Editores
Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Número 2 - Dezembro 2009

linguaMATICA

ISSN: 1647-0818

Universidade de Vigo
Universidade do Minho
25 Anos Associação Portuguesa Para a Inteligência Artificial

Número 2 – Dezembro 2009

LinguaMÁTICA

ISSN: 1647–0818

Editores

Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Conteúdo

Editorial

Os editores queren aproveitar a publicación deste segundo número de Linguamática para agradecer publicamente o labor dos membros do Comité Científico da revista, en especial dos que se incorporaron á nosa andaina recentemente (Aline Villavicencio, Ana Frankenberg-Garcia, Anselmo Peñas, Ferran Pla, Gerardo Sierra, Helena de Medeiros Caseli, Horacio Saggion, José Carlos Medeiros, Pablo Gamallo Otero, Susana Afonso Cavadas e Tony Berber Sardinha), mais tamén dos que xa levan connosco desde as orixes do proxecto (Alberto Álvarez Lugrís, Álvaro Iriarte Sanroman, Antón Santamarina, António Teixeira, Belinda Maia, Carmen García Mateo, Diana Santos, Gael Harry Dias, Iñaki Alegria, Joaquim Llisterri, José Paulo Leal, Joseba Abaitua, Lluís Padró, Maria Antònia Martí Antonín, Maria das Graças Volpe Nunes, Mercè Lorente Casafont, Mikel Forcada, Nieves R. Brisaboa e Salvador Climent Roca).

Linguamática é unha revista aberta sobre procesamento de linguaxes naturais, con especial atención ás linguas faladas na Peninsula Ibérica, como o portugués, o galego, o catalán, o español, o vasco, o mirandés ou o aranés. Os artigos publicados neste segundo número da revista tratan diversos aspectos do PLN das linguas catalá, española e portuguesa, incluíndo un artigo convidado para presentar monograficamente a investigación en extracción automática de definicións levada a cabo polo Grupo de Ingeniería Lingüística do Instituto de Ingeniería da Universidad Nacional Autónoma de México.

Canto á lingua usada para a escrita dos artigos, se no primeiro número da revista as linguas utilizadas foron o portugués (4 artigos), o galego (2 artigos) e o catalán (1 artigo), neste segundo número as linguas de redacción dos artigos son o español (4 artigos), o portugués (1 artigo) e o inglés (1 artigo). Linguamática suxire e recomenda o uso de portugués, galego, castelán ou catalán como lingua de redacción dos artigos enviados, mais non rexeita os artigos escritos en inglés sempre que os seus contidos concorden cos obxectivos da revista e os autores non sexan falantes nativos de ningunha das linguas recomendadas polos editores.

Finalmente, os editores queren manifestar o seu agradecemento a todas as persoas que contribuíron a esta publicación cos seus artigos e aos revisores e revisoras que leron e comentaron os traballos enviados.

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Comissão Científica

Alberto Álvarez Lugrís, Universidade de Vigo
Alberto Simões, Universidade do Minho
Aline Villavicencio, Universidade Federal do Rio Grande do Sul
Álvaro Iriarte Sanroman, Universidade do Minho
Ana Frankenberg-Garcia, ISLA e Universidade Nova de Lisboa
Anselmo Peñas, Universidad Nacional de Educación a Distancia
Antón Santamarina, Universidade de Santiago de Compostela
António Teixeira, Universidade de Aveiro
Belinda Maia, Universidade do Porto
Carmen García Mateo, Universidade de Vigo
Diana Santos, SINTEF ICT
Ferran Pla, Universitat Politècnica de València
Gael Harry Dias, Universidade Beira Interior
Gerardo Sierra, Universidad Nacional Autónoma de México
Helena de Medeiros Caseli, Universidade Federal de São Carlos
Horacio Saggion, University of Sheffield
Iñaki Alegria, Euskal Herriko Unibertsitatea
Joaquim Llisterri, Universitat Autònoma de Barcelona
José Carlos Medeiros, Porto Editora
José João Almeida, Universidade do Minho
José Paulo Leal, Universidade do Porto
Joseba Abaitua, Universidad de Deusto
Lluís Padró, Universitat Politècnica de Catalunya
Maria Antònia Martí Antonín, Universitat de Barcelona
Maria das Graças Volpe Nunes, Universidade de São Paulo
Mercè Lorente Casafont, Universitat Pompeu Fabra
Mikel Forcada, Universitat d'Alacant
Nieves R. Brisaboa, Universidade da Coruña
Pablo Gamallo Otero, Universidade de Santiago de Compostela
Salvador Climent Roca, Universitat Oberta de Catalunya
Susana Afonso Cavadas, University of Sheffield
Tony Berber Sardinha, Pontifícia Universidade Católica de São Paulo
Xavier Gómez Guinovart, Universidade de Vigo

Dossier

Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos

Gerardo Sierra
Universidad Nacional Autónoma de México
gsierram@ii.unam.mx

Resumen

La extracción automática de definiciones a partir de textos de especialidad es una tarea cada vez más demandante para diferentes aplicaciones del Procesamiento de Lenguaje Natural, tales como lexicografía computacional, extracción de información, semántica computacional, sistemas pregunta-respuesta, minería de textos, Web semántica y aprendizaje automático. Este artículo presenta un panorama de los trabajos realizados en el Grupo de Ingeniería Lingüística en el tema, desde los aspectos teóricos, la revisión del estado del arte, los estudios lingüísticos sobre definiciones y contextos definitorios, la metodología para la extracción automática y hasta diversas aplicaciones.

1. Introducción

Este artículo constituye una síntesis de la investigación realizada en el Grupo de Ingeniería Lingüística del Instituto de Ingeniería, UNAM, referente a la extracción automática de contextos definitorios en textos de especialidad en español, mediante el reconocimiento y análisis de patrones lingüísticos.

Esta investigación surge como parte de un proyecto central, que constituye la metodología para la creación de diccionarios onomasiológicos [23]. Para construir un diccionario de esta naturaleza, se debe contar con una base de conocimientos léxica lo suficientemente rica que contenga una diversidad de definiciones para cada uno de los términos que se están buscando. Para la obtención de dichas definiciones, además de las disponibles en los diccionarios, se puede acudir a los textos de especialidad, tales como artículos, reportes, tesis, etc., en donde los autores introducen los términos y, por tanto, proporcionan su definición. En este sentido, la investigación está orientada a la extracción automática de estas unidades del discurso utilizadas en los textos de especialidad donde se introduce un término y su definición, lo que aquí denominamos *contextos definitorios* (CDs).

Como parte del estado del arte, en la sección 2 tendremos una introducción a la extracción de información, en particular de la información terminológica y conceptual, campos donde se contextualiza esta investigación. Veremos las bases para el desarrollo de los sistemas de extracción de esta información y los trabajos realizados en la materia. Con ello concretaremos, en la sección 3, el concepto de CD para la terminología y específicamente para su extracción.

En la parte más descriptiva, iremos viendo que la base para la extracción radica en los patrones definitorios, los cuales detallaremos en la sección 4. En la sección 5 nos concretaremos en la definición, su tipología y el papel que juegan las predicaciones verbales en los tipos de definición. En la sección 6 continuaremos precisando los CDs sobre su extensión como unidad discursiva.

Con todos estos elementos, ya en la sección 7 veremos la metodología desarrollada para nuestro extractor de CDs. El extractor tiene como entrada una lista de patrones verbales definitorios y como salida los CDs clasificados por los tipos de definición. En la sección 8 veremos otra forma de agrupar los CDs por sus características semánticas, con lo que se mejora su extracción.

Como parte más aplicada, en la sección 9 describiremos el corpus utilizado a lo largo de nuestras investigaciones y la evaluación de algunos resultados obtenidos. En la sección 10 encontraremos como un resultado concreto la descripción del Corpus de Contextos Definitorios. Luego seguiremos, en la sección 11, con tres aplicaciones específicas de la utilización de CDs dentro del Grupo de Ingeniería Lingüística, entre las que encontramos el banco de conocimientos léxico para el diccionario onomasiológico y un sistema que aplica los resultados de esta investigación para realizar búsquedas en línea.

Cabe mencionar que la investigación en su conjunto forma parte de varios proyectos que han culminado en la publicación de algunos artículos y en diversas tesis, desde licenciatura hasta doctorado, en las áreas de lingüística, ingeniería de la computación y lingüística computacional.

Tendremos un reconocimiento a quienes han contribuido con sus estudios particulares y, para finalizar, las referencias, tanto las publicadas a lo largo de la investigación, como las que han sido base para el desarrollo de la misma.

2. La extracción de información terminológica y conceptual

Una de las áreas que dentro de la inteligencia artificial ha tenido un gran desarrollo en los últimos años, es la que se refiere al diseño de sistemas automáticos de extracción de información (EI). Este proceso, como señala Wilks [83], puede ser visto como el núcleo principal de las actuales tecnologías del lenguaje, de ahí que resulte necesario contar con sistemas de cómputo capaces de buscar, localizar y brindar información relevante de cualquier tipo a un usuario.

Se puede definir entonces a la EI como un proceso por el cual un sistema de cómputo busca de manera selectiva una serie de estructuras o combinaciones de datos, los cuales se encuentran, de manera explícita o implícita, dentro de un conjunto de textos. El resultado de lo anterior es la obtención de información específica que proporciona un conocimiento asociado a tales estructuras o combinaciones de datos [32].

En paralelo con la EI, se ha venido desarrollando un área de investigación enfocada al diseño de sistemas de cómputo capaces de generar y administrar conocimiento obtenido a partir de datos, tal área ha sido denominada *ingeniería del conocimiento* (IC). Una de los objetivos centrales de la IC, es la elaboración de *bases de conocimiento* (BC), las cuales funcionen como un repositorio organizado de información relevante susceptible de proporcionar conocimiento específico a un usuario sobre algún hecho dado.

Entre los aspectos que ha tomado gran relevancia dentro de la IC, cabe señalar la creación de sistemas de *extracción de información terminológica* y *conceptual* (EITC), proyectados para la elaboración de ontologías y diccionarios electrónicos. La generación de estos recursos es uno de los campos más relevantes en los cuales se ha aplicado la IC, en colaboración con otras disciplinas tales como la terminología y la lingüística [36].

La EITC puede ser definida como un conjunto de métodos y recursos tecnológicos orientados a la búsqueda, localización, almacenamiento y administración de términos y conceptos obtenidos de bases de textos relacionadas con un área de especialidad (ingeniería, computación, administración de empresas, periodismo, etc.). La información que se genera a partir de estas bases de texto permite diseñar glosarios, vocabularios y diccionarios electrónicos, herramientas para la traducción automática, sistemas de clasificación e indexación de textos, desarrollo de sistemas expertos y apoyo para labores terminológicas, y otros [36]. Por ello, de acuerdo con Jacquemin y Bourigault [45], se puede ver a la EICT como un área de investigación y aplicación particular y sumamente productiva de la EI.

2.1 Extracción de términos y de conceptos

Si bien existen métodos que han dado buenos resultados para los procesos de extracción terminológica [14, 30, 43], en el caso de la extracción de conceptos resulta un reto más complejo, debido sobre todo a la riqueza de relaciones que se dan a la hora de expresarlos en lengua natural. Las diferencias entre extracción terminológica y conceptual son en gran medida consecuencia de un cambio de paradigma entre una visión de índole normativa sostenida en el modelo propuesto por los diccionarios elaborados bajo los criterios de autoridades académicas [41] y una postura que tome en cuenta aspectos comunicativos y cognitivos subyacentes en la configuración de conceptos [28].

Para tener una distinción pertinente entre términos y conceptos, tomemos en cuenta que el término es una unidad de significación especializada, la cual cuenta con rasgos léxicos particulares (nombres, adjetivos, verbos o adverbios), capacidad referencial y nominativa concreta, así como un significado especializado en un dominio concreto [28, 35].

En contraste, un concepto puede ser visto como una unidad de conocimiento abstracto, la cual contiene una serie de rasgos, características o atributos propios de un objeto, un evento o una relación, con el fin de situarlo dentro del mundo [76]. Al nivel del lenguaje natural, esta unidad es representada por una definición [72, 84]. La definición, de acuerdo con la explicación tradicional de Aristóteles [56], se constituye a partir de dos elementos básicos: un *género próximo* y una *diferencia específica*. El género próximo o *genus* se entiende como un descriptor que hace referencia a la clase a la cual pertenece un objeto o evento, y la diferencia específica o *differentia* son la serie de rasgos propios que distinguen a dicho objeto o evento de los respecto a otros agrupados en su misma clase. En un nivel lingüístico, el género próximo se manifiesta, en el nivel sintáctico, a partir de unidades nominales tales como cuantificadores, determinantes o demostrativos; por su parte, la diferencia específica sería introducida por oraciones subordinadas compuestas

por frases nominales, frases adjetivas o frases prepositivas.

En el caso de los términos, su formación sintáctica implica sobre todo el uso de frases nominales, en particular nombres y adjetivos [43 44, 75], y en algunos casos, construcciones verbales en una función nominativa [49]. En las siguientes secciones veremos de qué manera estos rasgos lingüísticos marcan procesos de reconocimiento y extracción diferentes para términos y definiciones.

2.2 Marco teórico de EITC

Para el desarrollo de los sistemas de EITC es importante tomar en cuenta mecanismos de reconocimiento y extracción de información con determinadas características. En paralelo, de acuerdo con Jacquemin y Bourigault [45], deben considerarse los siguientes aspectos:

- Recopilación, organización y administración de corpus lingüísticos etiquetados, de modo que pueda reconocerse de manera automática ciertos patrones de datos (p.e., asociar términos con frases nominales).
- Implementación de bases de conocimiento léxico, las cuales permiten almacenar, administrar y suministrar conocimientos obtenidos del lenguaje natural, a partir de textos especializados.
- Diseño de sistemas de búsqueda, a partir del uso de lenguajes de programación lo suficientemente robustos como para hacer eficientes los procesos de reconocimiento de datos, la validación de los mismos y la adquisición de conocimiento.
- Empleo de métodos estadísticos, de modo que pueda evaluarse la eficacia o ineficacia de los sistemas de búsqueda de un modo formal. De este modo, el uso de esta clase de recursos es esencial para lograr un sistema de extracción óptimo y potente.
- Aplicación de modelos lingüísticos, los cuales brinden un marco teórico de interpretación pertinente para describir los patrones del lenguaje natural a buscar, así como su formalización de modo que puedan ser comprensibles para cualquier sistema de cómputo diseñado para esta clase de tareas.

Con base en estos puntos, es común que los sistemas de extracción de términos en corpus utilicen un método de aprendizaje automático que consiste en tomar en cuenta los patrones estructurales característicos que conforman tales términos [43]. Después de hacer una primera corrida en un conjunto de textos, con base en estos patrones previamente introducidos, los sistemas localizan y presentan una serie de candidatos posibles. Al final, el conjunto de candidatos es validado de forma manual por un grupo de expertos sobre el área de conocimiento a la cual pertenecen los textos, con miras a determinar cuán exitosa o no fue el proceso ejecutado por dicho sistema.

2.3 Trabajos en EITC

Un tipo de EITC en particular se ha enfocado a obtener información para la organización conceptual de unidades de conocimiento especializadas, así como para la descripción de sus significados. Este tipo de información terminológica suele denominarse *conocimiento definitorio* [7] y es un tipo de información que permite inferir el significado de los términos a partir de la descripción de sus atributos, características o relaciones semánticas [55]. Cabe distinguir dos tipos particulares de extracción automática de conocimiento definitorio.

Por un lado, la extracción de relaciones semánticas (p.e., hiperonimia, hiponimia, holonimia, meronimia, sinonimia, etc.), que en un principio se enfocaron en las definiciones obtenidas de diccionarios en formato electrónico [31, 67]. Posteriormente, buscaron extraer dichas relaciones de corpus lingüísticos tomando en cuenta patrones léxicos sintácticos [42] y luego mediante conceptos formales y el grado de subsunción [37].

Por otro lado, la extracción de contextos definitorios (CDs), con la cual no solo se permite recuperar relaciones semánticas específicas [20], sino también descripciones generales acerca del significado de los términos, y que pueden servir en la elaboración de diversos tipos de recursos terminológicos. A diferencia de la extracción de relaciones léxicas, la de contextos definitorios se realiza únicamente sobre corpus lingüísticos, no solo a partir de patrones léxicos sintácticos, sino de patrones tipográficos y pragmáticos, como veremos más adelante.

El estudio de Alarcón [7] presenta un estado del arte de la extracción de contextos definitorios, a la vez que realiza un análisis contrastivo de diez trabajos en este campo.

- Los trabajos de Rebeyrolle [68, 69] para el francés, que describen una metodología para la extracción de CDs a partir de patrones morfo-sintácticos y que presentan algunas consideraciones sobre la introducción de definiciones en textos de especialidad y el diseño de patrones para su extracción automática.
- El sistema DEFINDER desarrollado por Muresan y Klavans [58] para el inglés, con el fin de extraer definiciones de textos en-línea en el área de medicina mediante la búsqueda de patrones léxicos y tipográficos, en conjunto con una gramática de estados finitos.

- El trabajo de Saggion [73] para el inglés, enfocado a la extracción de definiciones para sistemas de pregunta-respuesta, usando una lista de 50 patrones definitorios.
- Los trabajos de extracción semi-automática de definiciones de la herramienta CORPÓGRAFO, para el alemán, español, inglés, italiano, francés y portugués. A partir de la extracción terminológica, cada término se combina con una serie de patrones definitorios típicos y se formulan así expresiones regulares de búsqueda [64].
- El estudio de Malaisé [52] para extraer lo que denominó definiciones formales, semi-formales e informales, para el francés, a partir de patrones léxicos, de la posición que guardan los términos con los patrones definitorios y de la categoría morfosintáctica de estos últimos.
- El trabajo aplicado de Sánchez y Márquez [74] para textos jurídicos en español, con el fin de extraer definiciones mediante la identificación de patrones verbales recurrentes.
- El estudio de Rodríguez [70], para el inglés, en el que mediante lo que denomina Operaciones Metalingüísticas Explícitas (OMEs), busca extraer unidades de conocimiento especializadas a partir de la detección de fragmentos metalingüísticos en textos de especialidad.
- El trabajo de Storrer y Wellinghoff [78], para el alemán, orientado a detectar y anotar automáticamente definiciones y sus componentes principales en textos técnicos a partir de verbos definitorios y patrones basados en la valencia de dichos verbos.
- El proyecto *Language Technology for eLearning* (LT4eL), patrocinado por la Unión Europea y coordinado por la Universidad de Utrecht, Holanda, en conjunto con 11 instituciones educativas. Una parte central del proyecto se enfocó en desarrollar metodologías para la extracción automática de definiciones para el alemán, búlgaro, checo, holandés, inglés, maltés, polaco, portugués y rumano, con el fin de proporcionar herramientas de ayuda en la elaboración de glosarios [57].
- La aplicación web para el inglés, GlossExtractor, de Navigli y Velardi [59], cuya función es extraer una lista de candidatos a definiciones sobre varios tipos de documentos en Internet.

De este estado del arte, Alarcón observó una similitud en las metodologías y consiste en que todas ellas parten de patrones definitorios para el reconocimiento automático de fragmentos con información definitoria. Resulta notable la coincidencia de usar patrones sintácticos y, en particular, la preferencia de los patrones verbales frente a construcciones sintácticas que incluyen palabras metalingüísticas pero no verbos. Asimismo, resaltó la coincidencia de recurrir no sólo a la búsqueda de patrones definitorios, sino también al uso de filtros de exclusión de contextos no relevantes, así como a la búsqueda y detección de los elementos constitutivos de los candidatos a CDs, es decir, los términos y las definiciones.

3. La noción de contexto definitorio

Para describir el concepto de CD, conviene retomar el estudio de Alarcón [7] sobre algunas aproximaciones de su uso en el ámbito de la terminología, lo cual nos servirá de base para entender lo que se pretende en la extracción automática.

3.1 Aproximaciones del concepto de CD en terminología

Alarcón establece, como punto de partida, lo que De Bessé [34] entiende por *contexto* y que constituye el punto de inicio de cualquier trabajo terminográfico. El contexto es el entorno lingüístico de un término conformado por un enunciado, es decir, las palabras o frases alrededor de dicho término, y que persigue dos funciones básicas: aclarar el significado de un término e ilustrar su funcionamiento. Por tanto, los contextos constituyen un elemento esencial para la descripción de un concepto y resultan indispensables para redactar una definición.

De Bessé distingue los CDs como aquellos contextos donde se aporta información sobre los atributos de los términos. Diferencia los contextos conceptuales como aquellos que se refieren a características sobre las relaciones conceptuales de los términos, en tanto los materiales proveen instrucciones sobre el alcance de los términos y la forma en que éstos operan en un contexto determinado.

Por su parte, Auger [26] divide los enunciados definitorios dependiendo de los tipos de verbos o formas lingüístico-sintácticas que se utiliza en ellos para vincular un término con su respectiva definición. Considera los *enunciados definitorios metalingüísticos* como los elementos que refieren al mismo lenguaje y que utilizan verbos o formas del tipo *llamarse*, *significar*, *el sustantivo*, *el sintagma*, etc. Los *enunciados definitorios lingüísticos*, por otro lado, son los que no se utilizan exclusivamente para referirse al propio lenguaje y se conforman por los verbos o formas lingüístico sintácticas que utilizan elementos del tipo *equivaler a*, *compuesto por*, *características*, *atributos*, etc.

Pearson [62] realiza también un estudio de cómo son empleadas las definiciones en las diversas situaciones comunicativas y describe la forma en que los actos performativos definitorios transmiten en mayor o menor grado cierto tipo de información metalingüística explícita o implícita, la cual provee datos sobre el contexto real de uso de los términos. Ciertamente, lo que clasifica Pearson no son todos los tipos de contextos de aparición de los términos, sino aquellos que incluyen un tipo específico de información definitoria. Dentro de este grupo de contextos clasifica dos clases generales, dependiendo de que la definición se presenta por primera vez, o, por el contrario, sea una reformulación de una definición previa.

Meyer [55] propone una categorización simple y más genérica de los tipos de contextos que contienen información conceptual. Meyer define los *contextos ricos en conocimiento* (CRCs) a aquellos contextos que indican por lo menos una característica conceptual del término, ya sea un atributo o una relación.

Con todo, cabe mencionar que las tipologías de Auger, Pearson y Meyer no incluyen una clasificación genérica sobre las clases de contextos que representan las ocurrencias simples de los términos, sino se ciernen a los contextos de textos especializados que informan sobre las características definitorias, conceptuales o metalingüísticas de un término.

3.2 El CD en el ámbito de la extracción de información

Con el objetivo de establecer las bases necesarias para la extracción automática de CDs, a lo largo de nuestra investigación hemos establecido que un CD es aquel fragmento textual donde se aporta información que permite comprender el significado de un término, de manera que la información contenida en el contexto puede proporcionar datos sobre sus características y atributos, así como funciones, partes o bien relaciones de éste con otros términos.

Así, delimitamos el prototipo de CD como la estructura discursiva conformada por dos elementos mínimos: un término (T) y una definición (D), los cuales se encuentran conectados entre sí mediante un *patrón definitorio* (PD). Además, los CDs pueden presentar otro tipo de información metalingüística y pragmática referente a la forma, las condiciones de uso o el alcance operativo de los términos. Dicha información corresponde a lo que denominamos un *patrón pragmático* (PPR). En el ejemplo 1 observamos la definición del término *logística*.

(Ej. 1) <PPR> Tradicionalmente </PPR>, <T> la logística </T> <PD> se define como </PD> <D> el arte militar que estudia el movimiento, transporte y estacionamiento de las tropas fuera del campo de batalla</D>.

Vemos que para conectar el término con la descripción de sus características distintivas (*el arte militar que estudia el movimiento…)*, el autor recurre al patrón definitorio que corresponde a la estructura *se define como*. Asimismo, podemos observar el patrón pragmático, *tradicionalmente*, que en este caso indica un matiz especial sobre el significado del término.

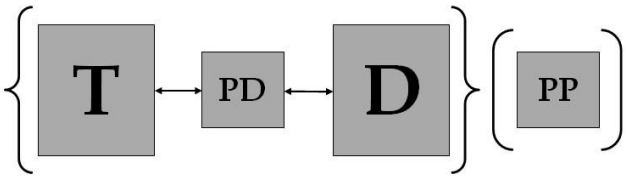

Figura 1: Estructura de un contexto definitorio

En resumen, representamos la estructura de los CDs con el esquema de la Figura 1, en donde los elementos mínimos constitutivos son el término T y la definición D junto con el patrón definitorio PD, que como unidad puede estar modificada por el elemento optativo PP.

3.2.1 Clasificación de CDs

Con base en amplias observaciones sobre la ocurrencia de CDs en diferentes tipos de textos, hemos realizado una clasificación de CDs, tomando en cuenta la presencia o ausencia de una serie de claves tipográficas y sintácticas recurrentes que se utilizan para conectar al término con la información definitoria que se introduce sobre ellos [18, 22].

CDs tipográficos. Los contextos más simples son aquellos que contienen sólo marcas tipográficas para unir al término con la definición, o bien cuya misma tipografía textual se usa para resaltar cualquiera de estos elementos. Este tipo de CDs ocurre tradicionalmente en diccionarios y glosarios, aunque, como refieren Pearson y Meyer, también es común encontrarlos en textos especializados.

(Ej. 2) Diseño: Desarrollo de configuraciones para la resolución de algún problema en base y sujetándose a sus restricciones.

(Ej. 3) *IMPACTOS AGREGADOS SOCIALES ¶* Los que impactan a la sociedad, produciendo, por ejemplo, la perturbación de las relaciones familiares.

En el ejemplo 2, el término *diseño* se presenta en negritas y se liga a la definición, en cursivas, mediante *dos puntos*. En el ejemplo 3, el término *impactos agregados sociales* se resalta en mayúsculas y cursivas, mientras que la liga a su definición se establece situando al término a modo

de título, seguido de un salto de párrafo que representamos con el símbolo ¶.

CDs sintácticos. Otro tipo de CDs igualmente simples son aquellos en donde el término se une a la definición mediante una estructura sintáctica, generalmente una frase verbal, aunque también es común encontrar marcadores reformulativos. En estos casos no se incluye ningún tipo de marca tipográfica para resaltar los elementos constitutivos de los CDs.

> **(Ej. 4)** De manera general, un Operador Logístico (OL) es una firma que realiza prestaciones logísticas en servicio público que adapta a necesidades específicas de cada cliente.
>
> **(Ej. 5)** Definimos un ramal como aquella sección del acueducto constituida por uno o más tubos interconectados y a lo largo de los cuales no existe derivación alguna, de manera que todos los tubos conducen un mismo caudal.

En estos dos ejemplos notamos que si bien no se recurre a la tipografía textual para resaltar la presencia del término o la definición, sí se utilizan otros patrones. El ejemplo 4 es prototípico del uso del verbo *ser* más un determinante, lo que se conoce como relación ISA, que aquí se usa para expresar la definición del término *operador logístico*. En el ejemplo 5 tenemos un caso en que para definir el término, *ramal*, se utiliza una estructura sintáctica formada por el verbo *definir* más el adverbio *como*.

CDs mixtos. Este tipo de patrones son una combinación de los dos anteriores, ya que se emplea una frase verbal o un marcador reformulativo como conector entre el término y la definición, pero además se resalta tipográficamente la presencia de cualquiera de estos dos elementos.

> **(Ej. 6) La energía primaria**, por definición, es aquel recurso energético que no ha sufrido transformación alguna, con excepción de su extracción.

En el ejemplo 6 observamos una estructura más sólida que en los ejemplos anteriores, pues aquí se utilizan elementos que permiten resaltar visual y gramaticalmente la presencia de un contexto con información definitoria.

CDs complejos. Estos representan los casos donde en un CD se definen dos o más términos.

> **(Ej. 7)** Por lo anterior, se llegan a distinguir dos tipos de sistemas interactuantes, responsables por la mayor parte de la problemática de desastres: el *afectable* y el *perturbador*. El primero, denominado **SA,** se define como el sistema donde pueden materializarse los desastres debido a la perturbación al que está expuesto; en términos generales, está integrado por la sociedad y los componentes que necesita para su subsistencia, incluyendo el medio ambiente; mientras que, en el contexto particular, puede ser una ciudad u obra civil. El otro, denominado **SP**, responsable por la perturbación, se define como el sistema capaz de producir calamidades, tales como sismos, incendios, explosiones, inundaciones y contaminación.

En 7 se muestra un tipo de casos que, si bien no ocurren en un gran porcentaje con respecto a los demás, nos permiten ver la complejidad de formas en que pueden introducirse CDs en textos especializados. En el párrafo podemos hallar dos términos: sistema afectable (SA) y sistema perturbador (SP), los cuales aunque no aparecen explícitos, se encuentran resaltados en cursivas y en negritas. Asimismo, encontramos la presencia de estructuras sintácticas que nos permiten inferir los términos, la relación entre ellos y las definiciones dadas por el autor. Aquí tenemos un caso claro de referencias anafóricas, la cual veremos a detalle más adelante.

4. Tipología de patrones definitorios

Un elemento clave en el proceso para reconocer CDs de forma automática lo constituye la identificación de los patrones que se emplean para conectar al término con su definición o para resaltar visualmente su presencia dentro del texto. Entre los elementos de CDs mencionamos estos patrones, llamados *patrones definitorios*.

Encontramos dos clases generales de patrones definitorios: los tipográficos y los sintácticos. En los últimos, y de acuerdo con los elementos que se presenten en el patrón, podemos encontrar patrones verbales y/o marcadores reformulativos. Recordemos que, con base en la clasificación de CDs, estos patrones no son excluyentes, puesto que pueden darse por separado o en conjunto.

4.1 Patrones tipográficos

La tipografía de un texto es un recurso que sirve como ayuda visual para identificar fácilmente los elementos importantes y diferenciarlos del resto del texto común. En muchos casos, los términos tienden a ser frecuentemente resaltados. Muchas veces ocurre que la definición también se encuentra señalizada con algún elemento tipográfico o con alguna tipografía específica. En este sentido, los patrones tipográficos se utilizan ya sea para resaltar a los elementos constitutivos mínimos de los CDs o bien para conectar dichos elementos.

> **(Ej. 8) Desastre**. *Perturbación de la actividad normal que ocasiona pérdidas o daños extensos o graves*.

(Ej. 9) MITIGACION: Disminuir los efectos de los impactos de las calamidades.

(Ej. 10) *Calamidad* ¶ Acontecimiento que puede impactar al sistema afectable y transformar su estado normal o deficiente en un estado de desastre.

En estos ejemplos, todos los términos están resaltados, ya sea en negritas, mayúscula o cursiva. En 8 y 9, el término se une a la definición a partir de un signo de puntuación, mientras que en 10 la definición aparece después de un salto de párrafo. En este último ejemplo, además de estar el término en cursivas, su presencia se hace más notoria por el hecho de aparecer en un párrafo anterior a modo de título.

Alarcón [6, 9, 24] encontró que las tipografías textuales más recurrentes para resaltar los elementos constitutivos mínimos de los CD son: cursivas, negritas, subrayados, mayúsculas, encabezados, viñetas y paréntesis. En cuanto al uso de signos de puntuación en los casos en los que se elide el verbo definitorio, encontró que los más usados son dos puntos, punto y guión, o punto y seguido.

4.2 Patrones sintácticos

Un camino para extraer CDs de manera automática en textos de especialidad consiste en identificar las estructuras sintácticas recurrentes tanto de los elementos mínimos constitutivos como de los conectores que unen a estos dos elementos. Alarcón [7] describe dos patrones sintácticos que sirven para conectar el término con su definición. Cuando dichos conectores tienen como núcleo un verbo, tenemos entonces un *patrón verbal definitorio* (PVD). Cuando se emplean otro tipo de formas sintácticas cuya finalidad es establecer una reformulación de una idea o concepto, y que se utilizan para esclarecer el significado de un término, tenemos *marcadores reformulativos*.

4.2.1 Patrones verbales definitorios

En CDs suelen utilizarse construcciones sintácticas verbales para unir a un término con su definición, a la vez de referir atributos y características conceptuales de dicho término [2]. Algunos de estos verbos son comúnmente considerados como verbos *metalingüísticos*, esto es, se emplean para referirse al propio lenguaje, como ocurre con *definir*, *entender* o *denominar*. También encontramos verbos muy comunes que podría decirse son de lengua general, empleados en diferentes situaciones comunicativas no solo definitorias, como los verbos *ser* y *considerar*.

Ocurren dos tipos de construcciones sintácticas verbales: En la más sencilla sólo se emplea un verbo de manera aislada, como *entendemos* o *definimos*. En la más compleja se recurre a una serie de partículas gramaticales, siendo de las más comunes el pronombre impersonal *se* en posición proclítica o enclítica en relación con el verbo definitorio, las preposiciones *a* o *por*, y el adverbio *como*. Algunas de las construcciones con estas partículas podrían ser: *se entiende por*, *se denomina a*, *definirse como*, etc.

(Ej. 11) En este sentido, el estado de un sistema se define como[1] una característica global que está determinada por un conjunto de valores en que se encuentran los parámetros relevantes para su funcionamiento en un momento dado.

(Ej. 12) Se denomina "equipo de salud" a todo el personal del hospital que tiene una función directa o indirecta para el paciente.

(Ej. 13) El tanque de almacenamiento es un recipiente en el cual se almacena el agua caliente para tenerla disponible a la hora que sea requerida su utilización.

En los ejemplos anteriores observamos que se introduce información definitoria a partir de los verbos *definir*, *denominar* y *ser*. Asimismo, la ocurrencia del pronombre *se* para los dos primeros verbos, *definir* y *denominar*, y el adverbio *como* y la preposición *a* para formar los patrones *se define como* y *se denomina a*. En el ejemplo 13, tenemos la combinación *ser* + *un*, estructura prototípica para definir un término.

4.2.2 Marcadores reformulativos

Al mismo nivel sintáctico e igualmente útiles para desarrollar una metodología de extracción automática de CDs, existe otro tipo de conectores que no consta de un núcleo verbal, pero que igualmente sirve para conectar al término con su respectiva definición. Este tipo de conectores o patrones sintácticos, que denominamos como *marcadores reformulativos*, conforman un proceso de reformulación en el que se explica el significado de un término a partir de estructuras sintácticas no verbales y, en el caso de los CDs, sirven para referirse a los términos como elementos del propio lenguaje.

Estos marcadores permiten retomar elemento de un discurso para presentarlo de otra forma, garantizan la cohesión textual y puntualizan el significado de algunos enunciados presentados anteriormente [27].

En el grupo de marcadores reformulativos podemos encontrar, entre otras estructuras: *por*

[1] Para distinguir de la tipografía original de los ejemplos, a partir de entonces utilizaré el subrayado para resaltar la parte de texto de interés.

ejemplo, *es decir*, *esto es*, *en otras palabras*, *dicho de otra manera*.

(Ej. 14) *El pronóstico de daños*, esto es, la cuantificación de la magnitud de las consecuencias o daños del fenómeno destructivo sobre el sistema afectable, conteniendo una relación de la cantidad de daños humanos, económicos, sociales y ecológicos que puede producir la calamidad.

(Ej. 15) El índice secundario es a menudo un índice denso, es decir, contiene todos los valores posibles de la clave primaria.

En 14 se utiliza el marcador *esto es* como conector entre el término *pronóstico de daños* con la definición. En 15 tenemos una reformulación para explicar que el término *índice secundario* implica que *contiene todos los valores posibles de clave primaria*.

4.3 Patrones pragmáticos

En textos especializados es común encontrar, además de la definición, otro tipo de información relevante para entender al término dentro del contexto en el cual aparece. Esta información describe el uso de los términos y manifiesta explícitamente las condiciones de uso o de alcance de dicho término, como son el ámbito temático, la ubicación geográfica, las instituciones que utilizan el término, el nivel de especialidad, o la frecuencia de uso, entre otras características pragmáticas [29].

Este tipo de patrones, que denominamos *patrones pragmáticos* (PPR), son muy útiles, junto con los patrones verbales, para identificar un posible CD dentro del texto cuando no existen patrones tipográficos. También nos permiten diferenciar fragmentos textuales donde el significado del verbo, por sí solo, no nos ofrece la seguridad de estar funcionando como un nexo entre un término y una definición.

Este tipo de patrones, que denominamos *patrones pragmáticos* (PPR), los dividimos en tres clases generales: patrones que corresponden al autor que propone la definición del término, patrones pragmáticos temporales y patrones pragmáticos instruccionales.

En los patrones pragmáticos de autor encontraremos patrones que hacen referencia directa al autor que propone el término. Estos patrones pueden ser sencillos, del tipo *Rosch* (nombre propio), o bien estructuras más complejas como: *los genetistas clásicos desde Mendel a Morgan*.

(Ej. 16) Inicialmente, Rosch definió el prototipo como el ejemplar que mejor se reconoce, el más representativo y distintivo de una categoría (…)

Los patrones pragmáticos temporales están en relación con la fecha de introducción o modificación del término, y ayudan por lo general a situar históricamente al término y su definición. Encontramos frases como *en 1889*, o bien estructuras más complejas como *a principios del siglo XX*.

(Ej. 17) Por ejemplo, la unidad de longitud – el metro - se definió en 1889 como la longitud de una determinada barra de platino iridiado (…)

Por último, los instruccionales consisten en estructuras que aportan matices diferentes para entender el término: *de manera general*, *desde un punto de vista práctico*, etc. Se denominan instruccionales ya que presuponen una condición de uso del término, es decir, el autor que introduce el CD aclara, mediante estas estructuras, cómo se debe entender el término o cuál es su alcance en un contexto determinado.

(Ej. 18) Desde el punto de vista genético, el desarrollo puede definir se como «un proceso regulado de crecimiento y diferenciación resultante de (…)

Es de reconocer que los patrones pragmáticos pertenecen a un paradigma estructural amplio, ya que su composición puede variar de acuerdo con formas estructurales o estilísticas utilizadas por cada autor. Con todo, podemos decir que las estructuras más recurrentes están conformadas por adverbios y frases adverbiales (*usualmente*, *de manera general*), frases prepositivas (*desde el punto de vista genético*), palabras simples (*definición*, *concepto*, *término*), y estructuras formadas por nombres propios (*Rosca*, *El norteamericano Instituto Nacional de la Salud*).

5. Análisis lingüístico de definiciones

El objetivo de extraer CDs es tener un repositorio de términos y sus correspondientes definiciones debidamente agrupadas según el tipo de información definitoria, lo que constituye la tipología de definiciones. Posteriormente veremos que esta tipología va íntimamente ligada con el patrón verbal definitorio, el cual presenta una estructura sintáctica precisa.

5.1 Tipología de definiciones

Nuestra tipología de definiciones identificables en CDs se sustenta en el modelo analítico [3, 24], en el hecho de que se haga explícito cuál es el género próximo y/o la diferencia específica, como se observa de la figura 2.

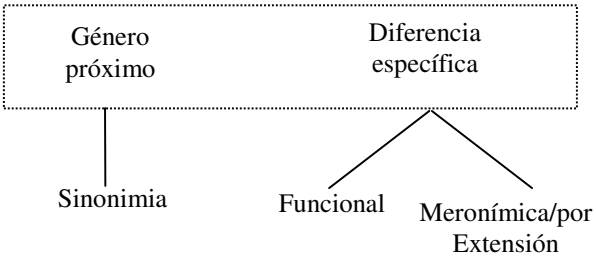

Figura 2: Tipología de definiciones

A partir de la relación observada entre la presencia y/o ausencia del género próximo y diferencia específica, así como entre el tipo de predicación que introduce y asocia a la definición con un término, se observan cuatro tipos de definiciones básicas con los siguientes rasgos:

- **Definición analítica o aristotélica:** se da una definición de este tipo cuando la predicación verbal introduce de manera explícita tanto el género próximo como la diferencia específica. El género próximo puede ser representado en forma de frase nominal, mientras que la diferencia específica puede expresarse en forma de algún tipo de frase (p.e., prepositiva, adjetiva o adverbial), o de oración subordinada introducida por alguna partícula de relativo (que/la cual/el cual/cuyo, quien, etc.). Por ejemplo: *Un algoritmo* es *un conjunto de instrucciones que se ocupa para una computadora*.
- **Definición sinonímica:** se da cuando la predicación introduce una definición que únicamente hace explícito el género próximo, sin considerar ningún tipo de diferencia específica, por lo que se establece una equivalencia conceptual con el término que es definido, p.e., *un maremoto* equivale a *un tsunami*.
- **Definición funcional:** se da cuando se reconoce únicamente la presencia explícita de la diferencia específica, la cual describe como rasgo distintivo de un objeto su función en un contexto dado. Por ejemplo: *una computadora* sirve para *procesar problemas y resultados lógicos, matemáticos y/o estadísticos*.
- **Definición extensional:** se presenta cuando la predicación introduce una definición en donde se explicita la diferencia específica (sin mencionar el género próximo). La clase de información conceptual asociada a estas definiciones puede ser de dos tipos: a) enumeración de las partes o componentes que integran un objeto; b) listado de todos aquellos objetos que conforman un conjunto. Por ejemplo: *una computadora* cuenta con *un hardware, un software, así como una serie de unidades periféricas.*

Definición	**Verbo**	**Adverbio o preposición**	**Unidades nominales**	**Predicación**
Analítica	Referir Representar Ser Significar	*A*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria
	Caracterizar Comprender Concebir Conocer Considerar Definir Describir Entender Identificar Visualizar	*Como* *Por*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Secundaria
Sinonímica	Denominar Equivaler Llamar Nombrar Ser	*También* *A* *Igual a* *Similar a*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria
Funcional	Emplear (se) Encargar Funcionar Ocupar Permitir Servir Usar Utilizar	*De* *Para*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria
Extensional	Componer Comprender Consistir Constar Contar Constituir Contener Incluir Integrar Es/son parte Es / son + **:** (dos puntos)	*De* *Por* *Con*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria

5.2 Sintaxis de las predicaciones verbales

En el estudio de Aguilar [1, 4, 5] se observó que las cuatro clases de definiciones anteriores mantienen estrecha relación con el verbo definitorio. Así, en función del verbo que opere como núcleo de una predicación, existe un patrón sintáctico en donde el género próximo y la diferencia específica se sitúan en posiciones de sujeto, objeto o predicado. Por ejemplo, en relación con el verbo *ser*, se observa un patrón sujeto + predicado, en donde el sujeto representa al término a definir y el predicado introduce la definición:

(Ej. 19) (Un algoritmo)$_{Suj}$ es (un conjunto de instrucciones para una computadora)$_{Pred}$

Tabla 1: Predicaciones verbales en CDs

(Ej. 20) (Turing)$_{Suj}$ define (algoritmo)$_{Obj}$ como (un conjunto de instrucciones para una computadora)$_{Pred}$

Aguilar analizó que los patrones predicativos que funcionan como conectores entre términos y definiciones en CDs muestran una constante frecuencia de uso a la hora de introducir el término y su definición.

En un plano general, existe una secuencia de organización sintáctica entre término, verbo y definición que se establece mediante el patrón predicativo: el término puede ocupar la posición de sujeto u objeto, el verbo como núcleo de la predicación, en tanto la definición es introducida por el predicado asociado al sujeto.

En un plano particular, los verbos que operan como núcleos de las predicaciones establecen una relación con la definición expresada por el predicado, de tal suerte que el verbo puede influir en la selección del tipo de definición que es introducida en un CD.

En un nivel de construcción sintáctica de un CD, una predicación organiza en qué posiciones pueden situarse el término y la definición en torno al verbo que opera como núcleo de dicha predicación. Entrando en mayores detalles, en este nivel se dan clases de secuencias de organización:

Una secuencia del tipo término + verbo + definición, en donde el término equivale al sujeto, el verbo funge como núcleo, y la definición es representada por el predicado que se asocia al sujeto, p. e.: *un error de programación es un fallo en la semántica de un programa*.

Una secuencia del tipo autor + término + verbo + definición, en donde el sujeto indica quién es el autor o los autores de una definición, el término equivale al objeto de la predicación, el verbo opera como núcleo, y la definición es introducida por el predicado asociado al objeto, p. e.: *Turing definió la inteligencia artificial como aquella inteligencia exhibida por artefactos creados por humanos*.

5.3 Variaciones tipológicas

En nuestra tipología de definiciones tenemos cuatro tipos. En la analítica se expresan los caracteres genéricos así como los diferenciales de una cosa, es decir, el género próximo y la diferencia específica. La extensional expresa las partes y componentes o el tamaño del término que se define. La funcional expresa la función, utilidad o el fin con el que se utiliza el concepto representado por el término en el CD. Tanto la extensional como la funcional tienen como rasgo característico compartido el carecer de género próximo.

Sin embargo, sucede en realidad que algunos CDs con definición de tipo analítica pueden tener como diferencia específica la extensión o la función del término que se define, ya que de esta manera la extensión o la funcionalidad de algún objeto permite tener un conocimiento más amplio del objeto que se está definiendo.

Sánchez [17] observó que la diferencia específica que expresa la función de un término definido puede ser introducida por una preposición o por el uso de sintagmas preposicionales. En particular, estudió la funcionalidad de un término introducida por la frase preposicional o patrón sintáctico *para* + infinitivo.

(Ej. 21) Así, un molino de viento es un artefacto útil para captar y aprovechar parte de esta energía

Del ejemplo 21 observamos que el término *molino de viento* tiene la función de *captar y aprovechar parte de esta energía*.

En términos generales, la preposición es una partícula o elemento sintáctico utilizado para establecer un tipo de relación entre un elemento A y un elemento B, donde A y B pueden ser oraciones o segmentos de una oración [38, 51, 61]. En función de la relación que establece con los términos que une, la preposición *para* es de tipo nocional [38], esto es, como su nombre lo indica, incluyen nociones como causa, finalidad, destinatario, instrumento, compañía, modo, etc.

En su investigación, Sánchez observó que la preposición *para* seguida de un verbo en infinitivo aporta, en un alto porcentaje, funcionalidad del término que se define, salvo las siguientes excepciones:

Caso 1.- El patrón *para* + *infinitivo* se encuentra fuera del CD. Como veremos en la sección 6.2, la extensión de un CD puede acabar antes del punto, por lo que hay que tomar en cuenta las reglas de delimitación para asegurar que el patrón se encuentra dentro de los límites del CD. Por ejemplo:

(Ej. 22) La máquina virtual es un ordenador con una pila sencilla; los programas están estructurados para permitir que los clientes verifiquen la existencia de referencias ilegales ni errores gramaticales en el código descargado

En 22 observamos que el CD termina en el punto y coma, por lo que el patrón no aporta información funcional al término *máquina virtual*, sino a *programas*.

Caso 2.- El patrón se encuentra alejado del término o del género próximo o del término por una sucesión en más de dos grados de sintagmas preposicionales (sp). Por ejemplo:

(Ej. 23) La impresora es el órgano típico $(\text{de salida})_{sp1}$ $(\text{de información})_{sp2}$ $(\text{del ordenador})_{sp3}$ para ser utilizada en la empresa

Caso 3.- Se encuentra separado el término o la diferencia específica del patrón mediante la introducción de una oración relativa; p.e.:

(Ej. 24) Un "analizador sintáctico" es un programa con el <u>que</u> se pueden comprobar series de caracteres <u>para ver</u> si son fórmulas bien formadas de un lenguaje dado.

Caso 4.- Existe un elemento que cambia la funcionalidad del patrón, ya sea modificándolo, mediante un adverbio de negación, o bien negándolo, mediante un adjetivo con carga semántica negativa; por ejemplo:

(Ej. 25) El problema de la luz es que es una mezcla de varias frecuencias, y por tanto <u>poco útil para ser</u> empleada como medio de comunicación, excepto si usamos una luz monofrecuencia obtenido por medio del láser y un conductor conocido como fibra óptica

Así, el estudio de Sánchez nos permitió observar que la tipología propuesta de definiciones es flexible y que el patrón sintáctico *para + infinitivo*, inserto en la definición de un CD de tipo analítico, aporta información de funcionalidad del término que se define. Asimismo, definió algunas reglas de exclusión a este patrón, con lo que es posible mejorar la extracción automática de CDs.

6. La extensión de un CD

Hemos visto los elementos constitutivos de los CDs y la forma en que se construyen. Sin embargo, falta mencionar la extensión de los CDs, esto es, los límites de inicio y finalización del fragmento textual que contiene la definición completa de un término. Como unidades discursivas, tienen estructuras distintas, sin tener un número de palabras fijo y con elementos que pueden presentarse en diferente orden.

En un principio puede considerarse el párrafo como la estructura textual para establecer la extensión de un CD, pero como veremos a continuación, cabe la posibilidad que la extensión vaya más allá de un párrafo, o bien que en el mismo párrafo exista más de un CD.

6.1 Las anáforas en la expansión de CDs

Un tema interesante, pero complejo tratándose de CDs, es la forma en que las relaciones anafóricas intervienen para su extracción. La anáfora es comúnmente el término que se emplea para hacer referencia a algo que anteriormente ya fue mencionado, y considera cualquier expresión, palabra o frase que recupera algo previamente enunciado. El análisis de relaciones anafóricas juegan un papel determinante para la obtención completa de un CD.

(Ej. **26)** <u>Este</u> consta de un banco de capacitores sumergidos en aceite en un recipiente de porcelana y conectados en serie (…)

En efecto, en (26) vemos un pronombre demostrativo en representación del término del contexto. Si solo extrajéramos este CD incompleto sería imposible determinar a qué término corresponde la definición: "un banco de capacitores sumergidos en aceite en un recipiente de porcelana (…)". Con base en lo anterior, es evidente la necesidad de una *extensión* de este tipo de casos. Por *extensión*, entendemos el tamaño del fragmento textual que contiene el CD completo, con término y definición, mientras que por *expansión* se comprenderá la pertinencia de acudir al documento de origen del contexto con el objetivo de verificar la extensión del CD.

Con la finalidad de resolver este problema, primero es necesaria la identificación de los tipos de relaciones anafóricas que operan con CDs. Con este fin, Benítez [13] realizó un estudio profundo donde se describen de manera completa relaciones anafóricas presentes. En dicho estudio Benítez encontró que, principalmente, son cuatro las expresiones más frecuentes en CDs.

En el primer grupo se encuentran algunos pronombres demostrativos (esto, aquellos, esta), personales (lo, le), relativos (la cual, lo cual, que) e impersonales (el primero). La frecuencia de esta clase de expresiones no es muy alta, pero son las más comunes en la ocurrencia de candidatos incompletos, como puede verse en el ejemplo 27, ya que la expresión apunta a un antecedente omitido en la extracción automática.

(Ej. 27) Esto es lo que se entiende por enfoque genético de la medicina o "medicina genética".

El segundo grupo abarca los sintagmas nominales con determinante demostrativo, los cuales son expresiones con valor anafórico porque refieren a una parte anterior en el texto. Por ejemplo:

(Ej. 28) <u>Estos elementos</u> son parte constitutiva de los compuestos que forman la base material para la vida (…)

El tercer grupo lo conforman las expresiones mixtas (pronombres y sintagmas nominales con demostrativo) en las que se muestran cadenas de anáforas o anáforas muy cerca de otras, es decir, que las cadenas de referencia se manifiestan con pronombres y sintagmas nominales que se encuentran en una relación anafórica.

(Ej. 29) <u>Esta concepción</u> es <u>lo</u> que se conoce con el nombre de materialismo histórico.

En 29 se oberva cómo la expresión anafórica, representada por el pronombre *lo* hace referencia al sintagma nominal con demostrativo *esta concepción,* que a su vez tiene como referente al verdadero término de la definición.

El último grupo está constituido por las expresiones ligadas a una entidad previamente enunciada, las cuales pueden ser sintagmas nominales, elipsis o marcadores discursivos.

(Ej. 30) El primer grupo es típico de los buques rápidos y consiste en olas de gran periodo, que sufren poca dispersión al alejarse del barco (…)

Una vez llevada a cabo la observación del corpus y después de realizar la clasificación de los elementos más frecuentes en las relaciones anafóricas, Benítez realizó el diseño de etiquetas XML para la identificación de relaciones anafóricas siguiendo los patrones de formación de las etiquetas ya establecidas para el CORCODE.

6.2 La delimitación del CD

Para la conformación de un sistema de extracción conceptual, es importante tener en cuenta que no todos los contextos definitorios son iguales, esto es, que no todos los CDs tienen una misma estructura en la que comienzan con el término y terminan en el primer punto después de la definición.

Para reconocer automáticamente la extensión de un CD dentro de un texto se tomó en cuenta un criterio básico inicial, que consiste en delimitar un contexto en el primer punto. Si bien este criterio es funcional en gran medida, no siempre obtiene buenos resultados como se muestra a continuación.

(Ej. 31) La "acción" es entendida como la conducta intencionada proyectada por el agente; en cambio el "acto" es definido como la acción cumplida.

En 31 podemos ver que la definición del término "acción" acaba antes del primer punto y antes de la introducción del término, "acto".

Con la finalidad de evitar información que no sea parte del CD y así mejorar el sistema de extracción, se requiere del planteamiento de reglas lingüísticas que permitan delimitar definiciones automáticamente cuando éstas terminan antes del primer punto.

Hernández [15] realizó un estudio para delimitar contextos en definiciones de tipo analítico; es decir, con género próximo y diferencia específica, debido a que cada tipo de definición requiere de un propio estudio y reglas particulares. En su investigación, observó y analizó dos tipos de patrones lingüísticos de delimitación.

6.2.1 Patrones que rompen con la definición

Un primer tipo de patrones lingüísticos que delimitan un CD tienen la característica de que lo que viene después del patrón rompe por completo con lo que se estaba expresando en la definición sobre el término, esto es, marcan la introducción de un nuevo término o foco dentro del discurso, el cual ya no pertenece al CD. Cinco de los patrones encontrados son:

Patrón 1.- Por tanto/por lo tanto. Este marcador discursivo, considerado como conector consecutivo [53], introduce una consecuencia o una conclusión en el elemento siguiente. Como podemos observar en (32), la intensión que se introduce en la definición se ve concluida con el enunciado posterior al patrón *por tanto*.

(Ej. 32) Finalmente, no debemos olvidar que el <T>dengue</T> <PVD>es</PVD> <D>un virus que puede replicarse en células de mamífero y en células de mosquito,</D> por tanto, los aspectos antes descritos para células de humano pueden también estar operando en el mosquito vector.

Patrón 2.- Sin embargo + FN. Este patrón se encuentra constituido por un marcador discursivo de tipo conectivo contra-argumentativo [53], pues vincula dos miembros, de tal modo que el segundo se presenta como supresor o atenuador de alguna conclusión que se pueda obtener del primero.

(Ej. 33) <PP>En general</PP>, el <T>ácido nucleico </T> <PVD> es </PVD> <D> una molécula única de hélice simple o doble</D>; sin embargo, ciertos virus tienen el material genético segmentado en dos o más partes.

Patrón 3.- En cambio + FN. Este patrón compuesto por un marcador conector contra-argumentativo seguido por una frase nominal muestra un contraste entre los términos que se definen. En el ejemplo 34 los términos "adenina" y "citosina" son contrapuestos semánticamente a través del marcador que funciona como conector.

(Ej. 34) La <T>adenina</T> y la <T>guanina </T> <PVD> son </PVD> <D>bases púricas </D>, en cambio la citosina y la timina son bases pirimidínicas.

Patrón 4.- Mientras que + FN. Con el marcador contra-argumentativo *mientras que* se oponen dos enunciados distintos y en nuestro caso los elementos contrapuestos son CDs. En 35, el patrón (mientras que + FN) se encarga de definir hasta dónde llega el primer CD cuyo término es "hiperalgesia primaria".

(Ej. 35) La <T>hiperalgesia primaria</T> <PVD> se concibe como </PVD> <D> el aumento de la respuesta al estímulo doloroso en la región de la lesión </D>, mientras que la hiperalgesia secundaria es aquella que se extiende para áreas adyacentes.

Patrón 5.- (En tanto/en tanto que) + FN. El marcador conector contra-argumentativo *en tanto* se encuentra funcionando de la misma forma que *en cambio* y *mientras que* cuando les sigue una frase nominal y están cerca de contextos definitorios en ámbitos de especialidad.

6.2.2 Patrones que continúan con información relevante

El segundo tipo de es aquel en donde la información que sigue a la regla o patrón lingüístico es pertinente para el CD, ya que amplía, reformula o explica la información definitoria del mismo término, pero ya no constituye ninguna de las partes formales de la definición analítica. El beneficio que aportan estos patrones consiste en que la información que se introduce puede ser parte o no del CD, según las necesidades y propósitos del sistema de extracción. Hay que tomar en cuenta que, al aportar información enriquecedora para el CD, no pueden ser considerados como patrones de delimitación como tal, sino más bien como indicadores del final de la diferencia específica.

(Ej. 36) La <T> adolescencia </T> <PVD> es definida como</PVD> <D>una etapa del ciclo vital entre la niñez y la adultez, que se inicia por los cambios puberales</D> y se caracteriza por profundas transformaciones biológicas, psicológicas y sociales, muchas de ellas generadoras de crisis, conflictos y contradicciones, pero esencialmente positivos.

En 36 podemos ver que el término "adolescencia" tiene dos definiciones. En la primera es definida como "una etapa del ciclo vital entre la niñez y la adultez, que se inicia por los cambios puberales" y en la segunda es caracterizada por "profundas transformaciones biológicas, psicológicas y sociales, muchas de ellas generadoras de crisis, conflictos y contradicciones, pero esencialmente positivos". El patrón delimita la extensión de la primera definición, aunque lo que viene después sigue siendo relevante para el CD y debe por tanto tomarse en cuenta.

Entre los patrones que podemos encontrar de este tipo tenemos: *por ejemplo*, *como por ejemplo*, *tal como*, *o sea*, *es decir*, *y*+PVD.

7. La extracción automática de CDs

El objetivo que perseguimos con el análisis previo es lograr la extracción automática de CDs en español a partir de textos de especialidad. Gracias al conocimiento lingüístico de la conformación de CDs nos fue entonces posible desarrollar la metodología pertinente.

Nuestra metodología para extraer CDs está basada en reglas lingüísticas y consiste en la búsqueda automática de ocurrencias de patrones definitorios, específicamente PVDs [10, 18, 19]. El Extractor de Contextos Definitorios (ECODE), que fue desarrollado por Alarcón [7], abarca un procesamiento automático de los candidatos a CDs: primeramente, un filtro de contextos no relevantes, esto es, aquellos contextos donde, a pesar de tener un PVD, no se define un término; luego, la identificación de los elementos constitutivos del CD, es decir el término y la definición; finalmente, una ponderación de resultados para determinar cuáles son los mejores CDs propuestos por el sistema.

Para obtener CDs se debe tener como entrada un corpus anotado con etiquetas de partes de la oración (POS). De ahí, el proceso general consiste en tres pasos: la extracción de candidatos, el análisis de candidatos y la evaluación de los resultados.

7.1 Extracción de candidatos

El proceso principal del ECODE lo constituye la extracción de candidatos, la cual requiere una gramática de PVD que contiene una serie de parámetros:

- Los verbos definitorios a buscar junto con los nexos que los acompañan, ya que un verbo puede estar acompañado o no de diferentes nexos para expresar definiciones de varios tipos. Por ejemplo, el verbo *conocer* asociado con el nexo *como* obtendrá por resultado CDs del tipo analítico, en tanto asociado con el nexo *también* nos dará un CD del tipo sinonímico.
- Las restricciones verbales referentes al tiempo y a la persona gramatical, ya que la información definitoria a recuperar depende del tiempo, de la forma verbal o de la persona gramatical para cada verbo. Como puede verse en los ejemplos 37 y 38, el verbo *definir* en primera persona de plural nos traerá información definitoria, pero no así el verbo *contar*.

(Ej. 37) La radiación provoca mutaciones, que definimos antes como cambios en la secuencia de las bases del ADN.

(Ej. 38) Cómo se va a regular la aplicación de estudios de escrutinio conforme contemos con el conocimiento de dichos genes?

- Los patrones contextuales, esto es, la delimitación de las posiciones en las que podrían aparecer el término y la definición respecto al verbo definitorio. Este parámetro es crucial para el ECODE, pues posteriormente será utilizado para identificar los elementos constitutivos de cada CD. Entre algunos de los patrones contextuales tenemos: T+PVD+D, VD+T+NX y PVD+T+D, como se observa en los siguientes tres ejemplos, respectivamente.

(Ej. 39) <T> La COMT </T> <PVD> <VD> es </VD> <NX> una </NX> </PVD><D> enzima de distribución amplia, presente tanto

en tejidos neuronales como en los no neuronales.</D>

(Ej. 40) <PVD> Se ha <VD> definido </VD> <T>el genotipo</T><NX>como<NX> </PVD> <D> la constitución genética del individuo en un locus. </D>

(Ej. 41) <PVD> Se denomina </PVD> <T> digestión </T> <D> al proceso por el cual las moléculas ingeridas son fraccionadas en otras más pequeñas mediante reacciones catalizadas por enzimas, bien en la luz o bien en la superficie orientada hacia la luz del tracto GI.</D>

- Restricciones de distancia entre el verbo y su nexo, pues entre ambos puede aparecer desde un adverbio o un término simple, hasta unidades más complejas como sería un término compuesto más una frase adverbial. Este parámetro debe analizarse con cuidado, pues de lo contrario puede causar que el extractor traiga mucho ruido. En el ejemplo 42 podemos observar un CD con una distancia de 8 palabras, mientras que en 42 tenemos que inclusive se rompe el CD entre el verbo definitorio y el nexo.

(Ej. 42) En 1977, Oshimura et al describieron las deleciones del brazo largo del cromosoma 6 como una anomalía recurrente en leucemias.

(Ej. 43) La clasificación de las distrofias musculares ha ido evolucionando con el tiempo: desde finales del siglo pasado y hasta los años cuarenta, las descripciones anatomoclínicas definían los criterios de clasificación; en una segunda etapa, los distintos patrones de herencia se contemplaron como parámetros a tener (...)

7.2 Análisis de candidatos

Una vez extraídos los candidatos a partir de los PVD y el empleo de la gramática, el análisis de los CDs incluye dos procesos principales: el primero consiste en eliminar los contextos no relevantes mediante reglas de filtrado y, el segundo, en la identificación de los elementos constitutivos.

El filtro de contextos no relevantes se basa en una serie de reglas lingüísticas y contextuales para determinar los casos en los que es probable que un patrón verbal no esté introduciendo información definitoria. Mientras existen verbos de carácter prototípicamente definitorios, otros se utilizan en una gran variedad de situaciones. Por ello, entre las reglas de filtrado, Alarcón [7, 8, 11] propuso una lista de restricciones basadas en ciertas partículas gramaticales (principalmente preposiciones, adverbios, pronombres y verbos en forma conjugada), y en la posición en las que pueden aparecer dichas partículas adyacentes o dentro del PVD. Por ejemplo:

(Ej. 44) <PVD><PR>Se</PR> <VD>conocen </VD> ya las secuencias de bases de muchos genes salvajes y mutantes, así<NX>como </NX></PVD> las secuencias de aminoácidos de las proteínas que codifican.

En 44 tenemos un contexto no relevante debido a la partícula *así* inmediatamente anterior al nexo *como*.

Ya con los candidatos que no fueron filtrados como excepciones, el siguiente paso consiste en la identificación de los elementos constitutivos, esto es, el término y la definición. Para ello, se utiliza un árbol de decisión (Fig. 3) que recurre igualmente a la gramática de patrones verbales. El árbol de decisión, a través de inferencias lógicas, asocia una serie de patrones contextuales para cada verbo definitorio, de forma que dichos patrones indican las posiciones en que puede aparecer el término y la definición.

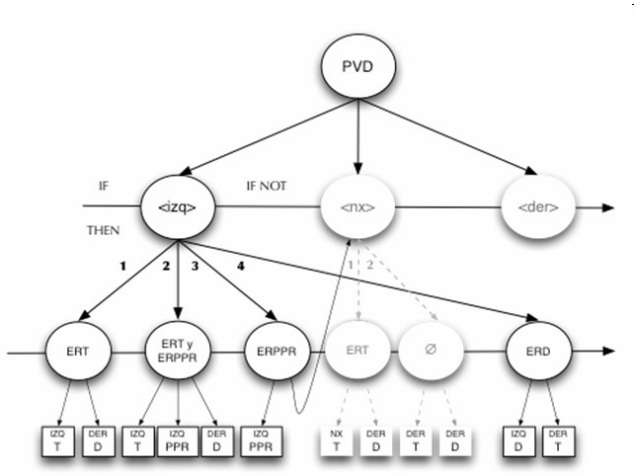

Figura 3: Árbol de decisión para el ECODE

De ahí el procedimiento busca asegurar, mediante el empleo de expresiones regulares, si el elemento se ajusta a la expresión de término, de definición o de patrón pragmático.

(Ej. 45) <IZQ>El turismo, en términos generales, </IZQ> <PVD> <AUX> ha sido </AUX> <VD> concebido </VD> <NX>como </NX> </PVD> <DER> la reproducción de los hábitos cotidianos en un ambiente diferente.</DER>

En el patrón contextual mostrado en el ejemplo 45 tenemos el PVD *ha sido concebido como,* además de una parte a la izquierda y otra a la derecha. La gramática de patrones verbales señala que, para el verbo concebir, el término puede encontrarse en la posición izquierda pero no en la posición derecha del verbo definitorio. Se asegura

que el término se encuentra en la posición izquierda por una expresión regular de término que pide la presencia de frontera *<IZQ>* seguida de un determinante, un sustantivo y todo lo que esté después hasta la siguiente frontera que es la etiqueta de cierre de la posición izquierda (*</IZQ>*). Continuando el análisis, con una expresión regular de patrón pragmático se identifica *en términos generales* porque empieza con una coma seguida de una preposición y hasta la frontera de cierre. Para el caso de la definición se tiene su expresión regular formada por determinante más sustantivo, delimitado por las fronteras de inicio y de cierre.

7.3 Ponderación de resultados

El tercero y último proceso del ECODE busca evaluar los CDs que resultan después del filtrado de excepciones de contextos no relevantes, y en particular los elementos constitutivos, para ponderar los mejores CDs según la estructura del contexto recuperado automáticamente. Se utiliza una serie de reglas heurísticas que comparan las estructuras sintácticas de los elementos etiquetados como término y definición con sus estructuras prototípicas. Se asigna un valor a cada elemento y un valor global a partir de las combinaciones encontradas. Los contextos que pasen de un umbral determinado serán los que el ECODE arroje como buenos CDs.

Si bien este último proceso permite obtener en primer lugar los mejores candidatos a CDs, cabe advertir que existen riesgos. Ya que este proceso de ponderación se basa en las estructuras sintácticas de los elementos que se van a ponderar, se puede asignar un valor equívoco en caso de también el etiquetado POS contenga errores. Además, las expresiones regulares de término y definición también pueden traernos elementos que no lo son pero que cumplen estructuralmente con las reglas de los buenos candidatos. Entre algunos casos erróneos, Alarcón llegó a encontrar los siguientes términos: *repetitivo*, *una vez*, *la molestia de la ropa interior teñida*.

8. Agrupamiento de CDs

El ECODE proporciona finalmente una lista de CDs asignados a alguno de los tipos de definiciones: analítica, extensional, funcional o sinonímica, y organizada según la probabilidad de que sean en mayor o menor medida mejores CDs.

Además de la clasificación de los CD por su tipo de definición, también pueden ser agrupados según sus características semánticas. Esto es, se pueden agrupar los CD polisémicos por sus diferentes significados o incluso por las características descritas en su definición. En el primer caso, tenemos por ejemplo el término virus, del cual se pueden tener por un lado los CDs correspondientes al área de informática y por otro lado los correspondientes al área de medicina o de biología. En el segundo caso, podemos encontrar por ejemplo los CDs con definiciones analíticas para el término gen, que por un lado lo describen como la unidad de la herencia y por otro como una secuencia de ADN.

Por esta razón, Molina [16] se ha dedicado a la tarea de desarrollar un algoritmo para poder llevar a cabo el agrupamiento automático de CDs según su significado, de tal forma que los resultados de la búsqueda de un término polisémico sean presentados mediante una clasificación semántica.

La ventaja más importante del algoritmo de agrupamiento es ser independiente del idioma, no requiere de ningún tipo de anotación lingüística, como etiquetas POS, tampoco requiere de un conjunto de entrenamiento previo, ni es necesario indicar el número de grupos a generar y, finalmente, a diferencia de otros algoritmos similares como *lingo* [60], el algoritmo aquí descrito es fácilmente configurable, pues depende únicamente de un parámetro: el valor de corte por distancia.

Para la realización del algoritmo de agrupamiento semántico se toman como base los resultados del sistema de extracción de contextos definitorios *ECODE*, el cual entrega un archivo de salida con CDs clasificados según el tipo de definición.

El algoritmo lleva a cabo tres grandes etapas. Dentro de la primera, el texto es procesado hasta llegar a la representación vectorial usando diversas técnicas de procesamiento del lenguaje natural. En la segunda, se calcula la distancia entre cada vector utilizando la matriz de energía textual y, en la última etapa, se aplica el agrupamiento jerárquico con el método de vecino más lejano.

De forma general, los pasos que realiza el algoritmo son los siguientes:

8.1 Preprocesamiento

Con la finalidad de reducir el tamaño del espacio vectorial, se procesa cada archivo en tres etapas: la transformación de signos de puntuación y diacríticos; la eliminación de palabras que no son de contenido y el truncamiento de las palabras.

El primer contacto del texto con el algoritmo será a través de un archivo en texto plano, sin etiquetas, ni ningún tipo de marcaje como XML y HTML. La primera etapa de preprocesamiento consiste en unificar la diversidad de los símbolos gráficos. Por ejemplo, con la intensión de reducir la diversidad de símbolos u, ü y ú, se unificarán bajo el símbolo u.

En la segunda etapa, los textos son filtrados con una lista de paro (stop list), lo cual reduce en gran medida el tamaño del diccionario generado por la colección. También, son eliminados todos los patrones verbales definitorios junto con el término, pues estos elementos aparecen en todas las definiciones de la colección y, por tanto, no contribuyen a constituir un criterio de agrupamiento.

La última transformación consiste en truncar las palabras mediante el algoritmo de Portero [65]. La intensión de esta transformación es unificar en un solo símbolo aquellas palabras que poseen en la misma raíz y que están relacionadas semánticamente. Por ejemplo, las palabras *vivo* y *viviente* se unifican bajo el mismo símbolo *viv*.

8.2 Construcción del espacio vectorial

En esta etapa se construye el espacio vectorial generado por las definiciones, es decir, una matriz concebida como un arreglo de vectores que representan documentos. Un documento es una cadena de longitud arbitraria pero finita de símbolos gráficos denominados entidades léxicas (EL). Entendemos como EL aquella que puede ser representada mediante un símbolo o la unión de varios de ellos. Así, la palabra *manzana* puede representar una EL, o bien una frase como *Estados Unidos Mexicanos*. Asimismo, una EL puede ser un símbolo ininteligible como *Viv* o *A4*. De esta manera, una colección es un conjunto de documentos y un diccionario es una lista de ELs únicas que aparecen en una colección.

8.3 Cálculo de la energía textual

Una vez generada una matriz binaria surge necesidad de comparar definiciones a partir de su representación vectorial. Para esto, es necesario tener un mecanismo de comparación entre textos que funcione como criterio para determinar los grupos semánticos. Con esta finalidad se optó por derivar una medida de distancia a partir de la matriz de *energía textual* propuesta por Fernández, San Juan y Torres Moreno [39]. Esta técnica resulta funcional porque fue concebida desde sus inicios como una aproximación teórica para ponderar las relaciones de significado en textos.

La distancia entre los vectores a partir de la matriz de energía textual se calcula a partir de la siguiente fórmula:

$$DistEner = \frac{máx(D_{ener}) - D_{ener}}{máx(D_{ener})}$$

DistEner es un arreglo que contiene la distancia entre cualesquiera dos documentos *i,j* de la colección dado que *E* es una matriz simétrica, esto es, $e_{ji}=e_{ij}$. De esta forma, hemos calculado la distancia entre documentos a partir de la matriz de energía textual. Tenemos, ahora, la posibilidad de utilizar un algoritmo de agrupamiento para generar una estructura de grupos utilizando esta distancia como criterio.

8.4 Agrupamiento de definiciones

Una vez que la proximidad entre textos es calculada, son generados los grupos por medio de un algoritmo jerárquico aglomerativo simple. Un algoritmo de tipo jerárquico ofrece la ventaja de que no requiere que el número de grupos sea especificado previamente.

El método utilizado para comparar los grupos en el algoritmo jerárquico es el método del vecino más lejano (*complete linkage*). Es preferible este método porque genera grupos pequeños, cohesivos y bien delimitados, brindando la posibilidad de mejorar la precisión de los grupos.

El criterio para determinar el número de grupos generados es un valor umbral de corte por distancia. Con dicho valor es posible indicar el valor máximo de distancia que puede haber entre dos grupos. Por ejemplo, si determinados que el valor umbral de corte por distancia es 0.1 significa que aquellos grupos cuya distancia es mayor a 0.1 nos son unificados. Además el algoritmo de agrupamiento jerárquico genera un dendograma que permite calcular coeficientes de comparación entre agrupamientos y representar gráficamente los resultados obtenidos de cada ejecución del módulo.

9. Resultados y evaluación

Hasta ahora he mostrado una síntesis de los estudios que se han realizado en el Grupo de Ingeniería Lingüística para analizar los CDs, clasificarlos, reconocer y precisar sus elementos constitutivos, delimitar su extensión, extraerlos y agruparlos. Hemos visto la metodología de cada uno de estos estudios, pero conviene ahora tener una síntesis de los resultados obtenidos y su evaluación.

9.1 El corpus de estudio

Para los diferentes estudios que se describen en este artículo, se trató de ser consistentes en el empleo de las mismas fuentes, con lo que conformamos los corpus de experimentación, de prueba y de evaluación. Los principales corpus utilizados son los siguientes:

- El Corpus Lingüístico de Ingeniería o CLI [54]. Se trata de un corpus en español orientado al área de ingeniería y desarrollado por el Grupo de Ingeniería Lingüística. Está

conformado por documentos en texto plano (extensión *.txt*), con alrededor de 500,000 palabras (tokens). Se trata de un corpus que reúne textos especializados del área de ingeniería, tales como tesis, artículos, informes, etcétera. Una de las ventajas de este corpus es que los textos usualmente incluyen apartados, ya sea introducción, presentación o bien un capítulo específico, que funcionan como marco teórico donde se definen los términos esenciales para la comprensión del contenido.

- El Corpus Técnico del Instituto Universitario de Lingüística Aplicada (CTIULA) de la Universidad Pompeu Fabra en Barcelona [80]. Este corpus cuenta con 9,542,000 palabras en su sección dedicada al español, al cual se puede acceder a través de su herramienta de búsqueda *BwanaNet*. Está etiquetado con partes de la oración y cuenta con tres opciones de búsqueda: básica, estándar y compleja.
- El Corpus Informático en Español o CIE [50], es un corpus técnico desarrollado para las áreas de informática y ciencias de la computación, con miras a la creación e implementación de un diccionario electrónico en español. Cuenta con alrededor de 500,000 palabras, divididas en 4 sub-corpus: de la revista *PC World Latinoamérica* (PCWLAF), revista *Guía Computación*, *WindowsTI Magazine*, y entradas obtenidas de la Wikipedia en español.

Fuente	Número de CDs
CLI	238
CTIULA	1,361
CIE	562
SkE	5
Google	49

Tabla 2: Corpus de CDs

En menor medida se utilizó el Spanish Web Corpus de la herramienta *Sketch Engine* (SkE) [47], y el motor de búsqueda *Google*.

Como resultado, en total se obtuvieron en total 2,215 contextos, como muestra la tabla 2.

9.2 Evaluación

Como medidas de evaluación se han usado principalmente las tradicionales para los sistemas de recuperación y extracción de información: precisión y cobertura. Como explican Jurafsky y Martin [46], la precisión es una medida que se utiliza para determinar cuánta información extraída automáticamente por el sistema es correcta, mientras que la cobertura es una medida para saber cuánta de la información relevante en el texto fue extraída automáticamente.

La precisión se representa entonces como la proporción del número de respuestas válidas propuestas por el sistema, del total de respuestas propuestas por el sistema. La cobertura queda como la proporción del número de respuestas válidas propuestas por el sistema, del total de respuestas del texto.

Cabe advertir que determinar las respuestas válidas resulta complicado en el caso de CDs. En el ámbito de la terminología y lexicografía resulta muchas veces un retos precisar los límites de una definición. Si bien Aguilar [1, 4] profundizó en el concepto de definición, en la práctica resulta muchas veces difícil llegar a consenso sobre el límite de la definición analítica o a precisar el género próximo de la misma.

(Ej. 46) En ecología, biomasa es el término usado para definir el volumen total de materia viva en forma de microorganismos, vegetales, animales, que soporta un ecosistema determinado.

Así, la definición del término biomasa es del tipo analítica, y como tal debe estar constituida por un género próximo y una diferencia específica. Sin embargo, es controversial precisar dónde termina el género próximo, si en total, en materia viva o en animales. Por esta razón, para la evaluación nos apoyamos en estudiantes involucrados en el área de terminología. Para resolver las dudas trabajamos en equipo y discutimos cada uno de los casos.

Como muestra de la evaluación, podemos mencionar la obtenida para el ECODE, en donde se consideró como CD cuando apareciera explícitamente el término y la definición. El corpus de evaluación quedó conformado por contextos definitorios y contextos no relevantes. Alarcón [7, 12] reporta que para la precisión dividió el número de CDs válidos propuestos por el sistema sobre el número de CDs propuestos por el sistema (1783/3309), con lo que quedó un valor de 0.53. Para la cobertura dividió el número de CDs válidos propuestos por el sistema sobre el número de CDs en el corpus (1783/2254), quedando un valor de 0.79. Esto es, se obtuvo una mejor cobertura frente a la precisión. Mientras que se recuperó el 80% de CDs presentes en el corpus, solo un poco más del 50% de lo obtenido era válido.

10. El corpus de CDs

A lo largo de la investigación hemos obtenido un acervo de CDs con lo que podemos construir el CORCODE o Corpus de Contextos Definitorios. Éste va más allá de ser un repositorio de documentos, pues constituye una herramienta valiosa para la terminología y la lexicografía, al

permitir facilitar el proceso de extracción de unidades tales como términos y definiciones.

El CORCODE es un corpus compuesto por CDs enfocados en áreas de especialidad. Actualmente puede consultarse en la página del Grupo de Ingeniería Lingüística un total de 127 CDs.[2] La interfaz de búsqueda permite realizar navegaciones a partir del tipo de término, tipo de definición, tipo verbo definitorio, de marcadores textuales definitorios (comas, dos puntos, comillas, etc.) y de los patrones pragmáticos (autoría, patrones temporales o instruccionales).

Este método de búsqueda se da a partir de un etiquetado en XML que facilita la identificación de las partes de los CDs. Estas etiquetas delimitan a cada CD de forma global, así como los elementos que los constituyen. En primera instancia, se configuró el encabezado del documento XML, que se muestra a continuación:

- Fuente. Indica la fuente original del documento (CLI, CTIULA, CIE, *Google*, SkE).
- Fecha. Indica la fecha del recopilado y del etiquetado del documento.
- Nombre. Contiene el nombre de la recopilación hallada en el documento, como puede ser "verbo definir".
- Verbo. Muestra el nombre del verbo definitorio que se analiza.
- Tipo. Se indica si el criterio de clasificación del documento es la *definición*. Estas pueden ser: analítica, funcional, extensional o sinonímica.
- Recopilador. Muestra el nombre de la persona que recopiló el documento.

El cuerpo del documento contiene los CDs etiquetados. Las etiquetas utilizadas se pueden apreciar en el siguiente cuadro.

- CD. Contexto Definitorio: Indica los elementos que constituyen al CD, dentro de ellos se encuentra el término, su definición, la predicación verbal y las relaciones de correferencia.
- TERM. Término: En su atributos se marca se trata de un término lingüístico o de uno no lingüístico (cifras, símbolos). Se toman en cuenta tres tipos de frase: *fn* (frase nominal, *fn Y fprep* (frase nominal seguida de frase prepositiva) y *fv Y fn* (frase verbal seguida de frase nominal).
- DEF. Definición: En ella se debe omitir cualquier texto complementario que de manera estricta no forme parte de dicha definición. Existen cinco tipos: *GD* (Género próximo/Diferencia específica), *FUN* (Funcional), *EXT* (Meronímica/Extensional), *Ges* (Género exclusivo) y *Sin* (Sinonímica que se marcan en los atributos.
- PVD. Patrón Verbal Definitoria: Contiene todos los componentes de un PVD, incluyendo el clítico *se*, el verbo auxiliar, el verbo definitorio y el nexo.
- VD. Verbo Definitorio: Cuenta con los atributos *lema*, *args* (marca los argumentos del verbo); *mod* (indica el modo verbal: infinitivo *inf*, gerundio *ger*, participio *part*, formas finitas o verbo conjugado *fin*).
- Semarc. Clítico *Se.* Indica su posición respecto al verbo. El atributo distingue entre *enclítico (enc)* cuando *se* es parte de la morfología verbal y está en posición final, y *proclítico (prec)* cuando el clítico está en posición preverbal.
- Vaux. Verbo Auxiliar. Contiene cualquier verbo auxiliar dentro de la PVD (p.e., se puede considerar como, se ha definido, se debe concebir como…).
- NX. Nexo: Señala la función que cumple un adverbio o preposición entre el verbo y la definición.
- MRD. Marcadores Reformulativos Definitorios: Abarcan estructuras sintácticas con la función de explicar el propio lenguaje, p.e.: es decir, por ejemplo, esto es, etc.
- MTD. Marcadores Tipográficos Definitorios: Señala cualquier signo de puntuación o marcadores tipográficos definitorios (MTD). Se distingue en dos tipos: 1) marcadores definitorios (mdef): unen a un término con su definición, sustituyendo o complementando la función de la PVD. En los atributos se señalan como mdef= dp, viñ, par, gui, cll. 2) marcadores tipográficos (mt): indicación de negritas, cursivas, subrayado y otras marcas que dan prominencia al término definido o a la definición, este caso se marca mt= neg,curs,subr,otr.
- PP. Patrones Pragmáticos: Dan información sobre el uso de los términos. Los tres patrones considerados en este rubro son: Autoría (Aut), instruccionales (Inst) y temporales (Temp).
- Cf. Correferencia: Contiene las relaciones de referencia que se dan dentro del CD. En los atributos se marca si la *Cf* se da con el término (TERM) o con cualquier otro elemento del CD que opere como referente (ORef). Se especifica si la *Cf* es una *frase nominal* (fn), *frase nominal con demostrativo* (frdem), o tiene otra estructura (otr). A partir de números se marca el *índice* de la Cf (idcf) que permite ligarla con su referente (REF).

[2] http://www.iling.unam.mx:8080/CorcodeAppV/

- Anf. Anáfora: Marca las anáforas dentro del CD. En los atributos se marca si la *Anf* se da con el término (TERM) o con cualquier otro elemento del CD que opere como referente (ORef); se especifica también el tipo de anáfora o tipo de pronombre. Igual que en el caso anterior, el *índice* para ligar con su referente, es señalado con números.
- REF. Referente: Contiene al referente (REF) o antecedente de las correferencias y a las anáforas presentes en el CD. En los atributos se señala como índice (indcf/indanf), si el término definido (TERM) es el referente o es cualquier otra entidad (ORef) del CD.

La estructura queda ilustrada jerárquicamente en la figura 4.

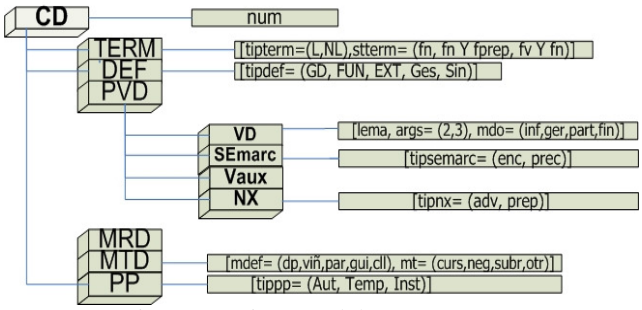

Figura 4: Etiquetas del CORCODE

11. Aplicaciones

Como hemos visto, las aplicaciones del empleo de la metodología aquí descrita para extraer CDs de textos de especialidad a partir de patrones verbales son diversas. En el grupo de Ingeniería Lingüística hemos trabajado en tres principales, las cuales describo a continuación.

11.1 Bancos de conocimiento

Como mencioné en la introducción, un aspecto relevante dentro de las investigaciones realizadas por el Grupo de Ingeniería Lingüística es el desarrollo de *bases de conocimientos léxico* (BCL) para diccionarios onomasiológicos electrónicos, las cuales incorporan de manera pertinente información lingüística, codificada en un nivel léxico, que ayuda a mejorar las consultas que hacen los usuarios. De manera general, las BCL son sistemas de bases de datos que almacenan, administran y proporcionan conocimientos obtenidos del lenguaje natural, a partir de textos tales como diccionarios, glosarios, artículos, etc. [63, 82].

El *diccionario onomasiológico* constituye un recurso léxico que permite a un usuario localizar la palabra adecuada para designar una idea que tiene en mente respecto a alguna cosa. En concreto, la intención de este diccionario es que a partir de conceptos o descripciones elaboradas por un usuario en lenguaje natural, el diccionario proporcione términos relacionados con dichas descripciones, en particular dentro de un dominio técnico o de especialización [77].

Un fenómeno que se ha observado a partir experimentos en torno a los modos de consulta onomasiológica en diccionarios electrónicos, es el amplio rango de posibilidades que tiene un usuario para codificar un concepto en una definición. Como señalan Lara [48] y Sager [71], existen diferentes métodos ofrecidos por el lenguaje natural para estructurar un concepto, más allá de la vía *Genus* y *Differentia* de la definición analítica. Se puede considerar entonces que los usuarios generan *definiciones libres*, las cuales se asocian a un término en particular; se trata de un proceso por el cual una persona, a partir de una idea, deduce la palabra que sirve para designar algo y que, en algún momento, se halla "en la punta de la lengua".

Dado que el diccionario onomasiológico arroja términos a partir de la descripción de los conceptos proporcionados por el usuario, la BCL requiere un módulo primario de adquisición de datos que concentre y amalgame la información conceptual que el usuario busca relacionar con una palabra específica. Para esto, es necesario considerar, además de la información contenida en diccionarios y enciclopedias, la información definitoria dada por los documentos de especialidad.

Por esta razón, la extracción de CDs resulta esencial, pues con la metodología mostrada se puede obtener cuatro tipos de definiciones: analíticas, funcionales, extensionales y sinonímicas. Además, todas ellas desde el punto de vista del experto que normalmente va más allá de la opinión del lexicógrafo.

11.2 Extracción de relaciones léxicas

Las relaciones léxicas (RLs) son un tipo de relación producida a partir del significado que contiene una palabra [66, 79]. El contenido de significación puede configurar dos tipos de situaciones:

Por un lado, como indica Fillmore [40], el contenido léxico de una palabra puede proyectar un escenario en donde se sitúan varios elementos que cumplen determinadas funciones acordes con dicho escenario. Por ejemplo, ciertos verbos de acción como *correr* configuran un escenario donde se necesita un agente que realice la acción, con una locación donde se lleve a cabo tal acto, una trayectoria que señale la ruta a seguir, una temporalidad que indique cuándo se realizó, etc.

Por otro lado, para Cruse [33] una palabra puede fijar una serie de relaciones con otras palabras que tengan un significado cercano a ésta. Por ejemplo, en el significado de un verbo como *correr* pueden encontrarse conceptos relacionados:

jerárquicamente superiores (p.e., *correr* es un tipo de acción); con un significado similar (trotar, acelerar); o con un significado contrario (*caminar*).

En el caso concreto de los lenguajes especializados, las RLs pueden servir para representar el sistema de conceptos de un campo de conocimiento específico. Dicho sistema constituye una especie de mapa donde se establece el lugar y la situación específica de un término frente a los demás de su mismo campo de conocimiento.

El desarrollo de un sistema de conceptos contempla la necesidad de conocer el significado de los términos. En el caso específico de los CDs, como unidades textuales que ayudan a describir el significado de un término, se pueden considerar como un repertorio de relaciones léxicas. En los CDs se establece una relación específica entre el término y su definición a partir del tipo de verbo definitorio que los une. Tal es el caso de las relaciones sinonímicas que se pueden distinguir con patrones verbales definitorios como *también llamado* o *también conocido como*. En otras situaciones, los verbos pueden indicar relaciones léxicas de función o extensión. Por ejemplo, en CDs con patrones como *consiste de*, *consta de*, *formado por*, *constituido por*, denotan una relación de extensión respecto al término que se define [21].

Las RLs son fundamentales para elaborar ontologías, tesauros, terminologías y otros recursos lingüísticos similares. Contar con herramientas para la identificación automática de relaciones léxicas permitirá su implementación en sistemas de pregunta-respuesta, web semántica, minería de textos e interfaces inteligentes, por mencionar algunos ejemplos. Desarrollar métodos automáticos con esta idea en mente implica crear perfiles sofisticados para repositorios de textos, los que serán necesarios en la siguiente generación de herramientas para el descubrimiento de recursos textuales tanto en Internet como en colecciones enormes de textos.

Si bien para el inglés existen varios sistemas de RLs, para el español son contados o casi nulos, y en general se trata de adaptaciones del inglés. Ahora, contar con una metodología y una herramienta para extraer relaciones léxicas que tome en cuenta el comportamiento lingüístico real del español tiene un impacto científico de gran valor para terminólogos y lexicógrafos, a la vez que permite la creación de otros recursos computacionales para nuestra lengua.

Ahora bien, es posible plantear la extracción de RLs a partir del análisis de los patrones verbales que aparecen como elementos constitutivos en definiciones localizadas en textos especializados.

Un hecho observado a raíz de esta investigación es la existencia de una relación estrecha entre el tipo de definición y el verbo que aparece como núcleo de un patrón verbal definitorio (PVD), lo que permite postular una taxonomía de cuatro tipos de definiciones basada en el tipo de PVD que aparece en el CD:

- Analítica: aquella definición que presenta de forma explícita un género próximo y una diferencia específica, por ejemplo: *una computadora es una máquina que resuelve operaciones lógicas*, donde el género próximo al que pertenece computadora es *máquina*, y las diferencias específicas son *que resuelve operaciones lógicas*.
- Sinonímica: aquella definición que manifiesta exclusivamente un género próximo, el cual establece una relación de equivalencia o sinonimia, por ejemplo: *un ordenador se llama también computadora*.
- Extensional: aquella donde se muestra una relación meronímica que enumera las partes que conforman una entidad, por ejemplo: *una computadora se compone de software, hardware y periféricos*.
- Funcional: aquella definición que describe la función o el uso de una entidad particular, por ejemplo: *una computadora sirve para resolver problemas lógicos, matemáticos y estadísticos*.

Esta clase de patrones, así como el comportamiento que presentan cuando aparecen ligados a una clase de definición específica, ha dado pie a que diferentes autores [25, 81, 84] reconozcan en ellos distintos tipos de RLs. Siguiendo la propuesta de Cruse [33], aquí se plantea la posibilidad de reconocer en los tipos de definiciones arriba expuestos las siguientes relaciones:

- Hiponimia-Hiperonimia: Una entidad hiponímica se deriva de un hiperónimo o elemento superior, por ejemplo: *una autobiografía es un libro*.
- Sinonimia: Dos entidades que mantienen cierta equivalencia a nivel cognitivo, por ejemplo: *Una mujer policía es un policía femenino*.
- Antonimia: Dos entidades que tienen un significado opuesto, por ejemplo: alto/bajo, computadora/calculadora, encender/apagar, entre otras.
- Individuación: Aquellas entidades donde aparece un cambio de individuación. Existen dos tipos de individuación: a) cantidad/masa, es decir, una relación entre una porción o una pieza y una cierta sustancia o entidad, por ejemplo: *Una hora es una porción de tiempo*; b) miembro/grupo, que es una relación entre una entidad que puede ser inherente a un grupo o colectivo, por ejemplo: *Un policía es un miembro de la fuerza policíaca*.

Así, se puede observar que con la clasificación de las relaciones posibles entre las definiciones, los patrones verbales asociados a definiciones y el agrupamiento automático, es posible la formulación de un algoritmo para la extracción automática de relaciones léxicas y, aun mejor, de definiciones.

11.3 El sistema Describe

Una aplicación directa del ECODE es el sistema denominado Describe® para la búsqueda, clasificación y agrupamiento de definiciones en la Web. La metodología parte de utilizar robots para indexar constantemente páginas que contengan alguno de los 2 millones de términos en el área de medicina. Estas páginas constituyen nuestra base de datos inicial para la extracción de contextos definitorios. Una vez extraídos los diferentes tipos de definiciones, éstos se clasifican según su tipo y se agrupan de acuerdo con el contenido semántico que en ellos se vincula.

Describe es una aplicación de arquitectura cliente-servidor orientada a Web, compuesta por varios módulos que permiten organizar la información disponible en Internet.

Del lado del servidor, el sistema está conformado por los módulos siguientes (Fig. 5):

- Extractor: Módulo encargado de extraer de Internet candidatos a CDs.
- Etiquetador: Permite etiquetar el texto de los candidatos a CDs proporcionados por el extractor.
- ECODE: Procesa el texto etiquetado e identifica los CDs finales, clasificándolos en los tres tipos de definición.
- Agrupamiento: Agrupa los CDs de acuerdo con sus características.

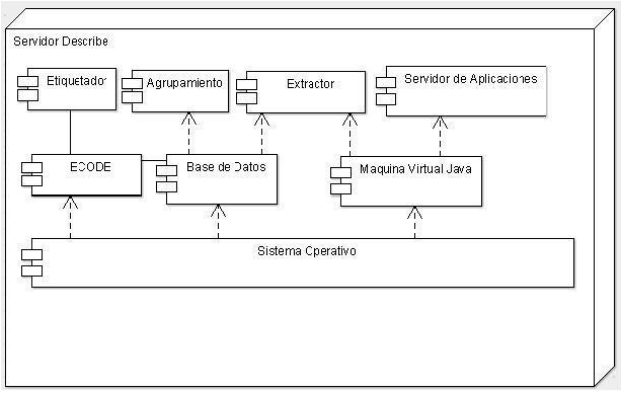

Figura 5: Diagrama del Describe

- Maquina Virtual de Java: Componente que permite ejecutar el Extractor de candidatos, independientemente de la plataforma o sistema operativo.
- Servidor de Aplicaciones: Permite al usuario interactuar con el sistema en ambiente Web.
- Sistema Operativo: Aplicación sobre la que se ejecutan todas las aplicaciones y módulos residentes en la máquina del servidor, permitiendo administrar y gestionar eficazmente sus recursos.

Dos Módulos vitales en el Describe son el ECODE y el de agrupamiento. Hemos visto en este artículo que el ECODE es un método satisfactorio para la extracción de definiciones en textos, además clasificadas en diferentes tipos. Este método, como se ha mostrado, sirve no sólo para el Describe, sino para otras aplicaciones, como la extracción de relaciones semánticas, elaboración de diccionarios semasiológicos y onomasiológicos, obtención de bases de conocimientos léxicas, etc.

El algoritmo de agrupamiento utilizado es un método novedoso que involucra una técnica adaptada de resúmenes automáticos y adecuada para fungir como medida de similitud. Este algoritmo, además de su uso para el Describe, será de utilidad para organizar los resultados (snippets) en motores de búsqueda.

El sistema Describe, de esta manera, apuesta a ser un buscador de definiciones con base en la web y será de gran utilidad tanto para especialistas como para individuos que deseen profundizar en el significado de un término especializado. Por ahora se trabaja en el área de medicina y se tiene contemplado ampliar el alcance de esta herramienta a otras áreas de conocimiento.

12. Agradecimientos

Esta investigación ha sido financiada por el Consejo Nacional de Ciencia y Tecnología, CONACYT, a través de los proyectos 46832 “Extracción de conceptos en textos de especialidad a través del reconocimiento de patrones lingüísticos y metalingüísticos”, 54616 “Análisis lingüístico de definiciones en contextos definitorios”, 82050 “Extracción de relaciones léxicas para dominios restringidos a partir de contextos definitorios en español” y de la beca doctoral CONACYT/Fundación Carolina 179210. Asimismo, bajo el patrocinio de DGAPA-UNAM, con el proyecto IN403108 “Extracción de relaciones semánticas a partir de definiciones en textos de especialidad”.

Un agradecimiento especial a los que en el marco de esta investigación realizaron estudios particulares y documentaron en su tesis, tanto a nivel de licenciatura como de maestría o doctorado: César Aguilar, Rodrigo Alarcón, Alberto Barrón, Valeria Benítez, Ariadna Hernández, Alejandro Molina y Octavio Sánchez. A Carme Bach que participó como codirectora de la tesis de doctorado de Rodrigo Alarcón. A los demás miembros del

Grupo de Ingeniería Lingüística que aportaron con su trabajo o en las discusiones: Edwin Aldana, Gabriel Castillo, Alfonso Medina, Víctor Mijangos y Carlos Rodríguez.

13. Referencias

13.1 Publicaciones del proyecto

[1] Aguilar, César. 2009. *Análisis lingüístico de definiciones en contextos definitorios*. Tesis de Doctorado, UNAM, México.

[2] Aguilar, César, Rodrigo Alarcón, Carlos Rodríguez y Gerardo Sierra. 2006. Reconocimiento y clasificación de patrones verbales definitorios en corpus especializados. En *La terminología en el siglo XXI: contribución a la cultura de la paz, la diversidad y la sostenibilidad*, editado por M. T. Cabré, R. Estopà, C. Tebé. Barcelona, IULA, Documenta Universitaria.

[3] Aguilar, César y Gerardo Sierra. 2008. Hacia una tipología de definiciones basada en el modelo analítico, *Memorias del XV Congreso Internacional ALFAL 2008*, Montevideo, Uruguay.

[4] Aguilar, César y Gerardo Sierra. 2009. Reconocimiento de definiciones asociadas a frases predicativas en contextos definitorios. *Procesamiento de Lenguaje Natural*, 43:151-158.

[5] Aguilar, César y Gerardo Sierra. 2009. A formal scope on the relations between definitions and verbal predications. *1st International Workshop on Definition Extraction*, Borovets, Bulgaria.

[6] Alarcón, Rodrigo. 2003. *Análisis de contextos definitorios en textos de especialidad*, Tesis de Licenciatura, UNAM, México.

[7] Alarcón, Rodrigo. 2009. *Extracción automática de contextos definitorios en corpus especializados*. Tesis de Doctorado, Universidad Pompeu Fabra, Barcelona.

[8] Alarcón, Rodrigo, Carme Bach C y Gerardo Sierra. 2008. Extracción de contextos definitorios en corpus especializados: Hacia una elaboración de una herramienta de ayuda terminográfica. *Revista Española de Lingüística* 37:247-278.

[9] Alarcón, Rodrigo y Gerardo Sierra. 2002. Hacia la extracción automática de conceptos. *Proc. VIII Simposio Iberoamericano de Terminología*. Red Iberoamericana de Terminología RITerm, Cartagena, Colombia.

[10] Alarcón, Rodrigo y Gerardo Sierra. 2003. El rol de las predicaciones verbales en la extracción automática de conceptos. *Estudios de Lingüística Aplicada*, 21(38):129-144.

[11] Alarcón, Rodrigo, Gerardo Sierra G y Carme Bach. 2008. ECODE: A Pattern Based Approach for Definitional Knowledge Extraction. *XIII EURALEX International Congress*, Barcelona.

[12] Alarcón, Rodrigo, Gerardo Sierra y Carme Bach. 2009. Description and Evaluation of Definition Extraction System for Spanish language. *1st International Workshop on Definition Extraction*, Borovets, Bulgaria.

[13] Benítez, Valeria. 2008. *Anáforas en la expansión de Contextos Definitorios: una propuesta de etiquetado.* Tesis de Licenciatura, UNAM, México.

[14] Barrón, Alberto. 2007. *Extracción automática de términos en contextos definitorios*. Tesis de Maestría, UNAM, México.

[15] Hernández, Ariadna. 2009. *Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática*. Tesis de Licenciatura, UNAM, México.

[16] Molina, Alejandro. 2009. *Agrupamiento automático de contextos definitorios*. Tesis de Maestría, UNAM, México.

[17] Sánchez, Octavio. 2009. *Análisis de relaciones léxicas en definiciones analíticas, extensionales y funcionales*. Tesis de Licenciatura, UNAM, México.

[18] Sierra, Gerardo y Rodrigo Alarcón. 2002. Identification of recurrent patterns to extract to definitory contexts. *Lecture notes in Computer Science* 2276:436-438.

[19] Sierra, Gerardo y Rodrigo Alarcón. 2003. The Role of Verbal Predications for Definitional Contexts Extraction. *TIA 2003*, Strasbourg: Université de Strasbourg.

[20] Sierra, Gerardo, Rodrigo Alarcón y César Aguilar. 2006. Extracción automática de contextos definitorios en textos especializados. *Procesamiento de Lenguaje Natural* 37:351-352.

[21] Sierra, Gerardo, Rodrigo Alarcón, César Aguilar y Carme Bach. 2008. Definitional verbal patterns for semantic relation extraction. *Terminology* 14(1):74-98.

[22] Sierra, Gerardo, Rodrigo Alarcón, Alfonso Medina, César Aguilar. 2004. Definitional contexts extraction from specialised texts. En *Practical Applications in Language and Computers*, editado por Barbara Lewandowska. Frankfurt: Peter Lang.

[23] Sierra, Gerardo, Gabriel Castillo, Antonio Reyes y Rodrigo Alarcón. 2001. Desarrollo de la Ingeniería Lingüística en la UNAM, México. *II Taller Internacional de Procesamiento Computacional del Español y Tecnologías del Lenguaje*. Jaén, España.

[24] Sierra, Gerardo, Alfonso Medina, Rodrigo Alarcón y César Aguilar. 2003. Towards the

extraction of conceptual information from corpora. *Proceedings of the Corpus Linguistics 2003 conference*, editado por Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. UCREL Technical Paper, No. 16, Lancaster University.

13.2 Bibliografía

[25] Alshawi, Hiyan. 1987. Processing Dictionary Definitions with Phrasal Pattern Hierarchies. *Computational Linguistics* 13(3-4):195-202.

[26] Auger, Alain. 1997. *Repérage des énoncé d'intér*êt *définitoire dans les bases de données textuelles*. Tesis de doctorado, Neuchâtel, Universidad de Neuchâtel.

[27] Bach, Carme. 2005. Los marcadores de reformulación como localizadores de zonas discursivas relevantes en el discurso especializado. *Debate Terminológico* 1.

[28] Cabré, Teresa. 1993. *La terminología. Teoría, metodología y aplicaciones*, Barcelona: Antártica.

[29] Cabré, Teresa. 1999. *La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

[30] Cabré, Teresa, Rosa Estopà y Jorge Vivaldi. 2001. Automatic term detection. A review of current systems. En *Recent Advances in Computational Terminology*, editado por Bourigault, D, Jaquemin, C, & L'Homme, M.C. Amsterdam: Benjamis.

[31] Calzolari, Nicoletta y Eugenio Picchi. 1988. Acquisition of Semantic Information from an On-Line Dictionary. *12th International Conference on Computational Linguistics*, *Coling'88*. Budapest.

[32] Cowie, Jim y Yorick Wilks. 2000. "Information extraction". En *Handbook of Natural Language Processing,* editado por R. Dale, H. Moisl and H. Somers. New York, Marcel Dekker.

[33] Cruse, D.A. 1986. *Lexical semantics.* Cambridge: Cambridge University Press.

[34] De Bessé, Bruno. 1991. Le Contexte Terminographique. *Meta* 26(1):111-120.

[35] Estopà, Rosa. 2001. Elementos lingüísticos de las unidades terminológicas para su extracción automática", en *La terminología científico-técnica*, editado por Cabré T, Feliu J., IULA-UPF, Barcelona.

[36] Estopà, Rosa, Jorge Vivaldi y Teresa Cabré. 1998. Sistemes d'extracció automática de candidats a terme. Estat de la qüestió. *Papers de l'IULA*, Série Informes, 22.

[37] Fajardo, Juan y Héctor Jiménez. 2003. Determinación de relaciones léxicas con base en el grado de subsunción. *Estudios de Lingüística Aplicada*, 22(38):81-87.

[38] Fernández, María del Carmen. 1999. *Las preposiciones en español. Valores y usos Construcciones Preposicionales*. Salamanca: Colegio de España.

[39] Fernández, Silvia, Eric San Juan y Juan Manuel Torres Moreno. 2008. Enertex: un sisteme basé sur l´energie textuelle. Traitement Automatique de la Langue Naturelle, Avignon.

[40] Fillmore, Charles. 1968. The case for case. En *Universals in Linguistic Theory,* Ediatod por Bach y Harms. New York: Holt, Rinehart and Winston.

[41] Haensh, Günther, Lothar Wolf, Stefan Ettinger y Reinhold Werner. 1982. *La lexicografía*, *de la lingüística teórica a la lexicografía práctica*. Madrid: Gredos.

[42] Hearst, Marti. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the 14th International Conference on Computational Linguistics, Coling'92*. Nantes.

[43] Heid, Ulrich, Susanne Jauss, Katja Krüger y Andrea Hohmann. 1996. Term Extraction with standard tools for corpus exploration". *4th International Congress on Terminology and Knowledge Engineering*, Viena.

[44] Jacquemin, Christian. 1996. A symbolic and surgical acquisition of terms through variation. En *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, editado por S. Wermter, E. Riloff y G. Scheler. Springer:Heidelberg.

[45] Jacquemin, Christian y Didier Bourigault. 2003. Term Extraction and Automatic Indexing. En *Handbook of Computational Linguistics*, editado por R. Mitkov, Oxford: Oxford University Press.

[46] Jurafsky, Daniel y James Martin. 2000. *Speech and Language Processing*. *An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Nueva Jersey: Upper Saddle River Prentice.

[47] Kilgarriff, Adam, Pavel Rychly, Pavel Smrz y David Tugwell. 2004. The Sketch Engine. *Proceedings of Euralex*, Lorient.

[48] Lara, Luis Fernando. 1997. *Teoría del diccionario monolingüe*, México: COLMEX.

[49] L'Homme, Marie-Claude. 2002. What can Verb and Adjectives tell us about Terms?. *Proc. Terminology and Knowledge Engineering, TKE 2002*. Nancy.

[50] L'Homme, Marie-Claude. 2005. Conception d'un dictionnaire fondamental de l'informatique

et de l'Internet : sélection des entrées, *Le langage et l'homme* 40(1):137-154.

[51] López, María López. 1972. *Problemas y métodos en el análisis de preposiciones*. Madrid: Gredos.

[52] Malaisé, Verónica. 2005. *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles á partir de corpus textuels*. Tesis de doctorado. Paris, Université Paris 7–Denis Diderot.

[53] Martín, María Antonia. 1999. Los marcadores del discurso. En *Gramática descriptiva de la lengua española,* editado por Bosque, I, Demonte, V. Madrid: Espasa.

[54] Medina, Alfonso, Gerardo Sierra, Gabriel Garduño, Carlos Méndez y Roberto Saldaña. 2004. CLI: An open Linguistic Corpus for Engineering. *Proc. Ibero-America Workshop on Artificial Intelligence,* Puebla, México.

[55] Meyer, Ingrid. 2001. Extracting a knowledge-rich contexts for terminography: A conceptual and methodological framework. En *Recent Advances in Computational Terminology,* editado por Bourigault, D.; Jaquemin, C. & L'Homme, M.C. Philadelphia: John Benjamins.

[56] Modrak, Deborah K.W. 2001. *Aristotle's Theory of Language and Meaning*, Cambridge: Cambridge University Press.

[57] Monachesi, Paola, Dan Cristea, Diane Evans, Alex Killing, Lothar Lemnitzer, Kiril Simov, Cristina Vertan. 2006. Integrating Language Technology and Semantic Web techniques in eLearning. Proc. ICL, Villach, Austria.

[58] Muresan, Smaranda y Klavans, Judith. 2002. A Method for Automatically Building and Evaluating Dictionary Resources. *Proc. 3th International Conference on Language Resources and Evaluation (LREC'02).* Las Palmas.

[59] Navigli, Roberto y Paola Velardi. 2007. GlossExtractor: A Web Application to Automatically Create a Domain Glossary. *Lecture Notes in Computer Science* 4733:339-349.

[60] Osinski, Stanis, Jerzy Stefano y Dawid Weiss. 2004. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. *Proc. Intelligent Information Systems*.

[61] Pavón, María Victoria. 1999. Clases de partículas: preposición conjunción y adverbio. En *Gramática descriptiva de la lengua española Vol 1. Sintaxis básica de las clases de palabras,* editado por Ignacio Bosque y Victoria Demonte. Madrid: Espasa.

[62] Pearson, Jennifer. 1998. *Terms in Context*, Philadelphia, John Benjamins.

[63] Pérez, Chantal. 2002. Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento, *Estudios de Lingüística Española* 18.

[64] Pinto, Ana Sofía y Oliveira, Débora. 2004. Extracção de Definições no Corpógrafo. Thecnical report. Faculdade de Letras da Universidade do Porto.

[65] Porter, Martin. 1980. An alghorithm for suffix stripping. *Readings in information retrieval,* San Francisco CA: Morgan Kaufmann Publisher Inc

[66] Pustejovsky, James. 1998. "Issues in text-based lexicon acquisition". *Corpus processing for lexical acquisition,* editado por B. Boguraev y J. Pustejovsky. Cambridge: The MIT Press.

[67] Pustejovsky, James, Sabine Bergler y Peter Anick. 1993. Lexical Semantic Techniques for Corpus Analysis. *Computational Linguistics* 19(2): 331-358.

[68] Rebeyrolle, Josette. 2000. *Forme et fonction de la définition en discours*, Tésis de doctorado, Université Toulouse-Le Mirail.

[69] Rebeyrolle, Josette y Ludovic Tanguy. 2000. Repérage automatique de structures linguistiques en corpus: le cas des énoncés définitoire. *Cahiers de Grammaire* 25:153-174.

[70] Rodríguez, Carlos. 2004. Metalinguistic Information Extraction from specialized texts to enrich computational lexicons. Tesis de Doctorado. Universitat Pompeu Fabra, Barcelona.

[71] Sager, Juan Carlos. 1990. *A Practical Course in Terminology Processing*, Philadelphia: John Benjamins.

[72] Sager, Juan Carlos. 2001. *Essays on Definitions*, Philadelphia: John Benjamins.

[73] Saggion, Horacio. 2004. Identifying Definitions in Text Collections for Question Answering. *Proc. 4th International Conference on Language Resources and Evaluation LREC2004,* Lisboa.

[74] Sánchez, A. y Melva Márquez. 2005. Hacia un sistema de extracción de definiciones en textos jurídicos. *Actas de la 1er Jornada Venezolana de Investigación en Lingüística e Informática*. Venezuela.

[75] Saurí, Roser. 1997. *Tractament Lexicogràfic dels Adjectius*, Sèries Monografies, IULA-UPF, Barcelona.

[76] Seiler, Bernhard y Wolfgang Wannenmacher. 1983. *Concept development and the development of the word meaning,* Berlin: Springer Verlag.

[77] Sierra, Gerardo y John McNaught. 2000. Design of an onomasiological search system: A concept-oriented tool for terminology. *Terminology,* 6(1): 1-34.

[78] Storrer, Angelika y Sandra Wellinghoff. 2006. Automated Detection and Annotation of Term Definitions in German Text Corpora. *Proc. 5th International Conference on Language Resources and Evaluation (LREC'06)*. Génova.

[79] Valero, Esperanza y Amparo Alcina. 2009. Linguistic realization of conceptual features in terminographic dictionary definitions. *Proc. 1st. International Workshop on Definition Extraction*. Borovets

[80] Vivaldi, Jorge. 1995. Proyectos del IULA: El corpus técnico, Simposio de Lingüística Hispánica. Instituto Cervantes y Universidad de Manchester, Manchester.

[81] Vossen, Piek y Ann Copestake. 1993. Untangling Definition Structure into Knowledge Representation. En *Inheritance, Defaults and the Lexicon.* Cambridge University Press.

[82] Walker, Donald y Robert Amsler. 1986. The Use of Machine-Readable Dictionaries in Sublanguage Analysis. En *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, Hillsdale: New Jersey.

[83] Wilks, Yorick. 1997. Information extraction as a core language technology. En *Information Extraction,* editado por M. T. Pazienza, Berlin: Springer.

[84] Wilks, Yorick, Brian Slator y Louise Guthrie 1996. *Electric Words. Dictionaries, Computers and Meaning*, MIT Press: Cambridge.

Artigos de Investigação

Kernels para la clasificación de preguntas en español y catalán

David Tomás y José L. Vicedo
Depto. de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{dtomas,vicedo}@dlsi.ua.es

Resumen

Este artículo presenta una aproximación a la clasificación automática de preguntas en español y catalán. El sistema de clasificación está basado en el algoritmo SVM y en el uso de diferentes funciones kernel, empleando únicamente características textuales superficiales que permiten la obtención de un sistema fácilmente adaptable a diferentes idiomas. Se ha realizado un estudio sobre el correcto ajuste de parámetros de los kernels, la precisión de los mismos, la definición de distintos vectores de características de aprendizaje y el rendimiento en función del idioma de trabajo. Adicionalmente, se ha experimentado con el algoritmo LIBLINEAR, aplicado aquí por vez primera a la tarea de clasificación de preguntas. Con este algoritmo, así como con los kernels definidos, se han obtenido valores de precisión por encima del 80 % para los dos idiomas tratados, superando a otros algoritmos tradicionales de clasificación. Para el entrenamiento y evaluación del sistema se ha desarrollado un corpus paralelo de 2.393 preguntas en inglés, español y catalán.

1. Introducción

Los sistemas de *búsqueda de respuestas* o *question answering* (QA) tienen como finalidad encontrar respuestas concretas a necesidades precisas de información formuladas por los usuarios en lenguaje natural. Estos sistemas se han convertido en objeto de amplio estudio en la última década, gracias en parte a los distintos foros internacionales desarrollados en este campo: TREC[1] (Voorhees, 2001), CLEF[2] (Vallin et al., 2006) y NTCIR[3] (Kando, 2005). Estos foros han marcado las pautas de desarrollo de estos sistemas, estableciendo los retos a superar y el marco para su evaluación.

La mayoría de sistemas de QA presentan una arquitectura común, organizando su funcionamiento en tres fases bien diferenciadas que habitualmente tienen lugar de forma secuencial (Voorhees, 1999): *análisis de la pregunta*, *selección de documentos o pasajes relevantes* y *extracción de la respuesta*. Dentro de la fase de análisis de la pregunta tiene lugar la *clasificación de preguntas*[4]. El objetivo de esta clasificación es identificar de forma automática qué se está preguntando, categorizando las preguntas en diferentes clases semánticas en función del tipo de respuesta esperado. Por ejemplo, ante una pregunta como "¿Quién es el presidente de los Estados Unidos?", un sistema de clasificación de preguntas detectaría que se está preguntando por una persona, mientras que para "¿Dónde está la Torre Eiffel?" identificaría que se está preguntando por un lugar. En estos ejemplos, *persona* y *lugar* representan la clase semántica de la respuesta esperada.

La clasificación de preguntas en los sistemas de QA tiene un doble propósito. En primer lugar, proporciona una restricción semántica a las respuestas esperadas, permitiendo filtrar un gran número de ellas durante la fase final de extracción. Por ejemplo, cuando se considera la pregunta "¿Cuál es la ciudad más grande de Alemania?", detectar que se está preguntando por un *lugar* permite descartar un gran número de respuestas candidatas, manteniendo únicamente aquellas que sean nombres de localizaciones. El segundo propósito de la clasificación es proporcionar información a los procesos subsiguientes del sistema de QA para establecer la estrategia de selección de respuestas, así como las bases de conocimiento que el sistema requiera para obtener la respuesta final.

La importancia de la clasificación de preguntas en el resultado global de los sistemas de QA ha quedado patente en diversos estudios. Radev et al. (2002) detectaron que una clasificación incorrecta de la pregunta conlleva que la posibilidad del sistema de obtener una respuesta correcta sea 17 veces menor. En otro análisis, realizado

[1] http://trec.nist.org.

[2] http://clef-campaign.org.

[3] http://research.nii.ac.jp/ntcir/.

[4] Algunos de los nombres que recibe esta tarea en la literatura anglosajona son *question classification*, *question categorization* y *answer type recognition*.

por Moldovan et al. (2003) sobre los errores ocurridos en un sistema de QA en domino abierto, se revela que más de un 35 % de éstos son directamente imputables al módulo de clasificación de la pregunta.

Al igual que sucede con el resto de tareas dentro del campo del *procesamiento del lenguaje natural* (PLN), la mayoría de sistemas de clasificación de preguntas están orientados al idioma inglés, siendo muy escasas las contribuciones para otros idiomas. En este trabajo presentamos un estudio sobre la clasificación automática de preguntas en español y catalán. Definiremos un sistema basado en corpus mediante el empleo de kernels y características textuales superficiales, dando como resultado una aproximación fácilmente adaptable a diferentes idiomas.

En el resto de este artículo, la sección 2 presenta una introducción a los métodos y funciones kernel y a su aplicación al PLN. La sección 3 describe los componentes de aprendizaje del sistema y los corpus de preguntas empleados en este trabajo. En la sección 4 se describen los experimentos llevados a cabo y los resultados obtenidos. La sección 5 muestra otras investigaciones relacionados con este trabajo. Finalmente, la sección 6 enumera las conclusiones y trabajo futuro derivado de esta investigación.

2. Métodos y funciones kernel

Los *métodos kernel* o *kernel methods* son un tipo de algoritmo de aprendizaje automático ampliamente utilizado en el campo del PLN debido a sus buenos resultados empíricos (Shawe-Taylor y Cristianini, 2004; Schölkopf y Smola, 2001). La estrategia adoptada por estos métodos consiste en dividir el problema de aprendizaje en dos partes: en primer lugar se trasladan los datos de entrada a un espacio de características adecuado, empleando seguidamente un algoritmo lineal para descubrir patrones no lineales en el espacio de características transformado. Esta transformación del espacio se lleva a cabo de forma implícita mediante una *función núcleo* (*kernel function*, *función kernel* o simplemente *kernel*). Una función kernel proporciona una medida de similitud entre los datos de entrada que depende exclusivamente del tipo de éstos y del dominio.

Formalmente, un kernel es una función simétrica $k : X \times X \rightarrow \mathbb{R}$ que toma como entrada dos objetos (por ejemplo, vectores, textos o árboles sintácticos) y obtiene como salida un número real caracterizando su similitud. Para todo vector de características $\mathbf{x}_i$ y $\mathbf{x}_j \in X$, definimos un kernel como

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle,$$

donde $\phi : X \rightarrow \mathcal{F}$ es una proyección explícita de X (el espacio de características original) a un nuevo espacio de características $\mathcal{F}$.

La selección del kernel apropiado para cada tarea resulta de gran importancia, ya que es éste el que define el espacio de trabajo transformado donde se llevará a cabo el entrenamiento y la clasificación. En este trabajo vamos a estudiar el rendimiento de cuatro funciones kernel, algunas de ellas empleadas de forma habitual en diferentes tareas de clasificación dentro del campo del PLN:

- **Lineal**

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

- **Polinómico**

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \langle \mathbf{x}_i, \mathbf{x}_j \rangle + r)^d, \gamma > 0$$

- **Radial Basis Function (RBF)**

$$k(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2), \gamma > 0$$

- **Sigmoide**

$$k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \langle \mathbf{x}_i, \mathbf{x}_j \rangle + r)$$

Las variables γ, r y d son los parámetros de los kernels. En la sección 4.2 hablaremos más en detalle de estos parámetros.

El kernel es el único elemento específico del dominio en el sistema de clasificación, mientras que el algoritmo de aprendizaje es un componente de propósito general. Potencialmente, cualquier función kernel puede trabajar con cualquier algoritmo basado en kernels, como *Support Vector Machines* (SVM) o *perceptrón*. En nuestros experimentos vamos a emplear SVM como algoritmo para trabajar con los kernels definidos más arriba. SVM ha sido aplicado con éxito en numerosos problemas de PLN, como el análisis sintáctico superficial (Kudo y Matsumoto, 2001) o la clasificación de textos (Joachims, 1998), demostrando su buen funcionamiento en espacios de alta dimensionalidad. En la actualidad, SVM se ha convertido en el algoritmo más popular de los empleados en clasificación de preguntas basados en corpus (Zhang y Lee, 2003). Para llevar a cabo nuestros experimentos hemos utilizado la implementación ofrecida en la librería LIBSVM (Chang y Lin, 2001).

3. Descripción del sistema

Como todo sistema de basado en corpus, nuestro clasificador de preguntas requiere la definición de una serie de elementos:

- Un conjunto de preguntas (corpus) para el entrenamiento y la evaluación del sistema.
- Una taxonomía de tipos de pregunta con la que queremos clasificar las entradas que lleguen al sistema.
- Un conjunto de características de aprendizaje extraídas del corpus que refleje la información relevante de cada ejemplo.
- Un algoritmo que aprenda a predecir la clase a la que pertenece cada nueva entrada a partir de las características de aprendizaje.

El algoritmo de aprendizaje ya fue descrito en el apartado anterior. En esta sección describiremos el resto de componentes del sistema.

3.1. Corpus

Al igual que sucede con otros recursos lingüísticos en el campo de PLN, los corpus de preguntas en idiomas diferentes al inglés son muy escasos. Existen algunos corpus disponibles en diferentes idiomas, como DISEQuA[5] (Magnini et al., 2003) (450 preguntas en holandés, italiano, español e inglés) o Multieight-04[6] (Magnini et al., 2005) (700 preguntas en alemán, inglés, español, francés, italiano, holandés y portugués). El problema de estos conjuntos de preguntas es su reducido tamaño, que los hace poco apropiados para el entrenamiento y evaluación de sistemas basados en corpus. Por ello, en este trabajo hemos desarrollado nuestros propios corpus para español y catalán, idioma este último en el que no hay constancia de conjuntos de preguntas disponibles para este tipo de sistemas.

El corpus en español se empleó por primera vez en (Tomás et al., 2005). Para formalizar este corpus se recopilaron las preguntas de evaluación en inglés definidas para la tarea de QA de las conferencias TREC, desde 1999 (TREC-8) hasta 2003 (TREC-12).[7] Una vez recopiladas las preguntas en inglés, se procedió a la traducción manual de las mismas. En el caso del español, se partió de las traducciones de las preguntas del TREC-8, TREC-9, TREC-10 y TREC-11 realizadas por el Grupo de Procesamiento del Lenguaje Natural de la UNED.[8] Para obtener el mismo corpus que en inglés se tradujeron las preguntas del TREC-12 y se revisaron todas las anteriores a fin de obtener una traducción uniforme. En el caso del corpus en catalán, se tradujeron íntegramente las preguntas en inglés, obteniendo finalmente un corpus paralelo de 2.393 preguntas en los tres idiomas. Este corpus, etiquetado con la taxonomía definida en el siguiente apartado, está libremente disponible para la comunidad científica.[9]

3.2. Taxonomía

Una vez recopiladas las preguntas, el siguiente paso consistió en el etiquetado manual de éstas con su correspondiente clase semántica. Las preguntas originales del TREC no presentan ningún tipo de etiquetado, ya que son los propios participantes de estas conferencias los que deciden la taxonomía de clases que más conviene a su sistema de QA. Al no existir ninguna taxonomía estándar en el campo de QA y de los sistemas de clasificación de preguntas, definimos una propia tomando como base la jerarquía extendida de entidades nombradas de Sekine (Sekine, Sudo, y Nobata, 2002). Esta jerarquía fue diseñada para cubrir aquellas entidades que, dicho de manera informal, aparecen habitualmente en textos periodísticos. El objetivo de Sekine era el de dar cobertura a entidades nombradas más específicas que las dadas habitualmente en los sistemas de extracción de información. De esta manera, no intenta cubrir ningún dominio particular, sino abordar el etiquetado de entidades en textos de carácter general en dominio abierto. La propuesta final de Sekine se traduce en una jerarquía de cerca de 150 tipos de entidades nombradas de carácter general. La intención al diseñarla de forma jerárquica era que pudiera ajustarse fácilmente a diferentes tareas según el grado de refinamiento de las entidades a detectar.

Entre los sistemas que inspiraron el diseño de esta jerarquía están aquellos empleados en la tarea de QA del TREC. Este detalle hace que la cobertura que proporciona esta taxonomía sea especialmente adecuada para etiquetar las preguntas que tienen lugar en nuestro corpus. Para este etiquetado utilizamos como base la jerarquía descrita por Sekine, centrándonos en las etiquetas que aparecen en el primer nivel. Sobre esta base se añadieron las clases *DEFINITION* y *ACRONYM*. Éstas no existían originalmente en la jerarquía de Sekine (ya que se centra en entidades) pero se han incluido para aumentar la cober-

[5] `http://clef-qa.itc.it/2004/down/DISEQuA\_v1.0.zip`.

[6] `http://clef-qa.itc.it/2005/down/corpora/multieight-04\_v1.2.zip`.

[7] `http://trec.nist.gov/data/qa.html`.

[8] `http://nlp.uned.es`.

[9] `http://www.dlsi.ua.es/~dtomas/resources/`.

tura de la taxonomía al ser dos tipos de pregunta que se dan de forma habitual en las conferencias TREC.

Una vez definida la taxonomía de clases se pasó al etiquetado de las 2.393 preguntas por parte de dos revisores, obteniendo un *índice kappa* o *kappa agreement* de 0,87. El acuerdo esperado se calculó según la descripción de (Fleiss, 1971) tomando como igual para los revisores la distribución de proporciones sobre las categorías. En caso de no haber acuerdo entre ambos revisores, una tercera persona intervino en la decisión final.

La figura 1 muestra la distribución de preguntas por clase. Se puede observar que para la clase *TITLE* no existe ninguna pregunta en el corpus, por lo que no se tendrá en cuenta en los experimentos, dando lugar a una taxonomía final de 15 clases.

3.3. Vector de características

Cada instancia del problema (pregunta) debe codificarse mediante un vector de características a partir del cual aprenderá el algoritmo de clasificación. Para mantener la independencia de nuestro sistema con respecto a otras herramientas o recursos lingüísticos, vamos a emplear como únicas características de aprendizaje los n-gramas obtenidos del propio corpus de entrenamiento:

- **Unigramas** (1-gramas). Se emplean los términos extraídos de la pregunta como componentes del vector de características.
- **Bigramas** (2-gramas). Representan todas las combinaciones de términos adyacentes en una pregunta como una secuencia de dos palabras.
- **Combinación** (1+2-gramas). Emplearemos combinaciones de unigramas y bigramas, buscando obtener una mejora con respecto a sus componentes individuales.

Vamos a utilizar una representación binaria para el vector de características, donde la aparición de una característica se representa con un 1 y la no aparición con un 0. En la tarea de clasificación de textos se emplea habitualmente la frecuencia de aparición del n-grama o el *tf-idf* para representar a cada término del vector indicando su peso en el documento. Sin embargo, en la tarea de clasificación de preguntas carece de sentido usar este tipo de representación para indicar el peso de los n-gramas en la pregunta, ya que la frecuencia de los términos raramente es superior a 1.

Para obtener los n-gramas de las preguntas hemos utilizado el CMU-Cambridge Statistical Language Modeling Toolkit,[10] un conjunto de herramientas para facilitar la construcción de modelos de lenguaje. El único preproceso llevado a cabo sobre el corpus ha sido la eliminación de signos de puntuación y la sustitución de todos los caracteres en mayúsculas por su equivalente en minúsculas.

4. Evaluación

Para la evaluación de nuestro sistema hemos planteado una serie de experimentos que pretenden cubrir cuatro aspectos fundamentales de nuestro sistema de clasificación:

- Influencia de los parámetros de los kernels.
- Comparación de precisión entre kernels.
- Comparación de los vectores de características.
- Comparación de rendimiento entre idiomas.

Todos los experimentos realizados se han evaluado mediante validación cruzada equilibrada en 10 particiones (*stratified 10-fold cross-validation*). De esta forma no es necesario dedicar una parte en exclusiva del corpus para la evaluación, pudiendo entrenar y evaluar con todo el conjunto de preguntas.

Para la verificación de los resultados hemos empleado *t-test* (Dietterich, 1998). Este test estadístico nos va a permitir saber cuándo las mejoras aportadas por unas configuraciones con respecto a otras son realmente significativas, minimizando la posibilidad de que la diferencia de precisión obtenida pueda ser fortuita. Tal y como indica (Sundblad, 2007), estas técnicas de corroboración de resultados no se han utilizado con asiduidad en los trabajos realizados en el campo de la clasificación de preguntas. Resulta difícil en estas condiciones certificar si las supuestas mejoras aportados por algunas de las aproximaciones en este campo son reales o no. El grado de confianza p que vamos a considerar en estos experimentos es $p < 0,05$ o $p < 0,01$, indicando que la diferencia obtenida entre los sistemas no se debe al azar con una seguridad del 95 % o del 99 % respectivamente.

4.1. Ajuste de parámetros

A la hora de llevar a cabo la clasificación mediante SVM y las funciones kernel, hay una serie de parámetros que gobiernan el proceso de entrenamiento que afectan profundamente al rendimiento final del clasificador. En este apartado

[10] `http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html`.

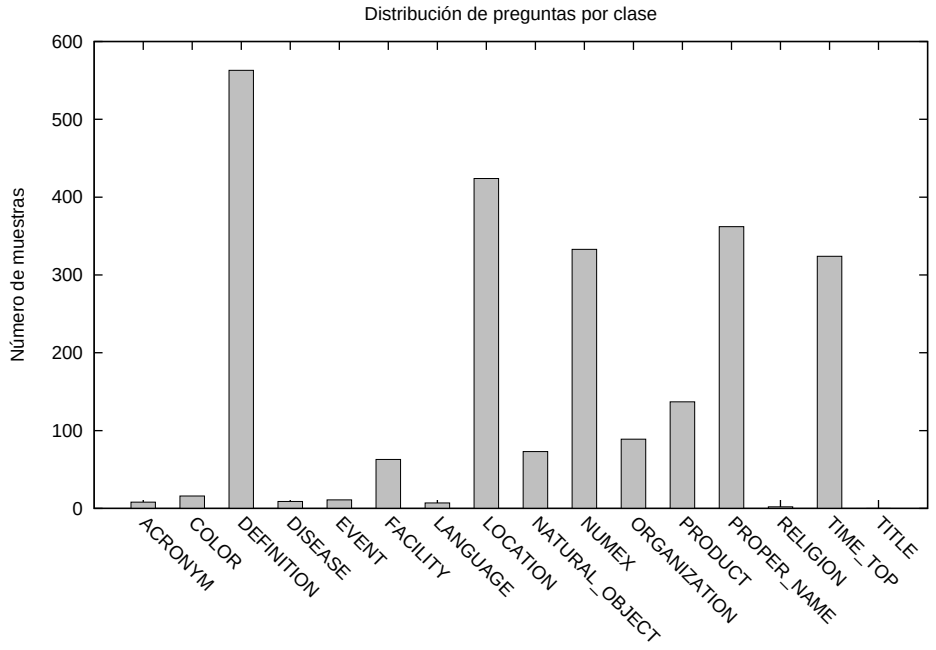

Figura 1: Número de preguntas en el corpus para cada una de las clases de la taxonomía.

queremos evaluar el efecto que tiene la correcta sintonización de estos parámetros en el rendimiento del sistema.

Cuando se emplea SVM es habitual que los conjuntos de datos con los que se trabaja no sean completamente separables mediante un hiperplano. Para estas situaciones, SVM posee un parámetro de coste C que permite la creación de márgenes suaves que permiten algunos errores de clasificación durante el entrenamiento. Un mayor valor de C implica una mayor penalización a los errores de clasificación, forzando la creación de modelos más ajustados a los datos, reduciendo su capacidad de generalización y viéndose afectados por el problema de sobreajuste (*overfitting*).

Además del parámetro C de SVM, cada uno de los kernels definidos en la sección 2, a excepción del lineal, tiene sus propios parámetros. Dichos parámetros son: γ, r y d para el kernel polinómico; γ para el RBF; y γ y r para el sigmoide.

No se conoce a priori cuál es la mejor selección de estos parámetros para un problema dado, por lo que debe llevarse a cabo algún tipo de búsqueda con el objetivo de identificar los mejores valores para la predicción de los datos de evaluación. Para realizar el ajuste de parámetros hemos llevado a cabo una búsqueda *grid* empleando validación cruzada en 5 particiones (*5-fold cross-validation*) sobre el conjunto de entrenamiento formado por los vectores de 1-gramas del corpus en español. Este tipo de validación cruzada permite evitar que haya problemas de sobreajuste al computar los mejores parámetros sobre un conjunto de datos (Hsu, Chang, y Lin, 2003). La tabla 1 muestra los parámetros por defecto empleados de forma habitual en estos kernels (Witten y Frank, 2005; Chang y Lin, 2001) junto con los parámetros optimizados obtenidos en nuestro estudio.

Kernel		C	γ	r	d
Lineal	Def	1,0	-	-	-
	Opt	0,8123	-	-	-
Polinómico	Def	1,0	0,01	1	3
	Opt	73,5167	0,0769	4,4444	3
RBF	Def	1,0	0,01	-	-
	Opt	12,9960	0,0474	-	-
Sigmoide	Def	1,0	0,01	1	-
	Opt	29,8571	0,0237	0,4444	-

Tabla 1: Parámetros por defecto (Def) y optimizados (Opt) para el clasificador SVM y cada uno de los kernels estudiados.

Una vez determinados los mejores parámetros para SVM y para cada uno de los kernels, hemos contrastado el rendimiento de éstos con respecto a la versión con los parámetros por defecto. Hemos llevado a cabo los experimentos en español y catalán, empleando los tres vectores de características comentados en la sección 3.3: unigramas, bigramas y la combinación de ambos. La tabla 2 muestra los resultados obtenidos para los kernels, así como los resultados para dos algoritmos adicionales (Naïve Bayes y LIBLINEAR)

que estudiaremos en el siguiente apartado. A simple vista se aprecia una considerable mejoría de rendimiento con los parámetros optimizados para los kernels polinómico, RBF y sigmoide. En el caso del kernel lineal se observa también una mejora para la mayoría de vectores de aprendizaje en ambos idiomas, pero ésta no es tan evidente y debe valorarse si es estadísticamente significativa.

En la tabla 3 se muestra la comparación estadística mediante *t-test* de los resultados obtenidos con los kernels optimizados y los kernels con parámetros por defecto. Los símbolos "$\gg$" y "$>$" indican que el *Kernel 1* es significativamente mejor que el *Kernel 2* con un grado de confianza $p < 0,01$ y $p < 0,05$ respectivamente. De forma equivalente, "$\ll$" y "$<$" indican que el *Kernel 1* es significativamente peor que el *Kernel 2*. El símbolo "$=$" indica que no hay una diferencia significativa de funcionamiento entre las dos aproximaciones.

A raíz de estos resultados, es evidente que el ajuste de parámetros en el caso de los kernels polinómico, RBF y sigmoide es absolutamente necesario. Las mejoras son evidentes y significativas en ambos idiomas y para todos los vectores de características estudiados. En el caso del kernel lineal se obtienen también mejoras en casi todos los experimentos (a excepción de 2-gramas en catalán), pero estos resultados no son estadísticamente significativos, por lo que no se puede asegurar que en este caso la correcta selección de parámetros del kernel pueda suponer una mejora real en el sistema.

4.2. Comparación entre kernels

En este apartado vamos a contrastar el rendimiento de los distintos kernels entre sí. Para ello vamos a comparar la precisión obtenida por la versión optimizada de cada uno de los kernels en los experimentos realizados en el apartado anterior. Adicionalmente, vamos a introducir dos algoritmos de aprendizaje más en esta comparativa. El primero de ellos, Naïve Bayes (NB) (Duda y Hart, 1973) es un clasificador estocástico usado de forma habitual en numerosas tareas de aprendizaje automático (Mitchell, 1997). Nos servirá como algoritmo de referencia para contrastar el rendimiento de SVM y los diferentes kernels. El segundo algoritmo que vamos a introducir es el clasificador lineal LIBLINEAR (Fan et al., 2008). Este clasificador resulta especialmente adecuado para tareas de aprendizaje en las que intervienen un gran número de instancias y características, como es el caso de nuestro problema (ver la tabla 6 para más detalles sobre el tamaño de los vectores de aprendizaje). Destacar que no existen evaluaciones previas de este algoritmo para la tarea de clasificación de preguntas.

La tabla 2 muestra los resultados obtenidos con NB y LIBLINEAR repitiendo los experimentos realizados en el apartado anterior con los cuatro kernels. En términos generales, se aprecia un buen rendimiento de LIBLINEAR (por encima del 80 % en la mayoría de casos), mientras que los resultados con NB resultan más modestos (por debajo del 70 % en todos los casos).

La tabla 4 muestra la comparación estadística de los resultados obtenidos con los distintos kernels y los dos algoritmos adicionales planteados. La comparación de precisión entre kernels revela que no existe una diferencia significativa entre el kernel lineal, el polinómico y el sigmoide para ninguno de los idiomas y vectores de características tratados. El kernel RBF consigue una precisión equivalente al resto de kernels para los vectores de 1-gramas y 2-gramas. Sin embargo, cuando se observa su funcionamiento con el vector de 1+2-gramas, su precisión es significativamente peor ($p < 0,1$) que los otros tres kernels para ambos idiomas. La pérdida de rendimiento en este caso se achaca a un peor funcionamiento de este kernel en espacios de alta dimensionalidad (Hsu, Chang, y Lin, 2003).

Por lo que respecta a la comparación con NB, queda patente la superioridad de SVM y los distintos kernels con respecto a este algoritmo. Para los cuatro kernels, los resultados obtenidos en los dos idiomas y con todos los vectores de características definidos son significativamente mejores ($p < 0,01$) que NB. En este caso, la tarea de clasificación de preguntas pone de manifiesto el problema del algoritmo NB para trabajar en espacios con un gran número de características de aprendizaje.

Por otra parte, los resultados obtenidos con LIBLINEAR son realmente prometedores. Los valores de precisión obtenidos con este algoritmo fueron significativamente mejores ($p < 0,01$ o $p < 0,05$) que el resto de configuraciones, tanto en español como en catalán, para los tres posibles vectores de características. Esto demuestra como, en situaciones en las que el número de características de aprendizaje es grande, la proyección realizada por los kernels a un espacio de dimensionalidad mayor puede no ser necesaria, obteniendo buenos resultados con un buen clasificador lineal. Estos resultados reafirman los ya obtenidos al evaluar los kernels, donde el kernel lineal obtenía resultados comparables al resto de kernels planteados.

Kernel		Español			Catalán		
		1-gramas	2-gramas	1+2-gramas	1-gramas	2-gramas	1+2-gramas
Lineal	Def	80,92	75,25	81,25	80,66	75,97	80,84
	Opt	81,07	75,27	81,25	80,74	75,91	80,87
Polinómico	Def	65,65	53,55	72,40	66,66	57,71	72,26
	Opt	80,86	74,76	80,69	80,91	75,56	80,53
RBF	Def	62,83	45,58	67,48	65,39	53,21	67,83
	Opt	81,13	74,53	79,01	80,88	75,62	78,87
Sigmoide	Def	43,20	23,53	48,25	53,02	24,06	57,20
	Opt	81,10	74,95	80,36	81,10	76,02	80,28
NB		69,82	58,13	68,89	69,42	61,35	68,68
LIBLINEAR		82,27	77,61	83,71	82,46	78,72	82,64

Tabla 2: Precisión obtenida por cada uno de los kernels con los parámetros por defecto (Def) y los parámetros optimizados (Opt). Las dos últimas filas muestran la precisión obtenida por los algoritmos Naïve Bayes (NB) y LIBLINEAR.

Kernel 1		Kernel 2		Español			Catalán		
				1-gramas	2-gramas	1+2-gramas	1-gramas	2-gramas	1+2-gramas
Lineal	Opt	Lineal	Def	=	=	=	=	=	=
Poli.	Opt	Poli.	Def	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
RBF	Opt	RBF	Def	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
Sigmo.	Opt	Sigmo.	Def	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$

Tabla 3: Comparación estadística entre la versión con parámetros optimizados (Opt) y con parámetros por defecto (Def) de cada uno de los kernels. El símbolo "=" indica que la diferencia no es estadísticamente significativa, mientras que "$\gg$" indica que el *Kernel 1* es significativamente mejor ($p < 0,01$) que el *Kernel 2*.

4.3. Comparación entre vectores de características

En este apartado vamos a comparar los distintos vectores de características definidos: 1-gramas, 2-gramas y 1+2-gramas. Para esta comparativa vamos a retomar los valores de precisión obtenidos en los dos apartados anteriores. La tabla 5 toma como base la precisión obtenida con el vector de 1-gramas y compara estos resultados con los obtenidos con los otros dos vectores para cada uno de los algoritmos e idiomas.

Los resultados revelan que para todos los algoritmos el vector de 1-gramas es significativamente mejor ($p < 0,01$) que el vector de 2-gramas, no existiendo diferencia con el vector de 1+2-gramas. Estos resultados reflejan que los bigramas por sí solos no resultan adecuados para la clasificación. Al ser combinados con los unigramas (1+2-gramas) se obtienen ligeras mejoras de precisión en muchos de los algoritmos tratados, pero estas mejoras no resultan ser significativas. En estos casos, el enriquecimiento de información que supone la incorporación de bigramas se ve lastrado por el aumento de atributos en el espacio de aprendizaje (ver tabla 6). La única excepción es el caso del algoritmo LIBLINEAR en español. En esta ocasión, la precisión obtenida con el vector de 1+2-gramas es significativamente mejor ($p < 0,01$) que para el resto de vectores. Se demuestra nuevamente el buen funcionamiento de este algoritmo con vectores de aprendizaje de gran tamaño.

4.4. Comparación entre idiomas

A continuación vamos a comparar el rendimiento del sistema en función del idioma de trabajo. Ya que no empleamos ningún tipo de herramienta o recurso lingüístico, a excepción de los n-gramas extraídos del corpus, este experimento nos va a permitir valorar si las características intrínsecas de cada idioma, como la flexión verbal y nominal, afectan a la tarea de clasificación de preguntas. El grado de flexión de cada lengua se va a ver reflejado en el tamaño de vocabulario del problema, tal y como se muestra en la tabla 6. Además de en español y en catalán, que poseen un grado similar de flexión, vamos a realizar el estudio comparativo con el corpus original de preguntas en inglés. Nos vamos a centrar en el kernel lineal para comprobar el rendimiento de éste en los tres idiomas propuestos.

		Español			Catalán		
Algoritmo 1	Algoritmo 2	1-gramas	2-gramas	1+2-gramas	1-gramas	2-gramas	1+2-gramas
Lineal	Polinómico	=	=	=	=	=	=
Lineal	RBF	=	>	≫	=	=	≫
Lineal	Sigmoide	=	=	=	=	=	=
Lineal	NB	≫	≫	≫	≫	≫	≫
Lineal	LIBLINEAR	<	≪	≪	≪	≪	≪
Polinómico	RBF	=	=	≫	=	=	≫
Polinómico	Sigmoide	=	=	=	=	<	=
Polinómico	NB	≫	≫	≫	≫	≫	≫
Polinómico	LIBLINEAR	≪	≪	≪	≪	≪	≪
RBF	Sigmoide	=	=	≪	=	=	≪
RBF	NB	≫	≫	≫	≫	≫	≫
RBF	LIBLINEAR	<	≪	≪	≪	≪	≪
Sigmoide	NB	≫	≫	≫	≫	≫	≫
Sigmoide	LIBLINEAR	≪	≪	≪	≪	≪	≪
NB	LIBLINEAR	≪	≪	≪	≪	≪	≪

Tabla 4: Comparación estadística de los resultados obtenidos con cada uno de los kernels y algoritmos definidos.

	Español		Catalán	
Algoritmo	2-gramas	1+2-gramas	2-gramas	1+2-gramas
Lineal	≫	=	≫	=
Polinómico	≫	=	≫	=
RBF	≫	≫	≫	≫
Sigmoide	≫	=	≫	=
NB	≫	=	≫	=
LIBLINEAR	≫	≪	≫	=

Tabla 5: Comparación estadística de los vectores de aprendizaje tomando como base la precisión obtenida con el vector de 1-gramas.

	Inglés	Español	Catalán
1-gramas	3.764	4.164	4.190
2-gramas	8.465	8.578	8.625
1+2-gramas	12.229	12.742	12.815

Tabla 6: Tamaño del vector de aprendizaje para cada uno de los vectores de características.

	Inglés	Español	Catalán
1-gramas	81,77	80,92	80,66
2-gramas	76,84	75,25	75,97
1+2-gramas	81,64	81,25	80,84

Tabla 7: Resultados de la comparación entre idiomas con el kernel lineal. No existen diferencias significativas de precisión entre los tres idiomas para ninguno de los vectores de aprendizaje.

La tabla 7 muestra los resultados obtenidos. La precisión ofrecida por el clasificador en inglés es ligeramente superior al de los otros dos idiomas para los tres vectores de aprendizaje propuestos. En el caso de español y catalán, el primero obtiene resultados ligeramente favorables para 1-gramas y 1+2-gramas con respecto al segundo. Sin embargo, ninguno de estas diferencias resulta ser significativa ($p < 0{,}01$ y $p < 0{,}05$).

Podemos concluir que las características propias de cada uno de estos tres idiomas no mejoran o empeoran el rendimiento final del sistema. El aumento de vocabulario que existe en español y catalán con respecto al inglés tampoco parece afectar al proceso de clasificación.

5. Trabajo relacionado

Pese a los numerosos trabajos realizados dentro del área de la clasificación de preguntas, son escasas las aproximaciones dedicadas a idiomas diferentes del inglés o su aplicación a sistemas multilingües.

Dentro de los sistemas en inglés, destacan los trabajos realizados por Xin Li y Dan Roth (2002; 2005), que sirvieron para establecer la tarea de clasificación de preguntas como una tarea inde-

pendiente de los sistemas QA y evaluable por sí misma. En estos trabajos desarrollaron un clasificador jerárquico de preguntas en inglés, basado en la arquitectura de aprendizaje SNoW (Sparse Network of Winnows) (Carlson et al., 1999).

En lo que respecta al uso de kernels para esta tarea, hay numerosos sistemas que han usado SVM en su forma más básica mediante el empleo de un kernel lineal, demostrando sistemáticamente el buen funcionamiento de este algoritmo con respecto a otros en la tarea de clasificación de preguntas. Entre los sistemas que emplean esta aproximación podemos destacar los de Hacioglu y Ward (2003), Krishnan et al. (2005), Skowron y Araki (2004), Nguyen et al. (2008), Day et al. (2007), Solorio et al. (2004), Bisbal et al. (2005) y Tomás et al. (2005).

Otro kernel muy extendido en clasificación de preguntas es el *tree kernel* (Collins y Duffy, 2001). Este kernel permite medir la similitud de dos árboles contando el número de ramas comunes. Zhang y Lee (2003) emplean este kernel para incorporar la información del árbol de análisis sintáctico de las preguntas al proceso de clasificación. Otra propuesta de este tipo es la desarrollada por Moschitti et al. (2007), donde definen una nueva estructura en forma de árbol basada en información sintáctica y semántica superficial codificada mediante estructuras predicativas (*Predicate-Argument Structures*). Este kernel permite explotar el poder de representación de dichas estructuras mediante un clasificador SVM. Pan et al. (2008) definen un *tree kernel* semántico que aprovecha distintas fuentes de información (relaciones de WordNet, listas manuales de palabras relacionadas y entidades) para incorporar información sobre la similitud semántica entre preguntas.

En el trabajo de Suzuki et al. (2003) se define un kernel denominado *Hierarchical Directed Acyclic Graph* (HDAG). Este kernel está diseñado para manejar con facilidad estructuras lingüísticas en el texto, como los sintagmas y sus relaciones, empleándolas como características de aprendizaje sin necesidad de convertir dichas estructuras a un vector de características de forma explícita.

Todos estos kernels han sido aplicados únicamente para inglés. En este caso, a la falta de corpus en otros idiomas se une la dependencia del sistema con respecto a los recursos utilizados, debido a que estos kernels sintáctico-semánticos requieren de herramientas de análisis lingüístico y bases de conocimiento que no existen o son difíciles de conseguir para otros idiomas.

Existen algunas aproximaciones que se han adentrado en idiomas diferentes al inglés, como por ejemplo el finés (Aunimo y Kuuskoski, 2005), el estonio (Hennoste et al., 2005), el francés (Feiguina y Kégl, 2005), el chino (Day et al., 2005; Lin, Peng, y Liu, 2006), el japonés (Suzuki et al., 2003), el portugués (Solorio et al., 2005) y el español (Ángel García Cumbreras et al., 2005; Tomás et al., 2005). Ninguno de estos sistemas ha afrontado la tarea en catalán, y son escasos aquellos que afrontan la tarea en más de un idioma (Solorio et al., 2005; Bisbal et al., 2005).

6. Conclusiones y trabajo futuro

En este trabajo hemos presentado una aproximación a la clasificación de preguntas basada en kernels para español y catalán, tratando de cubrir el vacío existente en el campo de la clasificación de preguntas para idiomas distintos del inglés. La ausencia de conjuntos de preguntas en estos idiomas nos ha llevado a desarrollar nuestros propios corpus de preguntas. Estos corpus han sido puestos libremente a disposición de la comunidad científica.

El sistema propuesto se basa en la utilización de SVM con diferentes kernels: lineal, polinómico, RBF y sigmoide. Para obtener el mejor rendimiento posible de estos kernels se ha realizado un ajuste de los parámetros para su funcionamiento. El ajuste correcto de los parámetros ha demostrado una mejora significativa para los kernels polinómico, RBF y sigmoide en ambos idiomas. En el caso del kernel lineal se ha conseguido mejorar los resultados, pero esta mejora no ha resultado ser estadísticamente significativa.

Una vez determinados los mejores parámetros para SVM y los kernels, hemos comparado el funcionamiento de éstos entre sí. Los resultados revelaron un funcionamiento similar de los cuatro kernels para los vectores de 1-gramas y 2-gramas, obteniendo resultados por encima del 80 % en ambos idiomas. La única diferencia significativa se da en el caso del kernel RBF y el vector 1+2-gramas, resultando significativamente peor que para el resto de kernels. Este resultado desaconseja el uso de RBF con vectores de aprendizaje de gran tamaño.

Además de la comparativa entre kernels, hemos analizado su rendimiento con respecto a otros dos algoritmos, Naïve Bayes y LIBLINEAR, este último especialmente adecuado para trabajos en espacios de alta dimensionalidad. Los resultados obtenidos con Naïve Bayes fueron estadísticamente inferiores a los obtenidos con los cuatro kernels empleados, demostrando el buen funcionamiento de éstos en espacios de alta di-

mensionalidad. Por otra parte, resulta destacable el rendimiento obtenido por LIBLINEAR, que superó a los cuatro kernels definidos en ambos idiomas. Este algoritmo no había sido aplicado con anterioridad a la tarea de clasificación de preguntas.

Por otra parte, hemos experimentado con diversos vectores de características. Los resultados revelan que al trabajar con kernels, el uso de 1-gramas mejora significativamente al de 2-gramas, y en ningún caso obtiene resultados significativamente inferiores a los obtenidos con la combinación de ambos (1+2-gramas). Sólo en el caso del algoritmo LIBLINEAR se obtiene una mejora significativa al emplear un vector de gran tamaño formado por 1+2-gramas.

Por último hemos realizado una comparación de rendimiento del clasificador entre idiomas, añadiendo en este caso el corpus en inglés a nuestro conjunto de experimentos. Los resultados obtenidos revelan que un kernel lineal aplicado sobre un corpus paralelo en inglés, español y catalán, obtiene resultados muy similares para los tres idiomas. En los experimentos realizados existe una aparente correlación entre el tamaño del vocabulario y el rendimiento del sistema, siendo éste ligeramente mejor en inglés que en español y en catalán. Sin embargo, ninguno de los resultados obtenidos para inglés es significativamente mejor que para el resto de idiomas. La aproximación basada en kernels y características textuales superficiales resulta, por tanto, adecuada para su aplicación en entornos multilingües.

Con el fin de obtener sistemas de alto rendimiento para español y catalán, se plantea como trabajo futuro el enriquecimiento del vector de características empleando información lingüística más profunda (usando herramientas como FreeLing (Atserias et al., 2006), que ofrece numerosas aplicaciones lingüísticas en los dos idiomas citados). Los resultados obtenidos con LIBLINEAR hacen de este algoritmo una apuesta destacable de cara al trabajo futuro con estos vectores enriquecidos.

Bibliografía

Ángel García Cumbreras, Miguel, Fernando Martínez Santiago, Luis Alfonso Ureña López, y Arturo Montejo Ráez. 2005. Búsqueda de respuestas multilingüe : clasificación de preguntas en español basada en aprendizaje. *Procesamiento del Lenguaje Natural*, (34):31–40, March. `http://www.sepln.org/revistaSEPLN/revista/34/03.pdf`.

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluis Padró, y Muntsa Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, páginas 48–55. `http://www.lsi.upc.edu/~nlp/papers/atserias06.pdf`.

Aunimo, Lili y Reeta Kuuskoski. 2005. Reformulations of finnish questions for question answering. En *Proceedings of the 15th NODALIDA conference*, páginas 12–21. `http://phon.joensuu.fi/lingjoy/01/aunimoF.pdf`.

Bisbal, Empar, David Tomás, Lidia Moreno, José L. Vicedo, y Armando Suárez. 2005. A multilingual svm-based question classification system. En Alexander F. Gelbukh Alvaro de Albornoz, y Hugo Terashima-Marín, editores, *MICAI 2005: Advances in Artificial Intelligence, 4th Mexican International Conference on Artificial Intelligence*, volumen 3789 de *Lecture Notes in Computer Science*, páginas 806–815. Springer, November. `http://www.springerlink.com/index/75jr3067j3472680.pdf`.

Carlson, Andrew, Chad Cumby, Jeff Rosen, y Dan Roth. 1999. The snow learning architecture. Informe Técnico UIUCDCS-R-99-2101, UIUC Computer Science Department, May.

Chang, Chih Chung y Chih Jen Lin, 2001. *LIBSVM: a library for support vector machines.* `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Collins, Michael y Nigel Duffy. 2001. Convolution kernels for natural language. En *Advances in Neural Information Processing Systems (NIPS14)*, páginas 625–632. MIT Press. `http://l2r.cs.uiuc.edu/~danr/Teaching/CS546-09/Papers/Collins-kernels.pdf`.

Day, Min-Yuh, Cheng-Wei Lee, Shih-Hung Wu, Chorng-Shyong Ong, y Wen-Lian Hsu. 2005. An integrated knowledge-based and machine learning approach for chinese question classification. *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering. IEEE NLP-KE '05*, páginas 620–625, October.

Day, Min-Yuh, Chorng-Shyong Ong, y Wen-Lian Hsu. 2007. Question classification in english-chinese cross-language question answering: An integrated genetic algorithm and machine learning approach. *IEEE International Conference on Information Reuse and*

Integration, 2007. IRI 2007, páginas 203–208, August.

Dietterich, Thomas G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923. http://web.engr.oregonstate.edu/~tgd/publications/nc-stats.ps.gz.

Duda, Richard O. y Peter E. Hart, 1973. *Pattern Classification and Scene Analysis*, páginas 98–105. John Wiley and Sons.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, y Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874. http://www.csie.ntu.edu.tw/~cjlin/papers/liblinear.pdf.

Feiguina, Olga y Balázs Kégl. 2005. Learning to classify questions. En *CLiNE 05: 3rd Computational Linguistics in the North-East Workshop*, August. http://www.crtl.ca/cline05/cline05_papers/FeiguinaKegl.pdf.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Hacioglu, Kadri y Wayne Ward. 2003. Question classification with support vector machines and error correcting codes. En *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, páginas 28–30, Morristown, NJ, USA. Association for Computational Linguistics. http://www.aclweb.org/anthology/N/N03/N03-2010.pdf.

Hennoste, Tiit, Olga Gerassimenko, Riina Kasterpalu, Mare Koit, Andriela Rääbis, Krista Strandson, y Maret Valdisoo. 2005. Questions in estonian information dialogues: Form and functions. En *Text, Speech and Dialogues*, volumen 3658, páginas 420–427. Springer Berlin / Heidelberg. http://www.springerlink.com/index/6C2298L0XC04T08B.pdf.

Hsu, Chih Wei, Chih Chung Chang, y Chih Jen Lin. 2003. A practical guide to support vector classification. Informe técnico, Taipei. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Joachims, Thorsten. 1998. Text categorization with support vector machines: learning with many relevant features. En Claire Nédellec y Céline Rouveirol, editores, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, numero 1398, páginas 137–142, Chemnitz, DE. Springer Verlag, Heidelberg, DE. http://www.joachims.org/publications/joachims_98a.ps.gz.

Kando, Noriko. 2005. Overview of the fifth ntcir workshop. En *Proceedings of NTCIR-5 Workshop*, Tokyo, Japan. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/NTCIR5-OV-KandoN.pdf.

Krishnan, Vijay, Sujatha Das, y Soumen Chakrabarti. 2005. Enhanced answer type inference from questions using sequential models. En *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, páginas 315–322, Morristown, NJ, USA. Association for Computational Linguistics. http://www.cse.iitb.ac.in/~soumen/doc/emnlp2005/382.pdf.

Kudo, Taku y Yuji Matsumoto. 2001. Chunking with support vector machines. En *NAACL*. http://www.aclweb.org/anthology/N/N01/N01-1025.pdf.

Li, Xin y Dan Roth. 2002. Learning question classifiers. En *Proceedings of the 19th international conference on Computational linguistics*, páginas 1–7, Morristown, NJ, USA. Association for Computational Linguistics. http://www.aclweb.org/anthology/C/C02/C02-1150.pdf.

Li, Xin y Dan Roth. 2005. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249. http://l2r.cs.uiuc.edu/~danr/Papers/LiRo05a.pdf.

Lin, Xu-Dong, Hong Peng, y Bo Liu. 2006. Support vector machines for text categorization in chinese question classification. En *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, páginas 334–337.

Magnini, Bernardo, Simone Romagnoli, Alessandro Vallin, Jesús Herrera, Anselmo Peñas, Víctor Peinado, Felisa Verdejo, y Maarten de Rijke. 2003. Creating the disequa corpus: A test set for multilingual question answering. En *Cross-Lingual Evaluation Forum (CLEF) 2003 Workshop*, páginas 311–320. http://www.springerlink.com/index/6l35q8cl7e864nmn.pdf.

Magnini, Bernardo, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas,

Maarten de Rijke, Paulo Rocha, Kiril Ivanov Simov, y Richard F. E. Sutcliffe. 2005. Overview of the clef 2004 multilingual question answering track. En Carol Peters Paul Clough Julio Gonzalo Gareth J. F. Jones Michael Kluck, y Bernardo Magnini, editores, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, volumen 3491 de *Lecture Notes in Computer Science*, páginas 371–391. Springer. http://www.springerlink.com/index/ebtpv2e71eg4txbu.pdf.

Mitchell, Tom M. 1997. *Machine Learning*. McGraw-Hill Science/Engineering/Math, March.

Moldovan, Dan, Marius Paşca, Sanda Harabagiu, y Mihai Surdeanu. 2003. Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems*, 21(2):133–154.

Moschitti, Alessandro, Silvia Quarteroni, Roberto Basili, y Suresh Manandhar. 2007. Exploiting syntactic and shallow semantic kernels for question answer classification. En *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, páginas 776–783, Prague, Czech Republic, June. Association for Computational Linguistics. http://www.ist-luna.eu/pdf/ACL07.pdf.

Nguyen, Tri Thanh, Le Minh Nguyen, y Akira Shimazu. 2008. Using semi-supervised learning for question classification. *Information and Media Technologies*, 3(1):112–130. http://www.springerlink.com/index/y85h00v3825r4081.pdf.

Pan, Yan, Yong Tang, Luxin Lin, y Yemin Luo. 2008. Question classification with semantic tree kernel. En *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 837–838, New York, NY, USA. ACM.

Radev, Dragomir, Weiguo Fan, Hong Qi, Harris Wu, y Amardeep Grewal. 2002. Probabilistic question answering on the web. En *WWW '02: Proceedings of the 11th international conference on World Wide Web*, páginas 408–419, New York, NY, USA. ACM. http://filebox.vt.edu/users/wfan/paper/www/www.pdf.

Schölkopf, Bernhard y Alexander J. Smola. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Sekine, Satoshi, Kiyoshi Sudo, y Chikashi Nobata. 2002. Extended named entity hierarchy. En *LREC 2002: Language Resources and Evaluation Conference*, páginas 1818–1824, Las Palmas, Spain. http://nlp.cs.nyu.edu/pubs/papers/sekine-lrec02.pdf.

Shawe-Taylor, John y Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, June.

Skowron, Marcin y Kenji Araki. 2004. Evaluation of the new feature types for question classification with support vector machines. *IEEE International Symposium on Communications and Information Technology, 2004. ISCIT 2004*, 2:1017–1022, October.

Solorio, Thamar, no Manuel Pérez-Couti Manuel Montes y Gémez, nor-Pineda Luis Villase y Aurelio López-López. 2004. A language independent method for question classification. En *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, páginas 1374–1380, Morristown, NJ, USA. Association for Computational Linguistics. http://www.aclweb.org/anthology/C/C04/C04-1201.pdf.

Solorio, Thamar, Manuel Pérez-Couti no, Manuel Montes y Gómez, Luis Villase nor Pineda, y Aurelio López-López. 2005. Question classification in spanish and portuguese. En *CICLing*, páginas 612–619. http://www.springerlink.com/index/46d3nw2qpe7tpx3f.pdf.

Sundblad, Håkan. 2007. Question classification in question answering. Master's thesis, Linköping University, Department of Computer and Information Science. http://liu.diva-portal.org/smash/get/diva2:23705/FULLTEXT01.

Suzuki, Jun, Hirotoshi Taira, Yutaka Sasaki, y Eisaku Maeda. 2003. Question classification using hdag kernel. En *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, páginas 61–68, Morristown, NJ, USA. Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W03/W03-1208.pdf.

Tomás, David, José L. Vicedo, Armando Suárez, Empar Bisbal, y Lidia Moreno. 2005. Una aproximación multilingüe a la clasificación de preguntas basada en aprendizaje automático. *Procesamiento del Lenguaje Natural*, (35):391–398. http://www.sepln.org/revistaSEPLN/revista/35/48.pdf.

Vallin, Alessandro, Bernardo Magnini, Danilo Giampiccolo, Lili Aunimo, Christelle Ayache, Petya Osenova, Anselmo Peñas, Maarten de Rijke, Bogdan Sacaleanu, Diana Santos, y Richard Sutcliffe. 2006. Overview of the clef 2005 multilingual question answering track. En Springer Berlin / Heidelberg, editor, *Accessing Multilingual Information Repositories*, volumen 4022 de *Lecture Notes in Computer Science*, páginas 307–331. `http://www.springerlink.com/index/dm61h684k55150p2.pdf`.

Voorhees, Ellen M. 1999. The trec-8 question answering track report. En *Eighth Text REtrieval Conference*, volumen 500-246 de *NIST Special Publication*, páginas 77–82, Gaithersburg, USA, November. National Institute of Standards and Technology. `http://trec.nist.gov/pubs/trec8/papers/qa_report.ps`.

Voorhees, Ellen M. 2001. The trec question answering track. *Natural Language Engineering*, 7(4):361–378.

Witten, Ian H. y Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edición.

Zhang, Dell y Wee Sun Lee. 2003. Question classification using support vector machines. En *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, páginas 26–32, New York, NY, USA. ACM. `http://www.comp.nus.edu.sg/~leews/publications/p31189-zhang.pdf`.

Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish

Hristo Tanev
JRC, Ispra, Italy
hristo.tanev@ext.jrc.ec.europa.eu

Vanni Zavarella
JRC, Ispra, Italy
vanni.zavarella@ext.jrc.ec.europa.eu

Jens Linge
JRC, Ispra, Italy
jens.linge@jrc.ec.europa.eu

Mijail Kabadjov
JRC, Ispra, Italy
mijail.kabadjov@jrc.ec.europa.eu

Jakub Piskorski
Polish Academy of Sciences
Jakub.Piskorski@ipipan.waw.pl

Martin Atkinson
JRC, Ispra, Italy
martin.atkinson@jrc.ec.europa.eu

Ralf Steinberger
JRC, Ispra, Italy
ralf.steinberger@jrc.ec.europa.eu

November 22, 2009

Abstract

We describe a multilingual methodology for adapting an event extraction system to new languages. The methodology is based on highly multilingual domain-specific grammars and exploits weakly supervised machine learning algorithms for lexical acquisition. We adapted an already existing event extraction system for the domain of conflicts and crises to Portuguese and Spanish languages. The results are encouraging and demonstrate the effectiveness of our approach.

1 Introduction

We present a multilingual methodology for building event extraction systems and describe its application for the Portuguese and Spanish languages. Formally, the task of event extraction is to automatically identify events in free text and to derive detailed information about them, ideally identifying *Who did what to whom, when, with what methods (instruments), where and why.* Automatically extracting events is a higher-level information extraction (IE) task (Appelt, 1999) which is not trivial due to the complexity of natural language and due to the fact that, in news, a full event description is usually scattered over several sentences and articles. In particular, event extraction relies on identifying named entities and relations between them. The research on automatic event extraction was pushed forward by the DARPA-initiated Message Understanding Conferences[1] and by the ACE (Automatic Content Extraction)[2] programme. Although, a considerable amount of work on automatic extraction of events has been reported, it still appears to be a lesser studied area in comparison to the somewhat easier tasks of named-entity and relation extraction.

First attempts to larger-scale event extraction systems were reported a decade ago, e.g., in (Aone and Santacruz, 2000). Some examples of the current functionality and capabilities of event extraction technology dealing with identification of disease outbreaks, conflict incidents and other crisis-related events are given in (Grishman and Yangarber, 2002),(Grishman and Yangarber, 2003), (King and Lowe, 2003), (Naughton and Carthy, 2006), (Ji and Grishman, 2008), (Yangarber, Rauramo, and Huttunen, 2005) and (Wagner and Baker, 2006).

We have created a multilingual event extraction system NEXUS, which is part of the Europe Media Monitor family of applications (EMM) (Steinberger, Pouliquen, and van der Goot, 2009). EMM performs automatic real-time gathering and analysis of online news in 45 languages. NEXUS aims at identifying vio-

[1] http://en.wikipedia.org/wiki/Message_Understanding_Conference

[2] ACE - http://projects.ldc.upenn.edu/ace

lent events, man made and natural disasters and humanitarian crises, in news reports. The information about such events is extremely important for better crisis management and for developing warning systems which detect precursors for threats in the fields of disaster and conflict.

Crucial information for all these events are the number and the description of the victims. Additionally, analysis of humanitarian crises requires identification of the number of the displaced and homeless people; analysis of the violent events requires identification of the weapons and the perpetrators.

Currently, NEXUS can handle 4 languages - English, French, Italian, and Russian. Within the EMM project, we aim at global monitoring of crisis and conflict events: at the same time, we also try to detect events with only national or local relevance. In this view, we decided to adapt Nexus to Portuguese and Spanish language so as to extend the coverage of our system to Latin American and African areas.

The architecture and the algorithms implemented in NEXUS are highly language-independent. The system involves the use of language-specific dictionaries and extraction grammars, which are plugged in as external resources; therefore, adding a new language to the system is possible without modifying the system itself. Moreover, the domain-specific grammars, which we use to extract event-specific entities, contain very few references to concrete words. Therefore, a grammar for one language can be reused without significant changes for another language, especially if they belong to the same language family. In our development cycle, we adapted an Italian grammar to other members of the Romance language family, namely French, Spanish and Portuguese.

In order to adapt our event extraction system to a new language, we adopted a multilingual methodology which is based on two semi-supervised machine learning algorithms and highly language-independent domain-specific grammars. Using this methodology, we were able to build event extraction systems for the Portuguese and Spanish languages with promising performances, which proved the viability of our approach.

In section 2 we outline the architecture of NEXUS and its integration within the European Media Monitor system; section 3 outlines the extraction grammar as it was adapted for the Portuguese and Spanish; then section 4 describes the machine learning algorithms we exploit and finally we present experiments and evaluation.

$$\begin{bmatrix} Date \\ Location \\ EventType \\ Dead \begin{bmatrix} Number \\ Description \end{bmatrix} \\ Injured \begin{bmatrix} Number \\ Description \end{bmatrix} \\ Kidnapped \begin{bmatrix} Number \\ Description \end{bmatrix} \\ Displaced \begin{bmatrix} Number \\ Description \end{bmatrix} \\ Arrested [Description] \\ Perpetrators [Description] \\ Weapons \end{bmatrix}$$

Figure 1: The output structure of the event extraction system

2 EMM and NEXUS

Europe Media Monitor (EMM) is an ongoing project, whose main outcome is a multilingual news gathering and analysis system which works for 41 languages, (Steinberger, Pouliquen, and van der Goot, 2009).

The NEXUS event extraction system takes on its input the information provided by other EMM modules, integrates it after performing validation and merging, in order to extract event report summaries. Before the proper event extraction process can proceed, news articles are gathered by dedicated software for media monitoring, that receives 90000 news articles from 2200 news sources in 41 languages each day. Next, the articles are grouped into news clusters according to content similarity. Subsequently, each cluster is geo-located.

For each such a cluster NEXUS tries to detect and extract only the main event by analyzing the title and first sentence of all of the articles in the cluster. For each detected violent and disaster event NEXUS produces a frame, whose main slots are shown in Figure 1.

In Figure 2, a sketch of the entire event extraction processing chain is shown. First, the full news article are scanned by EMM modules in order to identify entities and locations which are inserted as meta-data. These entities are typically separate from the ones deployed in the event extraction process proper. Next the articles are clustered and then geo-located according to extracted meta-data. Each article in the cluster is then linguistically preprocessed in order to produce a more abstract representation of its text. This encompasses the following steps: fine-grained tokenization, sentence splitting, domain-specific dictionary look-up (i.e. matching of key

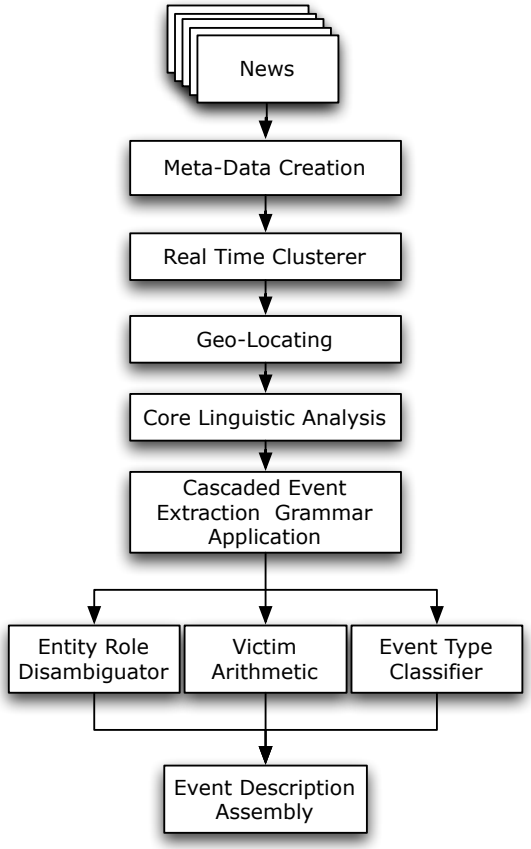

Figure 2: Event Extraction processing chain

terms indicating numbers, quantifiers, person titles, unnamed person groups like *civilians*, *policemen* and *Shiite*), and finally morphological analysis, simply consisting of lexicon look-up on large domain-independent morphological dictionaries. The aforementioned tasks are accomplished by CORLEONE (Core Linguistic Entity Online Extraction), our in-house core linguistic engine (Piskorski, 2008). Once the linguistic preprocessing is complete, a cascade of extraction grammars is applied on each article in order to identify phrases reporting about victims and entities and their participation in the event. For example, phrases like *"matou seis civis"* are parsed by the grammar cascade, extracting the *seis civis* as victims.

The news clusters contain reports from different news sources about the same fact. This redundancy mitigates the impact on system performance of linguistic phenomena which are hard to tackle, such as anaphora, ellipsis and long-distance dependency. Consequently, the system can process the first sentence and the title of each article, where the main facts are summarized in simple syntax (Bell, 1991), without significant loss in coverage.

On the other hand, contradictory information on the same story may occur at the cluster level; consequently, the last processing steps consist of cross-article information fusion in order to produce event descriptions. Namely, Nexus aggregates and validates information extracted locally from each single article in the same cluster. This process encompasses mainly three tasks, entity role disambiguation (as a result of extraction pattern application the same entity might be assigned different roles), victim counting and event type classification. An example of the system output as geolocated in a Google Map interface is shown in Figure 3.

3 Outline of the Extraction Grammars

The role of the grammars deployed in NEXUS is the recognition of phrases which introduce events participants. For example, in the text

Soldados israelenses matam palestino de 14 anos

the grammar should extract the phrase *Soldados israelenses* and assign to it the semantic role *perpetrator*, while the phrase *palestino de 14 anos* should be extracted as *dead victim* description. The extraction process is performed by devising a multi-layer grammar cascade in the EXPRESS formalism (Piskorski, 2007). EXPRESS is a finite state-based grammar formalism and pattern matching engine developed in-house which proved quite fast and efficient in real-time text processing.

3.1 Extraction Pattern Specification Language

An EXPRESS grammar consists of a cascade of pattern-action rules. The left-hand side (LHS) of a rule (the recognition part) is a regular expression over flat feature structures (FFS), i.e., non-recursive typed feature structures (TFS) without structure sharing, where features are string-valued and types are not hierarchically ordered (differening in this from traditional unification-based grammar formalisms). The right-hand side (RHS) of a rule (action part) consists of a list of FFS, which is returned in case the LHS pattern is matched. Variables can be associated to the string-valued attributes on the LHS of a rule in order to allow information transport into the RHS. Further, functional operators are allowed in the RHSs in order to form output slot values by string processing operations or specify constraints in the form of boolean-valued predicates. Rules can be associated with multiple ac-

Figure 3: A sample output of the event extraction system as shown in Google Map interface.

tions, i.e., producing multiple annotations (possibly nested) for a given text fragment. Finally, arbitrary processing resources can be integrated at any level of the grammar cascade. In our case, CORLEONE text processing modules are deployed. For more details on the EXPRESS formalism and its processing performance refer to (Piskorski, 2007).

3.2 Person and entity recognition

The lower levels of the grammar cascade contain patterns for recognition of named entities (e.g., person names), numbers, quantifiers, simple chunks representing unnamed person groups (e.g., *cinco policiais, milhares de portugueses, casi la mitad de los soldados extranjeros*); moreover, appositive and coordinated phrase composition are covered. On Figure 4 is presented a simplified example of a rule for detection of an unnamed person entity noun phrase such as *um jovem militante*. The rule matches a sequence consisting of an optional article, followed by a nationality noun, preceded and/or followed in its turn by an optional modifying adjective - so as to deal with relatively free position of modifiers in Romance language noun phrases. It produces a singular *person_group* type structure.

Expressions like *adjective* or *gazetteer* at the front of FFS's make reference to one of the output types of CORLEONE modules which are available at the level of the grammar cascade where the rule appears - in this case, morphological and domain-specific lexicon look-up. The symbol "&" links the name of the FFS's type with a list of constraints (in the form of attribute-value pairs) to be fulfilled, such as on grammatical number, morphological subtype, gender and so on.

Notice that the NAME value of the output structure is created by accessing the variables*#name0*, *#name1* and *#name2* on the LHS and concatenating them by calling a functional operator, while GENDER attribute value is read directly from grammatical gender of article and/or modifiers , if there are any; Gender agreement is then enforced by the string equality predicate *IsEqual()*.

Notice how such a simple rule would run almost identically for Portuguese, Spanish and other Romance languages, provided that suitable lexicons for domain specific categories such as *person* above, or for general grammatical categories (like *adjective*) were plugged into the system.

```
person_entity_sg :>
         ( determiner & [TYPE:"article", NUMBER:"sg", GENDER:#g0]?
           adjective & [NUMBER: "sg",GENDER:#g1, SURFACE: #name0]?
           gazetteer & [GTYPE: "person",NUMBER: "sg", GENDER:#g, SURFACE: #name1]
           adjective & [NUMBER: "sg", GENDER:#g2, SURFACE: #name2]? ):name
-> name: person_group & [NAME: #name, TYPE: "U_PER", GENDER:#g,
          AMOUNT:"1", NUMBER:"sg", RULE: "person_entity_sg"]
                & #name:=Concatenate(#name0,#name1,#name2)
                & IsEqual(#g0,#g1,#g,#g2).
```

Figure 4: Rule for detection of phrases referring to persons

More word token-level rules are also deployed in order to detect named person expressions in text such as *Doutor Eduardo R. Souza* etc. Finally, composition of *person* and *person_group* types into larger phrases is captured by higher level rules like the one shown on Figure 5, which matches appositional phrases like: *dois jovens de 20 anos, Paulo Souza e Gilberto Fernandez*.

Here, the constraints *IsNotUnspecified()* on the AMOUNT attributes are used to enforce the matching of person name coordinations and their appositive descriptions, while excluding underspecified quantifiers such as *algunas personas*.

All in all, person and entity recognition grammar is abstracting from surface forms and relies rather on: a) a number of fine-grained token classes (e.g., word-with-hyphen, word-with-apostrophe, all-capital-letters), which are to a large extent language-independent; b) person name and partial noun phrase syntactic structure; c) lexical resources for the target language. Because b) has low variation over Romance languages and only limited differences with respect to English, the process of person grammar porting onto Portuguese and Spanish was relatively straightforward and required limited level of linguistic expertise. Therefore, size and complexity of the grammars could be kept relatively low and the bulk of grammar development was mostly on providing suitable lexical resources.

We make use of two types of lexica:

1. morphological dictionaries for Portuguese and Spanish;

2. domain-specific lexicons, listing a number of (possibly multiword) expressions, subcategorized into semantic classes relevant for the domain of violent and disaster events, with limited or no linguistic annotation; classes range from person names, quantifying expressions (like *pelo menos dez*), through weapons and person positions (e.g. *grevistas*, *emigrante*, *passageiros*, *niños*, *mujer*).

As for the morphological dictionary, we make use of LABEL-LEX-sw electronic lexicon (Samuel et al., 1995), listing about 1M simple Portuguese wordforms. For Spanish we used MULTEXT, which encompasses 510K wordforms. It is important to note that we use MULTEXT (Erjavec, 2004) in order to perform morphological look-up, mainly due to the fact that MULTEXT tags are uniform for all languages. These resources are noticeably large, nonetheless we do not frequently make reference to abstract POS classes like nouns in our person recognition grammars as we noticed this highly exposes to the risk of overgeneralization and reduced accuracy. Consequently, we estimate a first potential bottleneck of the extraction process to be on the coverage and accuracy of domain-specific semantic classes; we will show in the next section how we generated and extended these resources.

3.3 Event triggering patterns

Prior to person recognition grammar application, event triggering linear patterns are matched on text for extraction of partial information on event roles, such as actors, victims, etc. These patterns are similar in spirit to the ones used in AutoSlog (Riloff, 1993). We use 1/2-slot surface level patterns like the following English and Portuguese samples, where role assignments are shown in brackets:

```
<DEAD> was shot by <PERPETRATOR>
police nabbes <ARRESTED>
<KIDNAPPED> has been taken hostage
<WOUNDED> was found injured
raptou <KIDNAPPED>
<DEAD> foram mortas
```

Note that the role slots (in brackets) can be filled by phrases referring to persons or person groups.

Patterns are stored in a domain-specific lexicon, each one associated with a type indicating the position of the pattern with respect to the slot to be filled (left or right), the event-specific semantic role assigned to the entity filling

```
person_group_apposition_rule :>
 (person_group & [NAME:#description, NUMBER:"p", AMOUNT:#amount1]
               token & [SURFACE: "," #com1]?
   person_group & [NAME:#name,NUMBER:"p",AMOUNT:#amount2]
   token & [SURFACE: ","]? ):noun_phrase
-> noun_phrase: person_group & [NAME:#final, AMOUNT:#amount1,
NUMBER:"p",RULE:"person_group_apposition_rule"]
& #final := ConcWithBlanks(#description,#com1,#name)
& IsEqual(#amount1,#amount2)
& IsNotUnspecified(#amount1)
& IsNotUnspecified(#amount2).
```

Figure 5: Rule for detection apposition phrases

the slot (e.g., DEAD, PERPETRATOR) and the grammatical number of the phrase which may fill the slot. For instance, the following represents the encoding for the surface pattern "foi sequestrada" detecting a kidnapped person:

```
foi sequestrada [
   TYPE: right-context-sg-and-pl,
   SURFACE: "foi sequestrada",
   SLOTTYPE: KIDNAPPED]
```

Through such a compact encoding, linear patterns can be then combined with detected person and person group entities at the top level of the grammar cascade via extraction rules like the simplified sample on Figure 6, which detects an Injuring event, extracting description and number of the victims.

These rules are meant to model simple domain-specific language constructions describing events, with extraction patterns being linearly non-overlapped with person phrases. For English language, strict word order and relatively simple morphology made such a surface level approach perform well in terms of both precision and recall (Piskorski, Tanev, and Wennerberg, 2007).

4 Semi-supervised resource acquisition

An important element in our approach is the usage of weakly supervised machine learning tools to acquire the language-specific resources which the system needs for processing the new languages. Namely, we use a news cluster based method for pattern learning, described in (Piskorski, Tanev, and Wennerberg, 2007) and we use a new weakly supervised approach (based on (Tanev and Magnini, 2006)) for learning of semantic categories, such as nouns, referring to people and weapons.

4.1 Ontopopulis - a system for learning of semantic categories

For each language our event extraction system should have among the other resources a list of phrases belonging to two semantic categories: weapons and persons. Event extraction uses this information in order to recognize entities mentioned in the articles (e.g. weapons) and also to parse noun phrases referring to specific semantic classes, such as people. We also learned several semantic categories which were used for event classification: vehicles, infrastructural objects, crimes, edge weapons and politicians.

There are different approaches for term extraction and categorization, however we have specific settings: First, we lack annotated data. On the other hand, we had available an unannotated corpus of Portuguese and Spanish news. Finally, we only had to learn few semantic classes. Considering this, we found quite relevant the weakly supervised term classification approach described in (Tanev and Magnini, 2006). Based on it and on its extention, presented by (Shi, Sun, and Che, 2007), we created our own term extraction and classification system - Ontopopulis.

Ontopopulis takes on its input a set of seed terms for each semantic category under consideration and an unannotated corpus of news articles. For example, for the category *weapons* in Portuguese we used terms like *arma branca*, *navalha*, *metralhadora*, etc. and for the category *persons*: *soldado*, *mulher*, *governador*, etc. The system performs two learning stages - Feature Extraction and Term Extraction:

4.1.1 Feature extraction and weighting

For each category (e.g. *weapons*), we consider as a context feature each uni-gram or bi-gram n which co-occurs at least 3 times in the corpus with any of the seed terms from this category (we have co-occurrence only when n is adjacent to a seed term on the left or on the right). The feature

```
injury-event :>
       ((person-group & [NAME: #name1, NUMBER: #num1]):injured1
         gazetteer & [POS: "conjuntion"]
        (person-group & [NAME: #name2, NUMBER: #num2]):injured2
         injured-phrase & [FORM: "passive"]
        ):event
-> injured1: victim & [NAME: #name1, NUMBER: #num1],
     injured2: victim & [NAME: #name2, NUMBER: #num2],
     event: injury & [VICTIM: #name, NUMBER: #count],
   & #name = Concatenate(#name1," & ",#name2)
   & #count = EstimateNumber(#num1," ",#num2).
```

Figure 6: Rule for detection of injury events

can not be composed only of stop words; we also do not consider words beginning with capitalized letters and numbers.

For each such a context feature n and a semantic category *cat* we calculate the score:

$$score(n, cat) = \sum_{st \in seeds(cat)} PMI(n, st)$$

where $seeds(cat)$ are the seeds terms of the category *cat* and $PMI(n, st)$ is the pointwise mutual information which shows the co-occurrence between the feature n and the seed term st.

At the end of this learning phase the user performs manual feature selection from a list of 250 best scored features, suggested by the system. This step guarantees high quality features which is very important for the accuracy of the final results. For example, some of the top ranking learned and approved features for *weapons* in our experiments are: *tiro de W*, *golpes de W*, *armado com W*, *ataque com W*, here *W* stands for the position where the weapon-terms should appear. Here are some examples of extracted features about for the class *vehicle*: *acidente com um V*, *bordo de um V*, *passageiros do V*.

4.1.2 Term extraction and weighting

The term extraction and learning stage takes the features, which were learned and manually selected for each category in the previous stage and extracts as candidate terms uni-grams and bigrams, which tend to co-occur with these features and which do not contain stop words, numbers or capitalized letters. Weighting of the candidate terms was carried out with the view to optimize the efficiency of the calculations. For this reason, we avoid to obtain the frequency of each candidate term in the corpus and we rather calculate the term feature vector in a non-standard way. It would be statistically more correct to use as a feature weight the pointwise mutual information between the term and the feature. However, this would require to collect statistics about the term frequency, which will decrease the algorithm speed.

We weight the term candidates, using the following algorithm:

1. For each category C we define a feature space, whose dimensions are only the features selected for this category
2. For each category C we define a *category feature vector* $\overrightarrow{C} = (wf_1, wf_2, wf_3, ..., wf_{nc})$ where wf_i are the weights of the category features, calculated as $wf_i = score(n_i, C)$, where n_i is the n-gram used as ith feature in our model; $score(n_i, C)$ is calculated with the pointwise-mutual-information based formula presented in the previous subsection.
3. We normalize each *category feature vector* $\overrightarrow{C}$ by dividing its coordinates with its length and obtain $norm(\overrightarrow{C})$.
4. Then, for each candidate term t for the category C we define a term feature vector $\overrightarrow{t_C} = (wt_1, wt_2, ..., wt_{nc})$ where $wt_i = \frac{ft_i}{ft_i+3}$, ft_i is the frequence with which the candidate term t appears with feature i.
5. Finally the weigth for each candidate term t for a category C is defined as a scalar product in the vector space defined for the category C, multiplied by the square root of the number of the non-zero features of the term feature vector: $weigth(t, C) = \overrightarrow{t_C}.norm(\overrightarrow{C}).\sqrt{NNZF(t_C)}$, where $NNZF$ is the number of the features with non zero weight.

Finally, the system orders the term candidates for each category by decreasing weight and filters out terms with a weight under a certain threshold. Then, the term list is given to the user for manual cleaning.

4.2 Learning linear patterns

In order to acquire the linear patterns for extraction of victims, perpetrators and arrested people, we implemented an iterative pattern acquisition algorithm, whose output is validated by a human moderator on each step. This algorithm was originally suggested by (Piskorski, Tanev, and Wennerberg, 2007). Their automatic approach takes on the input an annotated corpus and learns event specific templates. We modified the approach in such a way that it takes on its input a small set of seed patterns and a corpus without annotations. Then, as a first step we annotate the corpus using these patterns and then we run the original algorithm Here are the basic steps of the pattern learning algorithm:

1. For a specific role like dead victim, injured victim, perpetartor, etc. the user provides a small set of seed patterns. For example, let's consider the role dead victim and the small set of seed patterns: [PERSON] "mortas", [PERSON] "mortos", where [PERSON] matches any person description. We use the person recognition grammar and the semi-automatically learned list of person terms (see the previous sub-section), in order to extract phrases which refer to people, e.g. "cinco pessoas".
2. Annotate a corpus of news clusters, using these patterns. For example, if the text "cinco pessoas mortas" appears , the phrase "cinco pessoas" will be annotated as dead victim.
3. Propagate annotation inside the news clusters. At this step, if in a news cluster there is an annotated phrase, such as "cinco pessoas", then all the occurrences of this phrase inside the same cluster will be annotated with the same semantic role, e.g. dead victim. The assumption behind this step is that all the articles in a news cluster report about the same event, therefore equal phrases refer to the same entity which usually appears in the same semantic role across the whole cluster.
4. Learn automatically linear extraction patterns from the left and right contexts of the annotated phrases. For example, the phrase "cinco pessoas" and other annotated ones may appear systematically in phrases like"mata" [PERSON] as a result of the annotation propagation. As a consequence, such patterns will be added to the list of the learned ones. An entropy-based pattern extraction algorithm was used to perform this stage of the learning process (see (Piskorski, Tanev, and Wennerberg, 2007) for detailed description).
5. Manually filter out low quality patterns.
6. If the user estimates that the list of patterns is good enough, terminate.
   Otherwise, go to step 2.

We used successfully this algorithm for different languages, including the Portuguese and Spanish. Rarely, it was necessary to run more than two iterations. This approach facilitates the adaptation of the event extraction system to new languages by significantly decreasing the human efforts necessary to create language specific pattern libraries. Moreover, the algorithm does not need an annotated corpus. The human efforts are concentrated in the final step of each iteration, where the user is required to clean the output list of patterns, which in general requires less efforts than annotating a corpus. (We also experimented with manual corpus annotation and consequent pattern learning, however we found out that this approach is slower than the one presented here).

5 Experiments and Evaluation

We tested our methodology for Spanish and Portuguese. For each language we performed a series of resource-creation steps, which enabled NEXUS to extract event reports in the corresponding language:

1. Adapt the person recognition grammar from Italian
2. Run Ontopopulis to learn a dictionary of persons, weapons and other categories
3. Manually validate and clean the output of Ontopopulis
4. Create manually a small list of closed-class words and multiwords, such as quantifiers
5. Run the pattern learning algorithm for each of the following semantic roles: dead, wounded, kidnapped, perpetrator, and arrested
6. Manually clean the output of the pattern learning algorithm

5.1 Evaluation of Ontopopulis

The main purpose of the experiments was to evaluate the application of our methodology to Portuguese and Spanish. In this clue, there are two important parameters which can be used to estimate quantitatively the outcome of our experiments: First, the accuracy of Ontopopulis and

	person	weapon	politician	vehicle	watercraft	edged weapon	crime	building
seed terms	48	26	46	135	28	20	33	73
learned	930	122	990	315	173	45	911	1035
correct	473	44	226	123	39	4	397	360
precision	51%	36%	22%	39%	22%	8.8%	43%	34%
prec.top 20	90%	60%	75%	85%	70%	20%	85%	75%

Table 1: Evaluation of Ontopopulis for Portuguese

	person	weapon
seed terms	56	22
learned	578	900
correct	408	123
precision	71%	14%
prec.top 20	95%	60%

Table 2: Evaluation of Ontopopulis for Spanish

the linear pattern learning "per se" and second, the overall performance of NEXUS in terms of precision and recall.

We used Ontopopulis to learn several semantic classes. For each semantic class we manually filtered out the wrong terms before adding the list to the NEXUS resources. Note that in our experiments we limited the manual intervention to deleting while no adding or correction was allowed. In such a way we wanted to obtain a resource whose elements are all learned automatically.

We learned a dictionary of words and multiwords referring to people (e.g."enviado especial"). This dictionary is used intensively by the person recognition grammar. Moreover, it is the longest dictionary exploited by NEXUS and its manual creation would be quite time consuming. On this point, the application of Ontopopulis was very important. Another semantic class to learn was the class *weapons*, which NEXUS uses to detect the means by which violent acts were committed.

Additionally, we learned several other semantic classes to be used in the process of event classification, namely *politician*, *vehicle*, *watercraft*, *edged weapon*, *crime* and *building*. Event classification is performed by a set of over 30 event category definitions, which are composed of boolean operators over keywords. Category definition designing is usually a time consuming manual process which requires both domain knowledge and language competence. We tried to partially automatize this process by converting category definitions into more abstract boolean expressions over semantic classes, which we could learn by our semantic category learning algorithms. We do not report here about the performance of the overall event classification, but we show accuracy figures for the learning of these semantic classes.

For each semantic class, we provided a set of seed templates and run Ontopopulis. As training data we used two unannotated corpora - 3,4 million titles of news articles for Portuguese and 5,7 million news titles for Spanish. The results for Portuguese and Spanish are shown in Table 1 and Table 2, respectively.

For each semantic category we report the number of seed terms, the number of the new terms learned by the system, the number of correct learned terms, the overall precision and the precision in the top 20 ranked terms. The accuracy in the top 20 seems to be quite high for most of the categories with exception of *weapons* and its subclass *edged weapons*. The overall precision is lower, since the system threshold was set very low in order to increase the recall and add more resources for the event extraction system. This was safe since we manually clean the Ontopopulis output in a last step. However, the fact that the accuracy is relatively high in the top 20 shows that the system properly orders the learned terms by putting the most reliable ones on the top.

Another positive outcome of the application of Ontopopulis was that we increased the size of the term lists between 2 and 13 times for most of the categories, after manually validating the system output.

5.2 Evaluation of linear pattern learning

We run linear pattern learning in order to obtain linear patterns for extraction of several domain-specific semantc roles: DEAD, WOUNDED, KIDNAPPED, ARRESTED and PERPETRATOR. As an example, for the *dead* role one of the Portuguese patterns the system learned was

	dead	wounded	kidnapped	arrested	perpetrator
seed patterns	12	7	31	10	38
learned	382	104	178	78	113
correct	54	11	24	28	19
precision	14%	11%	13%	36%	17%

Table 3: Evaluation of pattern learning for Portuguese

	dead	injured	arrested
seed terms	22	25	15
learned	108	10	15
correct	30	5	9
precision	28%	50%	60%

Table 4: Evaluation of pattern learning for Spanish

	DEAD	WOUNDED	KIDNAPPED	ARRESTED
baseline Portuguese	0.62	0.53	0.54	0.29
target Portuguese	0.69	0.51	0.67	0.47
baseline Spanish	0.12	0	0	0.125
target Spanish	0.46	0	0	0.125

Table 5: Evaluation of extraction of different roles in terms of F1-measure

"assassinato do [PERSON]".

The experiments we report about here consisted of one learning iteration only. After that we manually filtered out unappropriate patterns. The results in Table 3 and Table 4 show the performance of the pattern learning algorithm for Portuguese and Spanish language, respectively - before the manual validation[3].

5.3 Evaluation of NEXUS

Test data were gathered by downloading EMM article clusters during 30 consecutive days in April 2009. The final test corpus was selected from these clusters as a sample of 100, which report about security and disaster-related topics.

On this corpus, we ran a baseline version of the system for both languages, namely the one based on seed linear patterns and seed dictionaries of persons and weapons. We also ran a target version in which we added to the seed resources the cleaned output of Ontopopulis for the classes *person* and *weapons* and the output of the pattern learning algorithm. We denote the baseline and target system with BL and TG, respectively.

Table 5 shows a comparative evaluation of the two baseline and target event extraction systems for Portuguese and Spanish.

In particular, we measured Precision (P), Recall (R) and F-measure for each role. We only show F-measure figures for a more compact comparison. Moreover, test data were slightly sparse, as some of the roles were not instantiated in text - namely RELEASED and PERPETRATOR - due to the relatively small corpus size. Therefore we do not report about them in the final evaluation.

Evaluation was done separately for each role, and data were collected cluster by cluster. Namely, for each cluster of articles we record if it contains a reference to the filler of a specific role; then we record if the system detected any filler whatsoever for that role, and finally, we record a correct detection if the returned role filler description equals at least one of the descriptions occurring in the cluster.

The comparative evaluation of the Portuguese baseline and target systems clearly shows that the target system performs better. On average, the F-measure improved by 0.09 in the target system. The maximal improvement was for the category ARRESTED - the F-measure improved from 0.286 for the baseline system to 0.47 for the target one. The average recall improvement was found to be 12%. The best improvement of the recall was for the role KIDNAPPED - from 60% to 80%. Moreover, the improvement in the recall was not at the cost of reduced precision, as on average the precision still improved by about 1%. Even if these results can be improved further, they demonstrate that machine learning algorithms bring improvement in the overall performance of the event extraction system. Data are less impressive for Spanish, and more sparse. Nonetheless, an even larger improvement in terms of F-measure could be recorded for the DEAD role.

[3]Results are only partial for Spanish due to data sparseness of the training corpus.

6 Conclusions

We presented a multilingual methodology for adapting an existing event extraction system to Portuguese and Spanish languages. The approach relies on weakly supervised learning of domain-specific lexicons, and requires minimal amount of domain and linguistic knowledge.

In our experimental settings, we only performed one learning stage, with no fine-tuning. Therefore, system performance in absolute terms was not excellent. Nonetheless, we believe that figures on the improved performance of the learned systems are encouraging, so that we plan to pursue in optimizing the development process. Moreover, the approach seems to be portable in the same way over semantic domains. One possible research direction would be then to test the methodology on adapting the event extaction system to new application domains.

The live event extraction system for Portuguese is publicly accessible at `http://press.jrc.it/geo?type=event&format=html&language=pt`. For the Spanish version change the value of the language attribute to es.

References

Aone, C. and M. Santacruz. 2000. Rees: A large-scale relation and event extraction system. In *Proceedings of ANLP 2000, 6th Applied Natural Language Processing Conference*, Seattle, Washington, USA.

Appelt, D. 1999. Introduction to information extraction technology. Tutorial held at IJCAI-99.

Bell, A. 1991. *The Language of News Media*. Blackwell.

Erjavec, Tomaz. 2004. Multext - east morphosyntactic specifications. http://nl.ijs.si/ME/V3/msd/html.

Grishman, R., Huttunen S. and R. Yangarber. 2002. Real-time event extraction for infectious disease outbreaks. In *Proceedings of Human Language Technology Conference*, San Diego, USA.

Grishman, R., Huttunen S. and R. Yangarber. 2003. Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4).

Ji, H. and R. Grishman. 2008. Refining event extraction through unsupervised cross-document inference. In *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio, USA.

King, G. and W. Lowe. 2003. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57:617–642.

Naughton, M., Kushmerick N. and J. Carthy. 2006. Event extraction from heterogeneous news sources. In *Proceedings of the AAAI 2006 workshop on Event Extraction and Synthesis*, Menlo Park, California, USA.

Piskorski, J. 2007. Express extraction pattern recognition engine and specification suite. In *Proceedings of the International Workshop Finite-State Methods and Natural language Processing*, Potsdam, Germany.

Piskorski, J. 2008. Corleone core linguistic entity online extraction. *Technical Report EUR 23393 EN*.

Piskorski, Jakub, Hristo Tanev, and Pinar Oezden Wennerberg. 2007. Extracting violent events from on-line news for ontology population. In *BIS*, pages 287–300.

Riloff, E. 1993. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the 11th National Conference on Artificial Intelligence*.

Samuel, E., E. Ranchhod, H. Freire, and J. Baptista. 1995. A system of electronic dictionaries of portuguese. *Lingvisticae Investigationes*, XIX:2.

Shi, Lian, J. Sun, and H. Che. 2007. Populating crab ontology using context-profile based approaches. *Knowledge Science, Engineering and Management, LNCS*.

Steinberger, R., B. Pouliquen, and E. van der Goot. 2009. An introduction to the europe media monitor family of applications. In *Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop*, Boston, USA.

Tanev, Hristo and B. Magnini. 2006. Weakly supervised approaches for ontology population. In *Proceedings of the European Chapter of the Association of Computational Linguistics*, Trento, Italy.

Wagner, E., Liu J. Birnbaum L. Forbus K. and J. Baker. 2006. Using explicit semantic models to track situations across news articles. In *Proceedings of the AAAI 2006 workshop on Event Extraction and Synthesis*, Menlo Park, California, USA.

Yangarber, R., Jokipii L., A. Rauramo, and S. Huttunen. 2005. Extracting information about outbreaks of infectious epidemics. In *Proceedings of the HLT-EMNLP 2005*, Vancouver, Canada.

Un algoritmo lingüístico-estadístico para resumen automático de textos especializados

Iria da Cunha,
IULA-Universitat Pompeu Fabra y
LIA-Université d'Avignon
iria.dacunha@upf.edu

Juan-Manuel Torres-Moreno,
LIA-Université d'Avignon y
Ecole Polytechnique de Montréal
juan-manuel.torres@univ-avignon.fr

Patricia Velázquez-Morales,
patricia_velazquez@yahoo.com

Jorge Vivaldi
IULA-Universitat Pompeu Fabra
jorge.vivaldi@upf.edu

Resumen

En este trabajo se presenta un nuevo algoritmo de resumen automático de textos especializados, en concreto del dominio médico, que aúna estrategias lingüísticas y estadísticas. La novedad del artículo radica en la correcta combinación de dichas estrategias de cara a demostrar que los sistemas híbridos pueden obtener mejores resultados que los sistemas estadísticos o lingüísticos por sí solos. Se aplica el algoritmo sobre un corpus de textos médicos y se evalúa siguiendo el protocolo de NIST y utilizando el paquete ROUGE. Se obtienen excelentes resultados en comparación con otros sistemas y se observa que los resúmenes realizados son muy similares a los de los especialistas del dominio.

1. Introducción

El resumen automático es actualmente un tema de investigación muy relevante. La investigación en esta área se inició en los años sesenta, empleando técnicas basadas en frecuencias de palabras (Luhn, 1959) o frases clave (Edmundson, 1969). Con el tiempo, estas técnicas han ido evolucionando y volviéndose más complejas. Podemos hacer una división general de estas técnicas en dos grupos principales: las técnicas estadísticas y las técnicas lingüísticas. En el primer grupo[1], encontramos, entre otros, trabajos que emplean modelos bayesianos (Kupiec, Pedersen, and Chen, 1995), la *Maximal Marginal Relevance* (Goldstein et al., 1999), técnicas de *clustering* (Radev, Jing, and Budzikowska, 2000), grafos (Radev et al., 2004; Vanderwende, Banko, and Menezes, 2004; Leskovec, Milic-Frayling, and Grobelnik, 2005) o aprendizaje automático (Kupiec, Pedersen, and Chen, 1995; Berger and Mittal, 2000; Marcu and Echihabi, 2002; Leskovec, Milic-Frayling, and Grobelnik, 2005; Barzilay and Lapata, 2005). En el segundo grupo, destacamos trabajos que explotan las posiciones textuales (Brandow, Mitze, and Rau, 1995; Lin and Hovy, 1997), la estructura del discurso (Ono, Sumita, and Miike, 1994; Marcu, 1998; Marcu, 2000; Teufel and Moens, 2002; Polanyi et al., 2004; Thione et al., 2004) o las cadenas léxicas (Barzilay and Elhadad, 1997; Silber and McCoy, 2000; Fuentes, 2008). Todos estos sistemas de resumen automático emplean estrategias que conllevan algún tipo de criterio lingüístico y estadístico, pero por lo general siempre hay una mayor proporción de uno u otro. Hay pocos trabajos en los que se combinan ambos criterios de una manera más igualitaria, por ejemplo Nomoto, T. and Nitta, Y. (1994), Aretoulaki (1996), Hovy, E. and Lin, C.Y. (1999), Alonso and Fuentes (2003) y Lacatusu, Parker, and Harabagiu (2003). En nuestro trabajo intentamos combinar técnicas estadísticas y lingüísticas de una manera equitativa y adecuada, para aprovechar las ventajas de ambas en la tarea de resumen automático. En concreto, hemos diseñado un algoritmo de resumen que combina los sistemas estadísticos CORTEX (Torres-Moreno, Velázquez-Morales, and Meunier, 2001) y ENERTEX (Fernández, SanJuan, and Torres-Moreno, 2007) y los sistemas lingüísticos YATE (Vivaldi, 2001) y DISICOSUM (da Cunha, 2008). En da Cunha et al. (2007) se realizó una primera aproximación, diseñándose un algoritmo que combinaba estos sistemas. En este nuevo trabajo se desarrolla y se refina ese algoritmo inicial, se realizan nuevos experimentos y se evalúan mediante el paquete ROUGE (Lin, 2004), usando un protocolo similar al de las evaluaciones del

[1] No pretendemos hacer aquí una revisión exhaustiva del estado de la cuestión en resumen automático. Para más información sobre técnicas y/o sistemas de resumen remitimos a los trabajos de (Spärck Jones, 2007; Mani, 2001; Mani and Maybury, 1999).

NIST[2]. La decisión de llevar a cabo estas mejoras se tomó porque el algoritmo inicial fue diseñado para resumir textos largos, de cuatro o cinco páginas y con diversos apartados, mientras que ahora nos planteamos obtener resúmenes de textos de contenido específico y más cortos, de una página aproximadamente.

Nuestra tarea consiste en resumir textos médicos especializados. Tal como se indica en Afantenos, Karkaletsis, and Stamatopoulos (2005), en el ámbito médico debe gestionarse una gran cantidad de textos y el resumen automático puede ayudar a procesar esta masa de documentos. Nosotros trabajamos con artículos médicos de investigación, ya que este tipo de textos se publican con sus correspondientes resúmenes escritos por los autores de los textos y esto nos permite compararlos con los resúmenes obtenidos por nuestro sistema y facilitar así su evaluación. En el futuro podría adaptarse el componente lingüístico de nuestro algoritmo para ser empleado en otras áreas similares, como la biología, la genómica, la química, etc., así como en otras lenguas próximas.

En la sección 2 detallamos los diversos componentes del algoritmo. En la sección 3 explicamos su arquitectura. En la sección 4 mostramos los experimentos realizados y su evaluación. Por último, en la sección 5, exponemos las conclusiones y algunas perspectivas.

2. Componentes del algoritmo

A continuación explicamos los cuatro sistemas que se han empleado como componentes del algoritmo.

2.1. CORTEX

Cortex (Torres-Moreno, Velázquez-Morales, and Meunier, 2001; Torres-Moreno, Velázquez-Morales, and Meunier, 2002) es un sistema de resumen automático basado en el Modelo de Espacio Vectorial (VSM) (Salton and McGill, 1983). Se trata de un sistema de resumen por extracción mono-documento que combina varias métricas sin aprendizaje. Estas métricas resultan de algoritmos de procesamiento estadísticos y de información sobre la representación vectorial del documento. La idea principal es la de representar un texto en un espacio vectorial adecuado y aplicar procesamiento estadístico. Con el fin de reducir la complejidad del espacio, se realiza un preprocesamiento del documento: se filtran y se lematizan las palabras del texto. La representación de *bolsa-de-palabras* produce una matriz de frecuencias/ausencias $S_{[P\times N]}$ de $\mu = 1, \cdots, P$ frases u oraciones[3] (filas) y un vocabulario de $i = 1, \cdots, N$ términos (columnas). Cortex puede emplear hasta $\Gamma = 11$ métricas para evaluar la pertinencia de las frases. Algunas métricas utilizan el ángulo entre el título y cada una de las frases, la matriz de Hamming (matriz donde cada valor representa el número de frases en las que uno de los términos i o j está presente), la suma de pesos Hamming de palabras por segmento, la entropía, la frecuencia o las interacciones, entre otras. El sistema asigna una puntuación a cada frase con el algoritmo de decisión que combina las métricas normalizadas. Se calculan dos promedios: una tendencia positiva $\lambda_s > 0{,}5$ y otra negativa $\lambda_s < 0{,}5$ (el caso $\lambda_s = 0{,}5$ es ignorado). El algoritmo de decisión que permite combinar el voto de Γ métricas es el siguiente:

$$\sum \alpha = \sum_{\nu=1}(||\lambda_s^\nu|| - 0{,}5); ||\lambda_s^v|| > 0{,}5 \quad (1)$$

$$\sum \beta = \sum_{\nu=1}(0{,}5 - ||\lambda_s^\nu||); ||\lambda_s^v|| < 0{,}5 \quad (2)$$

Γ es el número de métricas y ν es el índice de las métricas. El valor de λ fue normalizado en el rango $[0-1]$ para evitar diferencias de magnitud entre las métricas. El valor dado a cada frase s se calcula de la siguiente manera:

$$\text{IF } (\sum \alpha > \sum \beta) \text{ THEN Score}_s = 0{,}5 + \frac{\sum \alpha}{\Gamma}$$

$$\text{ELSE Score}_s = 0{,}5 - \frac{\sum \beta}{\Gamma}$$

La figura 1 muestra el esquema del resumidor Cortex. En los experimentos se usaron las métricas FAX (F=Frecuencia de términos, A=Ángulo entre el título y cada una de las frases y X=Posición de la frase en el documento, modelada con una función cuadrática que otorga una ponderación mayor a las primeras y últimas frases del texto).

2.2. ENERTEX

Enertex (Fernández, SanJuan, and Torres-Moreno, 2007; Fernández, SanJuan, and Torres-Moreno, 2008; Fernández, 2009) también es un sistema de resumen automático basado en VSM, pero en este caso se trata de un enfoque de redes de neuronas inspirado en la física estadística. El algoritmo modela los documentos como una red

[2]Más información sobre las campañas *Document Understand Conference* y *Text Analysis Conference* (DUC/TAC) puede encontrarse en el sitio web del NIST: `http://www.nist.gov/tac/`

[3]En función del segmentador empleado.

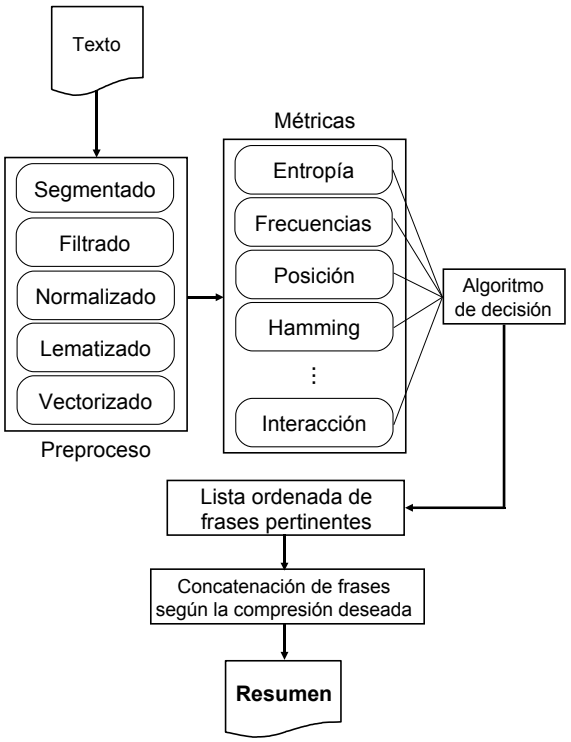

Figura 1: Arquitectura de CORTEX.

de neuronas de la que se estudia su energía textual. La idea principal es que un documento puede ser procesado como un conjunto de unidades interactivas (las palabras), donde cada unidad se ve afectada por el campo creado por las demás (véase figura 2).

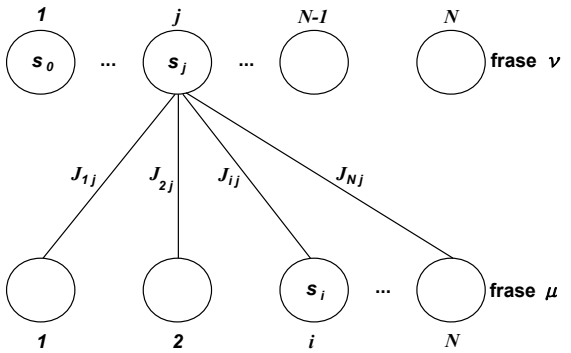

Figura 2: Campo creado por los términos de la frase μ que afecta cada término j de la frase ν.

La memoria asociativa de Hopfield (Hopfield, 1982) está basada en sistemas físicos, como el modelo magnético de Ising (formalismo que describe un sistema con dos estados, llamados *spines*, para construir una red de neuronas capaz de almacenar/recuperar patrones). El aprendizaje se realiza usando la regla de Hebb (Hertz, Krogh, and Palmer, 1991):

$$J_{i,j} = \sum_{\text{P frases}} s_i s_j \tag{3}$$

y la recuperación por minimización de la energía del modelo de Ising:

$$E^{\mu,\nu} = \sum s_i^{\mu} J_{i,j} s_j^{\nu} \tag{4}$$

La principal limitación de la red de Hopfield es su capacidad de almacenamiento: los patrones no deben estar correlacionados para poder obtener sin problemas el error de recuperación. Esto restringe en gran medida sus aplicaciones, pero ENERTEX se beneficia de este comportamiento. El VSM representa las frases del documento en vectores (partiendo de la vectorización producida por CORTEX) y estos vectores pueden estudiarse como una red neuronal de Hopfield. Las frases son representadas como una cadena (patrón) de N neuronas activas (término presente) o inactivas (término ausente) con un vocabulario de N términos por documento. Un documento de P frases está formado por P cadenas en un espacio vectorial de N dimensiones. Estos vectores están relacionados en función de las palabras compartidas. Si los temas están semánticamente próximos, es razonable suponer un alto grado de correlación. ENERTEX calcula la interacción entre los términos (3) y la energía textual entre las frases (4). La ponderación de las frases se obtiene utilizando sus valores absolutos de energía. El resumen está formado por las frases más importantes, es decir las que obtienen los valores más altos.

2.3. DISICOSUM

DISICOSUM (da Cunha and Wanner, 2005; da Cunha, Wanner, and Cabré, 2007; da Cunha, 2008) es un modelo de resumen automático de textos médicos que parte de la idea de que los profesionales de un dominio especializado emplean técnicas concretas para resumir los textos de su ámbito. Para diseñar este sistema se analizó un corpus de artículos médicos con sus correspondientes resúmenes para determinar cuáles son las estrategias que usan los médicos para resumir y cuál es la información que debe seleccionarse para realizar un resumen de este tipo. Otra de las aportaciones del modelo es la combinación de diversos criterios lingüísticos. Por lo general (ver sección 1) los sistemas lingüísticos de resumen automático suelen emplear un solo tipo de criterio (por ejemplo, frecuencias de palabras, oraciones clave, posiciones textuales, estructura discursiva, etc.). Sin embargo, en DISICOSUM se integran criterios basados en la estructura textual, en las unidades léxicas y en la estructura discursiva y sintáctico-comunicativa del texto. El modelo

está formado por reglas que se relacionan con estos criterios lingüísticos.

Con respecto a la estructura textual, DISICOSUM incluye una regla que asigna un peso adicional a las oraciones que se encuentren en las siguientes posiciones del texto: las tres primeras oraciones de la sección de *Fundamento*, las dos primeras oraciones de las secciones de *Pacientes y métodos* y *Resultados*, y las tres primeras y las tres últimas oraciones de la sección de *Discusión*.

En cuanto a las unidades léxicas, DISICOSUM incluye reglas de dos tipos:

- Reglas que otorgan más peso a las oraciones que contienen: 1) palabras del título principal del artículo (excepto *stopwords*), 2) formas verbales en primera persona del plural, 3) palabras incluidas en una lista que contiene verbos y sustantivos del dominio médico que pueden ser pertinentes para el resumen (por ejemplo, *analizar*, *evaluar*, *objetivo*, *estudio*, etc.) y 4) cualquier información numérica en las secciones de *Pacientes y métodos* y *Resultados*.
- Reglas de eliminación de oraciones que contienen unidades que se refieren a: 1) tablas o figuras, 2) aspectos estadísticos o computacionales, 3) trabajos anteriores y 4) definiciones.

Por último, el modelo incluye reglas discursivas y reglas que combinan algunos aspectos de la estructura discursiva con la estructura sintáctica y comunicativa (reglas DISICO). Para formalizar dichas reglas el modelo emplea la *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988) y la *Meaning-Text Theory* (MTT) (Mel'cuk, 1988; Mel'cuk, 2001). La RST es una teoría de organización del texto que caracteriza su estructura como un árbol jerárquico que contiene elementos (núcleos [N] y satélites [S]) ligados mediante relaciones discursivas (como, por ejemplo, Elaboración, Concesión, Antítesis, Condición, Contraste, *Background*, etc.). En la figura 3 se muestra un ejemplo de la representación discursiva de la RST en forma de árbol (con dos relaciones: Elaboración y *Background*).

La MTT es una teoría que integra diversos aspectos del lenguaje. DISICOSUM emplea, por un lado, elementos de la sintaxis de dependencias para representar una oración como un árbol donde las unidades léxicas son los nodos y las relaciones entre ellas son actanciales (ACT), atributivas (ATTR), apenditivas (APPEND) y coordinativas (COORD). Por otro lado, DISICOSUM emplea la distinción ente Tema y Rema, que es parte de

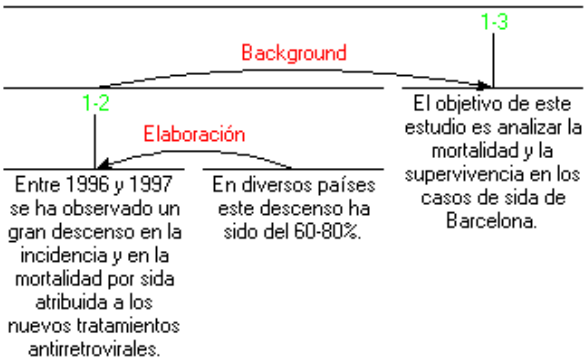

Figura 3: Ejemplo de árbol discursivo con relaciones de la RST.

la estructura comunicativa de la MTT. Algunos ejemplos de las reglas DISICOSUM son:[4,5]

- IF S is satellite$_{REFORMULATION}$ R
  THEN ELIMINATE S
  Ej. [Se incluyeron sólo pacientes estables.]$_N$ [*Es decir, se consideraron pacientes que no habían precisado cambiar su medicación habitual en los últimos 15 días y clínicamente no referían un empeoramiento importante.*]$_S$
- IF S is satellite$_{BACKGROUND}$ B
  THEN ELIMINATE S
  Ej. [*La quimioprofilaxis (QP) antituberculosa es una de las principales intervenciones en la cadena de actuaciones para la prevención de la tuberculosis (TBC).*]$_S$ [El objetivo de este estudio es conocer el grado de cumplimiento y la tolerancia terapéutica de la QP antituberculosa en nuestro medio, así como describir y analizar sus factores condicionantes.]$_N$
- IF S is satellite$_{ELABORATION}$ El
  AND S elaborates on the Theme of the nucleus of El
  THEN ELIMINATE S
  Ej. [Como grupo de control se empleó el formado por 377 mujeres sanas.]$_N$ [*Este grupo se obtuvo mediante selección aleatoria entre mujeres que entre 1989 y 1991 habían dado a luz en nuestro hospital.*]$_S$
- IF N is nucleus$_{LIST}$ L
  THEN KEEP N
  Ej. [La primera prueba de marcha (PM) se

[4]La lista de todas las reglas que conforman DISICOSUM puede observarse en (da Cunha, 2008).

[5]Los fragmentos de texto en itálica son eliminados por las reglas.

efectuó respirando aire sintético a 2 litros por minuto a través de gafas nasales.]$_N$ [La segunda PM se realizó con oxígeno continuo a 2 litros por minuto.]$_N$ [La tercera se llevó a cabo acoplando a la misma fuente una VAO a un flujo de 2 litros por minuto.]$_N$

DISICOSUM es un modelo de resumen semiautomático, debido a la carencia actual de analizadores automáticos discursivos para el castellano[6]. Así pues, los textos de entrada deben estar previamente anotados con etiquetas que reflejen las relaciones de la RST. En la figura 4 se muestra la arquitectura del modelo.

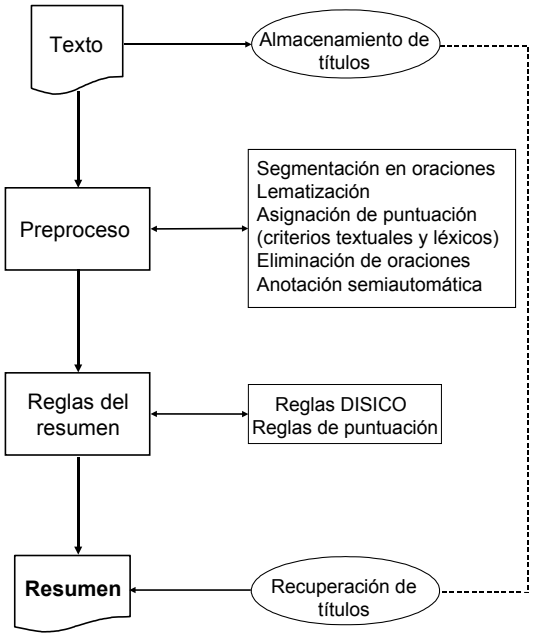

Figura 4: Arquitectura de DISICOSUM.

2.4. YATE

Otro de los sistemas empleados en este trabajo es el sistema YATE (Vivaldi, 2001; Vivaldi and Rodríguez, 2001; Vivaldi and Rodríguez, 2002). YATE es un extractor de términos híbrido cuyas características más relevantes son: el uso intensivo de información semántica junto con el uso de técnicas de combinación de los resultados obtenidos a partir de diferentes técnicas de extracción. Al igual que otros extractores de similares características ha sido desarrollado para el ámbito médico en castellano aunque está siendo adaptado con éxito a otros dominios (genómica, derecho, economía, informática y medio ambiente) y otras lenguas (catalán).

La figura 5 muestra el esquema general de YATE y permite apreciar los diferentes módulos de análisis que forman este extractor:

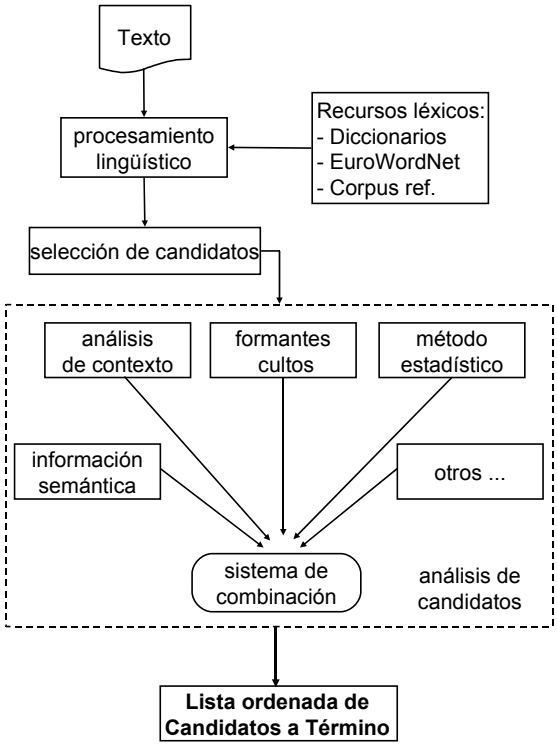

Figura 5: Arquitectura de YATE.

1. Información semántica: utiliza información obtenida a partir de EuroWordNet[7].
2. Análisis de contexto: evalúa cada candidato a término (CAT) en función de otros candidatos que aparecen en su contexto oracional.
3. Formantes cultos: descompone cada CAT en sus formantes aprovechando las características formales de algunos términos (sobre todo en el dominio médico).
4. Método estadístico: evalúa los CAT poliléxicos según su información mutua (u otra medida de asociación).

YATE puede ser eficazmente integrado como un componente de ponderación de palabras en un

[6] Actualmente hay analizadores discursivos para el japonés (Sumita et al., 1992), el inglés (Marcu, 2000) y el portugués de Brasil (Pardo, T. and Nunes, M. and Rino, M., 2004; Pardo, T. and Nunes, M., 2008). También existe un proyecto en curso en el *Laboratoire Informatique d'Avignon* para desarrollar un analizador discursivo para el castellano.

[7] `http://www.illc.uva.nl/EuroWordNet`

sistema de resumen, como CORTEX o ENERTEX por ejemplo, de manera simple. En nuestro trabajo, hemos usado la capacidad de detección de términos médicos deYATE para ponderar cada término del modelo vectorial proporcionalmente a su terminologicidad (valor comprendido entre 0 y 1). Así, las oraciones que contienen términos del dominio médico serán bonificadas en relación a las otras.

3. Arquitectura del algoritmo

El algoritmo desarrollado, que como hemos mencionado anteriormente parte del trabajo de (da Cunha et al., 2007), combina los cuatro sistemas detallados en la sección 2: CORTEX, YATE, ENERTEX y DISICOSUM.

El sistema propuesto tiene la arquitectura que se indica en la figura 6 y consta de varios resumidores autónomos que se combinan de manera equilibrada para formar un único resumidor híbrido. Algunos de los resumidores utilizan métodos numéricos (CORTEX y ENERTEX), otro resumidor tiene un carácter estrictamente lingüístico (DISICOSUM) y en los dos sistemas restantes las métricas estadísticas (de CORTEX y ENERTEX) se combinan con la información lingüística procedente del extractor de términos (YATE).

Tomando como entrada los resúmenes realizados por estos sistemas, el algoritmo de decisión para la selección de las oraciones del resumen incluye hasta cuatro fases (depende de los resultados de cada fase que sea o no necesaria la siguiente fase), que se aplican en el siguiente orden:

- *Fase 1.* ACUERDO: si una oración del texto es seleccionada por todos los sistemas, el algoritmo la mantiene.
- *Fase 2.* MAYORÍA: si una oración del texto es seleccionada por la mayoría de los sistemas, el algoritmo la mantiene.
- *Fase 3.* SCORE: si una oración es seleccionada solo por uno o dos sistemas, el algoritmo elije la que tenga asignado un mayor *score*.
- *Fase 4.* SCORE + ORDEN DE LAS ORACIONES EN EL TEXTO ORIGINAL: si se necesita una cantidad determinada de oraciones para el resumen y varias oraciones coinciden en su *score*, el algoritmo prioriza la que aparece en primer lugar en el texto original.

El algoritmo puede ser construido siguiendo diversas combinaciones de los sistemas. Sin embargo, buscando contar siempre con la presencia del sistema lingüístico, decidimos evaluar las combinaciones posibles de DISICOMSUM con los sistemas estadísticos (modificando o no el peso de los términos por YATE en CORTEX o ENERTEX). Las pruebas mostraron que la combinación de CORTEX, CORTEX+YATE, ENERTEX y DISICOSUM obtiene los mejores resultados. Por esto nos referiremos a esta combinación como el sistema o algoritmo híbrido lingüístico-estadístico[8].

Veamos un ejemplo para ilustrar el funcionamiento de este algoritmo. Imaginemos que necesitamos un resumen de 5 oraciones, a partir de un texto que contiene 10, como el mostrado en el cuadro 1. Las oraciones son numeradas de acuerdo al orden de aparición en el texto original.

Nº	Oración
1	En Ferrol las resistencias de M. tuberculosis son bajas, por lo que no está justificado un tratamiento inicial con 4 fármacos antituberculosos.
2	Los factores asociados a la presencia de resistencias son la edad y la existencia de tratamiento previo.
3	Nuestros valores de resistencias son similares o inferiores a las existentes en otras áreas geográficas.
4	No existieron diferencias en las resistencias primarias o secundarias según la presencia o no de infección por el VIH como en otros estudios, aunque algunos autores comunicaron mayor frecuencia de resistencias primarias y secundarias en pacientes positivos para el VIH.
5	Al comparar los resultados con otros estudios debe indicarse que no siempre utilizan la misma metodología, a veces no especifican el tipo de resistencias y algunos no aportan resultados según la presencia o no de infección por el VIH, o cuando lo hacen, el porcentaje de coinfección por el VIH en los pacientes con tuberculosis es muy variable.
6	La frecuencia de las resistencias de M. tuberculosis a los fármacos antituberculosos y su evolución en el tiempo debe ser un elemento de seguimiento epidemiológico de los programas de prevención y control.
7	Para que los resultados sean fiables y reflejen el funcionamiento de un programa, la muestra de pacientes estudiados debe ser representativa del área y se han de tener en cuenta los movimientos migratorios o la posibilidad de contaminaciones en el laboratorio.
8	El porcentaje de resistencias primarias nos ofrece una idea del funcionamiento de las medidas de prevención por cuanto reflejarían la mutación espontánea de M. tuberculosis y la transmisión de cepas resistentes desde pacientes bacilíferos.
9	Las resistencias secundarias reflejan el funcionamiento de las medidas de control, ya que indican la idoneidad y el buen cumplimiento de las pautas terapéuticas.
10	En nuestro medio se incumple el tratamiento en el 8,5 % de los casos y se emplean combinaciones farmacológicas desde hace años, lo que podría justificar la menor frecuencia de resistencias secundarias respecto a la de otras zonas geográficas con incumplimientos de hasta el 50 % de los casos.

Cuadro 1: Texto a modo de ejemplo con las oraciones divididas según su orden de aparición en el texto original.

El cuadro 2 muestra las oraciones seleccionadas por cada sistema para un resumen de 5 oraciones teniendo en cuenta el mayor *score* obtenido (indicamos solo el número de la oración en el texto original con su respectivo *score* normalizado).

A partir de estos datos de entrada el sistema pasa (en este caso) por todas las fases del algo-

[8]Constatamos que YATE apenas mejora el desempeño de ENERTEX por ejemplo, y resulta ligeramente contraproducente incluir YATE+ENERTEX en el algoritmo híbrido.

Cortex		Cortex+ Yate		Enertex		Disicosum	
Oración	*Score*	Oración	*Score*	Oración	*Score*	Oración	*Score*
4	0.9	4	0.9	1	0.9	1	0.9
2	0.4	5	0.9	10	0.8	10	0.8
1	0.3	2	0.8	5	0.7	2	0.6
3	0.1	1	0.7	6	0.6	8	0.5
6	0.1	6	0.5	7	0.3	9	0.4

Cuadro 2: Oraciones seleccionadas por los cuatro sistemas para un resumen de 5 oraciones del texto del cuadro 1 de acuerdo a los mayores *scores* obtenidos.

ritmo de decisión para seleccionar las oraciones que producirán el resumen final. El algoritmo incluirá en el resumen las cinco oraciones siguientes:

- oración 1 (seleccionada en la Fase 1)
- oración 2 (seleccionada en la Fase 2)
- oración 6 (seleccionada en la Fase 2)
- oración 4 (seleccionada en la Fase 3)
- oración 5 (seleccionada en la Fase 4)

Finalmente se reordenan estas oraciones en el orden del texto original para obtener el resumen final, que se presenta en el cuadro 3.

Nº	Oración
1	En Ferrol las resistencias de M. tuberculosis son bajas, por lo que no está justificado un tratamiento inicial con 4 fármacos antituberculosos.
2	Los factores asociados a la presencia de resistencias son la edad y la existencia de tratamiento previo.
4	No existieron diferencias en las resistencias primarias o secundarias según la presencia o no de infección por el VIH como en otros estudios, aunque algunos autores comunicaron mayor frecuencia de resistencias primarias y secundarias en pacientes positivos para el VIH.
5	Al comparar los resultados con otros estudios debe indicarse que no siempre utilizan la misma metodología, a veces no especifican el tipo de resistencias y algunos no aportan resultados según la presencia o no de infección por el VIH, o cuando lo hacen, el porcentaje de coinfección por el VIH en los pacientes con tuberculosis es muy variable.
6	La frecuencia de las resistencias de M. tuberculosis a los fármacos antituberculosos y su evolución en el tiempo debe ser un elemento de seguimiento epidemiológico de los programas de prevención y control.

Cuadro 3: Oraciones del resumen final obtenido mediante el algoritmo para el ejemplo presentado.

4. Experimentos y evaluación

El algoritmo de resumen automático desarrollado se aplicó sobre un corpus formado por 40 textos en castellano extraídos de la revista *Medicina Clínica*. Cada texto es un apartado de un artículo médico, de aproximadamente una página. Dependiendo del tipo de apartado del artículo se realizaron resúmenes de diferentes longitudes: 2 oraciones del apartado de *Fundamento*, 3 oraciones del apartado de *Pacientes y métodos*, 4 oraciones del apartado de *Resultados* y 2 oraciones

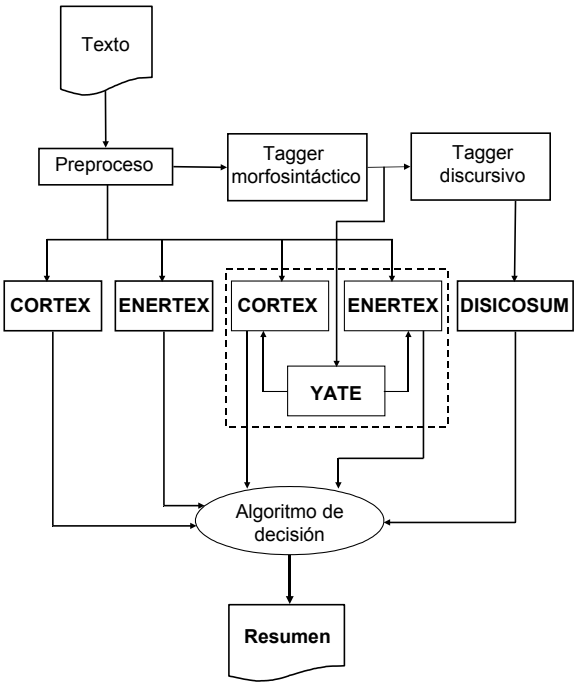

Figura 6: Arquitectura del sistema de resumen híbrido lingüístico-estadístico.

del apartado de *Discusión*. Para determinar este número, se calculó el promedio de oraciones incluidas en cada apartado de los *abstracts* de los autores. A continuación, se decidió incluir una oración adicional, ya que notamos que en muchas ocasiones en estos *abstracts* se fusionan en una los contenidos de dos o más oraciones de los artículos (veremos un ejemplo al final de este apartado, en el cuadro 5). En definitiva, fue una decisión empírica para no perder información.

Para evaluar los resúmenes se empleó un protocolo similar al usado por el NIST durante las campañas TAC/DUC. Este protocolo involucra el uso de resúmenes modelo o de referencia (escritos por personas) y el paquete Rouge, un sistema de evaluación de resúmenes que se basa en la co-ocurrencia de n-gramas entre resúmenes candidatos (los que se quiere evaluar) y resúmenes modelo. Rouge mide los máximos, los mínimos y el valor medio (reportado en este artículo) de la intersección de los n-gramas en los resúmenes candidatos y de referencia. Las campañas de evaluación del NIST han adoptado este test para medir la relevancia de los resúmenes. Para ser consistentes con la metodología del NIST, adoptamos el mismo protocolo en la evaluación de los resúmenes producidos por nuestro sistema. Sin embargo, hay otras pruebas estadísticas (como el test de Wilcoxon (Wilcoxon, 1945)) que podrían ser utilizadas para la evaluación. Deben llevarse

a cabo estudios más profundos en este sentido. Con el objetivo de guardar las mismas condiciones del protocolo NIST, se realizaron dos series de evaluaciones: la primera usando los resúmenes candidatos completos y la segunda usando los candidatos truncados a 100 palabras. En nuestro caso, analizamos bigramas (ROUGE-2) y bigramas separados por hasta 4 palabras (ROUGE-SU4). Para poder emplear ROUGE con textos en castellano, utilizamos un lematizador y una lista de *stopwords* en esta lengua.

Evaluamos el nuevo algoritmo lingüístico-estadístico utilizando diferentes combinaciones de los sistemas que lo forman (CORTEX, YATE, ENERTEX y DISICOSUM), así como cada uno de ellos individualmente para comparar sus resultados. Como comentamos en la sección 3, la mejor combinación fue la de CORTEX, CORTEX+YATE, ENERTEX y DISICOSUM.

Asimismo, realizamos resúmenes a modo de *baseline*. La primera (*Baseline*1) contiene oraciones del texto original seleccionadas de manera aleatoria. La segunda (*Baseline*2) incluye las primeras oraciones del texto original. El número de oraciones que incluyen estos resúmenes es el mismo que en todos los resúmenes candidatos.

Para fines de comparación se evaluaron resúmenes obtenidos con otros sistemas de resumen automático disponibles[9]: Microsoft Word, Pertinence[10], Swesum[11] y Open Text Summarizer (OTS)[12]. En principio la unidad lingüística para medir el tamaño de los resúmenes fue la oración (de nuevo, el número de oraciones que incluyen estos resúmenes es el mismo que el del resto de resuḿenes candidatos). Sin embargo los fragmentos segmentados automáticamente por estos sistemas no se corresponden necesariamente con una oración. Además, al no tener acceso a ciertos parámetros de estos sistemas, es imposible obtener exactamente la misma segmentación. Por ejemplo, algunos sistemas deciden arbitrariamente que los dos puntos son un separador de oraciones. Las abreviaturas o cifras decimales pueden causar separaciones que producirán oraciones incompletas o, en el caso contrario, oraciones unidas.

Finalmente, también incluimos como resúmenes candidatos resúmenes por extracción realizados por tres médicos que colaboraron en el experimento. Estos resúmenes se incluyeron en la evaluación para observar en qué medida los resúmenes producidos por nuestro algoritmo y por los demás sistemas se asemejan a resúmenes producidos por humanos especialistas del dominio. Los resúmenes de los médicos no tienen un número concreto de oraciones, como los demás candidatos, sino que este varía entre 1 y 5 oraciones del texto original (sin importar la sección), ya que estas fueron las instrucciones que se les dieron a los médicos.

Los resúmenes modelo para la evaluación con ROUGE son los *abstracts* de los propios autores de los artículos. Como ya hemos comentado, en este trabajo se resumen apartados de artículos médicos de la revista *Medicina Clínica*. Esta revista solicita a los autores que realicen sus resúmenes siguiendo la misma estructura del artículo (ya postulada por Swales (1990)) y, por tanto, estos también se dividen en los cuatro apartados antes mencionados. El siguiente fragmento refleja un *abstract* realizado por un médico:

Evaluación de las vías de acceso venoso innecesarias en un servicio de urgencias

FUNDAMENTO. Determinar la prevalencia de catéteres venosos periféricos innecesarios en un servicio de urgencias.

PACIENTES Y MÉTODOS. Estudio retrospectivo sobre una muestra de 1,113 pacientes del total (24,673) que acudieron a urgencias. Se revisaron las vías venosas canalizadas y si fueron o no utilizadas.

RESULTADOS. Se practicó acceso venoso en 202 pacientes (18.15 %). En 84 (41.6 %) no fue utilizado. El coste económico de las vías innecesarias ascendió a 10,264.33 euros.

DISCUSIÓN. Los accesos venosos periféricos frecuentemente son utilizados innecesariamente, generando un coste de "mala calidad".

Así, si necesitamos evaluar el resumen de un apartado de *Resultados*, por ejemplo, seleccionamos como resumen modelo únicamente el fragmento referido a este apartado en el *abstract* del autor. Al interpretar los resultados debe tenerse en cuenta que los resúmenes candidatos son resúmenes por extracción (incluidos los resúmenes de los médicos) mientras que los resúmenes modelo son resúmenes por abstracción.

Una vez obtenidos todos los resúmenes candidatos y modelo (tanto con oraciones completas como con oraciones truncadas), aplicamos ROUGE. Los resultados se muestran en el cuadro 4. Observamos que los resúmenes del algoritmo

[9] Otros sistemas disponibles, como Copernic Summarizer por ejemplo, no procesan textos en castellano y fueron descartados de este estudio.

[10] `http://www.pertinence.net/index.html`

[11] `http://swesum.nada.kth.se/index-eng.html`

[12] `http://libots.sourceforge.net/`

lingüístico-estadístico diseñado obtienen los mejores resultados (en negrita y subrayados), tanto con ROUGE-2 con oraciones completas (0.3307) y con oraciones truncadas (0.3163), como con ROUGE-SU4 con oraciones completas (0.3457) y con oraciones truncadas (0.3288). Así, el algoritmo obtiene mayores puntuaciones que cualquiera de los sistemas que incluye por separado, ya sean lingüísticos o estadísticos. Esto es una muestra clara de que el algoritmo se ha realizado de manera adecuada y de que la combinación de técnicas estadísticas y lingüísticas favorece los resultados de los sistemas de resumen automático. Es de destacar que los sistemas que obtienen los siguientes mejores resultados (en negrita) son CORTEX con ROUGE-2 con oraciones completas (0.3193) y truncadas (0.2927) y con ROUGE-SU4 con oraciones completas (0.3386), y la combinación de CORTEX+YATE con ROUGE-SU4 con oraciones truncadas (0.3152). Se observa, pues, que la combinación del extractor de términos YATE mejora los resultados de CORTEX por separado al evaluar las oraciones truncadas, lo cual es una evidencia más de las ventajas de la combinación de sistemas estadísticos y lingüísticos. El sistema lingüístico DISICOSUM está algo por debajo del sistema estadístico CORTEX, pero en cambio por encima del sistema estadístico ENERTEX. De todas maneras, el algoritmo propuesto así como cualquiera de los cuatro sistemas que lo componen (ya sean estadísticos o lingüísticos) obtienen resultados mejores tanto de las dos *baselines* y todos los otros sistemas de resumen automático evaluados (Microsoft Word, Pertinence, Swesum y OTS).

Es muy destacable el hecho de que los resúmenes de los médicos obtienen una puntuación muy similar a la de nuestro algoritmo, siendo incluso menor en ocasiones. Esto quiere decir que las oraciones seleccionadas por nuestro algoritmo son prácticamente las mismas que las seleccionadas por los especialistas del dominio. Ha de tenerse en cuenta, de todas maneras, que los resúmenes de los médicos son resúmenes por extracción, por lo que es imposible que su contenido (y por tanto sus n-gramas, que es lo que compara ROUGE) coincida totalmente con el de los *abstracts* de los autores.

Las figuras 7 y 8 reflejan de forma gráfica (medidas ROUGE-2 contra ROUGE-SU4) los resultados numéricos del cuadro 4.

Finalmente, a modo de ejemplo, mostramos en el cuadro 5 un resumen (por extracción) producido por el algoritmo lingüístico-estadístico, el resumen del autor del artículo (por abstracción) y el resumen de uno de los tres médicos que colaboraron en el experimento (por extracción). Como puede observarse, los resúmenes tienen diferentes longitudes. El resumen producido por el algoritmo incluye dos oraciones, ya que se trata del resumen del apartado de *Discusión* de un artículo médico de nuestro corpus. El resumen del autor, en cambio, solo contiene una oración y el resumen de uno de los médicos, cuatro. Es de destacar que las dos oraciones seleccionadas por el algoritmo (1a y 2a) coinciden con dos de las oraciones incluidas en el resumen del médico (3c y 4c). También es relevante el hecho de que, aunque el autor solo ofrezca una oración a modo de resumen (1b), esta refleje una fusión de las ideas expresadas en las dos oraciones que incluye el resumen del algoritmo (1a y 2a).

	ROUGE-2		ROUGE-SU4	
Sistema de resumen	**Frases**	100 **Palabras**	**Frases**	100 **Palabras**
Híbrido	**0.3307**	**0.3163**	**0.3457**	**0.3288**
CORTEX	**0.3193**	**0.2927**	**0.3386**	0.3105
CORTEX+YATE	0.3038	0.2913	0.3281	**0.3152**
ENERTEX	0.2552	0.2314	0.2830	0.2589
DISICOSUM	0.2851	0.2750	0.3053	0.2945
Baseline_1	0.1454	0.1428	0.1835	0.1808
Baseline_2	0.1931	0.1861	0.2333	0.2260
Word	0.1873	0.1857	0.2273	0.2245
Pertinence	0.1768	0.1606	0.2266	0.2095
Swesum	0.2026	0.1939	0.2382	0.2289
OTS	0.2337	0.2176	0.2675	0.2533
Médico 1	0.3329	0.3030	0.3414	0.3130
Médico 2	0.3230	0.2993	0.3374	0.3130
Médico 3	0.3099	0.2721	0.3201	0.2898

Cuadro 4: Comparación de los valores medios de ROUGE entre los diferentes resúmenes.

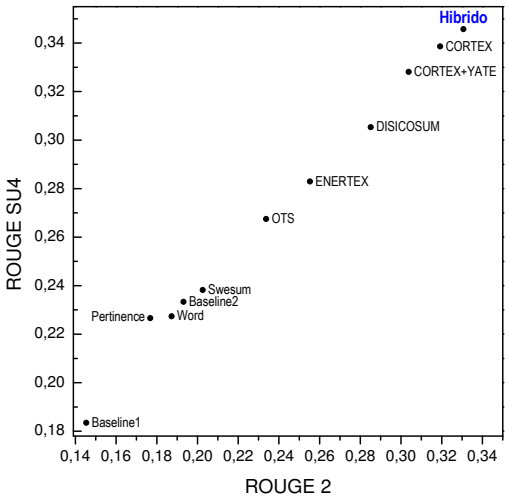

Figura 7: Resultados de la evaluación con ROUGE (resúmenes con oraciones completas).

5. *Conclusiones*

El método de resumen automático propuesto en este trabajo da como resultado un sistema híbrido de resumen que integra de manera equi-

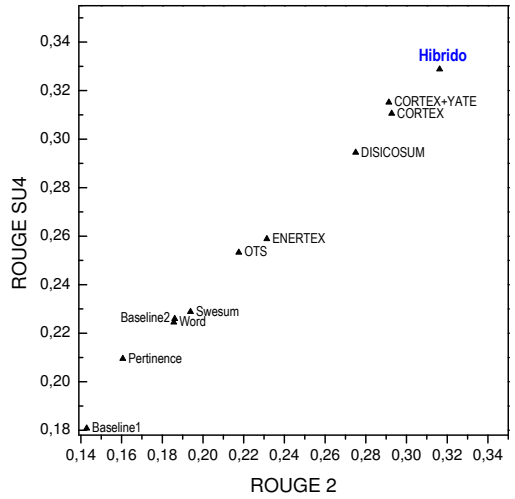

Figura 8: Resultados de la evaluación con ROUGE (resúmenes truncados a 100 palabras).

Oración	Resumen producido por el algoritmo
1a	El estudio genotípico de la región promotora del gen UGT-1 facilita la identificación de individuos con variantes alélicas asociadas a hiperbilirrubinemias no conjugadas y confirma el diagnóstico de síndrome de Gilbert que en la actualidad se realiza por exclusión de otros procesos patológicos.
2a	Estos resultados nos permiten plantear la idoneidad del escrutinio molecular para el síndrome de Gilbert como una prueba adicional, tanto en el protocolo diagnóstico de hiperbilirrubinemias no conjugadas, crónicas, de intensidad moderada, en ausencia de enfermedad hepática en el adulto, como en recién nacidos con ictericia neonatal prolongada.
Oración	Resumen del autor
1b	El porcentaje de alelos mutados detectados en la población analizada, similar al hallado en otras poblaciones caucásicas, plantea la inclusión del análisis genotípico del gen UGT-1 en el protocolo diagnóstico de hiperbilirrubinemias no conjugadas, crónicas, de intensidad moderada, en ausencia de hemólisis y de enfermedad hepática.
Oración	Resumen del médico 1
1c	Ante un paciente con episodios de hiperbilirrubinemia indirecta, habitualmente se efectúa el diagnóstico de síndrome de Gilbert después de excluir la existencia de una hepatopatía o de un síndrome hemolítico.
2c	Estas pruebas, además de ser incómodas para el paciente, son poco específicas.
3c	El estudio genotípico de la región promotora del gen UGT-1 facilita la identificación de individuos con variantes alélicas asociadas a hiperbilirrubinemias no conjugadas y confirma el diagnóstico de síndrome de Gilbert que en la actualidad se realiza por exclusión de otros procesos patológicos.
4c	Estos resultados nos permiten plantear la idoneidad del escrutinio molecular para el síndrome de Gilbert como una prueba adicional, tanto en el protocolo diagnóstico de hiperbilirrubinemias no conjugadas, crónicas, de intensidad moderada, en ausencia de enfermedad hepática en el adulto, como en recién nacidos con ictericia neonatal prolongada.

Cuadro 5: Ejemplo de resúmenes producidos por el algoritmo, el autor y un médico.

librada diversos sistemas lingüísticos y estadísticos ya existentes. En este trabajo hemos mostrado que, a diferencia de los resultados obtenidos en otras áreas del procesamiento del lenguaje, la combinación de métodos numéricos y simbólicos contribuye a mejorar los resultados de la tarea propuesta. Más concretamente, los resúmenes automáticos producidos por métodos estadísticos (CORTEX y ENERTEX) son similares a los producidos por métodos lingüísticos (DISICOSUM) y que un extractor de términos especializado como YATE contribuye a mejorar ligeramente la salida de un resumidor estadístico, al menos en los resúmenes truncados a 100 palabras. Sin embargo, la principal conclusión a la que llegamos es que la combinación de técnicas estadísticas y lingüísticas en el problema del resumen automático obtiene excelentes resultados, mejores que cualquiera de los sistemas evaluados por separado. Nuestro algoritmo lingüístico-estadístico consigue superar a los otros sistemas de resumen y a las *baselines* diseñadas. Además, los resúmenes que ofrece son muy similares a los realizados por médicos especialistas, lo cual es una evidencia clara de la calidad de los mismos.

En el futuro prevemos realizar experimentos con textos de otros dominios (genómica) y con otras lenguas (francés y catalán). Además, realizaremos un experimento en el que se use el sistema YATE para detectar automáticamente los términos del título del texto y posteriormente seleccionar para el resumen las oraciones que contengan términos relacionados semánticamente con ellos.

Agradecimientos

Parte de este trabajo ha sido financiado mediante una ayuda de movilidad posdoctoral otorgada por el Ministerio de Ciencia e Innovación de España (Programa Nacional de Movilidad de Recursos Humanos de Investigación; Plan Nacional de Investigación Científica, Desarrollo e Innovación 2008-2011) a Iria da Cunha.

References

Afantenos, S.D., V. Karkaletsis, and P. Stamatopoulos. 2005. Summarization of medical documents: A survey. *Artificial Intelligence in Medicine*, 2(33):157–177.

Alonso, L. and M. Fuentes. 2003. Integrating cohesion and coherence for Automatic Summarization. In *EACL'03 Student Session*, pages 1–8. ACL, Budapest.

Aretoulaki, M. 1996. *COSY-MATS: A Hybrid Connectionist-Symbolic Approach To The Pragmatic Analysis Of Texts For Their Automatic Smmarization.* Ph.D. thesis, University of Manchester, Institute of Science and Technology, Manchester.

Barzilay, R. and M. Elhadad. 1997. Using lexical chains for text summarization. In *Intelli-*

gent Scalable Text Summarization Workshop, ACL, Madrid, Spain.

Barzilay, R. and M. Lapata. 2005. Modelling local coherence: An entity-based approach. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 318–325.

Berger, A. and V. Mittal. 2000. A system for summarizing Web Pages. In *23rd Annual Conference on Research and Development in Information Retrieval*, pages 144–151. Atenas.

Brandow, R., K. Mitze, and L. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Inf. Proc. and Management*, 31:675–685.

da Cunha, I. 2008. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español.* Ph.D. thesis, IULA, Barcelona, España.

da Cunha, I., S. Fernández, P. Velázquez, J. Vivaldi, E. SanJuan, and J.M. Torres-Moreno. 2007. A new hybrid summarizer based on Vector Space Model, Statistical Physics and Linguistics. In *MICAI 2007: Advances in Artificial Intelligence. Lecture Notes in Computer Science*, pages 872–882. Gelbukh, A. and Kuri Morales, A. F. (eds.), Berlín: Springer.

da Cunha, I. and L. Wanner. 2005. Towards the Automatic Summarization of Medical Articles in Spanish: Integration of textual, lexical, discursive and syntactic criteria. In *Crossing Barriers in Text Summarization Research (RANLP-2005)*, pages 46–51. Saggion, H. and Minel, J. (eds.), Borovets (Bulgaria): INCOMA Ltd.

da Cunha, I., L. Wanner, and M. T. Cabré. 2007. Summarization of specialized discourse: The case of medical articles in Spanish. *Terminology*, 13(2):249–286.

Edmundson, H. P. 1969. New Methods in Automatic Extraction. *Journal of the Association for Computing Machinery*, 16:264–285.

Fernández, S. 2009. *Applications exploratoires des modèles de spins au Traitement Automatique de la Langue.* Ph.D. thesis, Université Henri Poincaré Nancy 2, France.

Fernández, S., E. SanJuan, and J. M. Torres-Moreno. 2007. Énergie textuelle de mémoires associatives. In *Traitement Automatique des Langues Naturelles*, pages 25–34. Toulouse, France.

Fernández, S., E. SanJuan, and J. M. Torres-Moreno. 2008. Enertex : un système basé sur l'énergie textuelle. In *Traitement Automatique des Langues Naturelles*, pages 99–108. Avignon, France.

Fuentes, M. 2008. *A Flexible Multitask Summarizer for Documents from Diferent Media, Domain, and Language.* Ph.D. thesis, UPC, Barcelona.

Goldstein, J., J. Carbonell, M. Kantrowitz, and V. Mittal. 1999. Summarizing text documents: sentence selection and evaluation metrics. In *22nd Int. ACM SIGIR Research and development in information retrieval*, pages 121–128. Berkeley.

Hertz, J., A. Krogh, and G. Palmer. 1991. *Introduction to the theorie of Neural Computation.* Redwood City, CA : Addison-Wesley.

Hopfield, J. 1982. Neural networks and physical systems with emergent collective computational abilities. *National Academy of Sciences*, 9:2554–2558.

Hovy, E. and Lin, C.Y. 1999. Automated text summarisation in SUMMARIST. In *Advances in automatic text summarisation (Ed. I. Mani and M. Maybury)*, pages 81–94. Cambridge, MA: MIT Press.

Kupiec, J., J. O. Pedersen, and F. Chen. 1995. A trainable document summarizer. In *SIGIR-95*, pages 68–73. New York.

Lacatusu, V.F., P. Parker, and S.M. Harabagiu. 2003. Lite-GISTexter: Generating short summaries with minimal resources. In *Proceedings of the DUC 2003*, pages 122–128.

Leskovec, J., N. Milic-Frayling, and M. Grobelnik. 2005. Impact of linguistic analysis on the semantic graph coverage and learning of document extracts. In *Proceedings of the AAAI 2005*, volume 3, pages 1069–1074.

Lin, C. and E. Hovy. 1997. Identifying Topics by Position. In *ACL Applied Natural Language Processing Conference*, pages 283–290. Washington.

Lin, C.Y. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26.

Luhn, H. P. 1959. The automatic creation of Literature abstracts. *IBM Journal of research and development*, 2(2).

Mani, I. 2001. *Automatic summarization.* Amsterdam: John Benjamins Publishing.

Mani, I. and M.T. Maybury. 1999. *Advances in automatic text summarization*. Cambridge, MA: MIT Press.

Mann, W. C. and S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marcu, D. 1998. *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. thesis, Dep. of Computer Science, University of Toronto.

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing Summarization*. Institute of Technology, Massachusetts.

Marcu, D. and A. Echihabi. 2002. An unsupervised approach to recognising discourse relations. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 368–375.

Mel'cuk, I. 1988. *Dependency Syntax: Theory and Practice*. Albany: State University Press of New York.

Mel'cuk, I. 2001. *Communicative Organization in Natural Language. The semantic-communicative structure of sentences*. John Benjamins, Amsterdam.

Nomoto, T. and Nitta, Y. 1994. A Grammatico-Statistical Approach to Discourse Partitioning. In *15th Int. Conf. on Comp. Linguistics*, pages 1145–1150. Kyoto.

Ono, K., K. Sumita, and S. Miike. 1994. Abstract generation based on rhetorical structure extraction. In *15th Int. Conf. on Comp. Linguistics*, pages 344–348. Kyoto.

Pardo, T. and Nunes, M. 2008. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing*, 15(2):43–64.

Pardo, T. and Nunes, M. and Rino, M. 2004. DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. In *XVII Brazilian Symposium on Artificial Intelligence - SBIA2004*, pages 224–234. São Luís.

Polanyi, L., C. Chris, M. van den Berg, G.L. Thione, and D. Ahn. 2004. A rule-based approach to discourse parsing. In *Proceedings of the fifth SIGdial workshop on discourse and dialogue*, pages 108–117.

Radev, D., T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. 2004. MEAD - a platform for multilingual summarisation. In *Proceedings of LREC 2004*.

Radev, D.R., H. Jing, and M. Budzikowska. 2000. Centroid-based summarisation of multiple documents: Sentence extraction, utility-based evaluation, and user Studies. In *Proceedings of the ANLP/NAACL-00*, pages 21–30.

Salton, G. and M. McGill. 1983. *Introduction to modern information retrieval*. Computer Science Series McGraw Hill Publishing Company.

Silber, H. Gregory and Kathleen F. McCoy. 2000. Efficient text summarization using lexical chains. In *Intelligent User Interfaces*, pages 252–255.

Spärck Jones, K. 2007. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481.

Sumita, K., K. Ono, T. Chino, T. Ukita, and S. Amano. 1992. A discourse structure analyzer for Japonese text. In *International Conference on Fifth Generation Computer Systems*, pages 1133–1140. Tokyo, Japan.

Swales, J. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge.

Teufel, S. and M. Moens. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4):409–445.

Thione, G.L., M. van den Berg, L. Polanyi, and C. Culy. 2004. Hybrid text summarisation: Combining external relevance measures with structural análisis. In *Proceedings of the ACL-04*, pages 51–55.

Torres-Moreno, J. M., P. Velázquez-Morales, and J. G. Meunier. 2001. Cortex : un algorithme pour la condensation automatique des textes. In *ARCo 2001*, pages 65–75. Lyon, France.

Torres-Moreno, J. M., P. Velázquez-Morales, and J. G. Meunier. 2002. Condensés de textes par des méthodes numériques. In *JADT 2002*, pages 723–734. St. Malo, France.

Vanderwende, L., M. Banko, and A. Menezes. 2004. Event-centric summary generation. In *Proceedings of the DUC 2004*, pages 76–81.

Vivaldi, J. 2001. *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona.

Vivaldi, J. and H. Rodríguez. 2001. Improving term extraction by combining different techniques. *Terminology*, 7(1):31–47.

Vivaldi, J. and H. Rodríguez. 2002. Medical term extraction using the EWN ontology. In *Terminology and Knowledge Engineering*, pages 137–142. Nancy.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics*, 1:80–83.

Hacia una semántica computacional de las anáforas demostrativas

Iker Zulaica-Hernández
The Ohio State University
248 Hagerty Hall
Columbus, OH USA
ikerzulaica@gmail.com

Javier Gutiérrez-Rexach
The Ohio State University
248 Hagerty Hall
Columbus, OH USA
gutierrez-rexach.1@osu.edu

Resumen

Los demostrativos exhiben una naturaleza dual en lo que respecta a su comportamiento discursivo. Por un lado, se comportan como elementos de referencia directa normalmente acompañados de un señalamiento en su uso canónico. Por otro lado, los hablantes utilizan los demostrativos para hacer referencia a una gran variedad de entidades que han sido previamente mencionadas en el discurso (anáfora del discurso), tales como eventos, proposiciones o cualquier otro tipo de entidad abstracta que carezca de anclaje espacio-temporal. En este artículo proponemos una caracterización de los determinantes y pronombres demostrativos del español como cuantificadores generalizados, que nos permitirá explicar su heterogénea naturaleza referencial así como las principales diferencias entre los diferentes elementos.

1. Introducción

El fenómeno de la anáfora del discurso constituye uno de los mayores desafíos a los que se enfrenta la disciplina denominada Procesamiento del Lenguaje Natural en la actualidad, tanto en lo que respecta a su grado de complejidad intrínseca como por la viabilidad de su implementación computacional. Es cierto que el número de estudios generales sobre la anáfora en el discurso se ha incrementado recientemente de forma considerable en las disciplinas de la Lingüística Teórica y Computacional durante las dos últimas décadas (véase entre otros Asher 1993, Grosz et al 1995, Webber 1979, 1988, 1991, Linde 1979, Passoneau 1989, Byron 2004, Poesio 2000, Poesio & Modjeska 2005). Sin embargo, actualmente todavía nos encontramos lejos de haber conseguido una teoría que explique de manera completamente satisfactoria los mecanismos subyacentes que gobiernan las dependencias textuales de largo alcance y de una caracterizción completamente satisfactoria de este fenómeno lingüístico.

Todos aquellos fenómenos englobados bajo la etiqueta de anáfora discursiva pueden asimismo analizarse como subcomponentes de teorías lingüísticas más generales sobre la coherencia y cohesión discursiva. Como es de sobra conocido, los hablantes de una lengua tienen a su disposición una gran variedad de mecanismos lingüísticos que les sirven de herramientas para integrar fragmentos de discurso en unidades mayores totalmente coherentes y cohesionadas. Entre dichos mecanismos podemos incluir los pronombres, los marcadores o partículas del discurso, las pautas de concordancia morfosintáctica, y otros tipos de mecanismos intra e ínter oracionales. En el presente trabajo, nos enfocaremos en varias propiedades de los pronombres demostrativos del español peninsular con el objetivo de proporcionar la caracterización más adecuada posible para estos elementos, especialmente en todo lo que respecta a su participación en procesos de anáfora del discurso. Sin embargo, los demostrativos no son los únicos elementos que desempeñan un papel importante en la anáfora del discurso, y esto es especialmente cierto cuando nos referimos a las fuentes de coherencia y cohesión discursiva globales. Otros mecanismos lingüísticos habituales, como el uso de pronombre fuertes y débiles (clíticos); la elipsis verbal, el vaciado o *gapping*, etc., también desempeñan un papel crucial en varias subclases de anáfora. Sin embargo, creemos que algunas de las ideas que presentamos en este trabajo podrían aplicarse a los fenómenos lingüísticos mencionados.

La lengua española posee un sistema demostrativo de tres elementos, *este*, *ese* y *aquel*, que poseen varición morfológica en género (masculino/femenino) y en número (singular/ plural). Estos tres elementos pueden funcionar como determinantes acompañando a un sustantivo (por ejemplo, *esta casa*, *este hombre*, *aquel planeta lejano*, etc.) y como pronombres. Existe asimismo un sistema de tres pronombres demostrativos llamados 'neutros' por sus propiedades referenciales pues el uso principal de estos pronombres es el de referir a objetos abstractos y entidades que no pertenecen a las categorías gramaticales Masculino

y Femenino. Los pronombres neutros son *esto*, *eso* y *aquello*. En casos de *deixis espacial*, los demostrativos del español se distinguen por el grado de proximidad del elemento demostrado con respecto al hablante. De este modo, el demostrativo *este* indica que el elemento demostrado es próximo con respecto al hablante, *aquel* indicaría lejanía del objeto con respecto al hablante y, finalmente, *ese* indicaría inespecificidad con respecto a la proximidad del objeto en relación con el hablante y el oyente. Existen varios usos derivados de este uso espacial deíctico básico. Así, podemos hablar por ejemplo de la *deixis temporal* en la que la dimensión espacial se reinterpreta metafóricamente como distancia/lejanía en el eje temporal (por ej. *aquellos maravillosos años*…), del uso de ciertos demostrativos como partículas del discurso (Zulaica-Hernández 2009), de las lecturas inespecíficas de los demostrativos (por ej. *aquel que saque un cinco en el examen, aprobará*), por sólo nombrar algunos de estos usos (Gutiérrez-Rexach 2002).

Sin embargo, en su empleo discursivo más habitual, los hablantes del español utilizan los demostrativos para referirse a una gran variedad de objetos de discurso tanto anafórica como catafóricamente. Los ejemplos (1)-(6) constituyen ejemplos reales documentados de habla coloquial, obtenidos del corpus CREA del español y nos sirven para ilustrar este punto.

(1) La alianza tiene mayoría. Y ***esto*** lo sabe todo el pueblo argentino.

En (1), el pronombre demostrativo *esto* en la segunda oración se utiliza anafóricamente para referir al hecho mencionado en la primera oración. Que la entidad denotada por la primera oración es un *hecho* (información factual) parece confirmarse por la presencia del típico predicado factivo *saber* ('to know') que acompaña al demostrativo y que comúnmente fuerza lecturas factivas. Un ejemplo similar es el que presentamos en (2), aunque en este caso el referente del pronombre demostrativo *eso* en la segunda oración parece ser un objeto dell tipo de los *eventos* (acción, evento, situación, logro, etc.) (Vendler 1957), que se ha descrito en la oración previa. En este caso, la presencia del predicado típico de eventos *suceder* fuerza una lectura eventiva de la entidad referida.

(2) El ejército fue instrumento de traición, de felonía y de hostilidad al pueblo. Pero cuando ***eso*** sucedió...

Existen dos modos, no mutuamente excluyentes necesariamente, de abordar una cuestión básica sobre la ontología y taxonomía de las expresiones referenciales en el fenómeno de la anáfora del discurso, a saber: la idoneidad de su consideración semántica o puramente morfosintáctica. Si aplicamos un criterio estrictamente formal (exclusivamente basado en la forma de las expresiones lingüísticas), no tiene ningún sentido caracterizar una expresión determinada como denotadora de un evento, de un hecho, de un individuo, etc., pues estos son conceptos estrictamente semánticos. Bajo este criterio, sólo es relevante la categoría gramatical a la que pertenece determinado referente (oración, sintagma nominal, sintagma verbal, etc.). Este criterio morfosintáctico y puramente textual es el que da lugar a conceptos formales como el de la correferencia entre antecedente y anáfora y tiene como principal ventaja el que sólo aquellos rasgos que tienen una correspondencia gramatical sean considerados (por ej. el género de un sustantivo). Su principal desventaja consiste en la imposibilidad de establecer una correferencia estricta entre determinadas expresiones. Así, por ejemplo, el pronombre demostrativo neutro *esto* en (1) parece tener como antecedente textual la oración anterior en su totalidad; sin embargo una oración y un pronombre no comparten rasgos morfosintácticos comunes que faciliten o expliquen el vínculo anafórico formalmente. Por otro lado tenemos el criterio semántico por el que se considera el tipo de entidad semántica que denota una expresión gramatical. Así, por ejemplo, decimos que un sustantivo denota un individuo, un adjetivo una propiedad, etc. (Kenan y Faltz 1985). La ventaja de este criterio radica en que permite ofrecer una explicación para los vínculos referenciales que son problemáticos desde el punto de vista formal estricto como el presentado en (1). Su principal desventaja la constituye el hecho de que algunas entidades semánticas son puramente intuitivas (establecer que un determinado sustantivo denota un *evento* no se basa en cuestiones gramaticales sino en otras puramente conceptuales.) Nosotros creemos que una combinación de ambos criterios es conveniente para el tratamiento de fenómenos lingüísticos como la anáfora del discurso. Es por ello que a lo largo del presente trabajo hablaremos de referentes de tipo *evento*, *hecho*, *situación*, etc. (Asher 1993) para los demostrativos, así como también utilizaremos el término antecedente en su concepción más amplia, es decir, no sólo como expresión lingüística identificable a través de rasgos morfológicos. Definir el concepto semántico de *evento* o *hecho* es una tarea compleja que está fuera del alcance de este trabajo y en la que

es necesario recurrir a cuestiones de semántica léxica, aspecto léxico (*aktionsart*), etc. Simplificando, podríamos definir un evento en general como un acontecimiento que ocurre en un lugar y tiempo determinados. Generalmente los verbos denotan eventos y gracias a la morfología verbal podemos situar tales acontecimientos en el pasado, presente o futuro con respecto al momento de habla. Existe toda una ontología eventiva que distingue entre acciones, logros, situaciones, estados, etc. basada en la micro estructura semántica del evento en cuestión y para cuya definición se recurre a cuestiones aspectuales como la telicidad o atelicidad, entre otras (véase Vendler 1957 y seguidores). Asimismo, se considera que ciertos sustantivos denotan eventos sobre todo cuando estos están acompañados por predicados de evento como *suceder*, *ocurrir*, etc. En cuanto a los hechos, podríamos definirlos como proposiciones que hacen referencia a objetos y acontecimientos del mundo empírico (por ej. *Juan tiene la gripe*) y a los que se puede asignar un valor de verdad. Al igual que con los eventos, ciertos verbos sirven como guías que fuerzan lecturas eventivas y nos permiten identificar con más precisión estos elementos (por ej. *saber*)

No sólo los pronombres demostrativos sirven como anáforas discursivas. Los determinantes demostrativos del tipo 'este N' desempeñan la misma función. Este punto se ilustra en (3), en el que el demostrativo distal o de lejanía *aquel* refiere anafóricamente al evento descrito en la oración anterior.

(3) El doce de octubre Cristóbal Colón ponía pie en América por primera vez. Dentro de cinco años se cumplen los quinientos años de ***aquel*** acontecimiento.

Por último, la oración de (4) sirve como ilustración de un caso típico de pronombre demostrativo en uso catafórico. A primera vista, el demostrativo *eso* parece tener una función presentacional en el discurso y su referente se encuentra inmediatamente a continuación del demostrativo. En este caso, el referente del demostrativo es la entidad denotada por la oración de infinitivo *ser un seductor*.

(4) ¿Sale muy caro ***eso*** de ser un seductor?

Pero, los antecedentes de los demostrativos no se restringen a entidades oracionales (oraciones completas, oraciones subordinadas, etc.) con las que no se produce correferencia morfosintáctica estricta. Al igual que los pronombres o las expresiones definidas, los demostrativos pueden tener un sintagma nominal como antecedente con el que comparten rasgos como género y número. En (5), el sintagma nominal *ese director* [GEN: masc., NUM: sing.] tiene como antecedente al sintagma nominal *Stanley Kubrick* [GEN: masc., NUM: sing.] de la frase adyacente. Dado que ambos elementos comparten los rasgos de género (GEN.) y número (NUM.) se produce correferencia entre ellos. El mismo subíndice en ambos elementos así lo indica.

(5) A Ana le gusta [Stanley Kubrick]$_j$ pero a mí no me gusta nada [***ese*** *director*]$_j$.

2. El problema de la referencia directa

En vista de los ejemplos anteriores, parece que un paso lógico y necesario para caracterizar la semántica de los demostrativos consiste en comenzar por establecer sus propiedades referenciales. El tratamiento de los demostrativos como expresiones de referencia directa[1] postula en esencia que estos elementos constituyen designadores rígidos[2]. Por ejemplo, Kaplan (1989) caracteriza los demostrativos como expresiones incompletas las cuales sólo se completan mediante un gesto demostrativo (señalar con el dedo índice, un movimiento de cabeza, etc.) De este modo, sólo conseguiremos una expresión demostrativa completa d[δ] cuando se ejecute un gesto de señalamiento [δ] de manera explícita al mismo tiempo que se emite el enunciado (d).

Las teorías de la referencia directa dan cuenta sin problemas de los usos canónicos de los demostrativos, es decir, aquellos casos en los que los demostrativos se utilizan como elementos deícticos para expresar la distancia física del objeto referido con respecto al hablante y al oyente y acompañado de un señalamiento explícito [δ] que servirá para fijar el referente pretendido en un contexto de habla particular. En la mayoría de las ocasiones, se trata de casos de deixis espacial en los que el objeto demostrado es una entidad física, tangible y tridimensional que se percibe por los sentidos en el entorno en el que se desenvuelven los

[1] Según el tratamiento Russelliano de la referencia directa (Russell 1905) el contenido de un nombre propio como *Juan* o el contenido de un elemento deíctico como *él* lo constituye, simple y llanamente, su referente. Dicho de otro modo, un término singular (nombre, pronombre, descripción definida, etc.) será *directamente referencial* si y sólo si su contenido fija de manera directa su extensión (su referente).

[2] Un *designador rígido* es un elemento que designa el mismo objeto en todos los mundos posibles en que tal objeto existe y nunca designa otro objeto distinto.

participantes en el acto de habla. Este caso se ilustra en (6)[3].

(6) [Juan y Pedro están hablando de sus coches favoritos. En un punto de la conversación Pedro señala en la dirección de un objeto en movimiento mientras dice:]

¡Mira, Juan! ***ese*** es mi coche favorito.

Sin embargo, existen usos frecuentes de demostrativos que no siguen esta pauta contextual. Entre estos usos podemos distinguir la deixis textual y la deixis discursiva/anafórica en sus variadas manifestaciones, como pudimos comprobar en la sección precedente. Además, existen casos en los que no hay un *demonstratum* (por ejemplo, una entidad u objeto señalado) en el sentido canónico del término o, dicho de otro modo, no existe una entidad concreta particular a la que el demostrativo pueda 'anclar' su referencia. Como señala King (1999), abundan los casos en los que parecen estar ausentes tanto el acto de demostración en sí como la referencia clara del hablante. Este punto se ilustra en (7).

(7) ¡Lo espantosa que debe ser la vida de ***ese*** hombre que va a una oficina donde se aburre!

En (7), el referente del sintagma demostrativo *ese hombre* no se ha mencionado previamente en el discurso. Una posible interpretación para este ejemplo sería aquella en la que el hablante lo único que conoce es la existencia de un individuo no específico o indeterminado que desempeña un aburrido trabajo en una oficina cualquiera. El hablante podría haber leído esta información en un periódico, habérselo contado alguien, o ser parte del conocimiento general compartido por la comunidad de habla que presupone que los individuos (en general) que trabajan en oficinas (en general) desempeñan (en general) trabajos aburridos. Lo que está claro es que el hablante no necesita tener un individuo específico en mente (por ejemplom, Juan) para proferir una oración como la de (7) que incluye la expresión *ese hombre*. Además, en este caso el hablante no lleva a cabo ningún acto de señalamiento explícito simplemente porque no se está refiriendo a nadie que se pueda encontrar físicamente presente en la situación de habla. Cualesquiera teorías que conciban los demostrativos como elementos puramente referenciales no podrán proporcionar una explicación factible para casos como éstos. Otro tipo de construcción bastante común en español moderno y que desafía los presupuestos de la teoría de la referencia directa is aquella en la que un demostrativo complejo contiene un pronombre ligado que, a su vez, se comporta como una variable ligada por un cuantificador externo.

(8) Todo lingüista recuerda ***aquel*** día en que presenta su primer trabajo en un congreso.

La razón por la que el demostrativo *aquel día* en (8) no se puede considerar un elemento de referencia directa se debe a que el demostrativo mismo se comporta como una variable. La interpretación más natural de esta oración es la de una afirmación genérica sobre lingüistas no específicos y trabajos de investigación no específicos que se presentan un día y en un congreso también no específicos. Existen asimismo casos en los que el demostrativo toma alcance o ámbito estrecho con respecto al cuantificador universal distributivo *cada*, de ahí que el demostrativo dependa de dicho elemento cuantificacionalmente.

(9) Sólo se ascenderá a ***aquel*** empleado con más experiencia de cada departamento.

El hecho de que el referente del sintagma nominal demostrativo varíe con respecto al valor que toma el cuantificador distributivo *cada* explica por qué una posible continuación para (9) sería una en la que se afirmara que *en total se promocionará a 10 trabajadores*, es decir, una promoción para un único trabajador de cada departamento independientemente del número de departamentos que haya. En este ejemplo es evidente que no se produce ningún gesto de señalamiento. De hecho, pronunciar la expresión demostrativa compleja *aquel empleado con más experiencia de cada departamento* acompañándola de un gesto de demostración produciría un resultado anómalo o no apropiado ya que la demostración explícita cancelaría el significado distributivo del cuantificador al fijarel referente en un individuo en particular. Evidencia adicional a favor del tratamiento de los demostrativos como elementos cuantificacionales y, en consecuencia, en contra de la consideración de la referencia directa en sentido estricto proviene de las oraciones denominadas *Bach-Peters*. En este tipo de construcciones aparecen dos elementos cuantificados, cada uno de ellos conteniendo una expresión pronominal ligada por la otra. Tradicionalmente se ha considerado a las

[3] No todos los casos de uso canónico del demostrativo deben acompañarse necesariamente de un gesto de señalamiento o demostración explícita. Cuando el objeto referido es lo suficientemente prominente en el contexto de habla tal señalamiento no es necesario.

oraciones *Bach-Peters* como prueba de los procesos de absorción del cuantificador experimentados por dos expresiones cuantificacionales (May 1985, 1989). El ejemplo (10) constituye un caso de oración *Bach-Peters*.

(10) Todo piloto que lo disparó alcanzó al Mig que lo estaba persiguiendo.

El punto crítico en estos casos es que la pauta de cruce anafórico se obtiene incluso cuando se sustituye el artículo definido *el* o el cuantificador universal *todo* por un demostrativo. Esto lo podemos observar en el siguiente ejemplo.

(11) ***Aquel*** estudiante tuyo que lo preparó aprobó aquel examen que tanto temía.

Por último, existen también casos en los que el antecedente del demostrativo es un sintagma nominal que se deduce mediante algún tipo de mecanismo inferencial. Estos casos pertenecen al fenómeno de la anáfora asociativa o anáfora indirecta (Hawkins, 1978) o, simplemente, 'bridging' como se ilustra en (12). Lo que realmente diferencia a la anáfora asociativa de otros tipos de anáfora del discurso es el hecho de que el vínculo entre el antecedente y la anáfora no es un vínculo de identidad. El referente de la expresión *aquella camarera* únicamente se identifica apelando al conocimiento general del mundo que tienen los interlocutores y la asociación que se obtiene entre un restaurante y la camarera de ese restaurante. De acuerdo con la terminología del propio Hawkins, el antecedente constituiría el disparador o catalizador y la anáfora su elemento asociado.

(12) Ayer cenamos en un restaurante japonés. ***Aquella*** camarera fue muy atenta.

En vista de los ejemplos presentados en esta sección, podríamos concluir inicialmente que los demostrativos del español deberían caracterizarse como elementos cuantificacionales; al menos en aquellos casos en los que los demostrativos no se muestran como elementos de referencia directa. Por desgracia el panorama no es tan sencillo ya que existen casos en los que los demostrativos funcionan claramente como designadores rígidos en el discurso (véase de nuevo el ejemplo (6). De este modo, ni la teoría de la referencia directa ni la caracterización en términos puramente cuantificacionales pueden explicar por si solas la variedad de usos observados. Por lo tanto podríamos proponer una explicación alternativa homogénea o bien aceptar una caracterización híbrida que explique el complejo comportamiento que muestran los demostrativos en el discurso hablado y escrito en el español moderno. Dada la heterogénea naturaleza de los demostrativos, proponemos en la siguiente sección una caracterización presuposicional para los pronombres y determinantes demostrativos del español que explica su comportamiento dual como cuantificadores y como elementos de referencia directa.

3. Los demostrativos como cuantificadores generales incrementados

Los determinantes demostrativos se pueden caracterizar como funciones determinantes que, al igual que los cuantificadores, introducen condiciones 'dúplex' en una Estructura de Representación del Discurso (Kamp y Reyle 1993); de aquí en adelante: ERD. De acuerdo con esta propuesta, los determinantes demostrativos constituirían funciones de conjuntos a cuantificadores generalizados (demostrativos) de tipo semántico: ⟨⟨e, t ⟩, ⟨⟨e, t ⟩, t ⟩⟩. Considérese el siguiente ejemplo.

(13) ***Este*** perro ladra mucho.

Por ejemplo, nótese la representación de la expresión compleja *este perro* de (13) en términos de una ERD que mostramos en la figura 1. El demostrativo de proximidad *este* introduce una condición dúplex de tipo Φ DEM Ψ, en donde Φ y Ψ son ERD's y DEM representa la fuerza cuantificacional del determinante demostrativo. Este punto se muestra en la ERD lineal (14) y su correspondiente notación de cajas en la figura 1.

(14) [y | [| perro(y)] **DEM** [| ladra(y)]]

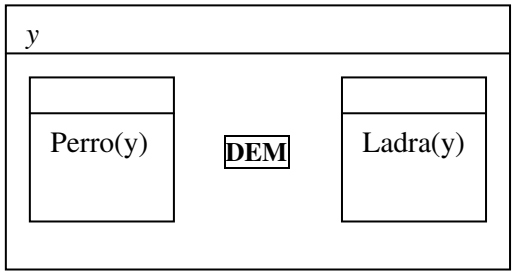

Figura 1: Condición dúplex forzada por el demostrativo

Sin embargo, la representación mostrada en (14) todavía no está completa pues todavía tenemos que introducir cierto material presuposicional con el objetivo de caracterizar plenamente el contenido de los demostrativos y poder establecer diferencias individuales entre los mismos. Con este propósito seguiremos algunas de las ideas que presentamos en la sección precedente junto con la propuesta de Zeevat (1999), Roberts (2002) y Gutiérrez-Rexach (2002) sobre el contenido presuposicional activado por los demostrativos. El contenido presuposicional al que nos referimos se puede representar en términos de una ERD presuposicional. El conjunto inicial de presuposiciones constituye el contexto de establecimiento inicial del demostrativo de tal manera que la ausencia de este contenido presuposicional significaría que el uso del demostrativo sería pragmáticamente inviable. La ERD presuposicional inicial del demostrativo anterior a la introducción de su fuerza cuantificacional es la siguiente: **[***e*, *x*, e_1, *t* | **agente** (*e*, *x*); **tiempo** (*e*, *t*); **señalamiento** (e_1); **agente** (e_1, *x*); **tiempo** (e_1, *t*)**]**, en donde una ERD presenta la forma básica **[**universo | **condiciones]** y los referentes de discurso *e*, *x*, e_1 y *t* representan, respectivamente, el evento de habla, el hablante, un gesto de señalamiento que identifica a una entidad de discurso α, y un tiempo que ancla los eventos de habla y señalamiento. En esta línea, una ERD enriquecida presuposicionalmente para la oración de (13) sería la que se muestra en (15), y de la que hemos suprimido la condición **tiempo** con el fin de simplificar su lectura. Las ERD's K_1 y K_2 son subordinadas de la ERD principal K y el símbolo flecha '⇒' representa la fuerza cuantificacional del demostrativo.

(15) [$_K$ *e*, *x*, e_1, *y* | **acto_habla** (*e*); **agente** (*e*, *x*); **señalamiento** (e_1) [$_{K1}$ | **perro** (*y*)] ⇒ ESTE [$_{K2}$ | **ladra** (*y*)]]
⟨anclaje (*y*,α)⟩

Nótese que hemos decidido mantener la condición **señalamiento** ya que asumimos que el demostrativo en (13) ha sido utilizado por el hablante acompañándolo con un gesto de señalamiento explícito (por ej. un dedo índice apuntando a la entidad referida). Llegados a este punto ya podemos ofrecer una caracterización de los tres demostrativos del español (*este*/*ese*/*aquel*). La caracterización es la que ofrecemos en (16)-(18).

(16) Este: [*e*, *x*, *y*, *t* | **acto_habla** (*e*); **agente** (*e*, *x*); **tiempo** (*e*, *t*); **señalamiento** (*x*, *y*); **proximal** (*y*, *x*)]
⟨anclaje (*y*, α)⟩

(17) Ese: [*e*, *x*, y, *x'*, *t* | **acto_habla** (*e*); **agente** (*e*, *x*); **tiempo** (*e*, *t*); **señalamiento** (*x*, *y*); **oyente** (*e*, *x'*); **inespecífico** (*y*, *x*); **inespecífico** (*y*, *x'*)]
⟨anclaje (*y*, α)⟩

(18) Aquel: [*e*, *x*, y, *x'*, *t* | **acto_habla** (*e*); **agente** (*e*, *x*); **tiempo** (*e*, *t*); **señalamiento** (*x*,*y*); **oyente** (*e*, *x'*); **distal** (*e*, *x*); **distal** (*e*, *x'*)]
⟨anclaje (*y*, α)⟩

Como se puede comprobar, los tres términos comparten prácticamente el mismo universo, que incluye un referente de discurso del tipo acto de habla (*e*), un agente o hablante (*x*), y un referente de discurso (*y*) que representa a una entidad discreta o de orden superior que puede estar presente en el contexto de habla o pertenecer al ámbito estrictamente textual. Los pronombres demostrativos *ese* y *aquel* incluyen un referente x' que representa al oyente. Obsérvense las diferencias entre los demostrativos en lo que respecta a las condiciones sobre los referentes de discurso mencionados. El demostrativo de proximidad *este* presupone que el objeto señalado (*y*) se encuentra próximo al hablante; el demostrativo medio *ese* es inespecífico con respecto a la condición de proximidad y el demostrativo de lejanía *aquel* presupone que el objeto señalado no está distante tanto del hablante como del oyente. En su uso más frecuente, los pronombres demostrativos neutros *esto*, *eso* y *aquello* sirven para hacer referencia a entidades más abstractas o de orden superior en el discurso (eventos, proposiciones, etc.), que suelen ser comúnmente introducidas en el discurso mediante oraciones completas u oraciones subordinadas[4].

Proponemos una caracterización para todos los sintagmas nominales encabezados por un demostrativo, es decir, para los pronombres demostrativos y los sintagmas nominales del tipo 'dem.+ N' como cuantificadores generalizados, o lo que es lo mismo, funciones de conjuntos a valores de verdad. Los cuantificadores generalizados son del tipo ⟨⟨e, t⟩, t⟩ y denotan familias de conjuntos. De manera alternativa, los cuantificadores generalizados se pueden representar mediante la expresión lógica Q(λx.P(x)), la cual sólo se convierte en una formula verdadera syss el conjunto

[4] Las entidades abstractas como eventos, acciones, hechos, etc. pueden también expresarse mediante sintagmas nominales. Así, por ejemplo, el sintagma *El asesinato de JFK* denotaría un evento pasado.

denotado por (λx.P(x)) pertenece a la denotación del cuantificador (Barwise y Cooper 1981). Así, por ejemplo, ESO (λx.P(x)) se considerará una fórmula verdadera syss el conjunto denotado por (λx.P(x)) es un miembro de la denotación del cuantificador ESO. Como consecuencia de su estatus pronominal, los pronombres demostrativos exhibirán un comportamiento similar al de otros pronombres aunque, ya que se utilizan comúnmente para referir a entidades abstractas o de orden superior, estarán condicionados por una operación de sustitución de variable abstracta ζ con una entidad abstracta (evento, situación, etc.). Dicha entidad satisfará otros requisitos conversacionales, como el ser prominente y mencionada previamente en el discurso. Considérese el siguiente fragmento de discurso (19) y su derivación en la figura 2.

(19) Juan vino. ***Aquello*** me sorprendió.

1. [Juan vino] → ζ
2. Aquello → λζ. Q(ζ)
3. Aquello → Q(ζ)
   * por conversión λ
4. λQ(Q(ζ))
5. sorprender → (**sorprender'**)
   * expresión de tipo ⟨e, t⟩
6. sorprender(Juan vino)
   * por conversión λ

Figura 2: derivación del discurso en (19)

El breve discurso presentado en (19) constituye un caso de anáfora ínteroracional que debemos concebir de manera dinámica. El discurso se construye de manera incremental y, por ello, la segunda oración de (19), la que contiene el pronombre, no se puede interpretar completamente sin la contribución de la primera oración en lo que respecta a la introducción de los referentes de discurso necesarios en el espacio/modelo cognitivo construido y compartido por los participantes de la conversación. La primera oración *Juan vino* constituye en sí misma una proposición completa. Asumamos que esta oración expresa un evento que representaremos con el símbolo de variable ζ para entidades de discurso abstractas. El referente de discurso de evento asociado a la oración *Juan vino* constituye un antecedente potencial para el referente de discurso con el que contribuye el pronombre demostrativo. Ya que estamos caracterizando a los pronombres demostrativos como conjuntos o paquetes de propiedades (las propiedades que tiene, por ejemplo, *aquello*), y dado que, desde el punto de vista de su extensión, las propiedades constituyen conjuntos de individuos, consideramos que la denotación del demostrativo es un conjunto de conjuntos o, dicho de otro modo, el conjunto de todos los conjuntos X tales que ζ es un miembro de X. Formalmente, [[AQUELLO]] = {X ⊆ U | ζ ∈ X}. El operador lambda λζ liga una variable de entidad abstracta (evento, situación, proposición, etc.).

Al igual que con los demostrativos complejos del tipo 'este N' tratados en (16)-(18), los pronombres demostrativos también requieren un contexto de establecimiento inicial de naturaleza presuposicional. La representación presuposicional que proponemos para estos pronombres es la que ofrecemos en (20)-(22).

(20) Esto: [*e*, *x*, ζ, *t* | **acto_habla**(*e*); **agente**(*e*, *x*); **tiempo**(*e*, *t*); **señalamiento**(*x*, ζ); **inespecífico**(ζ, *x*); **activado**(ζ)]

(21) Eso: [*e*, *x*, ζ, *t* | **acto_habla**(*e*); **agente**(*e*, *x*); **tiempo**(*e*, *t*); **señalamiento**(*x*, ζ); **inespecífico**(ζ, *x*); **activado**(ζ)]

(22) Aquello: [*e*, *x*, ζ, *t* | **acto_habla**(*e*); **agente**(*e*, *x*); **tiempo**(*e*, *t*); **señalamiento**(*x*, ζ); **distal**(ζ, *x*); **activado**(ζ)]

Como ya hemos comentado en este trabajo, en su uso más habitual los hablantes utilizan los pronombres demostrativos para la referencia anafórica a entidades de discurso abstractas. Como elementos anafóricos, los demostrativos no necesitan un gesto de señalamiento explícito. Por tanto, surge la cuestión de hasta qué punto es necesario o conveniente mantener un acto demostrativo o gesto de señalamiento presuposicional en la caracterización ofrecida para los pronombres demostrativos. En este trabajo defendemos que tal presuposición de señalamiento ha de entenderse no como un gesto demostrativo físico explícito sino como un mecanismo cognitivo asociativo que es activado o 'disparado' de modo inherente con el uso del pronombre y cuya función es la de dirigir o enfocar la atención del hablante y del oyente hacia un único demonstratum abstracto. Dicho demonstratum posee un estatus cognitivo 'activado', siguiendo, con ciertas reservas, la terminología de Gundel et al. (1993). Consideremos los siguientes ejemplos tomados de Wolter (2006) y que sirven para ilustrar esta idea.

(23) [Mary brings a large package into the room. Everyone stares at the package as it starts to tick and rock back and forth]
It's going to explode.

(24) [Mary brings a large package into the room. Only John notices as the package starts to tick and rock back and forth]
That's going to explode.

Los ejemplos (23) y (24) nos permiten observar la diferencia existente entre entidades con estatus cognitivo EN FOCO en contraposición a aquellas cuyo estatus es simplemente ACTIVADO. En (23), el uso del pronombre personal de tercera persona del inglés *it* es posible ya que el objeto referido es el centro de atención de los participantes en la situación de habla. Esto se debe a que ahora todos los participantes en la conversación han fijado su mirada en el paquete que *Mary* introdujo en la habitación en un momento dado. Por otro lado, el pronombre demostrativo *that* se utiliza en (24) porque el objeto en cuestión es sólo un objeto más del conjunto de todos los objetos presentes en la situación de habla. En (24), el paquete tiene el estatus de meramente ACTIVADO y por ello el uso del pronombre de tercera persona *it* no sería pragmáticamente adecuado en este caso. Es importante señalar que para Gundel et al. y sus seguidores no existen diferencias de estatus cognitivo entre los pronombre demostrativos. Es decir, el referente de un pronombre demostrativo siempre tendrá el estatus cognitivo de ACTIVADO independientemente del demostrativo utilizado. En contraposición, el referente de un pronombre personal siempre tendrá el estatus cognitivo EN FOCO.

Nosotros no coincidimos plenamente con la explicación ofrecida por Gundel et al. para explicar el estatus cognitivo que marcan estas expresiones y que se ilustra en (23)-(24), al menos no para el caso del español. En nuestra opinión, precisamente la función del pronombre demostrativo es la de 'promocionar' el estatus cognitivo de la entidad señalada/referida. Por tanto, una precondición necesaria para el uso pragmático adecuado o feliz de los pronombres demostrativos en español es que su referente potencial tenga un estatus igual o menor que ACTIVADO según la escala de estados de Gundel et al (*The Givenness Hierarchy*). Una vez que se utiliza el demostrativo para referir a dicho objeto, éste adquiere automáticamente el estatus EN FOCO. De este modo, los pronombres demostrativos del español tendrían la función de **marcadores de foco** o **focalizadores referenciales**. El mecanismo de señalamiento cognitivo propuesto en las caracterizaciones de (20)-(22) ayudaría a los interlocutores a buscar un tipo particular de entidad de discurso abstracta entre el conjunto de potenciales referentes activados en la situación de habla y que son candidatos teóricamente posibles a referente.

Por otro lado, como se puede observar en (22), hemos incluido la condición **distal**(ζ, x) en la caracterización del pronombre demostrativo *aquello* y que lo distingue de los otros dos elementos que son inespecíficos con respecto a la distancia. La explicación de por qué se debe mantener la condición de distancia para el demostrativo es la siguiente. Hay casos de referencia anafórica con pronombres demostrativos en los que el referente es una entidad neutra o abstracta de orden superior. En ellos el parámetro de distancia espacial que nos ayuda a definir los demostrativos usados deícticamente no parece, en principio, necesario para caracterizar los pronombres demostrativos pues algunos o todos los parámetros contextuales de la situación de habla (el *yo*, *aquí* y *ahora*) desaparecen al transferirse el ámbito referencial a lo textual discursivo. Ya no hay, por ejemplo, distancia física entre el hablante y el objeto referido. Sin embargo, existen casos de la llamada deixis temporal en los que referencia se hace a tiempos pasados o futuros. En estos casos de deixis temporal, parece existir un uso muy marcado del demostrativo distal *aquel* tanto en su forma determinante como pronominal para referir a tiempos pasados previamente introducidos en el discurso mediante verbos en pasado, expresiones temporales o sustantivos que denotan eventos del pasado (véase Fernández-Ramirez 1957 y Zulaica-Hernández & Gutiérrez-Rexach 2009). Esto es así hasta el punto de que resulta difícil encontrar casos del demostrativo *aquel* en situaciones de discurso en las que el marco temporal del mismo no sea un tiempo pasado. Por el contrario, los pronombres demostrativos *esto* y *eso* no participan de en la referencia a tiempos de una manera tan marcada. Es por ello que hemos decidido mantener una condición de lejanía en la caracterización del pronombre demostrativo *aquello*. En suma, la condición de proximidad o lejanía, basada en la referencia temporal, habría sido metafóricamente reinterpretada por los hablantes a partir del parámetro de distancia espacial típica de los usos demostrativos canónicos.

Llegados a este punto ya podemos ofrecer una representación final para el discurso (19) que presentamos en la figura 3. Los referentes de discurso *e* y e_2 se asocian con los predicados *venir* y *sorprender*, respectivamente. La variable de objeto abstracto ζ se identifica con la sub-ERD K_1 y el referente de discurso *z* se identifica con el

pronombre demostrativo *aquello* el cual, a su vez, se identifica con ζ. De esta manera queda resuelta la referencia ínteroracional.

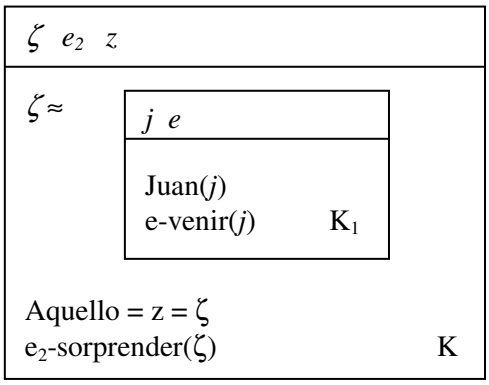

Figura 3: Representación del discurso (19)

4. Conclusión

La representación semántica de los demostrativos del español como catalizadores de ciertas condiciones presuposicionales que hemos propuesto en este trabajo nos permite explicar tanto los usos deícticos canónicos de los demostrativos como aquellos usos problemáticos en los que la denotación del demostrativo parece asemejarse más a la de una variable. Estos casos, frecuentes en los procesos de anáfora del discurso, representan un alto porcentaje del número total de usos de los demostrativos en el discurso hablado y escrito y, en consecuencia, no deben desatenderse en los estudios de procesamiento de lenguas naturales. Por desgracia, la falta de espacio nos impide explicar con más detalle algunos de los rasgos específicos de los demostrativos que hemos mencionado en este trabajo y que son igualmente relevantes para una mejor comprensión de cómo se procesan estos elementos. Entre estos rasgos se encuentran las condiciones de proximidad y su translación desde el espacio contextual al ámbito estrictamente textual, la presuposición de 'familiaridad', y las consecuencias que el contraste entre estas dimensiones tiene para la estructura particular de los sistemas demostrativos.

Referencias

Asher, N. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

Barwise, John y Robin Cooper. 1981. Generalized Quantifiers and Natural Language. *Linguistics and Philosophy* 4, pp. 159-219.

Byron, D. 2004. Resolving Pronominal Reference to Abstract Entities. Technical Report 815. The University of Rochester, NY.

Fernández-Ramírez, S. 1951. *Gramática española 3.2. El pronombre*. Madrid: Arco Libros.

Grosz, B., K. J. Aravind y S. Weinstein. 1995. Centering: A Framework for Modelling the Local Coherence of Discourse, *Computational Linguistics* 21(2): 203—226.

Gundel, Jeanette K., Hedberg, N. and R. Zacharski. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69, 274—307.

Gutiérrez-Rexach, Javier. 2002. Demonstratives in Context. In Javier Gutiérrez-Rexach (ed.) *From Words to Discourse. Trends in Spanish Semantics and Pragmatics*, Oxford/New York: Elsevier Science, pp.195—236.

Hawkins, J.A. 1978. Definiteness and Indefiniteness. Humanities Press, Atlantic Highlands.

Kaplan, D. 1989. Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and other Indexicals. In J. Almog et al (eds.), *Themes from Kaplan*. New York, Oxford University Press, pp. 481—564.

Kennan, Edward y Leonard Faltz. 1985. *Boolean Semantics for Natural language*. Dordrecht: Reidel.

King, J. 1999. Are Complex 'That' Phrases Devices of Direct Reference? *Noûs* 33, pp. 155—182.

Linde, C. 1979. Focus of Attention and the Choice of Referential Expressions in Discourse. In T. Givón (ed.): *Syntax and Semantics*, 12. Academic Press.

May, R. 1989. Interpreting Logical Form. *Linguistics and Philosophy* 12(4), pp. 387—435.

May, R. 1985. Logical Form: Its Structure and Derivation. *Linguistic Inquiry Monographs* 12. Cambridge, Mass.: MIT Press.

Passonneau, R. 1989. Getting at Discourse Referents. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* (ACL 89), pp. 51—59.

Poesio, M. y N. Modjeska. 2005. Focus, Activation and THIS-Noun Phrases. In Branco, A., T. McEnery and R. Mitkov (eds.). *Anaphora Processing*. John Benjamins, pp. 429—456.

Real Academia Española: Banco de datos (CREA). Corpus de referencia del español actual.<http://www.rae.es>

Roberts, C. 2002. Demonstratives as Definites. In K. van Deemter and R. Kibble, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. Stanford: CSLI Publications, pp. 89—196.

Russell, B. 1905. On Denoting. *Mind* 14: 479-493. Reprinted in Robert C. Marsh (ed.) *Logic and Knowledge: Essays 1901-1950 by Bertrand Russell, Allen & Unwin*, London 1956.

Vendler, Z. 1957. Verbs and Times. *The Philosophical Review* 66 (2), pp. 143—160.

Webber, Bonnie L. 1991. Structure and Ostension in the Intepretation of Discourse Deixis. *Natural Language and Cognitive Processes*, 6(2), pp. 107—135.

Webber, Bonnie L. 1988. Discourse Deixis and Discourse Processing. Technical Report. University of Pennsylvania, Philadelphia, PA.

Webber, Bonnie L. 1979. *A Formal Approach to Discourse Anaphora*. Garland, New York.

Wolter, Lynsey K. 2006. *That's That: The Semantics and Pragmatics of Demonstrative Noun Phrases*. Ph. D. Dissertation. University of California at Santa Cruz.

Zeevat, H. 1999. Demonstratives in Discourse. *Journal of Semantics* 16: 279—314.

Zulaica-Hernández, I. 2009. Demonstratives and the rhetorical structure of discourse. *Sintagma* 21, Universitat de Lleida.

Zulaica-Hernández, I. y J. Gutiérrez-Rexach. 2009. Tense, temporal expressions and demonstrative licensing in natural discourse. *Proceedings of SigDial 2009*, pp. 97-106.

Novas Perspectivas

Os dicionários onomasiológicos e as ontologias computorizadas

Patrícia Cunha França
Mestranda em Ciências da linguagem
(Área de Especialização em Língua e Tecnologias de Informação)
Universidade do Minho
pg10122@alunos.uminho.pt

Resumo

Este artigo pretende construir a ponte entre dicionários onomasiológicos e as recentes ontologias computorizadas ou formais.

São apresentados aqui os conceitos de onomasiologia e de dicionário onomasiológico, de forma a tomá-los como instrumentos auxiliares no trabalho que tem vindo a ser desenvolvido relativamente às ontologias. São expostas aqui também algumas das críticas, do ponto de vista prático e teórico, que esses dicionários mereceram aquando da sua publicação, de forma a que possam ser úteis à construção das ontologias modernas.

Farei ainda uma breve nota sobre o que está hoje a ser feito, na prática de elaboração de ontologias computorizadas, para ultrapassar algumas das limitações apontadas aos produtos lexicográficos onomasiológicos.

1. Introdução

> *Among the wide Spectrum of information representation and retrieval tools are thesauruses and ontologies, which are the most often linked in bibliography, even though they come from very different disciplinary areas*. (ARANO, 2005)

É comum falar-se de dicionários, muito particularmente de thesaurus[1], quando se lê sobre ontologias computorizadas[2]. Diversos autores, nomeadamente Arano (2005), Moreira, Alvarenga & Oliveira (2004), Hirst (2004), Oltramari & Vetere (2008) e Wielinga *et al*. (2001), têm escrito sobre o tema, esforçando-se por estabelecer pontes entre os dois instrumentos de representação.

Este artigo tem por finalidade contribuir para a definição dos conceitos de onomasiologia e de dicionário onomasiológico, de forma a que o trabalho desenvolvido nessa área possa constituir-se como um fundamento teórico capaz de auxiliar a construção de ontologias.

Para que se possa fazer a ponte entre dicionários onomasiológicos e ontologias é necessário começar por focar as principais semelhanças entre os dois instrumentos de representação. Assim, no ponto 2 deste artigo tentarei estabelecer essas semelhanças.

Partindo do pressuposto de que as modernas ontologias assentam no mesmo princípio teórico dos produtos lexicográficos onomasiológicos, passarei, no ponto 3, a descrever os conceitos de onomasiologia e dicionário onomasiológico, no contexto específico da Lexicografia. Ainda neste ponto são deixadas algumas das críticas, do ponto de vista prático e teórico, apontadas às obras lexicográficas onomasiológicas, muito particularmente a um dos exemplos mais paradigmáticos da história da Lexicografia onomasiológica: o *Sistema de Conceitos* de Hallig & Wartburg[3].

No ponto 4 deixo algumas notas sobre algumas propostas que estão hoje a ser consideradas, na prática de elaboração das ontologias computorizadas, para a resolução de alguns problemas teóricos apontados aos produtos da Lexicografia onomasiológica no passado.

2. De dicionários onomasiológicos e ontologias

Como referi acima, vários autores têm tentado estabelecer pontos de contacto entre thesaurus e

1 Um thesaurus pode ser entendido como um dicionário, embora com particularidades específicas. Faz parte da categoria de dicionários onomasiológicos.

2 Entende-se aqui uma ontologia computorizada como um artefacto usado para representação de conhecimento, que utiliza um determinado tipo de linguagem mais ou menos formal elaborado num contexto particular e para um fim específico. Numa ontologia estão representados os conceitos de um domínio particular bem como relações entre esses conceitos. Para uma explicação mais detalhada sobre o termo ontologia computorizada, ver FRANÇA, 2009a (Capítulo 3).

3 *Sistema de Conceitos* será usado como abreviatura para a obra de Hallig & Wartburg que tem como título original *Begriffssystem als Grundlage für die Lexikographie / Système Raisonné des Concepts pour Servir de Base à la Lexicographie.*

ontologias, argumentado que os thesaurus podem ser considerados ontologias simples. Uma das semelhanças entre os dois é que ambos usam termos e relações entre esses termos, termos estes que representam conceitos. Se estes conceitos, por sua vez, representam objectos linguísticos ou objectos do mundo real é uma pergunta cuja resposta depende da forma como nós concebemos uma ontologia e, em grande parte, depende também do propósito para o qual a construímos.

No mais, as relações hierárquicas entre conceitos estão presentes quer nos thesaurus, ou nos dicionários em geral, quer nas ontologias, na relação de subsunção 'is_a', ainda que ela esteja implícita num thesaurus ou num outro tipo de dicionário. Hirst, por exemplo, nota que um dicionário contém uma ontologia implícita, ou, pelo menos, uma hierarquia semântica ao apontar definições aristotélicas básicas:

> For example, if *automobile* is defined as a *self-propelled passenger vehicle that usually has four wheels and an internal-combustion engine*, then it is implied that *automobile* is a hyponym of *vehicle* and even that automobile IS-A VEHICLE; semantic or ontological part-whole relations are also implied (HIRST, 2004: 223)

Para além desta hierarquia semântica, o contexto em que as obras lexicográficas apareceram, o propósito para que foram construídas são elementos de contacto importantes com as ontologias actuais.

Não obstante estas semelhanças, Nickles *et al*. (NICKLES *et al.* 2007: 45) apontam três diferenças fundamentais a ter em conta:

(i) as ontologias usam linguagem formal[4], enquanto que os dicionários usam linguagem natural. Na verdade, uma das principais características das ontologias é conseguir usar uma linguagem capaz de ser processada por máquinas. Os dicionários destinam-se a ser interpretados por seres humanos. Como referem os autores, "nenhuma máquina é actualmente capaz de entender um dicionário num sentido realista da palavra 'entendimento'" (NICKLES *et al.,* 2007: 45);

(ii) o dicionário é descritivo, na medida em que fornece definições inseridas num determinado tempo específico, com anotações acerca da forma como as palavras são usadas num período de tempo específico. Uma ontologia formal computorizada é prescritiva e normativa; ela determina especificamente, numa linguagem formal, o que um dado termo significa;

(iii) um termo numa ontologia não é uma palavra, mas um conceito. Se é verdade que os termos numa ontologia podem receber nomes, que correspondem a palavras ou combinação de palavras, de forma a poderem ser facilmente entendidos por humanos, uma ontologia formal poderia perfeitamente substituir esses termos por códigos arbitrários, sem perder as suas propriedades formais[5].

Há ainda um outro argumento, que vem no seguimento deste último, desta vez elaborado não por Nickles *et al*., mas por Hirst (2004), e que sustenta que uma ontologia é radicalmente diferente de um thesaurus porque este último lida com palavras e não com objectos do mundo real[6], como acontece numa ontologia. Hirst defende que uma ontologia representa instâncias no mundo real e dificilmente podemos considerá-la um objecto linguístico. As relações ontológicas são, por isso, fundamentalmente diferentes das relações lexicais:

> An ontology […] is a set of categories of objects or ideas in the world, along with certain relationships among them; it is not a linguistic object. (HIRST, 2004: 8)

É necessário fazer aqui algumas objecções a estes três pontos. E partimos do princípio de que um dicionário onomasiológico cabe na categoria de dicionário proposta por Nickles *et al*. .

Relativamente ao ponto (i), se é verdade que as ontologias são construídas com linguagens formais, com o objectivo de serem processadas por computadores, também é verdade que as ontologias são lidas por seres humanos. O desafio que se coloca, quando se pensa em linguagens para definir ontologias, é precisamente esse: conseguir o poder

4 XML, UML ou OWL são algumas das linguagens de modelagem usadas hoje nas ontologias.

5 Como fazem notar Nickles *et al*., entre outros, os itens linguísticos usados nas ontologias formais não são signos linguísticos no seu sentido pleno, com forma e conteúdo, mas cadeias de bytes. (NICKLES *et al*, 2007: 32). De resto, como veremos adiante, Hallig e Wartburg também irão argumentar que as palavras usadas para representar os conceitos no seu *Sistema de Conceitos*, também não são signos linguísticos no seu sentido pleno.

6 Johansson parece-me esclarecedor neste ponto ao tentar esclarecer a distinção entre olhar para e olhar através das palavras. A linguagem pode ser usada como ferramenta para transmitir informação ou como ferramenta em si:

> When, for example, one is conveying or receiving information in a language in which one is able to make and understand language acts spontaneously, one is not looking at the terms, concepts […] in question […]. Rather, one looks through these linguistic entities in order to see the information (facts, reality, or objects) in question. We are looking at linguistic entities, in contrast, when fro example we are reading dictionaries and terminologies (JOHANSSON, 2008).

expressivo para descrever conteúdo processável por máquinas, mas, ao mesmo tempo, permitir que os humanos possam lê-las sem grande esforço. É isto que defende Lacy ao reportar-se à OWL[7]:

> Developers of Owl wanted to make the language intuitive for humans and to have sufficient power to describe machine-readable content needed to support Semantic Web applications. (LACY, 2005: 43)

Se olharmos para uma mesma conceptualização[8] representada num esquema UML ou em OWL verificamos que é muito mais fácil ler um esquema em UML do que em linguagem OWL. Não obstante, a capacidade representativa da OWL é superior à linguagem UML.

No que diz respeito ao argumento (ii), se teoricamente um dicionário assenta sobre um carácter descritivo, não podemos deixar de lhe apontar um carácter prescritivo. Em termos teóricos, um dicionário descreve a língua usada pelos falantes num dado momento e num dado espaço, mas que dizer às palavras de Green, quando se refere a Samuel Johnson e Noah Webster, lexicógrafos do século XVIII?

> What both men were doing, although neither articulated it as such, was playing God. Or if not God, then at least Moses, descending from Sinai with the tablets of the law. For them their role was not simply to select a word list, define it, and make it available to the reading public; in addition they took on the priestly task of revealing a truth, in this case a linguistic one. (GREEN, 1996: 5)

Relativamente ao argumento (iii), que defende que ontologias se separam dos dicionários por lidarem com conceitos, isto não é bem verdade. Os dicionários onomasiológicos são normalmente conhecidos por lidarem com conceitos, contrariamente ao que acontece com os dicionários semasiológicos, se bem que há questões importantes a serem esclarecidas no que respeita ao conceito CONCEITO[9].

E, com este argumento, podemos também objectar o argumento de Hirst. É que, embora possamos concordar que um dicionário semasiológico lida com palavras, entendidas como objectos linguísticos, ao contrário de uma ontologia, que toma essas palavras como itens representativos do mundo real, dificilmente podemos dizer a mesma coisa de um dicionário onomasiológico, que se detém no conceito e não na forma ou na palavra como objecto linguístico. De resto, não é correcto afirmar que uma ontologia lida apenas com conceitos e não com signos linguísticos. As ontologias linguísticas[10], como é o caso da WordNet, lida especificamente com signos linguísticos, dada a forma como os termos são trabalhados.

Considerar uma ontologia um objecto linguístico não depende da natureza da própria ontologia, mas do propósito para que é construída e da forma como olhamos para os termos com os quais queremos construir uma ontologia.

Tomemos um exemplo prático. A entrada '*cat*' na WordNet[11] aparece-nos como signo linguístico, na medida em que surge com várias acepções e é categorizada, logo à partida, gramaticalmente: são-nos dadas dez acepções, sendo que oito cabem na categoria 'nome' e duas na categoria 'verbo'. Também nos são dados exemplos de uso na língua e sinónimos. É comum ver-se esta informação linguística num qualquer dicionário de língua. No entanto, o que faz da WordNet uma ontologia[12] e não um simples dicionário são as relações semânticas que ela disponibiliza entre os termos, tais como hiponímia, hipernonímia, etc.

Mas há ainda uma outra objecção a ser feita ao argumento (iii) de Nickles *et al*., que vem no seguimento do que dissemos para o ponto (i). É que, embora uma ontologia seja feita para ser processada por computadores[13], ela deve permanecer inteligível para os seres humanos, de forma a poder ser usada por estes. E de facto, Nickles *et al*., referem que

7 A OWL, *Web Ontology Language*, é a linguagem para representar o conhecimento proposta pela W3C (World Wide Web Consortium http://www.w3.org/).

8 Uma conceptualização pode ser entendida como um conjunto de termos e relações entre termos independentemente da linguagem usada para os representar. Para uma discussão acerca do termo 'conceptualização', ver GUARINO, 1998.

9 Para um esclarecimento das diferentes interpretações para o conceito de CONCEITO no seio da Linguística, ver FRANÇA, 2009a (Capítulo 5).

10 Magnini & Speranza (2002) definem as ontologias linguísticas como recursos que olham para os seus itens como objectos linguísticos, embora com uma particular atenção aos conceitos. As ontologias linguísticas são definidas como

> large lexical resources that cover most words of a language, while at the same time also providing an ontological structure where the main emphasis is on the relations between concept; linguistic ontologies can therefore be seen both as a particular kind of lexical database and as particular kind of ontology. (MAGNINI & SPERANZA, 2002:43)

11 http://wordnet.princeton.edu/

12 Há que referir aqui que há autores que não consideram a WordNet uma ontologia, mas simplesmente uma base de dados lexicais.

13 Nickles *et al.* estão a referir-se, certamente, às ontologias elaboradas com linguagem formal, como a OWL.

uma das questões que se põem hoje aos projectos que estudam a linguagem e as ontologias é o estabelecimento de uma ligação satisfatória entre ontologias e as expressões linguísticas (NICKLES *et al*, 2007: 44).

Como vemos, há elementos de contacto entre thesaurus e ontologias. Uma análise das críticas que os thesaurus, e outros produtos da lexicografia onomasiológica, sofreram aquando da sua publicação merece especial atenção, uma vez que podem fornecer pistas importantes para melhores práticas.

Assim, o ponto seguinte é dedicado ao percurso do conceito de onomasiologia, tal como ele apareceu na Lexicografia, e explorar as visões críticas que ele mereceu. Isto porque, como refere Arano (2005), se o conceito de ontologia nasce no seio da Filosofia, o conceito de thesaurus nasce no seio da Lexicografia e pode entender-se como um produto daquilo que se designa por Lexicografia onomasiológica.

3. Do conceito de onomasiologia

De uma forma didáctica, Grzega e Schöner definem a onomasiologia como o ramo da Lexicologia que tem por finalidade "encontrar as formas linguísticas, ou as palavras, que podem estar em vez de um dado conceito/ideia/objecto (GRZEGA & SCHÖNER, 2007: 7). Para os autores, a onomasiologia pode também ser considerada como "o estudo das designações", mesmo quando o que se procura seja uma forma gramatical ("*How can I express future time?*") ou um padrão comunicacional ("*How can I greet somebody?*") (GRZEGA & SCHÖNER, 2007: 7).

Esta definição está em consonância com a tradicional distinção entre semasiologia e onomasiologia, proposta pelo *Dicionário de Linguística* de Dubois:

> *onomasiologia* é o estudo das denominações; ela parte do conceito e busca os signos linguísticos que lhe correspondem. […] A onomasiologia opõe-se à semasiologia, que parte do signo para ir em direcção à ideia. (DUBOIS *et al.*, 1998: **onomasiologia)**

Antes de explorarmos o conceito de onomasiologia, antes de entendermos melhor como a onomasiologia contrasta com a semasiologia, importa olharmos um pouco para trás e perceber onde surgiu a onomasiologia na história da Lexicografia e da Linguística.

3.1 Da origem da palavra na Lexicografia

A palavra 'onomasiologia' foi usada pela primeira vez, segundo Grzega (GRZEGA, 2002: 1022) e Casares (CASARES, 1992: 54), pelo alemão Zauner em 1902, num estudos sobre os nomes das partes do corpo em línguas românicas. E é precisamente na segunda metade do século XIX e inícios do século XX, com os trabalhos sobre as línguas românicas, que o interesse pela onomasiologia ganha força (HÜLLEN, 1999:16). Segundo Casares, o termo 'onomasiologia' surgiu, precisamente, no seio da Lexicologia, mais exactamente, a partir do termo "lexicologia comparada", usado por Tappolet em 1895[14]. Casares entende-a como a disciplina, no âmbito da Semântica, que

> partiendo de una cosa determinada, un objeto o una noción, se propone estudiar comparativamente los caminos que esa cosa ha seguido hasta encarnar en una palabra, y pretende reconstruir el proceso intelectual e imaginativo que determino tal encarnación. (CASARES, 1992: 54)

A obra de Tappolet foi apenas uma de entre as muitas que surgiram dos estudos sobre onomasiologia no âmbito da Lexicologia, nomeadamente a Lexicologia comparada[15]. As línguas românicas eram o objecto de eleição. Como refere Babini, isto deve-se ao facto de que tinham por origem o latim, o que permitia fazer o percurso histórico até às origens de determinados conceitos. Assim, partindo do latim e comparando diferentes línguas românicas, "foram analisadas dezenas de grupos de ideias, tais como as estações e os meses do ano, a flora, a fauna, os aspectos da vida humana etc." (BABINI, 2006: 38). Diferentes falantes de diferentes regiões atribuíam um nome a um mesmo conceito, através de questionários. Dos dados obtidos eram construídos mapas linguísticos, que, por sua vez, se constituíam em atlas (GRZEGA & SCHÖNER, 2007: 8).

14 Tappolet, E. (1895). *Die romanischen Verwandtschaftsnamen mit besonderer Berücksichtigung der französischen und italienischen Mundarten. Ein Beitrag zur vergleichenden lexikologie*. Estrasburgo: sem ed. a*pud* BABINI, 2006: 38.

15 Babini, por exemplo, refere a obra de Wartburg de 1928 [Wartburg, W. (von) (1928). *Französisches etymologisches wörterbuch* (FEW). 22 vol. Bonn; Klop; puis Leipzig et Berlin: Teubner; Bâle: Lichtenhahn] (BABINI, 2006: 41).

3.2 A onomasiologia na história da Lexicografia

Se a origem da palavra 'onomasiologia' surgiu com a Lexicologia, a verdade é que já muito tempo antes se havia desenvolvido o conceito na Lexicografia. Pode mesmo afirmar-se que a origem da Lexicografia onomasiológica pode ser contada a partir da própria história da Lexicografia, que remonta, provavelmente, aos séculos V a II A. C. (GREEN, 1996: 34). Mesmo que o termo 'onomasiologia' não existisse para designar nenhum tipo de obra lexicográfica, podemos inscrever o conceito de onomasiologia na história da Lexicografia desde então.

Ficando desde já precavidos para o facto de que a conquista da ordem alfabética não significa que a Lexicografia feita antes desse momento possa ser definida como onomasiológica, parece-nos interessante remontar a essa época para verificarmos em que termos se processava a Lexicografia antiga e perceber o motivo da importância dada à ordenação alfabética.

Num pequeno texto de dezasseis páginas intitulado "Petit histoire de la conquête de l'ordre alphabétique dans les dictionnaires médiévaux", Boulanger (BOULANGER, 2002) traça uma breve história acerca do contexto em que surgiu a importância da ordenação alfabética nos dicionários medievais. Como refere o autor, o que começou por ser apenas pequenas anotações feitas aos manuscritos e aos *códices* na Europa Medieval acabou por se transformar em verdadeiras compilações de palavras, ordenadas segundo o texto de onde procediam; uma ordenação em termos discursivos, portanto.

Boulanger escreve que as anotações aos manuscritos, feitas na Europa Medieval, se iam adensando no mesmo texto, o que obrigou a uma compilação e seriação (BOULANGER, 2002:11).

Desta forma surgiam as *glossae collectae*. Estas compilações constituíam-se como uma transposição das anotações feitas; as palavras e anotações estavam organizadas de acordo com a ordem em que apareciam nos manuscritos (BOULANGER, 2002: 11).

Não obstante esta ordenação prática, as anotações aumentavam e tornou-se crucial encontrar um novo sistema de compilação, que impedisse a repetição de anotações e poupasse espaço e tempo aos copistas. Daqui surge, então, a génese do que viria a transformar-se no método de ordenação alfabética, tomando por princípio de indexação a própria palavra, tida agora como uma entidade autónoma, independente do texto de onde provinha (BOULANGER, 2002: 12).

Começando por tomar como princípio de ordenação a primeira letra e, depois do século VIII e X, a segunda letra das palavras, a terceira letra passa a ser também, progressivamente, considerada. O que Boulanger enfatiza neste processo de compilação é, precisamente, a tomada em consideração dos signos linguísticos separados, retirados do texto e tomados por si mesmos.

Esta nova forma organizativa, diz Boulanger, aparece como "une révolution méthodologique et l'un des premiers pas vers la naissance de la linguistique." (BOULANGER, 2002: 14). É, precisamente, com a introdução da ordenação alfabética que se dá uma viragem importante sob um ponto de vista linguístico. Os signos-coisas[16] foram convertidos, pela ordenação alfabética, em signos-palavras (BOULANGER, 2002: 17), um passo extremamente importante, que vai impor-se decisivamente com a invenção da imprensa. A disposição alfabética contribuiu definitivamente para a uniformização dos critérios de indexação das palavras e é uma característica quase imprescindível nas obras que hoje chamamos dicionários.

Se não podemos dizer, em rigor, que um dicionário onomasiológico usa apenas o critério onomasiológico para a sua ordenação (seja ao nível macroestrutural, seja ao nível microestrutural), também o dicionário semasiológico não usa apenas a ordenação alfabética, para a ordenação da sua macroestrutura e microestrutura.

Como refere Haensch, ainda que a maior parte dos dicionários semasiológicos apresente as suas entradas ordenadas por ordem alfabética, há casos em que muitas palavras são agrupadas por famílias, combinando a ordem alfabética e o agrupamento por família (HAENSCH, 1982: 165). Esta combinação de critérios, porém, não é uniforme nem coerente na maior parte dos casos[17]. De resto, uma das desvantagens mais apontadas aos dicionários semasiológicos é, precisamente, a de separar palavras que morfologicamente ou semanticamente deveriam estar juntas.

---

16 Signos-coisas ou, segundo Russel palavras-objecto, i. e., palavras que dependem da nossa experiência do mundo, por oposição às palavras de dicionário, que podem ser definidas através de outras palavras de dicionário (B. Russel (1940). "The Object Language", *in* Allen & Unwin. *An Inquiry into meaning and Truth*, Londres *apud* ECO, 1995: 211-212).

17 Haensch dá o exemplo das palavras '*burgués*' e '*burguesía*' que no dicionário de uso de M. Moliner aparecem juntas, ao contrário do que acontece com as palavras '*aburguesado*', '*abuguesarse*' e '*aburguesamento*' (HAENSCH, 1982: 165).

Para Béjoint, a grande desvantagem na ordenação onomasiológica é a dificuldade da sua utilização, na medida em que, como diz, a organização do conhecimento é variável de autor para autor (BÉJOINT, 2004: 15). Para este autor, a ordenação onomasiológica surge na lexicografia como resposta a duas necessidades. Uma necessidade pedagógica, na medida em que apenas o dicionário onomasiológico fornecia ao usuário ajuda para encontrar uma palavra a partir de uma ideia. A segunda necessidade é ideológica: como refere Béjoint, em certa altura, em algumas sociedades, houve a vontade de pegar em todas as palavras de uma língua e construir com elas uma forma que fizesse sentido (BÉJOINT, 2004: 15). Esta foi, de resto, a ideia dos opositores à ordem alfabética já na Idade Média.

Mas Béjoint, que prefere apontar as vantagens da ordenação alfabética[18], responde a esta crítica explicando que a arbitrariedade da ordenação alfabética é facilmente ultrapassável na lexicografia moderna, uma vez que a ordenação da macroestrutura nos recentes dicionários tem tido em consideração as ligações semânticas na sua macroestrutura. Para além disso, surgiram recentemente as referências cruzadas, que são uma forma de ligar palavras que são semanticamente relacionadas (BÉJOINT, 2004: 17).

A forma que Béjoint encontrou, para valorizar a ordenação alfabética em detrimento da ordenação semântica ou onomasiológica, só abona em favor desta última. O que Béjoint está a dizer-nos é que tem havido, nas obras lexicográficas dos últimos anos, uma grande preocupação pela ordenação onomasiológica. Béjoint faz especial referência aos chamados dicionários combinatórios, desenvolvidos por Mel'čuk e seus colegas em que "each entry-word is the centre of a complex network of syntagmatically and paradigmatically related words" (BÉJOINT, 2004: 17).

Creio que, essencialmente, e do ponto de vista prático, ambos os tipos de ordenação são úteis e servem diferentes propósitos: enquanto os dicionários de orientação semasiológica resolvem o problema da descodificação, os dicionários de orientação onomasiológica têm como principal função codificar, elaborar mensagens. Como refere Fernandéz-Sevilla,

> No se trata, pues, de facilitar los medios para descifrar mensajes, como ha sido usual en la lexicografía tradicional, sino de proporcionar materiales para cifrar, para construir mensajes (FERNÁNDEZ-SEVILLA, 1974: 51)

Por esta mesma razão, talvez, Martínez de Sousa entende como sinónimos 'dicionários ideológicos' e 'dicionários cifradores' ou 'codificadores' (Martínez de Sousa, 1995: **diccionario cifrador**).

O ideal seria um dicionário misto. E, na verdade, muitas das obras concebidas originalmente como dicionários onomasiológicos, como por exemplo o *Thesaurus* de Roget, apresentam nas edições mais modernas um índex alfabético. Landau refere mesmo que este dicionário, na sua edição de 1977, aconselha desde logo o usuário a começar imediatamente pelo índex (LANDAU, 1989: 107). Béjoint vem juntar a este argumento o facto de que muitas das modernas variações do *Thesaurus* de Roget foram totalmente convertidas em ordenação alfabética (BÉJOINT, 2004: 16).

Com a Lexicografia computacional, não faz muito sentido falar de dicionários com ordenação alfabética ou onomasiológica, uma vez que ambas as possibilidades podem ser concebidas.

O que faz sentido, porém, é perguntar se é útil organizar o léxico tendo em conta o critério paradigmático, e, se ele é útil, de que forma essa organização deve ser feita. Este é um dos aspectos que veremos a seguir.

3.2.1 Um exemplo paradigmático de dicionário onomasiológico: o *Sistema de Conceitos* de Hallig & Wartburg

O *Sistema de Conceitos* de Hallig e Wartburg assenta no pressuposto de que é possível construir um produto lexicográfico representativo do "vocabulário como um todo organizado" (HALLIG & WARTBURG, 1963: 77). Como vem expresso na "Introdução", o *Sistema de Conceitos* assenta em dois princípios teóricos da teoria da linguagem de Humbolt:

(i) o princípio de que a língua é mais do que um meio de expressão ou de comunicação, uma vez que ela

> Crée un monde spirituel intermédiaire qui s'insère entre le moi et le monde extérieur, une «image du monde» qui est transmise à chaque représentant d'une communauté linguistique par l'enseignement et confirmée par l'emploi constant qu'il fait de la langue maternelle au cour de son existence. (HALLIG & WARTBURG, 1963: 77)

(ii) o princípio da «articulação», que vê todos os meios de expressão de uma língua como um

18 Para Béjoint, a grande desvantagem na ordenação onomasiológica é a dificuldade da sua utilização, na medida em que, como diz, a organização do conhecimento é variável de autor para autor (BÉJOINT, 2004: 15).

conjunto, "un système dans lequel chaque partie fait corps avec d'autres et est conditionnée par elles" (HALLIG & WARTBURG, 1963: 77-78). Afinal, este mesmo princípio parece encontrar-se já em Saussure quando o autor se refere a família associativa e afirma que um

> dado termo é como que o centro de uma constelação, o ponto para onde convergem os outros termos coordenados, cuja soma é indefinida. (SAUSSURE, 1995: 212)

Partindo destes dois pressupostos, Hallig e Wartburg expõem os quatro preceitos a serem observados, aquando da elaboração de um sistema de classificação, tendo em vista um dicionário descritivo:

(i) apenas os conceitos devem ser classificados;

(ii) estes conceitos que estarão na base do sistema devem ser pré-científicos, i. e., "ceux qui existent dans la langue avant l'introduction de la science" (HALLIG & WARTBURG, 1962: 82). Os conceitos científicos (que provêem das ciências) devem ser limitados e usados apenas quando os conceitos não científicos forem insuficientes;

(iii) é necessário seleccionar apenas alguns conceitos, i. e., a escolha por determinados conceitos em detrimento de outros baseia-se num princípio de economia;

(iv) os conceitos escolhidos serão classificados segundo uma visão de conjunto:

> Le classement doit être tel que le tout constitue un ensemble organisé. Les notions doivent se succéder selon la logique de la vie. Un lien interne doit être, autant que possible, maintenu afin que l'on puisse reconnaître la structure de l'ensemble, le système, la détermination d'une chose par une autre. (HALLIG & WARTBURG, 1962: 82)

E, como defendem os autores, como os conceitos provêem do uso da língua fora da ciência, é possível, através deles, construir uma ideia do mundo que reflicta a linguagem[19].

Aqui põe-se inevitavelmente duas questões fundamentais. A primeira questão prende-se com a natureza daquilo que se entende por conceito. A segunda questão, decorrente da primeira, prende-se com a ordenação desses conceitos.

Relativamente à primeira questão, é necessário saber como é possível representar um conceito. É que ainda que Hallig e Wartburg defendam que é necessário partir dos conceitos, eles afirmam mais adiante, na "Introdução", que

> comme le matériel utilisé pour notre système est emprunté à l'état «préscientifique» de la langue, il faut chaque fois partir du mot (HALLIG & WARTBURG, 1962: 82)

O que os autores nos fazem crer aqui é que uma análise dos conceitos implica partir da palavra, inevitavelmente. Isto porque, como dizem, e tomando a definição de Saussure, uma palavra é composta por duas partes inseparáveis, "um conceito e uma imagem acústica" (SAUSSURE, 1995: 122). Não obstante, o que Hallig e Wartburg tomam em consideração não é, necessariamente, a totalidade do signo linguístico, mas apenas o conceito.

Importa verificar que Hallig e Wartburg não conseguem desligar-se da palavra, a imagem acústica ou significante, inevitavelmente presente, para designar um conceito. Como já havia notado Wolf, os conceitos necessitam de uma língua qualquer para serem entendidos:

> ¿Qué se denomina, en qué lengua, y de qué manera? Ya que, mientras este 'qué', más concretamente el 'concepto? Sólo sigue siendo lo que ya está denominado en la lengua que va a investigarse, se crea un círculo vicioso. (WOLF, 1982: 340)

O melhor modo de se afastarem deste círculo vicioso foi socorrerem-se de uma língua estrangeira, no seu caso o francês. Assim, a língua francesa funciona como uma metalíngua, a partir da qual os ditos conceitos da língua materna dos autores (o alemão) são ordenados. Como os autores explicam em nota de rodapé, citando Trier,

> L'étude du vocabulaire et de son contenu ne peut commencer par un examen du système actuel des concepts de la langue maternelle; car on ne s'en tirerait pas. On doit étudier d'abord un autre système, un système étranger pour se rendre compte des différences et aiguiser son regard. (HALLIG & WARTBURG, 1962: 87)

Como bem notou Wolf, seguindo Heger[20] o conceito que serve de base à Lexicografia deve

19 Este apelo ao senso comum tem sido objecto de estudos por parte de ontologistas nos últimos anos. Ele tem sido reivindicada pela Semântica Cognitiva – ver Lakoff & Mark (1999), Teixeira (2001) – e tem servido de mote para estudos sobre ontologias do senso comum – ver Oltramari & Vetere (2008), Parslow *et al.* (2007).

20 K. Heger (1964). "Die methodologischen Voraussetzungen von Onomasiologie und begrifflicher Gliederung" en *Zeitschrift für romanische Philologie*, 80, pp. 486-516 *apud* WOLF (1982: 340).

exceder pelo menos o marco da língua individual (WOLF, 1982: 341), evitando assim a problemática da existência de conceitos extra-linguísticos.

A questão que se coloca aqui é a de saber que sistema estão os autores a construir: um sistema da língua francesa ou um sistema da língua alemã? Que sistema de conceitos é este? E esta questão é ainda mais pertinente quando sabemos que este *Sistema* que Hallig e Wartburg se propuseram construir tem a intenção de ser universal. De resto, são os próprios autores que colocam a questão: como poderiam representar os conceitos de outra forma que não a linguagem? Os autores descartam os signos extra-linguísticos, que nada têm a ver com a linguagem, com a única explicação de que

> il n'est pas donné á l'homme de faire connaître la pensée et les concepts autrement que par le langage (HALLIG & WARTBURG, 1962: 87-88)

Entendemos agora porque razão os autores defendiam a necessidade de recorrer à palavra para representar um conceito. Mas, porque razão usar um signo linguístico, que já traz arraigado a si uma determinada significação, ou valor, e usá-lo para representar aquilo que Hallig e Wartburg chamam 'conceito' apenas como etiqueta, desprovida de significado ou, nas palavras dos autores, signos " «convertis» d'une «valeur» dans une autre" (HALLIG & WARTBURG, 1962: 88). É como se quisessem partir de um conceito para designar algo que, por sua vez, já tem uma significação distinta daquela que lhe querem dar. Esta é também a crítica que se pode fazer às ontologias computorizadas. Também ali um termo não é um signo linguístico com o seu sentido pleno. É um termo usado para representar uma entidade num domínio (ou na realidade).

Tanto a língua francesa como estes termos das ontologias computacionais pretendem funcionar como aquilo que Wolf chama "*tertium comparationis*", que permite comparar não apenas línguas, mas também subsistemas delimitados por factores cronológicos, geográficos ou sociolinguísticos:

> Sin un 'tertium' siempre se usará una lengua como base de apreciación de otra o un subsistema se usará como punto de partida para apreciar otro, lo qual no cumple las exigencias de la metodología científica. (WOLF, 1982: 343)

Iriarte Sanromán chama a este *tertium* interlíngua e concebe-a como uma representação linguística abstracta, uma linguagem controlada[21] em que os elementos usados para a representação linguística abstracta dos conceitos são os descritores ou palavras-chave (IRIARTE SANROMÁN, 2001: § 4.5).

Resta saber ainda como respondem Hallig e Wartburg ao quarto preceito exigido para a elaboração do seu sistema e que corresponde à nossa pergunta posta acima: como irão os conceitos escolhidos ser classificados, tendo em conta uma visão de conjunto? A resposta a esta questão vem no seguimento do que ficou já esclarecido acima. Hallig e Wartburg defendem que essa resposta é dada no seguimento do que foi dito para os conceitos, i. e., da mesma maneira que apenas se interessam pelos conceitos pré-científicos, também irão adoptar um ponto de vista assente em considerações pré-científicas:

> C'est celui de l'individu moyen, intelligent, qui a une conception du monde fondée sur les concepts préscientifiques que la langue lui offre et qui considère le monde et les hommes avec un réalisme naïf. (HALLIG & WARTBURG, 1962: 88)

Mas que significa 'indivíduo médio'? Como se determina a concepção do mundo do indivíduo médio? Aquilo que Hallig e Wartburg definem como realismo ingénuo assenta numa visão fenomenológica, i.e., "le classement et l'assimilation d'objets empiriques, c'est-a-dire qui tombent sus le coup de l'expérience" (HALLIG & WARTBURG, 1962: 88). Isto significa que é uma visão orientada e limitada pelos objectos empíricos. E se se concebe a existência destes objectos, assume-se também a existência de um mundo exterior objectivo. Esta metodologia, assente na fenomenologia, tem ainda hoje adeptos no campo da ontologia. Esta ênfase posta na experiência encontra o seu argumento na assunção de que existe um tipo de conhecimento que se distingue da língua, e é designado conhecimento ontológico.

Consegue vislumbrar-se aqui alguma crítica a esta visão. Como bem fazem notar os autores do *Sistema de Conceitos*, esta classificação, como todas as classificações, tem um certo grau de subjectividade.

> Tout classement de ce genre est subjectif et conditionnée par les nombreux facteurs qui ont détermine la représentation que se fait son auteur

21 Por linguagem controlada, Iriarte Sanromán entende "um tipo de *linguagem documental* construída *a priori* […] em forma de *thesaurus de descritores* (listagem estruturada de conceitos) (IRIARTE SANROMÁN, 2001: § 4.5, nota 180).

> du monde et de la vie. (HALLIG & WARTBURG, 1962: 88)

Hallig e Wartburg terminam a "Introdução" garantindo que a pertinência da sua obra será julgada no uso. Assim se justifica qualquer imperfeição ou incoerência teórica com o pragmatismo.

3.3. Dos apontamentos teóricos à realidade prática

Os produtos lexicográficos onomasiológicos, para além das questões teóricas que foram abordadas acima, levantam também questões práticas. Baldinger aponta seis questões de aplicação prática ao Sistema de Hallig e Wartburg: (i) a hierarquia conceptual, (ii) a classificação lógica e associativa, (iii) a diferença entre língua geral e língua especializada, (iv) classificação científica e popular, (v) o carácter supranacional e (vi) o carácter supratemporal. As implicações práticas das questões (ii), (iii) e (iv) vão desembocar todas na mesma questão (i) e têm provavelmente a mesma resposta prática numa ontologia computorizada dos dias que correm. As questões (v) e (vi) foram tratadas no ponto anterior. Baldinger afirma que surgem muitas dificuldades quando se procede à tentativa de ordenar a totalidade numa hierarquia conceptual. O autor afirma que "nem na realidade nem na língua se dá uma divisão hierárquica total" (BALDINGER, 1977: 127). Isto torna-se evidente na dificuldade da elaboração prática de um sistema hierárquico conceptual pois, como refere Baldinger, partindo da análise do *Sistema* de Hallig e Wartburg, na prática, um sistema conceptual só pode dividir-se unilateralmente e não multilateralmente. Tomemos um dos exemplos apresentados por Baldinger:

> Tomemos el concepto *enfermedad*. Los hombres, los animales, incluso las plantas, pueden estar enfermos. Pero en el sistema conceptual de Hallig/Wartburg, las plantas se encuentran en A III, los animales en A IV y el hombre en B. Por eso, el concepto de enfermedad debe ser descompuesto en el sistema conceptual, porque el sistema de Hallig y Wartburg está concebido desde la contraposición Universo-Hombre. (BALDINGER, 1977: 127)

'*Enfermedad*' relaciona-se quer com as instâncias que pertencem a A, como com as instâncias que pertencem a B. Esta mesma questão foi levantada por Eco (ECO, 1995) no seu livro *A procura da Língua Perfeita*, que analisa as diferentes propostas de organização de conteúdo de línguas *a priori*, propostas por autores ao longo da história. No capítulo dedicado a John Wilkins, aquando da análise do organigrama representativo da tábua do mundo, Eco faz notar que a oposição VEGETATIVO/SENSITIVO, na tábua dos géneros, também aparece duas vezes. Como refere Eco, e como vemos na Ilustração I, a seguir, se a árvore de Wilkins ou o esquema de Hallig e Wartburg tivessem uma consistência lógica, se a intenção era uma organização conceptual, em que "cada entidade sua seja inequivocamente definida pelo lugar que ocupa na árvore geral das coisas" (ECO, 1995: 241), todas as instâncias que pertencem às classes que são subclasses de outras classes superiores, tinham, obrigatoriamente de pertencer a essas classes superiores.

Na Ilustração I, as instâncias das subclasses da classe VEGETATIVO que, por sua vez, são subclasses da classe CORPÓREO, pertenceriam também à classe ESPIRITUAL. Mas não é isso que se interpreta no esquema da tábua do mundo de Wilkins. As instâncias das classes MINERAIS, ERVAS e PLANTAS não fazem parte da classe ESPIRITUAL. Como bem refere Eco, a forma como a classe VEGETATIVO é entendida, no esquema onde pertence à classe ESPIRITUAL, é diferente da forma como é entendida no esquema da classe CORPÓREO.

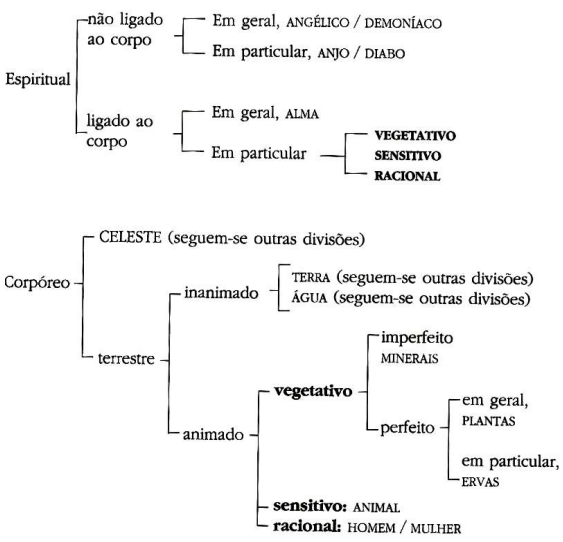

Ilustração I: Esquema da tábua do mundo de Wilkins (in ECO, 1995: 241)

Eco sublinha que estas subdivisões "são como os capítulos de uma grande enciclopédia capaz de reconsiderar a mesma coisa de diferentes pontos de vista" (ECO, 1995: 242). Um sistema que permitisse relações multilaterais, que representasse conceitos sobre diferentes pontos de vista responderia à necessidade de distinguir entre língua geral e língua especializada e, por consequência,

uma classificação científica e popular. De resto, é isso que acontece hoje com os dicionários de língua, que vão organizar as acepções tendo em conta as áreas de especialidade.

4. Dos dicionários onomasiológicos às ontologias

Nos dias de hoje a possibilidade de representar um conceito multilateralmente é possível com a noção de hipertexto e com as ferramentas informáticas que temos à nossa disposição. Tomando as palavras de Eco,

> Pode conceber-se um hipertexto sobre os animais que, a partir de 'cão', dê acesso a uma classificação geral dos mamíferos e insira o cão numa árvore de taxa que contenha igualmente o gato, o boi e o lobo. Mas, a partir desse nó, poderemos ser remetidos para um repertório acerca das propriedades do cão, e dos seus hábitos, e seleccionando uma outra ordem de informações poderemos ter acesso a uma resenha dos diversos papéis desempenhados pelo cão em diversas épocas históricas […], ou a um rol das imagens do cão na história da arte. (ECO, 1995: 243)

Passamos a ter uma rede de relações múltiplas, e já não relações hierárquicas. Mas nada saberíamos dessas relações. Elas seriam inferidas pela nossa capacidade cognitiva, mas não seriam relações explícitas e de nada serviriam para a representação do conhecimento. De resto, nesta rede de relações, perde-se o esquema global. E é este esquema global, esta plataforma de integração, que permite a compatibilidade entre, por exemplo, diferentes perspectivas científicas sobre um mesmo objecto, que constitui a base de uma ontologia. Este é o caso, por exemplo da BFO[22] (*Basic Formal Ontology*) (), que se constitui como uma ontologia de nível superior de suporte às ciências naturais, capaz de agregar diferentes ontologias de domínio específico.

Como vimos num artigo anterior (FRANÇA, 2009b), o conceito de ontologia assenta grandemente na explicitação das suas relações. Se olharmos para a definição tradicional de ontologia no seio da Filosofia, veremos que é assim:

> se a ontologia não quiser negar o carácter real da multiplicidade ôntica, ela terá de a obter através de uma síntese gradual, de uma construção progressiva, que deve a pouco e pouco **recompor o real segundo uma ordem, que estabeleça um laço de dependência e uma hierarquia entre os elementos componentes, dos mais simples aos mais complexos**.(BLANC, 1998: 49) [sublinhado meu]

Mas esta ordem, a que Blanc se refere parece sugerir uma divisão unilateral, assente na exclusão bipartida da taxonomia, baseada na estrutura lógica aristotélica do *genus proximum* e *differentia specifica,* que pode ser exemplificada pelo esquema seguinte:

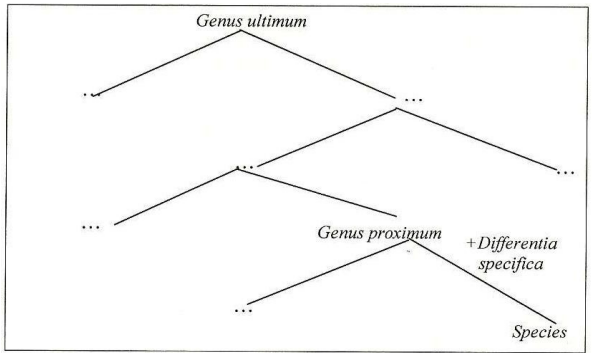

Ilustração II: A estrutura da árvore de Porfírio (in JANSEN, 2008: 164).

Mas a construção desta ordem não implica, inevitavelmente, uma divisão unilateral ou sequer a utilização de um sistema lógico. Sowa, por exemplo, defende que as ontologias computorizadas podem ser axiomatizadas ou baseadas na teoria dos protótipos[23] sem deixar de ser formais, i.e., sem perder o seu poder de representação:

> an *axiomatized ontology* distinguishes subtypes by axioms and definitions stated in a formal language, such as logic or some computer-oriented notation that can be translated to logic; a *prototype-based ontology* distinguishes subtypes by a comparison with a typical member or *prototype* for each subtype. (SOWA, s.d.[a])

Como refere Sowa, as grandes ontologias podem usar os dois tipos de métodos em que os axiomas formais e definições são usados, por exemplo, para termos da Matemática, da Física

22 http://www.ifomis.org/bfo/home

23 Para uma introdução à teoria do protótipo vd. Cuenca, M. J. & J. Hilferty, 1999. Sowa dá um exemplo de como a teoria dos protótipos pode ser usada numa ontologia:

> a black cat and an orange cat would be considered very similar as instances of the category Animal, since their common catlike properties would be the most significant for distinguishing them from other kinds of animals. But in the category Cat, they would share their catlike properties with all the other kinds of cats, and the difference in color would be more significant. In the category BlackEntity, color would be the most relevant property, and the black cat would be closer to a crow or a lump of coal than to the orange cat. (SOWA, s.d. [b])

(outras áreas especializadas)[24], ou para as categorias do nível superior, enquanto que os protótipos são usados para itens comuns, para os níveis inferiores de uma ontologia (SOWA, s.d.[b]). Sowa chama a estas ontologias ontologias mistas. Esta possibilidade iria garantir a representação, por exemplo, daquilo que Hirst descreve como exemplos de quase-sinónimos (HIRST, 2004: 216-21) e da polissemia.

5. Conclusão

Uma análise mais demorada pelo conceito de onomasiologia e pelos produtos da Lexicografia onomasiológica, nomeadamente uma atenção aos problemas práticos e teóricos que essas obras mereceram há quase um século atrás, pode revelar-se de importância crucial para a elaboração das ontologias modernas e computorizadas. Alguns dos pressupostos teóricos que serviram de base à Lexicografia onomasiológicas, como por exemplo a Semântica pré-estrutural analítico-referencial, com nomes como Humbolt e Hallig e Wartburg e os estudos sobre os campos semânticos e o conceito de onomasiologia, continuam, ainda hoje, a exercer influência na Semântica actual, em autores que se inserem já dentro do paradigma cognitivo, como Geeraerts (EERAERTS, 2006).

É necessário, como afirmou Smith, não reinventar a roda. Muitos dos desafios que se colocam hoje a um ontologista têm vindo a colocar-se desde, pelo menos, há dois milénios. Há questões que continuam as mesmas.

As ontologias computorizadas têm hoje um papel fundamental. Elas são instrumentos de trabalho *sine qua non* dos ontologistas hoje. São elas que permitem testar hipóteses e, como refere Johansson (JOHANSSON, 2008: 302), estas novas ferramentas trazem também novas formas de olhar os problemas, novas perguntas e novas soluções.

A elaboração de uma ontologia, tal como a elaboração de um dicionário, é uma tarefa essencialmente técnica e prática e é possível construí-la sem recurso a reflexões teóricas como as que ficaram expostas neste artigo. Quando este cuidado não é tido em consideração, o trabalho da engenharia corre sempre o risco de sofrer as mesmas críticas que durante muito tempo foram feitas à Lexicografia e que podemos resumir nas palavras de Wierzbicka:

> It has often been said that lexicographers are people who work hard but who can never escape having a guilty conscience, because lexicography has no theoretical foundations, and even the best lexicographers, when pressed, can never explain what they are doing or why (WIERZBICKA, 1995: 3)

24 Aqui parte-se do pressuposto que a terminologia de domínios técnicos apresenta menor ambiguidade; é mais precisa e clara. Como refere Hirst (2004: 222), "em alguns campos de estudo há uma autoridade reconhecida que mantém e publica uma categorização e a sua nomenclartura". É certo que isto não acontece em todas as áreas e domínios técnicos, mas é desejável que assim seja.

Referências

Arano, S. 2005. "*Thesauruses and ontologies" in Hipertext.net* [revista electrónica com o endereço: http://www.hipertext.net, num. 3, 2005. Disponível em http://www.hipertext.net/english/pag1009.htm [consult. 17-05-2009];

Babini, M.. 2006. "Do conceito à palavra: os dicionários onomasiológicos", *Revista Ciência e Cultura*, v. 58, n.2, São Paulo Abr./Jun. 2006, pp. 38-42. Disponível em http://cienciaecultura.bvs.br/pdf/cic/v58n2/a15v58n2.pdf [consult. 22-08-2009];

Baldinger, K.. 1977. *Teoria Semántica*. Madrid: Ediciones Alcala. ISBN: 84-7008-010-5;

Béjoint, H.. 2004. *Modern lexicography: an introduction*. Oxford: University Press. ISBN: 0-19-829951-6;

Blanc, M.. 1997. *Introdução à Ontologia*. Lisboa: Instituto Piaget. ISBN: 972-8407-67-X;

Boulanger, J.- C.. 2002. "Petit histoire de la conquête de l'ordre alphabétique dans les dictionnaires médiévaux" *in Cahiers de lexicologie: revue internationale de lexicologie et lexicographie*. Vol. 80, 2002. Paris: Didier Erudition, pp. 9-24. ISNN: 0007-9871;

Casares, J..1992. *Introducción a la Lexicografia Moderna*, Madrid: Consejo Superior de Investigaciones Científicas. ISBN: 84-00-07233-2;

Cuenca, M. J. & J. Hilferty. 1999. *Introducción a la lingüística cognitiva.* Barcelona: Editorial Ariel. ISBN: 84-344-8234-7;

Dubois, J. *et al*.. 1979. *Dicionário de Linguística (10ª edição)*. São Paulo: Editora Cultrix. ISBN: 85-316-0123-1;

Eco, U.. 1995. *A Procura da Língua Perfeita*. Lisboa: Editorial Presença. ISBN: 972-23-1996-5;

França, P.. 2009a. *Ontologia e ontologias: contributos teóricos para uma perspectiva transdisciplinar*. Tese de Mestrado. Braga: Universidade do Minho;

França, P. 2009b. "Conceitos, classes e/ou universais: com o que é que se constrói uma ontologia?" in *LinguaMática*, nº 1 - Maio 2009. ISSN: 1647-0818. Disponível em http://www.linguamatica.com/index.php/linguamatica;

Geeraerts, D.. 2006. *Words and other Wonders*. *Papers on Lexical and Semantic Topics.* Berlin/New York: Mouton de Gruyter. ISBN-13: 978-3-11-019042-7;

Green, J.. 1996. *Chasing the sun: dictionary makers and the dictionaries they made*. New York: Henry Holt and Company, Inc. ISBN: 0-8050-3466-8;

Grzega, J. & M. Schöner. 2007. *English and General Historical Lexicology*. Eichstätt: Katholische Universität Eichstätt-Ingolstadt. Disponível em http://www1.ku-eichstaett.de/SLF/EngluVglSW/OnOnMon1.pdf [consult. 31-05-2009];

Grzega, J.. 2002. "Some aspects of modern diachronic onomasiology" *in Linguistics*. Volume 40, Issue 5, pp. 1021–1045. Jul 2002. Berlin: Walter de Gruyer. Disponível em http://www.reference-global.com/doi/pdf/10.1515/ling.2002.035?cookieSet=1 [consult. 31-05-2009];

Haensch, G.. 1982. "Tipología de las obras lexicográficas" *in* Haensch, Wolf, Ettinger & Werner. 1982, pp. 95-187;

Haensch, Wolf, Ettinger & Werner. 1982. *La lexicografía. De la lingüística teórica a la lexicografía práctica.* Madrid: Gredos. ISBN: 84-249-0858-9;

Hallig, R. & W. Wartburg. 1963. *Begriffssystem als Grundlage für die Lexikographie / Système Raisonné des Concepts pour Servir de Base à la Lexicographie*. Berlin: Akademie-Verlag;

Hirst, G.. 2004. "Ontology and the lexicon" *in* Staab, S. & Studer, R.. 2004. *Handbook on ontologies*. Berlin: Springer. ISBN: 3-540-40834-7. Disponível em http://ftp.cs.toronto.edu/pub/gh/Hirst-Ontol-2003.pdf [cons. 20-09.09];

Hüllen, W. 1999. *English Dictionaries, 800-1700. The topical tradition*. Oxford: Oxford University Press. ISBN: 0-19-929104-7;

Iriarte Sanromán, Á.. 2001. A Unidade Lexicográfica. Palavras, Colocações, Frasemas, Pragmatemas. Braga: Centro de Estudos Humanísticos-Universidade do Minho. Disponível em https://repositorium.sdum.uminho.pt/bitstream/1822/4573/1/A_Unidade_Lexicografica.pdf [consult. 22-04-2007];

Jansen, L.. 2008 "Categories: The Top-Level Ontology" *in* Munn & Smith (eds.). 2008. pp. 173- 196;

Johansson, I.. 2008. "Bioinformatics and Biological Reality" *in* Munn & Smith (eds.). 2008;

Lacy, L..2005. *Owl: Representing Information Using the Web Ontology Language*. UK: Trafford. ISBN: 141203448-5;

Lakoff, G. & J. Mark. 1999. *Philosophy in the Flesh: the embodied mind and its challenges to the western thought*. New York: Basic Books. ISBN: 0-465-05674-1;

Magnini, B. & M. Speranza. 2002 "Merging Global and Specialized Linguistic Ontologies" In *Proceedings of the Workshop Ontolex-2002 Ontologies and Lexical Knowledge Bases*, LREC-2002, pp. 43-48. Disponível em http://multiwordnet.fbk.eu/paper/ontomerge-ontolex.pdf [cons. 20-09-09];

Martínez de Sousa, J. . 1995. *Diccionario de Lexicografía Prática*. Barcelona: Vox. ISBN: 84-7153-803-2;

Moreira, A., L. Alvarenga & A. Oliveira. 2004. "O nível do conhecimento e os instrumentos de representação: tesuarus e ontologias" *in* data GramaZero – Revista de Ciência da Informação – vol. 5, nº 6, Dezembro de 2004. Disponível em http://usuarios.cultura.com.br/eds/PDF/fasam.pdf [consult. 20-02-2009];

Munn, K. & B. Smith (eds.). 2008. *Applied Ontology. An Introduction*. Frankfurt/Paris/Lancaster/New Brunswick: Ontos Verlag. ISBN: 978-3-938793-98-5;

Nickles, M. *et al.*.2007. "Ontologies across disciplines" *in* Schalley, A. & D. ZaeffererR (eds.). 2007. *Ontolinguistics – How Ontological Status Shapes the Linguistic Coding of Concepts*, Berlin/New York: Mouton de Gruyer. ISBN: 978-3-11-018997-1;

Oltramari, A. & Vetere, G.. 2008. "Lexicon and Ontology Interplay in Senso Comune" *in Proceedings of OntoLex 2008* (Hosted by Sixth international conference on Language Resources

and Evaluation), Marrakech (Morocco). Disponível em http://www.loa-cnr.it/Papers/lexicon_oltramari-vetere.pdf [cons. 07-07-09];

Parslow, P. *et al.*. 2007. "Folksonomological Reification". Book chapter submittet and accepted to Social Software and Developing Community Ontologies Book. Disponível em http://www.lulu.com/items/volume_64/6043000/6043166/2/print/6043166.pdf [consult. 31-05-2009];

Saussure, F.. 1995. *Curso de Linguística Geral*. Lisboa: Publicações D. Quixote. ISBN: 972200056-x;

Smith, B.. 1998. "The Basic Tools of Formal Ontology" *in* N. Guarino (ed.). 1998. *Formal Ontology in Information Systems*. Amsterdam/Oxford/Washington DC: IOS Press, pp. 19-28. Disponível em http://ontology.buffalo.edu/smith/articles/fois1998.pdf [31-08-2008] ;

Sowa, J.. s.d.[a] "Ontology" Disponível em http://www.jfsowa.com/ontology/index.htm [consult. 07-09-2009];

Sowa, J.. s.d.[b.] "Glossary" Disponível em http://www.jfsowa.com/ontology/gloss.htm [consult. 07-09-2009];

Teixeira, J.. 2001. *A Verbalização do Espaço: modelos mentais de* frente/trás. Braga: Universidade do Minho, Centro de Estudos Humanísticos. ISBN: 972-98621-4-1;

Wielinga *et al.*.2001. "From thesaurus to ontology" *in International Conference On Knowledge Capture, Proceedings of the 1st international conference on Knowledge capture*, Victoria, British Columbia, Canada, pp.: 194 - 201. Disponível em http://www.cs.vu.nl/~guus/papers/Wielinga01a.pdf [11-05- 2009];

Wierzbicka, A.. 1995. *Lexicography and Conceptual Analysis*. S. l.: Karoma Publishers, Inc.. ISBN: 0-89720-069-1;

Wolf, L.. 1982. "Signo lingüístico y estruturas semânticas" *in* Haensch, Wolf, Ettinger & Werner (1982), pp. 329-358.

Chamada de Artigos

A revista Linguamática pretende colmatar uma lacuna na comunidade de processamento de linguagem natural para as línguas ibéricas. Deste modo, serão publicados artigos que visem o processamento de alguma destas línguas.

A Linguamática é uma revista completamente aberta. Os artigos serão publicados de forma electrónica e disponibilizados abertamente para toda a comunidade científica sob licença *Creative Commons*.

Tópicos de interesse:

- Morfologia, sintaxe e semântica computacional
- Tradução automática e ferramentas de auxílio à tradução
- Terminologia e lexicografia computacional
- Síntese e reconhecimento de fala
- Recolha de informação
- Resposta automática a perguntas
- Linguística com corpora
- Bibliotecas digitais
- Avaliação de sistemas de processamento de linguagem natural
- Ferramentas e recursos públicos ou partilháveis
- Serviços linguísticos na rede
- Ontologias e representação do conhecimento
- Métodos estatísticos aplicados à língua
- Ferramentas de apoio ao ensino das línguas

Os artigos devem ser enviados em PDF através do sistema electrónico da revista. Embora o número de páginas dos artigos seja flexível sugere-se que não excedam 20 páginas. Os artigos devem ser devidamente identificados. Do mesmo modo, os comentários dos membros do comité científico serão devidamente assinados.

Em relação à língua usada para a escrita do artigo, sugere-se o uso de português, galego, castelhano ou catalão.

Os artigos devem seguir o formato gráfico da revista. Existem modelos LaTeX, Microsoft Word e OpenOffice.org na página da Linguamática.

Datas Importantes

- Envio de artigos até: 31 de Março de 2010
- Resultados da selecção até: 15 de Abril de 2010
- Versão final até: 30 de Abril de 2010
- Publicação da revista: 15 de Maio de 2010

Qualquer questão deve ser endereçada a: `editores@linguamatica.com`

Petición de Artigos

A revista Linguamática pretende cubrir unha lagoa na comunidade de procesamento de linguaxe natural para as linguas ibéricas. Deste xeito, han ser publicados artigos que traten o procesamento de calquera destas linguas.

Linguamática é unha revista completamente aberta. Os artigos publicaranse de forma electrónica e estarán ao libre dispor de toda a comunidade científica con licenza *Creative Commons*.

Temas de interese:

- Morfoloxía, sintaxe e semántica computacional
- Tradución automática e ferramentas de axuda á tradución
- Terminoloxía e lexicografía computacional
- Síntese e recoñecemento de fala
- Extracción de información
- Resposta automática a preguntas
- Lingüística de corpus
- Bibliotecas dixitais
- Avaliación de sistemas de procesamento de linguaxe natural
- Ferramentas e recursos públicos ou cooperativos
- Servizos lingüísticos na rede
- Ontoloxías e representación do coñecemento
- Métodos estatísticos aplicados á lingua
- Ferramentas de apoio ao ensino das linguas

Os artigos deben de enviarse en PDF mediante o sistema electrónico da revista. Aínda que o número de páxinas dos artigos sexa flexíbel suxírese que non excedan as 20 páxinas. Os artigos teñen que identificarse debidamente. Do mesmo modo, os comentarios dos membros do comité científico serán debidamente asinados.

En relación á lingua usada para a escrita do artigo, suxírese o uso de portugués, galego, castelán ou catalán.

Os artigos teñen que seguir o formato gráfico da revista. Existen modelos LaTeX, Microsoft Word e OpenOffice.org na páxina de Linguamática.

Datas Importantes

- Envío de artigos até: 31 de marzo de 2010
- Resultados da selección: 15 de abril de 2010
- Versión final: 30 de abril de 2010
- Publicación da revista: 15 de maio de 2010

Para calquera cuestión, pode dirixirse a: `editores@linguamatica.com`

Petición de Artículos

La revista Linguamática pretende cubrir una laguna en la comunidad de procesamiento del lenguaje natural para las lenguas ibéricas. Con este fin, se publicarán artículos que traten el procesamiento de cualquiera de estas lenguas.

Linguamática es una revista completamente abierta. Los artículos se publicarán de forma electrónica y se pondrán a libre disposición de toda la comunidad científica con licencia *Creative Commons*.

Temas de interés:

- Morfología, sintaxis y semántica computacional
- Traducción automática y herramientas de ayuda a la traducción
- Terminología y lexicografía computacional
- Síntesis y reconocimiento del habla
- Extracción de información
- Respuesta automática a preguntas
- Lingüística de corpus
- Bibliotecas digitales
- Evaluación de sistemas de procesamiento del linguage natural
- Herramientas y recursos públicos o cooperativos
- Servicios lingüísticos en la red
- Ontologías y representación del conocimiento
- Métodos estadísticos aplicados a la lengua
- Herramientas de apoyo para la enseñanza de lenguas

Los artículos tienen que enviarse en PDF mediante el sistema electrónico de la revista. Aunque el número de páginas de los artículos sea flexible, se sugiere que no excedan las 20 páginas. Los artículos tienen que identificarse debidamente. Del mismo modo, los comentarios de los miembros del comité científico serán debidamente firmados.

En relación a la lengua usada para la escritura del artículo, se sugiere el uso del portugués, gallego, castellano o catalán.

Los artículos tienen que seguir el formato gráfico de la revista. Existen modelos LaTeX, Microsoft Word y OpenOffice.org en la página de Linguamática.

Fechas Importantes

- Envío de artículos hasta: 31 de marzo de 2010
- Resultados de la selección: 15 de abril de 2010
- Versión final: 30 de abril de 2010
- Publicación de la revista: 15 de mayo de 2010

Para cualquier cuestión, puede dirigirse a: `editores@linguamatica.com`

Petició d'articles

La revista Linguamática pretén cobrir una llacuna en la comunitat del processament de llenguatge natural per a les llengües ibèriques. Així, es publicaran articles que tractin el processament de qualsevol d'aquestes llengües.

Linguamática és una revista completament oberta. Els articles es publicaran de forma electrònica i es distribuiran lliurement per a tota la comunitat científica amb llicència *Creative Commons*.

Temes d'interès:

- Morfologia, sintaxi i semàntica computacional
- Traducció automàtica i eines d'ajuda a la traducció
- Terminologia i lexicografia computacional
- Síntesi i reconeixement de parla
- Extracció d'informació
- Resposta automàtica a preguntes
- Lingüística de corpus
- Biblioteques digitals
- Evaluació de sistemes de processament del llenguatge natural
- Eines i recursos lingüístics públics o cooperatius
- Serveis lingüístics en xarxa
- Ontologies i representació del coneixement
- Mètodes estadístics aplicats a la llengua
- Eines d'ajut per a l'ensenyament de llengües

Els articles s'han d'enviar en PDF mitjançant el sistema electrònic de la revista. Tot i que el nombre de pàgines dels articles sigui flexible es suggereix que no ultrapassin les 20 pàgines. Els articles s'han d'identificar degudament. Igualement, els comentaris dels membres del comitè científic seràn degudament signats.

En relació a la llengua usada per l'escriptura de l'article, es suggereix l'ús del portuguès, gallec, castellà o català.

Els articles han de seguir el format gràfic de la revista. Es poden trobar models LaTeX, Microsoft Word i OpenOffice.org a la pàgina de Linguamática.

Dades Importants

- Enviament d'articles fins a: 31 de març de 2010
- Resultats de la selecció: 15 d'abril de 2010
- Versió final: 30 d'abril de 2010
- Publicació de la revista: 15 de maig de 2010

Per a qualsevol qüestió, pot adreçar-se a: `editores@linguamatica.com`