0.32	**3.00**	**3.00**	**3.00**	**2.34**	**2.00**	**2.00**	**0.34**
Pr. a 15	**1.21**	**2.43**	**2.68**	**3.17**	**2.43**	**1.94**	**0.98**	**0.49**	**1.18**	**2.12**	**1.65**	**0.94**	**1.18**	**2.12**	**1.18**	**1.18**
Pr. a 20	**1.19**	**4.39**	**6.37**	**5.99**	**5.78**	**5.18**	**5.38**	**3.58**	**2.96**	**2.60**	**2.04**	**0.19**	**1.12**	**1.67**	**1.48**	**1.48**
Pr. a 30	-0.45	**6.03**	**6.62**	**6.31**	**5.72**	**5.58**	**5.28**	**3.23**	**2.73**	**4.11**	**3.70**	**3.02**	**2.19**	**1.92**	**1.79**	**1.38**
Pr. a 100	-2.03	**2.52**	**4.42**	**5.85**	**6.00**	**5.33**	**5.15**	**3.03**	-1.42	**1.65**	**2.96**	**3.02**	**3.22**	**2.54**	**2.17**	**0.80**
Pr. a 200	-2.64	**2.86**	**4.39**	**4.97**	**4.79**	**4.25**	**3.67**	**2.46**	**0.26**	**3.07**	**3.11**	**2.81**	**2.20**	**1.08**	**0.82**	**0.43**

Tabla F.18: Puesta a punto de los parámetros para la selección automática de dependencias, correspondientes a sintagmas nominales, mediante realimentación. Resultados para el Corpus CLEF 2001-02·A (número de documentos n'_1=10; de arriba a abajo, número de términos $t' \in \{5, 10, 15, 20, 30, 40, 50\}$)

	cortas								*largas*							
ω	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*
Pr. a 5	**3.28**	**3.28**	**1.32**	**1.97**	**1.97**	**1.97**	**1.97**	**1.32**	-0.63	**0.62**	**0.62**	**0.62**	-0.63	-0.63	-0.63	0.00
Pr. a 10	**1.72**	**3.10**	**2.07**	**1.38**	**1.38**	**1.38**	**1.03**	**0.68**	**1.33**	**0.34**	**1.01**	**1.01**	**0.67**	**0.67**	**0.34**	**0.34**
Pr. a 15	**2.19**	**0.98**	**0.73**	**0.73**	**0.49**	**0.73**	**0.24**	**0.24**	0.00	**1.18**	**1.41**	**1.18**	**0.71**	**0.71**	0.00	**0.24**
Pr. a 20	**2.99**	**2.39**	**2.19**	**2.19**	**1.78**	**1.98**	**1.40**	**0.99**	-0.36	0.00	**0.37**	**0.19**	**0.19**	**0.37**	0.00	0.00
Pr. a 30	**2.92**	**1.75**	**1.60**	**1.32**	**0.87**	**0.59**	**0.73**	**0.43**	**1.24**	**1.24**	**0.96**	**0.83**	**0.83**	**0.55**	**0.55**	**0.70**
Pr. a 100	**0.27**	**0.61**	**1.24**	**1.12**	**1.18**	**1.12**	**1.18**	**0.91**	**0.37**	**0.80**	**0.60**	**0.60**	**0.54**	**0.43**	**0.37**	**0.11**
Pr. a 200	**1.52**	**2.01**	**1.79**	**1.66**	**1.34**	**1.25**	**1.12**	**0.49**	**1.90**	**1.51**	**1.30**	**1.08**	**0.95**	**0.60**	**0.48**	**0.30**
Pr. a 5	**0.65**	**1.32**	**2.63**	**1.97**	**1.97**	**1.97**	**2.63**	**1.32**	0.00	**0.62**	**1.24**	**1.24**	**1.24**	**0.62**	**0.62**	**0.62**
Pr. a 10	**1.72**	**3.45**	**2.42**	**1.72**	**1.38**	**1.72**	**1.72**	**1.03**	**1.01**	0.00	**1.33**	**1.67**	**1.01**	**1.67**	**1.01**	**0.67**
Pr. a 15	**2.19**	**1.70**	**1.94**	**1.21**	**1.45**	**0.98**	**0.49**	**0.24**	**0.47**	**1.41**	**1.18**	**0.94**	**0.47**	**0.47**	-0.23	0.00
Pr. a 20	**3.38**	**2.59**	**2.19**	**1.98**	**1.78**	**2.19**	**1.40**	**1.40**	**1.67**	**1.31**	**1.48**	**1.12**	**1.31**	**0.56**	**0.19**	**0.37**
Pr. a 30	**3.96**	**3.23**	**3.23**	**2.64**	**2.19**	**1.01**	**1.18**	**0.43**	**2.60**	**2.06**	**2.19**	**1.51**	**1.38**	**0.83**	**0.55**	**0.70**
Pr. a 100	**0.73**	**1.39**	**2.12**	**1.73**	**2.03**	**1.45**	**1.45**	**1.12**	**0.11**	**0.68**	**0.91**	**0.91**	**0.74**	**0.54**	**0.48**	**0.43**
Pr. a 200	**1.12**	**2.55**	**2.19**	**1.88**	**1.52**	**1.48**	**1.34**	**0.72**	**1.64**	**1.43**	**1.21**	**1.08**	**1.04**	**0.82**	**0.56**	**0.35**
Pr. a 5	-0.67	**1.32**	**2.63**	**1.97**	**2.63**	**2.63**	**3.28**	**1.97**	-0.63	**0.62**	**1.24**	**1.24**	**1.24**	**0.62**	**0.62**	**0.62**
Pr. a 10	-1.05	**1.38**	**1.38**	**0.68**	**1.38**	**2.77**	**3.10**	**1.03**	-1.31	-0.66	**0.34**	**0.34**	**0.34**	**1.01**	**0.34**	**0.34**
Pr. a 15	**0.98**	**1.21**	**1.45**	**1.45**	**1.94**	**2.19**	**1.70**	**0.73**	-0.47	**0.94**	**1.18**	**1.18**	**0.47**	**0.24**	0.00	**0.47**
Pr. a 20	**2.39**	**2.59**	**3.38**	**3.58**	**2.59**	**2.39**	**1.78**	**1.78**	**0.37**	**0.75**	**1.31**	**1.12**	**1.12**	**0.56**	0.00	0.00
Pr. a 30	**3.81**	**3.81**	**3.51**	**2.92**	**2.78**	**2.05**	**1.75**	**0.73**	**2.06**	**2.06**	**1.92**	**1.24**	**1.09**	**0.83**	**0.70**	**0.55**
Pr. a 100	-0.45	**2.18**	**3.42**	**3.15**	**3.48**	**2.82**	**2.64**	**1.30**	**0.11**	**0.68**	**1.05**	**1.17**	**0.97**	**0.60**	**0.54**	**0.37**
Pr. a 200	**0.76**	**3.13**	**2.73**	**2.42**	**2.19**	**1.84**	**1.61**	**1.12**	**1.68**	**1.43**	**1.34**	**1.30**	**1.08**	**0.86**	**0.65**	**0.30**
Pr. a 5	**3.95**	**5.27**	**4.60**	**3.95**	**3.95**	**5.27**	**5.27**	**3.95**	**0.62**	**1.24**	**1.87**	**1.87**	**1.87**	**0.62**	**1.24**	**0.62**
Pr. a 10	-0.70	**1.72**	**2.42**	**1.72**	**2.07**	**2.42**	**2.77**	**1.38**	**0.34**	**0.67**	**1.67**	**1.33**	**1.33**	**0.67**	**1.01**	**1.01**
Pr. a 15	**0.49**	**1.21**	**1.94**	**2.43**	**1.70**	**1.94**	**1.70**	**0.49**	-0.69	**1.41**	**1.18**	**1.18**	**0.71**	0.00	0.00	**0.47**
Pr. a 20	**2.19**	**3.20**	**3.20**	**2.99**	**2.79**	**2.59**	**2.59**	**2.39**	**0.75**	**1.48**	**1.67**	**1.67**	**1.48**	**0.75**	**0.37**	**0.19**
Pr. a 30	**4.10**	**4.55**	**3.67**	**3.37**	**3.08**	**2.05**	**2.19**	**0.87**	**2.06**	**2.47**	**2.19**	**2.06**	**2.06**	**1.38**	**1.24**	**0.96**
Pr. a 100	**0.45**	**3.76**	**4.21**	**4.03**	**3.76**	**3.09**	**2.91**	**1.73**	-0.31	**0.68**	**1.37**	**1.37**	**1.28**	**0.60**	**0.48**	**0.43**
Pr. a 200	**1.16**	**3.45**	**3.18**	**2.86**	**2.64**	**2.33**	**2.06**	**1.43**	**2.07**	**2.03**	**1.86**	**1.47**	**1.08**	**0.86**	**0.65**	**0.26**
Pr. a 5	**5.27**	**7.88**	**7.88**	**6.57**	**7.23**	**4.60**	**4.60**	**3.95**	-0.63	**3.12**	**3.12**	**2.49**	**2.49**	**0.62**	**1.24**	0.00
Pr. a 10	**1.38**	**1.38**	**2.77**	**3.80**	**3.80**	**2.42**	**2.42**	**1.72**	-2.32	**1.01**	**1.33**	**1.01**	**1.33**	**1.33**	**1.67**	**0.67**
Pr. a 15	**0.49**	**1.70**	**1.70**	**1.21**	**2.19**	**1.70**	**1.70**	**0.73**	-1.63	-0.47	0.00	**0.71**	**0.24**	0.00	**0.24**	**0.47**
Pr. a 20	**2.79**	**4.59**	**3.78**	**3.78**	**3.38**	**4.19**	**4.39**	**3.98**	**0.19**	**1.12**	**1.12**	**1.12**	**0.75**	**0.94**	**0.75**	**0.56**
Pr. a 30	**3.81**	**3.96**	**4.10**	**4.55**	**4.40**	**2.92**	**2.50**	**1.46**	**1.51**	**2.47**	**2.34**	**1.92**	**2.06**	**1.24**	**0.96**	**0.55**
Pr. a 100	-1.91	**2.03**	**3.94**	**4.42**	**4.48**	**3.64**	**3.24**	**2.36**	-2.25	**1.05**	**1.54**	**1.28**	**1.37**	**1.05**	**1.05**	**0.74**
Pr. a 200	-1.61	**2.33**	**3.27**	**3.45**	**3.13**	**2.73**	**2.60**	**1.84**	**0.39**	**1.99**	**1.73**	**1.47**	**1.21**	**0.86**	**0.86**	**0.26**
Pr. a 5	**2.63**	**7.23**	**8.55**	**6.57**	**7.88**	**6.57**	**6.57**	**3.95**	**2.49**	**5.62**	**4.99**	**3.12**	**2.49**	**0.62**	**1.24**	0.00
Pr. a 10	-1.38	**3.10**	**4.14**	**5.52**	**4.84**	**3.10**	**2.42**	**2.07**	-0.32	**0.34**	**0.34**	**1.01**	**1.33**	**2.00**	**2.34**	**1.33**
Pr. a 15	-1.47	**1.94**	**2.68**	**2.19**	**2.19**	**1.70**	**1.94**	**0.98**	0.00	**0.94**	**0.94**	**1.18**	**1.41**	**0.94**	**1.18**	**0.71**
Pr. a 20	**0.20**	**3.78**	**4.39**	**5.78**	**4.79**	**5.18**	**5.18**	**4.39**	-0.92	**1.31**	**1.48**	**1.85**	**1.12**	**1.85**	**1.48**	**0.94**
Pr. a 30	**3.37**	**6.45**	**7.20**	**6.62**	**6.17**	**5.44**	**4.85**	**3.67**	**0.55**	**2.88**	**2.88**	**2.60**	**2.34**	**1.64**	**1.24**	**0.96**
Pr. a 100	-1.45	**3.15**	**4.61**	**5.67**	**5.45**	**5.00**	**5.00**	**2.97**	-2.68	**0.60**	**1.60**	**1.80**	**1.85**	**1.23**	**0.91**	**0.80**
Pr. a 200	-1.66	**2.24**	**3.31**	**3.72**	**3.94**	**3.58**	**3.27**	**1.92**	**0.22**	**2.03**	**2.25**	**2.25**	**1.86**	**1.17**	**1.12**	**0.26**
Pr. a 5	**2.63**	**6.57**	**7.88**	**7.88**	**9.20**	**6.57**	**5.92**	**4.60**	**4.99**	**3.74**	**5.62**	**4.99**	**4.37**	**1.87**	**1.87**	**0.62**
Pr. a 10	**0.33**	**4.49**	**5.87**	**5.19**	**5.52**	**4.49**	**3.80**	**2.42**	**1.67**	**3.00**	**2.00**	**1.67**	**1.67**	**1.67**	**2.34**	**2.00**
Pr. a 15	-0.74	**2.19**	**3.41**	**2.68**	**2.68**	**1.45**	**1.70**	**1.70**	**0.71**	**2.12**	**3.05**	**2.59**	**2.12**	**1.65**	**1.87**	**1.18**
Pr. a 20	-0.20	**3.58**	**4.79**	**6.57**	**6.37**	**5.99**	**5.58**	**4.98**	0.00	**1.48**	**1.67**	**1.85**	**2.60**	**2.04**	**1.67**	**1.48**
Pr. a 30	**1.01**	**4.85**	**6.17**	**6.45**	**5.86**	**5.72**	**5.28**	**4.10**	**0.96**	**3.02**	**3.43**	**3.15**	**3.43**	**2.47**	**1.92**	**1.24**
Pr. a 100	-1.64	**2.97**	**3.64**	**4.76**	**4.88**	**4.94**	**4.94**	**3.24**	-2.62	**0.31**	**1.11**	**1.91**	**1.74**	**1.60**	**1.42**	**0.68**
Pr. a 200	-2.64	**1.61**	**3.31**	**3.89**	**3.89**	**3.85**	**3.40**	**2.06**	-0.26	**2.68**	**2.72**	**2.55**	**2.07**	**0.99**	**0.86**	**0.30**

Tabla F.19: Puesta a punto de los parámetros para la selección automática de dependencias, correspondientes a sintagmas nominales, mediante realimentación. Resultados para el Corpus CLEF 2001-02·A (número de documentos n'_1=15; de arriba a abajo, número de términos $t' \in \{5, 10, 15, 20, 30, 40, 50\}$)

	cortas								*largas*							
ω	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*	*1*	*2*	*3*	*4*	*5*	*8*	*10*	*20*
Pr. a 5	**3.28**	**3.28**	**1.32**	**1.97**	**1.97**	**1.97**	**1.97**	**1.32**	-0.63	**0.62**	**0.62**	**0.62**	-0.63	-0.63	-0.63	0.00
Pr. a 10	**1.72**	**3.10**	**2.07**	**1.38**	**1.38**	**1.38**	**1.03**	**0.68**	**1.33**	**0.34**	**1.01**	**1.01**	**0.67**	**0.67**	**0.34**	**0.34**
Pr. a 15	**2.19**	**0.98**	**0.73**	**0.73**	**0.49**	**0.73**	**0.24**	**0.24**	0.00	**1.18**	**1.41**	**1.18**	**0.71**	**0.71**	0.00	**0.24**
Pr. a 20	**2.99**	**2.39**	**2.19**	**2.19**	**1.78**	**1.98**	**1.40**	**0.99**	-0.36	0.00	**0.37**	**0.19**	**0.19**	**0.37**	0.00	0.00
Pr. a 30	**2.92**	**1.75**	**1.60**	**1.32**	**0.87**	**0.59**	**0.73**	**0.43**	**1.24**	**1.24**	**0.96**	**0.83**	**0.83**	**0.55**	**0.55**	**0.70**
Pr. a 100	**0.27**	**0.61**	**1.24**	**1.12**	**1.18**	**1.12**	**1.18**	**0.91**	**0.37**	**0.80**	**0.60**	**0.60**	**0.54**	**0.43**	**0.37**	**0.11**
Pr. a 200	**1.52**	**2.01**	**1.79**	**1.66**	**1.34**	**1.25**	**1.12**	**0.49**	**1.90**	**1.51**	**1.30**	**1.08**	**0.95**	**0.60**	**0.48**	**0.30**
Pr. a 5	**0.65**	**1.32**	**2.63**	**1.97**	**1.97**	**1.97**	**2.63**	**1.32**	0.00	**0.62**	**1.24**	**1.24**	**1.24**	**0.62**	**0.62**	**0.62**
Pr. a 10	**1.72**	**3.45**	**2.42**	**1.72**	**1.38**	**1.72**	**1.72**	**1.03**	**1.01**	0.00	**1.33**	**1.67**	**1.01**	**1.67**	**1.01**	**0.67**
Pr. a 15	**2.19**	**1.70**	**1.94**	**1.21**	**1.45**	**0.98**	**0.49**	**0.24**	**0.47**	**1.41**	**1.18**	**0.94**	**0.47**	**0.47**	-0.23	0.00
Pr. a 20	**3.38**	**2.59**	**2.19**	**1.98**	**1.78**	**2.19**	**1.40**	**1.40**	**1.67**	**1.31**	**1.48**	**1.12**	**1.31**	**0.56**	**0.19**	**0.37**
Pr. a 30	**3.96**	**3.23**	**3.23**	**2.64**	**2.19**	**1.01**	**1.18**	**0.43**	**2.60**	**2.06**	**2.19**	**1.51**	**1.38**	**0.83**	**0.55**	**0.70**
Pr. a 100	**0.73**	**1.39**	**2.12**	**1.73**	**2.03**	**1.45**	**1.45**	**1.12**	**0.11**	**0.68**	**0.91**	**0.91**	**0.74**	**0.54**	**0.48**	**0.43**
Pr. a 200	**1.12**	**2.55**	**2.19**	**1.88**	**1.52**	**1.48**	**1.34**	**0.72**	**1.64**	**1.43**	**1.21**	**1.08**	**1.04**	**0.82**	**0.56**	**0.35**
Pr. a 5	-0.67	**1.32**	**2.63**	**1.97**	**2.63**	**2.63**	**3.28**	**1.97**	-0.63	**0.62**	**1.24**	**1.24**	**1.24**	**0.62**	**0.62**	**0.62**
Pr. a 10	-1.05	**1.38**	**1.38**	**0.68**	**1.38**	**2.77**	**3.10**	**1.03**	-1.31	-0.66	**0.34**	**0.34**	**0.34**	**1.01**	**0.34**	**0.34**
Pr. a 15	**0.98**	**1.21**	**1.45**	**1.45**	**1.94**	**2.19**	**1.70**	**0.73**	-0.47	**0.94**	**1.18**	**1.18**	**0.47**	**0.24**	0.00	**0.47**
Pr. a 20	**2.39**	**2.59**	**3.38**	**3.58**	**2.59**	**2.39**	**1.78**	**1.78**	**0.37**	**0.75**	**1.31**	**1.12**	**1.12**	**0.56**	0.00	0.00
Pr. a 30	**3.81**	**3.81**	**3.51**	**2.92**	**2.78**	**2.05**	**1.75**	**0.73**	**2.06**	**2.06**	**1.92**	**1.24**	**1.09**	**0.83**	**0.70**	**0.55**
Pr. a 100	-0.45	**2.18**	**3.42**	**3.15**	**3.48**	**2.82**	**2.64**	**1.30**	**0.11**	**0.68**	**1.05**	**1.17**	**0.97**	**0.60**	**0.54**	**0.37**
Pr. a 200	**0.76**	**3.13**	**2.73**	**2.42**	**2.19**	**1.84**	**1.61**	**1.12**	**1.68**	**1.43**	**1.34**	**1.30**	**1.08**	**0.86**	**0.65**	**0.30**
Pr. a 5	**3.95**	**5.27**	**4.60**	**3.95**	**3.95**	**5.27**	**5.27**	**3.95**	**0.62**	**1.24**	**1.87**	**1.87**	**1.87**	**0.62**	**1.24**	**0.62**
Pr. a 10	-0.70	**1.72**	**2.42**	**1.72**	**2.07**	**2.42**	**2.77**	**1.38**	**0.34**	**0.67**	**1.67**	**1.33**	**1.33**	**0.67**	**1.01**	**1.01**
Pr. a 15	**0.49**	**1.21**	**1.94**	**2.43**	**1.70**	**1.94**	**1.70**	**0.49**	-0.69	**1.41**	**1.18**	**1.18**	**0.71**	0.00	0.00	**0.47**
Pr. a 20	**2.19**	**3.20**	**3.20**	**2.99**	**2.79**	**2.59**	**2.59**	**2.39**	**0.75**	**1.48**	**1.67**	**1.67**	**1.48**	**0.75**	**0.37**	**0.19**
Pr. a 30	**4.10**	**4.55**	**3.67**	**3.37**	**3.08**	**2.05**	**2.19**	**0.87**	**2.06**	**2.47**	**2.19**	**2.06**	**2.06**	**1.38**	**1.24**	**0.96**
Pr. a 100	**0.45**	**3.76**	**4.21**	**4.03**	**3.76**	**3.09**	**2.91**	**1.73**	-0.31	**0.68**	**1.37**	**1.37**	**1.28**	**0.60**	**0.48**	**0.43**
Pr. a 200	**1.16**	**3.45**	**3.18**	**2.86**	**2.64**	**2.33**	**2.06**	**1.43**	**2.07**	**2.03**	**1.86**	**1.47**	**1.08**	**0.86**	**0.65**	**0.26**
Pr. a 5	**5.27**	**7.88**	**7.88**	**6.57**	**7.23**	**4.60**	**4.60**	**3.95**	-0.63	**3.12**	**3.12**	**2.49**	**2.49**	**0.62**	**1.24**	0.00
Pr. a 10	**1.38**	**1.38**	**2.77**	**3.80**	**3.80**	**2.42**	**2.42**	**1.72**	-2.32	**1.01**	**1.33**	**1.01**	**1.33**	**1.33**	**1.67**	**0.67**
Pr. a 15	**0.49**	**1.70**	**1.70**	**1.21**	**2.19**	**1.70**	**1.70**	**0.73**	-1.63	-0.47	0.00	**0.71**	**0.24**	0.00	**0.24**	**0.47**
Pr. a 20	**2.79**	**4.59**	**3.78**	**3.78**	**3.38**	**4.19**	**4.39**	**3.98**	**0.19**	**1.12**	**1.12**	**1.12**	**0.75**	**0.94**	**0.75**	**0.56**
Pr. a 30	**3.81**	**3.96**	**4.10**	**4.55**	**4.40**	**2.92**	**2.50**	**1.46**	**1.51**	**2.47**	**2.34**	**1.92**	**2.06**	**1.24**	**0.96**	**0.55**
Pr. a 100	-1.91	**2.03**	**3.94**	**4.42**	**4.48**	**3.64**	**3.24**	**2.36**	-2.25	**1.05**	**1.54**	**1.28**	**1.37**	**1.05**	**1.05**	**0.74**
Pr. a 200	-1.61	**2.33**	**3.27**	**3.45**	**3.13**	**2.73**	**2.60**	**1.84**	**0.39**	**1.99**	**1.73**	**1.47**	**1.21**	**0.86**	**0.86**	**0.26**
Pr. a 5	**2.63**	**7.23**	**8.55**	**6.57**	**7.88**	**6.57**	**6.57**	**3.95**	**2.49**	**5.62**	**4.99**	**3.12**	**2.49**	**0.62**	**1.24**	0.00
Pr. a 10	-1.38	**3.10**	**4.14**	**5.52**	**4.84**	**3.10**	**2.42**	**2.07**	-0.32	**0.34**	**0.34**	**1.01**	**1.33**	**2.00**	**2.34**	**1.33**
Pr. a 15	-1.47	**1.94**	**2.68**	**2.19**	**2.19**	**1.70**	**1.94**	**0.98**	0.00	**0.94**	**0.94**	**1.18**	**1.41**	**0.94**	**1.18**	**0.71**
Pr. a 20	**0.20**	**3.78**	**4.39**	**5.78**	**4.79**	**5.18**	**5.18**	**4.39**	-0.92	**1.31**	**1.48**	**1.85**	**1.12**	**1.85**	**1.48**	**0.94**
Pr. a 30	**3.37**	**6.45**	**7.20**	**6.62**	**6.17**	**5.44**	**4.85**	**3.67**	**0.55**	**2.88**	**2.88**	**2.60**	**2.34**	**1.64**	**1.24**	**0.96**
Pr. a 100	-1.45	**3.15**	**4.61**	**5.67**	**5.45**	**5.00**	**5.00**	**2.97**	-2.68	**0.60**	**1.60**	**1.80**	**1.85**	**1.23**	**0.91**	**0.80**
Pr. a 200	-1.66	**2.24**	**3.31**	**3.72**	**3.94**	**3.58**	**3.27**	**1.92**	**0.22**	**2.03**	**2.25**	**2.25**	**1.86**	**1.17**	**1.12**	**0.26**
Pr. a 5	**2.63**	**6.57**	**7.88**	**7.88**	**9.20**	**6.57**	**5.92**	**4.60**	**4.99**	**3.74**	**5.62**	**4.99**	**4.37**	**1.87**	**1.87**	**0.62**
Pr. a 10	**0.33**	**4.49**	**5.87**	**5.19**	**5.52**	**4.49**	**3.80**	**2.42**	**1.67**	**3.00**	**2.00**	**1.67**	**1.67**	**1.67**	**2.34**	**2.00**
Pr. a 15	-0.74	**2.19**	**3.41**	**2.68**	**2.68**	**1.45**	**1.70**	**1.70**	**0.71**	**2.12**	**3.05**	**2.59**	**2.12**	**1.65**	**1.87**	**1.18**
Pr. a 20	-0.20	**3.58**	**4.79**	**6.57**	**6.37**	**5.99**	**5.58**	**4.98**	0.00	**1.48**	**1.67**	**1.85**	**2.60**	**2.04**	**1.67**	**1.48**
Pr. a 30	**1.01**	**4.85**	**6.17**	**6.45**	**5.86**	**5.72**	**5.28**	**4.10**	**0.96**	**3.02**	**3.43**	**3.15**	**3.43**	**2.47**	**1.92**	**1.24**
Pr. a 100	-1.64	**2.97**	**3.64**	**4.76**	**4.88**	**4.94**	**4.94**	**3.24**	-2.62	**0.31**	**1.11**	**1.91**	**1.74**	**1.60**	**1.42**	**0.68**
Pr. a 200	-2.64	**1.61**	**3.31**	**3.89**	**3.89**	**3.85**	**3.40**	**2.06**	-0.26	**2.68**	**2.72**	**2.55**	**2.07**	**0.99**	**0.86**	**0.30**

Tabla F.20: Puesta a punto de los parámetros para la selección automática de dependencias, correspondientes a sintagmas nominales, mediante realimentación. Resultados para el Corpus CLEF 2001-02·A (número de documentos n'_1=20; de arriba a abajo, número de términos $t' \in \{5, 10, 15, 20, 30, 40, 50\}$)

Apéndice G

Puesta a Punto del Modelo Basado en Localidad

Tabla G.1: Reordenación mediante distancias de la lematización con realimentación (consultas cortas: α=0.8, β=0.1, γ=0, n_1=5, t=10). Resultados para el corpus CLEF 2001-02·A con consultas cortas y altura $h_t = f_{q,t} \cdot \frac{N}{f_t}$. Formas de función: constante (*cte*), triangular (*tri*), coseno (*cos*), circular (*cir*), arco (*arc*), exponencial (*exp*)

técnica	***lem***	***cte***	***%Δ***	***tri***	***%Δ***	***cos***	***%Δ***	***cir***	***%Δ***	***arc***	***%Δ***	***exp***	***%Δ***
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46k	46k	–	46k	–	46k	–	46k	–	46k	–	46k	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2767	2767	–	2767	–	2767	–	2767	–	2767	–	2767	–
Pr. no int.	.5220	.4454	-14.67	.4457	-14.62	.4448	-14.79	.4467	-14.43	.4462	-14.52	.4373	-16.23
Pr. doc.	.5784	.4805	-16.93	.4773	-17.48	.4772	-17.50	.4805	-16.93	.4792	-17.15	.4644	-19.71
R-pr.	.4990	.4402	-11.78	.4395	-11.92	.4392	-11.98	.4389	-12.04	.4402	-11.78	.4290	-14.03
Pr. a 0 %	.8221	**.8267**	**0.56**	**.8361**	**1.70**	**.8359**	**1.68**	**.8386**	**2.01**	**.8350**	**1.57**	**.8302**	**0.99**
Pr. a 10 %	.7490	.7333	-2.10	.7322	-2.24	.7241	-3.32	.7324	-2.22	.7310	-2.40	.7217	-3.64
Pr. a 20 %	.6866	.6314	-8.04	.6372	-7.19	.6391	-6.92	.6378	-7.11	.6355	-7.44	.6317	-8.00
Pr. a 30 %	.6573	.5715	-13.05	.5735	-12.75	.5695	-13.36	.5771	-12.20	.5758	-12.40	.5532	-15.84
Pr. a 40 %	.5997	.4874	-18.73	.4853	-19.08	.4880	-18.63	.4868	-18.83	.4865	-18.88	.4723	-21.24
Pr. a 50 %	.5456	.4296	-21.26	.4284	-21.48	.4303	-21.13	.4289	-21.39	.4278	-21.59	.4277	-21.61
Pr. a 60 %	.4994	.3818	-23.55	.3811	-23.69	.3809	-23.73	.3793	-24.05	.3800	-23.91	.3701	-25.89
Pr. a 70 %	.4375	.3359	-23.22	.3386	-22.61	.3386	-22.61	.3373	-22.90	.3386	-22.61	.3331	-23.86
Pr. a 80 %	.3661	.3014	-17.67	.2947	-19.50	.2947	-19.50	.3023	-17.43	.2980	-18.60	.2920	-20.24
Pr. a 90 %	.2939	.2260	-23.10	.2221	-24.43	.2206	-24.94	.2259	-23.14	.2228	-24.19	.2141	-27.15
Pr. a 100 %	.1547	.1235	-20.17	.1218	-21.27	.1199	-22.50	.1240	-19.84	.1226	-20.75	.1155	-25.34
Pr. a 5 docs.	.6609	.6391	-3.30	.6217	-5.93	.6217	-5.93	.6304	-4.61	.6261	-5.27	.6435	-2.63
Pr. a 10 docs.	.6457	.5891	-8.77	.6022	-6.74	.6043	-6.41	.5957	-7.74	.5978	-7.42	.5913	-8.42
Pr. a 15 docs.	.5884	.5623	-4.44	.5565	-5.42	.5565	-5.42	.5609	-4.67	.5565	-5.42	.5493	-6.65
Pr. a 20 docs.	.5630	.5228	-7.14	.5174	-8.10	.5141	-8.69	.5228	-7.14	.5163	-8.29	.5011	-10.99
Pr. a 30 docs.	.5225	.4601	-11.94	.4572	-12.50	.4551	-12.90	.4601	-11.94	.4572	-12.50	.4413	-15.54
Pr. a 100 docs.	.3507	.2957	-15.68	.2943	-16.08	.2970	-15.31	.2961	-15.57	.2946	-16.00	.2880	-17.88
Pr. a 200 docs.	.2348	.2110	-10.14	.2086	-11.16	.2091	-10.95	.2102	-10.48	.2087	-11.12	.2060	-12.27
Pr. a 500 docs.	.1122	.1102	-1.78	.1098	-2.14	.1098	-2.14	.1100	-1.96	.1101	-1.87	.1093	-2.58
Pr. a 1000 docs.	.0602	.0602	0.00	.0602	0.00	.0602	0.00	.0602	0.00	.0602	0.00	.0602	0.00

Tabla G.2: Reordenación mediante distancias de la lematización con realimentación (consultas cortas: α=0.8, β=0.1, γ=0, n_1=5, t=10). Resultados para el corpus CLEF 2001-02·A con consultas cortas y altura $h_t = f_{q,t} \cdot log_e(N/f_t)$. Formas de función: constante (*cte*), triangular (*tri*), coseno (*cos*), circular (*cir*), arco (*arc*), exponencial (*exp*)

técnica	***lem***	***cte***	**%Δ**	***tri***	**%Δ**	***cos***	**%Δ**	***cir***	**%Δ**	***arc***	**%Δ**	***exp***	**%Δ**
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46k	46k	–	46k	–	46k	–	46k	–	46k	–	46k	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2767	2767	–	2767	–	2767	–	2767	–	2767	–	2767	–
Pr. no int.	.5220	.4637	-11.17	.4662	-10.69	.4656	-10.80	.4668	-10.57	.4669	-10.56	.4567	-12.51
Pr. doc.	.5784	.5029	-13.05	.5015	-13.30	.5018	-13.24	.5039	-12.88	.5031	-13.02	.4882	-15.59
R-pr.	.4990	.4647	-6.87	.4659	-6.63	.4663	-6.55	.4651	-6.79	.4665	-6.51	.4533	-9.16
Pr. a 0 %	.8221	**.8806**	**7.12**	**.8791**	**6.93**	**.8794**	**6.97**	**.8835**	**7.47**	**.8812**	**7.19**	**.8824**	**7.33**
Pr. a 10 %	.7490	**.7839**	**4.66**	**.7790**	**4.01**	**.7761**	**3.62**	**.7870**	**5.07**	**.7853**	**4.85**	**.7648**	**2.11**
Pr. a 20 %	.6866	.6727	-2.02	.6845	-0.31	.6853	-0.19	**.6883**	**0.25**	.6862	-0.06	.6785	-1.18
Pr. a 30 %	.6573	.6067	-7.70	.6160	-6.28	.6099	-7.21	.6148	-6.47	.6111	-7.03	.6090	-7.35
Pr. a 40 %	.5997	.5254	-12.39	.5254	-12.39	.5255	-12.37	.5267	-12.17	.5283	-11.91	.5176	-13.69
Pr. a 50 %	.5456	.4656	-14.66	.4664	-14.52	.4675	-14.31	.4656	-14.66	.4668	-14.44	.4603	-15.63
Pr. a 60 %	.4994	.3999	-19.92	.3974	-20.42	.3987	-20.16	.4011	-19.68	.4001	-19.88	.3858	-22.75
Pr. a 70 %	.4375	.3331	-23.86	.3351	-23.41	.3363	-23.13	.3362	-23.15	.3363	-23.13	.3286	-24.89
Pr. a 80 %	.3661	.2794	-23.68	.2841	-22.40	.2838	-22.48	.2817	-23.05	.2814	-23.14	.2747	-24.97
Pr. a 90 %	.2939	.1954	-33.51	.1950	-33.65	.1934	-34.20	.1968	-33.04	.1965	-33.14	.1807	-38.52
Pr. a 100 %	.1547	.0960	-37.94	.0936	-39.50	.0916	-40.79	.0962	-37.82	.0956	-38.20	.0873	-43.57
Pr. a 5 docs.	.6609	**.6913**	**4.60**	**.6826**	**3.28**	**.6870**	**3.95**	**.6913**	**4.60**	**.6913**	**4.60**	**.6913**	**4.60**
Pr. a 10 docs.	.6457	.6435	-0.34	.6370	-1.35	.6413	-0.68	.6391	-1.02	.6413	-0.68	.6239	-3.38
Pr. a 15 docs.	.5884	**.5913**	**0.49**	.5841	-0.73	.5841	-0.73	**.5899**	**0.25**	.5884	0.00	.5812	-1.22
Pr. a 20 docs.	.5630	.5446	-3.27	.5467	-2.90	.5478	-2.70	.5446	-3.27	.5500	-2.31	.5391	-4.25
Pr. a 30 docs.	.5225	.4826	-7.64	.4848	-7.22	.4862	-6.95	.4848	-7.22	.4855	-7.08	.4754	-9.01
Pr. a 100 docs.	.3507	.3037	-13.40	.3041	-13.29	.3050	-13.03	.3052	-12.97	.3050	-13.03	.2963	-15.51
Pr. a 200 docs.	.2348	.2139	-8.90	.2135	-9.07	.2138	-8.94	.2145	-8.65	.2139	-8.90	.2086	-11.16
Pr. a 500 docs.	.1122	.1101	-1.87	.1102	-1.78	.1103	-1.69	.1100	-1.96	.1101	-1.87	.1096	-2.32
Pr. a 1000 docs.	.0602	.0602	0.00	.0602	0.00	.0602	0.00	.0602	0.00	.0602	0.00	.0602	0.00

Tabla G.3: Reordenación mediante distancias de la lematización con realimentación (consultas largas: α=1.2, β=0.1, γ=0, n_1=5, t=10). Resultados para el corpus CLEF 2001-02-A con consultas largas y altura $h_t = f_{q,t} \cdot \frac{N}{f_t}$. Formas de función: constante (*cte*), triangular (*tri*), coseno (*cos*), circular (*cir*), arco (*arc*), exponencial (*exp*)

técnica	***lem***	***cte***	***%Δ***	***tri***	***%Δ***	***cos***	***%Δ***	***cir***	***%Δ***	***arc***	***%Δ***	***exp***	***%Δ***
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46k	46k	–	46k	–	46k	–	46k	–	46k	–	46k	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2779	2779	–	2779	–	2779	–	2779	–	2779	–	2779	–
Pr. no int.	.5604	.4476	-20.13	.4471	-20.22	.4469	-20.25	.4476	-20.13	.4471	-20.22	.4458	-20.45
Pr. doc.	.5912	.4831	-18.28	.4831	-18.28	.4833	-18.25	.4837	-18.18	.4832	-18.27	.4776	-19.22
R-pr.	.5366	.4361	-18.73	.4390	-18.19	.4402	-17.96	.4386	-18.26	.4403	-17.95	.4357	-18.80
Pr. a 0 %	.8895	.8649	-2.77	.8645	-2.81	.8642	-2.84	.8689	-2.32	.8594	-3.38	.8722	-1.94
Pr. a 10 %	.8028	.7518	-6.35	.7476	-6.88	.7488	-6.73	.7512	-6.43	.7513	-6.42	.7436	-7.37
Pr. a 20 %	.7352	.6471	-11.98	.6346	-13.68	.6354	-13.57	.6383	-13.18	.6401	-12.94	.6416	-12.73
Pr. a 30 %	.6996	.5823	-16.77	.5768	-17.55	.5717	-18.28	.5812	-16.92	.5805	-17.02	.5721	-18.22
Pr. a 40 %	.6541	.4992	-23.68	.4900	-25.09	.4848	-25.88	.4928	-24.66	.4937	-24.52	.4796	-26.68
Pr. a 50 %	.6005	.4237	-29.44	.4322	-28.03	.4310	-28.23	.4263	-29.01	.4298	-28.43	.4237	-29.44
Pr. a 60 %	.5386	.3778	-29.86	.3810	-29.26	.3817	-29.13	.3763	-30.13	.3792	-29.60	.3818	-29.11
Pr. a 70 %	.4621	.3336	-27.81	.3350	-27.50	.3352	-27.46	.3331	-27.92	.3336	-27.81	.3388	-26.68
Pr. a 80 %	.3910	.2923	-25.24	.2974	-23.94	.2972	-23.99	.2967	-24.12	.2977	-23.86	.2943	-24.73
Pr. a 90 %	.3130	.2062	-34.12	.2099	-32.94	.2098	-32.97	.2079	-33.58	.2082	-33.48	.2103	-32.81
Pr. a 100 %	.1624	.1051	-35.28	.1084	-33.25	.1079	-33.56	.1072	-33.99	.1073	-33.93	.1091	-32.82
Pr. a 5 docs.	.6957	.6739	-3.13	.6739	-3.13	.6652	-4.38	.6739	-3.13	.6696	-3.75	.6522	-6.25
Pr. a 10 docs.	.6848	.5913	-13.65	.5957	-13.01	.5935	-13.33	.5870	-14.28	.5891	-13.97	.5935	-13.33
Pr. a 15 docs.	.6435	.5609	-12.84	.5652	-12.17	.5638	-12.39	.5638	-12.39	.5638	-12.39	.5522	-14.19
Pr. a 20 docs.	.6043	.5272	-12.76	.5217	-13.67	.5196	-14.02	.5228	-13.49	.5217	-13.67	.5152	-14.74
Pr. a 30 docs.	.5580	.4609	-17.40	.4645	-16.76	.4659	-16.51	.4667	-16.36	.4667	-16.36	.4652	-16.63
Pr. a 100 docs.	.3598	.2954	-17.90	.2976	-17.29	.2965	-17.59	.2963	-17.65	.2961	-17.70	.2941	-18.26
Pr. a 200 docs.	.2361	.2097	-11.18	.2101	-11.01	.2100	-11.05	.2093	-11.35	.2109	-10.67	.2104	-10.89
Pr. a 500 docs.	.1121	.1112	-0.80	.1108	-1.16	.1107	-1.25	.1110	-0.98	.1111	-0.89	.1107	-1.25
Pr. a 1000 docs.	.0604	.0604	0.00	.0604	0.00	.0604	0.00	.0604	0.00	.0604	0.00	.0604	0.00

Tabla G.4: Reordenación mediante distancias de la lematización con realimentación (consultas largas: α=1.2, β=0.1, γ=0, n_1=5, t=10). Resultados para el corpus CLEF 2001-02·A con consultas largas y altura $h_t = f_{q,t} \cdot log_e(N/f_t)$. Formas de función: constante (*cte*), triangular (*tri*), coseno (*cos*), circular (*cir*), arco (*arc*), exponencial (*exp*)

técnica	***lem***	***cte***	**%Δ**	***tri***	**%Δ**	***cos***	**%Δ**	***cir***	**%Δ**	***arc***	**%Δ**	***exp***	**%Δ**
#consultas	46	46	–	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46k	46k	–	46k	–	46k	–	46k	–	46k	–	46k	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2779	2779	–	2779	–	2779	–	2779	–	2779	–	2779	–
Pr. no int.	.5604	.4691	-16.29	.4731	-15.58	.4743	-15.36	.4714	-15.88	.4726	-15.67	.4700	-16.13
Pr. doc.	.5912	.5040	-14.75	.5063	-14.36	.5073	-14.19	.5064	-14.34	.5067	-14.29	.4967	-15.98
R-pr.	.5366	.4613	-14.03	.4661	-13.14	.4699	-12.43	.4652	-13.31	.4648	-13.38	.4610	-14.09
Pr. a 0 %	.8895	**.8974**	**0.89**	**.9015**	**1.35**	**.9023**	**1.44**	**.8979**	**0.94**	**.9022**	**1.43**	**.9003**	**1.21**
Pr. a 10 %	.8028	**.8127**	**1.23**	**.8160**	**1.64**	**.8124**	**1.20**	**.8143**	**1.43**	**.8159**	**1.63**	.7867	-2.01
Pr. a 20 %	.7352	.7014	-4.60	.6955	-5.40	.6984	-5.01	.7017	-4.56	.6946	-5.52	.6980	-5.06
Pr. a 30 %	.6996	.6032	-13.78	.6099	-12.82	.6118	-12.55	.6066	-13.29	.6077	-13.14	.6083	-13.05
Pr. a 40 %	.6541	.5373	-17.86	.5382	-17.72	.5389	-17.61	.5372	-17.87	.5409	-17.31	.5339	-18.38
Pr. a 50 %	.6005	.4668	-22.26	.4670	-22.23	.4674	-22.16	.4728	-21.27	.4713	-21.52	.4552	-24.20
Pr. a 60 %	.5386	.3916	-27.29	.4044	-24.92	.4070	-24.43	.3968	-26.33	.4024	-25.29	.3981	-26.09
Pr. a 70 %	.4621	.3314	-28.28	.3453	-25.28	.3463	-25.06	.3402	-26.38	.3415	-26.10	.3463	-25.06
Pr. a 80 %	.3910	.2780	-28.90	.2922	-25.27	.2932	-25.01	.2836	-27.47	.2887	-26.16	.2876	-26.45
Pr. a 90 %	.3130	.1835	-41.37	.1897	-39.39	.1891	-39.58	.1876	-40.06	.1878	-40.00	.1862	-40.51
Pr. a 100 %	.1624	.0855	-47.35	.0883	-45.63	.0880	-45.81	.0865	-46.74	.0872	-46.31	.0898	-44.70
Pr. a 5 docs.	.6957	**.7174**	**3.12**	**.7130**	**2.49**	**.7130**	**2.49**	**.7261**	**4.37**	**.7174**	**3.12**	**.7130**	**2.49**
Pr. a 10 docs.	.6848	.6522	-4.76	.6457	-5.71	.6435	-6.03	.6522	-4.76	.6522	-4.76	.6413	-6.35
Pr. a 15 docs.	.6435	.5986	-6.98	.6072	-5.64	.6043	-6.09	.5971	-7.21	.6029	-6.31	.5899	-8.33
Pr. a 20 docs.	.6043	.5620	-7.00	.5598	-7.36	.5652	-6.47	.5674	-6.11	.5620	-7.00	.5533	-8.44
Pr. a 30 docs.	.5580	.4935	-11.56	.4964	-11.04	.4942	-11.43	.4971	-10.91	.4971	-10.91	.4913	-11.95
Pr. a 100 docs.	.3598	.3015	-16.20	.3057	-15.04	.3061	-14.92	.3048	-15.29	.3048	-15.29	.3007	-16.43
Pr. a 200 docs.	.2361	.2125	-10.00	.2136	-9.53	.2142	-9.28	.2124	-10.04	.2128	-9.87	.2111	-10.59
Pr. a 500 docs.	.1121	.1106	-1.34	.1107	-1.25	.1107	-1.25	.1107	-1.25	.1109	-1.07	.1100	-1.87
Pr. a 1000 docs.	.0604	.0604	0.00	.0604	0.00	.0604	0.00	.0604	0.00	.0604	0.00	.0604	0.00

	Docs. relevantes							Docs. no relevantes			
K	$D \setminus L$	$L \setminus D$	$L \cap D$	R_{sup}	$Pr(L)$	$Pr(D)$	$Pr(L \cap D)$	$D \setminus L$	$L \setminus D$	$L \cap D$	N_{sup}
	corpus CLEF 2001-02·A: consultas cortas										
5	1.93	1.78	1.52	0.45	**0.66**	**0.69**	***0.80***	1.15	1.30	0.39	0.24
10	3.24	3.30	3.15	0.49	**0.65**	**0.64**	***0.76***	2.61	2.54	1.00	0.28
15	4.17	4.15	4.67	0.53	**0.59**	**0.59**	***0.72***	4.35	4.37	1.80	0.29
20	4.59	4.96	6.30	0.57	**0.56**	**0.54**	***0.72***	6.65	6.28	2.46	0.28
30	5.61	6.74	8.93	0.59	**0.52**	**0.48**	***0.68***	11.22	10.09	4.24	0.28
100	7.00	11.48	23.52	0.72	**0.35**	**0.31**	***0.49***	45.43	40.96	24.04	0.36
200	5.54	9.63	37.35	0.83	**0.23**	**0.21**	***0.36***	90.50	86.41	66.61	0.43
500	2.35	3.39	52.67	0.95	**0.11**	**0.11**	***0.16***	167.43	166.39	277.54	0.62
	corpus CLEF 2001-02·A: consultas largas										
5	2.11	1.96	1.52	0.43	**0.70**	**0.73**	***0.81***	1.02	1.17	0.35	0.24
10	3.13	3.46	3.39	0.51	**0.69**	**0.65**	***0.78***	2.54	2.22	0.93	0.28
15	3.98	4.67	4.98	0.54	**0.64**	**0.60**	***0.74***	4.30	3.61	1.74	0.31
20	4.54	5.28	6.80	0.58	**0.60**	**0.57**	***0.71***	5.91	5.17	2.74	0.33
30	5.07	6.89	9.85	0.62	**0.56**	**0.50**	***0.67***	10.20	8.37	4.89	0.34
100	6.26	11.70	24.22	0.73	**0.36**	**0.30**	***0.48***	43.02	37.59	26.50	0.40
200	5.13	9.87	37.35	0.83	**0.24**	**0.21**	***0.35***	87.37	82.63	70.15	0.45
500	2.70	3.39	52.67	0.95	**0.11**	**0.11**	***0.16***	165.46	164.76	279.17	0.63
	corpus CLEF 2001-02·B: consultas cortas										
5	1.73	1.71	1.27	0.42	**0.60**	**0.60**	***0.78***	1.64	1.67	0.36	0.18
10	3.02	3.16	2.42	0.44	**0.56**	**0.54**	***0.65***	3.22	3.09	1.33	0.30
15	3.91	4.16	3.76	0.48	**0.53**	**0.51**	***0.62***	5.07	4.82	2.27	0.31
20	4.29	4.67	5.36	0.54	**0.50**	**0.48**	***0.63***	7.16	6.78	3.20	0.31
30	5.24	5.93	7.40	0.57	**0.44**	**0.42**	***0.57***	11.73	11.04	5.62	0.33
100	6.84	8.47	20.96	0.73	**0.29**	**0.28**	***0.43***	44.73	43.11	27.47	0.38
200	5.78	6.89	32.73	0.84	**0.20**	**0.19**	***0.30***	86.67	85.56	74.82	0.46
500	2.60	2.53	46.49	0.95	**0.10**	**0.10**	***0.14***	164.00	164.07	286.91	0.64
	corpus CLEF 2001-02·B: consultas largas										
5	1.84	2.00	1.42	0.43	**0.68**	**0.65**	***0.81***	1.40	1.24	0.33	0.20
10	3.13	3.22	2.96	0.48	**0.62**	**0.61**	***0.72***	2.76	2.67	1.16	0.30
15	4.13	4.53	4.20	0.49	**0.58**	**0.56**	***0.66***	4.51	4.11	2.16	0.33
20	4.58	5.27	5.80	0.54	**0.55**	**0.52**	***0.64***	6.38	5.69	3.24	0.35
30	5.42	6.44	8.78	0.60	**0.51**	**0.47**	***0.60***	9.91	8.89	5.89	0.39
100	7.02	8.71	23.20	0.75	**0.32**	**0.30**	***0.43***	39.49	37.80	30.29	0.44
200	5.84	6.29	35.02	0.85	**0.21**	**0.20**	***0.30***	78.47	78.02	80.67	0.51
500	2.04	1.84	48.62	0.96	**0.10**	**0.10**	***0.14***	151.84	152.04	297.49	0.66
	corpus CLEF 2003: consultas cortas										
5	1.57	1.74	1.19	0.42	**0.59**	**0.55**	***0.73***	1.81	1.64	0.43	0.20
10	2.38	2.91	2.68	0.50	**0.56**	**0.51**	***0.73***	3.96	3.43	0.98	0.21
15	3.09	4.11	3.85	0.52	**0.53**	**0.46**	***0.66***	6.11	5.09	1.96	0.26
20	3.40	4.81	4.96	0.55	**0.49**	**0.42**	***0.61***	8.47	7.06	3.17	0.29
30	3.70	5.89	7.40	0.61	**0.44**	**0.37**	***0.59***	13.72	11.53	5.17	0.29
100	4.57	8.28	19.43	0.75	**0.28**	**0.24**	***0.42***	48.72	45.02	27.28	0.37
200	4.00	7.17	27.87	0.83	**0.18**	**0.16**	***0.28***	95.55	92.38	72.57	0.44
500	2.81	3.40	40.06	0.93	**0.09**	**0.09**	***0.12***	173.19	172.60	283.94	0.62
	corpus CLEF 2003: consultas largas										
5	1.40	1.72	1.38	0.47	**0.62**	**0.56**	***0.75***	1.77	1.45	0.45	0.22
10	2.19	3.09	2.79	0.51	**0.59**	**0.50**	***0.74***	4.06	3.17	0.96	0.21
15	2.68	3.96	4.30	0.56	**0.55**	**0.47**	***0.69***	6.09	4.81	1.94	0.26
20	2.98	4.96	5.57	0.58	**0.53**	**0.43**	***0.66***	8.53	6.55	2.91	0.28
30	3.53	6.19	7.81	0.62	**0.47**	**0.38**	***0.60***	13.53	10.87	5.13	0.30
100	3.91	8.43	20.13	0.77	**0.29**	**0.24**	***0.40***	46.19	41.68	29.77	0.40
200	3.45	6.77	29.06	0.85	**0.18**	**0.16**	***0.27***	88.30	84.98	79.19	0.48
500	2.17	3.00	40.57	0.94	**0.09**	**0.09**	***0.12***	167.40	166.57	289.85	0.63

Tabla G.5: Distribución de documentos relevantes y no relevantes tras la reordenación mediante distancias (función circular, altura $h_t = f_{q,t} \cdot log_e(N/f_t)$)

Tabla G.6: Reordenación mediante fusión por intersección de la lematización con realimentación (consultas cortas: α=0.8, β=0.1, γ=0, n_1=5, t=10; consultas largas: α=1.2, β=0.1, γ=0, n_1=5, t=10) . Resultados para el corpus CLEF 2001-02·A con consultas cortas y $K \in \{5, 10, 15, 20, 30\}$ (función circular, altura $h_t = f_{q,t} \cdot log_e(N/f_t)$)

técnica	***lem***	$K = 5$	%Δ	$K = 10$	%Δ	$K = 15$	%Δ	$K = 20$	%Δ	$K = 30$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2767	2767	0.00	2767	0.00	2767	0.00	2767	0.00	2767	0.00
Pr. no int.	.5220	**.5263**	**0.82**	**.5307**	**1.67**	**.5279**	**1.13**	**.5287**	**1.28**	**.5327**	**2.05**
Pr. doc.	.5784	**.5796**	**0.21**	**.5822**	**0.66**	**.5813**	**0.50**	**.5798**	**0.24**	**.5807**	**0.40**
R-pr.	.4990	.4971	-0.38	**.5082**	**1.84**	**.5073**	**1.66**	**.5069**	**1.58**	**.5126**	**2.73**
Pr. a 0 %	.8221	**.8444**	**2.71**	**.8502**	**3.42**	**.8592**	**4.51**	**.8435**	**2.60**	**.8386**	**2.01**
Pr. a 10 %	.7490	**.7705**	**2.87**	**.7777**	**3.83**	**.7791**	**4.02**	**.7837**	**4.63**	**.7758**	**3.58**
Pr. a 20 %	.6866	**.7058**	**2.80**	**.7229**	**5.29**	**.7224**	**5.21**	**.7313**	**6.51**	**.7193**	**4.76**
Pr. a 30 %	.6573	.6573	0.00	**.6756**	**2.78**	**.6744**	**2.60**	**.6740**	**2.54**	**.6844**	**4.12**
Pr. a 40 %	.5997	**.6042**	**0.75**	**.6021**	**0.40**	.5968	-0.48	**.6037**	**0.67**	**.6164**	**2.78**
Pr. a 50 %	.5456	.5453	-0.05	**.5466**	**0.18**	**.5487**	**0.57**	**.5461**	**0.09**	**.5644**	**3.45**
Pr. a 60 %	.4994	**.4997**	**0.06**	.4961	-0.66	**.5022**	**0.56**	**.5013**	**0.38**	**.5098**	**2.08**
Pr. a 70 %	.4375	.4366	-0.21	.4333	-0.96	.4257	-2.70	.4254	-2.77	**.4391**	**0.37**
Pr. a 80 %	.3661	.3657	-0.11	.3639	-0.60	.3592	-1.88	.3652	-0.25	**.3694**	**0.90**
Pr. a 90 %	.2939	.2934	-0.17	**.2945**	**0.20**	.2939	0.00	.2908	-1.05	.2919	-0.68
Pr. a 100 %	.1547	.1547	0.00	.1512	-2.26	.1516	-2.00	.1516	-2.00	.1545	-0.13
Pr. a 5 docs.	.6609	.6609	0.00	**.6913**	**4.60**	**.6870**	**3.95**	**.6913**	**4.60**	**.6739**	**1.97**
Pr. a 10 docs.	.6457	.6457	0.00	.6457	0.00	.6413	-0.68	**.6565**	**1.67**	**.6761**	**4.71**
Pr. a 15 docs.	.5884	**.6058**	**2.96**	**.6188**	**5.17**	.5884	0.00	**.6072**	**3.20**	**.6188**	**5.17**
Pr. a 20 docs.	.5630	**.5696**	**1.17**	**.5804**	**3.09**	**.5783**	**2.72**	.5630	0.00	**.5826**	**3.48**
Pr. a 30 docs.	.5225	**.5283**	**1.11**	**.5297**	**1.38**	**.5261**	**0.69**	.5196	-0.56	.5225	0.00
Pr. a 100 docs.	.3507	**.3511**	**0.11**	**.3528**	**0.60**	**.3524**	**0.48**	**.3524**	**0.48**	.3502	-0.14
Pr. a 200 docs.	.2348	**.2352**	**0.17**	**.2361**	**0.55**	**.2359**	**0.47**	**.2355**	**0.30**	**.2349**	**0.04**
Pr. a 500 docs.	.1122	.1122	0.00	.1122	0.00	.1122	0.00	.1122	0.00	.1122	0.00
Pr. a 1000 docs.	.0602	.0602	0.00	.0602	0.00	.0602	0.00	.0602	0.00	.0602	0.00

Tabla G.7: Reordenación mediante fusión por intersección de la lematización con realimentación (consultas cortas: α=0.8, β=0.1, γ=0, n_1=5, t=10; consultas largas: α=1.2, β=0.1, γ=0, n_1=5, t=10) . Resultados para el corpus CLEF 2001-02·A con consultas cortas y $K \in \{50, 75, 100, 200, 500\}$ (función circular, altura $h_t = f_{q,t} \cdot log_e(N/f_t)$)

técnica	***lem***	$K = 50$	%Δ	$K = 75$	%Δ	$K = 100$	%Δ	$K = 200$	%Δ	$K = 500$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2767	2767	0.00	2767	0.00	2767	0.00	2767	0.00	2767	0.00
Pr. no int.	.5220	**.5293**	**1.40**	**.5299**	**1.51**	**.5340**	**2.30**	**.5365**	**2.78**	**.5259**	**0.75**
Pr. doc.	.5784	**.5830**	**0.80**	**.5881**	**1.68**	**.5924**	**2.42**	**.5984**	**3.46**	**.5880**	**1.66**
R-pr.	.4990	**.5057**	**1.34**	**.5085**	**1.90**	**.5138**	**2.97**	**.5146**	**3.13**	**.5084**	**1.88**
Pr. a 0 %	.8221	**.8337**	**1.41**	**.8285**	**0.78**	**.8351**	**1.58**	**.8319**	**1.19**	**.8265**	**0.54**
Pr. a 10 %	.7490	**.7773**	**3.78**	**.7716**	**3.02**	**.7700**	**2.80**	**.7750**	**3.47**	**.7588**	**1.31**
Pr. a 20 %	.6866	**.7084**	**3.18**	**.7090**	**3.26**	**.7082**	**3.15**	**.7045**	**2.61**	**.6943**	**1.12**
Pr. a 30 %	.6573	**.6790**	**3.30**	**.6607**	**0.52**	**.6593**	**0.30**	**.6720**	**2.24**	**.6665**	**1.40**
Pr. a 40 %	.5997	**.6273**	**4.60**	**.6135**	**2.30**	**.6147**	**2.50**	**.6080**	**1.38**	**.6106**	**1.82**
Pr. a 50 %	.5456	**.5672**	**3.96**	**.5691**	**4.31**	**.5665**	**3.83**	**.5651**	**3.57**	**.5576**	**2.20**
Pr. a 60 %	.4994	**.5030**	**0.72**	**.5178**	**3.68**	**.5240**	**4.93**	**.5225**	**4.63**	**.5094**	**2.00**
Pr. a 70 %	.4375	**.4383**	**0.18**	**.4480**	**2.40**	**.4524**	**3.41**	**.4579**	**4.66**	**.4430**	**1.26**
Pr. a 80 %	.3661	**.3711**	**1.37**	**.3751**	**2.46**	**.3817**	**4.26**	**.3868**	**5.65**	**.3694**	**0.90**
Pr. a 90 %	.2939	.2794	-4.93	.2825	-3.88	**.2968**	**0.99**	**.3019**	**2.72**	.2810	-4.39
Pr. a 100 %	.1547	.1468	-5.11	.1413	-8.66	.1429	-7.63	.1516	-2.00	.1451	-6.21
Pr. a 5 docs.	.6609	.6609	0.00	**.6652**	**0.65**	**.6652**	**0.65**	**.6739**	**1.97**	**.6696**	**1.32**
Pr. a 10 docs.	.6457	**.6587**	**2.01**	**.6565**	**1.67**	**.6565**	**1.67**	**.6500**	**0.67**	**.6500**	**0.67**
Pr. a 15 docs.	.5884	**.6217**	**5.66**	**.6188**	**5.17**	**.6203**	**5.42**	**.6101**	**3.69**	**.5942**	**0.99**
Pr. a 20 docs.	.5630	**.5946**	**5.61**	**.5989**	**6.38**	**.6000**	**6.57**	**.5957**	**5.81**	**.5707**	**1.37**
Pr. a 30 docs.	.5225	**.5355**	**2.49**	**.5457**	**4.44**	**.5428**	**3.89**	**.5420**	**3.73**	**.5348**	**2.35**
Pr. a 100 docs.	.3507	.3433	-2.11	.3483	-0.68	.3500	-0.20	**.3552**	**1.28**	**.3535**	**0.80**
Pr. a 200 docs.	.2348	**.2354**	**0.26**	**.2367**	**0.81**	**.2372**	**1.02**	**.2349**	**0.04**	**.2350**	**0.09**
Pr. a 500 docs.	.1122	**.1125**	**0.27**	**.1127**	**0.45**	**.1127**	**0.45**	**.1134**	**1.07**	.1122	0.00
Pr. a 1000 docs.	.0602	.0602	0.00	.0602	0.00	.0602	0.00	.0602	0.00	.0602	0.00

Tabla G.8: Reordenación mediante fusión por intersección de la lematización con realimentación (consultas largas: α=1.2, β=0.1, γ=0, n_1=5, t=10) . Resultados para el corpus CLEF 2001-02·A con consultas largas y $K \in \{5, 10, 15, 20, 30\}$ (función circular, altura $h_t = f_{q,t} \cdot log_e(N/f_t)$)

técnica	***lem***	$K = 5$	%Δ	$K = 10$	%Δ	$K = 15$	%Δ	$K = 20$	%Δ	$K = 30$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2779	2779	0.00	2779	0.00	2779	0.00	2779	0.00	2779	0.00
Pr. no int.	.5604	.5598	-0.11	**.5630**	**0.46**	**.5624**	**0.36**	**.5650**	**0.82**	**.5607**	**0.05**
Pr. doc.	.5912	.5901	-0.19	.5903	-0.15	.5880	-0.54	.5912	0.00	.5900	-0.20
R-pr.	.5366	**.5425**	**1.10**	.5342	-0.45	**.5464**	**1.83**	**.5482**	**2.16**	**.5446**	**1.49**
Pr. a 0 %	.8895	**.8989**	**1.06**	**.9125**	**2.59**	**.9055**	**1.80**	**.9146**	**2.82**	**.9055**	**1.80**
Pr. a 10 %	.8028	**.8136**	**1.35**	**.8343**	**3.92**	**.8403**	**4.67**	**.8455**	**5.32**	**.8277**	**3.10**
Pr. a 20 %	.7352	**.7372**	**0.27**	**.7630**	**3.78**	**.7420**	**0.92**	**.7520**	**2.29**	**.7441**	**1.21**
Pr. a 30 %	.6996	.6993	-0.04	**.7249**	**3.62**	**.7093**	**1.39**	**.7038**	**0.60**	.6996	0.00
Pr. a 40 %	.6541	.6414	-1.94	.6415	-1.93	.6469	-1.10	.6479	-0.95	**.6585**	**0.67**
Pr. a 50 %	.6005	.5957	-0.80	.5968	-0.62	**.6050**	**0.75**	.6002	-0.05	.5979	-0.43
Pr. a 60 %	.5386	.5371	-0.28	.5198	-3.49	.5313	-1.36	.5344	-0.78	.5266	-2.23
Pr. a 70 %	.4621	.4558	-1.36	.4478	-3.09	.4533	-1.90	.4565	-1.21	.4577	-0.95
Pr. a 80 %	.3910	.3855	-1.41	.3823	-2.23	.3882	-0.72	.3889	-0.54	**.3916**	**0.15**
Pr. a 90 %	.3130	.3119	-0.35	.3065	-2.08	.3044	-2.75	.3058	-2.30	.3033	-3.10
Pr. a 100 %	.1624	.1617	-0.43	.1580	-2.71	.1598	-1.60	.1580	-2.71	.1596	-1.72
Pr. a 5 docs.	.6957	.6957	0.00	**.7348**	**5.62**	**.7478**	**7.49**	**.7435**	**6.87**	**.7348**	**5.62**
Pr. a 10 docs.	.6848	.6848	0.00	.6848	0.00	**.6913**	**0.95**	**.7022**	**2.54**	**.7043**	**2.85**
Pr. a 15 docs.	.6435	**.6522**	**1.35**	**.6478**	**0.67**	.6435	0.00	**.6580**	**2.25**	**.6580**	**2.25**
Pr. a 20 docs.	.6043	**.6141**	**1.62**	**.6098**	**0.91**	**.6109**	**1.09**	.6043	0.00	**.6239**	**3.24**
Pr. a 30 docs.	.5580	**.5609**	**0.52**	**.5587**	**0.13**	.5428	-2.72	.5486	-1.68	.5580	0.00
Pr. a 100 docs.	.3598	**.3607**	**0.25**	**.3604**	**0.17**	**.3604**	**0.17**	**.3602**	**0.11**	**.3607**	**0.25**
Pr. a 200 docs.	.2361	.2360	-0.04	**.2363**	**0.08**	**.2365**	**0.17**	**.2368**	**0.30**	**.2366**	**0.21**
Pr. a 500 docs.	.1121	.1121	0.00	.1121	0.00	**.1122**	**0.09**	**.1123**	**0.18**	**.1123**	**0.18**
Pr. a 1000 docs.	.0604	.0604	0.00	.0604	0.00	.0604	0.00	.0604	0.00	.0604	0.00

Tabla G.9: Reordenación mediante fusión por intersección de la lematización con realimentación (consultas largas: α=1.2, β=0.1, γ=0, n_1=5, t=10) . Resultados para el corpus CLEF 2001-02-A con consultas largas y $K \in \{50, 75, 100, 200, 500\}$ (función circular, altura $h_t = f_{q,t} \cdot log_e(N/f_t)$)

técnica	***lem***	$K = 50$	%Δ	$K = 75$	%Δ	$K = 100$	%Δ	$K = 200$	%Δ	$K = 500$	%Δ
#consultas	46	46	–	46	–	46	–	46	–	46	–
#docs. dev.	46000	46000	–	46000	–	46000	–	46000	–	46000	–
#rlvs. esp.	3007	3007	–	3007	–	3007	–	3007	–	3007	–
#rlvs. dev.	2779	2779	0.00	2779	0.00	2779	0.00	2779	0.00	2779	0.00
Pr. no int.	.5604	.5589	-0.27	.5586	-0.32	.5602	-0.04	.5602	-0.04	.5590	-0.25
Pr. doc.	.5912	**.5921**	**0.15**	**.5959**	**0.79**	**.6004**	**1.56**	**.5995**	**1.40**	**.5931**	**0.32**
R-pr.	.5366	**.5433**	**1.25**	.5280	-1.60	.5312	-1.01	**.5449**	**1.55**	**.5372**	**0.11**
Pr. a 0 %	.8895	**.9091**	**2.20**	**.9073**	**2.00**	**.9092**	**2.21**	**.8945**	**0.56**	**.8908**	**0.15**
Pr. a 10 %	.8028	**.8256**	**2.84**	**.8247**	**2.73**	**.8277**	**3.10**	**.8152**	**1.54**	**.8046**	**0.22**
Pr. a 20 %	.7352	**.7528**	**2.39**	**.7451**	**1.35**	**.7555**	**2.76**	**.7479**	**1.73**	**.7384**	**0.44**
Pr. a 30 %	.6996	.6922	-1.06	**.7070**	**1.06**	**.7072**	**1.09**	**.7052**	**0.80**	**.7049**	**0.76**
Pr. a 40 %	.6541	**.6610**	**1.05**	.6531	-0.15	**.6606**	**0.99**	**.6640**	**1.51**	**.6584**	**0.66**
Pr. a 50 %	.6005	**.6026**	**0.35**	.5998	-0.12	.5898	-1.78	.5996	-0.15	**.6028**	**0.38**
Pr. a 60 %	.5386	.5233	-2.84	.5222	-3.04	.5218	-3.12	.5325	-1.13	.5378	-0.15
Pr. a 70 %	.4621	.4554	-1.45	.4616	-0.11	.4550	-1.54	.4568	-1.15	.4604	-0.37
Pr. a 80 %	.3910	.3815	-2.43	.3872	-0.97	.3865	-1.15	.3725	-4.73	**.3933**	**0.59**
Pr. a 90 %	.3130	.2953	-5.65	.2974	-4.98	.3003	-4.06	.3005	-3.99	.2984	-4.66
Pr. a 100 %	.1624	.1537	-5.36	.1477	-9.05	.1507	-7.20	.1609	-0.92	.1457	-10.28
Pr. a 5 docs.	.6957	**.7217**	**3.74**	**.7217**	**3.74**	**.7130**	**2.49**	**.7043**	**1.24**	**.7000**	**0.62**
Pr. a 10 docs.	.6848	**.7065**	**3.17**	**.7109**	**3.81**	**.7152**	**4.44**	**.6978**	**1.90**	**.6870**	**0.32**
Pr. a 15 docs.	.6435	**.6449**	**0.22**	**.6580**	**2.25**	**.6551**	**1.80**	**.6565**	**2.02**	**.6449**	**0.22**
Pr. a 20 docs.	.6043	**.6185**	**2.35**	**.6163**	**1.99**	**.6163**	**1.99**	**.6141**	**1.62**	**.6054**	**0.18**
Pr. a 30 docs.	.5580	**.5652**	**1.29**	.5543	-0.66	**.5601**	**0.38**	**.5623**	**0.77**	**.5594**	**0.25**
Pr. a 100 docs.	.3598	.3539	-1.64	.3570	-0.78	.3593	-0.14	.3546	-1.45	.3563	-0.97
Pr. a 200 docs.	.2361	**.2380**	**0.80**	**.2375**	**0.59**	**.2371**	**0.42**	.2361	0.00	.2358	-0.13
Pr. a 500 docs.	.1121	**.1126**	**0.45**	**.1124**	**0.27**	**.1126**	**0.45**	**.1135**	**1.25**	.1121	0.00
Pr. a 1000 docs.	.0604	.0604	0.00	.0604	0.00	.0604	0.00	.0604	0.00	.0604	0.00

Bibliografía

[1] `http://trec.nist.gov` (visitada en febrero 2005).

[2] `http://www.clef-campaign.org` (visitada en febrero 2005).

[3] `http://snowball.tartarus.org` (visitada en febrero 2005).

[4] `ftp://ftp.cs.cornell.edu/pub/smart` (visitada en febrero 2005).

[5] `http://www.systransoft.com` (visitada en febrero 2005).

[6] `http://www.itl.nist.gov/iaui/894.02/works/papers/zp2/zp2.html` (visitada en febrero 2005).

[7] `http://www.dicesp.com` (visitada en febrero 2005).

[8] AGROVOC–Multilingual Agricultural Thesaurus. Food and Agricultural Organization of the United Nations, 1995. `http://www.fao.org/agrovoc` (visitada en febrero 2005).

[9] *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1997.

[10] Steven Abney. Parsing by chunks. In Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht, 1991.

[11] Steven Abney. Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4):337–344, 1997.

[12] Alfred V. Aho, Ravi Sethi, and Jeffrey D. Ullman. *Compilers: Principles, Techniques and Tools*. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, USA, 1986.

[13] Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation and Compiling*. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1972.

[14] Margarita Alonso. Presentación del diccionario de colocaciones y marcadores del español: estructura y objetivos. In María D. Muñoz, Ana I. Rodríguez-Piñeiro, Gérad Fernández, and Victoria Benítez, editors, *Actas del IV Congreso de Lingüística General*, volume II, pages 47–61. Servicio de Publicaciones de la Universidad de Cádiz, 2003. Demostración on-line disponible en [7].

[15] Margarita Alonso and Begoña Sanromán. Construcción de una base de datos de colocaciones léxicas. *Procesamiento del Lenguaje Natural*, 24:97–98, September 2000. Demostración on-line disponible en [7].

[16] Miguel A. Alonso. *Interpretación tabular de autómatas para lenguajes de adjunción de árboles*. PhD thesis, Departamento de Computación, Universidade da Coruña, Spain, 2000.

[17] Miguel A. Alonso, David Cabrero, Eric de la Clergerie, and Manuel Vilares. Tabular algorithms for TAG parsing. In *Proc. of EACL'99, Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 150–157, Bergen, Norway, June 1999. ACL.

[18] Miguel A. Alonso, Eric de la Clergerie, Jorge Graña, and Manuel Vilares. New tabular algorithms for LIG parsing. In *Proc. of the Sixth International Workshop on Parsing Technologies (IWPT 2000)*, pages 29–40, Trento, Italy, February 2000.

[19] Miguel A. Alonso, Eric de la Clergerie, and Manuel Vilares. Automata-based parsing in dynamic programming for Linear Indexed Grammars. In A. S. Narin'yani, editor, *Proc. of DIALOGUE'97 Computational Linguistics and its Applications International Workshop*, pages 22–27, Moscow, Russia, June 1997.

[20] Miguel A. Alonso, Víctor J. Díaz, and Manuel Vilares. Bidirectional automata for tree adjoining grammars. In *Proc. of the Seventh International Workshop on Parsing Technologies (IWPT-2001)*, pages 42–53, Beijing, China, October 2001. Tsinghua University Press.

[21] Miguel A. Alonso, Jesús Vilares, and Francisco J. Ribadas. Experiencias del Grupo COLE en la aplicación de técnicas de Procesamiento del Lenguaje Natural a la Recuperación de Información en español. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 8(22):123–134, 2004.

[22] Miguel A. Alonso Pardo, Vilares Ferro, and Víctor M. Darriba. On the usefulness of extracting syntactic dependencies for text indexing. In Michael O'Neill, Richard F. E. Sutcliffe, Conor Ryan, Malachy Eaton, and Niall J. L. Griffith, editors, *Artificial Intelligence and Cognitive Science*, volume 2464 of *Lecture Notes in Artificial Intelligence*, pages 3–11. Springer-Verlag, Berlin-Heidelberg-New York, 2002.

[23] Chinatsu Aone, Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz. SRA: Description of the IE^2 system used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[24] Avi Arampatzis, Th. P. van der Weide, P. van Bommel, and C.H.A. Koster. Linguistically-motivated information retrieval. In *Encyclopedia of Library and Information Science*, volume 69, pages 201–222. Marcel Dekker, Inc, New York-Basel, 2000.

[25] Lourdes Araujo. Part-of-Speech tagging with evolutionary algorithms. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 230–239. Springer-Verlag, Berlin-Heidelberg-New York, 2002.

[26] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley and ACM Press, Harlow, England, 1999.

[27] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983.

[28] Elena Bajo Pérez. *La derivación nominal en español*. Cuadernos de lengua española. Arco Libros, Madrid, 1997.

[29] J. K. Baker. The DRAGON system — an overview. *IEEE Transactions on Acoustic, Speech, and Signal Processing, ASSP*, 23(1):24–29, 1975.

[30] Fco. Mario Barcala, Eva M. Domínguez, Miguel A. Alonso, David Cabrero, Jorge Graña, Jesús Vilares, Manuel Vilares, Guillermo Rojo, M. Paula Santalla, and Susana Sotelo. El sistema ERIAL: LEIRA, un entorno para RI basado en PLN. In Emilio Sanchís, Lidia Moreno, and Isidoro Gil, editors, *Actas de las I Jornadas de Tratamiento y Recuperación de Información (JOTRI 2002). 4-5 de julio, Valencia*, pages 173–174, 2002.

[31] Fco. Mario Barcala, Eva M. Domínguez, Miguel A. Alonso, David Cabrero, Jorge Graña, Jesús Vilares, Manuel Vilares, Guillermo Rojo, M. Paula Santalla, and Susana Sotelo. Una aplicación de RI basada en PLN: el proyecto ERIAL. In Emilio Sanchís, Lidia Moreno, and Isidoro Gil, editors, *Actas de las I Jornadas de Tratamiento y Recuperación de Información (JOTRI 2002). 4-5 de julio, Valencia*, pages 165–172, 2002.

[32] Fco. Mario Barcala, Jesús Vilares, Miguel A. Alonso, Jorge Graña, and Manuel Vilares. Tokenization and proper noun recognition for information retrieval. In *3rd International Workshop on Natural Language and Information Systems (NLIS 2002), September 2-3, 2002. Aix-en-Provence, France*, Los Alamitos, California, USA, 2002. IEEE Computer Society Press.

[33] R. Beckwith, G.A. Miller, and R. Tengi. Design and implementation of the WordNet lexical database and searching software. Revised version of "Implementing a Lexical Network" in CSL Report 43, prepared by R. Tengi, August 1993.

[34] Donna Bergmark. Background readings for collection synthesis, 2002. Cornell Digital Library Research Group.

[35] J. M. Blecua, editor. *Diccionario Avanzado de Sinónimos y Antónimos de la Lengua Española*. Vox, Barcelona, Spain, 1997.

[36] T. L. Booth. Probabilistic representation of formal languages. In *IEEE Conference Record of the 1969 Tenth Annual Symposium on Switching and Automata Theory*, pages 74–81, 1969.

[37] Thorsten Brants. TnT - a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP'2000)*, Seattle, WA., 2000.

[38] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–566, 1995.

[39] Eric Brill. *Part-of-Speech Tagging*, chapter 17, pages 403–414. In Dale et al. [59], 2000.

[40] Chris Buckley. Implementation of the SMART information retrieval system. Technical report, Department of Computer Science, Cornell University, 1985. Source code available at [4].

[41] Chris Buckley. `trec_eval` software package, 1991. Source code available at [4].

[42] Chris Buckley, James Allan, and Gerard Salton. Automatic routing and ad-hoc retrieval using SMART: TREC 2. In D. K. Harman, editor, *NIST Special Publication 500-215: The Second Text REtrieval Conference (TREC-2)*, pages 45–56, Gaithersburg, MD, USA, 1993.

[43] Chris Buckley and Gerard Salton. Automatic retrieval with locality information using SMART. In D. K. Harman, editor, *NIST Special Publication 500-207: The First Text REtrieval Conference (TREC 1)*, pages 59–72, 1992.

[44] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic query expansion using SMART: TREC 3. In D. K. Harman, editor, *NIST Special Publication 500-225: Overview of the Third Text REtrieval Conference (TREC 3)*, pages 69–80, Gaithersburg, MD, USA, 1995. Department of Commerce, National Institute of Standards and Technology.

[45] Chris Buckley, Amit Singhal, Mandar Mitra, and Gerard Salton. New retrieval approaches using SMART: TREC-4. In D.K. Harman, editor, *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, pages 25–48, Gaithersburg, MD, USA, 1995. Department of Commerce, National Institute of Standards and Technology.

[46] David Cabrero. *Análisis eficaz de gramáticas de clásulas definidas.* PhD thesis, Departamento de Computación, Universidade da Coruña, Spain, 2002.

[47] David Cabrero, Jesús Vilares, and Manuel Vilares. Dynamic programming of partial parses. In *Actas de las Primeras Jornadas sobre Programación y Lenguajes. Almagro (Ciudad Real), Spain*, pages 63–76, 2001.

[48] David Cabrero, Jesús Vilares, and Manuel Vilares. Programación dinámica y análisis parcial. *Procesamiento del Lenguaje Natural*, 29:129–136, September 2002.

[49] Bob Carpenter. *The Logic of Typed Feature Structures.* Number 32 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge/New York/Melbourne, 1992.

[50] John Carrol, Ted Briscoe, and Antonio Sanfilippo. Parser evaluation: a survey and a new proposal. In *Proc. of the 1st International Conference on Language Resources and Evaluation*, pages 447–454, Granada, Spain, 1998.

[51] Jean-Pierre Chanod and Pasi Tapanainen. A non-deterministic tokeniser for finite-state parsing. In *ECAI '96 workshop on Extended Finite State Models of Language, August 11-12, Budapest, Hungary*, 1996.

[52] Eugene Charniak, Curtis Hendrickson, Neil Jacobson, and Mike Perkowitz. Equations for part-of-speech tagging. In *National Conference on Artificial Intelligence*, pages 784–789, 1993.

[53] Noam Chomsky. Three models for the description of language. *IRI Transactions on Information Theory*, 2(3):113–124, 1956.

[54] Noam Chomsky. On certain formal properties of grammars. *Information and Control*, 2(2):137–167, June 1959.

[55] K. W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proc. of the Second Conference on Applied Natural Language Processing*, pages 136–143. ACL, 1988.

[56] A. Colmerauer. Les systèmes-q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur. Internal Publication 43, Département d'Informatique de l'Université de Montréal, 1970.

[57] W. Bruce Croft. *Advances in Information Retrieval*, chapter Combining approaches to Information Retrieval, pages 1–36. Kluwer Academic Publishers, 2000.

[58] Robert Dale. *Symbolic Approaches to Natural Language Processing*, chapter 1, pages 1–9. In Dale et al. [59], 2000.

[59] Robert Dale, Hermann Moisi, and Harold Somers, editors. *Handbook of Natural Language Processing*. Marcel Dekker, Inc., New York & Basel, 2000.

[60] Víctor M. Darriba. *Corrección regional de errores con coste mínimo*. PhD thesis, Departamento de Computación, Universidade da Coruña, Spain, 2002.

[61] O. de Kretser and A. Moffat. Effective document presentation with a locality-based similarity heuristic. In *Proc. of SIGIR '99, August 15-19, Berkeley, CA, USA*, pages 113–120, 1999.

[62] O. de Kretser and A. Moffat. Locality-based information retrieval. In *Proc. of 10th Australasian Database Conference (ADC '99), 18-21 January, Auckland, New Zealand*, pages 177–188, 1999.

[63] Owen de Kretser. *Locality-Based Information Retrieval*. PhD thesis, Univeristy of Melbourne, 2000.

[64] Owen de Kretser and Alistair Moffat. SEFT: A search engine for text. *Software-Practice & Experience*, 34(10):1011–1023, August 2004. Source code available in: `http://www.cs.mu.oz.au/~oldk/seft/` (visitada en febrero 2005).

[65] Eric de la Clergerie. *Automates à Piles et Programmation Dynamique. DyALog : Une Application à la Programmation en Logique*. PhD thesis, Université Paris 7, Paris, France, 1993.

[66] Eric de la Clergerie, Miguel A. Alonso, and David Cabrero Souto. A tabular interpretation of bottom-up automata for TAG. In *Proc. of Fourth International Workshop on Tree-Adjoining Grammars and Related Frameworks (TAG+4)*, pages 42–45, Philadelphia, PA, USA, August 1998.

[67] J. Earley. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102, 1970.

[68] J.E. Ellman, I. Klincke, and J.I. Tait. Word sense disambiguation by information filtering and extraction. *Computers and the Humanities*, 34:127–134, 2000.

[69] J.L. Fagan. The effectiveness of a non-syntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40(2):115–132, 1989.

[70] Christiane Fellbaum. English verbs as a semantic net. *International Journal of Lexicography*, 3(4):278–301, 1990.

[71] Santiago Fernández Lanza. *Una contribución al procesamiento automático de la sinonimia utilizando Prolog*. PhD thesis, Universidade de Santiago de Compostela, Santiago de Compostela, Spain, 2001.

[72] Santiago Fernández Lanza, Jorge Graña Gil, and Alejandro Sobrino Cerdeiriña. A Spanish e-dictionary of synonyms as a fuzzy tool for information retrieval. In *Actas del XI Congreso Español sobre Tecnologías y Lógica Fuzzy (ESTYLF-2002)*, pages 31–37, León, Spain, September 2002.

[73] S. Fernández Ramírez. *La derivación nominal*. Number 40 in Anejos del Boletín de la Real Academia Española. Real Academia Española, Madrid, 1986.

[74] Carlos G. Figuerola, Raquel Gómez, Angel F. Zazo Rodríguez, and José Luis Alonso Berrocal. Stemming in Spanish: A first approach to its impact on information retrieval. In Peters [173].

[75] D. David Forney. The Viterbi algorithm. *Proceedings of the IEEE*, 61:268–278, March 1973.

[76] Edward A. Fox. Characterization of two new experimental collections in computer and information science containing textual and bibliographical concepts. Technical Report 83-561, 1983.

[77] Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In *Proceedings of 2nd Text REtrieval Conference (TREC-2), August 31-September 2, 1993, Gaithersburg, MD, USA*, pages 243–252. National Institute of Standards and Technology Special Publication 500-215, 1994.

[78] Sofia Natalia Galicia-Haro, Igor A. Bolshakov, and Alexander F. Gelbukh. A simple spanish part of speech tagger for detection and correction of accentuation errors. In V. Matousek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Text, Speech and Dialogue*, volume 1692 of *Lecture Notes in Computer Science*, pages 219–222. Springer-Verlag, Berlin-Heidelberg-New York, 1999.

[79] Pablo Gamallo, Alexandre Agustini, and Gabriel P. Lopes. Selections restrictions acquisition from corpora. In *10th Portuguese Conference on Artificial Intelligence (EPIA'01)*, Lecture Notes in Artificial Intelligence, pages 30–43. Springer-Verlag, 2001.

[80] Pablo Gamallo, Alexandre Agustini, and Gabriel P. Lopes. Clustering syntactic positions with similar semantic requirements. *Journal of Computational Linguistics*, 2004.

[81] Jesús Giménez and Lluís Márquez. Fast and accurate Part-of-Speech tagging: The SVM approach revisited. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, , Nicolas Nicolov, and Nikolai Nikolov, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP). September, 10-12 2003, Borovets, Bulgary*, pages 158–165, 2003.

[82] G. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP, Montreal, Canada*, pages 38–44, 1998.

[83] Jorge Graña. *Técnicas de Análisis Sintáctico Robusto para la Etiquetación del Lenguaje Natural*. PhD thesis, Departamento de Computación, Universidade da Coruña, A Coruña, Spain, December 2000.

[84] Jorge Graña, Gloria Andrade, and Jesús Vilares. Compilation of constraint-based contextual rules for part-of-speech tagging into finite state transducers. In Jean-Marc Champarnaud and Denis Maurel, editors, *Proceedings of Seventh International Conference*

on Implementation and Application of Automata (CIAA 2002). July, 3–5, Tours, France, pages 131–140, 2002.

[85] Jorge Graña, Gloria Andrade, and Jesús Vilares. Compilation of constraint-based contextual rules for part-of-speech tagging into finite state transducers. In Jean-Marc Champarnaud and Denis Maurel, editors, *Implementation and Application of Automata*, volume 2608 of *Lecture Notes in Computer Science*, pages 128–137. Berlin-Heidelberg-New York, 2003.

[86] Jorge Graña, Fco. Mario Barcala, and Miguel A. Alonso. Compilation methods of minimal acyclic automata for large dictionaries. In Bruce W. Watson and Derick Wood, editors, *Proc. of the 6th Conference on Implementations and Applications of Automata (CIAA 2001)*, pages 116–129, Pretoria, South Africa, July 2001.

[87] Jorge Graña, Fco. Mario Barcala, and Jesús Vilares. Formal methods of tokenization for part-of-speech tagging. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 240–249. Springer-Verlag, Berlin-Heidelberg-New York, 2002.

[88] Jorge Graña, Jean-Cédric Chappelier, and Manuel Vilares. Integrating external dictionaries into stochastic part-of-speech taggers. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nocolov, and Nokolai Nikolov, editors, *EuroConference Recent Advances in Natural Language Processing. Proceedings*, pages 122–128, Tzigov Chark, Bulgaria, 2001.

[89] Jorge Graña Gil, Miguel A. Alonso Pardo, and Manuel Vilares Ferro. A common solution for tokenization and part-of-speech tagging: One-pass Viterbi algorithm vs. iterative approaches. In P. Sojka, I. Kopeček, and K. Pala, editors, *Text, Speech and Dialogue*, volume 2448 of *Lecture Notes in Computer Science*, pages 3–10. Springer-Verlag, Berlin-Heidelberg-New York, 2002.

[90] B. B. Greene and G. M. Rubin. Automatic grammatical tagging of English. Technical report, Departament of Linguistics, Brown University, Providence, Rhode Island, 1971.

[91] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, 1994.

[92] Gregory Grefenstette. Light parsing as finite state filtering. In *Workshop on Extended finite state models of language (ECAI'96), Budapest, Hungary, August 11–12 1996.*, pages 20–25, 1996.

[93] Gregory Grefenstette, editor. *Cross-Language Information Retrieval*, volume 2 of *The Kluwer International Series on Information Retrieval*. Kluwer Academic Publishers, 1998.

[94] Gregory Grefenstette, Anne Schiller, and Salah Aït-Mokhtar. Recognizing lexical patterns in text. In Frank Van Eynde and Dafydd Gibbon, editors, *Lexicon Development for Speech and Language Processing*, volume 12 of *Text, Speech and Language*, pages 141–168. Kluwer Academic Publishers, Dordrecht/Boston/London, 2000.

[95] Gregory Grefenstette and Pasi Tapanainen. What is a word, what is a sentence? Problems of tokenization. In *3rd International Conference on Computational Lexicography and Text Research, COMPLEX'94, July 7-10, Budapest, Hungary*, pages 79–87, 1994.

[96] Ralph Grishman. The NYU system for MUC-6 or where's the syntax? In *Proc. of the Sixth Message Understanding Conference*. Morgan Kaufmann Publishers, 1995.

[97] Derek Gross and Katherine J. Miller. Adjectives in WordNet. *International Journal of Lexicography*, 3(4):265–277, 1990.

[98] D. K. Harman. Overview of the first text retrieval conference. In D. K. Harman, editor, *NIST Special Publication 500-207: The First Text REtrieval Conference (TREC-1)*, pages 1–20, Gaithersburg, MD, USA, 1992.

[99] Donna Harman. Relevance feedback revisited. In *Proceedings of the 15th annual international ACM SIGIR conference on research and development in Information Retrieval (SIGIR'92), June 21-24, Copenhagen, Denmark*, pages 1–10. ACM Press, 1992.

[100] Z. S. Harris. *String Analysis of Sentence Structure*. Mouton, The Hague, The Netherlands, 1962.

[101] Marti Hearst, Jan Pedersen, Peter Pirolli, Hinrich Schutze, Gregory Grefenstette, and David Hull. Xerox site report: Four TREC-4 tracks. In Harman, editor, *The Fourth Text REtrieval Conference (TREC-4)*, pages 97–119. U.S. Dept. of Commerce, National Institute of Standards and Technology, 1996.

[102] Marti A. Hearst. TextTiling: segmentating text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[103] Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In Enmanuel Roche and Yves Schabes, editors, *Finite-State Language Processing*. MIT Press, Cambridge, MA, USA, 1997.

[104] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Series in Computer Science. Addison-Wesley Publishing Company, Reading, Massachussetts, USA, 1979.

[105] David A. Hull. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70–84, 1996.

[106] David A. Hull, Gregory Grefenstette, B. Maximilian Schulze, Eric Gaussier, Hinrich Schütze, and Jan O. Pedersen. Xerox TREC-5 site report: Routing, filtering, NLP, and Spanish tracks. In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-238: The Fifth Text REtrieval Conference (TREC-5)*, pages 167–180, Gaithersburg, MD, USA, 1997. Department of Commerce, National Institute of Standards and Technology.

[107] W. John Hutchings and Harold L. Somers. *An Introduction to Machine Translation*. Academic Press, London and San Diego, 1992.

[108] E. Ide. *The SMART Retrieval System - Experiments in Automatic Document Processing*, chapter New experiments in relevance feedback, pages 337–354. In Salton [199], 1971.

[109] Tiago Ildefonso and Gabriel P. Lopes. Longest sorted sequence algorithm for parallel text alignment. In Quesada-Arencibia et al. [180], pages 51–54.

[110] Peter Jackson and Isabelle Moulinier. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*, volume 5 of *Natural Language Processing*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2002.

[111] Christian Jacquemin. Syntagmatic and paradigmatic representations of term variation. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Proceedings*, pages 341–348, Maryland, 1999.

[112] Christian Jacquemin. *Spotting and Discovering Terms through Natural Language Processsing*. The MIT Press, Cambridge, Massachusetts, USA, 2001.

[113] Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the european Chapter of the Association for Computational Linguistics (ACL-EACL'97), Barcelona, 7-10 July*, pages 24–31, Madrid, 1997.

[114] Christian Jacquemin and Evelyne Tzoukermann. NLP for term variant extraction: synergy between morphology, lexicon and syntax. In Strzalkowski [229], pages 25–74.

[115] F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–557, 1976.

[116] F. Jelinek and J. D. Lafferty. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3):315–323, 1991.

[117] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[118] Karen Sparck Jones. Retrieval system tests 1958 — 1978. In Karen Sparck Jones, editor, *Information Retrieval Experiment*, pages 213–255. Butterworths, London, 1981.

[119] Aravind K. Joshi, Leon S. Levy, and M. Takahashi. Tree adjunt grammars. *Journal of Computer and System Sciences*, 10(1):136–162, February 1975.

[120] Aravind K. Joshi, K. Vijay-Shanker, and David Weir. The convergence of mildly context-sensitive grammar formalisms. In P. Sells, Shieber S. M., and T. Warsow, editors, *Foundational Issues in Natural Language Processing*, pages 31–81. MIT Press, Cambridge, MA, USA, 1991.

[121] Daniel Jurafsky and James H. Martin. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, New Jersey, 2000.

[122] Ronald M. Kaplan. The formal architecture of lexical-functional grammar. In Mary Darymple, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, editors, *Formal Issues in Lexical-Functional Grammar*. Stanford University, 1994.

[123] T. Kasami. An efficient recognition and syntax algorithm for context-free languages. Scientific Report AFCRL-65-758, Air Force Cambridge Research Lab., Bedford, Massachussetts, 1965.

[124] M. Kaszkiel and J. Zobel. Effective ranking with arbitrary passages. *Journal of the American Society of Information Science*, 52(4):344–364, 2001.

[125] Martin Kay. Functional grammar. In *BLS'79*, pages 142–158, Berkeley, CA, USA, 1979.

[126] M. Shamim Khan and Sebastian Khor. Enhanced web document retrieval using automatic query expansion. *Journal of the American Society for Information Science and Technology*, 55(1):29–40, 2004.

[127] Judith L. Klavans, Christian Jacquemin, and Evelyne Tzoukermann. A natural language approach to multiword conflation. In *Proceedings of the 3rd DELOS Workshop on Cross-Language Information Retrieval, 5-7 March, Zurich, Switzerland. European Research Consortium on Information Management (ERCIM)*, 1997.

[128] S. Klein and R. F. Simmons. A computational approach to grammatical coding of English words. *Journal of the Association for Computing Machinery*, 10(3):334–347, 1963.

[129] Kimmo Koskenniemi. Two-level morphology: A general computational model for wordform recognition and production. Technical Report 11, University of Helsinki, Department of General Linguistics, 1983.

[130] Cornelis H. A. Koster. Head/modifier frames for information retrieval. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2945 of *Lecture Notes in Computer Science*, pages 420–432. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[131] Gerald Kowalski. *Information Retrieval Systems: Theory and Implementation.* The Kluwer international series on Information Retrieval. Kluwer Academic Publishers, Boston-Dordrecht-London, 1997.

[132] Wessel Kraaij and Renée Pohlmann. Comparing the effect of syntactic vs. statistical phrase indexing strategies for Dutch. In Christos Nicolaou and Constantine Stephanidis, editors, *Research and Advanced Technology for Digital Libraries*, volume 1513 of *Lecture Notes in Computer Science*, pages 605–614. Springer-Verlag, Berlin/Heidelberg/New York, 1998.

[133] Byung-Kwan Kwak, Jee-Hyub Kim, Geunbae Lee, and Jung Yun Seo. Corpus-based learning of compound noun indexing. In J. Klavans and J. Gonzalo, editors, *Proc. of the ACL'2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, October 8, Hong Kong*, 2000.

[134] F.W. Lancaster. *Information Retrieval Systems: Characteristics, Testing and Evaluation.* John Wiley & Sons, New York, 1968.

[135] Mervyn F. Lang. *Formación de palabras en español: morfología derivativa productiva en el léxico moderno.* Cátedra, Madrid, 1992.

[136] Joon Ho Lee. Analyses of multiple evidence combination. In *Proc. of SIGIR '97, July 27-31, Philadelphia, PA, USA*, pages 267–276. ACM Press, 1997.

[137] David D. Lewis. The TREC-4 filtering track. In D.K. Harman, editor, *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, pages 165–181, Gaithersburg, MD, USA, 1995. Department of Commerce, National Institute of Standards and Technology.

[138] F. Llopis. *IR-n: Un sistema de Recuperación de Información basado en Pasajes.* PhD thesis, Universidad de Alicante, 2003.

[139] Julie B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1–2):22–31, 1968.

[140] H.P. Luhn. The automatic creation of literature abstracts. *IBM journal of research and development*, 2(2):159–165, April 1958.

[141] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge (Massachusetts) and London (England), 1999.

[142] Solomon Marcus, Carlos Martín-Vide, and Gheorghe Păun. Contextual Grammars as generative models of natural languages. *Computational Linguistics*, 24(2):245–274, June 1998.

[143] A. A. Markov. Essai d'une recherche statistique sur le texte du roman "Eugene Onegin" illustrant la liaison des epreuve en chain. *Izvistia Imperatorskoi Akademi Nauk (Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg)*, 7:153–162, 1913.

[144] Nuno M.C. Marques and J. Gabriel P. Lopes. Tagging with small training corpora. In F. Hoffmann, D. Hand, N. Adams, D. Fisher, and G. Guimaraes, editors, *Advances in Intelligent Data Analysis*, volume 2189 of *Lecture Notes in Computer Science*, pages 62–72. Springer-Verlag, Berlin-Heidelberg-New York, 2001.

[145] I. Marshall. Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB corpus. *Computer and the Humanities*, 17:139–150, 1983.

[146] Carlos Martín-Vide, Victor Mitrana, and Gheorghe Păun, editors. *Formal Languages and Applications*, volume 148 of *Studies in Fuzziness and Soft Computing*. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[147] M. Masterman. The thesaurus in syntax and semantics. *Machine Translation*, 4(1):1–2, 1957.

[148] Yuji Matsumoto and Takehito Utsuro. *Lexical Knowledge Acquisition*, chapter 24, pages 563–610. In Dale et al. [59], 2000.

[149] Sharon McDonald and John Tait. Search strategies in content-based image retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 80–87. ACM Press, 2003.

[150] I. A. Mel'čuk. *Studies in Dependency Syntax*. Karoma Publishers, Ann Arbor, 1979.

[151] Rada Mihalcea. Diacritics restoration: Learning from letters versus learning from words. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 339–348. Springer-Verlag, Berlin-Heidelberg-New York, 2002.

[152] Andrei Mikheev. A knowledge-free method for capitalized word disambiguation. In *Proceedings of the 37th Annual Meeting of the ACL, 20-26 June, Maryland, USA*, pages 159–166, 1999.

[153] Andrei Mikheev. Document centered approach to text normalization. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR'00), Athens, Greece*, pages 136–143, 2000.

[154] Andrei Mikheev. Tagging sentence boundaries. In *Proceedings of the NAACL 2000, 1st Meeting of the North American Chapter of the Association for Computational Linguistics, April 29 - May 4, Seattle, Washington, USA*, pages 264–271, 2000.

[155] Andrei Mikheev. Periods, capitalized words, etc. *Computational Linguistics*, 28(3):289–318, September 2002.

[156] George A. Miller. Nouns in Wordnet: A lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264, 1990.

[157] George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[158] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[159] Ruslan Mitkov. *Anaphora Resolution.* Pearson Education, Harlow, UK, 2002.

[160] Ruslan Mitkov, editor. *The Oxford handbook of computational linguistics.* Oxford University Press, 2003.

[161] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *Proceedings of RIAO-97, 5th International Conference "Recherche d'Information Assistee par Ordinateur", June 25-27, Montreal, Canada*, pages 200–214, 1997.

[162] Markus Mittendorfer and Werner Winiwarter. A simple way of improving traditional IR methods by structuring queries. In *Proc. of the 2001 IEEE International Workshop on Natural Language Processing and Knowledge Engineering (NLPKE 2001) , October 7-10, Tucson, Arizona, USA*, 2001.

[163] Markus Mittendorfer and Werner Winiwarter. Exploiting syntactic analysis of queries for information retrieval. *Data & Knowledge Engineering*, 42(3):315–325, 2002.

[164] Enrique Méndez, Jesús Vilares, and David Cabrero. Cole at CLEF 2004: Rapid prototyping of a QA system for Spanish. In Peters and Borri [176], pages 413–418.

[165] Enrique Méndez, Jesús Vilares, and David Cabrero. COLE experiments at QA@CLEF 2004 Spanish monolingual track. In C. Peters, P.D. Clough, G.J.F. Jones, J. Gonzalo, M.Kluck, and B.Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, Lecture Notes in Computer Science. Springer-Verlag, Berlin-Heidelberg-New York, 2005.

[166] Richard Montague. The proper treatment of quantification in ordinary English. In R. Thomason, editor, *Formal Philosophy: Selected Papers of Richard Montague*, pages 247–270. Yale University Press, New Haven, CT, USA, 1973.

[167] Manuel Montes-y-Gómez, Aurelio López-López, and Alexander Gelbukh. Information retrieval with conceptual graph matching. In M. Ibrahim, J. Küng, and N. Revell, editors, *Database and Expert Systems Applications*, volume 1873 of *Lecture Notes in Computer Science*, pages 312–321. Springer-Verlag, Berlin/Heidelberg/New York, 2000.

[168] Masumi Narita and Yasushi Ogawa. The use of phrases from query texts in information retrieval. In *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–320, Athens, Greece, 2000.

[169] Carol Neidle. Lexical functional grammars. In *Encyclopaedia of Language and Linguistics*. Pergamon Press, New York, NY, USA, 1994.

[170] David D. Palmer. *Tokenisation and Sentence Segmentation*, chapter 2. In Dale et al. [59], 2000.

[171] Fernando C.N. Pereira and David H.D. Warren. Definite clause grammars for language analysis - a survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence*, 1980.

[172] Jose Perez-Carballo and Tomek Strzalkowski. Natural language information retrieval: progress report. *Information Processing and Management*, 36(1):155–178, 2000.

[173] Carol Peters, editor. *Results of the CLEF 2001 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2001 Workshop, 3 September, Darmstadt, Germany*, 2001.

[174] Carol Peters, editor. *Results of the CLEF 2002 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2002 Workshop, 19-20 September, Rome, Italy*, 2002.

[175] Carol Peters and Francesca Borri, editors. *Results of the CLEF 2003 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2003 Workshop, 21-21 August, Trondheim, Norway*, 2003.

[176] Carol Peters and Francesca Borri, editors. *Results of the CLEF 2004 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, UK*, 2004.

[177] Ulrich Pfeifer, Thomas Poersch, and Norbert Fuhr. Retrieval effectiveness of proper name search methods. *Information Processing and Management*, 32(6):667–679, 1996.

[178] Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. The University of Chicago Press, Chicago & London, 1994.

[179] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[180] A. Quesada-Arencibia, R. Moreno-Díaz, and J.C. Rodríguez, editors. *Extended abstracts of the 10th Internacional Workshop on Computer Aided Systems Theory (EUROCAST 2005). Las Palmas de Gran Canaria, Spain, February 2005*. IUCTC, Universidad de Las Palmas de Gran Canaria, 2005.

[181] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey, 1996.

[182] Francisco J. Ribadas. *Reconocimiento de patrones en bosques compartidos*. PhD thesis, Departamento de Computación, Universidade da Coruña, A Coruña, Spain, November 2000.

[183] Francisco J. Ribadas, Jesús Vilares, and Miguel A. Alonso. Integrating syntactic information by means of data fusion techniques. In Quesada-Arencibia et al. [180], pages 96–99.

[184] António Ribeiro, Gabriel P. Lopes, and João Mexia. Extracting translation equivalents from Portuguese-Chinese parallel texts. *Studies in Lexicography*, 11(1):181–194, 2001.

[185] Elaine Rich. *Artificial intelligence*. McGraw-Hill, Inc., 1983.

[186] Kelly Roach. Formal properties of Head Grammars. In Alexis Manaster-Ramer, editor, *Mathematics of Language*, pages 293–347. John Benjamins Publishing Company, Amsterdam/Philadephia, 1987.

[187] S.E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, (33):126–148, 1977.

[188] S.E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.

[189] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Sciences*, (27):129–146, May–June 1976.

[190] S.E. Robertson, M.E. Maron, and W.S. Cooper. Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Development*, (1):1–21, 1982.

[191] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241. Springer-Verlag New York, Inc., 1994.

[192] S.E. Robertson, S. Walker, and M.M. Hancock-Beaulieu. Large test collection experiments on an operational, interactive system: Okapi at TREC. *Information Processing and Management*, 31(3):345–360, 1995.

[193] S.E. Robertson, S. Walker, S. Jones, and M.M. Hancock-Beaulieu. Okapi at TREC-3. In D. K. Harman, editor, *NIST Special Publication 500-225: Overview of the Third Text REtrieval Conference (TREC 3)*, pages 109–126, Gaithersburg, MD, USA, 1995. Department of Commerce, National Institute of Standards and Technology.

[194] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-2. In D. K. Harman, editor, *NIST Special Publication 500-215: The Second Text REtrieval Conference (TREC-2)*, pages 21–34, Gaithersburg, MD, USA, 1993.

[195] J.A. Robinson. A machine-oriented logic based on resolution principle. *Journal of the ACM*, 12(1):23–49, Enero 1965.

[196] J.J. Rocchio. *The SMART Retrieval System - Experiments in Automatic Document Processing*, chapter Relevance feedback in information retrieval, pages 313–323. In Salton [199], 1971.

[197] Vitor Rocio. *Infra-estrutura para superação de falhas no processamento sintáctico de textos*. PhD thesis, Universidade Nova de Lisboa, 2002.

[198] Vitor Rocio and Gabriel P. Lopes. *Análise sintáctica parcial em cascata*, pages 235–251. Ediçoes Colibri/Associação Portuguesa de Linguística, 1999.

[199] G. Salton, editor. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[200] G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[201] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(1):288–197, 1990.

[202] G. Salton and M.E. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36, January 1968.

[203] Gerard Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, Massachusetts, USA, 1989.

[204] Gerard Salton. The SMART document retrieval project. In *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 356–358. ACM Press, 1991.

[205] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[206] Christer Samuelsson. Morphological tagging based entirely on bayesian inference. In Robert Eklund, editor, *Proceedings of the 9th Scandinavian Conference on Computational Linguistics (NODALIDA-93), Stockholm, Sweden*, pages 225–238, 1993.

[207] Christer Samuelsson, Pasi Tapanainen, and Atro Voutilainen. Inducing constraint grammars. In *ICGI '96: Proceedings of the 3rd International Colloquium on Grammatical Inference*, volume 1147 of *Lecture Notes in Computer Science*, pages 146–155. Springer-Verlag, Berlin-Heidelberg-New York, 1996.

[208] Mark Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, Dublin, Ireland*, pages 49–57. ACM Press, 1994.

[209] Octavio Santana, Francisco Javier Carreras, José Rafael Pérez, and Gustavo Rodríguez. Relaciones morfoléxicas sufijales del español. *Procesamiento del Lenguaje natural*, 30:1–73, March 2003.

[210] Tefko Saracevic and Paul Kantor. A study of information seeking and retrieving. III. Searchers, searches, overlap. *Journal of the American Society for Information Science*, 39(3):197–216, 1988.

[211] Jacques Savoy. Report on CLEF-2002 Experiments: Combining Multiple Sources of Evidence. In Peters [174], pages 31–46.

[212] Jacques Savoy. Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic Models for Effective Monolingual Retrieval. In Peters and Borri [175], pages 179–188.

[213] Yves Schabes and Aravind K. Joshi. An Earley-type parsing algorithm for tree adjoining grammars. In *Proc. of 26th Annual Meeting of the Association for Computational Linguistics*, pages 258–269, Buffalo, NY, USA, June 1988. ACL.

[214] Roger C. Schank. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3(4):pages 532–631, 1972.

[215] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, 1994.

[216] C. E. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64, 1951.

[217] Klaas Sikkel. *Parsing Schemata — A Framework for Specification and Analysis of Parsing Algorithms*. Texts in Theoretical Computer Science — An EATCS Series. Springer-Verlag, Berlin/Heidelberg/New York, 1997.

[218] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96). Zurich, Switzerland*, volume 32, pages 21–29, 1996.

[219] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. *Information Processing & Management*, 32(5):619–633, 1996.

[220] Amit Singhal, John Choi, Donald Hindle, David D. Lewis, and Fernando Pereira. AT&T at TREC-7. In D. K. Harman, editor, *NIST Special Publication 500-242: Overview of the Seventh Text REtrieval Conference (TREC 7)*, pages 239–252, Gaithersburg, MD, USA, 1998. Department of Commerce, National Institute of Standards and Technology.

[221] Alan F. Smeaton, F. Kelledy, and R. O'Donnell. TREC-4 Experiments at Dublin City University: Thresholding Posting lists, Query Expansion with WordNet and POS Tagging of Spanish. In D. K. Harman, editor, *NIST Special Publication 500–236: The Fourth Text REtrieval Conference (TREC-4)*, pages 373–390, Gaithersburg, MD, USA, 1995. Department of Commerce, National Institute of Standards and Technology.

[222] Alan F. Smeaton and Ian Quigley. Experiments on using semantic distances between words in image caption retrieval. In H. Frei, Donna Harman, P. Schauble, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland*, pages 174–180. ACM Press, 1996.

[223] F. Sánchez, J. Porta, J.L. Sancho, A. Nieto, A. Ballester, A. Fernández, J. Gómez, L. Gómez, E. Raigal, and R. Ruiz. La anotación de los corpus CREA y CORDE. *Procesamiento del Lenguaje Natural*, 25:175–182, 1999.

[224] Richard Sproat. *Lexical Analysis*, chapter 3, pages 37–57. In Dale et al. [59], 2000.

[225] M. Steedman. Combinators and grammars. In R. Oehrle, E. Bach, and D Wheeler, editors, *Categorial Grammars and Natural Language Structures*, pages 417–442. Foris, Dordrecht, 1986.

[226] Mark Stevenson. *Word Sense Disambiguation: The Case for Combinations of Knowledge Sources*. Studies in Computational Linguistics. CSLI, Stanford, 2003.

[227] Christopher Stokoe, Michael P. Oakes, and John Tait. Word sense disambiguation in information retrieval revisited. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 159–166. ACM Press, 2003.

[228] A. Stolcke. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–202, 1995.

[229] Tomek Strzalkowski, editor. volume 7 of *Text, Speech and Language Technology*. Kluwer Academic Publishers, Dordrecht/Boston/London, 1999.

[230] Tomek Strzalkowski and Jose Perez-Carballo. Recent developments in natural language text retrieval. In D. K. Harman, editor, *NIST Special Publication 500-215: The Second Text REtrieval Conference (TREC-2)*, pages 123–136, Gaithersburg, MD, USA, 1994. Department of Commerce, National Institute of Standards and Technology.

[231] Richard F.E. Sutcliffe, Igal Gabbay, and Aoife O'Gorman. Cross-Language French-English Question Answering using the DLT System at CLEF 2003. In Peters and Borri [175], pages 373–378.

[232] Paul Thompson and Christopher C. Dozier. Name recognition and retrieval performance. In Strzalkowski [229], pages 261–272.

[233] Evelyne Tzoukermann, Judith Klavans, and Christian Jacquemin. Effective use of natural language processing techniques for automatic conflation of multi-word terms: The role of derivational morphology, part of speech tagging, and shallow parsing. In *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27-31, 1997, Philadelphia, PA, USA*. ACM, 1997.

[234] H. Uszkoreit. Categorial unification grammar. In *Proc. of COLING'86*, pages 187–194, Bonn, Germany, 1986.

[235] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2^{nd} edition, 1979.

[236] José L. Vicedo. *Recuperación de información de alta precisión: los sistemas de búsqueda de respuestas*. Number 2 in Colección de Monografías de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), 2003.

[237] Jesús Vilares, Miguel A. Alonso, and Francisco J. Ribadas. COLE experiments at CLEF 2003 Spanish monolingual track. In Peters and Borri [175], pages 197–206.

[238] Jesús Vilares, Miguel A. Alonso, and Francisco J. Ribadas. COLE experiments at CLEF 2003 Spanish monolingual track. In Carol Peters, Martin Braschler, Julio Gonzalo, and Martin Kluck, editors, *Advances in Cross-Language Information Retrieval*, Lecture Notes in Computer Science. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[239] Jesús Vilares, Miguel A. Alonso, Francisco J. Ribadas, and Manuel Vilares. COLE experiments at CLEF 2002 Spanish monolingual track. In Peters [174], pages 153–160.

[240] Jesús Vilares, Miguel A. Alonso, Francisco J. Ribadas, and Manuel Vilares. COLE experiments at CLEF 2002 Spanish monolingual track. In *Advances in Cross-Language Information Retrieval*, volume 2785 of *Lecture Notes in Computer Science*, pages 265–278. Springer-Verlag, Berlin-Heidelberg-New York, 2003.

[241] Jesús Vilares, Fco. Mario Barcala, Santiago Fernández, and Juan Otero. Manejando la variación morfológica y léxica en la recuperación de información textual. *Procesamiento del Lenguaje Natural*, 30:99–106, March 2003.

[242] Jesús Vilares, David Cabrero, and Miguel A. Alonso. Applying productive derivational morphology to term indexing of Spanish texts. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2004 of *Lecture Notes in Computer Science*, pages 336–348. Springer-Verlag, Berlin-Heidelberg-New York, 2001.

[243] Jesús Vilares, David Cabrero, and Miguel A. Alonso. Generación automática de familias morfológicas mediante morfología derivativa productiva. *Procesamiento del Lenguaje Natural*, 27:181–188, September 2001.

[244] Manuel Vilares, Juan Otero Fco. Mario Barcala, and Eva Domínguez. Automatic spelling correction in Galician. In José Luis Vicedo, Patricio Martínez-Barco, Rafael Muñoz, and Maximiliano Saiz-Noeda, editors, *Computational Linguistics and Intelligent Text Processing*, volume 3230 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[245] Manuel Vilares, Víctor M. Darriba, and Jesús Vilares. Análisis sintáctico de sentencias incompletas. *Procesamiento del Lenguaje Natural*, 30:107–113, March 2003.

[246] Manuel Vilares, Víctor M. Darriba, and Jesús Vilares. Parsing incomplete sentences revisited. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2945 of *Lecture Notes in Computer Science*, pages 102–111. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[247] Manuel Vilares, Víctor M. Darriba, Jesús Vilares, and Francisco J. Ribadas. A formal frame for robust parsing. *Theoretical Computer Science*, 328:171–186, 2004.

[248] Manuel Vilares, Víctor M. Darriba, Jesús Vilares, and Leandro Rodríguez. Robust parsing using dynamic programming. In O. H. Ibarra and Z. Dang, editors, *Implementation and Application of Automata*, volume 2759 of *Lecture Notes in Computer Science*, pages 258–267. Berlin-Heidelberg-New York, 2003.

[249] Manuel Vilares, Jorge Graña, and Pilar Alvariño. Finite-state morphology and formal verification. *Journal of Natural Language Engineering, special issue on Extended Finite State Models of Language*, 3(4):303–304, 1997.

[250] Manuel Vilares, Juan Otero, and Jorge Graña. On asymptotic finite-state error repair. In Alberto Apostolico and Masssimo Melucci, editors, *String Processing and Information Retrieval*, volume 3246 of *Lecture Notes in Computer Science*, pages 271–272. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[251] Manuel Vilares, Juan Otero, and Jorge Graña. Regional finite-state error repair. In *Implementation and Application of Automata*, Lecture Notes in Computer Science. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[252] Manuel Vilares, Juan Otero, and Jorge Graña. Regional vs. global finite-state error repair. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 3406 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin-Heidelberg-New York, 2005.

[253] Manuel Vilares, Juan Otero, and Jorge Graña. Spelling correction on technical documents. In Quesada-Arencibia et al. [180], pages 73–76.

[254] Manuel Vilares, Francisco J. Ribadas, and Victor M. Darriba. Approximate pattern matching in shared-forest. In Mohamed T. Ibrahim, Josef Küng, and Norman Revell, editors, *Database and Expert Systems Applications*, pages 322–333, 2000.

[255] Manuel Vilares, Francisco J. Ribadas, and Victor M. Darriba. Approximate VLDC pattern matching in shared-forest. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 483–494, 2001.

[256] Manuel Vilares, Francisco J. Ribadas, and Jesús Vilares. Phrase similarity through the edit distance. In Fernando Galindo, Makoto Takizawa, and Roland Traunmüller, editors, *Database and Expert Systems Applications*, volume 3180 of *Lecture Notes in Computer Science*, pages 306–317. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[257] Manuel Vilares, Jesús Vilares, and David Cabrero. Dynamic programming of partial parses. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nocolov, and Nokolai Nikolov, editors, *EuroConference Recent Advances in Natural Language Processing. Tzigov Chark, Bulgaria, 5-7 September 2001*, pages 291–293, 2001.

[258] Jesús Vilares Ferro, Fco. Mario Barcala Rodríguez, and Miguel A. Alonso Pardo. Using syntactic dependency-pairs conflation to improve retrieval performance in Spanish. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 381–390. Springer-Verlag, Berlin-Heidelberg-New York, 2002.

[259] Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Inform. Theory*, 13:260–269, 1967.

[260] E. M. Voorhees and D. K. Harman. Overview of the 6th Text REtrieval Conference (TREC-6). In E. M. Voorhees and D. K. Harman, editors, *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC-6)*, pages 1–24, Gaithersburg, MD, USA, 1997. Department of Commerce, National Institute of Standards and Technology.

[261] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 61–69, Dublin, Ireland, July 1994. ACM.

[262] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 315–323. ACM, 1998.

[263] Piek Vossen, editor. *EuroWordNet. A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998. Reprinted from Computers and the Humanities, Volume 32, Nos. 2–3, 1998.

[264] A. Voutilainen and J. Heikkilä. An English constraint grammar (EngCG): a surface-syntactic parser of English. In Fries, Tottie, and Schneider, editors, *Creating and using English language corpora*. Rodopi, 1994.

[265] Joseph Weizenbaum. Eliza – a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[266] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes*. Morgan Kaufmann, 1999.

[267] W. A. Woods. *Semantics for a Question-Answering System*. PhD thesis, Harvard University, 1967.

[268] J. Xu and W.B. Croft. Query expansion using local and global analysi. In *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96). Zurich, Switzerland*, volume 32, pages 4–11, 1996.

[269] David Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proc. of the 32nd Annual Meeting of the ACL, 27 June – 1 July, Las Cruces, New Mexico, USA*, pages 88–95, 1994.

[270] David Yarowsky. A comparison of corpus-based techniques for restoring accents in Spanish and French text. In *Natural Language Processing Using Very Large Corpora*, pages 99–120. Kluwer Academic Publishers, 1999.

[271] D. H. Younger. Recognition and parsing of context-free languages in time n^3. *Information and Control*, 10(2):189–208, 1967.

[272] G. Zipf. *Human behaviour and the principle of least effort*. Addison-Wesley, 1949.

Índice alfabético