Definición y análisis de parámetros lingüísticos cuantitativos para herramientas automáticas de evaluación aplicables al Español como Lengua Extranjera

Paz Ferrero García de Jalón
Departamento de Filología Inglesa

Universidad Autónoma de Madrid

Memoria de Tesis Doctoral dirigida por
Rachel Whittaker y Teresa Bordón

Abril, 2011

A mi Odiseo y a nuestro retoño,
a quienes embarqué en esta aventura.

Agradecimientos

Durante mi periplo se han atado muchos cabos y, aunque algunos cantos de sirena me han desviado de la ruta, he podido retomar el rumbo gracias a mi familia y a mis bi-guías. Mientras, mis mayores, hermanos y amigos han esperado pacientes un lustro hasta que concluyese mi labor para reencontrarnos. Han sido mis fehacientes estudiantes los que me impulsaron a tender las velas de nuevo para llegar a puerto.

Mi Odiseo y retoño: Javier e Irene.
Mis bi-guías de la Universidad Autónoma de Madrid: Rachel y Teresa.
Mis mayores: mi madre Purificación, mi padre Eusebio, Santiago y Quiteria.
Mis hermanos: Puri, Use y Espe.
Mis amigos: Amber, Birgit, CarmenC, Caroline, Celia, Constanza, Elena, Estelle, Françoise, Gloria, Hala, Humi, Idel, Jenny, Josephine, Kari, Lara, Maggie, Malika, Marian, Maribel, May, MónicaD, MónicaP, Montse, Nuria, Paquita, Presen, Rosa, Sharon, Sutililla, Sylvia, Yolanda,...
Mis fehacientes estudiantes de la Universidad Popular de Alcobendas.

Siempre estaré muy agradecida a todos con quienes he compartido esta laboriosa travesía y recordaré con cariño a quienes han sido mis maestros y facilitadores.

Mis maestros y facilitadores: Sor María Luisa Abad, Isidro Aguillo, Javier Barreiro, Glenn Boreman, Elena Delgado, Mariví Espeso, Daniel Essig, Ana Fernández-Pampillón, James Ginn, Carmen de la Guardia, Leticia Herrero, Laura Hidalgo, Peter Kincaid, Rosario López, Clara Molina, Inmaculada Molina, Javier Ortiz, Lluís Padró, José Ramón Parrondo, Manuela Romano, Rosendo Tello, Rachel Whittaker, Fran Yoshiyama,...

.


TEST
EVALUATOR
GRADE
REPORT

Resumen

En este estudio, se ha desarrollado una serie de métodos automáticos para evaluar el nivel de competencia de textos escritos en castellano, según los seis niveles de referencia definidos por el Marco Común Europeo de Referencia para las Lenguas: Aprendizaje, Enseñanza, Evaluación (MCERL): A1-A2, B1-B2, C1-C2. Consideramos de forma separada las áreas léxica, sintáctica y semántica. Junto con el nivel de referencia se definen parámetros de confianza que miden la fiabilidad de la determinación del nivel.

La evaluación de la competencia léxica se basa en la clasificación de las palabras dada en el inventario del Plan Curricular del Instituto Cervantes (PCIC), considerándolo como la norma para este estudio. Para aquellas palabras no incluidas en el listado del PCIC, se han desarrollado otros criterios basados en la presencia de tales palabras en una combinación de varios glosarios. Además, las multipalabras son evaluadas para refinar la determinación del nivel del texto. Finalmente, el conjunto del texto se nivela cuantitativamente aplicando la ley de Zipf después de distribuir las palabras en los seis niveles del MCERL.

La evaluación del nivel sintáctico se calcula mediante la identificación, en el texto objeto de estudio, de estructuras sintácticas previamente niveladas. Para ello, se ha creado una numerosa colección de estructuras sintácticas niveladas basándonos en el PCIC. El nivel sintáctico de cada estructura se ha dado en función de la complejidad de la estructura. Estas estructuras son identificadas en el texto bajo análisis y los resultados cuantitativos de esta identificación se comparan frente a un corpus de referencia de un determinado nivel.

En el nivel semántico identificamos el contenido de un texto teniendo en cuenta los campos semánticos dados por el PCIC. Por ello, utilizando un conjunto apropiado de textos, es posible correlar cuantitativamente los campos semánticos del texto sometido al análisis con los del corpus. Es más, el Análisis Semántico Latente evidencia el agrupamiento de los textos en conjuntos que pueden ser fácilmente identificados y utilizados para la evaluación del nivel semántico.

La calificación automática léxica, sintáctica y semántica es posible tras procesar el texto mediante un analizador morfosintáctico (FreeLing). Los métodos desarrollados en esta tesis se han probado con 80 textos escritos por aprendices de español como segunda lengua con el fin de obtener el Diploma de Español como Lengua Extranjera (DELE). Estos textos ya han sido previamente calificados por personas evaluadoras oficiales. Los resultados obtenidos mediante el calificador automático muestran una buena correlación con los resultados dados por los evaluadores humanos.

Definition and analysis of quantitative linguistic parameters for automatic assessment tools applicable to Spanish as a Foreign Language

Abstract

In this study, a number of automatic methods have been developed to evaluate quantitatively the reference level of texts written in Spanish according to the six main levels given by the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFRL): A1-A2, B1-B2, C1-C2. Lexical, syntactic, and semantic areas are considered separately. The reference level goes with a confidence parameter that measures the reliability of the determination of the level.

The evaluation of lexical competence is based on the classification of lexical level given by the Plan Curricular del Instituto Cervantes (PCIC), as the norm for the study. For words not included in the PCIC list, other criteria have been developed, based on the presence of those words in a combination of selected glossaries. Moreover, multiword units are also evaluated to refine the determination of the level of the text. After this, the whole text is given a quantitative level by applying Zipf's law to the six levels of the CEFRL.

The evaluation of the syntactic level is calculated by the identification of categorized syntactic structures within the text under test. A wide collection of syntactic structures has been created and leveled based on the PCIC. The syntactic level has been given to these structures according to the complexity of the structure. These structures are identified in the text being tested and the quantitative result of this identification is compared against a reference corpus with a predetermined level.

At the semantic level, we identify the content of a text taking into account the semantic fields given by the PCIC. Then, after using customized corpora, it is possible to correlate quantitatively the semantic fields of the text being tested with the corpora. Besides this, Latent Semantic Analysis (LSA) provides evidence of the grouping of each text in semantic clusters that can be easily identified and used for the evaluation of the semantic level.

The lexical, syntactic, and semantic automatic assessments are possible after processing the text with a morphosyntactic analyzer (FreeLing). The methods developed have been tested on 80 texts written by learners of Spanish as a second language to achieve the Diploma de Español as Lengua Extranjera (DELE). These texts had been previously scored by official human evaluators. The results obtained using the automatic graders correlate well with the scores given by the human graders.

Índice general

Capítulo 1

Introducción

La automatización de procesos sistemáticos es una práctica para realizar tareas rutinarias de forma eficaz y obtener resultados inmediatos. La necesidad de automatizar tareas cotidianas en todos los ámbitos, incluso en el educativo, es una realidad. Concretamente, la investigación del Procesamiento del Lenguaje Natural (PLN) y el tratamiento automático de textos se lleva ya desarrollando con una gran actividad desde los años noventa, como muestra cada conferencia anual de Text REtrieval Conference (Voorhees *et al.*, 2005). En consecuencia, el procesado de textos mediante métodos estadísticos y herramientas computacionales ha facilitado el estudio y la obtención de resultados a gran escala y de forma casi inmediata. Este hecho ha ido generando productos comerciales muy interesantes. Además, se suma a esta actividad el interés de instituciones, investigadores y profesores por este tema. Por un lado, se aúnan esfuerzos por reagrupar y reutilizar herramientas y *corpora* como propone el grupo Common Language Resources and Technology Infraestructure (CLARIN) y, por otro, instituciones privadas compiten por describir y desarrollar criterios y métodos de evaluación computables, específicos y significativos, al tiempo que válidos y fiables para automatizar el aprendizaje y la autoevaluación.

Sin duda, la posibilidad de disponer de herramientas de autoevaluación o evaluación automática de textos ha generado nuevas áreas de trabajo e investigación en el área de lingüística y, más específicamente, en lingüística computacional. Entre los pioneros y patrocinadores de estos productos, en su mayoría comerciales, destacan entidades americanas como el Educational Testing Service (ETS) o las nuevas propuestas de exámenes del Test of English as a Foreign Language (TOEFL) como el New TOEFL. Incluso, cada día más, se crean espacios virtuales de trabajo con métodos y mecanismos de aprendizaje que contribuyen a integrar el proceso de escritura, lectura y evaluación para favorecer el autoaprendizaje, individual y colaborativo.

1.1. Revisión del tema

Existe un gran interés multidisciplinar por la automatización de procesos relacionados con la extracción automática de cantidad ingente de información para su posterior distribución, almacenamiento o catalogación (Fernández-Pampillón, 2010). No menos interés

suscita la posibilidad de evaluar textos escritos por estudiantes de idiomas y de proporcionar al estudiante herramientas para desarrollar su destreza durante su aprendizaje. El proceso de escritura automatizado, es decir, asistido durante el proceso de creación, corrección y evaluación, es una realidad.

Por un lado, se están desarrollando espacios virtuales para gestionar y registrar el proceso de composición combinado con la corrección de textos. Para ello, entidades privadas e instituciones académicas norteamericanas invierten en dicha investigación para crear productos comerciales como AutoTutor, entorno de aprendizaje que proporciona retroalimentación a los aprendices de forma inmediata (Graesser, 2008), o W-pal, entorno práctico asistido de escritura desarrollado en 2010 (Crossley *et al.*, 2010). Por ejemplo, editoriales estadounidenses de prestigio han creado espacios digitales específicos de aprendizaje del proceso de escritura y corrección automática de textos. Algunos de estos espacios virtuales de trabajo son WriteToLearn, de la editorial Pearson, y Eduspace o WriteSpace de Houghton Mifflin. En consecuencia, los profesores pueden asistirse con herramientas automáticas y registrar portafolios de textos, establecer los *rubrics* y tareas para sus estudiantes, además de obtener de forma automática la calificación de los textos y la progresión de sus estudiantes. A su vez, los estudiantes disponen de un espacio de aprendizaje del proceso de escritura con la posibilidad de revisar y de corregir lo escrito, así como de obtener retroalimentación a lo largo del proceso de escritura. Todos estos productos comerciales evalúan y puntúan textos escritos en inglés de forma inmediata.

Por otro lado, instituciones evaluadoras norteamericanas como el Educational Test Service (ETS), el Graduate Management Admission Council (GMAC) o el Center for Research on Evaluation, Standard, and Students Testing (CRESST) promueven el desarrollo de estas herramientas de procesado, evaluación y diagnóstico de textos integradas en espacios virtuales. ETS, por ejemplo, no sólo distribuye productos como Criterion SM, compuesto por *ETS e-rater* (Burstein *et al.*, 2004), *Stumping E-Rater*, *ScoreItNow!* (Enbar, 1999) sino que también apoya la investigación en este campo, forma investigadores y proporciona información de sus logros a nivel internacional por medio de seminarios y becas. Concretamente, *Criterion SM* es un producto *on-line* para el análisis de distintos tipos de textos y niveles. Además, es una plataforma educativa para la enseñanza-aprendizaje de las destrezas de la composición textual, según el nivel de aprendizaje del usuario. Entre las funciones que realiza, destacamos el análisis del contenido, la gramática y el estilo, y el diagnóstico del proceso de la expresión escrita de un estudiante, incluso de toda una clase. También, evalúa el portafolio del estudiante y registra estadísticamente los puntos fuertes y débiles de un estudiante o del grupo. Esta plataforma ofrece retroalimentación tanto al alumno durante el proceso de aprendizaje de escritura como al profesor acerca del desarrollo de las tareas. *Criterion SM*, compila en una plataforma dos herramientas de análisis patentadas por ETS, anteriormente conocidas como *E-rater* y *ScoreItNow!* para facilitar a los estudiantes el éxito en los Graduate Records Examinations (GRE®) y del New TOEFL. De forma similar, GMAC promueve su calificador *Pearson Test of English Academic* (PTE Academic) desde octubre de 2009. Este calificador es el resultado de las últimas investigaciones en evaluación objetiva de textos y en aplicación de las directrices del *Common European Framework of Reference* (CEFR).

Recientemente, también instituciones inglesas tan prestigiosas como Oxford University Press o Cambridge University Press presentan en el mercado este tipo de productos

que facilita el aprendizaje de la escritura y proporciona la evaluación o diagnóstico de textos escritos por el aprendiz. Por ejemplo, Oxford University Press en marzo de 2010 presentó en el mercado su diccionario virtual, *Oxford Advanced Learner´s Dictionary* en CD, con prestaciones similares a las herramientas estadounidenses expuestas más arriba. Esta herramienta se conoce como i-Writer. Además, cuenta con un tutor que ayuda en el proceso de escritura. Mientras, Cambridge Universite Press publica un manual que recoge propuestas y estudios presentados en el Cambridge Colloquium en diciembre de 2007 para la evaluación de lenguas siguiendo las directrices del Marco de Referencia Común Europeo (MCER) titulado *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual* (Martyniuk, 2010).

En cuanto al ámbito del castellano, en 2006 el Instituto Cervantes publicó tres manuales de referencia para la enseñanza y el aprendizaje del castellano, basándose en las recomendaciones y propuestas del *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. Las propuestas fueron desarrolladas y adaptadas al castellano en tres manuales titulados *Plan Curricular del Instituto Cervantes. Niveles de referencia para el español* (*PCIC*) que abarcan el ámbito gramatical, ortográfico, funcional, pragmático, discursivo, nocional y sociocultural, y desarrolla distintos aspectos y habilidades propias de cada nivel de referencia (Cervantes, 2006).

Partiendo de nuestra hipótesis, enunciada en el apartado 1.5, y para validarla, hemos considerado calcular varios índices cuantitativos que nos permitieran nivelar un texto de forma automática. La cantidad de datos numéricos que se generan al procesar muchos parámetros, nos ha hecho reconsiderar esta aproximación, limitando el número de parámetros que, aunque aportan información lingüística sobre un texto, no contribuyen a determinar un nivel. Gracias a la publicación de los volúmenes del *PCIC*, hemos recurrido a los niveles de referencia de los distintos apartados dedicados a la gramática y a las nociones ya que, al organizar el material de aprendizaje por niveles, sirven como referencia para valorar un texto. Para ello, nuestra evaluación se basa en procesar el contenido de un texto y calificarlo a nivel léxico, sintáctico y semántico. Sin duda, la publicación del triple manual *PCIC* se ha convertido para nosotros en la "norma de oro" de evaluación, del que partimos y por el que nos hemos guiado para evaluar los textos objeto de esta investigación.

1.2. Antecedentes

La sucesión de experiencias personales, formativas y profesionales de la autora han despertado el interés de esta investigación sobre el tratamiento automático de textos y, más ambiciosamente, sobre la evaluación de textos escritos por extranjeros. Entre las experiencias que han conducido a esta área de investigación, destacamos los cursos de doctorado orientados a la enseñanza de idiomas (2002-2004); la experiencia en la enseñanza de español a extranjeros (2002-2006) y de lengua castellana a personas adultas (2004-2006 y 2010-2011) en la Universidad Popular de Alcobendas (UPA) facilitando la interacción y el autoaprendizaje del castellano; la experiencia combinada de enseñar y aprender en plataformas de enseñanza virtuales como Moodle, WebCT y Learning Management System (LMS) (2004-2005); y el aprendizaje de extracción de textos para su

posterior tratamiento mediante el método Text Encoding Initiative (TEI) durante el curso de Lingüística Computacional impartido por la Dra. Ana Mª Fernández-Pampillón de la Universidad Complutense (2005-2006). Como resultado de estas experiencias, han surgido tres publicaciones relacionadas con entornos virtuales y tratamiento de textos presentados a tres Congresos del Campus Virtual de la Complutense (Ferrero y Alda, 2005; Alda y Ferrero, 2007b, 2009), una contribución en un libro homenaje al Dr. Manuel Quintanilla, catedrático de la Universidad de Zaragoza (Alda y Ferrero, 2007a) y la investigación previa a la obtención del Diploma de Estudios Avanzados (Ferrero, 2011).

1.3. Motivación

Como se expresa en los antecedentes, nuestro interés ha ido creciendo desde el tratamiento automático de textos hasta la evaluación o diagnóstico del nivel de conocimiento de una lengua. Fue un hecho decisivo conocer al Dr. Peter Kincaid para encaminar esta tesis (2006-2007). Él ha sido uno de los pioneros en la extracción de textos y en la creación de listados específicos y desambiguados de palabras para el control y elaboración de textos para estudiantes norteamericanos con carencias de conocimiento de la lengua inglesa. No menos importante fue hallar la libre disposición del analizador morfo-sintáctico de FreeLing 1.5 y la eficiente ayuda y perspectiva de su creador, el Dr. Lluís Padró (2006-2011). A pesar de la dificultad que supone tener en cuenta parámetros lingüístico-funcionales y evaluar el nivel de un texto de forma conjunta y automática, la experiencia y la dirección de la Dra. Rachel Whittaker y la Dra. Teresa Bordón han sido guías incondicionales para llevar a término este trabajo. No menos importante ha sido la gestión de José Ramón Parrondo Rodríguez, quien nos ha proporcionado los exámenes del Diploma de Español como Lengua Extranjera, las calificaciones y la documentación correspondiente para realizar la parte experimental de la investigación y comparar los resultados automáticos con los de los evaluadores expertos.

Por todo ello, la corrección de textos de forma automática se nos presenta a nosotros también como un gran reto. Reto que implica la suma de, al menos, dos disciplinas: la lingüística y la computación. Sin duda, la existencia de herramientas y espacios digitales descritos en el apartado anterior son un gran estímulo para investigar los criterios utilizados y para diagnosticar las características evaluables de un texto, en nuestro caso, en castellano. Además, este tipo de productos, comerciales o no, concebidos para ayudar en el proceso y mejorar la expresión escrita, contribuyen a la autoformación y facilitan el aprendizaje a lo largo de toda la vida. Es decir, se presentan como herramientas con estrategias y métodos estimulantes para el aprendiz autodidacta ya que proporcionan retroalimentación sobre los errores, así como la posibilidad de obtener un diagnóstico del texto prácticamente de forma inmediata.

En concreto, la evolución del TOEFL paper Based Test (pBT) a través del TOEFL internet Based Test (iBT) al New TOEFL supone la puesta en marcha de un modelo de procedimiento automatizado de evaluación interesante para otras instituciones evaluadoras similares. Además, el nuevo modelo facilita la prueba del test de escritura al combinar tres destrezas durante el proceso de composición de textos. Es decir, el alumno utiliza dos destrezas (lectora y auditiva) antes de demostrar su competencia en la tercera destreza,

la expresión escrita, durante la prueba de composición. Por ello, los investigadores que han propuesto el New TOEFL han estudiado la diferencia entre los dos tipos de tareas de escritura: la composición independiente y la composición integrada. La primera composición consiste en escribir sobre un tema, sin proporcionar ningún tipo de retroalimentación previa, confiando únicamente en los conocimientos académicos, culturales, gramaticales y discursivos previos del examinado. La segunda composición, sin embargo, previamente a la composición escrita, proporciona una lectura y una audición sobre el tema, integrando en el método de evaluación tres de las cuatro destrezas de comunicación: la comprensión lectora y auditiva, y la expresión escrita. Tras el estudio de estos dos modelos de evaluación, los resultados de las pruebas en el tipo de composición independiente y en el de composición integrada demuestran que los candidatos a los tests obtienen mejores resultados en la prueba integrada que en la independiente a nivel pragmático, discursivo y léxico (Cumming *et al.*, 2006).

Ante las posibilidades que ofrecen tales herramientas informáticas en idiomas como el inglés, consideramos que en castellano se pueden desarrollar herramientas que contribuyan a satisfacer prestaciones semejantes. Esta motivación es la razón por la que hemos utilizado herramientas de análisis sintáctico en castellano ya diseñadas, como FreeLing por el Dr. Padró, por la que hemos recurrido a bases de datos y por la que hemos desarrollado herramientas con el entorno de Matlab. Estas herramientas nos permiten procesar textos para computar características léxicas (Lexicator), sintácticas (Sintactor) y semánticas (Semantor) de un texto. En conjunto, nos permite dar un nivel a dichos textos (Evaluator), en función de los resultados obtenidos del análisis en los distintos módulos computacionales.

1.4. Herramientas y métodos utilizados

A partir de los resultados obtenidos al procesar un texto con el sistema de conocimiento de FreeLing 1.5, se ha elaborado otro sistema que denominamos "Evaluator". Este prototipo de sistema de conocimiento, carente actualmente de interfaz fácil para un usuario, es un compendio de módulos que van analizando los datos y aportando resultados.

1.4.1. Procesador de funciones (Analizator)

Este procesador ejecuta una serie de funciones diseñadas para computar cada elemento de un texto (ver las funciones en el apartado 3.2.4). Los resultados, a su vez, se procesan para obtener el mayor número de *items* de un texto y, luego, ser analizados en las herramientas correspondientes para diagnosticar el nivel de lengua del texto analizado.

1.4.2. Analizador léxico (Lexicator)

Esta herramienta, desarrollada por nosotros, aportará el nivel que tiene cada lema de un texto en función, por un lado, del nivel que le otorgue el "Índice de nociones generales y nociones específicas" del *Plan Curricular del Instituto Cervantes* y, por otro lado, de la

pertenencia a uno u otro diccionario en función de la combinación de diccionarios. Esto es, el léxico se puede considerar como básico si está incluido en todos los diccionarios y está en el diccionario de control (Kincaid 1992); como intermedio si aparece en el Wiktionary o Diccionario de FreeLing 1.5; como superior si aparece en el Diccionario de FreeLing 2.0 o esWordNet; incluso, si pertenece a un área temática concreta del diccionario de la Dra. Fuensanta López (López Martín, 1999), como específico. Lexicator funciona cuando el texto ha sido procesado y lematizado con el programa de FreeLing y sometido a las funciones de "Analizator".

1.4.3. Analizador sintáctico (Sintactor)

Es otra herramienta desarrollada también por nosotros. Procesa los datos del texto a partir de un archivo con más de 2.500 estructuras sintagmático-sintácticas, creadas bajo el criterio de complejidad sintáctica y calificadas con sus respectivos niveles de referencia. El fichero que contiene todas estas estructuras lo denominamos "complejidad_sintaxis.txt" y presentamos, como ejemplo, un fragmento del archivo en el apéndice A.6.1. Estas estructuras se han elaborado siguiendo dos criterios: el criterio de complejidad sintáctica y el criterio de nivel de referencia del *PCIC*.

El concepto de complejidad sintáctica, cuyos índices desarrolla la Universidad de Memphis en su herramienta Coh-Metrix, es un conjunto de parámetros desarrollado por el equipo de investigación de Memphis para medir la complejidad sintáctica de un texto a la que se enfrentan lectores de un cierto nivel académico (McNamara, 2002-2005). Estos índices cuantificables nos han servido de base y de reflexión para crear el fichero de estructuras y la implementación del archivo de multivocablos (ver apéndice A.6.2). Consideramos estos índices idóneos para cuantificar o diagnosticar la complejidad de un texto ya que dicho texto es el resultado de un proceso de escritura y, en consecuencia, de lectura.

En cuanto al criterio de nivel de referencia, nos hemos guiado por la clasificación que establece el *Plan Curricular del Instituto Cervantes* en los apartados dedicados a la Gramática (A1-A2, B1-B2 y C1-C2). Se han listado aquellas estructuras sintagmáticas o sintácticas que hemos podido formular para ser procesadas después. De manera que la herramienta, que hemos identificado como Sintactor, procesa estas estructuras e identifica porcentajes, frecuencias e índices de referencia de dichas estructuras dentro de los textos objetos de estudio. A partir de los resultados que nos proporciona Sintactor, primero, podemos calificar el nivel de las estructuras del texto y, después, hallar el nivel del texto basándonos en dos parámetros: número de estructuras por nivel de referencia y el nivel de referencia de las categorías gramaticales o *Part of Speech* (*PoS*) en función de su pertenencia a estructuras sintácticas de mayor nivel.

1.4.4. Analizador semántico (Semantor)

Semantor es el último programa desarrollado por nosotros para analizar el contenido semántico de un texto basado en el método de Principal Componentes Analysis (PCA), más conocido en Lingüística Computacional como Latent Semantic Analysis (LSA). Es

un método matemático desarrollado por varios investigadores en diversas áreas de conocimiento, pero aplicado al análisis de textos por el Dr. Landauer (Landauer *et al.*, 2003). Un trabajo realizado con este método de análisis de textos científicos se ha publicado en un capítulo conjunto en un libro homenaje al Dr. Quintanilla (Alda 2007). Este método, desarrollado en esta herramienta que denominamos Semantor, es el que utilizamos para el análisis semántico de los exámenes del Diploma de Español como Lengua Extranjera (DELE). Semantor nos propociona la semántica del texto basándose en todas las palabras del texto.

1.5. Hipótesis

La hipótesis de este trabajo se expresa de la siguiente manera:

> Si aceptamos que la variedad léxica, la complejidad sintáctica y el contenido semántico son criterios mínimos válidos para determinar el nivel de un texto escrito, es posible evaluar de forma automática el nivel de competencia comunicativa de un aprendiz en la composición escrita.

Además de los criterios lingüísticos que permiten a un evaluador juzgar la competencia escrita, éste debe considerar aspectos como la coherencia, la adecuación, el estilo, la función y los errores de un texto, entre otros. En este estudio, sin embargo, de forma sistemática se procesan parte de estos rasgos propios de un texto escrito ya que los métodos propuestos intrínsecamente detectan cierta coherencia y adecuación gracias a los glosarios, las estructuras sintácticas y los *corpora* de referencia utilizados. Es más, los errores ortográficos ni se procesan ni se computan, por ello ni otorgan nivel ni se consideran objeto de estudio en este trabajo.

1.6. Objetivos

Una vez definidos, probados e integrados un cierto número de parámetos cuantitativos computables en los distintos módulos y herramientas de análisis, seremos capaces de discriminar textos de nivel B2 y nivel C1 al calificar, por un lado, el nivel léxico y sintáctico y, por otro, al diagnosticar el mayor o menor contenido semántico esperable de cada texto.

Nuestra investigación consiste en probar que las herramientas diseñadas son capaces de analizar, evaluar y diagnosticar un total de ochenta exámenes escritos por aprendices extranjeros en las pruebas de expresión escrita para la obtención del Diploma de Español como Lengua Extranjera (DELE) para los niveles B2 y C1, según los criterios de nivel establecidos por el *Plan Curricular del Instituto Cervantes.*

Proponemos que, mediante unos índices generales, podemos probar el nivel de competencia escrita en el que se halla un aprendiz. Se quiere valorar empíricamente y de forma computacional aquellas realizaciones lingüísticas que un experto en lengua puede juzgar como acertadas, que un profesor de idiomas intuye como aceptables o que un grupo de

examinadores califica como propias de un determinado nivel en relación a una hoja-matriz de valoración con criterios de evaluación (*rubric*).

A continuación, de forma detallada, exponemos los distintos objetivos de este estudio.

1.6.1. Generales

- Especificar criterios lingüísticos y parámetros cuantitativos característicos de textos comunicativos e informativos en español.
- Aplicar los criterios y parámetros a las herramientas automáticas creadas para evaluar textos comunicativos e informativos.
- Utilizar herramientas computacionales que procesen gran cantidad de texto.
- Procesar y marcar con un nivel de referencia cualquier texto en castellano.

1.6.2. Concretos

- Definir parámetros procesables a nivel tipográfico, léxico, sintáctico y semántico que contribuyan a evaluar el nivel de un texto.
- Utilizar herramientas disponibles y probar las creadas, con índices de nivel de referencia integrados, en unos textos modelo (discursos navideños del Rey Juan Carlos I) para comprobar la eficiencia de los parámetros léxicos, sintácticos y semánticos propuestos.
- Nivelar textos de españoles nativos de forma automática con estas herramientas.
- Utilizar *corpora* de referencia que permitan marcar un nivel de referencia.
- Comprobar automáticamente el nivel de referencia de textos-lecturas previamente asignadas con un nivel por el Instituto Cervantes (IC).
- Comparar la calificación automática a nivel léxico, sintáctico y semántico de exámenes del DELE con parámetos nivelados automáticamente con la calificación otorgada por dos correctores especializados del IC.
- Comprobar que la evaluación de textos en castellano de forma automática es factible y validar los criterios.

1.7. Estructura del trabajo

A lo largo de este trabajo, estructurado en 8 capítulos y un apéndice, introducimos el tema de investigación, algunos antecedentes y exponemos el proceso de la investigación desde la motivación inicial y el tema de estudio al planteamiento de una hipótesis y los

objetivos en el capítulo primero. Consecuentemente, se prueba la hipótesis en el resto de los capítulos.

En el capítulo segundo presentamos el estado de la cuestión, el marco teórico lingüístico-computacional que encuadra nuestra tarea y algunos métodos de análisis lingüístico-computacional y de evaluación de textos que se aplican en la actualidad.

Como contribución original, ya desde el capítulo tercero, presentamos los materiales, herramientas y métodos de análisis que nos "sirven de lentes" (comunicación personal del Dr. Padró, 2007) para observar, probar y demostrar lo que postulamos a nivel léxico (cap. 4), sintáctico (cap. 5) y semántico (cap. 6).

En el capítulo séptimo, se han aplicado las herramientas y los métodos de análisis. Por un lado, se analizan textos de un castellano estándar cuya estructura y contenido son entendidos por todos los hispanohablantes, concretamente por los españoles. Los textos estándar seleccionados son los discursos navideños del Rey Juan Carlos I. A su vez, estos análisis del castellano se contrastan con textos escritos por estudiantes adultos que, en su afán por obtener el Graduado en Educación Secundaria Obligatoria, intentan mejorar su nivel de competencia escrita en castellano. Por otro lado, se aplicarán todas las herramientas desarrolladas para analizar textos de exámenes escritos por extranjeros y evaluados por expertos del Instituto Cervantes. Este material es clave para la comparación entre los resultados obtenidos en este trabajo y los procedentes de la corrección por los expertos. En este mismo capítulo se presentan los resultados y la interpretación de los mismos. Por una parte, proporcionamos los valores de la evaluación automática de los textos en los tres niveles de lengua y, por otra, la comparación de la evaluación automática con la de los expertos. La evaluación de los expertos ha sido previa e independiente a esta investigación. Los dos correctores son expertos en evaluación y certificación del Instituto Cervantes.

Se finaliza el trabajo de investigación en el capítulo octavo con la enumeración de las conclusiones y futuras propuestas de investigación en esta misma área.

En el apéndice se incorporan aquellos módulos, tablas y materiales que se han creado para esta investigación, no están sujetos a derechos de autor o sirven como ejemplo.

1.8. Siglas

AWL: Academic Word List

CEFR: Common European Framework of Reference

CL: Computational Linguistics

CVC: Centro Virtual Cervantes

DELE: Diploma de Español Lengua Extranjera

DTC: Derivational Theory of Complexity

E/LE: Español Lengua Extranjera

esWN: esWordnet (base de datos y sistema de conocimiento léxico-semántica en castellano)

ETS: Educational Testing Service

GRE: Graduate Record Examination (2011: GRE revised General Test)

IC: Instituto Cervantes

LMS: Learning Management System

LSA: Latent Semantic Analysis

LSC: Latent Semantic Components

LSI: Latent Semantic Indexing

L1: Lengua materna

L2: Segunda lengua

MCER: Marco Común Europeo de Referencia para las lenguas: aprendizaje, enseñanza, evaluación

MD: Marcadores del Discurso

NE Nociones Específicas

NG: Nociones Generales

NL: Natural Language

NLP: Natural Language Processing

PCA: Principal Component Analysis

PCIC: Plan Curricular del Instituto Cervantes

PCIC. A1-A2: Plan Curricular del Instituto Cervantes. Nivel A1-A2

PCIC. B1-B2: Plan Curricular del Instituto Cervantes. Nivel B1-B2

PCIC. C1-C2: Plan Curricular del Instituto Cervantes. Nivel C1-C2

PLN: Procesamiento del Lenguaje Natural

SEAN: Simplified English Analyser

TEI: Text Encoding Initiative

TOEFL: Test of English as a Foreign Language

TOEFL-iBT: Test of English as a Foreign Language-internet Based Test

TOEFL-pBT: Test of English as a Foreign Language-paper Based Test

TREC: Text REtrieval Conference

Unidad-t: unidad terminal u oración

Capítulo 2

Estado de la cuestión y marcos teóricos

2.1. Estado de la cuestión

Desde el comienzo de esta investigación en 2006 hasta su lectura en 2011 hemos seguido el desarrollo e investigación de la evaluación automática de varios autores e instituciones. Durante este tiempo se han obtenido resultados y creado herramientas y productos innovadores por instituciones, empresas y universidades aplicando conocimientos interdisciplinares que abarcan a la lingüística, la computación, la psicología, las matemáticas, la estadística, la evaluación y la enseñanza de idiomas.

2.1.1. Sistemas de evaluación

Muchas instituciones, investigadores y profesores tratan de describir y desarrollar criterios de evaluación específicos y significativos, viables y computables, además de válidos y precisos para calificar niveles de aprendizaje. La investigación sobre la evaluación automática de la información, en general, y de textos en particular, comenzó en 1966 con Ellis Page. Con su Project Essay Grade (PEG) demostró que se podía calificar la calidad de un escrito de forma automática y detallada frente a la calificación holística de los correctores humanos (Page, 1966, 1994; Page *et al.*, 1997).

Desde entonces se han desarrollado nuevas herramientas de trabajo e investigación en la misma área. Entidades educativas estadounidenses como Educational Testing Service (ETS) han hecho nuevas propuestas de exámenes, como el New TOEFL, para poder procesar los datos con sistemas de evaluación automática (Cumming *et al.*, 2006).

Otro paso para adaptar los sistemas educativos a la tecnología es la experiencia realizada por el National Assessment of Educational Progress (NAEP). En 2010 se examinaron estudiantes norteamericanos de 8^{o} a 12^{o} curso y se evaluaron sus ensayos previamente diseñados para ser realizados en el ordenador. Luego, manualmente, unos expertos calificaron aspectos como el desarrollo de ideas, la organización de éstas, la destreza del lenguaje y el uso de convencionalismos (Gorman, 2010). No obstante, aunque el resultado de algunas herramientas de evaluación automática (Automated Writing Evaluation: AWE) como MY Access!® se han criticado por la comunidad educativa del sur de California, los estudios

sobre este tipo de sistemas demuestran que favorecen el rendimiento escolar (Grimes y Warschauer, 2010).

Sin embargo, a nivel oficial y educativo ya se realizan evaluaciones automáticas. Por ejemplo, el Educational Testing Service utiliza *e-Rater*, desarrollado por Jill Burstein, para calificar textos (Burstein *et al.*, 2004). Otro ejemplo concreto del estado de la cuestión en la evaluación textual son las herramientas elaboradas en el ámbito educativo norteamericano por MetaMetrics®, Inc.. Éstas están recomendadas también por el Educational Testing Service para trabajar las destrezas de la lectura y la escritura. Las herramientas son Lexile Framework for Reading y Lexile Framework for Writing. La primera se conoce como MyWritingWeb, que ayuda en el proceso de escritura y evalúa un escrito; la segunda es Lexile® Measure, que diagnostica y mide la dificultad de lectura de un texto, y con esa medida se puede ajustar el nivel del lector al nivel de la lectura. Además, esta herramienta aporta ciertos parámetros que miden la complejidad de lectura de un texto. Como ejemplo, se ha procesado un texto en castellano con la herramienta Spanish Lexile Analyzer y que se observa en la figura 2.1. En esta figura se reproducen los parámetros y valores correspondientes al texto Penpal_1_1, redactado por un estudiante adulto nativo español de destreza escrita media-baja. Los parámetros que facilita este analizador son: 800L como nivel *Lexile* de lectura, 15, 00 como la longitud media de una frase, 4,97 como la media del logaritmo de la frecuencia de palabras y 105 es el recuento de palabras que identifica el texto.

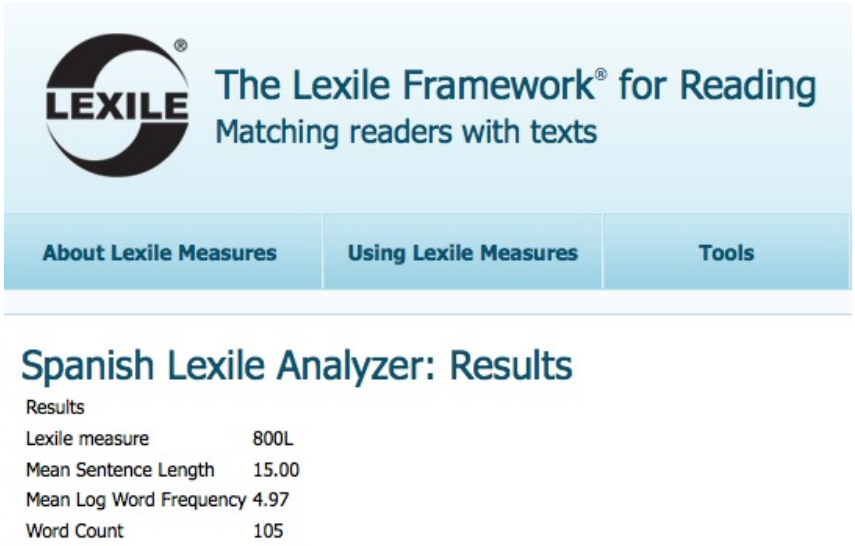


Figura 2.1: Resultado de Lexile® Measure en el cálculo de los parámetros del texto de penpal_1_1 de 105 palabras.

En la figura 2.2 se reproducen los mismos parámetros pero, en este caso, referimos los valores correspondientes a un fragmento de "Un sueño muy extraño", lectura intermedia de la colección "Lecturas paso a paso" del Centro Virtual Cervantes en formato electrónico. Los valores *lexile* de esta lectura son muy semejantes a los de la figura 2.1. Cada vez son más numerosos los libros estadounidenses destinados a la educación que llevan un identificador del grado de lectura o *lexile* (L).

Nuestro trabajo se enmarca también en esta línea de investigación sobre la generación

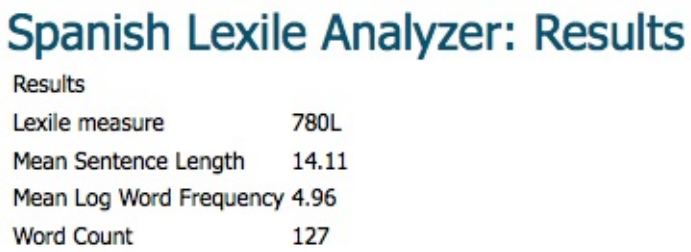


Figura 2.2: Resultado de Lexile® Measure en el cálculo los parámetros de la lectura CVC_1_1.

de sistemas de evaluación del nivel de aprendizaje. En el ámbito del castellano existen algunos entornos virtuales que permiten el autoaprendizaje, como el Centro Virtual Cervantes o plataformas educativas virtuales con autoevaluación de tareas prediseñadas y cerradas. Sin embargo, apenas existen sistemas de conocimiento que permitan evaluar textos con los niveles de aprendizaje del Marco Común Europeo de Referencia (MCER). Un ejemplo de evaluación cerrada es DIALANG Test Server. El departamento de idiomas de la Universidad Nacional de Educación a Distancia (UNED) presenta DIALANG como una prueba de nivel. Se describe como un recurso que

> "ofrece tests de diferentes habilidades lingüísticas, cuidadosamente diseñados y validados, junto a una extensa gama de comentarios (*feedback*) y consejos especializados sobre como mejorar sus conocimientos" (UNED).

DIALANG es el resultado de un proyecto conjunto de universidades europeas operativo desde 2001, sin actualizar desde el 2009 y mantenido de forma altruista. Desde que fue saboteada la página web DIALANG.org, DIALANG Test Server está ubicado en un servidor de la Universidad de Lancaster, y sus consultas diarias oscilan entre 500 y 1000.

Según Alderson, DIALANG se caracteriza por ser un sistema de diagnóstico en 14 idiomas que identifica el nivel de aprendizaje y de las destrezas de una lengua, dentro de los postulados del MCER. Una de las características de sus tests es que presentan tareas indirectas en todas las destrezas. Por ejemplo, en relación a la escritura, el aprendiz que realiza el test debe elegir segmentos de lengua o *items* preestablecidos para completar ejercicios de rellenar huecos. Este procedimiento permite medir la precisión gramatical, la ortografía, la adecuación del registro y la organización textual. En 2005 se dieron a conocer los estudios y resultados de DIALANG en los que se expresaba la necesidad de validar los resultados de DIALANG con otros sistemas evaluadores y de realizar tests directos de escritura ya que es una de las carencias que resaltan sus investigadores (Alderson, 2005).

Un año después, en 2006, el Instituto Cervantes publica el *Plan Curricular del Instituto Cervantes (PCIC).* Esta publicación se convierte en texto clave para la realización de esta tesis dado que el *PCIC* describe los niveles de referencia para el castellano. El *PCIC* se fundamenta en las indicaciones del *Threshold Level Series* (1978, 1990), en las recomendaciones del *Marco Común Europero de Referencia para las lenguas: aprendizaje, enseñanza, evaluación* (MCER 2001) y en la *Guía para la elaboración de descripciones de niveles de referencia para las lenguas nacionales y regionales* (2005) del Departamento de Política Lingüística del Consejo de Europa (Cervantes, 2006, 21). Sus criterios de

selección de material y establecimiento de niveles se han basado fundamentalmente en el conocimiento y experiencia de expertos y profesionales de la enseñanza del español como lengua extranjera (Bordón Martínez, 2004; Sánchez Lobato y Santos Gargallo, 2004; Bordón Martínez, 2006). Sin embargo, no utilizan "inventarios de frecuencia estadística [...] ni los datos lingüísticos producidos en situaciones de evaluación" (Cervantes, 2006, 26), como considera recomendable la *Guía para la elaboración de descripciones de niveles de referencia para las lenguas nacionales y regionales* del Departamento de Política Lingüística del Consejo de Europa (de Europa, 2005, 6). Por ello, nuestro estudio y aportación, por un lado, complementa todo esto ya que gran parte de la investigación de este trabajo se basa en análisis estadísticos realizados a varios textos de distinta naturaleza y a exámenes, producidos por aprendices y evaluados automáticamente. Por otro lado, contribuye con sus resultados porque ratifica y valida cuantitativamente las propuestas del *PCIC*.

2.1.2. Sistemas de extracción y procesamiento de datos: productos comerciales

En lo que se refiere a productos comerciales, existen entidades como Daedalus que trabajan para ofrecer herramientas y servicios relacionados con el procesado de textos y la recuperación de información. Daedalus es una empresa española que participa en Cross Language Evaluation Forum (CLEF), el foro europeo más importante para la investigación y el desarrollo respecto a la extracción de información multilingüe y la reutilización de datos para posteriores investigaciones y la evaluación de sistemas de extracción de información multilingüe y multimedia. Este foro es equivalente al norteamericano Text REtrieval Conference (TREC) y al asiático National Institue of Informatics Test Collection for Information Retrieval (NIITCIR). A diferencia de TREC, que se interesa por las técnicas de recuperación de información en idioma inglés, y de NIITCIR, centrado en los idiomas asiáticos, CLEF da acceso a toda información multilingüe. Por ejemplo, entre las áreas temáticas de investigación, destacan la recuperación de información multilingüe y multimedia (ImageCLEF), la búsqueda de textos concretos y de información geográfica (GeoCLEF), de información en la web (WebCLEF) y la búsqueda de respuestas (QA@CLEF).

Entre las herramientas desarrolladas por Daedalus, disponibles en Internet y de interés para un usuario del idioma español, destacan el corrector de textos de Stilus, que proporciona un informe de corrección y la herramienta que permite hacer análisis morfológicos y conjugar verbos. Además, han desarrollado herramientas que permiten encontrar palabras en un diccionario inverso, aunque no se sepa cómo se escriben las palabras, e identificar vocablos con el etiquetador morfo-sintáctico, incluso extraer resúmenes de forma automática, o hacer rastreos de contenido en la *web* con la herramienta de búsqueda borrosa.

En esta misma línea, se aúnan esfuerzos interdisciplinares también para crear potentes herramientas basadas en diccionarios y bases de datos en internet. Al igual que Daedalus, otras empresas que trabajan con el idioma español son la ecuatoriana Signum o la española Molino de ideas. Por su parte, Signum ofrece servicios lingüísticos como tratamiento de textos, módulos de diccionarios, tesauros o correctores ortográficos para sistemas in-

formáticos o usuarios particulares. La otra empresa emergente es Molino de ideas. Dentro de sus productos, destaca Onoma, un conjugador inteligente de verbos capaz de conjugar de forma automática verbos en infinitivo que, aunque no existan, reconoce el "prototipo" de verbo español. Entre sus proyectos, desea generar herramientas como Plakton, que recolecte textos de Internet para crear un *corpus* de español y que permita hacer análisis sintácticos; Ashmera, que deconstruye la palabra distinguiendo la palabra raíz y los morfemas (flexivos, prefijos, sufijos y afijos); Ishmov, buscador que responderá a preguntas sobre un tema y remitirá, a su vez, a páginas web que traten sobre el tema de interés para el usuario; Coral, servirá para identificar información contenida en un conjunto de palabras; y, por último, Azrael, reconocerá el idioma español por su estructura silábica.

Como queda expuesto, todos estos productos son comerciales y la investigación que realizan está en función de desarrollar herramientas competitivas con aplicaciones novedosas.

2.1.3. Proyectos y grupos de investigación

En el área investigadora de las universidades, destacamos el equipo de Coh-Metrix en la Universidad de Memphis, que desde los años 90 lleva desarrollando índices de medida y herramientas para medir la dificultad de los textos, bien para ser leídos (Graesser *et al.*, 2010) o bien como resultado del proceso de escritura (Crossley *et al.*, 2010). El trabajo de este grupo se basa en estudios lingüísticos para determinar parámetros de medida y aplicar diferentes métodos para el análisis textual. Desarrollamos explícitamente su trabajo en la página 28, en la tabla 2.1 y en el apartado 3.2.2.

Entre las universidades españolas relacionadas con el Procesamiento del Lenguaje Natural, destacan trabajos realizados con métodos de estadística multivariante (LSA, LSC, PCA). Por ejemplo, el Grupo de Tecnología del Habla de la Universidad Politécnica de Madrid investiga la identificación de grupos de palabras o *clusters* cuya semántica se aplique a dispositivos activados por la voz (Lucas Cuesta *et al.*, 2010); también, el Grupo de Herramientas Interactivas Avanzadas de la Escuela Politécnica Superior de la Universidad Autónoma de Madrid con tesis doctorales como la de la Dra. Pérez Marín sobre la generación de modelos conceptuales de estudiantes a partir de textos evaluados automáticamente sobre temas de respuesta libre (Pérez Marín, 2007).

Es importante citar el trabajo de colaboración entre la Universitat Pompeu Fabra, la Universitat de Barcelona y el Laboratoire Informatique d'Avignon para diseñar un Segmentador Discursivo (DisSeg) en español (da Cunha *et al.*, 2010a). Este segmentador sintáctico y discursivo utiliza el analizador morfosintáctico de FreeLing y se basa en reglas léxicas y sintácticas del castellano, fundamentalmente en marcadores del discurso, conjunciones, formas verbales y signos de puntuación (da Cunha y Torres-Moreno, 2010c). Esta herramienta, DisSEg, se presenta como única en el área de análisis sintáctico-discursivo dentro del marco de la Rhetorical Structure Theory (RST) propuesto por Mann y Thompson (1988). Dadas las características de segmentación, también puede ser muy útil para analizar la complejidad sintáctica por segmentos dentro de una cláusula u oración (da Cunha *et al.*, 2010b). Además, al segmentar el texto en oraciones marcadas por las conjunciones y marcadores del discurso, esta herramienta permitiría identificar y com-

parar las relaciones sintáctico-discursivas dentro del mismo texto y, a su vez, en relación con otros textos. Esta aplicación se considera muy útil para poder evaluar la adecuación y estructura de un texto, aspectos que no calificamos en este trabajo pero que sí evalúan los expertos del Instituto Cervantes en los exámenes de DELE.

2.2. Marcos teóricos lingüístico-computacionales

El ámbito general en el que se desarrolla este trabajo, el Procesamiento del Lenguaje Natural, involucra otras áreas tan diversas y multidisciplinares como la Lingüística Aplicada, la Lingüística Computacional, la Semántica Computacional y la Evaluación Computacional. A su vez, éstas son áreas amplias en las que, en general, encuadramos nuestro trabajo para la compilación, la extracción y el procesamiento de textos, el análisis de datos y la evaluación automática de dichos textos. Esta concurrencia de disciplinas supone que la computación y el diseño de modelos de la realidad permitan, a través de datos, procesar, caracterizar y reproducir dicha realidad con las propiedades que la configuran (Griffiths *et al.*, 2009; Sanborn *et al.*, 2011). En el caso del Procesamiento del Lenguaje Natural se recurre a modelos de Psicolingüística y procesos cognitivos que no pueden deslindarse de la Lógica-Matemática y la Estadística para calcular y correlacionar datos. No menos importantes son las aplicaciones de muchas de estas disciplinas al desarrollo de sistemas basados en Inteligencia Artificial, uso de heurísticas y de representación del conocimiento. Concretamente, las disciplinas y marcos que configuran el desarrollo de nuestra investigación se exponen a continuación.

2.2.1. Lingüística aplicada

Distintos marcos de Lingüística Aplicada se utilizan en el desarrollo de modelos y teorías para el análisis de textos como, por ejemplo, la gramática generativo-transformacional (Chomsky, 1989) que estudia la variedad de transformaciones de los sintagmas, los principios formales de la Gramática de Interdependencias sintáctico-semánticas (Bornkessel *et al.*, 2006), la Teoría de los Elementos Relacionales de Chaffin y Herrmann (Díez Orzas, 1999) o las Relación de Inferencias (Lin y Pantel, 2001a,b).

La Gramática Funcional (Dik, 1997; Connolly y Dik, 1989) es otro entorno teórico para el análisis de textos al igual que la Gramática Sistémico-Funcional de Halliday (Halliday, 1970, 1974, 1985). Halliday, como otros autores interesados en el análisis de las variaciones discursivas (Conrad y Biber, 2009), estudia criterios que contribuyen a comprender y mejorar el proceso de comunicación, atendiendo al género, a la función del texto y su significado, a la coherencia y al contexto. En Lingüística Aplicada también se incluyen las contribuciones de la Gramática formalista o transformacional-generativa porque se considera importante una buena estructura de la frase y su cohesión, aunque no el valor semántico o social del texto (Carpenter, 2005, 181). Además, con las técnicas computacionales, cada vez más criterios lingüísticos pueden transformarse en parámetros procesables para analizar el texto. Por ejemplo, los marcadores del discurso son elementos de gran interés en Gramática Estructural (Martín Zorraquino, 1998), Lingüística Fun-

cional (Pons Bordería, 2006), Lingüística Computacional (Alonso *et al.*, 2002a,b,c,d,e) o Lingüística Pragmática (Romero Trillo, 2007).

2.2.1.1. Cohesión y coherencia

La cohesión y la coherencia son criterios genéricos pero muy relevantes para la evaluación semántica y funcional de un texto. Louwerse identifica dos dobles categorías para la comprensión o elaboración de un texto: la coherencia y cohesión global y local del texto, y la coherencia y cohesión gramatical y semántica (léxico) (Louwerse, 2004; Louwerse *et al.*, 2004). No obstante, según un reciente trabajo del equipo de Louwerse, por los resultados obtenidos en el procesamiento de diversos tipos de texto, afirma que la cohesión es satisfactoria, pero que no se puede validar una cohesión alta o baja en los textos analizados porque no se ha contrastado todavía la fiabilidad de los resultados con otro modelo de medida (McNamara *et al.*, 2010).

La cohesión identificada con la Gramática Formal

Por una parte, los componentes discursivos, las estructuras textuales globales o macro-estructuras como títulos, subtítulos, resumen, conclusión, idea principal, ideas secundarias, etc., en definitiva, la distribución de la información en párrafos acordes a lo que se comunica, contribuyen a la coherencia y cohesión global. Por otra parte, la cohesión local se obtiene mediante la mecánica gramatical o cohesión gramatical entre palabras funcionales (conectores, artículos, pronombres, verbos auxiliares, etc.) y los componentes de escritura, estructuras textuales locales o micro-estructuras, como palabras contenido, tiempo y modo verbal, marcadores del discurso y sus tipos de cohesión (referencial, causal, espacial, temporal, aditiva, etc.).

La coherencia asociada a la Gramática sistémico-funcional

Se consideran conceptos globales a la estructura y al género discursivo (tema, registro y función comunicativa), y conceptos locales, al léxico. Por ello, el contenido semántico del vocablo aporta la coherencia local del léxico y, la global, la relación del léxico con el resto de los vocablos. Igualmente la cohesión semántica local se produce cuando el léxico pertenece al mismo campo semántico o a la misma intención de significado. Se alcanza la coherencia semántica o adecuación cuando el estilo, el tema y el tipo de discurso se conjugan con la semántica léxica.

Según la teoría de que muchos géneros discursivos presentan esquemas estructurales y de contenido semejantes, los textos académicos, el género epistolar o el correo-e son modelos de estructuras modulares (López Alonso, 2006) y, puesto que muchos enunciados se producen en diferentes ámbitos de la actividad humana, son relativamente estables (Bakhtin, 1998). Ese postulado permitirá aplicar modelos estadísticos, matemáticos, computacionales o estocásticos a frases, párrafos y textos diferentes pero coherentes entre sí para analizar su semántica (Landauer *et al.*, 1998a).

2.2.1.2. Análisis del discurso

Una de las secciones de trabajo de este estudio ha sido listar, categorizar y marcar la función de los marcadores del discurso (MD) como categorías gramaticales y como elementos significantes del discurso, aplicando nociones cualitativas o gramaticales.

Terminológicamente, utilizaremos el apelativo de marcadores del discurso. Desde una perspectiva estructural se consideran marcadores porque indican la relación semántica-estructural de un texto aunque, desde una perspectiva de la gramática tradicional, se les denomina conectores porque señalan la interrelación sintáctica entre segmentos oracionales.

Para el análisis del discurso se proponen distintos métodos de aproximación a estas partículas o marcadores (Fischer, 2006). Destacamos el capítulo de Pons donde presenta el debate sobre la denominación entre marcadores del discurso y conectores (Pons Bordería, 2006). Con vistas a la divulgación y aprendizaje del español como una L2, autores como Briz y Pons no comparten la distinción entre conectores, marcadores o modalizadores. Como ejemplo, en la configuración de su diccionario electrónico, opta por una forma más "procedimental que conceptual", denominando "partículas" a cada entrada, sea vocablo simple o multivocablo, acompañada de su descripción. El Diccionario electrónico de partículas del discurso es el producto del grupo de investigación Val.Es.Co del Departamento de Filología Hispánica de la Universidad de Valencia. La concepción de este diccionario, disponible en Internet, es útil como consulta de partículas documentadas pero no para su reutilización en computación, ya que no se dispone de etiquetas procesables que marquen las peculiaridades que describe cada partícula del discurso.

En nuestro estudio, también optamos por esta economía terminológica denominando multivocablos a todos los marcadores, aunque distinguimos gramaticalmente adverbios, conjunciones, locuciones preposicionales o adverbiales. Es más, no entramos en conflicto con la definición de frase, proposición ni oración, ni en la clasificación de oraciones coordinadas ni subordinadas. Simplemente procedemos a calificar los marcadores compuestos o multivocablos en función del nivel o etapa de aprendizaje que se propone en el *PCIC*. Esta aproximación global nos permite también considerar al mismo tiempo aspectos gramaticales y no gramaticales (Pons Bordería, 2006, 80).

Efectivamente, los marcadores del discurso y los conectores aportan cohesión y coherencia a un texto. Por tanto, son parámetros para medir la coherencia. En investigaciones iniciales, Louwerse clasificó unos pocos marcadores, y muy poco ambiguos, en una taxonomía caracteriza por tipo (aditivos, temporales y causales), polaridad (negativos o positivos) y dirección (anticipativos, retrospectivos y bi-direccionales) (Louwerse, 2001). Según este estudio, su interés por los conectores del discurso era medir el tiempo y el movimiento de ojos de un lector en relación a los conectores que se encontraba para dotar al conector y a la lectura de cierto grado de dificultad. Se observa que el lector invierte más o menos tiempo en establecer la coherencia entre frases dependiendo del tipo de conector, de su polaridad y de su posición en el frase. A partir de estos datos, se constata empíricamente que los conectores son buenos índices para medir la complejidad de un texto, que su combinación con categorías léxicas negativas dificulta la comprensión de un texto y que su posición dentro de la frase permite medir la complejidad de la estructura

del texto en relación a la coherencia que busca establecer el lector para comprender el texto.

De manera que, en estudios posteriores, Lowerse amplía los tipos de marcadores en relación a la cohesión que establecen en el texto (aditiva, causal, lógica y temporal), y su complejidad aumenta. Además, ya no sólo procesan conectores y conceptos simples como *because* sino compuestos de varios elementos como *a consequence of* (McNamara *et al.*, 2006).

En nuestra investigación, nosotros no vamos a medir la coherencia de la frase sino que computamos y nivelamos aquellos marcadores que aparecen en el texto. Obviamos que cohesionan y generan la complejidad sintáctica de la frase y del texto. Por ello, en este estudio se procesan varios tipos de marcadores organizados por su tipología y funcionalidad (aditivos, adversativos, explicativos, causales, condicionales, concesivos, comparativos, consecutivos, temporales, finales, locativos, relativos, referenciales, conclusivos, subjetivos, etc.) sin considerar si se comportan como marcadores discursivos u oracionales. En nuestro estudio se califican a nivel léxico y sintáctico. En la taxonomía listada y nivelada de multivocablos no se incluyen marcadores o conectores simples como "cuando", sino que fundamentalmente se registran y nivelan marcadores multivocablo como "después de que".

Desde una perspectiva semántico-discursiva, la taxonomía o clasificación de marcadores no sólo distingue la tipología semántica y funciones sintácticas o discursivas de cada marcador sino su grafía, posición inicial, final, intermedia o intercalada en el discurso o en la frase, es decir, el nivel de posicionamiento. Esta clasificación agrupada con tantas variables permite un alto grado de desambiguación de los marcadores y, por tanto, una identificación muy precisa en su computación (Alonso *et al.*, 2002a,e,b,d,c).

2.2.1.3. Estructuras gramaticales procesables

La *Derivational Theory of Complexity* (DTC) propuesta por Miller se funda en bases psicológicas para postular que cuanto más compleja se va haciendo una estructura sintáctica más complejo es el proceso de comprensión y reproducción de esa estructura (Miller, 1962). Posteriormente, también Hunt postuló (Hunt, 1965) esta teoría en sus investigaciones, pero hoy está sometida a revisión y crítica por considerarse que la investigación empírica no probaba la teoría (Pfau, 2000; Checa García y Lozano, 2002).

En esta misma línea teórica se enmarca también la metodología de Véliz. Metodología que se basa de nuevo en los índices de Hunt y que nos parece aplicable también para aprendices de E/LE de un nivel B2-C1. Véliz afirma que los índices de madurez sintáctica se corresponden con la etapa de adquisición en la que se encuentra el aprendiz (Véliz, 1999). No obstante, qué índices se corresponden con qué etapas y cuántas son las etapas de aprendizaje es un trabajo pendiente de investigación para probarse empíricamente.

Por nuestra parte, consideramos que los criterios que propone Hunt resultan útiles. Estos criterios permiten analizar de forma automática estructuras fijas. Luego, se puede diagnosticar el nivel de un texto al haber calificado cada estructura con diferentes niveles, según los criterios de nivel de *Plan Curricular del Instituto Cervantes* (*PCIC*).

2.2.2. Lingüística computacional

En esta disciplina se enmarca la creación de *parsers* y *taggers*. Éstos son los que luego permiten configurar sistemas de análisis y etiquetados de textos y *corpora* para luego procesarlos. Existen muchos grupos de investigación que precisan de conocimientos lingüísticos para la identificación de las bases del lenguaje y, luego, para el desarrollo de bases de datos, herramientas, sistemas y modelos de procesamiento que sirven para la extracción y compilación de información de textos o para la traducción. Uno de los grupos más activos en España es el grupo del Centro de Tecnología y Aplicación del Lenguaje y del Habla (TALP) y el grupo Natural Language Processing (NLP), ambos de la Universidad Politécnica de Cataluña, grupos a los que pertenece el Dr. Padró, desarrollador de FreeLing; el grupo del Centre de Llenguatge i Computació CLiC desarrolla *corpora* consultables y analizadores morfológicos. También destaca el grupo IXA de la Universidad del País Vasco, quienes crean sistemas orientados a la traducción. No menos importante es el grupo de Natural Language Processing de la Universidad Nacional de Educación a Distancia NLP UNED, que destaca por trabajar en la minería de datos: búsqueda, seguimiento, evaluación y almacenamiento de información.

2.2.3. Lexicografía de *corpus*

En este ámbito se crean y utilizan los *corpora* por la gran cantidad de léxico y propiedades léxico-sintácticas que aporta el estudio de un *corpus* (Aarts *et al.*, 1992; Baker, 1993; Pérez Hernández, 1994; Moreno Ortiz, 1998; Biber *et al.*, 1998). Se obtienen datos sobre colocaciones, distribuciones, frecuencias, etc. que luego, mediante técnicas cuantitativas y estadísticas, se aplican a diseños de lexicones (Baayen *et al.*, 2007) y, a su vez, a sistemas de compilación de *corpora* (España-Bonet *et al.*, 2009). Las aplicaciones más comunes se orientan a la traducción (Moreno Ortiz, 2000; Alegria *et al.*, 2005), a la enseñanza de idiomas (Sánchez *et al.*, 1995; O'Donnell *et al.*, 2009), o a la extracción de información, análisis y caracterización temática de textos por métodos de Análisis del Contenido (Krippendorff, 2003; Krippendorff y Bock, 2009), por ejemplo. En este ámbito destaca otro grupo de investigación denominado GRIAL, acrónimo de "Grup de Recerca Interuniversitari en Aplicacions Lingüístiques", cuyas líneas de investigación son la sintaxis y semántica verbales, la lexicografía, la anotación de *corpora* y la obtención de información.

2.2.4. Semántica computacional

Esta disciplina trata del estudio del léxico y se basa en fuentes pluridisciplinares como la Semántica Léxica, que contribuyó con sus aportaciones teóricas (Cruse, 1986; Lyons, 1980; Jackendoff, 1990, 1992). Es útil para organizar el léxico, procesarlo y representarlo mediante estructuras de significado. Entre las aportaciones de otras disciplinas a este campo destacamos la Semántica Léxica de la Gramática Funcional y Modelo relacional de Anclaje (Vossen, 1994) en lo que concierne a la relaciones de sinonimia, antinomia, hiponimia e hiperonomia que se establecen entre vocablos. Por ejemplo, el modelo relacional de Wordnet se basa en las redes semánticas de Miller (Miller, 1986, 1990, 1991), que ha derivado en Eurowordnet (Atserias *et al.*, 1997; Díez Orzas, 1999; Climent Roca, 2000)

y, a su vez, ha dado como fruto la versión española esWordnet y catalana caWordnet, desarrolladas ambas por los grupos de investigación TALP, CLiC y PNL UNED. También destacamos la Euskal Wordnet, desarrollada por el grupo IXA (Agirre *et al.*, 2006) y operativa en inglés, italiano, castellano y catalán.

En cuanto a la organización del léxico, considerando la semántica y la función de los vocablos en la elaboración de lexicones (Moreno Ortiz, 1998, 2000), muchos estudios siguen criterios propuestos por autores reconocidos (Coseriu, 1977; Dik, 1980) dentro de lo que se denomina Gramática Léxico Funcional (Faber y Mairal, 1998). Además, aplican modelos psico-cognitivos, como la Teoría de Prototipos, para el estudio y clasificación de locuciones (Ruiz Gurillo, 1997), por ejemplo.

2.2.5. Evaluación linguístico-computacional

Identificamos esta área como un ámbito en expansión dentro del marco lingüístico-computacional para análisis y evaluación del nivel de competencia de discursos de una lengua. Desde hace décadas se ha tratado de valorar la legibilidad, coherencia y comprensibilidad de un texto (Klare, 1974-1975) con métodos menos sofisticados que las propuestas actuales. Desde entonces, las investigaciones avanzan y generan sistemas de conocimiento (McNamara *et al.*, 2006) para poder evaluar automáticamente la calidad de un documento (Dufty *et al.*, 2004) y analizar automáticamente textos especializados (Graesser y Petschonek, 2005). Por ejemplo, la Universidad de Memphis computa discursos y determina valores indicativos del grado de complejidad de los textos analizados. Incluso, el equipo de Memphis es capaz de determinar la riqueza de vocabulario de un texto mediante la herramienta Measurement of Text, Lexical Diversity (MTLD) (McCarthy, 2005); de probar las diferencias entre el inglés americano y el inglés de Gales (Hall *et al.*, 2006); de evaluar la legibilidad de un texto (Crossley *et al.*, 2008) con sistemas de conocimiento y evaluación, conjugando muchas de las disciplinas anteriores o analizando diferencias de competencia escrita entre un grupo de lengua materna (L1) y lengua extranjera (L2) (Crossley y McNamara, 2009). Éstas son algunas de las funciones o aplicaciones de las herramientas que cada día se desarrollan para automatizar tareas rutinarias u obtener resultados cuantitativos de lo que, en múltiples ocasiones, se intuye. Muchos de estos trabajos son el resultado del proyecto de investigación Coh-Metrix, sistema que se articula dentro de los marcos teóricos multidisciplinares mencionados anteriormente.

En la línea propiamente dicha de evaluación automática (Calfee, 2000; Hirschman *et al.*, 2000; Kukich, 2000; Landauer *et al.*, 2003), destaca el equipo de investigación liderado por Landauer, de la Universidad de Colorado, en el ámbito de la calificación de ensayos de forma automática, aplicando métodos estadísticos como el LSA (Landauer y Dumais, 1997; Landauer *et al.*, 1998a,b, 2004), e implementando el rendimiento, por ejemplo, de este método para crear KAT Engine, tecnología que subyace en los productos comerciales conocidos como *Summary Street*, *Intelligent Essay Assesor* (IEA) (Foltz *et al.*, 2000; Streeter *et al.*, 2002; Landauer *et al.*, 2000) y *WritetoLearn* (Landauer *et al.*, 2009). El Institute of Cognitive Science de la Universidad de Colorado ha evaluado el producto "Summary Street" durante los cinco años de proyecto subvencionado por la Interagency Education Research Initiative, y patrocinado por la National Science Foundation, el U.S.

Department of Education and the National Institutes of Health. La creación de estos productos, atractivos y diseñados dentro de criterios de viabilidad, fiabilidad y efectividad en el aprendizaje, se presentan no como un desafío al corrector humano sino como un complemento de economía de trabajo y de objetividad en el proceso de corrección. Los estudios realizados para comprobar la correlación de los resultados de puntuación de un corrector humano con la herramienta "Intelligent Essay Assessor" muestran tener una correlación más alta que la calculada entre correcciones de dos expertos humanos (Group, 2007, 6). Uno de los métodos de evaluación de los textos se basa en cálculos lógico-probabilísticos conocidos como Rasch Lexile Method, aplicado a respuestas cerradas o esperadas.

Entre las universidades norteamericanas destacamos la Universidad de Memphis y la Universidad de Colorado, en la medida de nuestro conocimiento, como punteras en el desarrollo de ingenierías lingüístico-computacionales orientadas a la evaluación textual.

En general, no menos importantes son las investigaciones realizadas por grupos de investigadores de la University of Maryland-College Park, Cisco Systems, Inc., Educational Testing Service, Arizona State University, University of Texas en Austin y la Florida State University para fundamentar la importancia del diseño de sistemas evaluadores (Assessment Design) y de modelos de conocimiento (Knowledge Representations) cuyos entornos, tareas y contenidos evaluables representen modelos reales (Mislevy *et al.*, 2010). Es más, en cuanto a la calificación automática de textos, se propone diseñar variables, ponderadas o no, dependiendo de que se ajusten a un contexto concreto u otro para evaluar automáticamente tareas en función de un *rubric* determinado. Según Bennett,

> "the work on automated essay scoring, for example, has capitalized well on theory in computational linguistics (e.g., Burstein, Kukich, Wolff, Lu, & Chodorow, 1998) and on the psychology of knowledge representation (e.g., Landauer & Dumais, 1997). But there has been little, if any, cognitive writing theory incorporated in any of the automated essay scoring work even though these systems typically issue scores for writing proficiency" (Bennett, 2004).

Efectivamente, el acto cognitivo de escribir puede representarse por un modelo en el que confluyen tres elementos básicos: el entorno de la tarea, el conocimiento-memoria del escritor y el proceso de escritura (planificación, composición y revisión) (Flower y Hayes, 1981; Cassany, 1993). Este modelo está cada vez más presente en los diseños de herramientas comerciales de autoaprendizaje y evaluación, planificando cada fase y procedimiento (tema-audiencia, investigación-documentación, organización de ideas, etc.). Además, en estas herramientas se programa una serie de requisitos y se contempla la toma de decisiones (destrezas retórico-lingüísticas, adecuación de forma y contenido, etc.) para revisar y, por último, evaluar el producto final.

Para terminar esta sección, resaltamos el método aplicado a numerosos exámenes procedentes del Graduate Record Examination (GRE) y del Test of English as a Foreign Language-Internet based Test (TOEFL iBT) por Attali en las últimas investigaciones sobre evaluación automática de textos, cuyos resultados se comparan con los dados por los correctores humanos de dichas pruebas. La finalidad de la investigación de Attali ha sido evaluar tales exámenes con la herramienta *E-rater* versión 2 para probar su validez. Para

ello, analiza tres grandes grupos de textos, atendiendo al tipo de ensayo, para disponer de una muestra significativa. Son analizados los ensayos mediante unas variables genéricas, por un lado y, por otra, mediante unas específicas. Las variables específicas son ocho formales y dos de contenido. Las formales son:

1. gramática: concordancia sujeto-verbo;
2. ortografía: errores ortográficos;
3. uso: errores del artículo y de palabras homófonas;
4. repetición de palabras;
5. organización: introducción, ideas principales e ideas secundarias;
6. desarrollo: número de palabras por cláusula;
7. nivel de vocabulario;
8. ratio de la longitud de las palabras.

Las dos variables de contenido son:

1. la similitud interna del vocabulario del ensayo;
2. la similitud externa del vocabulario del ensayo con el vocabulario de los ensayos del mismo grupo de exámenes.

Se concluye en dicho análisis que el método genérico-formal es apropiado para la evaluación de textos a gran escala, que es un método que se basa en criterios uniformes para cada texto y que permite comparar ensayos por su forma, en vez de por su contenido, característica menos objetiva esta última. Los resultados automáticos que se obtienen muestran que son coherentes con los de los correctores humanos (Attali *et al.*, 2010).

2.3. Métodos lingüístico-computacionales de investigación y análisis

2.3.1. Métodos estadísticos: PCA / LSA

Debido a la importancia que tiene en nuestro estudio el método Latent Semantic Analysis (LSA), especificamos que este método y teoría aplicado a textos se utiliza para extraer y representar el contenido semántico de las palabras en su contexto en un gran *corpus*. Este método estadístico se fundamenta en los principios matemáticos del Análisis por Componentes Principales (PCA). Por ello, nosotros en distintas ocasiones nos referimos a PCA o LSA indistintamente. El método, en cualquier caso, se describe con detalle en el apartado 3.2.3.

En realidad, este modelo computacional estocástico aplicado a textos convierte los vocablos en dígitos dentro de matrices procesables. Los pioneros en aplicar este método a la extracción de contenido de los textos afirman que la representación semántica de los vocablos se puede considerar como nodos en un espacio dimensional (Landauer y Dumais, 1997) o como una red de puntos interconectados (Collins y Loftus, 1975) que, con modelos probabilísticos adecuados, pueden ser capturados (Griffiths y Steyvers, 2002).

El uso de este método se ha ido haciendo muy popular desde que se patentó en 1989 por los propios investigadores (Deerwester *et al.*, 1989) para la extracción de información textual, visual o sonora. Actualmente, algunos de estos mismos investigadores ya han desarrollado productos comerciales para la calificación de textos en inglés (Landauer *et al.*, 2009) o para extraer información con el mismo método, pero conocido como Latent Semantic Indexing (LSI) (Dumais, 1993). No obstante, Biber también aplicó este método estadístico para estudiar las correlaciones entre el discurso escrito y hablado en distintos tipos de géneros. Aunque, en vez de referirse en su estudio a Principal Components Analysis (PCA), lo identifica como Principal Factor Analysis (PFA) (Biber, 1988, 79-97). En definitiva, es un método tan susceptible de ser aplicado en tantas disciplinas que, incluso, se han hecho meta-estudios de LSA sobre los artículos que tratan de este método (Tonta y Darvish, 2010).

En castellano, son cada vez más las aplicaciones de este método. Destaca la evaluación de textos realizada por Venegas en la Pontificia Universidad Católica de Valparaíso (Chile) (Venegas, 2009). También el Grupo de Interés en el Análisis de la Semántica Latente del departamento de Psicología de la Universidad Autónoma de Madrid investiga las conexiones en distintos textos y discursos, en un contexto e, incluso, sus aplicaciones en la creación de autotutores en el ámbito académico con reconocimiento interactivo de la voz (Jorge-Botana *et al.*, 2007) o la evaluación de resúmenes comparados entre la LSA y jueces expertos, con resultados satisfactorios (Olmos *et al.*, 2009).

2.3.2. Sistemas de conocimientos y bases de datos

En este trabajo se utilizan módulos diseñados con funciones y algoritmos computacionales que procesan plantillas y bases de datos concebidas con criterios de Psico-Lingüística y de Semántica Computacional. Por ejemplo, al considerarse el lenguaje como "interconnected nodes in a semantic network" (Collins y Loftus, 1975), se configuran diccionarios con esas características. El diccionario ontológico esWordnet es una base de conocimiento diseñada como un listado de lemas con relaciones semánticas entre muchos de su vocablos que permite no sólo relacionar los lemas entre sí sino hallar el grado de relación de esos lemas en otros contextos.

Las bases de datos son componentes fundamentales de los que se nutren los sistemas de conocimiento. En nuestro estudio las bases de datos son las que llamamos indistintamente lexicones, diccionarios, glosarios y listados de lemas con información distinta glosada en cada uno de ellos. Dentro de los sistemas de conocimiento y base de datos que destacamos para la elaboración de esta tesis son dos:

- *FreeLing*: sistema de conocimiento desarrollado por el Dr. Padró y su equipo de

la Universidad Politécnica de Cataluña (Atserias *et al.*, 2006). FreeLing procesa y analiza morfo-sintácticamente textos y etiqueta vocablos y sintagmas gracias a librerías, diccionarios y módulos estadísticos. Este sistema tiene una interfaz con una demo en internet para realizar los análisis en línea por cualquier usuario. Además, al ser un programa descargable, puede funcionar en un ordenador personal una vez instalado. Tras más de cinco años de ser un sistema de acceso libre y disponible para utilizarse en trabajos de investigación, FreeLing ha incorporado un mayor diccionario, mejoras en sus funcionalidades y aplicaciones, incluso ha introducido nuevos idiomas (Padró *et al.*, 2010a).

- esWordnet: base de datos ontológica desarrollada conjuntamente por el grupo del Centro de Tecnología y Aplicación del Lenguaje y del Habla (TALP), el Centre de Llenguatge i Computació CLiC y el grupo de Natural language Processing de Universidad Nacional de Eduación a Distancia NLP UNED. Es el Wordnet castellano.

Tomando la distinción terminológica que hace Moreno Ortiz entre base de datos y base de conocimientos (Moreno Ortiz, 2000), podemos afirmar que, por ejemplo, el programa de FreeLing, con su lexicón, funciona como una base de conocimiento porque se conforma y recurre a reglas basadas en conocimientos fonéticos, morfológicos, sintácticos, e incluso, semánticos y pragmáticos para el análisis morfo-sintáctico. Nosotros utilizamos dos de sus lexicones, FreeLing 1.5 y FreeLing 2.1. Además de estos dos lexicones, utilizamos un conjunto de diccionarios como base de datos para el análisis y nivelación del léxico, con el objetivo de convertir nuestro sistema de nivelación, a su vez, en un sistema o base de conocimiento.

2.3.3. Métodos lingüísticos de tratamiento y análisis de textos

- Lexicografía computacional. El uso de un buen lexicón en todo procesamiento del lenguaje se basa en dos principios básicos: la multifuncionalidad y la reutilización ya que es un componente básico de cualquier sistema de procesamiento del lenguaje (Moreno Ortiz, 1998). En nuestro estudio, nos fundamentamos, sobre todo, en un buen lexicón cuyos vocablos o entradas están vinculados a un lemario para la identificación y desambiguación morfosintáctica y semántica. Dicho lexicón, el más extenso, es el que denominamos diccionario FreeLing 2.1. Además, se ha dispuesto de un conjunto de lexicones-lemarios como el del Dr. Kincaid, traducido por nosotros para servir de control de nivel; el Wiktionary, lematizado por nosotros para este trabajo; el específico de la Dra. Fuensanta López, digitalizado para procesar vocablos y para diferenciar áreas temáticas de textos; el glosario nivelado del *Plan Curricular del Instituto Cervantes*, que se ha digitalizado; y el listado de conjunciones o multivocablos de FreeLing 1.5, que se ha implementado. Por último, el diccionario de redes semánticas, identificado como esWordnet, ha sido útil también por el gran número de entradas que registra y por sus características intrínsecas.

- Lexicografía de *corpus*. Se han utilizados diferentes *corpora* en formato de texto plano, lematizado con FreeLing para ejecutar diferentes funciones. Estos *corpora*

son:

1. Discursos navideños del Rey Juan Carlos I (1975-2010): 36 discursos que conforman un *corpus* de referencia.

2. Textos académicos escritos por personas adultas nativas españolas que estudian para la obtención del Graduado en Educación Secundaria Obligatoria. Aunque hay estudiantes de dos niveles (I y II), su producción escrita se configura como un único *corpus* formado por 115 textos, clasificados en cartas (nivel I: 21 textos y nivel II: 23 textos), reclamaciones (nivel I: 15 textos y nivel II: 18 textos) y descripciones (nivel I: 13 textos y nivel II: 25 textos).

3. Un número de 23 lecturas niveladas del Centro Virtual Cervantes (CVC) correspondientes a las "Lecturas paso a paso": 11 de nivel intermedio y 12 de nivel avanzado. Además, incluimos un breve texto en gíglico de Julio Cortázar (Cortázar, 1986).

4. Exámenes de candidatos al DELE intermedio (B1-B2) y superior (C1-C2): 20 exámenes por nivel suponen 80 textos en total.

De todos estos textos, unos serán útiles como referencia (1, 2), otros como comprobación, análisis y prueba del sistema (3) y otros, los exámenes del DELE, para ser evaluados (4).

- *Coh-Metrix*: La Universidad de Memphis desarrolla esta herramienta para medir la complejidad de un texto. Contiene sesenta índices cuantitativos que exponemos en la tabla 2.1 y que nos han permitido reflexionar sobre índices equivalentes en castellano. Además de los índices de Coh-Metrix desarrollamos otros que exponemos en el apartado 3.2.2. Coh-Metrix 2.0 es un programa en red desarrollado por el Departamento de Psicología de la Universidad de Memphis para calcular el grado de cohesión y de coherencia de un texto. Mediante la aplicación de modernas herramientas lingüístico-computacionales, se pueden medir unos parámetros establecidos que caracterizan el texto, atendiendo a la sintaxis y cohesión textual. Los datos numéricos que se extraen al procesar el texto permiten determinar la complejidad de lectura de un texto y calibrar a qué grado de competencia lectora se ajusta un cierto texto. En definitiva, Coh-Metrix identifica la legibilidad de un texto. Desde una perspectiva práctica, y en su fase de prueba e investigación, Coh-Metrix calcula la dificultad que puede presentar un texto a unos lectores nativos mediante el cómputo de unos criterios lingüísticos concretos.

  Efectivamente, un texto siempre contiene información con un mayor o menor grado de dificultad conceptual o sintáctica. Por ello, creemos que se produce un fenómeno similar en el proceso de creación de un texto. Si la dificultad de leer un texto puede medirse mediante unos parámetros de forma automática con una herramienta como la de Coh-Metrix, también se podrá otorgar un nivel de complejidad a un texto producido por un aprendiz, ya que se trata de evaluar concretamente dicho texto.

Es sabido que un aprendiz de una segunda lengua, cuando escribe, quiere comunicar y trata de hacerlo de una manera correcta, más o menos compleja, aunque posiblemente se exprese de una forma que "suene mal" a los hablantes nativos debido al fenómeno conocido como variedad de la L2 o interlengua. Sin embargo, respecto a la mecánica del funcionamiento de una lengua, eso no es un indicador de que el aprendiz no sepa gramática o vocabulario para expresarse por escrito, sino que existe una transposición de la mecánica de la propia lengua, una falta de asimilación o desconocimiento del uso de unas estructuras o colocaciones propias de los nativos. Por ello, el modelo de análisis que nosotros presentamos se basa también en la idea de que si es medible la dificultad de lectura de un texto (Simón Granda, 1992), sus parámetros serán extrapolables para evaluar la complejidad de un texto escrito. Según los autores del manual *Saber escribir*, existe una interrelación estructural entre ambas destrezas al afirmar que "la complejidad producida en la lectura de un texto es similar a la que se produce en el proceso de construcción textual" (Cervera *et al.*, 2006, 338).

Como argumento a favor de la relación tan estrecha entre los parámetros medibles de lectura y escritura, hay estudios realizados por MetaMetrics®, Inc. que comparan las dos destrezas: la lectora y la escritora, y miden su diferencia en Lexiles (L) como se ve en la figura 2.3. Estos estudios confirman, con similares parámetros, que la distancia entre la destreza lectora de un sujeto suele ser de 350L mayor frente a la escritora. Es decir, sabiendo que *Lexile*, L, es la unidad de medida de la complejidad de los textos en esta escala de medición, el índice de medida es generalmente menor en la destreza escritora que en la lectora. Como dato de interés, la escala de medida de la complejidad lectora oscila en un rango de 200L (nivel muy bajo) a 1700L (nivel muy alto) (Smith III, 2009).

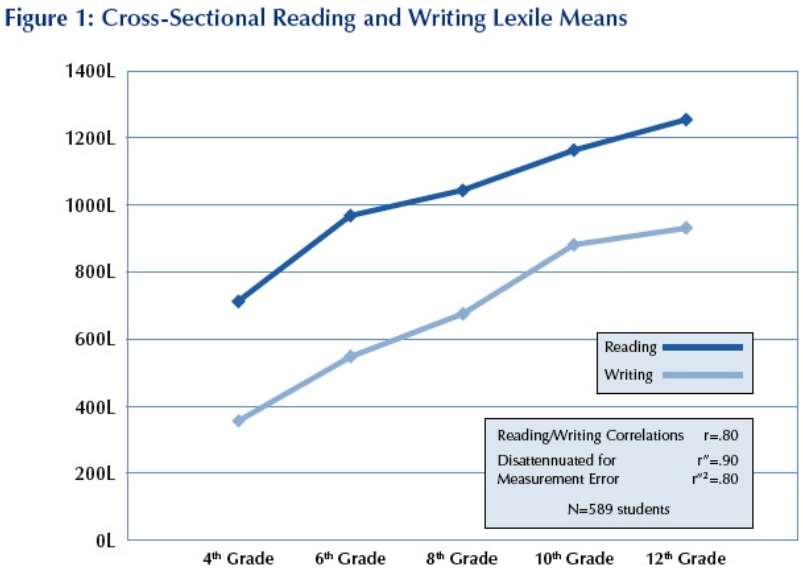


Figura 2.3: Escala de medidas en *Lexiles* para marcar la diferencia entre el nivel de lectura y de escritura en función de los cursos académicos estadounidenses. Figura reproducida de MetaMetrics®, Inc. (Smith III, 2009).

Por ello, apoyándonos en esta concepción postulamos que muchos de los paráme-

tros propuestos en Coh-Metrix para calcular la coherencia y complejidad de una lectura son válidos también para un texto-resultado de un proceso de escritura. En definitiva, se cree que los parámetros de Coh-Metrix son índices medibles y, por tanto, factibles para evaluar cualquier texto porque, igualmente, pueden ajustarse al estudio de la complejidad sintáctica de un texto producido por un aprendiz de una lengua. Concretamente, Coh-Metrix registra, como puede verse en la tabla 2.1, sesenta parámetros para ser procesados de tal manera que sus resultados numéricos luego deben ser interpretados.

Nº	**Índices**
1	Título
2	Género
3	Fuente
4	Código del trabajo
5	Uso de LSA
6	Fecha
7	Referentes anafóricos adyacentes
8	Referentes anafóricos
9	Superposición de palabras-raíz adyacentes
10	Superposición de argumentos
11	Superposición de argumentos adyacentes
12	Superposición de palabras-raíz
13	Superposición de palabras contenido
14	LSA entre oraciones adyacentes
15	LSA entre todas las oraciones
16	LSA entre párrafos
17	Pronombres personales
18	Relación de pronombres con sintagmas nominales
19	Relación de tokens con palabras contenido
20	Contenido causal
21	Cohesión causal
22	Contenido final o intencional
23	Cohesión final o intencional
24	Estructuras sintácticas adyacentes similares
25	Estructuras sintácticas similares en todos los párrafos
26	Estructuras sintácticas similares entre y dentro de los párrafos
27	Cohesión temporal
28	Cohesión espacial
29	Nº de marcadores
30	Marcadores condicionales
31	Marcadores aditivos positivos
32	Marcadores temporales positivos
33	Marcadores causales positivos
34	Marcadores lógicos positivos

Nº	Índices
35	Marcadores aditivos negativos
36	Marcadores temporales negativos
37	Marcadores causales negativos
38	Marcadores lógicos negativos
39	Marcadores lógicos: "y", "o", "entonces" y "si", entre otros
40	Media de frecuencias de palabras contenido (0-1 mill. en la base de datos CELEX)
41	Logaritmo, media de frecuencia de palabras contenido (0-6 en la base de datos CELEX)
42	Mínimo de palabras contenido concretas por frase (0-1 mill. en la base de datos CELEX)
43	Logaritmo, mínimo de palabras contenido por frase (0-6 en la base de datos CELEX)
44	Media de palabras contenido concretas
45	Mínimo de palabras contenido concretas por frase
46	Media de hiperónimos nominales
47	Media de hiperónimos verbales
48	Nº de negaciones
49	Nº de sintagmas nominales por cada 1000 palabras
50	Media de modificadores por sintagama nominal
51	Media de constituyentes de nivel más alto por palabra
52	Media de palabras delante del verbo principal en una oración compuesta
53	Nº de palabras
54	Nº de frases
55	Nº de párrafos
56	Promedio de sílabas por palabra
57	Promedio de palabras por frase
58	Promedio de frases por párrafo
59	Test de Flesch (Reading Ease Score)
60	Test de nivel de Flesch-Kincaid (Cursos 1-12)

Tabla 2.1: Índices de Coh-Metrix.

Como resumen de este apartado, consideramos que el método y herramienta Coh-Metrix utilizados por el equipo de Louwerse para interpretar la complejidad de textos en la compresión lectora son válidos para el estudio de la complejidad de cualquier texto escrito. Es más, la aplicación de sus criterios convertidos en parámetros cuantificables aportan un serie de datos numéricos que, aunque permiten diagnosticar la complejidad y características generales de un texto, e incluso determinar la coherencia a partir de los elementos cohesivos del texto, no nos aportan ningún nivel de lengua concreto del texto, objetivo de esta investigación.

Capítulo 3

Herramientas y materiales para el análisis

3.1. Glosarios

Para procesar el léxico de los exámenes de DELE de los niveles intermedio (B2) y superior (C1) se recurre a distintos diccionarios como medida de control del léxico, identificación de vocablos y reconocimiento de nivel. En general, se les va a llamar indistintamente glosarios y diccionarios porque, aunque son listas de vocablos y lemas, además, llevan información diversa asociada al vocablo.

Tipos de glosarios
Los glosarios electrónicos que utilizamos en nuestra investigación son ocho y los enumeramos a continuación:

1. El glosario del Dr. Kincaid: en principio, es un listado original de 2.000 lemas en inglés desambiguados gramaticalmente pero que, traducido y adaptado al castellano, hemos aumentado a 2.022 lemas.

2. El Spanish Wiktionary o Frequency Spanish List es un listado obtenido a partir de un *corpus* de 27.417.111 vocablos de 6.527 subtítulos de series de televisión. Hemos seleccionado los 10.000 vocablos más frecuentes relacionados con sus lemas correspondientes pero, después de suprimir los lemas repetidos, se han reducido a 5.207 lemas.

3. El glosario de la Dra. Fuensanta López (5.273) o "Vocabulario Básico de Orientación Didáctica (VBOD)" de 5.261 vocablos. La Dra. Fuensanta López reúne sustantivos y verbos relacionados con el vocabulario académico, obtenido a partir de un *corpus* compuesto por textos de libros de bachillerato de distintas materias.

4. El diccionario morfológico de FreeLing 1.5, compilado por el Dr. Padró y su equipo, lista 7.156 lemas y sus formas flexivas. Es un glosario clave porque contiene términos muy frecuentes y es de menor tamaño que el FreeLing 2.1.

5. El diccionario morfológico de FreeLing 2.1-β con 76.214 lemas desambiguados es una ampliación del FreeLing 1.5, realizado también por el Dr. Padró y su equipo de investigación. Este diccionario de lemas, con sus correspondientes formas flexionadas, recoge 556.213 vocablos.

6. El esWordnet utilizado es la versión de esWordnet 1.6 en castellano de noviembre de 2006. Los listados de palabras identificados como "esWN-variants" y "esWordnet-relations" son los que vamos a procesar para el estudio del vocabulario y para sus relaciones semánticas. Este glosario consta de 93.425 lemas y multi-vocablos. Contiene nombres, verbos, adjetivos y lexías nominales, verbales y adjetivales, pero carece de adverbios. Disponemos de este glosario porque nos ha sido cedido por el Dr. Padró para realizar esta investigación.

7. El glosario del *Plan Curricular del Instituto Cervantes* (*PCIC*) o "Índice de nociones generales y nociones específicas" del *PCIC* consta de 8.662 lemas repetidos identificados por niveles. Algunos vocablos han sido revisados manualmente y desambiguados, en cuanto a su categoría gramatical, por nosotros automáticamente.

8. El glosario de Locuciones o multivocablos, basado en el listado original de locuciones de FreeLing 1.5 (locucions.dat) con 1.480 multi-vocablos y de FreeLing 2.1-β (locucions-extended.dat) con 3.158, se ha ampliado hasta 5.868 entradas. Después se ha nivelado y, siempre, es susceptible de ser mejorado.

Características generales de cada glosario

Atendiendo a las distintas características de cada glosario (número de vocablos, especificidad, desambiguación gramatical y nivel de aprendizaje), consideramos que se puede evaluar el léxico de un texto. Es decir, son los criterios intrínsecos a la concepción de cada glosario los que nos van a servir para controlar el léxico y evaluarlo. Por ejemplo, el número de entradas es un rasgo distintivo de cada glosario.

El principio básico de todos estos glosarios es que listan las vocablos más frecuentes a partir de un *corpus* concreto, obteniendo el vocabulario más frecuente y, por tanto, caracterizando al glosario por su tamaño o especificidad. Creemos que el número de glosarios que utilizamos para este trabajo nos permite disponer de un vocabulario básico (Kincaid), cotidiano (Wiktionary), general (FreeLing 1.5, FreeLing 2.1 y esWordnet 1.6), académico (Fuensanta) y específico (Cervantes y Locuciones). Todos ellos son muy significativos para el objetivo de este trabajo. Es decir, según la combinación y pertenencia de un vocablo a uno u otro glosario aplicadas al análisis, la mayor o menor extensión del glosario y la existencia de vocablos desambiguados o específicos, podríamos identificar un determinado vocablo, el nivel del vocabulario de un texto e, incluso, la especificidad del texto.

La oportunidad de poder procesar textos con glosarios de naturaleza diferente nos permite identificar cierto número de vocablos, más o menos frecuentes y más o menos específicos. Por ejemplo, cuanto más comunes o frecuentes son las palabras que contiene un glosario reducido, más básico es el glosario. En consecuencia, cuando tenemos varios diccionarios, generales y específicos, con un número de vocablos menor o mayor, mejor se puede calcular el nivel y tipo de léxico de un texto. Es decir, utilizar varios glosarios con unas características concretas y diferentes entre sí permite reconocer un mayor número de

vocablos de un texto e identificar el tipo de vocablo gracias a la concepción específica del diccionario. Incluso, la combinación de todos ellos permite determinar el nivel de un texto en función del número diferente de vocablos que están o no en uno u otro diccionario. Además, el método de utilizar diferentes diccionarios y combinarlos evita una división arbitraria del vocabulario. Como apunta Nation, la frecuencia, la representatividad y rango de vocablos se presentan en los textos como un *continuum* (Nation y Kyongho, 1995, 37). Por ello, un método para estimar el nivel léxico es recurriendo a los diccionarios como referentes (Nation, 1993).

Función de los glosarios

En concreto, el uso de uno o varios glosarios de autoridad, extensos o breves, específicos o generales, como los glosarios *ad hoc* empleados para esta investigación son importantes porque permiten utilizar sus frecuencias y rangos para posteriores análisis y, sobretodo, controlar e identificar el máximo número de vocablos que se procesen en un texto. Por ejemplo, la propuesta del Dr. Kincaid de crear un glosario básico o de control para generar textos normalizados y diagnosticar el nivel de un texto, nos ha llevado a sugerir que, mediante el procesamiento de un texto con un buen glosario básico, se puede identificar, diagnosticar y nivelar gran parte de las palabras de un texto. Es decir, al procesar un texto mediante un glosario básico podemos identificar cierto porcentaje de sus vocablos que nos indica que el texto tiene, al menos, un nivel de lengua B1 y que, cualquier otro vocablo, implica una riqueza de vocabulario que apunta hacia un nivel B2 y C1. No obstante, un porcentaje alto de repetición y la especificidad del léxico en un texto nos indicaría que dicho texto tiende a un nivel C2.

Además, muchos autores se basan en diccionarios con un número acotado de vocablos para diagnosticar el nivel de un texto. Algunos recurren a listados de mil o dos mil lemas (Laufer y Nation, 1995) o a un determinado Academic Word List (AWL) (Morris y Cobb, 2004) para medir el nivel léxico con fines académicos; otros, como la Flesch-Kincaid Grade Level Formula o la Dale-Chall Formula, utilizan un listado de 3000 lemas para identificar distintos grados de lectura (Chall y Dalee, 1995). Siguiendo este mismo método para diagnosticar el nivel léxico de un texto, el evaluador Oxford, conocido como Oxford 3.000 TM Text Checker de Oxford University Press, también basado en un listado electrónico de 3000 lemas, identifica de forma automática el nivel de lengua de un texto en inglés. Con este diccionario se califica el nivel de un texto en relación al porcentaje de vocablos que coinciden con el listado de 3000 palabras. El Oxford 3000TM Text Checker propone que:

- Un texto propio de un nivel intermedio bajo [B1] reflejaría casi un 100 % de los vocablos del diccionario Oxford 3000TM Text Checker.
- Un texto propio de un nivel intermedio alto [B2] utilizaría entre un 90 %-95 % de los vocablos del Oxford 3000TM Text Checker.
- Un texto propio de un nivel avanzado [C1] recogería entre un 75-90 % de los vocablos del Oxford 3000TM Text Checker.

La evaluación léxica automática de un texto, realizada por el Oxford 3000 TM Text

Checker, se basa en una lista de 3000 lemas desambiguados. Esta lista de vocablos funciona como una lista de control similar a la que elaboró el Dr. Kincaid en los años setenta a partir de un gran *corpus* de textos muy variados para hallar las palabras más frecuentes. Análogamente, creemos que nuestra lista de control, basada en la original de 2000 lemas del Dr. Kincaid, traducida y adaptada al castellano con 2022 lemas, podría funcionar de forma similar a la de Oxford 3000 o VocabProfile. Este último programa, también para el inglés, apunta el nivel léxico de un texto según criterios de frecuencia y aparición de un cierto número de vocablos en dicho texto. Ha sido desarrollado por Laufer y Nation y conocido con el nombre de WebVP v3 Classic, practicable en internet, o de *Range*, como descargable y ejecutable en un ordenador personal (Goodfellow *et al.*, 2002).

3.1.1. Glosario del Dr. Kincaid

El vocabulario simplificado en inglés del Dr. Kincaid, originariamente, consta de 2000 palabras y están desambiguadas respecto a su categoría gramatical. Los 2000 lemas del Dr. Kincaid fueron seleccionados, siguiendo el criterio de frecuencia, de un gran *corpus* periodístico de temática muy diferente. Este glosario lo tradujimos al castellano, en principio, de forma literal e intuitiva. Posteriormente, se ha afinado la traducción y, en agradecimiento por la cesión de su glosario y permiso para utilizarlo en este trabajo, se le ha entregado traducido al propio Dr. Kincaid. Sin embargo, su listado en inglés circula en internet sin ser reconocida su autoría.

3.1.1.1. Origen del glosario del Dr. Kincaid

El glosario del Dr Kincaid es un diccionario simplificado, básico que sirvió de control del vocabulario de manuales de estudio y mantenimiento aeroespacial con el fin de estandarizar y simplificar la información con un vocabulario tan frecuente como básico. La necesidad de formar personas para estudiar con facilidad en un entorno militar motivó a un equipo de investigación dirigido por el Dr. Kincaid a crear un vocabulario mínimo que fuera útil para elaborar manuales sencillos y aptos para trabajadores de mantenimiento de origen y cultura diferente a la norteamericana, y con una competencia lectora media-baja. La creación de un vocabulario básico fue una parte importante del proyecto dentro del área de investigación que buscaba un uso simplificado de la lengua para leer fácilmente instrucciones de manuales de mantenimiento. Este área se conoce actualmente como *Simplified (Thecnical) English* o, también, *Controlled English*, campo de investigación en el que el Dr. Kincaid fue uno de los pioneros en los años 70. Posteriormente, el Dr. Kincaid ha seguido trabajando en este área llegando a crear el programa *Simplified English Analyser* (SEAN) de manera que traduce textos de inglés estándar a inglés simplificado (Thomas *et al.*, 1992) y es útil en la elaboración de modelos de actuación y mantenimiento para casos de emergencia (Jenvald *et al.*, 2001).

Se considera muy interesante este vocabulario para este trabajo porque es un léxico seleccionado a partir de vocablos muy frecuentes, porque cada vocablo está desambiguado al estar etiquetado con su correspondiente categoría gramatical y porque su extensión es breve. Una utilidad más que le otorgamos nosotros a su concepción original es que sirve

también de control, pero de forma similar a la del diccionario de Oxford 3000, esto es, que diagnostiquemos el nivel de un texto a partir de este glosario de control como se ha expuesto en la página 35.

En consecuencia, el glosario del Dr. Kincaid permitirá ejecutar el "método de diagnóstico de control" sobre un texto cualquiera para determinar si contiene un léxico básico. Al mismo tiempo, al aplicar el método de "combinación de diccionarios", obtendremos los límites entre uno y otro glosario (Nation y Kyongho, 1995). Además, podremos distinguir la especificidad de discursos por el porcentaje de vocabulario perteneciente a un listado de cierto tamaño o a glosarios más especializados, como el de la Dra. Fuensanta López o el "Índice de nociones generales y nociones específicas" del *Plan Curricular del Instituto Cervantes* (*PCIC*).

3.1.1.2. Adaptación del glosario del Dr. Kincaid al castellano

Para configurar el glosario procesable actual del Dr. Kincaid se han realizado una serie de pasos. En un principio, se ha traducido cada lema respetando fielmente la categoría gramatical identificada por el Dr. Kincaid. A continuación, se ha etiquetado cada vocablo con su *PoS* (*Part of Speech*), se ha verificado y rectificado manualmente el lema resultante después de procesarlo con el glosario de FreeLing 1.5. Lógicamente, hemos adoptado la nomenclatura de FreeLing para etiquetar los lemas del glosario del Dr. Kincaid. Por lo tanto, dadas las características del castellano y la posibilidad de etiquetar cada vocablo con FreeLing, se ha implementado el *PoS* del listado del Dr. Kincaid con el género y número de determinantes, sustantivos y adjetivos.

Posteriormente, para adaptar la versión inglesa del Dr. Kincaid a la castellana, se ha aplicado metodológicamente una serie de criterios que detallamos a continuación.

Desdoblamiento o reducción de dobletes

A pesar de ser un vocabulario de 2000 lemas, se ha tenido que trabajar con los vocablos porque, en la revisión, aparecían traducidos dobletes léxicos etiquetados por el Dr. Kincaid como "some" determinante [det] y pronombre [pron] equivalente a "alguno". Para este vocablo, al traducirlo al español sin contexto sintáctico, hay que distinguir además entre "algún" [det] y "alguien" [pron]. Por el contrario, ha sido menor el número de términos que se han simplificado a uno sólo. Este sería el caso de "chairman" y "president" que los reducimos sólo al nombre de "presidente" [n].

Posteriormente, cuando se procede a ordenar el léxico castellano alfabéticamente, se observa léxico repetido. Es decir, hay vocablos distintos en inglés que, al traducirse, coinciden en un mismo significante en castellano, como por ejemplo, "output" [n] y "outcome [n]. Entonces, se mantiene la categoría gramatical de "resultado" [n] como nombre pero se incluye también "resultado" [v] como participio de "resultar".

En consecuencia, cuando se dan repeticiones y hay varias posibilidades para traducir un vocablo, el método que se ha seguido para elegir un término u otro es el de consultar el índice de usabilidad calculado por la Dra. Fuensanta López para nombres y verbos y el de optar por el término con mayor índice de "usabilidad". También se han consultado

aquellas palabras que aparecen en el diccionario de FreeLing 1.5 ya que, aunque tiene un léxico más amplio, recoge aquellas palabras más frecuentes obtenidas también de un gran *corpus* creado a partir de muchos y distintos tipos de textos. Además, como en la lista de la Dra. Fuensanta López no se recogen adjetivos, adverbios ni conjunciones se ha optado por los que registra FreeLing 1.5.

Aunque en un principio la traducción del Vocabulario Básico del Dr. Kincaid fue intuitivamente literal, posteriormente se ha modificado la traducción de ciertos términos dadas las peculiaridades lingüísticas del castellano frente al inglés y las diferencias entre la cultura anglosajona y la española. Por este motivo, el número de palabras del listado original ha aumentado ligeramente a una veintena de palabras más debido, por ejemplo, a casos como "home" [adv] que se ha traducido con dos lexías "en_casa" y "a_casa" manteniendo su valor adverbial. A su vez, en el léxico inglés el determinante "this" [det], que sólo aparece como determinante, se ha traducido con dos categorías propias del castellano: éste [pron] y "este" [det] en sus dos posibles grafías en castellano del pronombre "este" [pron] y éste [pron], con y sin tilde; igualmente para "kilómetro", introducimos las dos abreviaciones de "km" [n] y "kms"[n].

Criterio de selección por sinonimia y usabilidad

A lo largo del proceso de traducción, también se han descartado algunos posibles significados de palabras sustituyéndolos por un sinónimo que no estuviera indexado todavía o bien tuviera un índice de usabilidad alto, según el vocabulario de la Dra. Fuensanta López. Por ejemplo, hubo dificultad para elegir la traducción de "way" [n]. Aunque al principio se tradujo como "camino" y "manera", luego se optó por "vía" [n] (14,52). Además, también la existencia en el vocabulario del Dr. Kincaid de "shape" [n] como "forma" (196,80), "form"[n] como "modo" [n] (131,79), "manner" [n] como "manera" (48,98) y "path" [n] como "camino" (20,16) dificultaba la elección de un determinado vocablo para "way" [n].

Sin duda, ante dos posibilidades de selección entre "talla" (3,3) o "tamaño" (24,79) para "size" [n], se opta por el término con mayor usabilidad. De forma semejante, se prescinde de "membership" con el significado de "membresía". Éste es un nombre específico y abstracto que no figura en el léxico de la Dra. Fuensanta López ni siquiera en FreeLing 1.5. Por ello, se opta por "miembro" (24,79) en vez de "socio" (1,32) aunque la lista de vocabulario del *Plan Curricular del Instituto Cervantes C1-C2* incluye uno con un nivel C1 (Cervantes, 2006, C1-C2: 748) y otro con un nivel B2 (Cervantes, 2006, B1-B2: 759), respectivamente. Para el verbo "understand" se puede traducir por "entender" (22,40) y "comprender" (29,40). En este caso, optamos por incorporar a la lista los dos vocablos por su proximidad de uso. Ante los nombres "talk" [n] y "speech" [n] se puede traducir por "discurso" (2,16) o "charla" indistintamente. Se ha indexado "discurso" para "talk" [n] porque es el que tiene un coeficiente en la lista de la Dra. Fuensanta López, mientras que "charla" no se registra, al parecer, por no ser una palabra tan usada en textos académicos. Además, a favor de esta elección, en la lista de vocabulario del *Plan Curricular del Instituto Cervantes C1-C2*, se califica "charla" con un nivel de vocabulario C1 (Cervantes, 2006, C1-C2: 727) y "discurso" con un nivel B2 (Cervantes, 2006, B1-B2: 734).

El listado del Dr. Kincaid recoge términos como "policía" (3,28) con índices de usabilidad menores que otros que ni siquiera lista como "riqueza" (9,10) o "contradicción"

(4,80), pero debemos seguir el listado elaborado por el Dr. Kincaid basado en su criterio de listado básico lo máximo posible. Cuando ocasionalmente se ha probado nuestra intuición con el índice de usabilidad, la elección intuitiva de un vocablo ha resultado favorable en la mayoría de las ocasiones. Por ejemplo, para "fuel" [n] se optó por "gasolina" (1,85) en vez de "combustible" (4,00). Como es lógico, el glosario de la Dra. Fuensanta López registra un vocabulario más académico que básico o cotidiano. Es más, FreeLing 1.5 no recoge el término "combustible" pero sí "gasolina" porque el criterio de indexación de palabras se basa, como el del Dr. Kincaid, en la frecuencia de aparición de términos en distintos tipos de textos académicos, periodísticos y divulgativos.

Paralelismo
Con este criterio nos referimos a vocablos que directamente se asocian con otros por ser complementarios o duales. Palabras como "left" [adj] que se traducirían como "dejado" o "izquierdo", al poder ser ambos adjetivos en castellano, se opta por el nombre "izquierda" (33,37) pero con lema "izquierdo" [adj]. Además, "dejado" [adj] se procesaría con el lema "dejar" (69,70), que ya queda indexado en la lista del Dr. Kincaid como "leave" [v]. También, por paralelismo con "right" que aparece como [adj] y [adv], "right" se traduce como "derecho" [adj] y "a_la_derecha" [adv]. Este último lema se recoge como un multi-vocablo ya que, aunque el listado del Dr. Kincaid no contiene multi-vocablos, este criterio de traducción se hace imprescindible para traducir al castellano ciertos adverbios del inglés.

Incorporación de multi-vocablos
Se han creado multi-vocablos que se corresponden con una sola palabra en el listado original del Dr. Kincaid pero que en castellano es preciso uno o más vocablos para identificar su vocablo correspondiente en inglés. Por ejemplo, los vocablos "weekend" [n] y "tonight" [adv] se convierten en lexía o locución al ser "fin_de_semana" y "esta_noche" respectivamente. Semejantes ejemplos son los pronombres "myself", "yourself", "himself", "ourselves" o "themselves" como "yo_mismo", "tú_mismo", "él_mismo", etc.; la conjunción "once" [conj] se correspondería con "una_vez_que" mientras que "somewhere" y "sometimes" se han traducido como "alguno_sitio" [adv] y "alguna_vez" [adv].

Otra licencia que nos hemos tomado en la traducción ha sido la de los modales "may", "can", "could", "might". Consideramos que estos verbos no pueden traducirse simplemente como el verbo "poder". Así que se han traducido respectivamente como multi-vocablo o locución de posibilidad como siguen: "es_posible_que" [conj], "poder" [v] y "puede_que" [conj], "ser_capaz" [v] y "ser_capaz_de [v], "poder_ser" [v] y "puede_ser_que" [conj] y "podría_ser_que" [conj]. Por ejemplo, la locución verbal "<poder>_ser_que" se ha listado como multi-vocablo ya que el programa FreeLing identifica la conjugación, el tiempo y la persona de las formas verbales de estas estructuras perifrásticas y conjuntivas.

Reducciones
Se han reducido palabras inglesas como "other" [det], [pron], [n] y "another" [det] y se han identificado ambas con "otro". A pesar de que el Dr. Kincaid distingue "other" [det], [pron] y [n], en la traducción castellana se diferencia sólo "otro" como [det] y [pron]. En

cuanto al significante de algunos vocablos, se han reducido los alófonos "organization" / "organisation", "realise" / "realize" y "not" / "no" pero, a su vez, se han introducido en la versión castellana otros fenómenos semejantes como "quizás" / "quizá" [adv], a nivel gráfico, o "capital" [n] con género masculino y femenino, a nivel gramatical.

Como no se distingue en castellano la diferencia entre las preposiciones inglesas "on" / "in", se reduce la entrada a la preposición "en" [prep]. Se asocia el término "más" [adv] a "plus" [prep], "more" [adv] y [det] y "else" [adv]; y "menos" [adv], a "less" [adv] y [det] y "least" [adv]. Es decir, solamente se han traducido como un adverbio, a pesar de las diferentes categorías gramaticales que distingue el listado del Dr. Kincaid.

Omisiones
Se elimina "to" como partícula formadora de infinitivo ya que en castellano no es un elemento gramatical independiente del verbo.

Repeticiones
En el caso de los pronombre flexivos, se repiten varias veces lemas en el listado del Dr. Kincaid y su traducción. Por ejemplo, con el lema "yo" se asocian las formas flexivas como "nosotros", "nosotras" o "nos", incluso en mayor número en castellano que en inglés. Estos vocablos, al ser traducidos y listados, han aumentado el listado de vocablos y se ha repetido varias veces el mismo lema.

Modificaciones gramaticales
Respecto a adjetivos en inglés que se identifican con un participio en español, en algunos casos se ha optado por listarlos con el lema del verbo. Por ejemplo, el adjetivo "busy" [adj] "ocupado" se procesaría con el verbo "ocupar" (75,24) en vez del adjetivo-participio "ocupado". En otros casos en que coincide un vocablo en su forma de participio y de nombre, se ha mantenido la categoría gramatical que tiene en el listado del Dr. Kincaid. Un ejemplo es "defendant" [n], "acusado", que se ha mantenido como nombre en vez de como participio del verbo "acusar". El glosario de la Dra. Fuensanta López, una vez más, ha permitido decidir la validez de la elección para la elaboración del listado del Dr. Kincaid ya que el verbo "acusar" (5,32) tiene una baja usabilidad.

Desdoblamiento
Nos hemos encontrado con términos que apuntaban a dos sinónimos o requerían dos significantes al ser traducidos al castellano. Por ejemplo, el verbo "achieve" [v] al poder ser traducido indistintamente por "lograr" (24,00) y "conseguir" (24,48) y tener un índice de usabilidad similar y alto, se han listado los dos verbos.

Gramaticalmente, se produce otro desdoblamiento al traducir "whom" al castellano. Se precisa de la preposición "a" o "para". En este caso, se indexa "who" como pronombre nominativo "quien" [pron] y "whom" [pron] con dos multi-vacablos en dativo: "a_quien" y "para_quien".

Respecto el verbo "welcome" [v], para evitar un multi-vocablo "dar_la_bienvenida", introdujimos el sustantivo "bienvenida" [n] y el verbo "saludar" (0,64) por mantener la

categoría del listado del Dr. Kincaid, en detrimento del sustantivo "saludo" [n] que propone el *PCIC* (Cervantes, 2006, C1-C2: 758).

En inglés, algunos significantes son distintivos sólo por la categoría a la que pertenecen. Igualmente en castellano, algunos nombres, bien nombres o bien participios, coinciden en su significante, pero tienen categorías y significado diferentes. Por ejemplo, el nombre "encargado" [n] se corresponde con "manager" [n] y el verbo "encargarse" con "handle", aunque en castellano es pronominal. Otro ejemplo es el nombre "treaty" [n] equivalente a "tratado" [n] y el verbo "treat" [v]. Por lo tanto, ambos términos se han desambiguado gramaticalmente ya que al procesarse de forma automática iban a coincidir con el lema verbal "encargar" y "tratar" respectivamente.

Cuasi-sinónimos
Hay palabras homónimas que se procesan con ambigüedad semántica al tener un significante idéntico pero un significado diferente. Un ejemplo son aquellos verbos que requieren la partícula pronominal "se" para precisar el significado como "dar_cuenta" por "darse_cuenta" . Incluso, "acordar" por "acordarse" referido a "remember" entra en conflicto con "agree", así que traducimos éste como "estar_de_acuerdo", con un multi-vocablo.

Hay algunos nombres en inglés que están próximos en su significado como "move" [n], "movement" [n], "motion" [n]. Así que en vez de traducirlas todas por "movimiento" (157,85), se ha optado por otros vocablos con alto índice de usabilidad como "desplazamiento" (13,68) y "circulación" (8,19) ya que se recogen también en FreeLing 1.5.

Debido a la utilidad y frecuencia del término "item", ha sido difícil encontrarle una traducción cuando ya están listados "part" como "parte" (280,56), "element" como "elemento" (154,07), "thing" como "cosa" (59,95), "stuff "[n] como "utensilio" (3,85), "component" como "componente" (20,70), "bit" como "trozo" (9,80) "section" como "sección" (16,08), "piece" como "pieza" (13,86) o "article" como "artículo" (8,06). Así que, entre otras traducciones, dada la posibilidad de optar por los sustantivos "apartado" [n] (6,50) o "pedazo" (0,62), se opta por "apartado" para "item".

Igualación de categorías gramaticales
El inglés presenta más vocablos de pronombres indefinidos con diferentes categorías que el castellano en términos como "nadie" [pron] y "nada" [pron] y [adv]. En inglés se listan los siguientes significantes: "no-one" [pron], "nobody" [pron], "anybody" [pron] y "none" [pron], y "nothing" [pron], respectivamente. Al contrario, el inglés sólo tiene la forma de determinante "the" [det] y, aunque en castellano existen "el", "la", "los" y "las", registramos sólo el lema "el" [det]. Lo mismo ocurre con las formas pronominales "le", "la", "los" y "las", porque con el lema "él" [pron] se computan el resto de las formas flexivas del pronombre de tercera persona de forma automática.

En el listado inglés se distinguen distintas categorías gramaticales para un mismo término como "very" [adv] y [pron], pero el castellano no registra la del pronombre. Así que, en nuestro listado adaptado y traducido, se mantiene "muy" como adverbio [adv] pero se excluye la categoría de pronombre [pron]. A la inversa, ocurre con "all" [adv] y [det]. El listado del Dr. Kincaid no contempla la categoría de pronombre. El léxico gramatical

de FreeLing distingue en "todo" las categorías de determinante, pronombre y nombre. Por tanto, ampliamos "all" [pron] como pronombre y, además, se incluyen dos adverbios multi-vocablos: "al_completo" o "por_completo", aunque con significado algo diferente con respecto a "all" [adv] adverbio. En algunos casos, ante la falta de índices de usabilidad para adjetivos y adverbios en el glosario de la Dra. Fuensanta López, porque sólo registra nombres y verbos, se ha recurrido al glosario de FreeLing 1.5 para optar por un vocablo o una categoría determinada.

Adaptación cultural
Vocablos como "farmer" o "farm", traducidas como "granjero" o "granja", no son frecuentes en castellano, no aparecen en el glosario de la Dra. Fuensanta López ni en el de FreeLing 1.5. Así que "granja" se ha traducido por "huerto" [n] (0,92). La adaptación de otro término fue "yard", traducido como "metro" (81,95) o "patio" (0,47). Se resolvió introduciendo los dos términos por tener distintos significados.

3.1.2. Glosario de Wiktionary

El glosario Wiktionay de español está disponible en Internet. Matthias Buchmeier lo elaboró en 2008 con un *corpus* de 6.527 subtítulos de películas y series de televisión compilando un total de 27.417.111 vocablos. A su vez, este listado está dividido en diez sub-listados de 1000 vocablos cada uno; en cada sub-listado, cada vocablo está numerado por el rango, le sigue el número de frecuencia de aparición de dicho vocablo en el *corpus* y, por último, el lema. Aunque cada vocablo va acompañado de su lema, éste no está desambiguado por categoría gramatical. A pesar de que el listado íntegro consta de 225.000 vocablos, nosotros optamos por el listado de las 10.000 primeras palabras más frecuentes de manera que, simplificando los lemas repetidos y limitando a 5 el número de frecuencia de palabras por millón (ppm), su frecuencia abarca de 32.894 hasta 5 ppm.

Se muestra en la tabla 3.1 cómo se organizan los vocablos en el listado de Wiktionary, a modo de ejemplo.

rango	vocablo	frecuencia (ppm)	lema
1	que	32.894	que
...			
35	como	3.760	como comer
...			
68	fue	1.811	ser ir
69	ser	1.782	ser
...			
1000	carrera	88	carrera

Tabla 3.1: Organización de vocablos en el listado de Wiktionary.

A la hora de configurar el glosario del Wiktionary se han tomado varias decisiones. Por un lado, cuando un vocablo aparece asociado a dos lemas, estos se desdoblan y se registran dos lemas separados. Esto es, si "fue" tiene dos lemas, uno "ser" y otro "ir", se

computan uno y otro lema por separado (este ejemplo figura en la tabla 3.1). Por otro lado, se han eliminado los lemas repetidos obteniendo un listado definitivo de 5.207 lemas distintos. Por ejemplo, cada forma conjugada del verbo, como "tengo", "tenido", "tendré", se asocia al lema del verbo "tener". Así se eliminan todas las realizaciones de los vocablos y listamos sólo un lema, de manera que se suma la frecuencia de los vocablos flexivos correspondientes a un lema y se computa un solo lema. También, se han suprimido los nombres propios como "Juan" basándonos en el criterio de mayor frecuencia y menor rango del vocablo en el listado.

No obstante, otra carencia de este glosario es que no está desambiguado en cuanto a categorías gramaticales. Sin embargo, nos parece de interés porque incluye más lemas que el glosario del Dr. Kincaid, porque se puede utilizar este listado para hacer una comparativa con el resto de glosarios que utilizamos en nuestro análisis y porque incluye palabras no sólo frecuentes en el lenguaje hablado sino incluso más cotidianas e informales. Por ejemplo, el Wiktionay de español recoge el vocablo "camino" que no sólo consideramos más frecuente y estándar respecto a "sendero" sino que "sendero" es un vocablo que no se lista. Además, este glosario incluye índices numéricos de frecuencias que nos van a ser útiles para determinar el rango que ocupa cada palabra cuando apliquemos la ley de Zipf (Zipf, 1932, 1949).

El motivo que nos lleva a utilizar estos glosarios, aparentemente poco significativos, es que nos resultan útiles porque nos van a permitir distinguir mejor entre niveles intermedios (B1-B2) y superiores (C1-C2) al diferenciar qué vocablos son propios de un nivel u otro cuando se comparan y combinan los glosarios entre sí y, posteriormente, cuando hallemos la calificación de un texto.

3.1.3. Glosario de la Dra. Fuensanta López

El supuesto de Atienza de que el nivel de los textos de B2-C1 de aprendices de español se aproxima al nivel lingüístico que pueda tener un estudiante de Educación Secundaria (Atienza-Cerezo, 1992) es uno de los motivos por el que consideramos interesante contar con el glosario de la Dra. Fuensanta López, creado a partir de un *corpus* de libros de texto de bachillerato (López Martín, 1999).

Este glosario utilizado para nuestro análisis es el resultado del trabajo de investigación realizado por la Dr. Fuensanta López en el departamento de Lengua Española y Lingüística General de la Universidad de Murcia en 1999 y dirigido por el Dr. Ramón Almela Pérez. En su tesis propone un vocabulario básico para bachillerato partiendo de la reducción y actualización de dos registros de vocabulario anteriores compilados ambos por el profesor García Hoz (López Martín, 1999, 181 y ss): uno, realizado entre 1970-1973 y titulado Vocabulario General de Orientación Científica (VGOC) con un registro de 25.402 términos (nombres, determinantes, verbos, adjetivos, adverbios y preposiciones); y otro, el Vocabulario General de Mayor Frecuencia (VGMF) realizado entre 1990-1994 (López Martín, 1999, 203-204).

Ambos índices se crearon a partir de un *corpus* de textos de bachillerato de trece materias (Matemáticas, Física, Química, Biología, Zoología, Botánica, Geología, Literatura, Lengua, Historia, Geografía, Historia de la Filosofía y Fundamentos de la Filosofía)

correspondientes a los planes de enseñanza de los años 70 y 90.

La Dra. Fuensanta López en su tesis, además de comparar los resultados de la frecuencia y dispersión de los términos en trabajos anteriores, ha reducido los registros. Si García Hoz aportó un vocabulario de 25.402 vocablos en el registro de 1970-1973, la Dra. Fuensanta López lo redujo a 7.582 términos, indexando únicamente nombres y verbos.

Después, uniendo los registros de 1970-1973 y de 1990-1994 y basándose en el "criterio de frecuencia" (204), optó por reducirlo hasta obtener un total de 8.080 de los que 6.561, un 81,20 %, son sustantivos y 1.519, un 18,79 %, son verbos. Luego, reorganizó las trece materias en cuatro áreas (1= matemáticas, 2= ciencias naturales, 3= ciencias sociales, 4= lengua y literatura).

A continuación, redujo más el número de vocablos aplicando el principio de efectividad. Este criterio supone que unos 4.000 vocablos son suficientes, o 3.000, entre sustantivos y verbos, según Rivenc (López Martín, 1999, 224), para crear un núcleo de vocabulario que el alumno domine, con el fin de comprender un texto científico y de expresar un pensamiento lógico y científico (228).

De manera que, aplicando el criterio de dispersión, la Dr. Fuensanta López redujo la lista de palabras de 8.080 a 5.012 vocablos aunque, actualizando la terminología, agregó otros nuevos términos léxicos que marcó con un asterisco hasta obtener 5.261 lemas. Aunque en nuestro cómputo, después de contabilizado el listado definitivo de la Dra. Fuensanta López, hay 5.273 lemas indexados según su usabilidad (63 y 79).

En definitiva, la Dra. Fuensanta López comprobó que existe un vocabulario común básico científico para expresar nociones generales o fundamentales (de medida, peso, relación, etc.) y métodos de pensamiento lógico de un contexto científico, además de un vocabulario básico y permanente en el tiempo dentro de cada disciplina.

Finalmente, la Dra. Fuensanta López organizó el listado definitivo de manera que cada vocablo va acompañado por 5 índices:

- Frecuencia relativa: aparición de un vocablo por cada 100.000 vocablos.
- Dispersión: medida de la aparición de los vocablos entre diversos tipos de textos. Su definición está ligada a la desviación estándar. Valores altos indican una distribución homogénea entre áreas, y valores bajos representan una acumulación de las apariciones en una o dos áreas.
- Uso: permite ponderar la frecuencia con la dispersión. Su valor se obtiene al multiplicar la frecuencia por la dispersión.
- Áreas: indica las áreas en las que aparece el vocablo.
- Materias: indica las materias en las que aparece el vocablo.

Además, en su listado se marca con un asterisco aquel término novedoso que no estaba incluido en el listado del Dr. García Hoz. Por otro lado, cada vocablo se distribuye en uno de los cinco niveles establecido por criterio de dispersión, como se especifica en la tabla 3.2

Nivel	**Lema** (* nuevo)	**Frecuencia relativa**	**Dispersión**	**Uso**	**Áreas**	**Nº de materias**
I (100-80)	acción	119	81	96,39	1.2.3.4	13
II (70-60)	abandonar	14	76	10,64	1.2.3.4	13
III (50-40)	abandono	3	56	1,68	2.3.4	5
IV (30-20)	clan*	2	27	0,54	2.4	2
V (10-1)	aberración	2	5	0,10	1.2	2

Tabla 3.2: Estructura del glosario de la Dra. Fuensanta López.

Los niveles están organizados en función del índice de dispersión en vez del índice de frecuencia o usabilidad. Por un lado, este parámetro de dispersión describe la distribución por materias de una palabra. Además, muestra "el grado de interrelación que presentan los términos con vistas a la integración de las enseñanzas; aparecen así desde palabras con significación más amplia o general hasta las más especializadas con significación concreta" (214). Por otro lado, el coeficiente de usabilidad (U) $U = F \times D/100$ es el producto de la frecuencia (F) por la dispersión (D), y resulta de gran utilidad para la elaboración de un diccionario básico de una lengua y para una planificación léxica curricular (222-223).

Gracias a todos los datos numéricos que nos proporciona el glosario de la Dra. Fuensanta López, en nuestro trabajo se ha utilizado no sólo el parámetro de frecuencia, que contribuye a la distribución de vocablos por niveles sino el parámetro de distribución de un vocablo en cierto nº de materias para identificar un determinado vocablo en una área temática. Ambos parámetros son procesables para aplicarlos a nuestro estudio.

Entre algunas aplicaciones generales que señala el trabajo de la Dr. Fuensanta López, especifica que su vocabulario puede "ser utilizado como material base para la enseñanza del castellano a estudiantes extranjeros [...] y para futuros trabajos de investigación" (López Martín, 1999, 193). Concretamente, en nuestro estudio, el glosario de la Dra. Fuensanta López ha sido de gran utilidad. Por un lado, nos ha ayudado a optar por uno u otro vocablo en la traducción del listado básico castellano de 2022 vocablos a la hora de traducir del listado básico inglés de 2000 vocablos del Dr. Kincaid. Por otro lado, queremos aplicar el parámetro de dispersión y el criterio de distribución de un vocablo en cierto número de materias para probar si se puede identificar, por una parte, el nivel del vocabulario y, por otra, el tipo de vocabulario. Es decir, queremos comprobar primero el nivel de un vocablo en relación al *PCIC* y luego tipificar el vocabulario de la Dra. Fuensanta López para distinguir un vocabulario cotidiano-básico, general y académico o específico en relación con el glosario del *PCIC*, del Dr. Kincaid y del Wiktionary, FreeLing 1.5 y 2.0 y esWordnet.

Efectivamente, la concepción de cada uno de los glosarios es diferente. Sin embargo, queremos probar si se complementan. Si el *PCIC* está orientado a la enseñanza del español para extranjeros, el de la Dra. Fuensanta López presenta un vocabulario básico académico para la enseñanza de españoles, el del Dr. Kincaid es un glosario básico de control y el Wiktionary es un glosario a caballo entre el lenguaje hablado y escrito, queremos probar estadísticamente, en primer lugar, la similitud de niveles y, en segundo lugar, la existencia de diferentes tipos de vocabulario en todos ellos. Por ejemplo, al elegir un vocablo como

"rey" comprobamos que este lema existe en el "Índice" del *PCIC* con un nivel A1 y C2 dentro del concepto NE (Noción Específica) y en el de la Dra. Fuensanta López en un nivel 4 distribuido en 6 de las 13 materias. Nos preguntamos qué nivel tiene realmente dicho vocablo. En el procesamiento automático de ese vocablo dentro del *PCIC*, Lexicator, el nivelador de los vocablos, le concedería un nivel B1-B2. Según el criterio de combinación de pertenencia o no a un glosario, "rey" es un vocablo básico puesto que está indexado en el listado básico del Dr. Kincaid y en todos los demás. Por tanto, se corresponde con un vocablo de tipo básico o general.

Para ello, primeramente, realizamos una comparación basándonos en los distintos niveles establecidos por la Dra. Fuensanta López y los del *PCIC* como se muestra en la figura 3.1. Al hacer esta comparativa entre los niveles que propone la Dra. Fuensanta López en función de la usabilidad, la dispersión y la frecuencia de lemas con los lemas de los niveles que establece el *PCIC*, observamos que no existe una correlación clara entre la clasificación de niveles propuesta por la Dra. Fuensanta López y el *PCIC*. Más bien, los índices de error (marcados con las barras en color azul en la figura 3.1) son muy grandes y los niveles de la Dra. Fuensanta López parecen tener todos una relevancia semejante. Efectivamente, en la concepción del glosario de la Dra. Fuensanta López la proporción de lemas en todo los niveles es similar. Lógicamente, existen lemas menos frecuentes en los niveles más altos (ver la gráfica de la derecha de la figura 3.1).

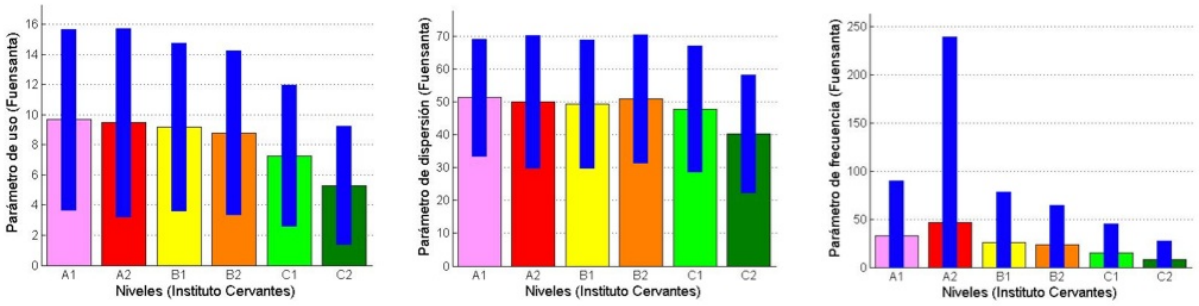


Figura 3.1: Comparativa del glosario de la Dra. Fuensanta López con el "Índice" del *PCIC*. De izquierda a derecha se representa el uso, la dispersión y la frecuencia, distribuidos por niveles. Cada color muestra un nivel: rosa para A1, rojo para A2, amarillo para B1, naranja para B2, verde claro para C1 y verde oscuro para C2. Las barras azules son los errores de cada uno de los parámetros.

Comprobada la poca funcionalidad del glosario de la Dra. Fuensanta López para calificar el vocabulario, lo vamos a descartar. Tampoco lo incluiremos al proponer el método de la combinación de diccionarios para la calificación del léxico cuando el *PCIC* no califica ciertos lemas. Este método se describirá en el apartado 4.1.4. No obstante, el glosario de la Dra. López lo tendremos en cuenta para el análisis semántico y el estudio de áreas temáticas de los textos, según se describe en el apartado 6.1.4.

3.1.4. Glosario de FreeLing 1.5 y Glosario de FreeLing 2.1

El diccionario castellano de lemas de FreeLing ha sido elaborado por el equipo del Dr. Padró de la Universidad Politécnica de Barcelona (Padró, 2006). Este diccionario es fun-

damental para identificar un alto porcentaje de vocabulario que, junto la automatización del reconocimiento de palabras por sus afijos, permite al analizador FreeLing etiquetar vocablos con un 99 % de aciertos. Por ello, el diccionario de lemas desambiguados con sus correspondientes vocablos, bien flexivos y conjugables o bien fijos, convierte a FreeLing en un analizador sintáctico que procesa, etiqueta los vocablos y analiza sintácticamente textos (Atserias *et al.*, 2006).

El diccionario morfológico de FreeLing ha sido creado a partir de los lemas más frecuentes en castellano. Aunque en un principio empezamos a trabajar con el diccionario morfológico de la versión 1.5 de FreeLing, posteriormente, como se han ido publicando nuevas versiones de FreeLing, todos los textos de nuevo se han analizado con el diccionario de la versión FreeLing 2.1, diccionario mucho más extenso que el de la anterior versión 1.5. Según nuestros cálculos, si el diccionario morfológico de la versión 1.5 listaba 7.156 lemas con sus respectivos vocablos flexivos o fijos, la versión 2.1-β indexa 76.214 lemas con sus 556.213 vocablos correspondientes.

El diccionario de la versión 1.5 se diferencia de la versión 2.1 exclusivamente en el mayor número de vocablos listados. Sin embargo, el poder disponer de los dos diccionarios es muy útil para este trabajo. Esta diferencia de extensión nos va a permitir distinguir también niveles de lengua. Esto es, si la versión 1.5 contiene vocablos propios de B2-C1, la versión 2.1 beta contiene más vocablos y, por consiguiente, incluye vocablos propios de un nivel C2.

La utilidad de ambos glosarios de FreeLing para la realización de este trabajo es clave, no sólo porque el más extenso optimiza el análisis y etiquetado morfológico y puede identificar más vocablos de los textos objeto de estudio sino porque ambos glosarios nos sirven como referencia para el análisis léxico. Por ejemplo, su lemario gramatical y desambiguado nos ha permito validar lemas de algunos de los otros glosarios utilizados como el del Dr. Kincaid (2.022 lemas), el de la Dra. Fuensanta López (5.273 lemas) o el del Wiktionary (5.207 lemas), incluso comprobar los niveles de los vocablos del "Índice" del *PCIC* (8.662 lemas) o el esWordnet (93.425 lemas) como se representa en la figura 4.3 del siguiente capítulo.

El formato de ambos diccionarios es un modelo de archivo indexado tipo BerkeleyDB (Padró, 2009, 15 y ss.). Es decir, en un renglón o línea se escribe el vocablo, el lema y su etiqueta y, así sucesivamente, se repiten tantos lemas y etiquetas de categoría como precise un determinado vocablo. Sirvan de ejemplo algunos vocablos con sus lemas flexivos o no, desambiguados gramaticalmente, que reproducimos en la tabla 3.3.

Vocablo	**Lema**	**PoS**	**Lema**	**PoS**	**Lema**	**PoS**
abajo	abajo	I	abajo	RG		
contribuya	contribuir	VMSP1S0	contribuir	VMSP3S0		
controles	control	NCMP000	controlar	VMSP2S0		
caracoles	caracol	NCMP000	caracoles	I		
cual	cual	DD0CS0	cual	PI0CS000	cual	PR0CS000

Tabla 3.3: Ejemplo de vocablos del glosario de FreeLing.

3.1.5. Glosario de esWordNet

El glosario esWordnet 1.6 (esWN) es la versión del *Wordnet 1.6* español-catalán que utilizamos en nuestra investigación. La versión tanto en castellano como catalán ha sido desarrollada conjuntamente por el grupo de investigadores CLIC de la Universidad de Barcelona, por el Centro de Investigación TALP de la Politécnica de Barcelona y por el Grupo de Procesamiento Natural de la Universidad Nacional de Educación a Distancia. La versión que se utiliza en este trabajo se corresponde con la versión castellana de noviembre de 2006.

Al igual que todas las versiones que se han ido desarrollando de la American Wordnet (Miller, 1990) o EuroWordnet (Association, 1999), esWordnet 1.6 se estructura de forma similar a todas ellas aunque con carencias respecto a las versiones inglesas y otras versiones desarrolladas a nivel comercial (Daudé *et al.*, 2001). En cuanto a su tamaño, la esWordnet 1.6 tiene 93.425 lemas en castellano. El léxico de esWN recoge tres categorías gramaticales: nombres, adjetivos y verbos. Los nombres y verbos están estructurados en jerarquías, mientras que los adjetivos y adverbios en *clusters* aunque, a diferencia de las versiones inglesas, la esWN 1.6 no tiene listados adverbios.

Por una parte, el glosario de esWordnet 1.6 nos va a ser útil como otro diccionario más para identificar los lemas de un texto. Por otra parte, vamos a establecer relaciones paradigmáticas, sintagmáticas o jerarquizadas a distintos niveles utilizando esWordnet como una taxonomía de relaciones ideológicas o semánticas, ya que su verdadera y originaria concepción es la de funcionar como una ontología. Esto es, cada vocablo, identificado con un código numérico, se relaciona con otros vocablos dentro del glosario y, a su vez, con otros Wordnets multilingües.

3.1.5.1. Configuración del esWordnet

En cuanto a su configuración, la base de datos de esWordnet, al igual que la de Wordnet, consta de tres listados interrelacionados entre sí: esWordnet *variant*, esWordnet *relation* y esWordnet *synset*, y se identifica cada vocablo en función de unos atributos concretos.

- En primer lugar, desde un punto de vista léxico, vamos a utilizar el esWordnet *variant*. Consta de 93.425 vocablos y nos son útiles para identificar vocablos en un texto por su lema, tanto sustantivos como verbos o adjetivos. En cuanto a los atributos correspondientes con sus vocablos, se distinguen: 1º, categoría gramatical; 2º, *synset* o número de identificación de una palabra; 3º, vocablo; 4º, número de sentido; 5º, índice de confianza; y 6º, un campo nulo (-) en la versión castellana de esWordnet, como se observa en la tabla 3.4.

a|00003057|naciente|3|99|-
n|00003095|célula|1|99|-
v|00003430|emanar|2|99|-

Tabla 3.4: Organización del fichero esWordnet *variant*.

En cuanto a los atributos, estos se expresan en inglés con las siglas: <POS>| <SYN>| <WORD>| <SENSE>| <CS>| <->, que describimos en la tabla 3.5.

Etiquetas de cada campo	**Descripción**
<POS>is the part-of-speech	la categoría gramatical o *part-of-speech: n (nombre), a (adjetivo), v (verbo)*
<SYN>is the ILI synset id in WN1.6	el número de identificación del *Inter-Lingual-Index* de un vocablo en WN1.6.
<WORD>is the Spanish synonym word	el vocablo
<SENSE>is the corresponding sense number	el número correspondiente al sentido o *sense* de la palabra
<CS>is the confidence score	el índice de confianza o *confidence score*
<>always "-" or null in the current version	campo siempre "-" nulo en la versión actual del castellano esWordnet 1.6

Tabla 3.5: Identificación de las etiquetas de esWordnet *variant*.

- En segundo lugar está el esWordnet *synset* que consta de 105.516 vocablos relacionados con sus hipónimos e hiperónimos. Este listado consta de seis campos: 1º, la categoría gramatical; 2º, el número de identificación del vocablo; 3º, el número de identificación de relación de hiponimia; 4º, un campo nulo; 5º, la verificación humana del término ("i") o la no existencia del término en castellano ("n", "–", "in") ; y 6º, la glosa o definición de la palabra en castellano.

  En la tabla 3.6, se muestran unos ejemplos de cómo se configura este archivo de esWordnet *synset*.

```
n|09185137|09185137|0|i-|
n|09185279|09185279|0|-n| Período de mayor producción y prosperidad
n|09185628|09185628|0|in| Tiempo en que transcurre algo de principio a fin
```

Tabla 3.6: Organización del fichero esWordnet *syn*set.

Los campos en los que se organiza la información se corresponden en inglés con las siguientes etiquetas: <SPOS>| <SSYN>| <NDES>| <STAT>| <SDEF>, que describimos en la tabla 3.7.

- Finalmente, en el esWordnet *relation* cada palabra o multi-palabra está registrada con un número de identificación o *synsets* que permite establecer relaciones semánticas de meronimia, sinonimia, antonimia e hiponimia con otros *synsets* o vocablos numerados. Este glosario de relaciones distingue seis campos: 1º, relación de hiponimia/hiperonimia descrita en EuroWordnet; 2º, categoría gramatical del término fuente; 3º, número de identificación del término fuente; 4º, categoría gramatical del

Etiquetas de cada campo	**Descripción**
<SPOS>is the Spanish synset part-of-speech	Categoría gramatical en castellano
<SSYN>is the ILI synset id in WN1.6	Número de identificación del *Inter-Lingual-Index* del vocablo en esWordnet 1.6
<NDES>is the number of hyponym synsets	Número de identificación de un hipónimo de dicho vocablo
<STAT>is the status of the synset.	Estado del synset: "i-" significa que el synset está verificado. "-n" significa que no hay vocablo para ese concepto. "in" significa que se ha verificado que no hay concepto para cierto vocablo.
<SDEF>is the synset gloss/definition in Spanish	La glosa o explicación del significado del vocablo

Tabla 3.7: Identificación de las etiquetas de esWordnet *synset*.

pertains_to|a|00051548|n|09729055|90
xpos_near_synonym|n|50004272|a|50000785|99
xpos_near_synonym|n|10442122|v|00246253|99

Tabla 3.8: Organización del fichero esWordnet *relation*.

término de destino; 5º, es el número de identificación del término de destino; y 6º, el índice de confianza.

Un ejemplo de la configuración del glosario del esWordnet *relation* se muestra en la tabla 3.8.

En inglés, las etiquetas correspondientes a cada campo son: <EWNR>| <SPOS>| <SSYN>| <TPOS>| <TSYN>| <CS>, que se especifican en la tabla 3.9.

En la tablas 3.10 y 3.11 se describen las distintas relaciones semánticas que se establecen entre los vocablos en esta base de datos léxica esWordnet. En realidad, los tipos de relaciones descritas en un estudio de EuroWordnet son más numerosas (Díez Orzas, 1999), y muchas de ellas no aparecen en esWordnet 1.6.

Etiquetas de cada campo	Descripción
<EWNR>is one of the EuroWordNet relations	Identifica el tipo de relación que tiene un vocablo con otro
<SPOS>is the source part-of-speech	La categoría gramatical del vocablo fuente
<SSYN>is the source ILI synset id in WN1.6	Número de identificación del *Inter-Lingual-Index* del vocablo fuente en el *esWordnet 1.6*
<TPOS>is the target part-of-speech	La categoría gramatical del vocablo diana
<TSYN>is the Target ILI synset id in WN1.6	Número de identificación del *Inter-Lingual-Index* del vocablo diana en el *esWordnet 1.6*
<CS>is the confidence score for the relation	Índice de confianza de la relación

Tabla 3.9: Identificación de las etiquetas en esWordnet *relation*.

Relaciones en esWordnet	**Definición**
Relaciones sintagmáticas o coordinadas	
near_antonym	Casi-antónimo
near_synonym	Casi-sinónimo

Tabla 3.10: Relaciones semánticas sintagmáticas del esWordnet.

Relaciones en esWordnet	**Definición**
Relaciones paradigmáticas o subordinadas	
role_ agent	Relación entre nombre de cosa o persona y verbos relacionados semánticamente
role_ instrument	Relación entre nombre de objeto y verbos relacionados semánticamente
role_ location	Relación entre nombre de lugar y verbos relacionados semánticamente
xpos_ fuzzynym	Existe cierta relación semántica entre distintas categorías gramaticales
xpos_ near_ synonym	Tiene un casi-sinónimo con otra categoría gramatical o *PoS* diferente a la suya

Relaciones en esWordnet	**Definición**
pertains_to	Se corresponde con un adjetivo relacionado semánticamente con un nombre en su *item* o grupo nominal
has_derived	Es un adjetivo relacionado semánticamente con un nombre en su raíz
has_subevent	Es un verbo relacionado con otro con el que se especifica la acción: e.g. "respirar-inhalar"
has_hyponym	Tiene un hipónimo
has_holo_member	Tiene un hiperónimo tipo miembro-colección
has_mero_member	Tiene un merónimo tipo miembro-colección
has_holo_part	Tiene un hiperónimo tipo componente-objeto
has_mero_part	Tiene un merónimo tipo componente-objeto
has_xpos_hyponym	Tiene un hipónimo con otra categoría gramatical o *PoS* diferente a la suya
has_holo_madeof	Tiene un hiperónimo que relaciona substancia-objeto
be_in_state	Nombre relacionado con un adjetivo
see_also_wn1.5	Relaciona adjetivos entre sí y verbos entre sí con una relación o contexto semántico similar

Tabla 3.11: Relaciones semánticas paradigmáticas del esWordnet.

3.1.5.2. Debilidades de esWordnet

Según terminología de Daudé, está por determinar el índice de error entre las taxonomías, redes semánticas o jerarquías de esta base de conocimiento (Daudé, 2005). Estas taxonomías están organizadas en "estructuras anidadas o frames que se relacionan entre sí" (Díez Orzas, 1999). Fundamentalmente esta base de datos relacional se concibió como una herramienta para la traducción. Son muchas las mejoras que se podrían aportar a esta base léxica ya que las taxonomías actuales impiden acceder a *clusters* coherentes. No obstante, las posibilidades que ofrece esta base de datos son útiles también para su explotación léxica y semántica fundamentalmente. Las taxonomías, tal y como existen en estos momentos, nos pueden informar sobre cierto grado de coherencia de un texto.

Por un lado, observamos imprecisiones en la identificación de los términos traducidos del inglés al castellano. Por ejemplo, la palabra *fungus* del inglés se ha traducido como "hongos" al castellano. Esto implica que nuestro programa nunca procesaría "hongos" porque es el plural del lema "hongo". En consecuencia, no se establecería la relación entre "hongo" y "seta", por ejemplo. Precisamente este ejemplo concreto se ha corregido aunque intuimos que existen deficiencias debidas a falsos amigos, a errores ortográficos e, incluso, a "realidades culturales" distintas, como apuntan otros autores en su contribución a la Wordnet 3.0 (Fernández-Montraveta *et al.*, 2008).

Se ha comprobado que muchas relaciones lógicas e inmediatas que establece nuestra experiencia, intuición o sentido común no están establecidas entre los vocablos en un primer grado de parentesco. Por ejemplo, "olor" - "olfato" es una relación que no se halla en esWordnet. Sin embargo, sí está la relación de "olor" - "perfume". Otro ejemplo similar a la falta de relación entre campos semánticos, ciertas jerarquías o *clusters* es la inexistencia de una relación de meronimia, hiponimia o hiperonimia como para "huerto" - "cebolla". Sin embargo, sí está relacionado "jardín" - "hortaliza" - "cebolla". Este último ejemplo muestra la traducción literal generada entre "garden" y "jardín". No obstante, en la base de datos de esWordnet, el lema "huerto" está y es un sinónimo de "jardín" pero no se relaciona con los vocablos anteriores, de manera que la relación entre los dos lemas "hortaliza" - "cebolla" permanece en ese primer nivel sin expandirse ni poder relacionarse con "huerto".

Se deduce que, para obtener el máximo rendimiento de esta base de conocimiento, hay que establecer una relación de más de un nivel para hallar una mayor correspondencia entre vocablos, que no siempre está establecida en primer grado. Es más, no todas las relaciones establecidas en la tabla 3.11 se dan en el glosario castellano.

A pesar de las debilidades de esta base de conocimiento, esWordnet es una herramienta muy útil en toda nuestra investigación.

3.1.6. "Índice de nociones generales y nociones específicas" del *PCIC*.

El "Índice de nociones generales y nociones específicas" del *Plan Curricular del Instituto Cervantes* es un glosario de 6.745 vocablos clasificados en tres apartados según se

correspondan con:

- el uso de la palabra dentro de un entorno de Noción General (NG) o Noción Específica (NE).
- el nivel de aprendizaje (A1, A2, B1, B2, C1, C2).
- alguno de los veinte campos semánticos diferenciados en los seis niveles (ver sección A.5 en el apéndice).

El "Índice de nociones generales y nociones específicas" del *Plan Curricular del Instituto Cervantes* o "Índice" del *PCIC,* como lo denominamos de forma abreviada, está publicado en formato papel en el tomo tercero del *Plan Curricular del Instituto Cervantes. Niveles de Referencia para el español C1-C2* (Cervantes, 2006, 719-764). Este glosario se ha convertido a formato electrónico para poder procesarlo. Además, en distintas etapas del trabajo, lo hemos implementado añadiendo la categoría gramatical y lematizando las entradas. También hemos separado aquellas entradas del glosario que tenían más de un nivel de lengua pasando de las 6.745 entradas originales no repetidas a un desdoblamiento de 8.662 lemas repetidos o diferenciados. Estos cambios favorecen el procesamiento de los datos y su tratamiento para posteriores análisis de modo que ahora, al desambiguar un vocablo, éste se diferencia no sólo por el nivel sino también por la categoría gramatical. Por tanto, generamos cinco campos por entrada: tipo de noción, nivel de lengua, campo semántico, lema y *PoS* o categoría gramatical.

3.1.6.1. Lematización y transformación del "Índice" del *PCIC*

Consideramos que el "Índice" del *PCIC* es una herramienta muy específica y valiosa en esta investigación. No obstante, aunque no se contempla en el trabajo de esta tesis desambiguar perfectamente la morfología del "Índice de nociones generales y nociones específicas" del *Plan Curricular del Instituto Cervantes*, se ha tratado de optimizar el rendimiento de la forma más eficaz realizando algunos cambios. Por ello, hemos lematizado el "Índice" desambiguando parte de los vocablos con la etiquetación de categorías gramaticales y adaptándolo a nuestro trabajo.

El listado de cada vocablo en el "Índice" del *PCIC* se presenta como sigue, por ejemplo, para el vocablo "diario":

- diario: NG, B2, 4.3.12.; NE, B1, 3.3; NE, B2, 18.4

Sin embargo, como esta forma de registro es muy profusa para procesar los datos tal cual, se ha procedido a desdoblar cada entrada por niveles y por categorías, como se muestra en la tabla 3.12.

Para el análisis del texto es útil distinguir en un mismo vocablo distintos apartados:

- a nivel semántico:
  - la noción general (NG) o específica (NE)

Vocablo	Noción	Nivel	Campo Semántico	Lema	*PoS*
diario	NG	B1	4.3.12	diario	AQ0MS0
diario	NE	B1	3.3	diario	NCMS000
diario	NE	B2	18.4	diario	NCMS000

Tabla 3.12: Organización de los vocablos en el "Índice" del *PCIC*.

- el campo semántico en el que se enmarca un lema se indica como 1.1: Existencia, Inexistencia; 2.1: Cantidad numérica; 3.1: Localización; etc. Concretamente para el ejemplo "diario", el código 4.3.12 corresponde a "Frecuencia"; el 3.3 incluye "Objetos personales", y el 18.4 se refiere a "Literatura".

- a nivel léxico:
  - el nivel de aprendizaje se marca con los niveles A1, A2, B1, B2, C1, C2
  - el lema, se desambigua con la categoría gramatical o *PoS* al distinguir "diario" como adjetivo calificativo masculino singular (AQ0MS0) y como nombre común masculino singular (NCMS000).

Cada apartado contribuye a desambiguar el significado y el nivel del vocablo. Desde el punto de vista del nivel y significado, se han desdoblado aquellas palabras identificadas con más de un nivel de lengua o campo semántico para poder procesarlas de forma independiente y precisa, obteniendo el máximo rendimiento e información del "Índice" del *PCIC*.

Efectivamente, existe una imprecisión con los niveles si no se ajusta el vocablo desdoblado a su correspondiente significado o contexto, y esto requiere un tratamiento minucioso y manual. Por ello, inmediatamente nos preguntamos ¿cómo resolvemos qué palabra procesar cuando un vocablo tiene, según su contexto o campo semántico al que pertenece, varios lemas con un mismo significado pero con un nivel diferente? En el "Índice" en formato papel cada vocablo registra tres campos: noción general o específica, nivel y campo semántico. En el de formato digital, el sistema automático puede procesar los datos de dos formas diferentes: una, se dará sólo un nivel a un vocablo con varios niveles mediante el cálculo de la media de todos los niveles que le da el "Índice" del *PCIC* y se le otorga esa media; otra, se le otorgaría un nivel a un vocablo de un texto cuando coincidiera el lema y la categoría gramatical del "Índice" con el lema y la categoría gramatical del texto que se analiza.

Como indicamos al inicio de este apartado, el "Índice" del *PCIC* es un listado muy valioso por varias razones: primero, porque, conforme a unos criterios establecidos de noción y contexto, un número importante de vocablos tiene un nivel de lengua concreto; segundo, porque es un listado-patrón de referencia de nivel que nos va a permitir dar un nivel a los otros glosarios; y tercero, porque, a pesar de que el "Índice" del *PCIC* nivela entre un 40-50 % de los textos que analizamos, al combinar entre sí todos los glosarios (Kincaid, Wiktionary, FreeLing 1.5, FreeLing 2.1, esWordnet y Locuciones) y comparar los resultados con los vocablos nivelados del *PCIC*, vamos a poder dar nivel a un total de 75-80 % del vocabulario de un texto.

No obstante, entre las futuras implementaciones del "Índice" del *PCIC* para mejorar el procesado, destacamos dos: especificar los lemas, etiquetándolos con su *PoS* correspondiente, y otorgar un único nivel por cada lema. Por un lado, como no se registran las categorías gramaticales o *PoS* de los lemas en el "Índice" del *PCIC*, los lemas no están desambiguados. Por ejemplo, "mejor" puede ser adverbio o adjetivo. Cuando se procesa un texto con FreeLing, se diferencia en "mejor" la categoría gramatical. Sin embargo, al tener sólo una entrada como adverbio y un nivel en el "Índice", "mejor" no se nivelaría en el texto cuando su *PoS* fuese un adjetivo. Por ello, no se ha nivelado atendiendo a ambos criterios del léxico, lema y *PoS*, porque si no reduciríamos el número de términos nivelados. Por otro lado, el siguiente inconveniente es la existencia en el "Índice" del *PCIC* de un mismo lema con varios niveles. Para este trabajo, todo lema con varios niveles se ha desdoblado. Por tanto, se ha obtenido dos o más lemas repetidos y cada uno, por su contexto o función, conserva su nivel. ¿Qué nivel tiene un lema con varios niveles, entonces? Se ha resuelto hallando la media de los niveles de esos mismos lemas con niveles diferentes. Por ejemplo, la preposición "en" tiene dos niveles A1 y B1, por tanto se le da un nivel A2. En definitiva, consideramos que tanto la falta de desambiguación léxica como la doble o triple nivelación de un mismo vocablo son debilidades del "Índice" del *PCIC* que podrían resolverse para aplicaciones y trabajos futuros.

3.1.6.2. Fases de adaptación del "Índice de nociones generales y nociones específicas" del *PCIC* a glosario electrónico procesable

La adaptación del "Índice" del *PCIC* en formato papel al electrónico se ha ido realizando en distintas fases. La fase previa fue reproducir todo el índice, y su posterior revisión y corrección gráfica. En una segunda fase, se lematizaron los vocablos y se adaptaron los lemas conforme al criterio de lema del diccionario de FreeLing. En una tercera fase, después del proceso de indexación de multi-vocablos que lista tanto el "Inventario de nociones generales" como el "Inventario de nociones específicas" del *PCIC*, se ha detectado que el "Índice de nociones generales y nociones específicas" del *PCIC* no registra algunos términos que aparecen en sus inventarios como locución o como vocablo independiente. Por ejemplo, "peligro" (Cervantes, 2006, B1-B2: 491), "auge" (Cervantes, 2006, C1-C2: 515), "parco" o "pródigo" (402) no se listan en el listado que denominamos como "Índice" del *PCIC*.

Finalmente, cuando hemos sometido el "Índice" del *PCIC* a tratamiento electrónico para probar su rendimiento y para procesarlo con los vocablos de todos los diccionarios mencionados más arriba, se han tenido que reajustar más de mil vocablos. Por razones gráficas, morfológicas o léxicas, quedaban sin computar estos mil vocablos del "Índice" del *PCIC*.

3.1.6.3. Justificación de los cambios, inclusiones, adaptaciones y registro de algunos vocablos

Se han hecho algunas adaptaciones al listado del "Índice" del *PCIC* desde el punto de vista del significante, siguiendo los criterios del analizador de FreeLing para poder procesar el máximo de palabras. En primer lugar, se han cambiado algunos vocablos a su

forma básica o lema. Esto es, se han lematizado aquellos sustantivos que aparecen en el glosario en plural. Es decir, sustantivos como "ahorros", "gambas" o "llaves" se cambian a su forma básica "ahorro", "gamba" o "llave". También se ha eliminado el pronombre de aquellos verbos listados con la partícula pronominal "se" en verbos como "cargar(se)", "divorciar(se)", etc. Se procesarán como "cargar" o "divorciar". Efectivamente, el hecho de que un verbo sea pronominal, reflexivo o no, según su contexto o nivel, cambia el significado de dicho verbo. Por ejemplo, "cargar" como noción general de un nivel B2 significa "llevar peso" y "cargarse (algo)" como noción general de un nivel C2 significa "estropear". Esta distinción semántica, de momento, no podemos diferenciarla con las herramientas que contamos a nivel léxico, pero si se detectaría en el procesamiento de Análisis Semántico Latente.

Por otro lado, aquellos vocablos dobles indexados en el "Índice" del *PCIC* como "todavía no" o "cubrir gastos" los hemos convertido en multi-vocablos, procesando "todavía_no" como un adverbio y "cubrir_gastos" (515) como un multi-vocablo verbal procesable en el glosario de locuciones. Sin duda, el programa de FreeLing 2.1 podría procesar realizaciones distintas de locuciones verbales similares que pudieran darse como "cubrir ningún gasto", "cubrir los gastos" o "cubrir todos los gastos", pero siempre que se listases previamente.

A continuación, señalamos el número [entre corchetes] de vocablos y casos más significativos que tuvimos que reajustar o dejar como aparecen en el formato original, y exponemos las razones por las que los adaptamos o permanecen como los lista el "Índice de nociones generales y nociones específicas" del *PCIC*.

- El "Índice" del *PCIC* registra muchos adverbios [127] acabados en –mente. Este tipo de vocablo derivado no siempre se suele registrar en los diccionarios. Entre algunos de los adverbios listados, el "Índice" del *PC*IC indexa incluso "(usual)mente" o "todavía (no)". Estos son vocablos con una grafía imposible de procesar si no se eliminan los paréntesis. Concretamente, para los adverbios que se listan acabados en –mente, se ha programado identificarlos, eliminar el sufijo –mente, procesarlos de nuevo y nivelarlos con su correspondiente adjetivo homónimo.

- Adjetivos acabados en –able o –ible [2] como "ensalzable" (de "ensalzar") e "intragable" (de "tragar"), bien transformados en adverbios terminados en –mente [2], como "increíblemente" (de "creíble" o "creer") e "incuestionablemente" (de "cuestionable" o "cuestionar"), o bien derivados de verbos o adjetivos con prefijos no se incluyen en la comparación ni nivelación automática. Es decir, no se ha programado que se procesen para relacionarlos con el vocablo raíz o con su adjetivo homónimo en los otros glosarios sino que, al ser un número muy reducido, se ha decidido listarlos manualmente en el glosario de multi-vocablos con el nivel correspondiente que se otorga en el "Índice" del *PCIC*.

- Los nombres propios [13] como "América", "África", "Asia", "Europa", "Historia" [2], "Internet" [4], "Oceanía", "Sabbat" y "Satanás" no los registran los diccionarios de FreeLing aunque sí Wiktionary y esWordnet.

- Las siglas [4]: RENFE, IBERIA, DVD, UVI, unas se listan en el *PCIC* mientras que otras como ADN y TAC están en Wiktionary y esWordnet.

- Los multi-vocablos [6] "análisis_clínicos" (nivel C1), "anhídrido_carbónico" (nivel C2), "frenos_ABS" (nivel C2), "lenguaje_HTML" (nivel C2), "mueble_bar" (nivel C2), "franja_horaria" (C2) no se procesan con ningún glosario, por ello los hemos incluido en el glosario de locuciones manualmente con su nivel correspondiente.

- En cuanto a los prefijos "pos-" [2], "post-" [2], "pre-" y "re-", los sufijos "-ote" y "–ón", y los vocablos con afijos [15] como "preadolescencia", "abuelote", "involucionar", "peliculón", "rellenito", "repatear" y "rollazo", aunque se identifique el lema, la categoría gramatical y su nivel, no se computan tal cual, con sus afijos, al compararlos con los demás diccionarios.

- Cierto número de nombres y algún verbo [33] quedan también fuera en el proceso de comparación referencial para la identificación de nivel por ser:
  - Vocablos cotidianos o familiares [4]: "amuermar", "curre", "encrispado" y "tapear".
  - Extranjerismos adaptados [9]: "apartotel", "bungaló", "campin" (vs. "camping" en FreeLing 2.1), "chat", "chatear", "chef", "estresar", "jacuzzi", "márquetin" (vs. "marketing" en FreeLing 2.1).
  - Neologismos [3]: "deshumidificar", "emoticón" y "cibercafé", que todavía no están registrados en el diccionario de la RAE).
  - Vocablos relacionados con el folklore español [3]: "aurresku", "bailaor" y "muiñeira" (vs. "muñeira" en FreeLing 2.1).
  - Específicos [4]: "cognitivismo", "entradilla", "patronaje" y "succionar".
  - Vocablos compuestos [8]: "autolavado" , "arcoíris", "hidromasaje", "hipervínculo", "pasicorto", "teleadicto", "telebasura" y "treintañero".

Comprobamos que ninguno de los otros diccionarios con los que trabajamos registra los vocablos de arriba por las características propias de cada palabra, bien por pertenecer a contextos muy concretos o no ser muy frecuentes. Es más, otra razón por la que se diferencian los vocablos es la grafía específica adaptada a la lengua gallega como "muiñeira" (nivel C2) o a la castellana como "campin" (nivel A1) y "márquetin" (nivel B1). Aunque estos vocablos los registra FreeLing 2.1, aparecen con la grafía castellana, como "muñeira", o inglesa, como "camping" y "marketing". Sin embargo, el hecho de que algunos vocablos estén registrados por FreeLing 2.1 nos ha dado licencia para incluir en su lemario sólo aquellos vocablos que tenían grafías distintas por estar adaptadas a las normas del castellano.

También hemos adaptado aquellos vocablos que se indexan en el "Índice" del *PCIC* como *singularia tantum* y en los otros glosarios aparecen indexados en *pluralia tantum* [3] como "vacación" vs. "vacaciones" (registrada por FreeLing 2.1 en singular y plural, pero sólo en plural por FreeLing 1.5); "pasapuré" (registrada por el *PCIC* y por el Diccionario de la RAE en la 23ª edición junto con "pasapurés") vs. "pasapurés" (registrada por FreeLing 2.1 en plural); "finanza" vs. "finanzas" que registra FreeLing 1.5 en singular y plural. Sin embargo, FreeLing 2.1 la registra en plural sólo. Plural que no existe como entrada en el diccionario de la RAE. Así que tanto los vocablos como los lemas se han adaptado y ampliado en los dos diccionarios de FreeLing en función de las entradas propuestas por el "Índice" del *PCIC* con el objetivo de nivelar el máximo número de vocablos.

Al igual que en el caso anterior, se ha procedido de forma similar con los vocablos

de diferente grafía [2]: "postdata" vs. "posdata" (registrada esta última por FreeLing 2.1 y esWordnet, aunque ambas entradas figuran en el *Diccionario electrónico de la RAE*); "sándwich" (vs. "sandwich", registrada por FreeLing 2.1 sin tilde y en el Diccionario electrónico de la RAE con tilde); y "fríjol" vs. "frijol". Los tres se han modificado y ampliado en los dos diccionarios de FreeLing ya que FreeLing 1.5 y FreeLing 2.1 los registraban, pero con diferente grafía.

En los casos en que el "Índice" del *PCIC* lista la forma flexionada de un vocablo [7] con un género o número diferente al que es propio de su lema, se ha modificado en el "Índice" electrónico implementándolo con un lemario paralelo al listado en el "Índice" del *PCIC* en formato papel y siguiendo los criterios de indexación de FreeLing. Por ejemplo, el lema de un adjetivo estará en masculino y singular; y el lema de un nombre, en singular. Se ignora el vocablo y se opta por el lema. Entre los vocablos lematizados destacamos "carácter" vs. "caracteres", "arrendador" vs. "arrendadora", "esquí" vs. "esquíes", "fauce" vs. "fauces", "monje" vs. "monja", "solariego" vs. "solariega", "vetusto" vs. "vetusta".

Otro ajuste realizado con un módulo específico ha permitido nivelar unos centenares de participios que se registran en el "Índice" del *PCIC* pero que no se procesaban. En un principio, estos participios quedaban sin identificar porque los comparábamos con los lemas verbales de los otros diccionarios. Ahora, estos participios han quedado identificados con el lema verbal por haberlos identificado a través de las formas flexivas del verbo, como es el participio. Este ajuste ha permitido procesar más vocablos del "Índice" del *PCIC* y, en consecuencia, nivelar más lemas de los otros diccionarios.

Una vez realizados todos estos ajustes y creado un "Índice" del *PCIC* electrónico y procesable, aún quedan sin procesar 94 vocablos de los listados en el "Índice de nociones generales y nociones específicas" del *PCIC*, como describimos más arriba y se observa en la figura 4.4 del capítulo siguiente, por ser palabras específicas, llevar afijos o tener su propia idiosincrasia y no tener referente en otro glosario. Una vez más, el afán por adaptar el máximo número de vocablos tiene como objetivo computar y calificar el mayor número de vocablos de un texto.

3.1.7. Glosario de Locuciones

Otro glosario específico es el diccionario de locuciones. Este diccionario está conformado por 5.688 multi-vocablos en su totalidad. Aunque es un lemario del que se extraen aquellos multi-vocablos que únicamente están indexados, permite identificar una gran variedad de lexías flexionables tanto nominales, adverbiales y conjuntivas como, en un menor número, adjetivales, verbales, determinantes y pronominales. Este glosario está integrado en el programa analizador de FreeLing. Es fácilmente ampliable y su buen funcionamiento depende del diccionario de FreeLing 2.1 ya que, en función de este gran diccionario de lemas, se identifican individualmente aquellos vocablos que conforman las lexías listadas. Queremos decir que el mayor o menor éxito de la identificación de las lexías depende de que, a su vez, los vocablos flexivos de las lexías se hallen lematizados en el diccionario de FreeLing 2.1.

Este glosario de locuciones lo utilizamos concretamente para identificar lexías o multivocablos. La otra aplicación importante del glosario de locuciones es que permite listar

muchas lexías con los niveles establecidos por el *PCIC* y nivelar aquellas que no se les ha marcado con un nivel previamente. Es decir, se ha programado otorgarles un nivel automáticamente a aquellas lexías que no tienen un nivel otorgado por el *PCIC*, basándonos en los criterios que detallamos en el apartado 4.1.6.

Dentro de nuestro estudio del léxico, hemos implementado el listado de lexías aumentando el número de entradas en un tamaño superior al que configura el archivo de "locuciones.dat" del analizador sintáctico de FreeLing. Así que, de forma sistemática, hemos indexado nuevas locuciones siguiendo la estructura establecida por el Dr. Padró de tal manera que se pueda computar toda la información añadida a cada locución, incluso, a nivel sintáctico. Consecuentemente, el listado de locuciones se ha convertido en un glosario porque, además de añadir información sintáctico-semántica a la categoría gramatical o *PoS* (*Part of Speech*), también se ha otorgado un nivel de lengua a todos los multi-vocablos. Además, al considerar que el uso de locuciones en una segunda lengua puede ser un buen indicador del nivel de escritura, nos hemos empeñado en trabajar exhaustivamente con el glosario de locuciones para incrementar su listado y asignar un nivel de referencia a cada una de las locuciones indexadas. Habitualmente, la enseñanza de locuciones se empieza a introducir tímidamente en el nivel B1 (Peramos Soler *et al.*, 2010), pero no empiezan a sistematizarse hasta un nivel B2-C1, según se observa en el *PCIC*. Esto implicaría su aparición en la escritura en los niveles más altos.

La gramática tradicional denomina locuciones nominales, adjetivas, verbales, adverbiales, pronominales, determinativas, preposicionales y conjuntivas a un conjunto de palabras desgramaticalizadas que, al combinarse, han cambiado de significado y dejado de tener su categoría gramatical original. Sólo entonces, este conjunto se clasifica dentro de la categoría gramatical de locución y, de nuevo, adquiere otra vez una de estas categorías, dentro de la categoría de nombre, adjetivo, verbo, adverbio, pronombre, determinante o conjunción. Las unidades léxicas indexadas se han obtenido de diversas fuentes como de los niveles B1-B1 y C1-C2 del *Plan Curricular del Instituto Cervantes*, de la edición digital íntegra del archivo de Salvador Fernández Ramírez o Archivo Gramatical de la Lengua Española (AGLE) ubicada en el Centro Virtual Cervantes, del *Diccionario práctico de gramática* de Óscar Cerrolaza Gili (Cerrolaza-Gili, 2005) y de *Redes. Diccionario combinatorio del español contemporáneo* de Ignacio Bosque (Bosque, 2004).

En nuestro estudio, denominamos de forma general multi-vocablos a todo el conjunto de palabras que son propiamente locuciones, lexías, colocaciones o incluso estructuras partitivas. También se listan algunos verbos preposicionales como multi-vocablos. No obstante, todos ellos se procesan etiquetando su categoría gramatical mediante un *PoS* y, en algunos casos, se destaca la especificidad semántico-funcional de algunas categorías. Es más, cada multi-vocablo se glosa con un nivel, bien en función de la longitud del multi-vocablo en sí, bien por ser un multi-vocablo flexivo que cambia en relación a su lema o bien porque tiene un nivel ya establecido por el *PCIC*. Esto es, utilizando la metodología y los resultados que se obtienen al procesar un texto con el programa FreeLing del Dr. Padró, la versatilidad del programa y la posibilidad de ampliar el glosario de "locuciones" nos ha permitido no sólo implementar el glosario del Dr. Padró con más de 2000 nuevos multi-vocablos sino añadir información semántica a algunas etiquetas o *PoS* y fijar un nivel de aprendizaje a aquellos multi-vocablos que ha determinado el *PCIC* previamente.

En cuanto a la metodología de etiquetación de los multi-vocablos para ser procesados y de identificación de cada multi-vocablo con un PoS o categoría del multivocablo, nos basamos en el protocolo y método de etiquetación establecido por el Dr. Padró en FreeLing. Por ejemplo, cuando listamos colocaciones como las lexías nominales, éstas pueden ser bien flexivas o bien fijas. En el caso de una lexía flexiva, aquellos vocablos susceptibles de ser flexivos aparecen entre los signos convencionales de mayor y menor (<>). Concretamente, dentro de una locución nominal, cuando el nombre común (NC) aparece colocado en primera posición y se flexiona, se marca dicha posición con el símbolo convencional "1$" ($1:NC). Igualmente, cuando el nombre ocupa una segunda posición, dicha posición se marca con el símbolo "2$" ($2:NC).

En cambio, si una lexía es fija, se lista el multi-vocablo marcando su categoría de nombre (N), adjetivo (A), adverbio (R), determinante (D), pronombre (P), numeral (Z) o interjección (I) seguido de las iniciales que identifican sus características gramaticales (común o propio, género y número) o semánticas (nombre de persona, lugar, objeto, etc.), pero sin posibilidad de ser flexionado ni el género ni el número. Por ejemplo:

Locución nominal fija:

"equilibrio ecológico": su *PoS*, NCMS000, identifica que es Nombre, Común, Masculino, Singular, con clasificación semántica nula (-00-) y grado nulo (-0).

Locución adjetival fija (según el Diccionario electrónico de la R.A.E.):

"de armas tomar": su *PoS*, A0ACN0, indica que es una locución de Adjetivo (A), no Calificativo (-0-), de grado Aumentativo (-A-), género neutro (-C-), número invariable (-N-) y sin función (-0).

Locución adverbial fija:

"una vez más": su *PoS*, RG_TP, indica que es Adverbio General (RG), Temporal (TP).

Locución determinativa fija:

"una tal": su *PoS*, DI3FS0, marca que es Determinante, Indefinido, 3ª persona, Femenino, Singular, no Poseedor.

Locución pronominal fija:

"lo mío": su *PoS*, PX1NS0S0, indica que es Pronombre, Posesivo, 1ª Persona, Neutro, Singular, Caso -0-, Poseedor Singular, no de Cortesía.

Locución numeral fija:

"un cuarto de": su *PoS*, Zp_Partitivo, identifica que es Numeral fraccionario, Partitivo.

Mientras que el objetivo del Dr. Padró con las locuciones es analizar sintácticamente un texto con la máxima precisión, el nuestro es, además del análisis morfo-sintáctico, también procesar la información de cada vocablo para otorgarle un nivel de lengua al texto objeto de estudio. En esta etapa de calificación de los multi-vocablos, se considera la categoría gramatical de los multi-vocablos, la longitud, la información semántica del *PoS* o especificidad y el nivel de referencia previamente fijado ya en ciertos multi-vocablos por el *PCIC* o, *a posteriori*, por los criterios anteriores o por analogía con estructuras similares.

3.1.7.1. Lexías nominales

En cuanto a las locuciones nominales, la etiquetación gramatical puede variar dependiendo de que el nombre sea o no flexivo. Como ejemplo, sirvan las siguientes lexías nominales:

Lexía nominal fija: "número_de_identificación_fiscal" NCMS000
Lexía nominal flexiva: "<estación>_de_tren" $1:NC

Además de la etiquetación morfológica, etiquetar con cierta información semántica es interesante porque dotamos al multi-vocablo de unos datos semánticos procesables que pueden aportarnos más informaciónn sobre los textos que vamos a analizar. En el diccionario de FreeLing 2.1 elaborado por el Dr. Padró hallamos algunos nombres con información complementaria en los penúltimos ceros (-00-) o en el cero final (-0) del *PoS* de algunos sustantivos tales como:

NP000G0: con la "-G-" se expresa que es un Nombre Propio de Lugar; e.g. "Barcelona".
NP000O0: con la "-O-" se expresa que es un Nombre Propio de Organización; e.g. "COI".
NP000P0: con la "-P-" se expresa que es un Nombre Propio de Persona; e.g. "Pedro".
NCMS00D: con la "-D" se expresa que es un Nombre Común Masculino Singular Diminutivo; e.g. "gatito".
NCMS00A: con la "-A" se expresa que es un Nombre Común Masculino Singular Aumentativo; e.g. "casona".
NCMN000: con "-MN-" se expresa que es un Nombre Común Masculino Invariable; e.g. "cortapapeles".
NCFN000: con "-FN-" se expresa que es un Nombre Común Femenino Invariable; e.g. "tesis".

También, en los ceros (-00-) del *PoS* de algunos multi-vocablos nominales, hemos decidido insertar alguna información semántica más. Es decir, con la inserción de unas iniciales, hemos etiquetado algunos nombres con la noción de lugar (-LG-), de tiempo presente

(-TP-), de futuro (-TF-) y de pasado (-TS-). Incluso etiquetamos algunos términos relacionados con la mecánica gramatical (-MC-) o con características gramático-semánticas explícitas. El objetivo es poder procesarlas para futuras investigaciones. Se muestran a continuación algunos multi-vocablos morfológicamente etiquetados que ejemplifican estos tipos de locuciones. Locuciones a las que se les ha añadido información semántica complementaria e independiente del análisis morfo-sintáctico, de manera que el *PoS* contiene información semántica en los dos penúltimos ceros (-00-) de un nombre (NCMS000): Por ejemplo:

"día_de_acción_de_gracias" NCMSTP0
"esa_misma_ocasión" NCFSTS0
"año_que_viene" NCMSTF0
"pasado_año" NCMSTS0
"punto_final" NCMSMC0

Incluso podemos etiquetar información más específica bajo un criterio sintáctico-semántico:

"mitad_de" NCFS000_Partitivo
"mitades_de" NCFP000_Partitivo (Cervantes, 2006, B1-B2: 404).

3.1.7.2. Lexías adjetivales

Respecto a las locuciones adjetivales, éstas se etiquetan de manera similar a las nominales. Dentro de los multi-vocablos con función adjetiva, diferenciamos varios tipos: la locución adjetiva flexiva con preposición ($1:AQ) o la locución fija (AQ0CN0) con valor predicativo/atributivo (AQ_P/A), partitivo (AQ0CS0_Partitivo), aumentativo (AQACN0) y superlativo absoluto (AQSCN0_Absoluto) o relativo (AQSCN0_Relativo). Sirvan de ejemplo:

AQ_P/A es Adjetivo calificativo, con función de Predicado o Atributo; e.g. "por las nubes".
$1:AQ es Adjetivo calificativo con <término flexivo>en 1ª posición; e.g. "<ligero>de cascos".
AQ0CN0 es Adjetivo calificativo, sin grado (-0-), género Común o neutro, Número invariable, sin función (-0); e.g. "de_confianza".
AQACN0 es Adjetivo calificativo, grado Aumentativo, género Común o neutro, número Invariable, sin función (-0); e.g. "a precio de oro".
AQ0CN0_Partitivo es Adjetivo, partitivo, grado -0-, género Común o neutro, Número invariable con función de Partitivo; e.g. "doble de" (Cervantes, 2006, B1-B2: 404).
AQSCN0_ Absoluto es Adjetivo, Calificativo, de grado Superlativo, con género Común o neutro, Número invariable, con función de Absoluto; e.g. "la flor y nata".

3.1.7.3. Lexías verbales

En relación a las locuciones o lexías verbales, procesamos locuciones verbales propiamente dichas, verbos preposicionales, expresiones verbales y algunas perífrasis. Estos multi-vocablos se indexan igualmente con símbolos convencionales adoptados por el Dr. Padró en su programa de FreeLing. Respecto a las locuciones verbales, la estructura formular $1:V se utiliza para marcar la primera posición de la forma flexiva o conjugable de un verbo. Cuando el verbo es predicativo, se representa con una "M" ($1:VM), con una "S" ($1:VS) para representar el verbo atributivo "ser", y con una "A" ($1:VA) para marcar el verbo auxiliar "haber". Unos ejemplos de cómo se listan las lexías verbales se muestran a continuación:

"<estar>_de_vuelta_de_todo" $1:VM
"<dejar>_de_lado" $1:VM
"<dejar>_de" $1:VM
"<ser>_<capaz>" $1:VS

Efectivamente, estas locuciones verbales son fácilmente desmembrables. En cuanto el autor de un texto sea creativo con una locución verbal, insertando otro elemento no indexado, no se procesará dicha locución. Por ejemplo, al implementar la locución con un adverbio como en "estar ya de vuelta de todo", el programa automático no va a procesar la expresión verbal. Para procesar esta posibilidad se precisa la implementación de un algoritmo para el módulo de Sintactor, pero lo postergamos y no lo aplicamos en este trabajo.

Sin embargo, sí que se puede indexar y procesar algunas lexías verbales con los pronombres enclíticos en los infinitivos, gerundios y los imperativos. Sin embargo, se procesarán sólo aquellos verbos que previamente hemos indexado. Por ejemplo, en la locución verbal "dar cuerda (a alguien)", se procesa la locución completa si lleva el pronombre personal proclítico al verbo, es decir, por un lado, la locución verbal "<dar>_cuerda" se procesaría en "le da cuerda" o "les dio cuerda" con el pronombre proclítico y la forma verbal; por otro lado, también se procesa el complemento de persona enclítico o insertado en la siguiente locución verbal si lo hemos configurado previamente como "<dar>_<él>_cuerda". De esta manera, FreeLing reconoce las formas enclíticas "dándole cuerda" (gerundio), "darse cuerda" (infinitivo) o "dame cuerda" (imperativo).

3.1.7.4. Lexías adverbiales

Dentro de las lexías adverbiales incluimos aquellas con función de adverbios e, incluso, aquellas que están a caballo entre conjunción o adverbio o entre adverbio de modo o (adjetivo) predicativo. Así, algunos multi-vocablos que funcionan como conectores unas veces tienen un *PoS* de adverbio (RG) y otras un *PoS* de conjunción coordinante (CC) o subordinante (CS), sin distinguir si son marcadores discursivos u oracionales. El criterio de indexación y clasificación que hemos seguido para marcar los multi-vocablos ha sido, en primer lugar, mantener los multi-vocablos listados ya por el Dr. Padró en su glosario de locuciones. Además, por otra parte, se han listado varios multi-vocablos conforme a

nuestro conocimiento, siguiendo la nomenclatura y clasificación semántica de(Martín Zorraquino, 1998, 62-63), (Cerrolaza-Gili, 2005),(Fernández-Ramírez, 2009), el Vademécum de Fundéu, y recurriendo además al diccionario de María Moliner y el diccionario electrónico de la Real Academia Española para desambiguar o ampliar información. Por otra parte, se han indexado, con sus niveles, muchas locuciones que se recogen en los dos volúmenes del *PCIC* en los capítulos de "Nociones generales" y "Nociones específicas" de los niveles B1-B2 y C1-C2.

Incluimos también. dentro de las locuciones propiamente adverbiales, frases cliché o *chunks* y las representamos a todas con el mismo *PoS* de adverbio (RG). Además de la nomenclatura para marcar la categoría gramatical de adverbio, nosotros implementamos la información de la lexía añadiendo un guion bajo seguido de la clase a la que pertenece la lexía o función (RG_). No obstante, la mayoría de locuciones latinas originales registradas por el Dr. Padró han quedado marcadas simplemente con RG. Un ejemplo es *"ex abrupto"*, etiquetada sólo como RG. A continuación, se presenta una muestra de cómo marcamos funcionalmente los *PoS* de las nuevas locuciones adverbiales que hemos registrado en el glosario de locuciones:

RG_AD: locución adverbial de adición o de culminación (Martín Zorraquino, 1998, 63), (Moliner, 2007); e.g. "por añadidura".
RG_AF : locución adverbial de afirmación o refuerzo argumentativo; e.g. "por descontado".
RG_DV: locución adverbial adversativa para expresar distanciamiento o contra-argumentos; e.g. "por el contrario".
RG_CT: locución adverbial de cantidad; e.g. "en demasía".
RG_CL: locución adverbial de causa o justificación; e.g. "por esa razón".
RG_CD: locución adverbial de condición-prevención (Martín Zorraquino, 1998, 63), (Moliner, 2007); e.g. "por si acaso".
RG_CC: locución adverbial concesiva; e.g. "a pesar de los pesares".
RG_CN: locución adverbial conclusiva-recapitulativa; e.g. "para finalizar".
RG_CS: locución adverbial consecutiva; e.g. "por eso".
RG_DG: locución adverbial de digresión; e.g. "a propósito".
RG_DD: locución adverbial de duda; e.g. "tal vez".
RG_ES: locución adverbial especificativa, aclarativa o de particularización (Martín Zorraquino, 1998, 62), (Moliner, 2007); e.g. "ante todo".
RG_XP: locución adverbial explicativa, aclarativa o de particularización; e.g. "en concreto".
RG_IL: locución adverbial ilativa; e.g. "pues bien".
RG_ID: locución adverbial indefinida; e.g. "menos de".
RG_XC: locución adverbial de exclusión positiva; e.g. "a cambio".
RG_XP: locución adverbial explicativa-rectificativa-ilativa-ejemplificativa; e.g. "a decir verdad".
RG_FN: locución adverbial final; e.g. "para que".
RG_GN: locución adverbial de generalización; e.g. "por norma general".
RG_HP: locución adverbial de hipótesis-preventiva (Martín Zorraquino, 1998, 63), (Moliner, 2007); e.g. "sea cual fuere".
RG_NT: locución adverbial de introducción-tema (Bosque y Demonte, 1999);

e.g. “de otra parte”.
RG_IN: locución adverbial interrogativa; e.g. “qué tripa se le ha roto”.
RG_LG: locución adverbial de lugar; e.g. “de la ceca a la meca”.
RG_MD: locución adverbial de modo; e.g. “de golpe”.
RG_MV: adverbial de movimiento-dirección; e.g. “de derecha a izquierda”.
RG_RF: locución adverbial de referencia, focalización o de operador discursivo; e.g. “al respecto”.
RG_SQ: locución adverbial secuencial; e.g. “en última instancia”.
RG_TP: locución adverbial temporal; e.g. “de hoy en adelante”.
RG_TC: locución adverbial temporal-causal; e.g. “tras haber estado”.
RG_OP: locución adverbial de opinión; e.g. “a mi juicio”.
RG_UB: locución adverbial de ubicación en el tiempo o en el espacio; e.g. “al final”.
RG_Registro-Epistolar; e.g. “muy señor mío”.
RG_Particularizador; e.g. “especialmente” (Cervantes, 2006, B1-B2: 291, 300, 404).
RG_Focalizador/TP: enfatiza y refuerza; e.g. “justamente ahora” (Cervantes, 2006, B1-B2: 300) o “justo delante” (Cerrolaza-Gili, 2005, 173).
RG_Catafórico; e.g. “así de”.
RG_Cortesía; e.g. “por favor”.
RG_Coloquial; e.g. “así así”.
RG_Intensificador_cuantitativo (Cervantes, 2006, B1-B2: 300); e.g. “absolutamente”.
RG_Incluyente (Cervantes, 2006, B1-B2: 300); e.g. “incluso aún”.

Por otro lado, también se listan locuciones adverbiales propiamente negativas. Éstas locuciones se marcan con RN: la “R” por ser adverbio y la “N” por ser negación. Es más, se etiquetan, si se considera significativo, con la misma información que se añade a las locuciones adverbiales positivas o afirmativas (RG). Por ejemplo:

RN_MD: locución adverbial negativa de modo; e.g. “en modo alguno”.
RN_CT: locución adverbial negativa de cantidad; e.g. “ni fu ni fa”.
RN_XC: locución adverbial de exclusión negativa; e.g. “ni aún”.
RN_Valoración-negativa (Cerrolaza-Gili, 2005, 176); e.g. “ligeramente”.

Se han listado algunas locuciones con formato adverbial (RG_), aunque gramatical o semánticamente sean adjetivas, por poder equivaler bien a un adjetivo o bien a un adverbio. Por ejemplo:

AQ_P/C: locución Adjetiva Predicativa o Copulativa; e.g. “<hecho>y <derecho>”.
AQACN0: locución Adjetiva Aumentativa, Común, Neutra; e.g. “a precio de oro” o “fuera de lo común”.

Por ello, muchas locuciones que tienen el formato RG_MD son propiamente adverbiales pero otras, en realidad, son adjetivales ¿Por qué se etiquetan de forma diferente a su categoría? Porque como locución adverbial la estructura es susceptible de ser muy

flexiva y, por tanto, más versátil. En cambio, el registro de estas locuciones dentro de una categoría adjetiva es más restrictiva al comportarse, en muchas ocasiones, como una locución fija. Otra de las razones por las que no se indexan estas locuciones con un verbo determinado es por su vulnerabilidad a desmembrarse como se explica a continuación.

Por ejemplo, la locución "de_broma" es adverbial (RG). Desempeña función de predicativo con un verbo predicativo pero atributiva con verbo copulativo. Por lo expuesto anteriormente, se opta por indexarla como una locución adverbial (RG). Otro ejemplo es la lexía "en_blanco_y_negro". Según el contexto gramatical, funcionaría como un complemento predicativo o atributivo. Sin embargo, en el ejemplo "la película está en_blanco_y_negro", la lexía funcionaría como atributo.

Además, de momento, se lista la lexía verbal de forma independiente, sin acompañarla de un verbo para facilitar que, si el aprendiz inserta otra categoría dentro de la lexía verbal indexada, al menos se procese la locución adverbial y se registre su nivel. Esta forma de indexar nos otorga más flexibilidad y nos abre la posibilidad de procesar en un futuro este tipo de estructuras. Esto es, se propone computar estas estructuras adverbiales cuando aparezcan con verbos que llamamos polisemi-comodines. Algunos ejemplos de lexías verbales creadas *ad hoc* con lexías adverbiales como "en_juego" y con este tipo de verbos (poner, estar, quedar, andar, etc.) son: "poner en_juego", "poner siempre en_juego", "estar en_juego" o "estar de_nuevo en_juego".

En definitiva, por razones procesables y cambiantes debido a la mecánica de la gramática, es más productivo el tratamiento automático de lexías indexadas con la categoría de adverbio, de momento, para los fines de esta investigación. Aunque, muchas de estas "locuciones adverbiales de modo" (RG_MD) deberían ser revisadas y rescatadas del "cajón de sastre" en el que están ubicadas (Bosque, 1989, 26). Como proponen algunos autores (de Miguel, 2009), para las locuciones se debería determinar el co-texto donde adquieren una función u otra. Sin duda, esta peculiaridad podría ser otro rasgo distintivo de nivel en futuras aplicaciones.

3.1.7.5. Lexías preposicionales

Las locuciones prepositivas (SPS_) se marcan igualmente seguidas de una nomenclatura de clasificación semejante a la del adverbio, según su función semántico-gramatical. A este tipo de locuciones Fernández Ramírez lo identifica como "adverbio con complemento prepositivo" (Fernández-Ramírez, 2009). Un ejemplo de locuciones preposicionales indexadas son:

SPS_AD: locución preposicional de adición; e.g. "amén de".
SPS_AF : locución preposicional de afirmación o refuerzo argumentativo; e.g. "no hay duda de".
SPS_DV: locución preposicional para expresar contra-argumentos; e.g. "por oposición a".
SPS_CT: locución preposicional de cantidad; e.g. "a base de".
SPS_CL: locución preposicional de causa o justificación; e.g. "con el motivo de".

SPS_CD: locución preposicional de condición; e.g. "con la condición de".
SPS_CC: locución preposicional concesiva; e.g. "a pesar de".
SPS_CS: locución preposicional consecutiva; e.g. "por gentileza de".
SPS_DG: locución preposicional digresiva; e.g. "a propósito de".
SPS_DS: locución preposicional de deseo; e.g. "con ganas de".
SPS_XP: locución preposicional explicativa, e.g. "por lo que respecta a".
SPS_ID: locución preposicional indefinido; e.g. "de menos de".
SPS_XC: locución preposicional de exclusión positiva; e.g. "a excepción de".
SPS_FN: locución preposicional final; e.g. "a fin de".
SPS_HP: locución preposicional hipotética; e.g. "en el supuesto de".
SPS_NT: locución preposicional introductoria de tema; e.g. "en lo que respecta a".
SPS_LG: locución preposicional de lugar; e.g. "de la base de".
SPS_MD: locución preposicional de modo; e.g. "a golpe de".
SPS_MV: locución preposicional de movimiento-dirección; e.g. "hacia adelante de".
SPS_NG: locución preposicional de negación; e.g. "ni hablar de".
SPS_RF: locución preposicional de referencia; e.g. "con el de".
SPS_TP: locución preposicional temporal; e.g. "desde antes de".
SPS_TC: locución preposicional temporal-causal; e.g. "en vida de".
SPS_OP: locución preposicional de opinión; e.g. "según mi modo de".
SPS_PS: locución preposicional de posesión; e.g. "en poder de".
SPS_UB: locución preposicional de ubicación en tiempo o en espacio; e.g. "a caballo entre".
SPS_CM: locución preposicional de compañía; e.g. "con apoyo de".
SPS_CP: locución preposicional de comparación; e.g. "en comparación de".
SPS_LG/DEIC: locución preposicional locativa-deíctica; e.g. "al norte de".
SPS_Partitivo: locución preposicional de proporción; e.g. "en grupos de".

3.1.7.6. Lexías conjuntivas

A las locuciones conjuntivas se las ha etiquetado con un *PoS* diferente dependiendo de si son conjunciones coordinantes (CC_) o subordinantes (CS_), seguidas de una nomenclatura de clasificación similar a la que se utiliza para las locuciones adverbiales y preposicionales. Por ejemplo:

CC_DV: locución conjuntiva adversativa; e.g. "no obstante".
CC_DV/CR: locución conjuntiva adversativa/correlativa; e.g. "sino que".
CS_TP: locución conjuntiva temporal; e.g. "para cuando".

3.1.7.7. Lexías determinativas

Locuciones determinativas y pronominales son aquellas que funcionan en la frase igual que un determinante o un pronombre. Dentro de estas locuciones mostramos algunos ejemplos de locuciones indefinidas, posesivas y demostrativas, tanto fijas como flexivas:

Locuciones determinativas fijas:

DI00000: locución determinante indefinida; e.g. “bien de”.
DI00000: locución determinante indefinida; e.g. “no más de”.

Locuciones determinativas y pronominales flexivas:

$1:DD: locución determinante demostrativa; e.g. “<tal>o <cual>”.
$1:PI: locución pronominal indefinida; e.g. “<alguno>de”.
$1:PI: locución pronominal indefinida; e.g. “<mucho>de”.
$1:PP: locución pronominal posesiva; e.g. “<él>mismo”.

3.1.7.8. Lexías exclamativas

Otro tipo de lexía es la locución exclamativa (I). Efectivamente, un vocablo o multi-vocablo se puede listar con los signos exclamativos. Si no se precisan los signos porque en sí el multi-vocablo es exclamativo, lo marcamos con o sin los signos exclamativos, indicando si es una forma flexiva o fija:

I: locución flexiva de exclamación ; e.g. “pobre de <yo>” .
I: vocablo fijo exclamativo fijo; e.g. “¡vale!”.

3.1.7.9. Lexías niveladas

Además de toda la información gramático-semántica que acompaña a las locuciones, se ha implementado el glosario de locuciones con la información de un nivel de lengua. Éste oscila entre un B1, B2, C1 y C2 según el *Plan Currricular del Instituto Cervantes* (*PCIC*). El *PCIC* basa sus niveles, por un lado, en criterios funcionales (Cervantes, 2006, B1 B2: 183-275; C1-C2: 179-257), gramaticales (Cervantes, 2006, B1-B2: 43-111; C1-C2: 51-112) y pragmáticos (Cervantes, 2006, B1 B2: 279-307; C1-C2: 261-294); por otro, en criterios nocionales generales (Cervantes, 2006, B1 B2: 398-441; C1-C2: 393-456) y específicos (Cervantes, 2006, B1 B2: 445-511; C1-C2: 463-533). Aunque desarrollamos en el apartado 4.1.6 los criterios para evaluar por niveles las locuciones, exponemos a continuación unos ejemplos de locuciones ya niveladas en los inventarios del *PCIC*:

Colocación nominal: e.g. “año sabático” C1 (Cervantes, 2006, C1-C2: 486).
Locución adjetival: e.g. “de nueva planta” C1 (525).
Locución verbal: e.g. “pasar lista” B2 (Cervantes, 2006, B1-B2: 464).
Locución preposicional: e.g. “a principio de” B1 (415).
Locución adverbial: e.g. “por la espalda” B2 (410).

En resumen, con este diccionario de locuciones y su implementación, proponemos que el uso de multi-vocablos o lexías sea otro índice medible que nos informe sobre el nivel de un texto y de aprendizaje del idioma. Además, siguiendo el criterio que han aplicado

también los creadores de Coh-Metrix para medir la dificultad lectora que tiene un texto, el coeficiente o valor otorgado a cada lexía se basa en los dos criterios del Test de Flesch-Kincaid, esto es, cuanto más extensa es una palabra o una frase, mayor complejidad supone reproducirla correctamente. Por ello, consideramos que cuanto más extensa sea una lexía más alto es su valor y, en consecuencia, indica que el nivel de lengua del aprendiz es más alto.

Todas estas locuciones nivelas son procesadas, y a las no niveladas, mediante un módulo auxiliar, se les ha calculado el nivel en función de los criterios desarrollados en el apartado 4.1.6.

3.2. Herramientas de análisis lingüístico-computacional

Tanto la identificación de los índices de medida propuestos para esta investigación y computados por las herramientas como los resultados de los análisis de tales índices están en función de herramientas o sistemas auxiliares computacionales desarrollados por diferentes autores (FreeLing o Wordnet) y por nosotros (todos los módulos que componen Evaluator).

3.2.1. FreeLing 1.5: Analizador morfológico

El analizador morfo-sintáctico FreeLing, desarrollado en la Universidad Politécnica de Cataluña (Padró 2006), es una herramienta clave en nuestro estudio. Desde su creación (Carreras *et al.*, 2004), esta herramienta está siendo constantemente actualizada por su autor, el Dr. Lluís Padró y su equipo (Padró *et al.*, 2010a). En un principio este analizador constaba de un diccionario morfológico de 7.156 lemas. Actualmente el diccionario de FreeLing se ha aumentado a 76.214 lemas. La versión FreeLing 1.5 ya cubría un 90 % de las palabras de un texto de una temática general. El 10 % restante se resolvía con un analizador estadístico de sufijos que proporciona un 98 % de acierto a la hora de identificar las categorías gramaticales de un texto. En nuestro estudio, se ha utilizado como analizar morfológico el amplio diccionario de la versión FreeLing 2.1 con sus 76.214 lemas, proporcionando un acierto en el etiquetado de categorías *o* "*Part of Speech*" (*PoS*) de un 99,8 % de los vocablos y un 98 % en la identificación de lemas.

Para nuestro estudio, la parte que vamos a utilizar de esta herramienta es la morfológica. Aunque es una pequeña aportación, se ha conseguido mejorar el proceso de análisis morfo-sintáctico de algunas categorías gramaticales o *PoS* al introducir cambios en el archivo de reglas gramaticales "constr-gram.dat" de FreeLing 1.5. No obstante, el Dr. Padró ha introducido otras mejoras en la versión 2.1 pero no son las utilizadas por nosotros para analizar los textos objeto de estudio. En consecuencia, se han utilizado las reglas gramaticales del archivo "constr-gram.dat" de FreeLing 1.5 tanto para los análisis morfo-sintácticos de los discursos de Navidad del Rey desde 1975, que sirve de prueba de las herramientas, como para los textos de los exámenes de DELE.

También, se ha utilizado en el análisis de los textos el diccionario del archivo "locucions.dat" de FreeLing 1.5 que consta de un lexicón complementario de locuciones conjun-

Concepto	Medida	Descripción
Título del texto	Título del texto	Título del texto
Género	Género	Epístola (personal o formal) Ensayo (descriptivo, narrativo, expositivo, argumentativo, ...)
Nivel	B1, B2, C1	Identificación del nivel de aprendizaje
Código del texto	Código	Identificación del texto

Tabla 3.13: Índices generales del texto.

tivas, adverbiales, preposicionales y verbales, además de lexías y colocaciones clasificadas por categorías gramaticales (Alonso *et al.*, 2002a). Asimismo, como se ha otorgado un nivel de aprendizaje a cada una de las locuciones, bien de forma manual siguiendo los criterios del *PCIC* o de forma automática, según los criterios de longitud del multivocablo propuestos por Kincaid y Coh-Metrix, se obtiene un vocabulario etiquetado con su *PoS* correspondiente y con información específica de cada uno de los multi-vocablos.

3.2.2. Coh-Metrix: Analizador de textos

Concepto	Medida	Descripción
Dimensión del texto	N_PALABRAS	Número de palabras en el texto
Número de frases	N_FRASES	Número de frases en el texto
Nº de párrafos	N_PARRAFOS	Número de párrafos (no se computa)
Nº de Palabras/Frase	N_PAL/FRASE	Número de palabras por frase
Nº de Frases/Párrafos	N_FRASE/PARR	Número de frases por párrafo
Número de Palabras Antes del Verbo Principal	NPAVP	Número de palabras antes del verbo en una frase
Numero de Palabras Después del Verbo Principal	NPDVP	Número de palabras después del verbo en una frase

Tabla 3.14: Índices numéricos del texto.

Como apuntábamos en el apartado 2.3.3, esta herramienta de análisis elaborada por la Universidad de Memphis (Graesser *et al.*, 2004) define una gran variedad de parámetros cuantificables para evaluar el grado de dificultad de lectura de un texto. Este sistema de análisis a nosotros nos ha servido de reflexión. Es más, muchos parámetros se podrían computar, pero apenas unos pocos índices se aplican en nuestro estudio, porque se precisa

desarrollar pequeños módulos no disponibles en español, aunque sí existen en inglés. Además, si seguimos la evolución de los estudios realizados con Coh-Metrix por el equipo de Memphis, los índice más significativos de la investigación en el estudio de textos de alto nivel son la complejidad sintáctica, la variedad de vocabulario y la frecuencia de palabras (Crossley *et al.*, 2010, 7).

Concepto	Medida	Descripción
Mayor Frecuencia de Palabras Contenido	MYFRPC	Las palabras contenido son nombres, adjetivos, verbos y locuciones nominales, adjetivales, adverbiales, conjuntivas y preposicionales
Menor Frecuencia de Palabras Contenido	MNFRPC	Indica qué categorías de palabras contenido son menos frecuentes
Mayor Peso de Palabras Contenido	MYPSPC	Indica qué categorías de palabras contenido son más frecuentes
Menor Peso de Palabras Contenido	MNPSPC	Indica qué categorías de palabras contenido predominan
Coseno de Palabras Contenido	CSPC	Correlación entre palabras contenido
Porcentaje de Palabras Diferentes	PPD	Indica cuántos lemas contenido hay diferentes
Porcentaje de Palabras Iguales	PPI	Indica cuántos lemas contenido hay iguales

Tabla 3.15: Índices estadísticos léxicos.

No obstante, nosotros recreamos muchos de sus criterios cuantificables, de carácter general en la tablas 3.13 y 3.14. Otros, más específicos, pueden medir la riqueza léxica (ver tabla 3.15), la cohesión morfológica (ver tabla 3.16), sintáctica (ver tabla 3.17) y semántica. También son destacables los índices de madurez sintáctica (ver tabla 3.18) y la complejidad (ver tabla 3.19), además de las relaciones léxico-semánticas (ver tabla 3.20), la coherencia (ver tabla 3.21) y los distintos modelos situacionales. Las tablas, en cuyas celdas se registran estos índices, se organizan en tres columnas para expresar el "Concepto" textual o gramatical; el código de "Medida" de frecuencias, pesos e índices; y la "Descripción" del concepto o de las herramientas, métodos, *PoS* o categorías gramaticales objeto de análisis en un texto.

Concepto	Medida	Descripción
Nº de Sintagmas Nominales	NSN	N* en 1ª, 2ª o 3ª posición del fichero de estructuras

Concepto	Medida	Descripción
Nº de Sintagmas Adverbiales	NSRG	RG*
Nº de Sintagmas Verbales	NSV	VM* , VA*, VS*
Nº de Sintagmas Preposicionales	NSPREP	SPS*
Nº de Modificadores por SN	NMSN	D*/D* D* + N*
Nº de Verbos + Preposición	NVPREP	VMI* /VMS*/VAI*/VAS*/VSI*/VSS* + SPS00 / SPS_*
Nº de Pronombres	NPRON	PP*, PO*, PD*, PX*, PI*, PT*, PR*, PN*, PE*
Nº de Determinantes	NDETER	DD*, DI*, DP*, DT*, DE*, DA*
Índice de tiempos verbales	ITV	VMIP*,VMII*,VMIF*,VMIS*,VMIC*, VMS*,VMSI*,VMSF*,VMIS*, VAIP*,VAII*,VAIF*,VAIS*,VAIC*, VASP*,VASI*,VASF*,VASS*, VSI*, VSII*,VSIF*,VSIS*,VSIC*, VSS*,VSSI*,VSSF*,VSSS*
Índice de modos verbales	IMV	VMI*, VMS*, VMM*, VAI*, VAS*, VAM*, VSI*, VSS*, VSM*
Índice de formas verbales compuestas	IFVC	VAI*+VMP*, VAS*+VMP*, VAI*+VSP*+VMP*, VAS*+VSP*+VMP*
Índice de formas verbales no personales	IFVI	VSN*, VSG*, VSP*, VMN*+VMN*, VMN*, VMG*, VMP*, VAN*+VMP, VAN*+VSP*+VMP*, VAG*+VMP*, VAG*+VSP*+VMP*,
Índice de multi-vocablos o Locuciones	MULTI_V	RN*, RG*, SPS_*, CC*, CS*, N*, D*,V* con guion _ en el lema
Longitud del multi-vocablo	LONG_MULTI_V	Longitud dependiendo del número de guiones bajos

Tabla 3.16: Índices morfo-sintácticos.

Concepto	Medida	Descripción
Índice de conjunciones coordinantes	INDMARCCOOR	CC, CC*
Índice de conjunciones subordinantes	INDMARCSUB	CS, CS_*

Concepto	**Medida**	**Descripción**
Marcadores lógicos Positivos	MARLOG_P	CC, CC_*, CS, CS* "y, e, o, entonces" para expresar razonamientos lógicos
Marcadores lógicos Negativos	MARLOG_N	RN, RN_* , CC, CC_* CS, CS* con adverbio negativo (RN) incluido en el marcador "no, ni" para expresar razonamientos lógicos
Marcadores Argumentativos Positivos	MARARG_P	*_XC/DV, *_ES/DG, *_DV exclusivos, adversativos y digresivos
Marcadores Argumentativos Negativos	MARARG_N	RN + *_XC, *_XC/DV/, *_ES/DG, *_DV
Marcadores Subjetivos Positivos	MARSUBJ_P	CS* + VSS*, VMS*, VAS*
Marcadores Subjetivos Negativos	MARSUBJ_N	RN + CS* + VSS*, VMS*, VAS*
Marcadores Referenciales Positivos	MARREF_P	*_RF*
Marcadores Referenciales Negativos	MARREF_N	RN + *_RF*
Marcadores Ejemplificadores	MAREJEM	*_XP
Marcadores Conclusivos Positivos	MARCONCL_P	SPS_CN, RG_CN, CS_CN
Marcadores Conclusivos Negativos	MARCONCL_N	RN + CS_CN, SPS_CN, RG_CN
Marcadores Aditivos-Introductorios Positivos	CONADIT_P	RG_AD, RG_AD/CP, RG_AD/NT, RG_IN/DEIC
Marcadores Aditivos –Introductorios Negativos	CONADIT_N	RN + RG_AD, RG_AD/CP, RG_AD/NT, RG_IN/DEIC
Marcadores Temporales	MARCTEMP_P	RG_TP, CS_TP, *TP*
Marcadores Temporales Negativos	MARCTEMP_N	RN + RG_TP, RN + CS_TP, RN + *TP*
Conectores Adversativos Positivos	CONADV_P	CS_DV

Concepto	**Medida**	**Descripción**
Conectores Adversativos Negativos	CONADV_N	RN + CS_DV
Conectores Causales Positivos	CONCAU_P	CS_CL, CS_CP/CL
Conectores Causales Negativos	CONCAU_N	RN + CS_CL, CS_CD/CL/RN
Conectores Completivos	CONCOMP	CS
Conectores Condicionales	CONCOND	CS_CD, CS_HP, CS_CS/CD, PR3MOS0/CD
Conectores Condicionales Negativos	CONCOND_N	RN + CS_CD, CS_CD/CL/RN, CS_CD/RN
Conectores Concesivos	CONCONC	CS_CC
Conectores Comparativos	CONCOMP	RG_CP
Conectores Correlativos Positivos	CONCORR_P	CS_CR
Conectores Correlativos Negativos	CONCORR_N	RN + CS_CR
Conectores Consecutivos Positivos	CONCONS_P	CS_CS
Conectores Consecutivos Negativos	CONCONS_N	RN + CS_CS
Conectores Finales Positivos	CONFIN_P	CS_FN
Conectores Finales Negativos	CONFIN_N	RN + CS_FN
Conectores Locativos	CONLOC	CS_LG, CS_UB

Tabla 3.17: Conectores.

Concepto	Medida	Descripción
Índice de Complejidad Sintáctica en estructuras	INDCOMSIN_S	Número y variedad de categorías en una estructura sintáctica
Índice de Complejidad Sintáctica en Frases	INDCOMSIN_F	Número de frases y conectores en una frase: CC*, CS*, SPS_* + CS, PR* + VMI*, VMS, VSI*. VSS*, VAI*, VAS*
Índice de Complejidad Sintáctica en Párrafos.	INDCOMSIN_P	Número de frases: Nº de Fp (.), Fc (,), Fia y Fit (¿?), Faa y Fat (¡!), Fd (:), Fpa y Fpt (( )), Fra y Frt (“ ”) -Nº conectores en un párrafo
Índice de Diversidad Sintáctica	IDIVSIN_T	Variedad de estructuras en todo el texto: entre frases y párrafos

Tabla 3.18: Madurez sintáctica.

Concepto	Medida	Descripción
Índice causal	IND_CL	Marcadores, conectores y locuciones adverbiales y preposicionales: CS_CL, RG_CL, SPS_CL CS_TC, RG_TC, SPS_TC
Índice de Cohesión Causal (textos expositivos y narrativos) + tiempos verbales	INDCOHCA	Ratio entre los marcadores y conectores causales CS_CL, RG_CL, SPS_CL, CS_TC, RG_TC, SPS_TC + acciones en todos los modos y tiempos: VMIP*, VMII*, VMIF*, VMIS*, VMIC*; VMSP*, VMSI*; VMP*, VMSF*, VMSC*, VMSS*, VSIP*, VSII*, VSIF*, VSIC*, VSIS*, VSSP*, VSSI*, VSSF*, VSSC*, VSSS*, VAIP*, VAII*, VAIF*, VAIC*, VAIS*, VASP*, VASI*, VASF*

Concepto	**Medida**	**Descripción**
Índice temporal	IND_TP	W (fechas, horas, días, meses) + acciones en todos los modos y tiempos: VMIP*, VMII*, VMIF*, VMIS*, VMIC*, VMSP*, VMSI*, VMP*, VMSF*, VMSC*, VMSS*, VSIP*, VSII*, VSIF*, VSIC*, VSIS* VSSP*, VSSI*, VSSF*, VSSC*, VSSS* VAIP*, VAII*, VAIF*, VAIC*, VAIS*, VASP*, VASI*, VASF*
Índice de Cohesión Temporal (textos narrativos) + tiempos verbales	INDCOHTP	Ratio entre los marcadores y conectores temporales CS_TP, RG_TP, SPS_TP CS_TC, RG_TC, SPS_TC + acciones en todos los tiempos y modos: VMIP*, VMII*, VMIF*, VMIC*, VMIS*, VMSP*, VMSI*, VMP*, VMSF*, VMSC*, VMSS* VSIP*, VSII*, VSIF*, VSIC*, VSIS* VSSP*, VSSI*, VSSF*, VSSC*, VSSS* VAIP*, VAII*, VAIF*, VAIC*, VAIS* VASP*, VASI*, VASF*
Índice Intencional	IND_INT	Contenido final y consecutivo: SPS_CS SPS_FN CS_CS CS_FN RG_CS RG_FN Modo subjuntivo y tiempo futuro: VMS*, VSS*, VAS*, VMIF*, VSIF*, VAIF* Lemas: "ir_a" o perífrasis: "habría/habrá que" VAIC*/VAIF* + CS Nombres de futuro NCMSSP0, NCMPSP0, NCFSSP0, NCFPSP0 Verbos de deseo o intención: "decidir, desear, intentar, permitir, querer"

Concepto	**Medida**	**Descripción**
Índice de Cohesión Intencional	INDCOHIN	Ratio entre los marcadores y conectores finales y verbos en subjuntivo *_FN* + VMS* VSS* VAS*, y entre los conectores consecutivos y verbos en indicativo *_CS* + VMI* VSI* VAI*
Índice Espacial	IND_ESP	Preposiciones de lugar: "en, entre, bajo, ante" Preposiciones de movimiento: "a, hacia, de, desde" Locuciones: *_LG* o *_UB* Verbos de movimiento: "abandonar, acercar, andar, bajar, caer, caminar, conducir, desplazar, empujar, entrar, ir, llegar, llevar, meter, mudar, poner, quitar, recorrer, regresar, sacar, salir, saltar, subir, traer, transladar, trasladar,venir, viajar, volver" Tiempo de perfecto: VA* y VMP*, VAI* y VSP*
Índice de Cohesión Espacial	INDCOHES	Media entre la ratio de lugar y de movimiento

Concepto	Medida	Descripción
Índice de Cohesión Dialógica (Voces)	INDCOHDI	- Personal: Verbos de dicción: "comentar, contar, decir, explicar, preguntar" en *1S0, *1P0, *2S0, *2P0, *3S0, *3P0 - Pronombres personales PP1CSN00, PP2CSN00, PP3MS000, PP3FS000, PP1MP000, PP1FP000, PP2MP000, PP2FP000, PP1CSN00, PP3MP000, PP3FP000 PP2CS00P, PP2CP00P Nombres propios (NP* y NC*) "sujeto" antes u "objeto-agente" después del verbo (SPS00 NP*/ SPS00 D* N*) - Impersonal: "se": P0000000 + V* Indefinido Pronombre (PI*) o Determinante (DI*) + V*

Tabla 3.19: Índices situacionales o circunstanciales.

Concepto	Medida	Descripción
Índice de Correferencia Nominal en la Frase	INDCORRF_NF	Referencia anafórica a nombres (N*) anteriores con pronombres (P*) coincidentes en persona y número o referencia catafórica a nombres posteriores dentro de la frase.
Índice de Correferencia Nominal en el Párrafo	INDCORRF_NP	Referencia anafórica a nombres (N*) anteriores con pronombres (P*) coincidentes en persona y número o referencia catafórica a nombres posteriores dentro del párrafo.

Concepto	Medida	Descripción
Índice de Correferencia Nominal en el Texto	INDCORRF_NT	Referencia anafórica a nombres (N*) anteriores con pronombres (P*) coincidentes en persona y número o referencia catafórica a nombres posteriores dentro del texto.
Índice de Correferencia Radical en la Frase	INDCORRF_RF	Recurrencia de palabras con la misma raíz (stem) o "lema" en la unidad-t
Índice de Correferencia Radical en el Párrafo	INDCORRF_RP	Recurrencia de palabras con la misma raíz (stem) o "lema" en el párrafo
Índice de Correferencia Radical en el Texto	INDCORRF_RT	Recurrencia de palabras con la misma raíz (stem) o "lema" en la unidad-texto
Grado de Similitud Léxico-Semántica en la unidad-t: Meronimia	GSLSU_F	Grado de similitud en relación a un repositorio de palabras similares en la unidad-t (Venegas, 2006); por su relación de meronimia en esWordnet de la Universidad Politécnica de Cataluña, similar a la base de conocimiento Eurowordnet, desarrollada por Thera-Clic de la Universidad de Cataluña
Grado de Similitud Léxico-Semántica en párrafo: Meronimia	GSLSU_F	Grado de similitud en relación a un repositorio de palabras similares en el párrafo aplicando el LSA (Venegas, 2006); por su relación de meronimia en esWordnet
Índice de Palabras Contenido con Significado Común en todo el texto: Meronimia	INDPCOCO_T	Grado de similitud en relación a un repositorio de palabras similares en la unidad-texto por su relación de meronimia

Tabla 3.20: Cohesión semántica formal.

Como escribíamos al principio de este apartado, casi todos estos índices son computables para el español actualmente, pero no vamos a computarlos porque obtendríamos una gran cantidad de datos numéricos que luego precisan de una interpretación y que no aportarían una nivelación. Sin embargo, todos estos índices nos han servido para reflexionar sobre aquellos elementos esperables en las estructuras que hemos creado en el fichero

Concepto	Medida	Descripción
LSA entre oraciones adyacentes	LSA_OAD	Semantor procesa la semántica entre oraciones dentro de la frase
LSA entre todas las frases	LSA_O	Semantor procesa la semántica entre oraciones dentro de la frase
LSA entre párrafos	LSA_P	Semantor procesa la semántica entre frases dentro del párrafo/texto

Tabla 3.21: Coherencia semántica.

de estructuras para el análisis sintáctico de los textos. Es más, todos estos índices son de gran interés para obtener datos cuantitativos sobre las características de distintos niveles, tipos y funciones de los textos.

3.2.3. Latent Semantic Analysis: Identificador semántico

Uno de los métodos de análisis semántico aplicados a este trabajo se basa en técnicas matemáticas de estadística multivariante automatizadas. Es decir, métodos de frecuencias e indexación basados en matrices (Konchady, 2006). El análisis semántico latente, conocido en inglés como *Latent Semantic Analysis* (LSA), es un método de estudio de los contenidos semánticos de un texto comparado con otros textos o, mejor aún, con un *corpus* de textos de una cierta materia o nivel. El objetivo fundamental de este método es extraer el contenido semántico de un texto mediante el análisis de la correlación estadística de las frecuencias de aparición de las palabras con respecto a las frecuencias de aparición en otros textos o *corpus*. Los niveles de comparación también pueden ser intra-textuales, al comparar el contenido semántico entre frases, párrafos o entre unidades de mayor o menor extensión dentro de un texto. Podemos asumir como cierto que, si varios textos tienen contenidos semánticos relacionados, éstos utilizarán un conjunto de palabras similares y sus frecuencias de aparición serán semejantes.

Desde un punto de vista básico, el LSA se basa en el método de análisis por componentes principales, también conocido en inglés como *Principal Component Analysis* (PCA). El método PCA ha sido utilizado con éxito en multitud de campos: ciencias experimentales (López-Alonso *et al.*, 2002; López-Alonso y Alda, 2004, 2002; López-Alonso *et al.*, 2004; Ferrero *et al.*, 2007), de la salud (Kintsch, 2002; Heinzle y Haynes, 2009), ciencias sociales (Dunteman, 1989; Jorge-Botana *et al.*, 2007) y humanidades (Sigley, 1997; Alda y Ferrero, 2007a; Olmos *et al.*, 2009). Esta diversidad en la aplicación parte del carácter puramente estadístico de sus fundamentos, lo cual lo convierte en un método tremendamente versátil. Con el fin de mejorar la comprensión de este método, vamos a tratar de explicarlo paso a paso.

Para comenzar a aplicar el método debemos obtener las frecuencias de aparición de los lemas que tienen importancia semántica en el texto. Por ello, tiene sentido que no se consideren aquellos lemas que aparecen en un sólo texto, ya que lo que se pretende con este estudio es conocer la conexión semántica entre textos. Este recuento de frecuencias de aparición se organiza en una matriz de números en filas y columnas. Las filas de esta matriz describen las frecuencias de aparición de los lemas contenidos en un texto. Las columnas contienen las frecuencias de aparición de un determinado lema en todos los textos analizados. En nuestro caso, obtenemos muchos más lemas que textos analizados por lo que la matriz tiene más columnas que filas.

En este punto nos encontramos con diversas posibilidades de "modular" los valores de las frecuencias, utilizando diversas transformaciones matemáticas. Landauer propone calcular el logaritmo de las frecuencias (Landauer *et al.*, 2007), lo cual parece interesante cuando se analiza un *corpus* de textos muy extenso donde los valores de las frecuencias son altos. En ese caso, la utilización de la función logarítmica reduce el rango de valores y permite compensar ciertos problemas. Otras transformaciones pueden aplicarse para normalizar los valores de frecuencias de un lema a lo largo de todo un texto o los valores de frecuencias de los lemas en un mismo texto (Ziempekis y Gallopoulos, 2006).

Una vez obtenida la matriz de frecuencias de aparición, o la matriz de frecuencias modificada según los criterios de normalización expresados en el párrafo anterior, es posible

aplicar el método PCA sobre esta matriz. Este método identifica las relaciones entre textos. Una de las métricas más habituales para realizar este procedimiento es la que se obtiene calculando la matriz de varianza-covarianza. Este cálculo consiste en comparar dos textos viendo lo correladas que están las frecuencias de aparición de los lemas en los dos textos. Para un conjunto de N textos se obtienen $N \times (N-1)/2$ valores de covarianza (relación entre dos textos) y N valores de varianza (relación de un texto consigo mismo). Todos estos valores se organizan en una matriz $N \times N$ que contiene los valores de varianza (comparación de un texto consigo mismo) en una diagonal y los de covarianza (comparación de dos textos distintos) fuera de ella. Si los textos no tuviesen nada que ver entre sí, podemos intuir que al compararlos por parejas su covarianza (su relación) sería nula. Sin embargo, esto no es lo que ocurre en el análisis de textos y la matriz de varianza-covarianza, porque siempre hay términos fuera de la diagonal, indicando que los textos están conectados.

Sin embargo, es interesante poder obtener un conjunto de textos no relacionados entre sí de los que se puede describir de forma fiable el contenido semántico de todos los textos originales. Esto es precisamente lo que se consigue con la aplicación del método PCA y el LSA. En nuestra nomenclatura denominaremos a estos textos como "textos sintéticos"; otros autores los denominan "pseudo-textos" (Jorge-Botana *et al.*, 2010).

Desde un punto de vista matemático el procedimiento consiste en diagonalizar la matriz de varianza-covarianza original. Debemos recordar que este proceso de diagonalización va a crear una nueva matriz con ceros fuera de la diagonal, es decir, los "textos sintéticos" obtenidos no tienen relación entre ellos. Realmente, podemos ver esta transformación como un cambio de base de nuestros datos. En origen, cada texto estaba descrito por las frecuencias de aparición de los lemas, y estos textos originales tenían relación mutua. Ahora, después de aplicar este procedimiento, obtenemos unos "textos sintéticos", caracterizados por unos valores de frecuencia distintos a los originales pero que describen, independientemente unos de otros, el mismo contenido semántico que el conjunto de textos originales.

Por tanto, otro resultado importante son los vectores o autovalores obtenidos de este procedimiento de cambio de base necesario para pasar desde los textos originales a los textos sintéticos. Este cambio de base se describe mediante un conjunto de N vectores de N componentes. Por ejemplo, el vector #1 nos dice cómo se reparte el primer texto sintético entre el resto de textos. A la vez, en este proceso de diagonalización nos aparece otro objeto matemático que denominamos como autovalor. Existirán tantos autovalores como textos originales, en nuestro caso N. Estos N autovalores describen la importancia que tienen los N textos sintéticos a la hora de explicar la distribución de frecuencias de aparición de los lemas. Además, debido al procedimiento matemático empleado, el autovalor #1 es el que mayor importancia tiene. El resto de autovalores se ordenan sucesivamente de mayor a menor importancia. De hecho, llega un momento en el que dos autovalores consecutivos no pueden distinguirse entre sí debido a la naturaleza estadística del procedimiento empleado. En ese caso, diremos que los autovalores están conectados estadísticamente y no podemos considerarlos por separado. Sólo con el LSA los primeros autovalores son independientes entre sí, estando el resto sucesivamente conectados entre ellos en lo que típicamente podríamos considerar como "ruido semántico". Por ello, en muchas aplicaciones del método PCA se consigue reducir la dimensionalidad del proble-

ma. En nuestro caso, el conjunto de autovalores empleados se reduce a dos, el segundo y el tercero (ver la figura 3.2 correspondiente a los discursos del Rey y la figura 7.13 del capítulo 7).

El LSA utilizado en esta memoria ha sido empleado para identificar semánticamente la pertenencia de un texto a un determinado grupo o conjunto de textos. Sin embargo, consideramos que todavía es posible profundizar mucho más en la comprensión semántica de los textos utilizando este método. Es posible, por ejemplo, identificar aquellos lemas cuyo reparto entre los textos no es comparable al reparto del resto de lemas. Estos lemas "anómalos" pueden servir para identificar peculiaridades de textos. También es posible "filtrar" un determinado conjunto de textos para eliminar el "ruido" producido por la aparición de lemas en "textos sintéticos" no independientes entre sí.

Con el fin de expresar mejor el procedimiento empleado, podemos utilizar los 36 discursos navideños del Rey como ejemplo de análisis. En esta ocasión tenemos $N = 36$ por lo que el número de textos sintéticos obtenibles será también de 36. Lo primero que se debe hacer es construir la matriz de frecuencias de aparición. En este caso hemos modificado esta matriz de frecuencia obteniendo el logaritmo de las frecuencias.

En la figura 3.2 hemos representado los pesos (la importancia) de cada texto sintético obtenido. Podemos ver que el primer texto sintético ya es capaz de explicar el 61.71 % de la variabilidad de los 36 textos. A la vez, sólo los textos sintéticos #1, #2 y #3 (círculos rojos) son independientes entre sí, estando el resto conectados en un conjunto de "ruido semántico" (circunferencias).

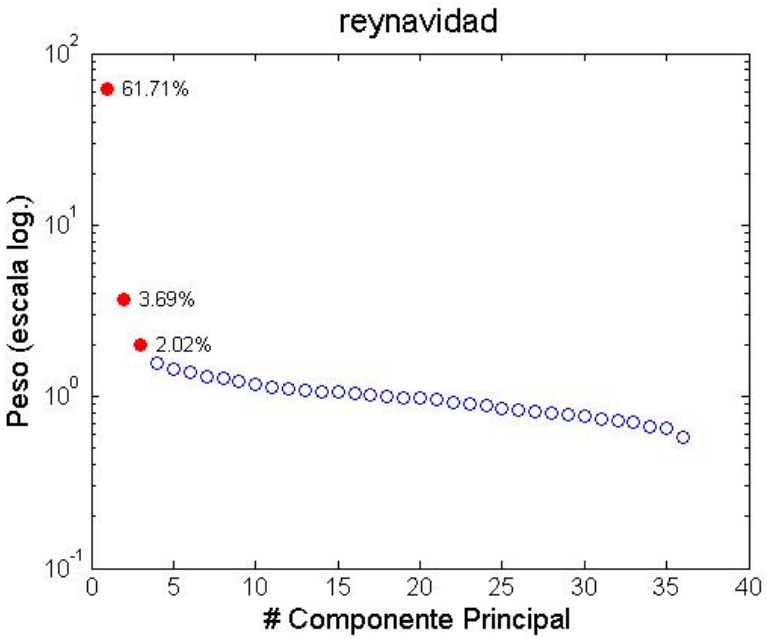


Figura 3.2: Pesos de los "textos sintéticos" obtenidos al aplicar el método LSA.

En la figura 3.3 hemos representado los vectores #1, #2 y #3 correspondientes a los tres primeros textos sintéticos respectivamente. Vemos que el vector #1 toma valores muy parecidos para todos los textos (simbolizados por cuadrados que se aproximan a una línea horizontal) mientras que los vectores #2 y #3 se distribuyen entre los textos de forma

distinta.

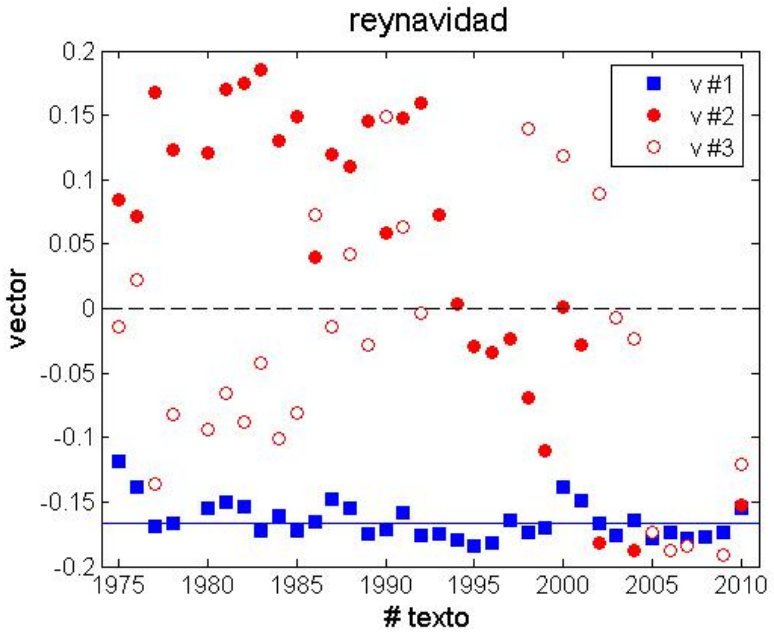


Figura 3.3: Representación de los vectores #1 (cuadrados azules), #2 (círculos rojos), y #3 (circunferencias rojas). Los valores de los vectores indican el reparto del texto sintético correspondiente entre los textos originales.

Finalmente, la representación de las coordenadas del vector #2 y del vector #3 nos permite identificar la "evolución" semántica de los textos. En la figura 3.4 se muestran las componentes de estos dos vectores para cada uno de los 36 textos analizados. A su vez, estos textos están conectados, cronológicamente, de forma sucesiva por una línea recta. Es llamativa la distribución de los discursos en décadas: de los 70 y 80 (parte derecha), de los 90 (parte superior central) y de la primera década del 2000 (inferior izquierda). El alejamiento del discurso 1979 muestra que es distinto a los demás. En principio, el discurso 1979 es el más extenso de todos.

Como observación, señalamos que se procesan todas las palabras contenido (nombre, adjetivo, verbo y adverbio) y las palabras función (determinantes, preposiciones y conjunciones simples) de cada discurso. Aunque muchas palabras se repiten y aumentan el porcentaje de lemas procesados en el vector #1, este hecho no influye en la representación de los datos. Los vectores significativos suelen ser el #2 y el #3 incluso, a veces, un vector #4 para poder visualizar mejor el contenido semántico en forma de *clusters*. La representación de los vectores #2 y #3 son los que se representan en la figura 3.4.

Para llegar a la afirmación de que no importa que un gran porcentaje de palabras, muchas de ellas palabras función, se repite en todos los textos y que este hecho no afecta a los resultados, sometimos los discursos al método. Aplicamos al conjunto de discursos una *stoplist* con la que eliminamos en los discursos aquellos vocablos más frecuentes entre las conjunciones, preposiciones, algunos pronombres, etc. En la tabla 3.22 representamos el peso porcentual de los tres primeros componentes principales. Aunque varían mucho sus valores, sobre todo para el componente principal #1, esta circunstancia no afecta al resultado final, ya que son el #2 y el #3 generalmente los que nos informan del contenido

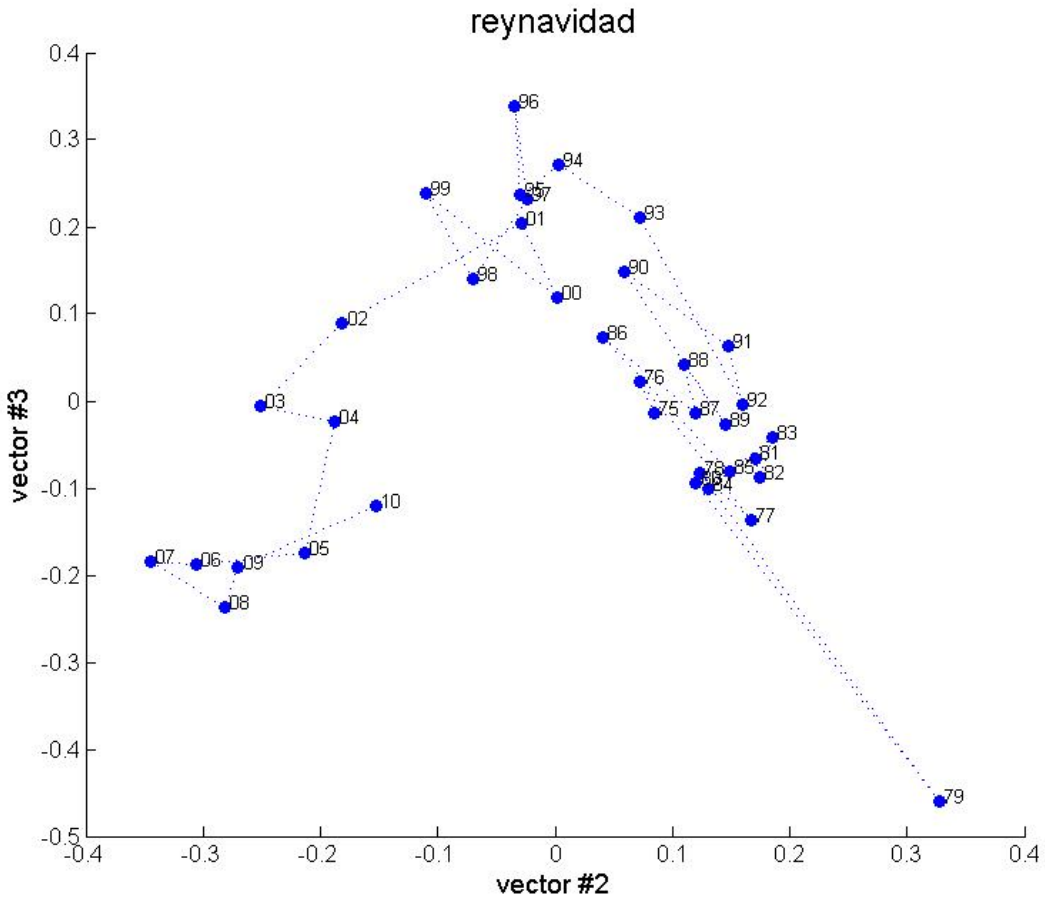


Figura 3.4: Representación de las componentes de los vectores #2 y #3. Cada punto corresponde a un discurso y las líneas de puntos trazan la conexión cronológica entre los discursos para su mejor visualización.

semántico de un texto.

3.2.4. Módulos de evaluación: Evaluator

Como ya adelantábamos en el apartado 1.4, el resultado del análisis morfo-sintáctico de FreeLing es utilizado por un conjunto de funciones y módulos de cálculo que, a su vez, utilizan la información de diccionarios y de *corpora* de referencia para obtener el resultado que genera Evaluator. En la figura 3.5 se puede ver un esquema general del proceso y flujo de información que se activan en este trabajo. Se puede apreciar que el esquema de la figura representa el proceso de nivelación desde el *input* del texto original hasta el *output* o calificación del mismo en distintos niveles lingüísticos y con diferentes métodos.

En este apartado vamos a describir someramente los módulos, programas y funciones que se han desarrollado en el ámbito de esta tesis. Éstos han sido escritos y ejecutados en Matlab (R2008b). Mediante estos módulos y funciones hemos aplicado algoritmos y procesado bases de datos para el análisis léxico, sintáctico y semántico de los discursos,

Vector	Discursos completos	Discursos aplicando la *stoplist*
#1	61, 71 %	45,52 %
#2	3,69 %	5,11 %
#3	2,02 %	2,84 %

Tabla 3.22: Peso porcentual de los tres primeros componentes principales en los discursos navideños antes y después de aplicar la *stoplist*.

textos y exámenes. El entorno Matlab ha sido utilizado mediante una licencia de uso concurrente propiedad de la Universidad Complutense de Madrid.

Exiten cuatro grandes módulos de procesamiento que se aplican sobre los textos ya procesados por FreeLing. Al conjunto de estos módulos, junto con todas las funciones que están integrados en ellos, lo denominamos "Evaluator". Los cuatro módulos que componen "Evaluator" han sido denominados como "Analizator", "Lexicator", "Sintactor" y "Semantor". De estos módulos, los tres últimos se ejecutan independientemente pero después de obtener los datos del primero. El módulo primero es "Analizator" e incluye diversas funciones que permiten obtener información de gran importancia para la calificación posterior. Los otros tres módulos utilizan la información recogida en los glosarios, calificados o no, y en los ficheros de estructuras y de relaciones de esWordnet. Aunque el módulo "Analizator" no se representa en la figura 3.5 por simplicidad, se situaría entre los resultados de FreeLing y los otros tres módulos: "Lexicator", "Sintactor" y "Semantor".

Analizator

Es un programa que recoge el fichero procesado por FreeLing y lo somete a diversas funciones o rutinas de cálculo y procesamiento para el análisis de diversos aspectos del texto. El objetivo de este módulo es obtener un fichero de resultados para cada uno de los textos analizados.

Estas rutinas son:

fhallanorepetidasyfrecuenciasentexto

```
function [pnr fpnr lnr flnr posnr fposnr] =
fhallanorepetidasyfrecuenciasentexto(palabratexto,lematexto,postexto);
```

Se trata de una función en la que se obtienen las palabras, lemas, y *PoS* no repetidos junto con sus frecuencias de aparición en el texto. Dentro de esta función se utiliza la función `frecuencimetro2` que procesa como separador de unidades de recuento el marcador 'BLANK', y que FreeLing introduce tras cada punto o signo de puntuación al término de una oración. Para hallar las frecuencias y los elementos no repetidos (ya sean vocablos, lemas, o *PoS*), la función `frecuencimetro2` se describe como:

```
function [cosanr_n,matrizfnr_n] =
frecuencimetro2(cosanr_v,matrizfnr_v,cosa_n);
```

Esta función halla las frecuencias de aparición de cualquier elemento (signos de puntuación, vocablos o lemas) dentro de las unidades realizadas en un texto. Los resultados obtenidos aquí serán utilizados profusamente en todo el análisis posterior; por ejemplo,

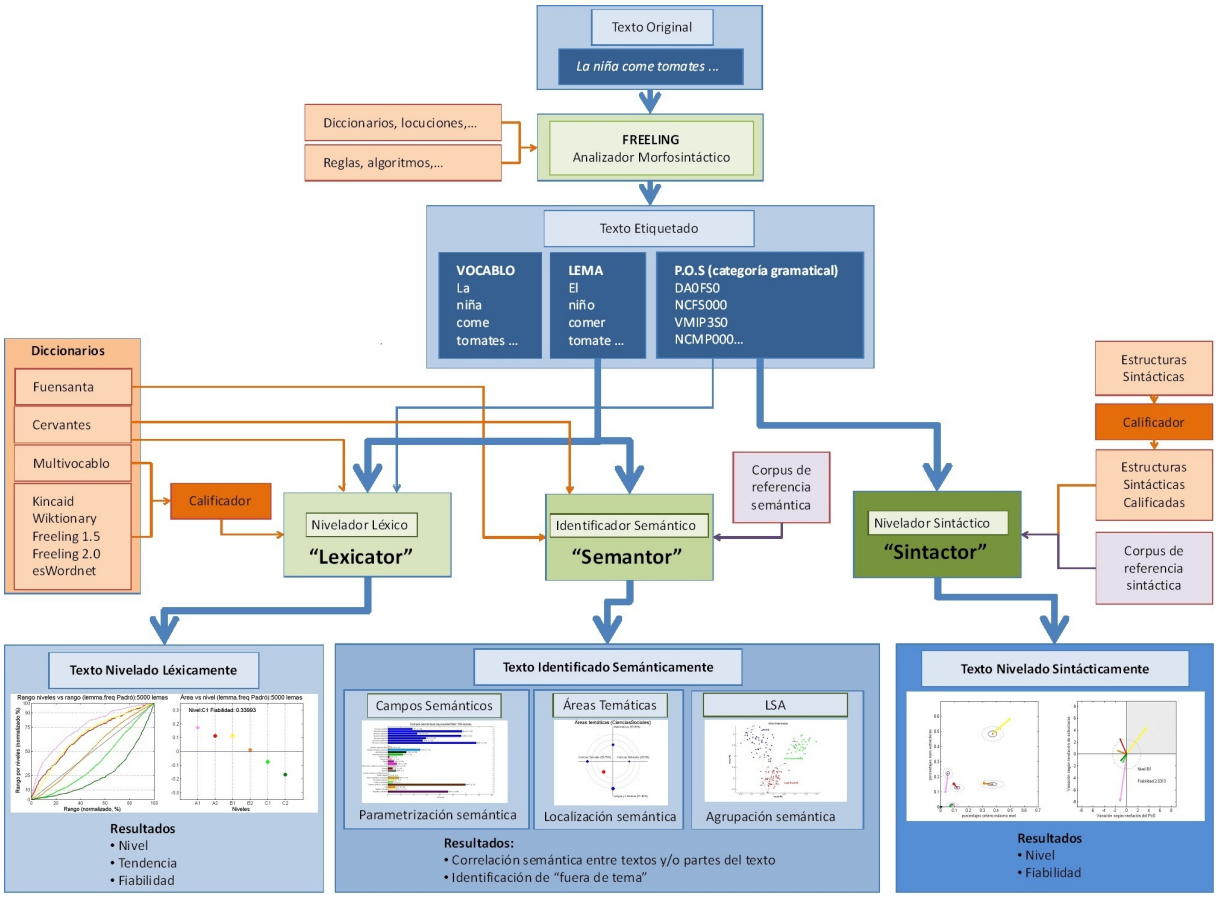


Figura 3.5: Esquema general del flujo de información desde el texto original a la calificación léxica y sintáctica, y a la identificacion semántica.

en la identificación semántica mediante el Análisis Semántico Latente (LSA), descrito en el apartado 3.2.3.

fnivelator

```
function [es,esta,nivel] =
fnivelator(palabratexto,lematexto,postexto);
```

Mediante esta función se califica por niveles léxicos a cada uno de los lemas. Es una función clave para la calificación léxica del texto. Esta función procesa todos los diccionarios disponibles: Cervantes, Kincaid, Wiktionary, FreeLing 1.5, FreeLing 2.1, esWordnet y multi-vocablos. La función genera tres variables para cada una de los *items* del discurso. La variable `es` permite determinar si el elemento es un signo de puntuación, una línea en blanco, una palabra o una locución. La variable `esta` permite determinar en qué diccionario se ha identificado cada elemento, incluyendo la posibilidad de que un elemento pueda estar en varios diccionarios. La variable `nivel` da el nivel del lema.

Por una parte, cuando se trata de un lema, la calificación se realiza en primer lugar mediante el nivel otorgado por el diccionario del Instituto Cervantes, que se toma como referencia autorizada. Si el lema no está en este diccionario, se pasa a continuación a la calificación mediante el criterio de combinación de diccionarios. Esto se realiza mediante

la función `fcalificador_estanoesta` que toma la forma:

```
function [nivel] =
fcalificador_estanoesta(esta,valornivelpalabratexto);
```

Por otra parte, los multivocablos se califican mediante su identificación en el "Índice" del Instituto Cervantes. Si el multivocablo no está en ese diccionario, que es lo más probable porque hay menos de una decena, se utiliza el fichero de locuciones calificadas que permite otorgar un nivel a dicho multivocablo.

Después de esta primera vuelta pueden existir palabras no calificadas, pero que pueden serlo identificando sus lemas o su categoría gramatical. Esto ocurre, por ejemplo, con las palabras terminadas en "-mente". Una vez determinados sus lemas correspondientes, se vuelven a someter a una calificación idéntica a la aplicada en la primera vuelta.

Una última función de este módulo, denominada `ftipovocabulario`, sirve para distinguir el tipo de vocabulario entre básico, cotidiano, académico o específico. Esto se realiza en función de la pertenencia a diccionarios y utilizando también los parámetros de frecuencia y diversidad del glosario de la Dra. Fuensanta López.

Estos resultados se utilizan en el módulo "Lexicator".

fcalificaestructuras

```
function [mapaestructuras,maximonivelposid,nestruentexto] =
fcalificaestructuras(palabratexto,lematexto,postexto);
```

Con esta función se controlan las estructuras gramaticales que aparecen en el texto. Para ello, se identifican estas estructuras gramaticales ya definidas y calificadas por niveles (desde A1 a C2). En este proceso de identificación se tiene en cuenta la definición de caracteres comodín que permiten procesar categorías flexivas. Dentro de esta función se asigna un nivel sintáctico a cada *PoS* mediante el criterio del máximo nivel (este criterio se explica con detalle en el apartado 5.2.3). Los resultados de esta función son de gran importancia para la calificación sintáctica del texto. Esta calificación se completa en el módulo "Sintactor".

fcamposyareas

```
function [cs,cs2,cstexto,cstexto2,cstexto_ponderado,af,af2,aftexto,
aftexto2,aftexto_ponderado] =
fcamposyareas(lnr,flnr,lematexto,postexto,iiseparador);
```

Esta función produce la identificación de los lemas en los campos semánticos definidos por el *Plan Curricular del Instituto Cervantes* y en las áreas temáticas definidas en el glosario de la Dra. Fuensanta López. Para ello, carga los glosarios del Instituto Cervantes y de la Dra. Fuensanta López. Los diversos métodos de recuento, incluyendo las repeticiones y las ponderaciones para la asignación semántica en función del entorno, son aplicados para obtener los resultados calculados. En esta función de cálculo se han eliminado ciertos lemas en función de su categoría gramatical, es decir, se aplica una *stoplist* al recuento para simplificar vocablos función como preposiciones y conjunciones simples, determinantes y los verbos "ser" y "estar".

Una vez realizada esta identificación, se agrupan los datos para conformar los histo-

gramas de aparición de los lemas identificados en campos semánticos y áreas temáticas. Estos resultados serán utilizados para la identificación semántica del texto mediante el módulo "Semantor".

frelacioneswn

```
function [matrizrelacion,lnrwn] =
frelacioneswn(lematexto,postexto, nivelrelacionmaximo);
```

Esta función permite obtener la relación entre dos lemas en función de las conexiones necesarias para llegar de uno a otro a través de las relaciones semánticas definidas en esWordnet. En esta función se utilizan las categorías gramaticales para limitar la búsqueda de relaciones a los sustantivos, verbos y adjetivos.

Esta función utiliza otras funciones auxiliares para crear las generaciones ascendientes y descendientes de una palabra. De éstas, varias actúan de forma sucesiva. Empezamos por la función `fcreageneraciones`:

```
function [lnrwn, ggg] =
fcreageneraciones(lematexto,postexto,nivelmaximo);
```

En esta función se utilizan los ficheros de identificación de lemas en esWordnet y también el que establece relaciones ascendentes y descendentes entre palabras. A su vez, esta función utiliza otra función denominada `generaciones` en donde se definen todos los descendientes y ascendientes de un lema.

```
function ggg =
generaciones(palabra,pospalabra,nivelmaximo,word,pos,syn,tsyn,ssyn);
```

A su vez, esta función utiliza otra denominada `padresehijos`

```
function [padres,hijos] =
padresehijos(nucleo,ssyn,tsyn);
```

Dentro de esta función se llama a otra denominada `buscarelaciones`

```
function [ssynpalabra,tsynpalabra] =
buscarelaciones(synpalabra,ssyn,tsyn);
```

que identifica la existencia o no de un lema que conecte los lemas comparados en un determinado nivel de relación.

Una vez obtenidas estas generaciones, se comparan entre sí para conocer el grado de relación (número de conexiones entre lemas) que une un lema con otro. Puesto que esta labor, tal y como está programada, es extremadamente costosa computacionalmente, se ha limitado el número de conexiones entre lemas para asegurar un tiempo de ejecución razonable.

flsa

```
function [palabraaevaluar,m,pc,vector,lambda,distancia,grupo,
cosenovectores] =
flsa(lnr,flnr,opciones):
```

Mediante esta función se calculan los parámetros de interés en el análisis semántico latente. Este análisis se realiza en el interior de cada texto. Las unidades de análisis se

determinan mediante el separador 'BLANK', que corresponde a la separación de oraciones en el texto. Por tanto, podemos decir que esta función proporciona los datos para la realización de un análisis semántico latente (LSA) en el interior de un texto. Esta función utiliza el método de análisis por componentes principales (PCA), adaptado a diversas opciones que pueden controlarse fácilmente. De esta manera se pueden utilizar, o no, pesos locales de las frecuencias, tal y como propone Landauer, por ejemplo, hallando el logaritmo de la frecuencia (Landauer *et al.*, 1998a). También se pueden eliminar las palabras que sólo aparecen una vez en el texto. Se pueden incluir pesos globales a las frecuencias de cada texto, mediante criterios asociados a la entropía o a distribuciones estadísticas diversas. Además se pueden normalizar las frecuencias de una misma palabra a lo largo de los diversos textos para matizar su peso o importancia. El resultado de esta función es un conjunto de variables que permiten identificar semánticamente el contenido de un texto.

Esta función utiliza, a su vez, la función `fclasifica_lsa`

```
function grupo =
fclasifica_lsa(lambda,npixeles,confianza);
```

Esta función permite la agrupación de componentes principales, o discursos sintéticos, con el fin de identificar aquellos que son estadísticamente independientes para separarlos de los que se agrupan en procesos de ruido.

Los resultados de este módulo aplicado a un conjunto de textos se utilizan en el módulo "Semantor" para la identificación semántica de un conjunto de textos.

En definitiva, el programa "Analizator" crea un fichero de resultados para cada uno de los discursos. Estos resultados son utilizados por los otros tres programas para la presentación de los resultados globales del análisis y para la nivelación léxica, sintáctica y la identificación semántica.

Lexicator

Este módulo utiliza los resultados obtenidos mediante el módulo "Analizator" para cada uno de los textos.

En particular, en este módulo se utilizan los valores de los niveles otorgados por los diccionarios y los valores de las variables `es` y `esta` definidas por la función `fnivelator`.

En este programa se ha incluido una función denominada `fcuentaletrasyvocales`

```
function [nletras,vocales,nvocales,vocalestilde,nvocalestilde,guionbajo,
nguionbajo] =
fcuentaletrasyvocales(palabra);
```

Mediante esta función es posible conocer la distribución estadística del número de letras en las palabras, la aparición de vocales, vocales con tilde y la presencia de guiones bajos que indican que el vocablo es una locución.

El propósito fundamental de este programa es la nivelación léxica del texto. Para ello, se utiliza el criterio de nivelación basado en la regla de Zipf (ver apartado 4.2.1) y su

seccionamiento en los seis niveles de referencia (A1, A2, B1, B2, C1, C2). Por tanto se ordenan los lemas en función de su frecuencia, se hallan sus rangos en esa ordenación y se repite esa ordenación por niveles. A partir de aquí se obtienen las áreas positivas o negativas de esta distribución de rangos y subrangos para cada nivel, y se aplica el criterio de las áreas según explicamos en el apartado 4.4.2. Como fruto de este criterio, se obtiene la tendencia de esta calificación y su fiabilidad. La fiabilidad está calculada en función del decrecimiento esperado en el cálculo de áreas por niveles.

Este cálculo está resuelto por dos funciones, la función `ffigurasreglazipf`

```
function [ii_a1,ii_a2,ii_b1,ii_b2,ii_c1,ii_c2] =
ffigurasreglazipf(f,nivellema);
```

que permite obtener gráficas para la interpretación de los resultados, y la función `distancia_bisectrizkarela`

```
function [dd_nivel] =
distancia_bisectrizkarela(vector);
```

que permite calcular el área de cada nivel con respecto a la diagonal en el diagrama subrango-rango.

Además de la nivelación producida en este programa, también se pueden obtener los histogramas de identificación por diccionarios, y los histogramas de los seis niveles de referencia del Instituto Cervantes.

Todos estos datos, cuando se realizan para un conjunto de textos con una conexión común, permiten obtener figuras y datos comparativos donde se expresan los niveles calificados. Estas representaciones son realizadas por el programa `representaresultadoslexico`.

Ejemplos de estas representaciones conjuntas se muestan a lo largo del capítulo 7, y sus datos están recogidos en el apéndice A.

Sintactor

El programa sintactor utiliza, como hemos expuesto arriba, los resultados obtenidos mediante el programa "Analizator". Especialmente se utilizan los valores ligados a la calificación sintáctica de cada elemento del discurso y a la identificación de las estructuras en el texto. Además, Sintactor utiliza un conjunto de textos de referencia o *corpus* para la comparación sintáctica de sus estructuras identificadas con otros textos. El *corpus* corresponde a los 36 discursos navideños del Rey desde 1975.

Por un lado, los resultados sintácticos de estos discursos permiten identificar la distribución de los *PoS* del discurso en los 6 niveles de referencia. Para ello, se utiliza el criterio de máximo nivel que exponemos en el apartado 5.2.3 y la distribución de las estructuras identificadas en los 6 niveles. Por tanto, para cada discurso, cada nivel tiene dos variables asociadas que pueden representarse en un diagrama bidimensional.

Por otro, los discursos del *corpus* de referencia permiten definir una posición media y un error para la localización de cada nivel. Esta posición media y error delimitan un umbral de identificación para la calificación de cualquier texto que se analice. La posición de los

6 niveles del texto, en relación a la posición de los niveles del *corpus* de referencia, define el nivel sintáctico del texto junto con un parámetro de fiabilidad en esa determinación.

El resultado es representado gráficamente para una mejor comprensión en las figuras 5.3, 5.4 y 5.5.

Semantor

El programa Semantor utiliza, una vez más, los resultados obtenidos en el módulo "Analizator". En principio, realiza un recuento de los campos semánticos distinguidos por el Instituto Cervantes en el *PCIC* y de las áreas temáticas de acuerdo a las especificadas por la Dra. Fuensanta López en su glosario.

Los campos semánticos se presentan utilizando los datos léxicos de Nociones Generales (NG) y Nociones Específicas (NE) ponderados. Se crea un vector de 28 componentes, tantos como campos semánticos ha definido el Instituto Cervantes en el *PCIC*. Este vector de cada texto es comparado con el resto de textos para conocer su coherencia semántica con el asunto tratado.

De forma similar, y más general, se ha procedido con las áreas temáticas en donde se ha preferido definir una localización para cada texto comparando y seccionando las áreas en cuatro: Matemáticas, Lengua y Literatura, Ciencias Naturales y Ciencias Sociales en un diagrama bidimensional de fácil interpretación.

Por último, se realiza un análisis por componentes principales (PCA) de varios discursos entre sí. Este análisis permite agrupar conjuntos de textos a modo de "nubes" o *clusters* en función de su contenido semántico. Para ello, se recurre a un *corpus* maestro con el nivel de referencia precisado o a varios *corpora* de referencia semántica (éste último es nuestro caso), que permiten definir en un diagrama regiones de los vectores propios del análisis por componente principales (PCA), según explicamos en el apartado 3.2.3. Además, es posible identificar cuantitativamente la importancia de los textos sintéticos obtenidos en el análisis. El programa está preparado para realizar un análisis semántico latente con gran flexibilidad y eficacia.

Finalmente, en este estado del sistema de evaluación y debido al alto coste computacional, no hemos incluido los resultados obtenidos al procesar los textos con esWordnet. Sin embargo, consideramos que, tras una implementación más eficaz de los algoritmos de identificación y una vez superadas sus debilidades, el uso de esWordnet puede producir valiosos resultados semánticos.

Capítulo 4

Análisis léxico

4.1. Nivelación del léxico

4.1.1. Nivelación mediante el "Índice de Nociones Generales y Nociones Específicas" del *PCIC*

Si el *Plan Curricular del Instituto Cervantes* (*PCIC*) "marca los niveles de referencia para la enseñanza del español" (Cervantes, 2006, A1-A2: 8), el "Índice de nociones generales y nociones específicas" del *PCIC* se puede definir como el glosario patrón para el diagnóstico y nivelación del léxico de textos escritos por estudiantes y aprendices del español como lengua extranjera.

En consecuencia, nuestro modelo de propuesta de nivelación automática del léxico en textos escritos por aprendices de español se basa fundamentalmente en el "Índice de nociones generales y nociones específicas" del *PCIC*. Con ello, queremos contribuir con nuestra investigación y resultados a poner en disposición de los aprendices y evaluadores recursos que les permitan diagnosticar el nivel de un texto automáticamente. Queremos, además, aportar nuestra contribución para promover el autoaprendizaje y la autogestión de sus textos.

La dificultad de nuestra investigación en realidad estriba en cómo se nivela un texto. El procedimiento para nivelar el léxico de un texto lo exponemos a continuación.

Primero, debemos disponer de un texto en castellano en formato plano para poder lematizarlo y desambiguarlo gramaticalmente. La lematización se hace con el programa de FreeLing. Luego, el texto lematizado se procesa con el "Índice" del *PCIC* también lematizado. De forma automática, con el programa Lexicator, nivelamos los lemas del texto dándoles el nivel de los lemas del "Índice" del *PCIC*. Esto es, si los lemas del texto coinciden con los lemas del "Índice" se clasifican independientemente de su categoría gramatical o *PoS*. Al procesar los textos con los niveles del "Índice" del *PCIC*, obtenemos una gráfica de nivel de un texto que muestra el histograma de frecuencia de aparición de los vocablos del texto en el color correspondiente a los colores de los tomos del *PCIC*: Nivel A1-A2 en rojo, nivel B1-B2 en naranja y nivel C1-C2 en verde. Nosotros hemos desdoblado los niveles por colores para su mejor representación, resultando un nivel A1

rosa, un A2 rojo, un B1 amarillo, un B2 naranja, un C1 verde claro y un C2 verde oscuro, como se muestra en la figura 4.1. Esta figura representa los resultados para los discursos navideños del Rey Juan Carlos I. Las barras azules representan la dispersión en la determinación de las frecuencias de aparición.

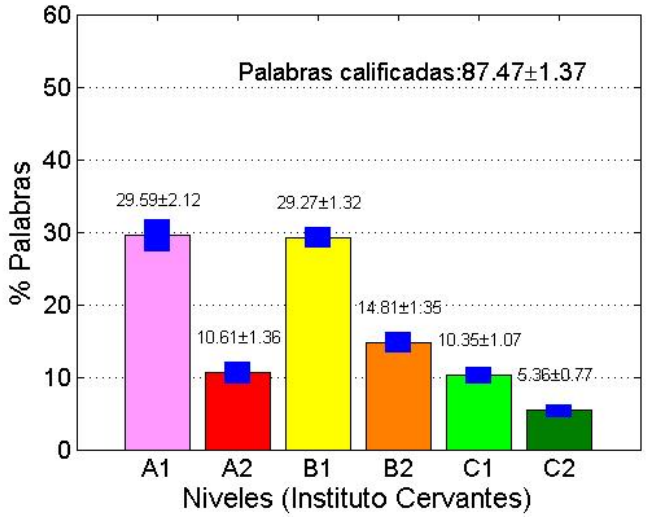


Figura 4.1: Representación de la distribución del léxico en el conjunto de discursos navideños, según los niveles de referencia del *PCIC*.

Como resultado, obtenemos un texto cuyo léxico está clasificado por niveles. Puesto que en cualquier texto encontramos léxicos de casi todos los niveles, inmediatamente nos cuestionamos dos preguntas clave: ¿Qué nivel tiene el texto?¿Qué determina que un texto sea de un nivel B1, B2, C1 o C2? Estas cuestiones las resolvemos en el apartado 4.4.

4.1.2. Nivelación de diccionarios mediante el "Índice" del *PCIC* como modelo referente

Un método aproximado de nivelación léxica es la de nivelar algunos de los glosarios de los que disponemos mediante el "Índice de nociones generales y nociones específicas" del *PCIC*. En la figura 4.2 se representa gráficamente la distribución de niveles por diccionarios. Como es lógico, se observa que vocablos de niveles A2-B1 muy frecuentes y funcionales se hallan dentro de un diccionario simplificado y básico como el del Dr. Kincaid (Kc); vocablos frecuentes de un lenguaje familiar y habitual con niveles A1-B2 están en un glosario más cotidiano como el Wiktionary (Wk); vocablos entre niveles B1-B2, incluso C1, se encuentran en un glosario general como FreeLing 1.5 (F1); vocablos formales y menos frecuentes de un C1 e incluso C2 se listan en un glosario amplio como FreeLing 2.1 (F2); y vocablos más especializados de nivel C2 se indexan en un glosario más extenso como esWordnet (Wn).

Como se observa en la tabla 4.1, el contenido del "Índice" del *PCIC* está prácticamente contenido en FreeLing 2.1 y en el esWordnet. La especificidad del glosario de la Dra. Fuensanta López es el menos representado dentro del "Indice" y el glosario del Dr. Kincaid se manifiesta como un glosario muy básico pero menor que el Wiktionay español.

Nº lemas	Glosarios	Cervantes	Kincaid	Wiktionary	Fuensanta	FreeLing 1	FreeLing 2	Wordnet	Locuciones
8662	Cervantes	x	23,47%	44,53%	41,91%	54,70%	94,84%	90,82%	0,15%
2022	Kincaid	63,45%	x	79,77%	60,58%	84,32%	88,87%	84,72%	6,58%
5207	Wiktionary	50,01%	29,15%	x	43,46%	65,55%	87,13%	88,48%	0,25%
5273	Fuensanta	47,39%	22,51%	42,94%	x	60,97%	98,07%	92,89%	0,00%
7156	FreeLing 1	47,74%	25,06%	51,80%	46,59%	x	96,49%	86,21%	0,10%
76214	FreeLing 2	9,39%	2,66%	7,00%	7,40%	10,00%	x	40,29%	0,01%
93425	Wordnet	21,15%	8,88%	19,01%	17,31%	23,35%	60,92%	x	0,40%
5868	Locuciones	0,20%	2,37%	0,24%	0,00%	0,07%	0,12%	5,71%	x

Tabla 4.1: Porcentaje de lemas de un diccionario (de la fila) que está en otro diccionario (de la columna). El número total de lemas de cada diccionario está listado en la primera columna de la tabla.

Además de los vocablos autorizadamente nivelados por el *PCIC*, estos diccionarios nos sirven de referencia para nivelar el léxico que no nivele el *PCIC*. Para ello proponemos el método de superposición o combinación de diccionarios que desarrollamos en el apartado 4.1.4 como otro método para nivelar más lemas que los que nivela el "Índice" del *PCIC*.

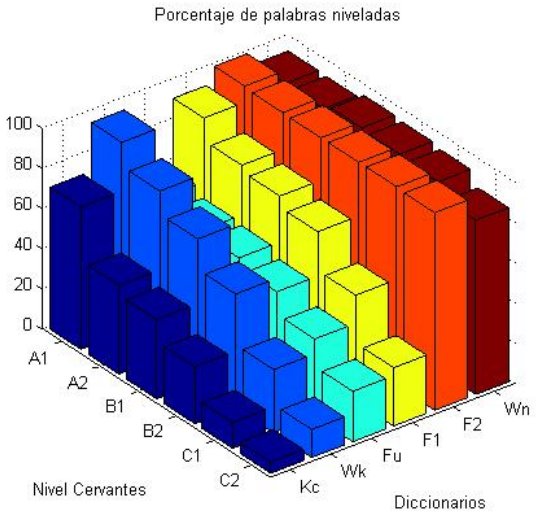


Figura 4.2: Distribución de niveles de referencia del *PCIC* en los distintos diccionarios.

4.1.3. Identificación del nivel del vocablo no procesado por el listado del Dr. Kincaid

Una de las características de los glosarios es la información específica registrada en ellos o la propia concepción del glosario. Por un lado, el listado del Dr. Kincaid es un listado básico desambiguado, adaptado al castellano, y nos sirve de lista de control para identificar casi el 100 % de vocablos de un nivel B1. De manera que el léxico no controlado por el glosario del Dr. Kincaid pertenece ya a niveles B2. Consecuentemente, los diccionarios de FreeLing 1.5 y FreeLing 2.1 se muestran como listados generales que nos permite identificar la mayor parte de los vocablos de un texto. Por un lado, el glosario de FreeLing 1.5 tiene

81.826 vocablos relacionados con 7.156 lemas que nos permiten identificar entre un 75 %-90 % de palabras dentro de un texto de nivel B2 y C1. Por otro, el glosario de FreeLing 2.1, con un glosario de 556.213 vocablos relacionados con 76.214 lemas, nos permitirá identificar el léxico restante de nivel más elevado, esto es, el 20 %-35 % de vocablos de C1 e incluso C2, según el mayor o menor número de vocablos empleados en el texto que se evalúa.

De manera semejante, esWordnet 1.6 se utiliza como lista de identificación de vocablos y como diccionario referente de nivel. Es un glosario de 93.425 entradas desambiguadas. Su tamaño y configuración no sólo amplía la posibilidad de procesar mayor número de vocablos específicos sino incluso de identificar el nivel de aquellos vocablos que el *PCIC* no identifica.

No obstante, se ha observado que el hecho de que un glosario tenga un mayor número de palabras no significa que se obtenga un mayor porcentaje de palabras identificadas en un texto sino que, cuanto mayor sea el contenido de vocablos específicos y menos frecuentes, se van a procesar menos palabras de nivel B1 en un texto. Por una parte, el glosario de esWordnet 1.6 es posible que identifique un menor número de vocablos de B1 que el del Dr. Kincaid. Por otra parte, el glosario de la Dr. Fuensanta López no es muy extenso pero sí muy específico. Al ser un listado de palabras relacionado con las matemáticas, las ciencias naturales, las ciencias sociales, la antropología y la lingüística, el porcentaje de identificación de vocablos de B1 puede ser muy bajo.

El glosario Wiktionary español es otra herramienta válida para identificar y nivelar un gran porcentaje de vocablos básicos. Este listado nos confirma la fiabilidad del listado de 2.022 lemas relativos a las palabras más frecuentes registradas por el Dr. Kincaid, ya que ambos coinciden en muchos lemas listados por su alta frecuencia. Sin embargo, aunque parece coincidir el criterio de clasificación léxica por niveles de las palabras más frecuentes del Wiktionary o del Kincaid con el criterio de niveles del glosario de *PCIC*, al procesar las frecuencias del Wiktionary, la dispersión entre las frecuencias y el nivel de lengua es tan alta que este método no ha funcionado. Es decir, se quería asociar la frecuencia de los vocablos con los distintos niveles de lengua del glosario del *PCIC* ya que parecía un método lógico para otorgar un nivel a cada lema del Wiktionary de español y, así, probar el nivel básico tanto del Wiktionary de español como del glosario del Dr. Kincaid. Se planteaba la hipótesis de que el criterio de mayor frecuencia de una palabra implicaría la pertenencia a un nivel A1-A2 y el de menos frecuencia apuntaría a un nivel B2 o C1. Sin embargo, la configuración del "Indice" de *PCIC*, no sólo está articulado en función del criterio de calificación de los niveles de referencia y adaptación a la enseñanza de idiomas sino también del criterio de clasificación. El *PCIC* organiza el vocabulario, no sólo en función de niveles y adaptación a la enseñanza de idiomas sino también en función de la clasificación del registro más o menos formal o de la noción más o menos específica dentro de un campo semántico y de la combinación léxica de un vocablo con otros (Cervantes, 2006, C1-C2: 719). En definitiva, cada glosario aplica criterios diferentes en la elección del vocabulario. Además, si el Wiktionary y el Kincaid indexan muchas palabras función y propias del lenguaje básico y muy frecuente, el *PCIC* indexa muchas palabras contenido de un lenguaje estándar y funcional.

No obstante, al procesar un texto mediante distintos tipos de glosarios, se busca extraer

y controlar el mayor número de palabras correctamente escritas en dicho texto para poder identificar el tipo, variedad de léxico y nivel. Además, el uso de varios diccionarios nos puede aportar otros índices evaluables: la riqueza del vocabulario del texto, la repetición, la especificidad o, incluso con el "Índice" del *PCIC*, la identificación de un vocablo de noción general o específica.

4.1.4. Identificación del nivel de vocablos por combinación de cinco glosarios

Por una parte, suponiendo que un diccionario tiene más nivel cuanto más entradas registra, hemos expuesto en el apartado 4.1.2 que el diccionario del Dr. Kincaid se corresponde con un nivel B1-B2 y el Wiktionary se corresponde mejor con el nivel B2; el de FreeLing 1.5 está entre un nivel B2-C1, y el de FreeLing 2.1 y esWordnet están entre un nivel C1-C2. Por otra parte, como el "Índice" del *PCIC* no nivela la totalidad del léxico de los textos, vamos a aplicar el criterio de que el hecho de que unos vocablos estén en un glosario pero no en otro, es un indicador de nivel del diccionario y este criterio nos va a servir para poder disponer de más lemas nivelados mediante el método de combinación de diccionarios.

Como manejamos dos conceptos, extensión del glosario y nivel de los vocablos, se ha hecho la comprobación de nivelación de los glosarios, en el primer caso, y la combinación de glosarios, en el segundo. En ambos procesos, se ha tenido como patrón referente de nivel al "Índice" del *PCIC* para comprobar las dos hipótesis con los resultados obtenidos.

Antes de proceder a comprobar la primera hipótesis, para computar el mayor número de términos de todos los glosarios, tuvimos que procesar, por un lado, los dos diccionarios morfológicos de FreeLing y, por otro, ajustar de nuevo ciertos vocablos del glosario del *PCIC* que no se procesaban, como participios o adverbios acabados en "-mente", y ajustar vocablos a sus lemas (ver apartado 3.1.6). Al hacer la comprobación con los diccionarios de FreeLing, se decidió procesar los vocablos en vez de los lemas. Esta decisión mejoró el procesamiento de los datos con el glosario del *PCIC* porque, si sólo contrastábamos los lemas de FreeLing y los lemas del *PCIC*, más de 1.200 términos del *PCIC* quedaban sin procesar. El nuevo reajuste en el procesamiento de los diccionarios de FreeLing y del glosario del *PCIC* ha permitido recuperar y autentificar el nivel de unos 1.140 términos, útiles por ser igualmente términos-patrón de nivel de referencia para la identificación de niveles de los lemas de los glosarios y de otros lemas en la combinación entre glosarios.

A continuación, procesamos automáticamente todos los glosarios tomando como referente nivelador el "Índice" del *PCIC*. Por un lado, nivelamos los cinco glosarios con el "Índice" del *PCIC* como mostramos en el apartado 4.1.2 y en la figura 4.2. Por otro lado, vamos a combinar los cinco glosarios y nivelar los vocablos para utilizar, en subsiguientes nivelaciones, los resultados de las combinaciones de vocablos más productivas junto con los vocablos nivelados con el "Índice" del *PCIC*.

Respecto a la primera propuesta de dar un nivel a cada glosario, teniendo como referencia el "Índice" del *PCIC* y el de la Dra. Fuensanta López, debemos señalar que, según los datos procesados, el glosario de la Dra. Fuensanta López no es muy significativo

con respecto a los otros glosarios de manera que lo descartamos en lo que respecta a la combinación para dar un nivel léxico a los lemas de un texto.

Después de procesar todos los glosarios, a excepción del de la Dra. Fuensanta López, los resultados numéricos obtenidos que se plasman en la figura 4.3 confirman nuestra primera hipótesis: el nivel de un glosario es mayor en función del número de términos. Esto se observa al identificar los niveles del *PCIC* y hallar en qué porcentaje se distribuyen los lemas en cada nivel. Por ejemplo, en el glosario del Dr. Kincaid, un 71,04 % de lemas de nivel A1 están en el *PCIC*; contiene además un 44 % de nivel A2 y un 39,59 % de vocablos de nivel B1, mientras que un 28,38 % de vocablos son de nivel de B2, un 17,96 % de nivel C1 y un 5,57 % de nivel C2. En el listado de Wiktionary, respecto al *PCIC*, contiene un 95,32 % de vocablos de A1, un 83,42 % de nivel A2, un 71,87 % de vocablos de B1, un 56,77 % de vocablos de B2, un 31,48 % de nivel C1 y un 13,56 % de nivel C2.

Como se puede observar en esta misma figura 4.3, tanto el glosario del Dr. Kincaid, el Wiktionary y el FreeLing 1.5 tienen una distribución semejante. Aunque cada glosario es mayor que el anterior, la relación de pertenencia de vocablos a los niveles más altos es similar. Como es de esperar, al aumentar el número de vocablos en Wiktionary (5.208) y FreeLing 1.5 (7.156 lemas) respecto al Kincaid (2.022), también va aumentando el número de vocablos en los niveles más altos. Lógicamente, los glosarios más extensos, FreeLing 2.1 (76.214) y esWordnet (93.425), registran no sólo la mayoría de los vocablos del Dr. Kincaid y FreeLing 1.5 (ver tabla 4.1) sino que contienen la mayoría de los vocablos en todos los niveles identificados en el "Índice" del *PCIC*.

Concluimos que nuestro supuesto de que cada glosario tiene un nivel en función de su tamaño es acertado al haber sometido los glosarios a la distribución de los vocablos por niveles del *PCIC* y al comprobar que el porcentaje de distribución de vocablos por niveles va en aumento. Consecuentemente, su nivel es mayor a medida que el diccionario es más extenso.

También para la segunda hipótesis, proponemos seis combinaciones entre los cinco glosarios. A los glosarios los vamos a identificar con las iniciales "Kc" para el glosario del Dr. Kincaid, "Wk" para el listado de Wiktionay, "F1" para el diccionario de FreeLing 1.5, "F2" para el diccionario de FreeLing 2.1 y "Wn" para esWordnet.

En la combinación de glosarios el dígito "1" significa que en cierto glosario hay unos vocablos con un nivel determinado, mientras que el dígito "0" indica que en cierto glosario no hay vocablos de dicho glosario.

En una primera aproximación proponemos las siguientes combinaciones y sus niveles:

Combinación 11111 = Kc, Wk, F1, F2, Wn; B1
Combinación 01111 = Wk, F1, F2, Wn; B2
Combinación 00111 = F1, F2, Wn; C1
Combinación 00011 = F2,Wn; C1
Combinación 00010 = F2; C2
Combinación 00001 = Wn; C2

Esta propuesta sobre las combinaciones y sus niveles se procesa con el "Índice" del *PCIC* como *nivel patrón* y se obtienen los resultados que se representan gráficamente en

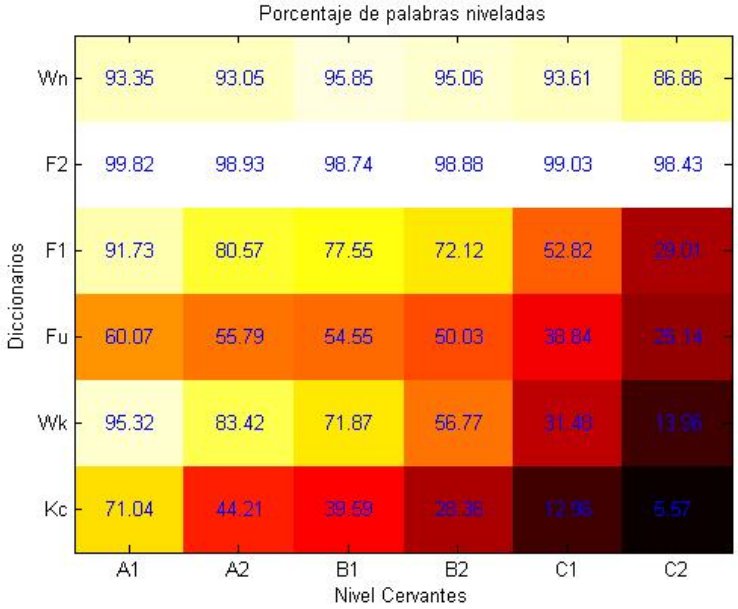


Figura 4.3: Distribución de vocablos por niveles en cada glosario. Los números expresan el porcentaje de palabras del diccionario del *PCIC* con el nivel correspondiente que tienen los vocablos en cada uno de los diccionarios analizados.

la figura 4.4 y en la figura 4.5.

Al analizar la figura anterior 4.4, comprobamos que se valida la nivelación según el criterio de combinación de diccionarios. En esta figura se representa el número de palabras en función del nivel otorgado por ambos métodos (el criterio de combinación y el criterio del *PCIC*). Si el criterio de nivelación por la combinación de diccionarios fuera perfecto, el nivel otorgado sería el mismo que el dado por el *PCIC*. En ese caso, cada palabra contribuiría a aumentar el valor de la diagonal de esta gráfica ya que en esta diagonal se representan las palabras que tienen el mismo nivel por ambos métodos.

Sin embargo, al nivelar mediante nuestra combinación de diccionarios, no todas las palabras aparecen en la diagonal. En la figura 4.4 se observa que el máximo número de palabras sí se sitúa sobre la diagonal salvo un par de excepciones. Las palabras calificadas por el *PCIC* con un nivel C1 aparecen mayoritariamente en un nivel C2, según la combinación de diccionarios, es decir, nuestra combinación sobrevalora la nivelación del *PCIC,* mientras que las palabras calificadas con un B1, según nuestro criterio de combinación, son calificadas por el *PCIC* mayoritariamente como nivel B2. Por lo tanto, nuestra nivelación léxica va a ser un poco inferior en los análisis y se observará en los resultados ya que nos centramos en los niveles B2 y C1 en este estudio. Como en nuestra nivelación no consideramos posibles combinaciones de A1-A2, puesto que no es el objetivo de esta tesis, las columnas de "0" marcan esa ausencia de combinación de diccionarios.

Posteriormente, después de procesar automáticamente los datos, se revelan dos combinaciones más de interés que no habíamos considerado y que se observan en la figura 4.6. Estas combinaciones son:

Combinación 10111 = Kc, F1, F2, Wn; C1

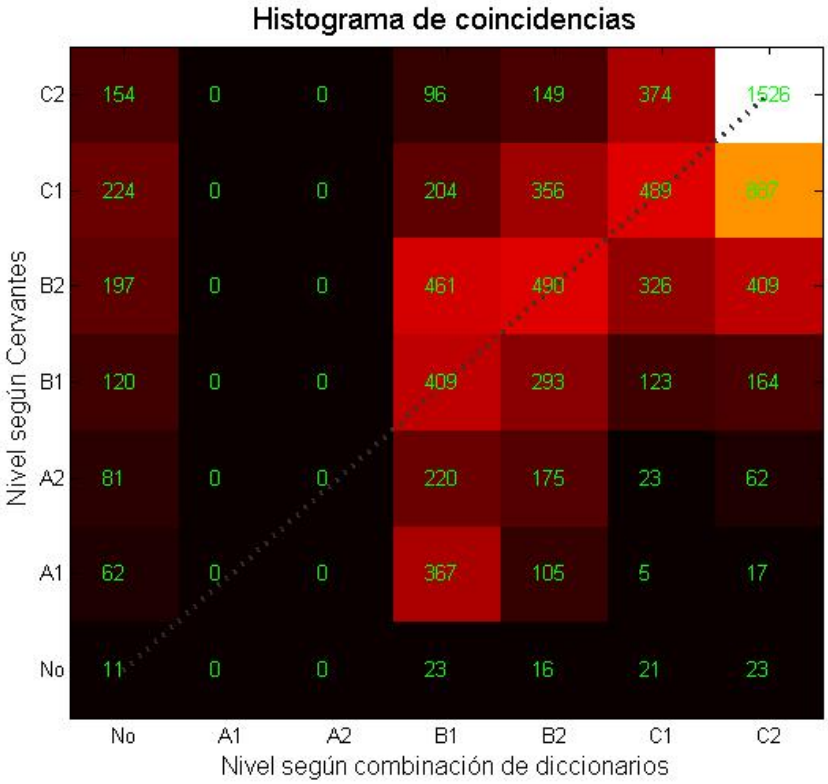


Figura 4.4: Mapa de combinación de diccionarios.

Combinación 01011 = Wk, F2, Wn; C1

En resumen, de todas las posibles combinaciones, en relación al mayor número de vocablos calificados con un nivel, destacamos las siguientes por orden de relevancia:

1ª 11111 = Kc, Wk, F1, F2, Wn: B1-B2
2ª 01111 = Wk, F1, F2, Wn: B2
3ª 00111 = F1, F2, Wn: C1
4ª 00011 = F2, Wn: C1
5ª 00010 = F2: C2
6ª 00001 = Wn: C2
7ª 10111 = Kc, F1, F2, Wn: C1
8ª 01011 = Wk, F2, Wn: C1

En la figura 4.6 estas combinaciones de glosarios, limitadas entre líneas paralelas, muestran la distribución de sus vocablos y el nivel predominante en cada combinación. Además, la figura refleja el número de vocablos mediante una gama de colores del negro al blanco para indicar desde ausencia de vocablos sin nivelar, en color negro, hasta alta presencia de vocablos con el nivel más alto, en color blanco.

A continuación, en la tabla 4.2 observamos la similitud de niveles que se obtiene a partir de la combinación propuesta de los diccionarios.

Al final de este análisis podemos plantear y responder las siguientes preguntas:

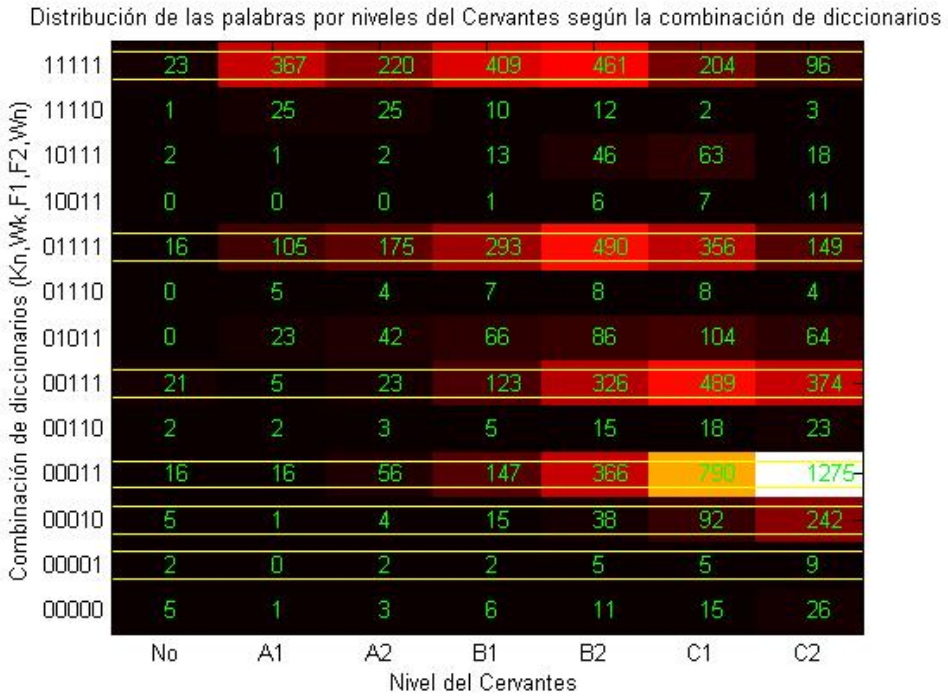


Figura 4.5: Gráfica con varias combinaciones de diccionarios. Se han marcado con líneas horizontales las seis combinaciones elegidas de diccionarios y sus niveles.

- ¿Qué obtenemos de estas combinaciones de glosarios al tomar como referente el glosario del *PCIC* para darles un nivel?
- Obtenemos la distribución de los vocablos de los glosarios por niveles según los niveles marcados por el Instituto Cervantes.

- ¿Qué aplicación tiene el método de la combinación de diccionarios?
- Podemos nivelar entre un 30 %-40 % más de vocablos en un texto. Es decir, si, por ejemplo, con el glosario del *PCIC* nivelábamos un 40 %-45 % de lemas, ahora podemos nivelar un 70 %-80 % del léxico de un texto.

#	**Combinación propuesta**	**Nivel**	**Combinación automática**	**Nivel**
1ª- 11111	Kc, Wk, F1, F2, Wn	B1-B2	Kc, Wk, F1, F2, Wn	B1-B2
2ª- 01111	Wk, F1, F2, Wn	B2	Wk, F1, F2, Wn	B2
3ª- 00111	F1, F2, Wn	C1	F1, F2, Wn	C1
4ª- 00011	F2, Wn	C1	F2, Wn	C1
5ª- 00010	F2	C2	F2	C2
6ª- 00001	Wn	C2	Wn	C2
7ª- 10111	-	-	K, F1, F2, Wn	C1
8ª- 01011	-	-	Wk, F2, Wn	C1

Tabla 4.2: Comparativa de la combinación de diccionarios y los niveles.

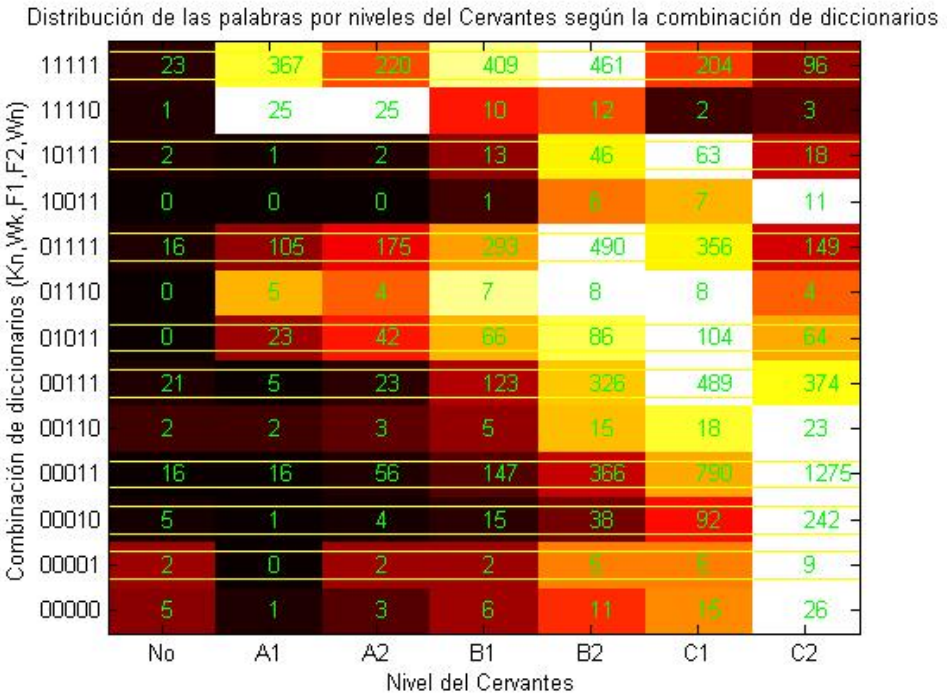


Figura 4.6: Gráfica con ocho combinaciones procesadas de diccionarios y sus niveles.

4.1.5. Identificación del tipo de vocablo por ubicación en un determinado glosario

Otro glosario específico que permite diferenciar el tipo de vocabulario de un texto es el Vocabulario Básico de Orientación Didáctica de la Dra. Fuensanta López. Este glosario lista sólo vocablos nominales y verbales propios de temas académicos. Los vocablos se estructuran en cinco niveles en función del índice de dispersión de cada vocablo. Como el porcentaje de coincidencia de nivel del glosario de la Dra. Fuensanta López y el nivel de aprendizaje del "Índice" del *PCIC* es muy bajo, no vamos a considerar esta comparación. La distribución de los niveles de la Dra. Fuensanta López no coinciden con los niveles de aprendizaje del "Índice" del *PCIC*. No obstante, podemos identificar de forma automática qué tipo de vocabulario predomina en un texto. Con ello podemos probar que el glosario de la Dra. Fuensanta López es una herramienta de referencia para determinar qué porcentaje de vocabulario es académico.

Además, basándonos en los niveles del "Índice" del *PCIC* y en el criterio de ubicación de un vocablo en uno u otro diccionario, podemos señalar el porcentaje de vocablos básicos, general o especializado de un texto siempre que sus vocablos se hallen en el listado de lemas del Dr. Kincaid, del Wiktionary y FreeLing 1.5, o del FreeLing 2.1 y esWordnet, respectivamente. Por ejemplo, el glosario de la Dra. Fuensanta López lista vocablos específicos, de un nivel 4 o nivel 5 (según los niveles definidos por la Dra. Fuensanta-López), que el "Índice" del *PCIC* no contiene. Este hecho nos permite identificar no el nivel concreto de la palabra sino qué tipo de vocabulario tiene el glosario. Por ello, tipificamos que un vocabulario es básico cuando dicho vocablo tiene en el "Índice" del *PCIC* un nivel entre A1-B1 o se halla en el glosario del Dr. Kincaid; un vocabulario general sería aquel que

contiene términos que el "Índice" del *PCIC* califica de B1-B2 y se hallan en el Wiktionary y FreeLing 1.5; un vocabulario académico (B2-C2) sería el de la Dra. Fuensanta López; y específico (C2) el que contiene lemas que se hallan en los lemarios de FreeLing 2.1 y esWordnet. Según el criterio de frecuencia y el criterio de distribución de vocablos en áreas temáticas, dos de los criterios que utiliza la Dra. Fuensanta López para establecer sus niveles, se podría decir que pertenecen al tipo específico aquellos términos de muy baja frecuencia y poca distribución en algunas materias del glosario de la Dra. Fuensanta López.

Para identificar el tipo de vocabulario de cada glosario comparamos el glosario de la Dra. Fuensanta López con el "Índice de nociones generales y nociones específicas" del *PCIC*, con el del Dr. Kincaid, con el Wiktionary, con los dos glosarios de FreeLing y con el esWordnet. En definitiva, consideramos tres variables para la clasificación del tipo de vocabulario: la ubicación o no del término en uno u otro glosario, la frecuencia y el rango de frecuencia, y el número de materias en las que aparece el término dentro del glosario de la Dra. Fuensanta López. Como una identificación general, se podría programar un módulo en el que se cumplieran las condiciones que exponemos a continuación para distinguir entre un vocabulario básico, genérico, específico o académico que permitiera identificar los textos analizados dentro de un registro discursivo (Nation y Kyongho, 1995).

- Básico (B1): Consideramos que un vocablo es básico cuando es un vocablo que se aprende en las primeras etapas de aprendizaje. Por ello, estará listado en el *PCIC* con un nivel A1 o A2 o B1, en el del Dr. Kincaid y tendrá mayor frecuencia en el glosario de la Dra. Fuensanta López.
- Genérico (B2-C1): Incluimos en esta clasificación a aquel vocabulario propio de los niveles intermedios del aprendizaje. Por tanto, el término podrá estar en el *PCIC* con un nivel B1 o B2 o en el FreeLing 1.5 o en el Wiktionary.
- Académico (B2-C2): Un vocablo, nominal o verbal, se considerará académico cuando sea propio de la temática de la enseñanza reglada. El hecho de estar en el glosario de la Dra. Fuensanta López, dentro de un rango de distribución de frecuencia, es un referente válido.
- Específico (C2): Se clasificará como específico a aquel vocablo con una frecuencia muy baja y rango alto, que esté en el *PCIC* con un nivel C2, en FreeLing 2.1 o en esWordnet pero que no esté en FreeLing 1.5 ni en el Wiktionary con una frecuencia alta.

4.1.6. Nivelación de locuciones o multi-vocablos por niveles de aprendizaje

El método que hemos seguido para procesar y nivelar las locuciones ha sido minucioso. En primer lugar, antes de otorgar un valor a las locuciones, para poder procesar cualquier texto con el programa de FreeLing, se ha etiquetado manualmente cada multi-vocablo, se le ha asociado un lema y se le ha identificando con un *PoS* o etiqueta que marca la categoría gramatical. Una vez que cada multi-vocablo se ha marcado y configurado

en el glosario de locuciones para que el programa de FreeLing lo analice e identifique, dicho multi-vocablo está listo para ser procesado por Analizator con la función que nivela (`fnivelator`) a partir de los criterios de nivel que hemos elaborado. No obstante, el glosario de multi-vocablos es un listado cerrado pero susceptible de ser ampliado de forma manual constantemente.

4.1.6.1. Criterios de asignación de nivel

Entre nuestros objetivos está nivelar todos los multi-vocablos registrados, de forma manual o automática. Por ello, por una parte, se ha marcado manualmente con un determinado nivel aquella locución que recoge alguno de los inventarios del *PCIC* o el "Índice de nociones generales y específicas" del *PCIC*. Por ejemplo, se ha marcado con un nivel B2 en el glosario de locuciones a aquella lexía que el *PCIC* lista con ese nivel B2. Esto es, se ha añadido manualmente el nivel de lengua apuntado por el *PCIC* tanto a las locuciones ya indexadas, por el Dr. Padró o por nosotros, como a las nuevas propuestas por el *PCIC* y que hemos incorporado. Esto ha sido posible gracias a la publicación de los inventarios del *PCIC* que organizan vocablos, colocaciones o expresiones por niveles en función de criterios gramaticales, pragmáticos y funcionales, según la etapa de enseñanza y aprendizaje del castellano en la que se considera propio su aprendizaje.

Por otra parte, hemos nivelado automáticamente aquellas locuciones listadas a las que no hemos otorgado un nivel previo porque no las indexan ni los inventarios ni el "Índice de nociones generales y nociones específicas" del *PCIC* expresamente. Para ello, utilizamos un módulo de nivelación en Analizator a través de una subrutina o función llamada `fnivelator`. Analizator es uno de los módulos desarrollado durante la elaboración de esta tesis para nivelar tanto vocablos como multi-vocablos que no nivela el *PCIC*. Esta herramienta reconoce y otorga automáticamente un nivel de lengua a partir de los criterios de nivel propuestos por nosotros a todos los multi-vocablos indexados sin marca de nivel de lengua. Concretamente, el nivel que se da a los multi-vocablos en algunas ocasiones coincide con los criterios del *PCIC* y en otras no. No obstante es destacable que, con respecto a la calificación de los multi-vocablos por niveles del *PCIC*, nuestra herramienta evalúa más bajo que el *PCIC*.

Efectivamente, apuntamos a continuación criterios de evaluación del léxico que el *PCIC* aplica tanto para vocablos como para multi-vocablos. El *PCIC* marca con un nivel u otro ciertos vocablos según la etapa de aprendizaje en la que se considera apropiado aprender dicho vocablo; califica con mayor o menor nivel a aquel vocablo que se enmarca dentro de unos criterios gramaticales, funcionales o pragmáticos determinados o bien dentro de una noción específica o general. El *PCIC* apenas considera ninguna locución adverbial dentro del nivel B1 (Cervantes, 2006, B1, B2: 74) y eleva el nivel de un vocablo al combinarse con otro vocablo o al convertirse en una colocación léxica. Consecuentemente, nosotros también hemos seguido esos criterios concretos del *PCIC* para identificar los niveles de algunas locuciones pero, además, hemos aplicado criterios computables para el resto de aquellas locuciones que carecen de un nivel previo y se pueden nivelar automáticamente.

Para establecer los criterios de nivel nos hemos basado también en criterios formales, gramaticales y léxicos del multi-vocablo. Así se ha creado un módulo que otorga un nivel

(B1, B2, C1 o C2) de forma automática a todos aquellos multi-vocablos no marcados de antemano con el nivel de lengua que marca el *PCIC*. Concretamente distinguimos, por un lado, como criterios formales, aquellos elementos que nos permiten medir la locución de forma distintiva tales como guiones bajos (_), que sirven para unir vocablos creando multi-vocablos y para marcar algunas de sus especificaciones funcionales; símbolos de mayor y menor (<>) para marcar que un vocablo, dentro del multi-vocablo, es flexivo; asteriscos (*) para generalizar o simplificar características gramaticales; barras inclinadas (/) en los *PoS* para marcar doble funcionalidad gramatical de algunas locuciones.

Por otro lado, consideramos unos criterios gramático-léxicos. En general, diferenciamos rasgos distintivos en cada tipo de categoría gramatical. Computamos que la categoría gramatical tenga exponentes fijos o flexivos y valoramos la posición del elemento flexivo dentro de la lexía como veremos en el apartado 4.1.6.3.

4.1.6.2. Definición de criterios de nivelación de multi-vocablos

A continuación describimos como índices medibles la longitud del multi-vocablo, la diferencia entre el multi-vocablo y el lema, la especificidad del *PoS* por algún rasgo gramatical distintivo y, por último, la ubicación del vocablo contenido que forma lexía con un elemento función o preposición dentro de uno u otro tipo de diccionario.

Longitud del multi-vocablo

Siguiendo criterios utilizados por anteriores autores que han realizado trabajos sobre la evaluación de textos para identificar qué texto se adapta mejor a qué lector (Thomas *et al.*, 1992; Graesser *et al.*, 2004), consideramos, en nuestro caso, que la longitud de los multi-vocablos es un criterio importante y distintivo para nivelar los multi-vocablos en el glosario de locuciones. Por ello, diferenciamos fundamentalmente la longitud de aquellos multi-vocablos por el número de vocablos que conforman el multi-vocablo nominal (N), adjetival (A), determinante (D), pronominal (P), verbal (V), adverbial (R), preposicional (SPS00) y conjuntivo (C). Puesto que este criterio es muy productivo y concreto, se desarrolla detalladamente dentro del criterio general de los distintos tipos de categorías gramaticales en el apartado 4.1.6.3.

Diferencia entre el vocablo y el lema

Este criterio matiza más el criterio anterior. El multi-vocablo a veces se diferencia del lema bien porque dentro del multi-vocablo hay algún vocablo que se flexiona y entonces su estructura se presenta de forma diferente al lema o bien el multi-vocablo no coincide exactamente con su lema porque se inserta algún signo diacrítico o de puntuación en el multi-vocablo. En los dos primeros casos se puntuará el multi-vocablo por encima de B2 porque la diferencia entre el vocablo y el lema muestra un conocimiento de ciertas estructuras, reglas gramaticales o de cómo se articula una lengua. Es decir, la utilización de tildes o signos de puntuación como la coma o los dos puntos dentro de los multi-vocablos son un indicador de conocimiento de una lengua. Por ejemplo, podemos distinguir el simple adverbio "además" del conector discursivo adverbial "además," al considerar con la coma la posición inicial del marcador en la frase o inserto en ella.

Un ejemplo claro de diferencia entre el multi-vocablo y el lema son las locuciones numerales. En los casos en que el multi-vocablo se refiere a números cardinales (Z), ordinales (AO0*), fracciones y porcentajes (Zp), partitivos (Zd), distancias, pesos y medidas (Zu) o monedas (Zm), tal multi-vocablo no coincide con el lema. Por ello, otorgamos distintos niveles en función del criterio de diferencia entre el multi-vocablo y lema, y su longitud. Por ejemplo:

- Un multi-vocablo distinto del lema cuyo *PoS* es Z es propio de un nivel A1: "veintitrés".
- Un multi-vocablo con un _ guion bajo, distinto del lema, cuyo PoS es Z, es propio de un nivel A2 (Cervantes, 2006, A1-A2: 118): "ciento_dos".
- Un lema cuyo *PoS* es Zm es un A2: "dos_euros".
- Un lema cuyo *PoS* es Zp es un B1: "tercio".
- Un lema con un _ guion bajo cuyo *PoS* es Zp es un B1: "un_tercio".
- Un lema con dos _ _ o más guiones bajos cuyo *PoS* es Zp_ Partitivo es propio de un nivel B2: "un_quinto_de".

La mayoría de este tipo de multi-vocablos numéricos no se encuentra en el glosario de las locuciones que hemos nivelado sino en librerías aparte previamente indexadas por el equipo del Dr. Padró. Estas librerías, insertas dentro del programa FreeLing, se procesan automáticamente y de forma independiente del glosario de multi-vocablos. Este tipo de vocablos son susceptibles de aparecer en los textos que analizamos. No obstante, tanto estos multi-vocablos y sus lemas, que se refieren no sólo a números cardinales y ordinales (Z) sino a fraccionarios como "un tercio" (Zp), partitivos como "un cuarto de" (Zd), nomenclaturas monetarias (Zm), fechas (W) o incluso nombres propios, son todos ellos procesables por el módulo Analizator para que los nivele con los criterios establecidos por nosotros arriba.

Especificidad en el PoS

Este índice nos sirve para detallar cierta información y matizar mejor el nivel en locuciones fijas adverbiales y preposicionales, incluso algunas nominales, determinantes y adjetivas. Se procesan símbolos que añaden información a los *PoS* de cada multi-vocablo tales como el guion bajo (_) que separa entre categoría y función, y la barra (/) que especifica la función. El formato del glosario se configura con la siguiente disposición: el multi-vocablo, el lema, la categoría gramatical, el *PoS* y su especificidad, fin del proceso de FreeLing (I), y, por último, el nivel de referencia, como queda ejemplificado en los siguientes multi-vocablos indexados:

a) en_la_medida_de_lo_posible en_la_medida_de_lo_posible RG_HP/CD/CL I
b) por_ciento_de por_ciento_de SPS_Partitivo I B1
c) lote_de lote_de NCMS000_Partitivo I C1

d) una_barbaridad_de una_barbaridad_de DI0000_Partitivo I B2
e) sin_otro_particular sin_otro_particular RG_Registro-Epistolar I B2

Gracias a la posibilidad de añadir cierta información funcional, gramatical y nocional-específica a los multi-vocablos del glosario de locuciones en su *PoS*, podemos computar información morfo-sintáctico-semántica válida para conocer la estructura sintáctica en un texto y también, en futuras investigaciones, por ejemplo, para identificar qué tipo de marcadores utiliza un determinado tipo de texto.

Aunque en los ejemplos anteriores, referidos a diferentes categorías gramaticales, la mayoría de multi-vocablos ya tiene un nivel prefijado, el ejemplo a), por su especificidad en el *PoS* y por su longitud, tendría un nivel C2. Es decir, el criterio de especificidad funcionaría en casos de lemas cortos o largos sin nivelar previamente. Por ello, proponemos el siguiente criterio para evaluar la especificidad del *PoS*:

PoS con una / tendría como mínimo un nivel B2.
PoS con dos / tendría un nivel C1.
PoS con dos / y lema con 3 _ _ _o más sería un C2.

Ubicación del vocablo contenido

Los niveles del *PCIC* son los validados, pero para aquellos vocablos que están sin nivelar, recurrimos a un módulo desarrollado para esta tarea. En este caso dicho módulo va a identificar el nivel de los vocablos dentro de un diccionario determinado, discriminando diccionarios o, lo que es lo mismo, distinguiendo qué vocablo está dentro de un glosario y qué nivel le correspondería a cierto vocablo por estar en un glosario y no en otro dentro de la combinación. De esta forma vamos a identificar, mediante uno u otro diccionario, el nivel de aquellos vocablos que no están nivelados en el "Índice de nociones generales y nociones específicas del *PCIC*".

Este método de nivelación también lo vamos a aplicar a aquella locución adverbial que no esté nivelada por el *PCIC* y que conste de dos vocablos solamente, y que de los vocablos uno sea, concretamente, una preposición. Por ejemplo, un multi-vocablo de dos términos sin nivel en nuestro glosario de locuciones es "en peligro". Otorgaremos automáticamente qué nivel tiene "peligro". En este caso no sirve el "Índice de nociones generales y nociones específicas" del *PCIC* ya que no está el vocablo "peligro", aunque sí "peligroso" formando una colocación con "curva" y "cruce" (Cervantes, 2006, B1-B2: 491). Como en el glosario del Dr. Kincaid sí está este término, "peligro" tendría como mínimo un nivel B1 o, por su longitud, si lleva una preposición, un máximo de B2. Esto, según el criterio del *PCIC* de que las locuciones adverbiales son de uso incipientes en el nivel B1 y se sistematizan en el nivel B2 (Cervantes, 2006, B1-B2: 207), nivelaríamos este vocablo con nivel B1 como mínimo pero, al aparecer combinado como multi-vocablo, ascendería a un nivel B2.

Sirva para ilustrar nuestro argumento, el contra-ejemplo del multi-vocablo "en auge". Éste se registra con un nivel C1 en el inventario de "Nociones específicas (C1-C2)" (Cervantes, 2006, C1-C2: 515). Por el contrario, el término "auge" no está listado en el "Índice de nociones generales y nociones específicas" del *PCIC*. Comprobamos en los otros glosarios que el término "auge" no está ni en el del Dr. Kincaid ni en FreeLing 1.5, pero sí

en FreeLing 2.1 y en esWordnet, por tanto no puede ser menos de un nivel C1 que, al aparecer combinado, le otorgaríamos manual y automáticamente un nivel C2.

El hecho de que estos multi-vocablos tengan sólo un guion bajo y se nivelen con el criterio de longitud, el multi-vocablo no alcanza un nivel elevado pero sí, como mínimo, puede obtener el nivel que se le otorgue en la combinación de diccionarios. En el caso de las locuciones, este criterio de pertenencia a un glosario y, por tanto, de un nivel asociado a la combinación de glosarios lo vamos a utilizar sólo para aquel multi-vocablo adverbial que únicamente tenga un guion bajo. Principalmente este criterio de nivelación es aplicable para aquellos casos en los que una locución adverbial de dos elementos no tenga un nivel previo dado por el *PCIC*.

Aplicando este criterio de refinamiento de calificación, la recalificación se ha hecho con 964 locuciones que tenían un nivel de A1, A2 y B2. Tras procesarlas, se han elevado de nivel 261 locuciones adverbiales. Sin embargo, este afinamiento no ha sido muy productivo porque la segunda parte de la mayoría de locuciones adverbiales contienen vocablos en plural, compuestos, diminutivos o vocablos tan poco frecuentes o inexistentes en nuestros glosarios como "batiburrillo". Por lo tanto, todavía un cierto número de locuciones adverbiales queda infra-nivelada.

En otro tipo de locuciones, como las verbales, podríamos proceder de manera similar en futuros ajustes. Por ejemplo, consideremos una locución registrada y nivelada con el *PCIC* como "poner(se) rojo". Tanto "poner" como "rojo" son dos vocablos básicos que juntos, si aplicásemos un futuro criterio de combinación de términos, alcanzarían un nivel máximo de B1 ya que ambos son vocablos de nivel A2 y A1 respectivamente. Sin embargo, el *PCIC* otorga a "me pongo rojo" un nivel B2 (Cervantes, 2006, B1-B: 233) y a "ponerse rojo de ira/de rabia" un nivel C2 (Cervantes, 2006, C1-C2: 468). La expresión "me pongo rojo" es equivalente a "avergonzarse", vocablo/concepto que en el diccionario del *PCIC* tiene también un nivel B2 (Cervantes, 2006, B1-B2: 452). Con este ejemplo, comprobamos que tanto el multi-vocablo como la noción son ya propios del nivel B2, lo que confirma el criterio expresado más arriba de la sistematización de ciertas locuciones en el nivel B2 (Cervantes, 2006, B1-B2: 207) y de su longitud.

En conclusión, al comprobar en ambos casos que la combinación de dos vocablos básicos, uno contenido y otro funcional, se convierten en una lexía de nivel intermedio B2, observamos que la mayoría de locuciones verbales de este tipo se compone de un verbo comodín, pronominal o no, y que el criterio de ubicación de los vocablos en ciertos diccionarios junto con el criterio de combinación de vocablos para otorgar niveles a los vocablos coincide en muchos casos con el nivel que otorga el *PCIC*.

Aunque la combinación de vocablos es una aproximación de análisis léxico-sintáctico que no hemos puesto en practica, sí que proponemos computar multi-vocablos cuando éstos aparezcan en los textos con verbos comodín. Tradicionalmente, se ha llamado verbos comodín a "ser, haber, hacer, estar, decir, poner, tener, ver" (Paredes Chavarría, 2006, 161 y ss.) porque van acompañados de otra categoría gramatical modificando el significado de ambos. Otros autores los denominan polisémicos e implementan la lista incluyendo "dar" (Sánchez Lobato, 2006, 204-206). Dentro de este criterio de combinación, no menos interesante es poder computar verbos que combinen sistemáticamente con determinadas preposiciones, es decir, los verbos preposicionales, apartado relevante en la enseñanza de

E/LE.

4.1.6.3. Locuciones y multi-vocablos nivelados

Hemos organizado, de forma general, los multi-vocablos por categorías en función de si son estructuras fijas o flexivas. Además, se aplican algunos de los índices expuestos más arriba. Esto es, vamos a poder marcar automáticamente locuciones que varían desde un nivel B1, B2 a un nivel C1, C2, dependiendo de la longitud del lema del multi-vocablo, del *PoS*, de la flexividad o no del multi-vocablo, y del léxico concreto que conforma la lexía.

Categorías gramaticales con exponentes fijos

Estas estructuras son fórmulas fijas o *chunks* que se realizan como los nombres en *singularia tantum* o *pluralia tantum* (Cervantes, 2006, C1, C2: 124), es decir, sólo en singular o en plural, respectivamente. De una parte, se da un valor inicial básico a aquellas locuciones o "exponentes fijos" (Cervantes, 2006, B1-B2: 177) que se aprenden o reproducen tal cual. De otra parte, según la longitud del lema o la especificidad del *PoS*, se elevará el nivel del multi-vocablo. Es decir, dependiendo de la longitud del multi-vocablo, más o menos extenso gracias al número de guiones bajos en el lema o barras inclinadas en el *PoS*, se puede clasificar con un menor o mayor nivel. De momento, los criterios establecidos para los multi-vocablos, hasta ahora, se han aplicado sólo al fichero de locuciones y, por tanto, a los textos analizados en este trabajo.

Lexías nominales fijas

Estas locuciones nominales se expresan, siguiendo la nomenclatura del Dr. Padró, con la etiqueta de nombre (N), seguida del tipo común (NC) o propio (NP), del género masculino (NCM), femenino (NCF) o neutro (NCC), y del número tanto singular (NCMS, NCFS, NCCS) como plural (NCMP, NCFP, NCCP). Por ejemplo, las siguientes estructuras se presentan siempre en plural o singular:

[los] juegos_olímpicos juegos_olímpicos NCMP000
[las] horas_de_oficina horas_de_oficina NCFP000
[el, la] policía_judicial policía_judicial NCCS000

Tanto para locuciones equivalentes a nombres comunes (NC) como propios (NP), partimos de un nivel B1 hasta un nivel C2. Ambos tipos de locuciones están indexadas bien dentro del glosario de las locuciones propiamente dicho o dentro de otras librerías de FreeLing menos visibles pero igualmente computables.

NCM*, NCF*, NCC* B1 si tiene un _ en el lema.
NCM*, NCF*, NCC* B2 si tiene un _ en el *PoS*.
NCM*, NCF*, NCC* B2 si tiene 2 _ _ en el lema.
NCM*, NCF*, NCC* C1 si tiene 3 _ _ _ en el lema.
NCM*, NCF*, NCC* C2 si más de 3 _ _ _ _ en el lema.

Dentro de los nombres propios (NP), FreeLing incluye algunas lexías dentro de ciertas clases o categorías como organizaciones e instituciones (NP00O00), personas (NP0SP00), lugares (NP0LG00) o crea un cajón de sastre que el Dr. Padró denomina varios (NP00V00). Indistintamente de la categoría o del tipo de locución nominal, el nivel de la locución estaría en función de su longitud y especificación del *PoS*.

NP* B1 si tiene un _ en el lema.
NP* B2 si tiene un _ en el lema y un *PoS* con NP0SP00, NP0LG00, NP00O00 o NP00V00.
NP* B2 si tiene 2 _ _ en el lema.
NP* C1 si tiene 2 _ _ en el lema y un *PoS* con NP0SP00, NP0LG00, NP00O00 o NP00V00.
NP* C1 si tiene 3 _ _ _ en el lema.
NP* C2 si tiene más de 3 _ _ _ _ en el lema.

Por ejemplo, la fórmula Príncipe_Felipe es un nombre propio con su epíteto, pertenecientes ambos a la categoría de persona (NP00SP0) tendría un nivel B2, igualmente la lexía nominal común como comunidad_económica_europea NP00O00 tendría un nivel B2.

Incluso, según la estructura y significado, FreeLing permite indexar algunas locuciones nominales con grados de comparación semejantes a los que identifican a los adjetivos. A estas locuciones nominales equivalentes a adjetivos aumentativos (NCCN00A) se les confiere un nivel predeterminado porque son también fórmulas fijas con una longitud determinada y una característica específica. Por ejemplo, el multi-vocablo "fuera_de_serie", con *PoS* NCCN00A, sería de un nivel C2.

Lexías determinantes fijas

Los determinantes se nivelan según el tipo al que pertenecen y su longitud. Es lógico que el aprendiz en un nivel inicial (A2) maneje con soltura el determinante artículo (DA), demostrativo (DD) y posesivo (DP). Sin embargo, combinar determinantes entre sí es un indicador de un nivel superior (Cervantes, 2006, B1-B2: 50). Por ello, proponemos los siguientes niveles según el tipo y estructura del determinante:

DI* B1.
DI* B2 si tiene 2_ _ en el lema.
DD* B2 si tiene un _ en el *PoS*.
DI* B2 si tiene un_ en el *PoS*.

Lexías pronominales fijas

En cuanto a las locuciones pronominales, se sigue el criterio de las locuciones determinantes con exponente fijo. Nivelamos según el tipo de locución pronominal indefinida (PI), relativa (PR), interrogativa (PT) o posesiva (PX), y según su longitud y especificidad en el *PoS*:

PI*, PR*, PT B2.
PI*, PR* C1 si tiene 2 _ _ o más en el lema.
PI*, PR*, PT * C1 si tiene un _ en el *PoS*.
PX* C1.

Lexías adjetivas fijas

Al igual que las categorías nominales, las locuciones adjetivas pueden ser fijas o flexivas, o tener más o menos extensos sus lemas. Por ello, según su estructura y especificidad semántica, les damos un nivel u otro:

AQ0CN0_Partitivo B1 si tiene un_en el *PoS*.
AQ0C* B2 si tiene 2_ _ en el lema.
AQ_P/A C1 si se etiqueta en el *PoS* tanto valor predicativo (P) como atributivo (A).
AQ_Preposicional B2 por tener un _ en el *PoS*.
AQ_Superlativo B2 por tener un _ en el *PoS*.
AQACN0 B2 si tiene un _ en el lema. AQACN0 etiqueta a la locución adjetiva aumentativa.
AQACN0 C1 si tiene 2_ _ en el lema.
AQDCN0 B2. AQDCN0 etiqueta a la locución adjetiva diminutiva.
AQSCN C2. AQSCN etiqueta a la locución adjetiva superlativa.
AQS* C2 con un _en el *PoS*. AQS* etiqueta a la locución adjetiva superlativa o relativa.

Lexías conjuntivas fijas

Independientemente de la complejidad sintáctica que supone optar por conjunciones, conectores, marcadores del discurso o locuciones preposicionales asociadas a un modo verbal o a una *consecutio temporum*, nosotros otorgamos un nivel B2 como mínimo a todas las locuciones conjuntivas y preposicionales. Como en el caso de las locuciones adverbiales, las locuciones conjuntivas coordinantes (CC) y subordinantes (CS) se presentan para su aprendizaje en un nivel B1 (Cervantes, 2006, B1-B2: 80-81), aunque se sistematizan algunas en el nivel B2 (Cervantes, 2006, 98 y ss.) y se fijan en el nivel C1 (Cervantes, 2006, C1-C2: 97 y ss.). Por ello, a aquellas locuciones a las que no hemos indexado un nivel según el *PCIC*, aplicaremos una vez más el criterio de longitud y especificidad en el *PoS*, ya que consideramos que ambos criterios se muestran como indicadores de conocimiento y habilidad de la expresión escrita.

- Conjunción Coordinante (CC)

  Asumiendo que la conjunción como vocablo simple (CC) tienen un valor propio de un nivel B1, consideramos que todas aquellas conjunciones coordinantes como multi-vocablos (CC*) recogidas en el glosario de locuciones son propias de un nivel mínimo de B2, como expresamos a continuación:

  CC* B2.
  CC* C1 con 2_ _ o más en el lema.

CC* C2 con una / o más en el *PoS*.

- Conjunción subordinante (CS)

  Con respecto a la conjunción subordinante de un sólo vocablo (CS) del tipo *cuando, porque, como, que,* etc. no otorgamos explícitamente el nivel B1, sino que sólo las nivelamos si son multi-vocablos (CS*) y partimos de un nivel B2 como en el caso de las locuciones prepositivas, adverbiales y conjuntivas coordinantes:

  CS* B2.
  CS* C1 con 2_ _ o más en el lema.
  CS* C2 con una / o más en el *PoS*.

Lexías preposicionales fijas (SPS*)

En el supuesto de que una preposición simple (SPS00) tuviera un valor A1, A2 o B1, entonces las preposiciones agrupadas (Sánchez Lobato, 2006, 185) y las locuciones preposicionales (SPS_*), al igual que las locuciones adverbiales, serían susceptibles de tener un nivel mínimo de B2. Por ello, de nuevo, en función de la longitud del multi-vocablo o de la especificidad del *PoS*, otorgaríamos estos niveles:

SPS_* B2.
SPS_* C1 con 2_ _ o más en el lema.
SPS* C2 con una / o más en el *PoS*.

Lexías adverbiales (RG o RG*)

Las locuciones adverbiales pueden ser positivas (RG*) o negativas (RN*). Ambos tipos de locuciones, a excepción de aquellas expresiones latinas que en su mayoría aparecen en el glosario de locuciones indexadas como RG y de aquellas que previamente se han marcado con el nivel que les otorga el *PCIC*, se regirán fundamentalmente por los índices de longitud, especificidad del *PoS* y flexividad como sigue:

RG* B2 si el lema es simple o al menos contiene un _ en el lema.
RG* B2 si contiene un _ en el *PoS*.
RG* C1 si contiene 2 _ _ en el lema.
RG* C2 si contiene una / en el *PoS*.
RG* C2 si contiene más de 3 _ _ _ en el lema.
RG C2 porque suelen ser locuciones latinas o muy poco frecuentes en textos de E/LE.

Como apuntábamos más arriba, para aquellas locuciones con uno o más vocablos sin contenido, es decir, preposiciones, se aplicará el nivel procesando de forma automática al vocablo más extenso y, por tanto, con contenido, en función del "Índice" del *PCIC* o de la combinación de diccionarios en la que se encuentre dicho vocablo. Por ejemplo, la locución adverbial "de_un_voleo", será "voleo" el vocablo que determine el nivel para precisar entre los niveles B2, C1, C2. Proponemos el siguiente nivel para las locuciones adverbiales positivas:

RG* con un _ o 2 _ _ en el lema entre cuyos guiones hallamos preposiciones o determinantes:
- si la palabra más extensa está en el Dr. Kincaid la locución será un B2.
- si la palabra más extensa está en el FreeLing 0.5 la locución será un C1.
- si la palabra más extensa está en el FreeLing 2.0 la locución será un C2.

En cuanto a las formas negativas, nivelamos en función de la especificidad y la longitud del multi-vocablo:

RN B2.
RN* C1.
RN* C2 con más de 2 _ _ en el lema.

También en este apartado de locuciones adverbiales flexivas, positivas o negativas, vamos a nivelar aquellos multi-vocablos adverbiales semi-fijos o semi-flexivos, ya que requieren una nivelación concreta bien por el criterio de diferencia entre el multi-vocablo y el lema o bien por la naturaleza del multi-vocablo, cuya estructura tiene uno o dos vocablos susceptibles de flexión.

Por ejemplo, en algunos multi-vocablos es distintivo el rasgo de número singular o plural. Por ejemplo, la locución adverbial "por_<pareja>" indexa en su mismo lema el multi-vocablo "por_pareja" o "por parejas" dependiendo del vocablo o del co-texto en que se realice la locución. Es decir, la expresión adverbial difiere, por ejemplo, en "dar una entrada por pareja" que "organizarse por parejas". Mientras que otras veces la distinción de número es indiferente como es el caso de "sin cambio" o "sin cambios" en "quedarse sin cambios" o "estar sin cambio". Por ello, consideramos que, si el multi-vocablo difiere del lema, la flexividad del multi-vocablo es un indicador distintivo de la versatilidad de algunas locuciones adverbiales y debería medirse. Es decir, aquellas locuciones adverbiales con algún elemento flexivo, verán incrementado el nivel a manera de *bonus* a favor de un nivel C1, ya que seguimos el criterio del *PCIC* de que con las locuciones adverbiales se parte de un nivel B2. Así que formulamos que un RG* es C1 si contiene un elemento flexivo o más en el multi-vocablo.

Lexías numéricas o cuantificadoras, fijas o semi-fijas

Nos referimos a aquellas lexías que, aunque no están indexadas por nosotros en el glosario de locuciones, son también multi-vocablos significativos y computables. Por este motivo, hemos desarrollado otra función en Nivelator para que nivele aquellos multi-vocablos que no están registrados en el glosario de locuciones pero que registra y procesa el programa de FreeLing en su análisis morfológico. Entre ellos, están los multi-vocablos relacionados con cifras y cantidades, tales como los partitivos (Zd), los pesos, medidas y distancias (Zu), las fracciones y porcentajes (Zp), las horas y las fechas (W) e incluso los adjetivos numerales ordinales (AO). De estos, ponemos como ejemplo de multi-vocablos los numerales partitivos que se nivelan automáticamente atendiendo a los siguientes criterios de longitud, diferencia y especificidad del lema. Se exponen a continuación los criterios de nivel con algunos ejemplos de partitivos:

Zp B1 multi-vocablo con un _ guion, multi-vocablo distinto del lema.
Zp B2 multi-vocablo con dos _ _ guiones, multi-vocablo distinto del lema (Cervantes, 2006, B1-B2: 59).
Zp C1 multi-vocablo con 3 _ _ _ guiones, multi-vocablo distinto del lema.
Zp C2 multi-vocablo con más de 3 _ _ _ guiones, distinto del lema (Cervantes, 2006, C1-C2: 60).

También cabe destacar los numerales ordinales, cuyos niveles describimos en función del criterio de longitud del multi-vocablo y la diferencia gráfica del ordinal:

AO* A2 para ordinales $<10^{o}$, como vocablo (Cervantes, 2006, A1-A2: 118).
AO* B2 para ordinales $>10^{o}$ y $<20^{o}$, como vocablo (Cervantes, 2006, B1-B2:59).
AO* C2 para ordinales $>20^{o}$, como vocablo (Cervantes, 2006, C1-C2: 60).

Por último, las fechas y horas se simbolizan en FreeLing con la letra W en el *PoS*. Esta etiqueta "W" identifica el *PoS* tanto de cifras como de vocablos y multi-vocablos relacionados con conceptos de tiempo. Para nivelar los multi-vocablos que procesa FreeLing etiquetándolos con este *PoS*, hemos aplicado a algunos multi-vocablos el nivel y criterio del inventario de "Nociones Generales" del PCIC correspondiente al apartado "4. Nociones temporales" y sus subapartados "4.1. Referencias generales" y "4.2. Localización en el tiempo" de los diferentes niveles. Muchos multi-vocablos etiquetados con el *PoS* "W" se corresponden con niveles iniciales (Cervantes, 2006, A1-A2: 319 y ss.), mientras que otra mayoría de multi-vocablos referidos a conceptos de tiempo se identifican con los niveles intermedios (Cervantes, 2006, B1-B2: 414 y ss.) y superiores (Cervantes, 2006, C1-C2: 417 y ss.). Los parámetros aplicados se identificarían con los siguientes niveles:

W A2 si tiene cifras y letras en el multi-vocablo.
W B1 si tiene menos o 2 _ _ en el multi-vocablo.
W B2 si tiene más de 2 _ _ en el multi-vocablo.

Categorías gramaticales con exponentes flexivos

Respecto a la característica flexiva, distinguimos la nominal (género y número del determinante, sustantivo, pronombre y adjetivo) y la verbal (persona y número), además de su posición dentro del multi-vocablo. Ambos características se marcan de forma distintiva. Por un lado, la flexividad se marca encuadrando cada término entre los símbolos de mayor y menor, <>. Por otro, la primera posición o segunda posición del término flexivo se marca con el símbolo dólar, un numeral y dos puntos, "$1": o "$2:". Así, distinguimos aquellos multi-vocablos concordables o conjugables, como exponemos a continuación.

Lexías nominales flexivas

En esta categoría se incluyen, entre otros, algunos sustantivos que son susceptibles de aparecer en singular o en plural, acompañados de nombres aposición o adjetivos calificativos o atributivos. Es más, consideramos que la posición del sustantivo en la lexía es otro

indicador de madurez sintáctica ya que la posición habitual de un sustantivo en castellano es la de anteponerse al adjetivo. La anteposición de un adjetivo relega al sustantivo a segunda posición y, generalmente, dota de otro matiz o significancia al sustantivo. Por ello, consideramos que la concordancia, supeditada a la flexividad y a la posición primera ($1:) o segunda ($2:N) del elemento flexivo, son rasgos distintivos para otorgar un nivel.

$1:NC B2 el nombre del multi-vocablo es flexivo en primera posición.
$2:NC C1 el nombre del multi-vocablo es flexivo en segunda posición.

Lexías determinantes flexivas

También podemos encontrar un multi-vocablo determinante flexivo en posición inicial al combinarse con otro determinante o con la preposición "de":

$1:DI B2 determinante indefinido.
$1:DD C1 determinante demostrativo.

Igualmente, también se hallan determinantes flexivos en segunda posición combinados con otro determinante o con la preposición "de":
$2:DI C1 determinante indefinido.

Lexías pronominales flexivas

Semejante recurrencia se da con algunos multi-vocablos pronominales cuyos exponentes flexivos en primera posición se combinan con otro pronombre o con la preposición "de":

$1:PP B2.
$1:PI B2.

Se procesan también algunos pronombres con exponentes flexivos computables en segunda posición que aparecen generalmente precedidos por la negación "no" o seguidos de otro pronombre o preposición como no_<todo>:

$2:PR B2.
$2:PI C1.
$2:PT C1.

Lexías adjetivas flexivas

Al igual que las anteriores, los exponentes flexivos dentro de una locución adjetiva suelen ser dos vocablos. En el caso de los adjetivos, la categoría adjetiva se encuentra, unas veces, en primera posición ($1):

$1:AQ B2 adjetivos seguidos de una preposición.

Otras veces, los que se encuentran en segunda posición ($2) son colocaciones donde los adjetivos van precedidos de un determinante y / o un nombre seguido de una locución adjetival como <el>_<bueno>_de.

$2:AQ B2

Multi-vocablo conjugable
Dentro de este apartado, se nivelan algunas locuciones verbales y algún verbo preposicional.

Lexías verbales
Se nivelarán en función del índice general de longitud y flexividad. Además, dependiendo de la posición en qué aparece el verbo, se especifica más el nivel.

Con el verbo copulativo (VS) o predicativo (VM) en primera posición de la locución ($1:V):

$1:VS* B1 si es el verbo "ser" con la preposición "de" o "para".
$1:VS* B2 si tiene un _ en el lema y otro elemento flexivo (<>) en el vocablo.
$1:VS* B2 si tiene 2 _ _ en el lema.
$1:VM* B1 si es un verbo comodín o seguido de una preposición concreta (parámetro que no se aplica en este estudio).
$1:VM* B2 si tiene 2 _ _ en el lema.
$1:VM* C1 si tiene 3 o más _ _ _ en el lema.

Generalmente aparece el adverbio negativo en primera posición seguido de un verbo copulativo ($2:VS) o predicativo ($2:VM) en segunda posición ($2) de la locución:

$2:VS* B2
$2:VS* B2 si tiene 2 _ _ en el lema.
$2:VS* B2 si tiene un _ en el lema y otro elemento flexivo (<>) en el vocablo.
$2:VM* B2
$2:VM* C1 si tiene 2 _ _ en el lema.
$2:VM* C1 si tiene 3 o más _ _ _ en el lema.

Efectivamente, en muchas ocasiones algunas locuciones en los textos pueden aparecer desmembradas al introducirse un adverbio intensificador de un adjetivo o de un verbo. No obstante, apenas se han indexado locuciones verbales ya que, por un lado, son muy abundantes y, por otro, porque son fácilmente desmembrables al poderse insertar adverbios y pronombres personales entre el verbo y el resto de los vocablos que conforman la locución. Como ejemplo, recogemos algunas locuciones que pueden fácilmente desmembrarse:
(no) tener_ni_idea >"(no) tener ni *pajorera /la más mínima /la menor* idea".
poner_de_patitas_en_la_calle >"poniéndo*(nos)* de patitas en la calle".

Sin duda, se podría programar que aquellas locuciones verbales susceptibles de desmembrarse se computasen a caballo entre el nivel léxico y sintáctico. Esto es, si entre las locuciones indexadas se insertara otro vocablo, se podría seguir procesando el multivocablo. Para ello, podríamos establecer unas reglas mínimas productivas para las locuciones verbales ($1:V*, $2:V*) concretamente. Por ejemplo, podría procesarse una locución verbal desmembrada aunque inmediatamente después del verbo se inserte un adverbio (RG), una locución adverbial (RG*) o un pronombre personal (PP*). Ponemos un ejemplo de frecuente desmembramiento con la locución verbal "echar de menos":
Inserción de un adverbio (RG): echar mucho de menos.

Inserción de una locución adverbial (RG*) : echar de vez en cuando de menos.
Inserción de un pronombre enclítico (PP*): echarte de menos.

Claramente, esta posibilidad nos presentaría una locución más compleja y dinámica y sería un buen indicador de nivel de conocimiento de una segunda lengua (L2). Sin duda, la implementación de estas peculiaridades del castellano contribuirían a detectar y precisar mejor los niveles de lengua más elevados.

4.2. Definición de índices léxicos

- Índice de mayor frecuencia de lemas.
  Se consideran todas las categorías: nombres (N*), adjetivos calificativos (AQ*) y numerales (A0*), verbos predicativos (VM*) y auxiliares (VA*, VS*), determinantes (D*), pronombres (P*), adverbios (RG*), preposiciones (SPS*), numerales (Z) y fechas (W).

- Índice de frecuencia de locuciones.
  Consideramos dentro de este indicador todos los vocablos que conforman las locuciones independientemente de su categoría gramatical.

- Índice de lemas diferentes.
  Este indicador nos informa de la riqueza de vocabulario del texto. Un texto con varios lemas repetidos muestra un nivel básico o la especificidad del léxico en un texto. Este criterio es relativamente objetivo ya que sólo es válido cuando medimos textos extensos y temática semejante (McCarthy, 2005).

- Índice de longitud de vocablos.
  Este índice, también denominado de *Flesch-Kincaid Grade Level*, sirve para medir la longitud y el uso de palabras más o menos largas en un texto (McNamara *et al.*, 2006). Dado que en la reproducción de palabras más largas se corre el riesgo de escribirlas erróneamente, el hecho de poder computarlas significa que están correctamente escritas. Consideramos que la longitud de un vocablo esperado en un nivel B2 tendría una mediana de 5 letras. Si el valor es superior a 5 letras, el nivel se decantaría por el nivel C1.

- Índice de categorías repetidas.
  Nos aportaría un valor acerca de las distintas categorías gramaticales repetidas.

- Índice de vocablos con tilde.
  Cuando se procesan las palabras para etiquetarlas morfológicamente, FreeLing sólo reconoce las palabras correctamente escritas por tanto, con este indicador, sabremos cuántas palabras con tilde ha empleado el aprendiz en su texto correctamente puesto que las erróneas no se computan, y sin tilde se registran erróneamente.

- Índice de palabras no computadas.
  Nos indica qué palabras no se han computado bien por no estar bien escritas por el aprendiz o por ser vocablos que no se registran en ninguno de los glosarios de referencia en este estudio.

- Índice de calificación de lemas.
  Según la complejidad de un texto, con los criterios del *PCIC* se consigue calificar entre un 40 %-45 % de los lemas, pero se puede llegar a calificar en torno a un 90 % de lemas con la combinación de diccionarios y el glosario de locuciones. Este índice nos va a permitir decidir qué nivel tiene un texto.

- Índice de Kincaid.
  Nos basamos en el criterio de intervalos de nivel del glosario del Dr. Kincaid para confirmar el nivel del léxico. Si el léxico del texto está en un casi 100 % en el glosario del Dr. Kincaid, nos hallamos ante un texto de nivel B1. Si los lemas del texto están entre un 90-95 %, podemos diagnosticar que el texto tiende a un nivel B2, y cuando el porcentaje esté entre un 95-70 %, el texto ya apunta a un léxico de nivel C1. Estos son porcentajes aproximados y semejantes a los que propone Oxford, el diagnosticador del nivel (Oxford, 2010).

- Índice de la ley de Zipf.
  Este índice nos propociona la relación que existe entre los lemas más frecuentes y su rango de aparición en un texto. Dado que en este índice se fundamenta nuestra nivelación del léxico para decidir el nivel léxico de un texto, lo exponemos detalladamente en el apartado 4.2.1.

4.2.1. Ley de Zipf

En estudios lingüísticos, la ley de Zipf se utiliza para mostrar el orden de las palabras en función de su frecuencia en un *corpus* de cualquier lengua. Zipf la utilizó para el estudio del vocabulario de *Ulises* de James Joyce. Sus discípulos (G. Altmann, K. H. Best, L. Hrebícek, R. Köhler, V. Kromer, O. Rottmann, A. Schulz, G. Wimmer y A. Ziegler) le consideran uno de los fundadores de la Quantitative Linguistics y lo han materializado con la creación de la revista Glottometrics. Su propuesta es aplicable actualmente a estudios sociológicos (Adamic y Huberman, 2002), de Lingüística Computacional (Li, 1992; Ha *et al.*, 2002; Baayen, 2008) o de Semántica Computacional (Griffiths y Steyvers, 2002). Una consecuencia de la ley de Zipf es la denominada Ley de Heap (1978), utilizada para el estudio de diversidad léxica, porque establece que cuanto más extenso va siendo un texto menor es el número de apariciones de nuevos vocablos (McCarthy, 2005, 3). La ley de Zipf es una ley definida en 1932 por el lingüista George Kingsley Zipf. La aplicó concretamente en lingüística para demostrar que mientras sólo unas pocas palabras son utilizadas con mucha frecuencia, otras muchas apenas son usadas (Zipf, 1932), aunque también la hizo extensiva a la Historia para explicar fenómenos sociológicos (Zipf, 1949). La ley de Zipf se expresa mediante la fórmula:

$$P_r = \frac{C}{r^{\alpha}}. \tag{4.1}$$

Zipf propuso la anterior relación matemática por la que la frecuencia de una palabra P_r es inversamente proporcional a su rango r, elevado a un cierto exponente α, que es aproximadamente igual a 1, siendo C la frecuencia de la palabra en el rango $r = 1$. Esto significa que, suponiendo que $\alpha = 1$, los vocablos se ordenan en función de su frecuencia

de manera que el segundo vocablo aparece la mitad de veces que el primero, el tercero una tercera parte de veces que el primero, y así sucesivamente.

La representación gráfica habitual para la ley de Zipf suele ser una curva o línea de puntos (equivalentes a lemas) que presenta el logaritmo de la fórmula de Zipf:

$$\log_{10} P_r = \log_{10} C - \alpha \log_{10} r. \tag{4.2}$$

En nuestro caso, la gráfica 4.7 muestra la distribución logarítmica descendente de 5.207 lemas del glosario Wiktionary al filtrarlo por el "Índice de nociones generales y nociones específicas" del *PCIC*. Los puntos negros son aquellos lemas que no están en el "Índice de nociones generales y nociones específicas" del *PCIC* y que se han desplazado en la gráfica para una mejor visualización. Los distintos colores significan que cada punto se identifica con un lema y un nivel. Como se observa en la figura 4.7, los parámetros que se representan son el rango y la frecuencia de los lemas pero utilizando un parámetro más: su nivel. Esta gráfica prueba no sólo que una palabra menos frecuente tiene un rango mayor sino que, además, un lema menos utilizado corresponderá más probablemente a un nivel más alto.

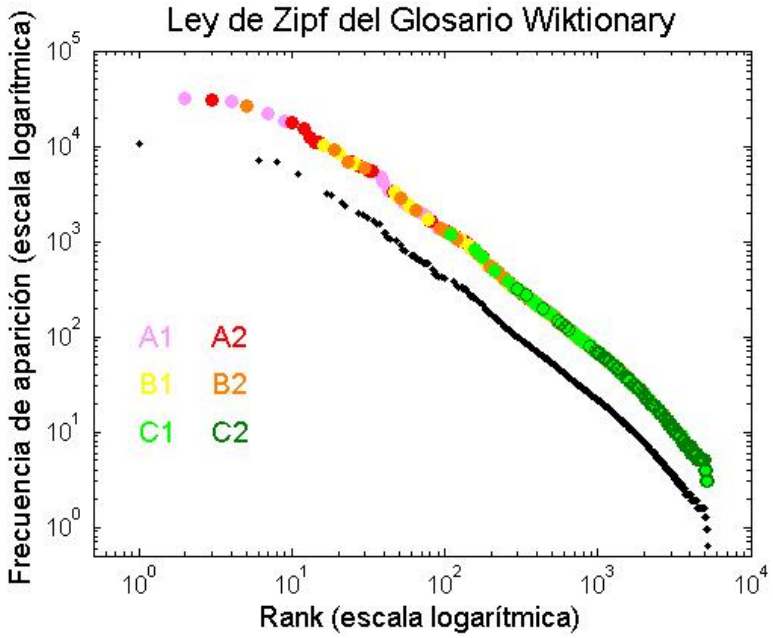


Figura 4.7: Representación log-log de la ley de Zipf para las palabras incluidas en el glosario Wiktionary. Los colores de los puntos indican el nivel de esos vocablos identificados y calificados por el *PCIC*. Los vocablos no calificados se han representado como puntos negros y se han desplazado verticalmente de su lugar en la gráfica para mejorar la visibilidad de la figura.

Después de ajustar la relación porcentaje-vocablos y los lemas repetidos del Wiktionary, se han reagrupado lemas y frecuencias para probar un listado de lemas con sus frecuencias en español. Se decide comprobar una vez más esta ley con un listado más fiable. Se recurre al listado de 97.910 lemas creado por el Dr. Padró y disponible en su página web en el apartado de *Natural Language Processing (NLP) >Some useful things >Several frequency counts of word forms, lemmas, and PoS from a 5.5 million words Spanish corpus of unrestricted text* (Padró, 2011a). No obstante, adaptamos este listado. Se reduce

sus lemas de 97.910 a 47.825, tras eliminar los lemas de frecuencia 1. Una vez procesados sus lemas y frecuencias, observamos en la figura 4.8 que el resultado correspondiente a este listado, aplicando la ley de Zipf y el glosario del *PCIC*, es similar al del Wiktionay. Esto es, se distribuyen los rangos por niveles en función de su frecuencia, ajustándose a una distribución lineal (en representación logarítmica).

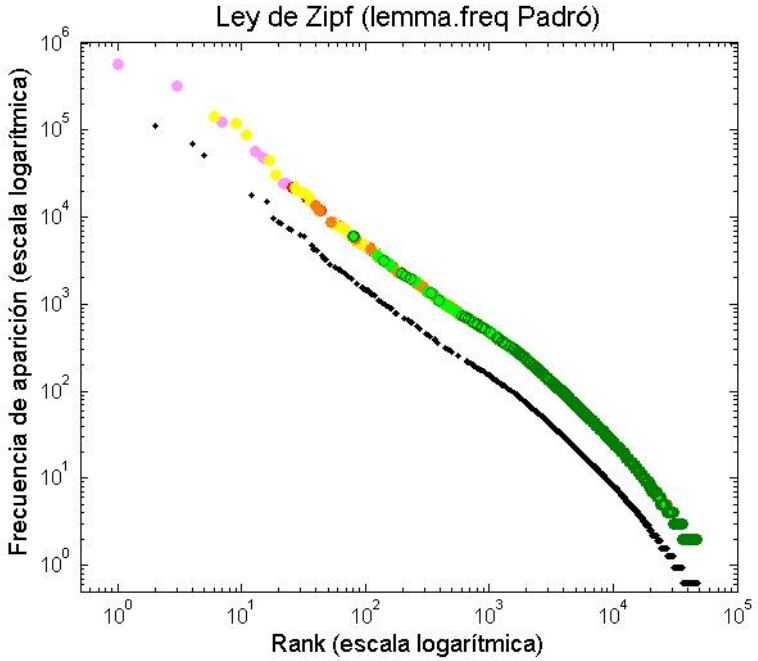


Figura 4.8: Ley de Zipf aplicada al listado de frecuencias del Dr. Padró.

Esta ley de Zipf, como apuntábamos en el apartado 4.2 donde definimos los índices léxicos, nos va a ayudar a definir la nivelación léxica de un texto. Se aplica esta ley de Zipf en el método de nivelación que describimos en el apartado 4.4.2.

4.3. Herramienta de análisis léxico: Lexicator

Lexicator es una herramienta automática que hemos diseñado durante la elaboración de esta tesis para identificar los lemas y los niveles de los lemas que aparecen en los textos.

Esta herramienta trabaja sobre textos procesados previamente por FreeLing. Por tanto, funciona con lemas y categorías gramaticales etiquetadas. En un primer paso, Lexicator identifica lemas y los compara con lemas que ya están listados en los diccionarios. ¿Qué interés tiene identificar lemas que estén en uno u otro diccionario? Saber que todos los lemas del texto existen. La identificación de un lema en un texto y su correspondencia con un lema en un diccionario es una garantía de la validez formal del vocablo.

En un segundo paso, Lexicator nivela los lemas que coinciden con los lemas nivelados del "Índice" del *PCIC*. Aquellos lemas que quedan sin nivelar por el *PCIC* (entre un 50 %-60 %) se nivelan ahora, según la pertenencia a una combinación de diccionarios, mediante el módulo Analizator, con una función desarrollada para este proceso de nivelación. Mó-

dulo que, después de tener nivelado cierto porcentaje de lemas con el "Índice" del *PCIC*, vuelve a nivelar después de identificar de nuevo más lemas. Generalmente estos nuevos lemas son adverbios acabados en "–mente" o, incluso, los participios que aparecen en los glosarios.

Detallamos el caso de los adverbios acabados en "–mente". Como la mayoría de diccionarios, nuestros glosarios apenas registran adverbios de este tipo ya que regularmente no son entradas en los diccionarios. Por ese motivo, y para identificar el máximo de vocablos de un texto, se ha decidido procesar y evaluar los adverbios acabados en "-mente". Efectivamente, estos vocablos no se han nivelado durante el segundo paso de Lexicator así que, en un tercer paso, Lexicator disocia el sufijo "–mente" y "-amente" del vocablo y busca un adjetivo o término del que derive dicho adverbio sufijado con "-mente" o "-amente". Una vez identificado con un lema, de nuevo, en un segundo paso, dará a dicho adverbio acabado en "-mente" el nivel que tenga el lema que le corresponde. Una vez más, para nivelar en función de la combinación de diccionarios, se busca un lema determinado primero en el "Índice" del *PCIC* y, si no está, entonces se recurre de nuevo a la combinación de glosarios del Dr. Kincaid, Wiktionary, FreeLing 1.5, FreeLing 2.1 y esWordnet, que servirán como referentes para dar un nivel de lengua a aquellos vocablos de los que derivan los adverbios acabados en "-mente".

Para el caso de los participios, en un principio Lexicator no los reconocía como lemas propiamente sino que, al vincularlos el analizador de FreeLing a sus respectivos lemas verbales, estos participios, sí nivelados en el *PCIC*, quedaban sin computar, tanto el vocablo como el nivel. Sin embargo, después de los ajustes hechos en el módulo adicional, éste procesa de igual manera para que Lexicator les otorgue un nivel. De forma similar al proceso que se sigue para los adverbios acabados en "-mente", en un primer paso, Lexicator identifica estos participios en los textos como lemas y, en un segundo paso, les da el nivel del *PCIC*. Esto ha sido posible, al haber identificado previamente los participios a través de los vocablos o *tokens* del verbo en el diccionario de FreeLing 2.1.

Por otro lado, este proceso de re-nivelación nos ha permitido calcular los niveles de los multi-vocablos basándonos en los criterios de nivel que exponemos en el apartado 4.1.6. Lexicator otorga automáticamente un nivel de lengua a todos los multi-vocablos ya que éstos no se indexan en el "Índice de nociones generales y nociones específicas" del *PCIC*. Es más, este módulo mantiene el nivel otorgado por el "Índice" del *PCIC* para aquellos multi-vocablos existentes y nivelados previamente en los inventarios de distintos niveles del *PCIC* y que hemos registrado y nivelado manualmente en el archivo de "locuciones.dat" del programa FreeLing.

Una vez nivelados todos los multi-vocablos restantes con nuestros criterios de nivel, se observa que el nivel que se da a varios multi-vocablos en algunas ocasiones coincide con los criterios del *PCIC,* pero en otras no. Sin embargo, en este último caso, al comparar los resultados de nuestros multi-vocablos nivelados con respecto a la calificación de niveles del *PCIC*, nuestra herramienta evalúa generalmente un nivel más bajo que el *PCIC*.

Otros multi-vocablos que no se encuentran en el glosario de locuciones pero que nivela Lexicator son las lexías numéricas y cuantificadoras. Son lexías de interés en el estudio de una segunda lengua (L2) por ello se procesan y computan. Como el analizador de FreeLing, al procesar un texto, las identifica y las etiqueta, consideramos que deben puntuar a la

hora de evaluar un texto. Estas lexías, a las que aplicamos los mismos criterios de nivel que los multi-vocablos, quedarían sin calificar si no fuera porque se ha programado a Lexicator para que las nivele con los criterios de nivel en un módulo auxiliar.

4.4. Método de nivelación léxica del texto

La identificación del nivel de un texto es uno de los apartados más discutibles de nuestro trabajo porque aplicamos unos índices de medida, proponemos unos métodos de evaluación y determinamos el nivel de un texto según unos niveles de referencia determinados. Además, toda evaluación precisa de unos índices y métodos fiables y válidos. De momento, se nivela un texto mediante la medida patrón que nos proporciona el "Índice" del *PCIC*, la combinación de glosarios y el glosario de locuciones conjuntamente. De este modo, obtenemos lemas calificados en todos los niveles. Por ello, para determinar qué nivel tiene un texto donde aparecen lemas de todos los niveles, hay que marcar unos límites entre los niveles e identificar cuál es el nivel de un texto. A continuación, proponemos tres métodos cuantitativos:

- Método K-2000 por intervalos de frecuencias.
- Método de combinación de niveles y frecuencias por cálculo de área.
- Método de intervalos según extensión y nivel de un texto.

4.4.1. Método K-2000 por intervalos de porcentajes

Siguiendo el método de diagnóstico del Oxford 3000TM Text Checker, una herramienta automática que evalúa un texto basándose en su diccionario conocido como el Oxford 3000 vocablos clave, nosotros también podemos evaluar un texto con el mismo criterio pero con el glosario de 2022 lemas del Dr. Kincaid. El Oxford 3000 TM Text Checker diagnostica el nivel de un texto según el porcentaje de vocablos del diccionario Oxford 3000 que contenga el texto objeto de estudio. El Oxford 3000 TM Checker propone que un texto cuyo vocabulario pertenezca en casi un 100 % al listado del Oxford 3000 se identifica como de nivel intermedio bajo B1; el que contenga entre un 90-95 % será de un nivel intermedio alto (B2) y el que contenga entre un 75-90 % sería de un nivel avanzado C1-C2.

Este método de listar palabras más frecuentes, como registro indicador del léxico básico, lo utilizan los analizadores electrónicos de textos como Lextutor que, basándose en listados que contienen entre 1000 y 5000 vocablos más frecuentes, dan el perfil de un texto en relación a un vocabulario básico, académico o específico. Una vez más, estos glosarios de palabras listadas por su frecuencia, rango o cobertura del texto, son útiles como listados para el diagnóstico léxico como propone Nation tanto para la enseñanza como para el aprendizaje (Nation y Kyongho, 1995) o, como proponemos nosotros, para la evaluación.

En nuestro caso, incluso con un número menor de palabras que el medidor de Oxford (3000) y el de Lextutor (entre 2000 y 5000 lemas y familias), proponemos diagnosticar el

nivel de un texto de forma semejante con el listado de 2022 palabras del Dr. Kincaid. Según Paul Nation, un listado de 2000 palabras es una buen límite para cubrir el vocabulario básico. Es más, un listado con un mayor número de palabras ya empieza a contener un vocabulario académico, aunque un vocabulario de más baja frecuencia puede ser cotidiano e, incluso, específico (Nation y Kyongho, 1995, 35 y ss.).

Tomando el glosario del Dr. Kincaid y basándonos en el intervalo de porcentaje de frecuencias determinado por Oxford 3000, proponemos dar un nivel léxico a un texto según se expresa en los porcentajes de la tabla 4.3.

Niveles	**A1-A2**	**B1**	**B2**	**C1**	**C2**
Cobertura del texto	95 %-100 %	90 %-95 %	80 %-90 %	75 %-80 %	<75 %

Tabla 4.3: Nivel léxico con el glosario del Dr. Kincaid.

Por ello, basándonos en los criterios de porcentaje propuestos por otros analizadores, nosotros postulamos que:

• Aquellos textos que incluyan en su vocabulario entre un 90 %-95 % de los lemas del Dr. Kincaid, los evaluaríamos de un nivel B1.

• Aquellos textos que incluyan en su vocabulario entre un 80 %-90 % de los lemas del Dr. Kincaid y el resto entre el Wiktionary, el de la Dra. Fuensanta López y FreeLing 0.5, los evaluaríamos con un nivel B2.

• Aquellos textos que incluyan en su vocabulario entre un 75 %-80 % de los lemas del Dr. Kincaid y el resto entre Wiktionary, Fuensanta, FreeLing 1.5, FreeLing 2.1 o esWordnet 1.6, los evaluaríamos con un nivel C1.

• La especificidad y alto nivel (C2) de un texto vendría marcada cuando el porcentaje del vocabulario utilizado en un texto superase el 30 % en cualquiera de los glosarios más grandes como FreeLing 2.1 o esWordnet 1.6.

Adicionalmente al nivel que nos diagnostique el glosario del Dr. Kincaid, en nuestro análisis de los textos obtendremos el porcentaje de vocablos en cada uno de los glosarios como indicador de cuántos vocablos y qué nivel tienen los vocablos que cubre cada uno de los glosarios.

Además, contamos con el referente del glosario del Dr. Kincaid en comparación con los otros glosarios, que podrán corroborar los niveles mediante el criterio de combinación de diccionarios.

4.4.2. Método de combinación de niveles y frecuencias con cálculo de área

El deseo de querer nivelar más del 40-50 % de vocablos de un texto, tal y como hace el "Índice" del *PCIC*, nos indujo a considerar la posibilidad de nivelar mediante la combinación de diccionarios aquel número de vocablos no nivelados por el "Índice" del *PCIC*. De esta manera podemos tener nivelados en torno a un 90 % de vocablos conjuntamente

con el "Índice" del *PCIC*, la combinación de glosarios y el glosario de locuciones. Una vez nivelado el léxico de un texto, precisamos determinar el nivel de dicho texto.

Como hemos podido observar en todos los procesos de prueba de nivelación, tanto de los textos como de los glosarios, se confirma que en todos los textos o glosarios confluyen lemas de todos los niveles, por tanto es preciso conocer cómo y qué determina el nivel de un texto una vez nivelado su léxico.

Para ello el método que exponemos ahora exige que procesemos el texto objeto de análisis. Por un lado, hallamos la frecuencia y rango de los vocablos en el texto. Por otro lado, calificamos los lemas teniendo como referencia los niveles del "Índice" del *PCIC* y los niveles de la combinación de glosarios. Al procesar los lemas nivelados no repetidos, obtenemos su frecuencia y rango y, a su vez, su distribución por niveles. Por tanto, cada lema nivelado tiene un valor de rango (número de orden en función de la frecuencia) y un valor de subrango (definido como el rango del lema en el subconjunto de lemas del mismo nivel). Al nivelar con 6 niveles de referencia, se obtendrán 6 curvas resultantes de unir los puntos correspondientes a los lemas de cada nivel. Además, con el fin de ofrecer una representación válida para cualquier longitud de texto, se han normalizado los rangos y subrangos desde 0 % al 100 %. Al procesar cualquier glosario nivelado, se obtiene una gráfica semejante a la forma de la vaina del cacao o a una "karela" (fruto hindú) por sus curvas rugosas. Como ejemplo, sirvan las representaciones del glosario del Wiktionary y del glosario de frecuencias del Dr. Padró que se observan en las figuras 4.9 y 4.10.

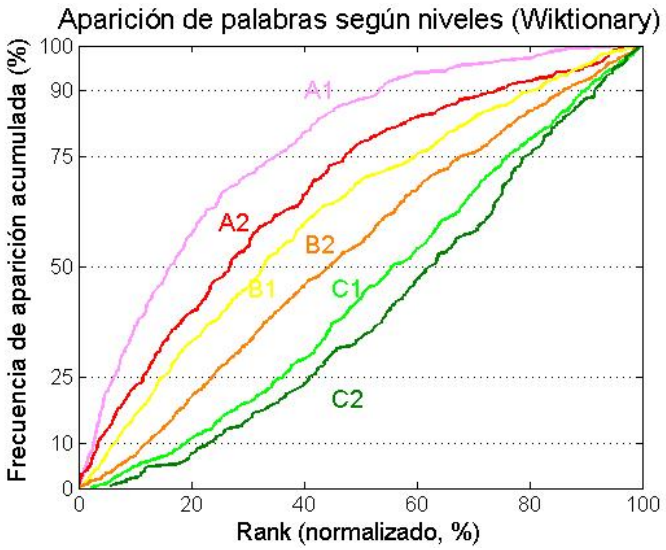


Figura 4.9: Representación subrango-rango de los lemas del glosario de Wiktionary. El subrango se define como el rango para cada uno de los niveles.

En la figura 4.9 se representa el rango de las palabras en función de su frecuencia de aparición. El eje horizontal se ha normalizado al rango total de todas las palabras. El eje vertical se ha normalizado para cada uno de los rangos de cada nivel. Los ejes han sido normalizados. Los 6 niveles de lengua ya se han representado anteriormente de forma separada y codificada en colores tal y como aparecen en las figuras 4.7 y 4.8. En la figura 4.9 se observa que la distribución de rangos de nivel empieza con el nivel A1 y siguen los demás niveles. Además de que el orden de aparición es el esperado, se observa que el

ascenso de los primeros niveles A1, A2 y B1 son inmediatamente los primeros en aparecer y anteriores a los niveles B2, C1 y C2, que van apareciendo inmediatamente después y ascendiendo de forma más lenta. Esta imagen no sólo confirma la experiencia de que en los niveles iniciales el aprendizaje de vocabulario es más rápido sino que la distribución nos permite observar claramente los niveles del léxico de un texto.

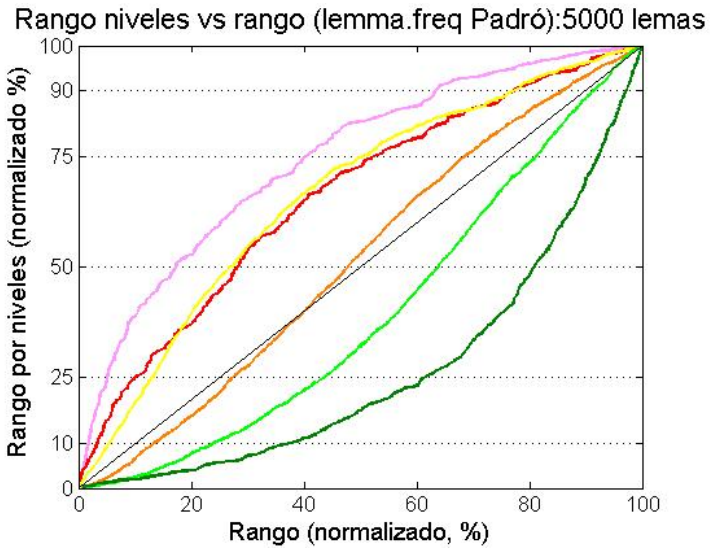


Figura 4.10: Representación subrango-rango de los lemas del glosario del Dr. Padró. El subrango se define como el rango para cada uno de los niveles.

4.4.2.1. Evaluación del nivel léxico de un texto

Sin embargo, llegados a este punto, aún no sabemos cuál es el nivel de un texto ni conocemos si tal calificación es fiable. Para ello, proponemos calcular el área entre la curva subrango-rango y una línea de referencia. La línea de referencia se define como aquella en la que el subrango coincide con el rango. Gráficamente es una línea recta que une el punto (0,0) y el punto (100,100) tal como se representa en la figura 4.10. Esta línea correspondería con la representación de esta gráfica en el caso de que sólo existiese un nivel y, por lo tanto, el subconjunto que define el subrango es todo el conjunto de lemas del texto. El área encerrada entre la línea de referencia y la curva subrango-rango se define como positiva si la curva se sitúa sobre la línea de referencia, y como negativa si se sitúa por debajo. Una curva subrango-rango que tenga partes por encima y por debajo de la línea de referencia dará lugar a un área que será la suma algebraica (teniendo en cuenta los signos positivos y negativos situados por encima de la línea de referencia y por debajo de ésta, respectivamente) de las áreas definidas por los cortes. Tal y como se observa en la figura 4.10, las curvas correspondientes a los niveles A1, A2 y B1 se sitúan por encima de la línea de referencia, mientras que B2, C1 y C2 están por debajo. Por tanto las áreas serán positivas para A1, A2, B1 y parte de B2 y negativas para el resto. Una vez definido este parámetro ya podemos establecer nuestro criterio de nivelación del texto que se enuncia como sigue:

El nivel de un texto es el mínimo nivel entre aquellos que poseen un área

negativa, siempre y cuando el nivel anterior sea positivo y el posterior (si lo hay) sea negativo.

En la figura 4.11 se representa el nivel que correspondería al glosario de frecuencias del Dr. Padró (Padró, 2011b) procesando los 5.000 primeros lemas de un total de 47.825.

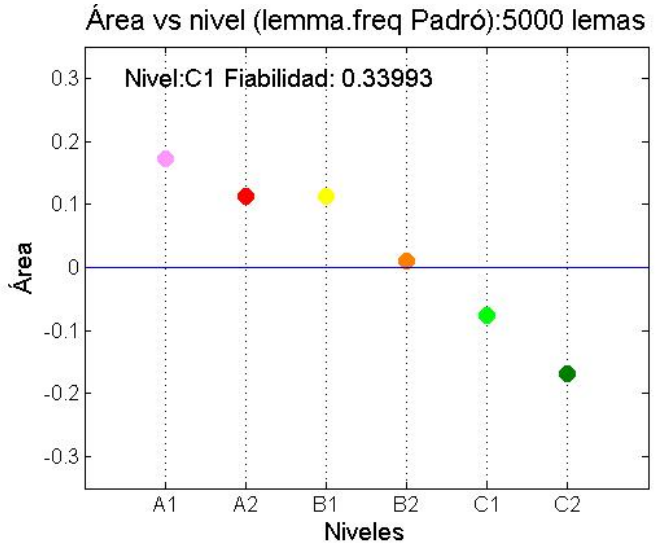


Figura 4.11: Área de los niveles para los 5.000 lemas más frecuentes del glosario del Dr. Padró.

4.4.2.2. Tendencia del nivel léxico de un texto

Tras aplicar esta definición de nivel, obtenemos un valor cuantificado a los 6 niveles de referencia, sin escalones intermedios. Sin embargo, sería apropiado saber si el texto nivelado se aproxima más al nivel superior o al inferior respecto al calculado. Para ello podemos definir otro parámetro que denominamos como tendencia. Para su cálculo tenemos en cuenta la recta que une el valor de área negativo que define el nivel (según el criterio anterior) y el valor positivo inmediatamente anterior. Esta línea recta corta al eje horizontal en un determinado punto situado entre los dos niveles. Si este punto está más cerca del nivel de área positiva diremos que la tendencia es hacia un nivel inferior, mientras que si el punto de intersección está más cerca del nivel de área negativa, diremos que la tendencia es hacia un nivel superior. Además, este parámetro tiene un carácter cuantitativo por lo que podemos utilizarlo para comparar resultados de distintos textos.

En la figura 4.11 podemos ver que el nivel del conjunto de vocablos analizado es C1, ya que el primer punto de área negativa (color verde claro), después de un punto de área positiva (color naranja), es el correspondiente al C1. Si trazamos una línea recta entre los puntos correspondientes a los niveles B2 y C1, podemos ver que el punto de intersección con la línea horizontal (de área nula) está más cerca del nivel B2 que del C1. En ese caso diremos que el nivel es C1 con una tendencia a B2. Cuantitativamente, al realizar los cálculos, diremos que el conjunto de vocablos tiene nivel C1 con tendencia $-0{,}87$ hacia B2 (con valor negativo para expresar que vamos hacia un nivel inferior). Este parámetro

de tendencia será utilizado en la calificación léxica de los textos presentada en el capítulo 7.

4.4.2.3. Fiabilidad del nivel léxico de un texto

Una vez obtenido el nivel del texto, es posible, a través de este cálculo, estimar la fiabilidad de esta determinación del nivel. Para ello, se ha realizado la identificación del nivel para el glosario del Dr. Padró, ya que contiene datos de frecuencia (ver un ejemplo de este glosario en el apartado A.8 del apéndice). Se ha observado que, al aumentar el número de lemas en el glosario, el área definida previamente decrecía conforme aumentábamos de nivel. Sin embargo, este decrecimiento monótono se ve perturbado cuando el número de lemas es bajo. Lo mismo ha ocurrido en el análisis de textos en función de la longitud del texto. Es decir, se ha comprobado que conforme aumenta el número de palabras de un texto, la calificación es más fiable. Atendiendo a este comportamiento, hemos definido un parámetro de fiabilidad cuantitativo que se define como:

> Fiabilidad: es la suma algebraica de las diferencias sucesivas de los valores de área para cada uno de los niveles.

Es decir, con una nivelación en 6 niveles existirán 5 diferencias sucesivas. Por ejemplo, si el área del nivel A1 es mayor que la del nivel A2, su contribución al parámetro de fiabilidad será positivo y viceversa. Por tanto, un valor mayor de este parámetro de fiabilidad se corresponderá con una nivelación más acorde con la evolución esperada y, por lo tanto, más fiable. El valor de la fiabilidad se encuentra en el rango $[-0{,}5, 0{,}5]$.

Como en todos los casos y fenómenos de la naturaleza, en el estudio del lenguaje también es importante el estudio de muestras con numerosos sujetos o elementos, porque ello aumenta la fiabilidad. No es menos significativo este hecho para el estudio del léxico de un texto ya que, cuantos más vocablos tenga un texto, más fiable es la determinación del nivel de dicho texto. Consideramos que un texto de 250 palabras es muy breve para que la fiabilidad del diagnóstico sea buena. Cuando sometemos un texto breve a esta distribución de niveles, se observa que la inestabilidad de los niveles es alta. Es decir, un nivel se estabiliza antes cuanto más extenso es el texto procesado. Por ello, consideramos que el método desarrollado en el siguiente apartado 4.4.3 es otra herramienta tan auxiliar como precisa para determinar el nivel exacto de un texto.

Deducimos, entonces, que con la ley de Zipf se establece una buena relación entre frecuencia y rango de las palabras cuando la extensión del texto es grande. Ese mismo inconveniente nos hemos encontrado cuando calificamos textos muy breves (menos de 300 vocablos). Aunque hay autores que proponen un "Half-rational model" más refinado, matizando la ley de Zipf, para evitar que la longitud del texto sea un inconveniente (Debowski, 2002), nosotros no lo hemos aplicado sino que se ha sometido el método de combinación de niveles y frecuencias al proceso del "cálculo de áreas" para determinar con más precisión el nivel que tiene un texto independientemente de su extensión.

Sin embargo, la falta de fiabilidad del diagnóstico en textos muy breves de este método, nos condujo a buscar otro método que permitiese definir el nivel en función de un número

bajo de vocablos en un texto como explicaremos en el método siguiente. Este método, desarrollado en el apartado 4.4.3, se puede generalizar para nivelar cualquier tipo de texto siempre que subyazca en el módulo que computa el nivel unos glosarios de referencia que den un nivel a los lemas. Por ello, es tan importante que contemos con el listado nivelado del "Índice" de *PCIC* y, en su defecto, con la combinación de glosarios, método que se ha explicado aplicando los criterios de nivelación que exponemos en el apartado 4.1.4.

4.4.3. Método de intervalos de porcentajes según extensión y nivel del texto

Mediante este método calculamos unos intervalos de porcentajes de lemas (equivalentes a la extensión del texto) por niveles para un texto. Este método se hace en función del número de palabras de un texto y una vez conocido previamente el nivel de los lemas. Este método de intervalos supone que cada texto tendrá unos valores diferentes para cada nivel y para cada extensión de texto.

Como se observa en la tabla 4.4, los valores de cada intervalo difieren. Estos valores se han obtenido de un *corpus* de nivel C1 al procesar y nivelar los discursos de Navidad del Rey Juan Carlos I desde 1975 hasta 2010. Postulamos que es un texto de nivel C1 ya que está escrito por un nativo competente con un vocabulario estándar y común para los receptores nativos españoles.

Nº lemas	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
200	10.2 - 16.1	6.4 - 10.7	28.3 - 35-4	19.2 - 26.8	12.2 - 18.7	5.3 - 10.4
300	9.4 - 14.3	6.3 - 10.0	28.3 - 34.7	21.0 - 27.1	13.2 - 19.0	6.0 - 10.5
500	8.4 - 11.9	6.3 - 8.7	28.2 - 33.2	23.0 - 27.1	13.2 - 19.0	6.0 - 10.5
1000	7.4 - 9.7	6.0 - 7.6	28.1 - 31.3	24.9 - 28.0	16.4 - 20.0	8.3 - 11.1
1500	6.7 - 8.6	5.7 - 7.0	26.7 - 30.5	25.0 - 29.4	17.6 - 20.7	9.5 - 11.5

Tabla 4.4: Intervalos de frecuencias en función del número de lemas de un texto.

Se observa en el desarrollo del método que los niveles de un texto se estabilizan a medida que el texto es de mayor tamaño. Es decir, conforme mayor es el número de lemas de un texto, mayor es la fiabilidad de evaluar el nivel de un texto. En la tabla 4.4 vemos que la diferencia numérica entre los intervalos de frecuencia de todos los niveles va disminuyendo a medida que aumenta el número de lemas en el texto.

De manera gráfica, en la figura 4.12 se representa el *corpus* del Rey con dos tipos de curvas para cada nivel. Por un lado, las líneas continuas de color más anchas o *curvas medias* representan los porcentajes medios de cada nivel. Por otro lado, las líneas de puntos de color representan el rango de error.

La gráfica 4.12 obtenida al procesar el *corpus* de los discursos navideños del Rey al que le otorgamos, por definición, un nivel C1, nos ha permitido probar este método, su fortaleza y dificultad. Por una parte, su fortaleza permite que el método sea más fiable cuanto más extenso es el texto. Por otro, su dificultad nos obliga a establecer varias tablas con unos intervalos de frecuencia a partir de un *corpus* de un nivel determinado. Es decir,

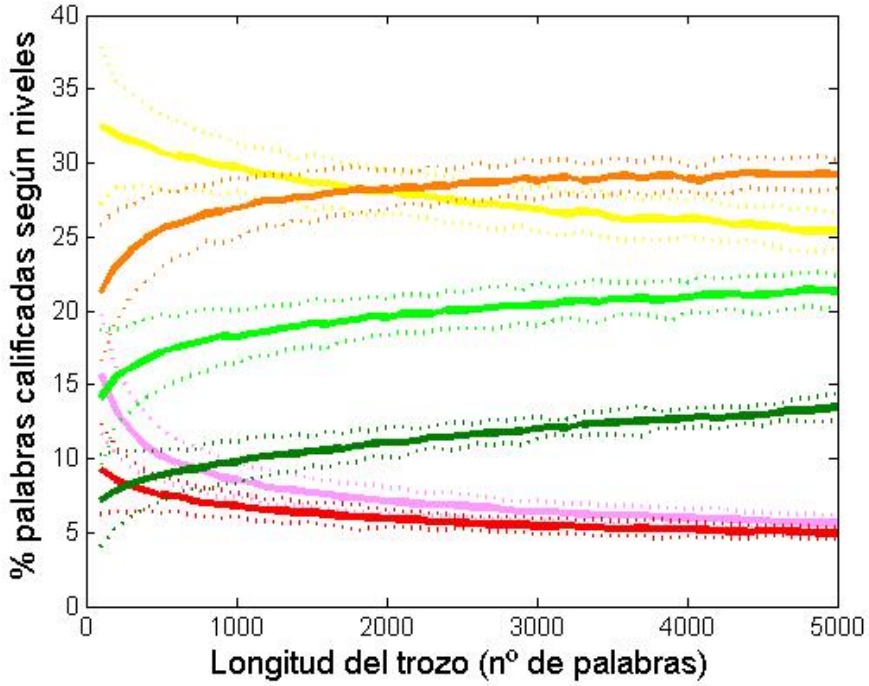


Figura 4.12: Rangos de porcentaje por niveles en función de la longitud del texto. Los textos han sido obtenidos del conjunto de los discursos navideños del Rey.

con este método no se obtiene de forma automática un valor numérico concreto para cada nivel que, a su vez, nos permita calcular el nivel del texto como en el método anterior. Sin embargo, este tercer método es tan dependiente como sistemáticamente aplicable una vez que se obtienen las tablas de intervalos para cada nivel a partir de un *corpus* nivelado como B1 o B2 o C1 o C2. Esto es, un *corpus* nivelado nos va a servir como referente para nivelar un texto ya que dicho *corpus* nos va permitir establecer unos intervalos para cada nivel en función del número de lemas.

Como afirmamos más arriba, procesar textos con un número de vocablos reducido va a aumentar el error de diagnóstico. Este error de diagnóstico, no obstante, ya se puede predecir al observar en la gráfica 4.12. Las *curvas de media* que representan los niveles más bajos A1 (rosa), A2 (rojo) y B1 (amarillo) comienzan a estabilizarse en torno a las 2000 palabras distintas, mientras para el intermedio B2 (naranja) y los superiores C1 (verde claro) y C2 (verde oscuro) la estabilidad de los niveles empieza desde las 2500 palabras distintas.

Así como podemos predecir la estabilidad de los niveles en función del número de lemas de un texto, la fiabilidad es otro criterio que se podría extraer de esta gráfica. Sin duda, la fiabilidad es muy grande en textos de 5000 lemas como se observa cuando las *curvas de error* y las *curvas de media* están más próximas entre sí. En nuestro estudio, para la defensa de este trabajo, vamos a utilizar textos entre 200 y 300 vocablos que se corresponden con los tests de escritura de los Diplomas de Español como Lengua Extranjera (DELE). Por ello, el método de intervalos de porcentajes parece ajustarse bien a textos breves.

Capítulo 5

Análisis sintáctico

5.1. Nivelación de estructuras sintácticas

5.1.1. Generación de estructuras sintácticas y criterios de calificación

Partimos de la creación de unas estructuras sintácticas a nivel sintagmático que hemos organizado por niveles de referencia, como las presenta el *Plan Curricular del Instituto Cervantes* (*PCIC)*. Estas estructuras las hemos reproducido en un fichero *ad hoc* que denominamos en lo sucesivo "fichero de estructuras". Construyendo el fichero a partir de estructuras sencillas, se ha ido aumentando el nivel de la estructura en función de criterios de complejidad para cada una de las estructuras. Por ello, para su desarrollo y configuración, hemos ido aplicando distintos principios como el de combinación de *PoS* o categorías gramaticales y de sintagmas. Se ha aplicado el criterio de variedad de transformaciones de los sintagmas, de longitud de las estructuras o sintagmas y de inserción de más categorías como adjetivos, participios e intensificadores adverbiales dentro de estructuras menores, según postula la gramática generativo-transformacional (Miller, 1962; Chomsky, 1989). En total, se han generado 2.586 estructuras no repetidas, de las cuales un 2,09 % pertenecen al nivel A1, un 6,88 % al nivel A2, un 31,40 al nivel B1, un 38,82 % al nivel B2, un 19,22 % al nivel C1 y un 1.59 % al nivel C2. En la tabla 5.1 presentamos el número de estructuras y su porcentaje por niveles.

Niveles	**# 2.586**	**%**
A1	54	2,09 %
A2	178	6,88 %
B1	812	31,40 %
B2	1004	38,82 %
C1	497	19,22 %
C2	41	1,59 %

Tabla 5.1: Distribución de estructuras sintácticas por niveles.

A continuación, especificamos algunos de los criterios aplicados para configurar dichas estructuras:

- Incremento de categorías verbales con pronombres personales (PP*) o relativos (PR*), adverbios afirmativos (RG) o negativos (RN), locuciones adverbiales (RG*, RN*), preposiciones (SPS00), locuciones preposicionales (SPS_*), conjunciones (CS), locuciones conjuntivas (CS*), para conformar sintagmas verbales diferentes. Además, se amplían las estructuras distinguiendo verbos predicativos (VM*) o copulativos (VS*) y marcando modos verbales en indicativo (VSI*, VAI*, VMI*), subjuntivo (VSS*, VMS*, VAS*) o imperativo (VSM*, VMM*, VAM*); también, las formas compuestas de perfecto en indicativo (VAI* VMP*) y subjuntivo (VAS* VMP*) o de formas pasivas (VAI* VSP* VMP*) se registran como incremento de una forma verbal.

- Combinación de categorías funcionales. Se intercalan adverbios positivos (RG*) antes o después de un verbo (V*), de una conjunción (CS, CS*), de otro adverbio (RG RG*) o de una preposición (SPS_*). Se disponen adverbios negativos (RN) ante verbo (V*); se considera la ubicación de una preposición (SPS00, SPS*) u otro adverbio (RG, RG_*) antes o después de un infinitivo (VMN*, VAN*, VSN*), de una conjunción (CS, CS*) o del verbo copulativo "ser" (VS*), auxiliar "haber" (VA*) o predicativo (VM*).

- Incremento de estructuras nominales. Se introducen elementos insertables que matizan al nombre o lo intensifican mediante adverbios (RG, RG*, RN), determinantes (D*) o pronombres (P*). Se amplía la estructura mediante el uso de la preposición (SPS*) ante nombres o adjetivos que requieren preposición.

  RG ante NC*, AQ*, P*, D*
  RN ante D*, P*
  SPS00 antes y después NC*, A*

- Combinación de categorías con los mismos *PoS* o elementos generando estructuras de significado equivalente. Se crean grupos de sintagmas similares o estructuras ecuacionales diferentes. Por ejemplo:

  P0000000 PP3* VM* SPS00 NP* "se lo doy a Juan"
  P0000000 PP3* VM* DA* NC* SPS00 NP* "se lo doy el libro a Juan"
  P0000000 PP3* VM* SPS00 NP* DA* NC* "se lo doy a Juan el libro"

- Ampliación de sintagmas nominales (N*, A*, P*) que se incrementan con cláusulas de relativo (PR*, Fc PR*). Además, la oración de relativo puede tener un verbo simple (VM*, VS*, VA*) o compuesto, seguido de un participio (VM*, VS*, VA* + VMP*) o de dos participios (VA* + VSP* + VMP*) en la estructura pasiva.

  PR* después de N*, A*, P*

- Variación de sintagmas verbales. Se generan estructuras con verbos en modo indicativo (VSI*, VMI*) o en subjuntivo (VSS*, VMS*), en tiempos compuestos en indicativo (VAI* VMP*) o en subjuntivo VAS* VMP*), en voz pasiva en indicativo

(VAI* VSP* VMP*), en subjuntivo (VAS* VSP*, VMP*) e imperativo (VMM*, VSM*).

- Agrupación de categorías gramaticales iguales o diferentes. Se configuran estructuras similares a las locuciones. Es decir, se crean estructuras con cuantificadores formados con locuciones adverbiales, preposicionales, pronominales o partitivas; también combinables con colocaciones nominales o estructuras con verbos preposicionales. Por ejemplo, al haber agrupado la conjunción "ni" (CC) más un adverbio "siquiera" (RG) en la locución adverbial "ni siquiera" (RG*), disponemos de este tipo de estructuras RG* que son muy rentables y combinables dentro de otras estructuras sintagmáticas. Además, estas agrupaciones se identifican fácilmente a nivel discursivo y oracional.

- Combinación de pronombre distintivos (catafórico acusativo) o no (catafórico dativo) en estructuras pronominales. Sólo el acusativo de tercera persona es formalmente distintivo: "lo" PP3MSA00, "los" PP3MPA00, "la" PP3FSA00, "las" PP3FPA00. Esta distinción permite generar estructuras muy concretas y claramente especificadas. En nuestras estructuras, los demás pronombres acusativos se registran de forma genérica (PP3*):

  PP* PP3* VM* SPS00 PP* "te lo hago a ti"
  PP* PP3* VA* VMP* SPS00 PP* "te lo ha hecho a ti"

5.1.2. Estructuras sintácticas de B2

Teniendo en cuenta todos estos principios de complejidad sintáctica, hemos creado varias ecuaciones sintácticas computables. Para ello, nos hemos basado en los criterios expuestos arriba y en las estructuras determinadas y niveladas en el *Plan Curricular del Instituto Cervantes* (*PCIC*). Si las estructuras registradas en nuestro fichero existen en el texto, se procesarán porque se corresponderán con las estructuras niveladas que se registran en el fichero de estructuras y, por lo tanto, las estructuras procesadas obtendrán un nivel. Si no están registradas dichas estructuras, quedarán sin identificar en el texto y, en consecuencia, sin nivelar. Posteriormente, se precisa aplicar el método que nos permita diagnosticar el nivel de un texto después de computar todas las estructuras procesadas.

A continuación, reproducimos algunas estructuras extraídas del *PCIC*. Se cita entre paréntesis el tomo con el nivel de referencia que les corresponde y la página dónde se hallan ubicadas.

- Estructura de anafóricos, dativos posesivos o de interés con verbos predicativos o copulativos que tienen un nivel A2 (Cervantes, 2006, B1-B2: 180) pero que al sistematizarse se convierten en B1 (Cervantes, 2006, B1-B2: 64):

  SPS00 PP* PP* VM* "a mí me gusta" con nivel B1
  SPS00 NP* PP* VM* "a Juan le gusta" con nivel B1
  SPS00 D* NP* PP* VM* "a los niños les gusta" con nivel B1

- Procedemos de tal manera que, cuando esa misma estructura es susceptible de que se le introduzca un cambio, le añadimos un nuevo *PoS*, elevándose su nivel. En el proceso de ampliación, introducimos una estructura con verbo copulativo:

  SPS00 PP* PP* VS* AQ* " a mí me es igual" con nivel B2
  SPS00 NP* PP* VS* AQ* "a Juan le es igual " con nivel B2
  SPS00 D* NP* PP* VS* AQ* "a los niños les es igual " con nivel B2

- En este caso podemos introducir también el adverbio negativo:

  SPS00 PP* RN PP* VS* AQ* " a mí no me es igual" con nivel B2
  SPS00 NP* RN PP* VS* AQ* "a Juan no le es igual " con nivel B2
  SPS00 D* NP* RN PP* VS* AQ* "a los niños no les es igual " con nivel B2

- También podemos computar estructuras de posesivos anafórico-dativos o de interés de 3ª persona con nivel B1 como P0000000 PP3* VM* "se lo doy" (Cervantes, 2006, B1-B2: 64) que, al ampliarse la estructura o tener un carácter ecuacional, adquiere un nivel superior en nuestras estructuras:

  P0000000 PP3MSA00 VM* RG* "se lo doy también" con nivel B1
  P0000000 PP3MSA00 VM* SPS00 NP* "se lo doy a Juan" con nivel B1
  P0000000 PP3MSA00 VM* RG* SPS00 NP* "se lo doy también a Juan" con nivel B2

- Igualmente procedemos con la estructura pronominal del catafórico dativo y del catafórico acusativo (PP*) como complemento directo del nombre (NC*) al que se refiere, en función de complemento directo también. Se computaría entonces para los casos en que se especifica claramente el pronombre personal acusativo como "lo" PP3MSA00, "los" PP3MPA00, "la" PP3FSA00, "las" PP3FPA00. Este tipo de estructura es interesante porque revela la cohesión sintáctica dentro del sintagma verbal al tener que coincidir el género y el número del pronombre con el nombre al que hace referencia. Partimos de la estructura inicial de B1 PP* PP3* VM* "se lo di" (Cervantes, 2006, B1-B2: 64). Luego, las estructuras se sistematizan y aumentan el nivel de nuevo cuando aparecen estructuras del tipo:

  PP* PP3**A00 VM* SPS00 PP* "te lo hago a ti" con nivel B1
  PP* PP3**A00 VM* SPS00 PP* D* NC* "te lo hago a ti el peinado" con nivel B2
  PP* PP3**A00 VA* VMP* SPS00 PP* D* NC* "te lo ha hecho a ti el peinado" con nivel C1

- Otro ejemplo de transformación es la estructura combinada del determinante catafórico-anafórico "lo" (DA0NS0) con otras categorías cuando se parte de un nivel B2 (Cervantes, 2006, B1-B2: 55):

  DA0NS0 AQ* "lo bueno" con nivel B2
  DA0NS0 D* AQ* "lo bastante grande" con nivel B2
  DA0NS0 RG AQ* "lo más interesante" con nivel B2

DA0NS0 AO* "lo primero", deducimos que también tendrá un nivel B2.

De ahí que, si para DA0NS0 AQ* CS "lo bueno que" el *PCIC* marca un nivel C1 (Cervantes, 2006, C1-C2: 56), cada una de las siguientes estructuras mantiene el nivel C1 como DA0NS0 AO* CS "lo primero que" pero, al introducirse otra categoría o cambiar el orden del *PoS* en la estructura, se nivela como sigue:

DA0NS0 AQ* CS_CL "lo primero es que" que se eleva a un nivel C2
DA0NS0 AQ* V* CS "lo interesante es cuando" que también se eleva a un nivel C2

En definitiva, muchas de las estructuras se han creado partiendo de las propuestas por el *PCIC*. Una gran parte ellas son el producto de haberse ampliado estructura menores e incrementado el nivel original.

5.1.3. Estructuras sintácticas de C1

Uno de los criterios que el *PCIC* aplica en varias ocasiones para la nivelación de estructuras es marcarlas como propias de un nivel aunque se desarrollan y vuelven a apuntar en otro nivel más elevado donde se considera que se sistematizan (Cervantes, 2006, A1-A2: 25). Por lo tanto, deducimos que la aparición de estructuras de B2 utilizadas por un aprendiz de forma sistemática y con variaciones son aquellas que ya están asimiladas y se consideran con un nivel C1. Este tipo de estructuras generalmente se identifican con aquellas que sus categorías se amplían o se combinan. Por ejemplo, cuando tenemos estructuras preposicionales incipientes en un nivel B1 como adjetivos (AQ* SPS*), nombres (NC* SPS*) o verbos (V* SPS00) acompañados de una preposición, éstas se aprenden en un nivel B2 pero se sistematizan en el nivel C1. Otro caso de nivel avanzado (C1) creemos que es la inserción sistemática de dos adverbios (RG RG) o de locuciones adverbiales (RG* RG*) con sintagmas verbales.

La agrupación de dos categorías iguales o con una preposición entre ellas es el caso de lo que hemos categorizado como locuciones determinativas (D* SPS00 D*) o pronominales (P* SPS00 P*) que la gramática tradicional denomina determinantes o pronombres indefinidos, numerales o adverbios de cantidad, y que el *PCIC* clasifica como "cuantificadores" desde una perspectiva más generativista (Cervantes, 2006, A1 A2: 105). Por ejemplo:

- "cualquiera de" está clasificado como pronombre indefinido con un nivel B2.
- "alguno de" igualmente se lista como pronombre indefinido de nivel B2.
- "alguno que otro" también se indexa como pronombre indefinido, pero con un nivel C2.

Con respecto a las lexías verbales, hay varias locuciones verbales listadas y niveladas en el glosario de locuciones (locuciones.dat) y, dada la abundante existencia de este tipo de estructuras en castellano, es un apartado que se podría implementar. En el nivel B1 se introduce parte de estructuras verbales como las perífrasis (V* SPS00 V*) y los verbos

preposicionales (VMP* SPS*). Estas estructuras verbales corresponden a un B2, nivel donde se practican y amplían, por ejemplo, las perífrasis verbales, pero se sistematizan en el nivel C1.

Sin embargo, no se han listado muchas locuciones verbales porque algunas son vulnerables a la inserción de adverbios, sintagmas nominales o pronombres. Este fenómeno provoca que se desmembre la locución verbal y, entonces, no se pueda computar. Este caso de desestructuración de la locución merece un estudio aparte.

Con respecto a las locuciones preposicionales (SPS_*) y locuciones adverbiales (RG*), se justifica el nivel dado a cada tipo de locución en la sección 4.1.6. En este apartado se exponen los dos criterios, en general, de nivelación que se han seguido: el del *PCIC* para las locuciones que listan sus glosarios y el de la longitud de la locución; concretamente, se describen otros criterios más específicos. Hemos considerado este glosario de locuciones una parte fundamental de esta investigación ya que es un glosario que favorece el etiquetado de los textos y permite no sólo procesar grupos de palabras sino recategorizarlas gramaticalmente al agruparlas y otorgarles un nivel de aprendizaje, para luego ser todo procesado.

Tanto en el glosario de locuciones como en el fichero de estructuras se recogen estructuras que combinan categorías gramaticales, que en ocasiones se duplican pero que no se procesan dos veces sino que cada una se computa en el nivel léxico o en el sintáctico, respectivamente. Por ejemplo, en el fichero de multivocablos registramos "lo más" como una locución adverbial comparativa (RG_CP) con un nivel B2; en la plantilla de estructuras, el sustantivador "lo" (DA0NS0) en combinación con un adverbio (RG), "lo más" (DA0NS0 RG) tiene un nivel B2. El mismo nivel B2 se le da a la misma estructura con un adverbio más (DA0NS0 RG RG) "lo más pronto" y con un adjetivo (DA0NS0 RG RG AQ*) "lo más pronto posible" pero, al introducir ahora una conjunción subordinante (DA0NS0 RG CS) "lo más que", le corresponde un nivel C1 (Cervantes, 2006, C1-C2: 56).

Por otro lado, en el glosario de locuciones recogemos dos vocablos con dos categorías diferentes, un determinante (DA0NS0) y un pronombre posesivo (PX). Ambos se conforman como una locución con una sola etiqueta, por ejemplo, el posesivo sustantivado de primera persona PX1NS0S0 "lo_mío" (Cervantes, 2006, C1-C2: 59).

Aunque algunas estructuras, como el relativo referencial DA0NS0 PR0CN000 "lo que" (Cervantes, 2006, B1 B2: 66, C1-C2: 68 y 95), se podrían haber representado con los *PoS* diferenciados, en el fichero de estructuras se han indexado como una locución conjuntiva referencial: CS_PR0CN0RF "lo_que" porque procesamos un relativo compuesto específico con la función de ser un nexo referencial (RF). Consecuentemente, mantenemos el nivel C1 para las siguientes estructuras que registramos en el fichero de estructuras:

- CS_PR0CN0RF RG "lo_que más"
- CS_PR0CN0RF DI* "lo_que tantos"

Se clasifican con nivel C1 aquellas estructuras compuestas por un determinante demostrativo más un pronombre relativo, e incrementadas, para mostrar distancia, cercanía o lejanía al hablante. Estas estructuras podemos expresarlas como:

- PD* PR0CN000 “aquel que” (Cervantes, 2006, C1-C2: 67)
- DD* NC* PR0CN000 “aquel muchacho que”
- DD* NC* AQ* PR0CN000 “aquel muchacho rubio que”
- DD* NC* RG AQ* PR0CN000 “aquel muchacho tan rubio que”

Otra de las estructuras propias del nivel C1 son las conjunciones o locuciones conjuntivas (CS, CS*) por introducir cláusulas subordinadas en posición antepuesta al sintagma verbal, en *RDP* (*Right Detach Position*), o postpuesta al sintagma verbal, en *LDP* (*Left Detach Position*). Es decir, en este nivel nos encontramos con sintagmas verbales simples en indicativo (VSI*, VMI*, VAI*), en subjuntivo (VMS*, VSS*, VAS*) o compuestos de indicativo (VAI* VMP*, VAI* VSP*), de subjuntivo (VAS* VMP*, VAS* VSP*) o de imperativo (VMM*, VSM*) que conforman cláusulas circunstanciales con conector o locución conjuntiva (CS*) antes o después del verbo principal. Sin embargo, esta propiedad de ubicación tan característica de un nivel de referencia de C1, no lo computamos porque no reproducimos todas las posibilidades, a pesar de que es un buen índice para distinguir entre el nivel B2 y C1. Dado que la extensión de la unidad de frase, la distribución de los *PoS* y de las estructuras en la oración reflejan una complejidad mayor en textos de mayor nivel, estos índices de extensión, distribución y posicionamiento de los *items* son implementables, computablemente independientes de las estructuras y ajustables al nivel, porque son índices genéricos (Attali *et al.*, 2010).

5.2. Definición de índices sintácticos

Para definir los índices sintácticos partimos de los parámetros propuestos por Kellog Hunt y su concepto de “T-unit” o unidad mínima terminal (Hunt, 1977, 101). Estos índices son útiles para analizar la madurez sintáctica en discursos escritos en lengua inglesa y, en nuestro caso concretamente, en español. Destacamos los propuestos por Véliz (Véliz, 1999, 1988) y Torres (Torres González, 1993) y consideramos las "unidades sintácticas estructuralmente complejas" como criterios susceptibles de medición. Miller y Mckean en su Derivational Theory of Complexity (DTC) propusieron unos parámetros similares basándose en medir la complejidad sintáctica a partir del número de transformaciones que pueden producirse en una cláusula o una unidad básica (Miller y McKean, 1964). Se comprobó que la complejidad sintáctica en una L1 aumentaba con la edad del aprendiz, es decir, con la experiencia formativa.

En primer lugar, exponemos los índices primarios y secundarios propuestos por Hunt (Hunt, 1965) y también aplicados por Véliz (Véliz, 1999, 2004); después, los criterios cuantificativos procesables que hemos aplicado a nuestras estructuras sintácticas niveladas y, por último, el criterio de máximo nivel, cuyas variables nos permiten reidentificar las estructuras de los textos analizados por sus valores cuantitativos.

5.2.1. Índices propuestos por Hunt y Véliz

Los índices propuestos por Hunt y Véliz para medir la madurez sintáctica que nos orientan para nuestro estudio son tres: longitud de la unidad mínima terminal, longitud de la cláusula e índice de cláusula subordinada. Por ello, el promedio de longitud de la U-terminal, el promedio de longitud de las cláusulas y el promedio de cláusulas por Unidad-terminal son índices calculables, medibles (Hunt, 1977; Véliz, 1999; Olloqui de Montenegro, 1991) y buenos indicadores. La identificación de la unidad mínima terminal (U-t) de Hunt es la que consta de una cláusula principal y todas las subordinadas dependientes de ella, mientras que las oraciones coordinadas y yuxtapuestas se consideran independientes y marcan el límite de la U-t (Rodríguez Fonseca, 1999; Martín Úriz *et al.*, 2005). Sin embargo, en los siguientes apartados adaptamos el concepto de U-t al estudio como unidad terminal gráfica para poder procesarlo como exponemos a continuación.

5.2.1.1. Índices primarios

1. Longitud promedio de la unidad mínima terminal. En nuestro caso se calcularía dividiendo el número total de palabras por el total de unidades terminales del texto. Estas unidades terminales las identificamos por estar finalizadas con un punto gráfico (.).

2. Longitud promedio de la cláusula. Se obtendría al dividir el número total de palabras por el número total de cláusulas del texto. Las cláusulas se marcarían por los verbos en forma personal.

5.2.1.2. Indices secundarios

Índices no clausales

- Promedio de sintagmas por unidad terminal. Se calcula dividiendo el número total de sintagmas o *S-constituents* por el total de unidad terminal. Tales componentes sintagmáticos son sintagmas nominales por unidad terminal, sintagmas adjetivos por unidad terminal, sintagmas adverbiales por unidad terminal, sintagmas preposicionales por unidad terminal, sintagmas de formas verbales no finitas por unidad terminal.

Índices clausales

Índices de subordinación

Permiten calcular el número de marcadores subordinantes de los *S-constituents* e identificar el tipo de subordinación. Según, Véliz este índice aumenta en los textos argumentativos y expositivos, y disminuye en los narrativos (Véliz, 1999).

1. Promedio de cláusulas adjetivas por unidad-terminal. Se halla al dividir el total de cláusulas adjetivas por el total de unidades terminales.

2. Promedio de cláusulas completivas por unidad terminal. Se halla al dividir el total de cláusulas completivas por el total de unidades terminales.

3. Promedio de cláusulas circunstanciales por unidad terminal. Se halla al dividir el total de cláusulas preposicionales por el total de unidades terminales.

Índices de categorías gramaticales

Según Hunt, el mayor o menor uso de categorías como adjetivos, gerundios, etc. indica la madurez de un texto. Incluso las puntuaciones de distintos niveles de madurez de los textos estudiados mostraron ser semejantes en distintas lenguas (Hunt, 1977, 101). Por ello, promedios como los que siguen también podrían computarse:

4. Promedio de calificativos por unidad terminal.

5. Promedio de posesivos por unidad terminal.

7. Promedio de aposiciones por unidad terminal.

8. Promedio de infinitivos por unidad terminal.

9. Promedio de gerundios por unidad terminal.

10. Promedio de participios-adjetivos por unidad terminal.

11. Promedio de participios-predicativos por unidad terminal.

12. Promedio del parámetro interrogativo (e.g. qué, quién, cuándo, cómo, cuánto, dónde, porqué).

Con los recursos que disponemos nosotros, sería posible tanto identificar los índices primarios como secundarios. Por ejemplo, los índices clausales 1, 2 y 3 se obtendrían al calcular el promedio de las cláusulas adjetivas mediante los relativos (PR*), de las cláusulas sustantivas mediante los marcadores completivos (CS) y de las cláusulas circunstanciales mediante los marcadores adverbiales (RG*) y preposicionales (CS*). Efectivamente, resulta muy económico computar los índices de promedios del 4 al 12 porque se trabaja con lemas o sintagmas fácilmente procesables. Es más, el cálculo de tantos índices aporta muchos datos numéricos sobre sintaxis oracional o sobre categorías gramaticales en el texto pero no información del nivel propiamente dicho. Una vez más, nuestro objetivo es dar un nivel sintáctico de referencia a un texto.

En nuestro planteamiento inicial, y posterior diseño del fichero de estructuras, aplicamos el supuesto de que un mayor nivel del texto implica una mayor complejidad de sus estructuras sintácticas o sintagmáticas y de que el promedio entre los índices primarios y secundarios (Torres González, 1993, 15-16) es una medida complementaria para reconocer el nivel de un texto. Hay autores que han relacionado los distintos niveles con "etapas de adquisición" (Checa García y Lozano, 2002) y que, en realidad, como las etapas de adquisición están vinculadas a los niveles, éstos niveles se pueden reproducir en algunos esquemas sintácticos en función de las etapas o el nivel como se representan y reproducen en el *PCIC*. Podemos deducir, entonces, que la mayor o menor complejidad sintáctica se identifica con una etapa del aprendizaje. Es decir, cierta etapa en el desarrollo de aprendizaje y producción de una lengua se corresponde con un determinado nivel de adquisición alcanzado (Wolfe-Quintero *et al.*, 1998). Consecuentemente, la madurez sintáctica depen-

de, a su vez, de la representación mental que el aprendiz tenga primero en la L1 como luego en la L2.

Sin duda, la complejidad sintáctica en ocasiones implica incoherencia en el texto ya que las oraciones se insertan y suceden, y se puede perder la "eficacia comunicativa" (Véliz, 1999). A pesar de que la coherencia semántica es un importante indicio de madurez en un aprendiz de L2, ésta se puede detectar en los textos objeto de estudio con el método de LSA. No obstante, en nuestro estudio apuntamos la posibilidad de aplicar un criterio de cohesión, al menos, en las categorías nominales y verbales con el glosario de esWordnet.

5.2.2. Índices procesables adaptados al estudio

A continuación, exponemos un conjunto de índices primarios y secundarios inspirados en Hunt (Hunt, 1965, 1977) pero adaptados a nuestra investigación. Por un lado, describimos los "índices generales marcadores de nivel" que se corresponden con los índices del 1 al 9, asociados a la oración o al texto. Por otro, dentro de los sintagmas o grupos de categorías gramaticales del texto que Hunt relacionaría con sus índices secundarios, consideramos de interés, destacar los "índices específicos o auxiliares" del 1 al 11. Entre la gran variedad de índices que nos informan qué características tiene un texto de un cierto nivel, presentamos una serie de ellos en los siguientes apartados.

5.2.2.1. Índices generales indicadores de nivel

Son índices que se obtienen a partir de la aplicación de unos criterios sintácticos computables definidos a partir del *PCIC* y que fundamentan la robustez de nuestra hipótesis. Los cálculos se basan en criterios que han configurado un conjunto de estructuras identificables y extraíbles en un texto. Entre los índices destacamos:

1. Longitud del texto. Nosotros analizamos textos con un número de palabras entre 150-250.

2. Longitud de la unidad-mínima-terminal o unidad-t procesable: Desde la letra mayúscula con la que empieza el texto hasta el punto, incluidas subordinadas, coordinadas o yuxtapuestas, si las tiene.

3. Longitud de la cláusula: oración independiente con verbo en forma personal con una longitud de X elementos pero con dos o más verbos no-finitos (verbos predicativos: VMI*, VMS*, VMM*; verbos copulativos: VSI*, VSS*, VSM*; y verbos auxiliares VAS*, VAI*, VAI*). En este índice se incluyen cláusulas completivas (CS), coordinadas (CC) o subordinadas adverbiales (CS*).

4. Índice de coordinación.
   Mide el número de vocablos coordinantes (CC) en la frase y en el texto.

5. Índice de subordinación.
   Calcula el número subordinadas conjuntivas (CS) y pronominales de relativo (PR*) en la frase y en el texto.

6. Número de estructuras sintácticas diferentes.
   Este índice informa de cuántas estructuras diferentes indexadas en el fichero de estructuras reproduce el texto analizado.

7. Frecuencia de distintas estructuras.
   Permite conocer la variedad de estructuras que crea el aprendiz.

8. Índice de repetición de estructuras iguales dentro de la frase.
   Éste es un parámetro indicador tanto de paralelismo o especificidad de un tipo de texto, como de inmadurez sintáctica si se repiten sistemáticamente estructuras semejantes.

5.2.2.2. Índices específicos o auxiliares

Estos índices informan sobre las características sintácticas de un texto sin diagnosticarle la calificación o nivel. Aunque estos índices son aplicables a cualquier texto, según el tipo de texto que se analice, es importarte diseñar los índices acorde con la tipología discursiva del texto. Por ello, ciertos textos precisarán índices específicos (Biber *et al.*, 1998) . Por ejemplo, en una carta personal o informal esperaremos la presencia de agentes animados informales como "tú" mientras que en una carta formal hallaremos el uso de "usted".

1. Índice de variedad verbal modal, temporal y aspectual del texto (modos, tiempos y agrupaciones verbales como la estructura pasiva o la perífrasis).
   La medición de este índice permitirá conocer la destreza del aprendiz al utilizar las numerosas estructuras verbales del castellano en:

   - Modo: indicativo (V*I), subjuntivo (V*S), imperativo (V*M), infinitivo (V*N), gerundio (V*G) y participio (V*P).
   - Tiempo simple: presente (V**P), imperfecto (V**I), futuro (V**F), pasado (V**S) y condicional (V**C); tiempo compuesto: presente (VA*P V*P), imperfecto (VA*I V*P), futuro (VA*F V*P), pasado (VA*S V*P) y condicional (VA*C V*P).
   - Locuciones verbales: V* SPS00 V*P, V* CS V*N, VM* V*G.
   - Agrupaciones verbales: V* SPS00 V*, V* V* SPS00 V*, V* SPS00 V* V*, VMN* V*P, VMG* V*P.
   - Pasiva: simple (VS* V*P) o compuesta (VA* VS* V*P).

2. Índice de cohesión nominal y verbal.
   Para hallar este índice se tiene en cuenta la concordancia de género y número para las categorías nominales, y la de persona y número para la categoría verbal. Por ello, el nombre (N*) concordará con el determinante (D*), el adjetivo (A*) y el participio al menos en uno de los dos rasgos: en género masculino (M) o femenino (F), o en número singular (S) o plural (P). Mientras, los verbos (V*) concordarán con los pronombres (P*), dentro de la unidad de frase, en número singular (S) o plural

(P), o en persona (1ª, 2ª , 3ª); y con los adjetivos (AQ*) o participios (VMP*), inmediatamente seguidos al verbo, concordarán en número.

3. Índice de negación (RN*).
   El uso de la negación es otro índice de madurez sintáctica que interesa registrarlo y computarlo, según los creadores de Coh-Metrix (McNamara *et al.*, 2006).

4. Índice de persona.
   Es de interés para textos de carácter narrativo o persuasivo. Se computa el pronombre personal o presencia de persona gramatical en el pronombre:

   - PP1CSN00 yo
   - PP1MP000 nosotros
   - PP1FP000 nosotras
   - PP2CS000 tú
   - PP2MP000 vosotros
   - PP2FP000 vosotras
   - PP3MS000 él
   - PP3MP000 ellos
   - PP3FS000 ella
   - PP3FP000 ellas
   - PP2CS00P usted
   - PP2CP00P ustedes

   Se puede registrar también la forma verbal personal de las 1ª y 2ª personas del singular y plural de un verbo predicativo (VM*), copulativo (VS*) o auxiliar (VA*):

   - VM **1S0, VM**2S0, VM**3S0, VM**1P0, VM**2P0, VM**3P0
   - VS**1S0, VS**2S0, VS**3S0, VM**1P0, VS**2P0, VS**3P0
   - VA**1S0, VA**2S0, VS**3S0, VA**1P0, VA**2P0, VA**3P0

5. Índice de reflexividad.
   Los pronombres que marcan la reflexividad, según las estructuras sintácticas que lo acompañen, pueden convertirse en dativo de interés. También este índice, una vez más, apunta destreza en el uso de pronombres personales de 1ª y 2ª persona del singular y plural:

   - PP1CS000 V***1S0 “me lavo”
   - PP2CS000 V***2S0 “te lavas”
   - PP1CP000 V***1P0 “nos lavamos”
   - PP2CP000 V***2S0 ”os laváis”

6. Índice de estructuras con "se".
   Es útil este índice para detectar la expresión de persona sin responsabilidad, por tanto, impersonalidad o noción de pasiva-refleja o dativo de interés (Cervantes, 2006, C1-C2: 65). Ejemplos de este tipo de estructura son:

   - P0000000 VM**3S0 “se dice” (impersonal, pasivo-reflejo)
   - P0000000 VM**3P0 “se caen” (reflejo)
   - P0000000 VA**3S0 VMP* “se ha dicho” (impersonal, pasivo-reflejo)
   - P0000000 VA**3P0 VMP* “se han caído” (reflejo)

7. Índice de dobletes de pronombres átonos.
   Estos dobletes nos muestran la destreza del aprendiz en el uso y en el orden de los pronombres en función de acusativo, dativo de interés o enfático, tanto en posición proclítica con las formas no-finitas del verbo como enclíticas con las formas finitas:

   - Verbos no-finitos con formas proclíticas:
     - P0000000 PP1CS000 V* “se me cayó”
     - P0000000 PP2CS000 V* “se te cayó”
     - P0000000 PP3CS000 V* “se le cayó”
     - P0000000 PP1CP000 V* “se nos cayó”
     - P0000000 PP2CP000 V* “se os cayó”
     - P0000000 PP3CP000 V* “se les cayó”
     - P0000000 PP3FSA00 V* “se la doy” (dativo-acusativo)
     - P0000000 PP3FPA00 V* “se las doy (dativo-acusativo)
     - P0000000 PP3MSA00 V* “se lo doy (dativo-acusativo)
     - P0000000 PP3MPA00 V* “se los doy” (dativo-acusativo)
   - Verbos finitos con formas enclíticas de la partícula “se” y un pronombre personal (PP*):
     - en infinitivo V*N P0* PP* “comprárselo”
     - en gerundio V*G P0* PP* “comprándoselo”
   - Doblete de pronombres personales.
     - Pronombres personales (PP*) antepuestos a las formas no-finitas de verbos predicativos (VMI*), pasivos (VSI*) y auxiliares (VAI*) en indicativo y subjuntivo (VMS*, VSS*, VAS*):
       - P* P* ante verbo en formas no-finitas en indicativo (VMI* VSI* VAI*) y en subjuntivo (VMS*, VSS*, VAS*)
     - Pronombres personales (PP*) postpuestos a formas no-finitas como el imperativo:
       - P* P* después de la forma no-finita de imperativo (VMM*, VAM*)
       - P* P* después de las formas finitas de infinitivo (VMN000, VAN000) y gerundio (VMG000, VAG000)

8. Índice de modalidad.
   Se miden adverbios (RG), locuciones (RG*) modales (RG_MD) o predicativas insertadas en las estructuras como adverbios intensificadores, complementos de modo o predicativos.

   - RG ante categorías como AQ*, NC*, VMP*
   - RG* entre verbos VM* RG* VM*, VS* RG* VMP*, VA* RG* VMP*
   - RG* ante el verbo RG* VS*, RG* VM*, RG* VA*, RG* VA* VM*
   - RG* después del verbo VS* RG*, VM* RG*, VA* RG*, VA* VM* RG*

9. Índice de cantidad.
   Se miden los cuantificadores, locuciones determinativas o pronominales indefinidas. Las etiquetas de este tipo de estructuras referidas a la noción de cuantificadores son muy variadas. Según la gramática tradicional, a los vocablos cuantificadores los categorizamos en nombres partitivos (NCMS000_Partitivo), determinantes (DI*), pronombres indefinidos (PI*), numerales (Z*), locuciones con determinantes y pronombres indefinidos combinados (PI0FS000_Combinado) o locuciones preposicionales (SPS_CT) y adverbiales de cantidad (RG_CT), incluso locuciones negativas de cantidad (RN_CT). Como podemos procesar todas estas etiquetas registradas en el fichero de locuciones, es posible obtener este índice cuantificador, aunque no distinguimos las categorías del *PCIC* de estos cuantificadores con las clasificaciones de propios, universales, de grado, focales, etc. A continuación, exponemos algunos ejemplos de etiquetas computables:

   - NCFS000_Partitivo “la mayoría de”
   - NCMS000_Partitivo “resto de”
   - PI0FS000_Combinado “otro cualquiera”
   - PI* “alguno que otro”
   - PT* “cuánto”
   - DI* “algunos”
   - RN_CT “ni siquiera”
   - RG_CT “al menos”
   - RG_MD/CT “a cántaros”
   - RG_CT/MD “a discreción”
   - RG_CT/Coloquial “algo es algo”
   - SPS_LG/TP/CT “alrededor de”
   - SPS_CT “a menos de”
   - Z* para cifras o fracciones y porcentajes (Zp), monedas (Zm), medidas de magnitud (Zu), etc.

10. Índice de deíxis.
    En un texto este índice puede expresarse mediante diferentes categorías:

- PD0NS000 “esto”, “eso”, “aquello”
- PP3NS000 “ello”
- PD3NS0RF “eso_de” (Cervantes, 2006, B1-B2: 183)
- SPS_RF “lo_de” (195)

Algunas deíxis “ad oculos” computables son:

- RG_MD_Deíctico “así_de”
  - RG_MD_Deíctico AQ* “así_de pequeño”
  - RG_MD_Deíctico VMP* “así_de torcido”

Sin embargo, otras, como la deíxis "en fantasma", son construcciones difíciles de detectar sin el contexto semántico en estructuras sencillas formadas por un demostrativo (DD*) y un nombre (NC*):
DD* NC*.

11. Índice de circunstancia.
Este índice es procesable a partir de locuciones adverbiales y preposicionales. Dentro de este indicador se incluyen distintas nociones como la cuantitativa (CT), temporal (TP), modal (MD) o espacial (LG, UB). La concepción espacial y temporal se especifica en el apartado de los campos semánticos de la Nociones Generales del *PCIC* mediante adverbios simples (RG: aquí, ahí, allí; debajo, arriba, encima, cerca, lejos, etc.) y otras categorías gramaticales pertenecientes a estas nociones. Estas categorías son procesables por campos semánticos como términos simples pero no como vocablos a nivel léxico-gramatical de forma distintiva. Es decir, a nivel sintáctico, nosotros podemos medir el número de estructuras listadas como locuciones adverbiales (RG*), preposicionales (SPS*) y conjuntivas cuando estén etiquetadas con el concepto de lugar (LG), ubicación (UB), tiempo (TP), modo (MD), causa (CL), consecuencia (CS), finalidad (FN) o concesión (CC) como por ejemplo:

- Lugar: RG_LG* “de_ahí” , SPS_LG* “cerca_de”, CS_LG* “dondequiera que”
- Tiempo: RG_TP* “ahora_mismo”, SPS_TP* “a finales de”, CS_TP* “a la vez que”
- Ubicación: RG_UB* “al_final”, SPS_UB* “al final de”
- Modo: RG_MD* “a pies juntillas”, SPS_MD* “con forma de”, CS_MD* “tal y como”
- Causa: RG_CL “después de todo”, SPS_CL “en compensación de”, CS_CL “en vista de que”
- Consecuencia: RG_CS, SPS_CS, CS_CS
- Finalidad: RG_FN, SPS_FN, CS_FN
- Concesión: RG_CC, SPS_CC, CS_CC

Un ejemplo de los resultados que se obtienen al procesar algunos de estos índices se observa en las tablas 5.2, y 5.3, 5.4 y 5.5. Como muestra de variedad y complejidad medible (Martín Úriz *et al.*, 2001; Bloor y Bloor, 2004), se han analizado los sintagmas nominales en dos textos seleccionados al azar, un discurso del Rey (1992) y un texto breve redactado por un nativo español adulto (Penpal_1_16) como se ve en la tabla 5.2. De los índices que medimos, los datos más destacables son los porcentajes de palabras contenido y el índice de palabras diferentes, incluyendo el promedio de determinantes en el sintagma nominal. En esta tabla 5.2, en los índices nº 4 y 6 del texto breve, Penpal_1_16, se observa que los porcentajes son mayores que los del discurso de Navidad de 1992. Por el contrario, en el discurso navideño, el valor del índice 9, que representa el promedio de determinantes en el sintagama nominal, es mayor en el discurso navideño (1,55) que en el de Penpal_1_16 (1,00). Por lo que deducimos que, en el estudio de la complejidad sintáctica, unos índices son más significativos que otros, independientemente de la longitud del texto.

Nº	**Concepto**	**Discurso 1992**	**Penpal_1_16**
1	Tokens	1629	150
2	Nº de frases	58	8
3	Nº de palabras/frase	24,91	16,37
4	Nº de palabras contenido	796 (48,86 %)	86 (57,33 %)
5	Nª de palabras función	580 (35,60 %)	34 (22,67 %)
6	Porcentaje de palabras distintas	38,62 %	59,54 %
7	Nº de locuciones	81 (4,97 %)	8 (5,33 %)
8	Nº de Sintagmas Nominales	170 (10,44 %)	14 (9,33 %)
9	Promedio de determinantes en el SN	1,55	1

Tabla 5.2: Índices comparados entre el discurso navideño de 1992 y Penpal_1_16

A continuación, vamos a proceder a analizar los nueve índices anteriores en todos los discursos del Rey (36 discursos) y los textos de *penpal* de dos grupos de estudiantes adultos nativos (43 textos). Este análisis y la comparación nos va a permitir comprobar si los porcentajes de los índices difieren o son significativos. Una vez procesados todos los textos, los navideños y los de *penpal*, cuyos niveles de madurez sintáctica predecimos que difieren entre los dos grandes bloques de textos, comprobamos los valores de los índices representados en las tablas 5.3, 5.4 y 5.5. Concretamente en las tablas 5.4 y 5.5, los índices nº 4 y 6 de todos los textos de *penpal* presentan unos porcentajes también mayores que en los discursos navideños. Por el contrario en la tabla 5.3, el valor del índice 8, que representa el número de sintagmas nominales, y el 9, que indica el promedio de determinantes en el sintagama nominal, son mayores en los discursos navideños (1,63) que en los de Penpal (1,39).

Consecuentemente, por una parte, deducimos que los resultados de los índices 4 y 6 ratifican que una mayor longitud de un texto no implica mayor número de palabras significativas y distintas sino que se repiten más las palabras, y que un texto más elaborado, en este caso los discursos navideños, tiene más elementos por frase y está compuesto por estructuras con más elementos. En concreto, se hace evidente que el número de sintagmas nominales (índice 8) y el promedio de determinantes (índice 9) en los sintagmas nominales está directamente relacionado con el estadio de madurez de un texto.

Nº	Concepto	36 Discursos
1	Tokens	1406 ± 343
2	Nº de frases	49 ± 14 (3.49 % ± 0.53 %)
3	Nº de palabras/frase	21,34
4	Nº de palabras contenido	714 ± 171 (50.88 % ± 1.53 %)
5	Nª de palabras función	38.69 % ± 2.39 %
6	Porcentaje de palabras distintas	496 ± 125 (35.29 % ± 1.46 %)
7	Nº de locuciones	64 ± 19 (4.50 % ± 0.65 %)
8	Nº de Sintagmas Nominales	154 ± 35 (11.08 % ± 0.81 %)
9	Promedio de determinantes en el SN	1.63 ± 0.08

Tabla 5.3: Nueve índices calculados en los discursos navideños del Rey.

Nº	Concepto	23 textos de penpal_2
1	Tokens	155 ± 31
2	Nº de frases	8.7 ± 2.7 (5.61 % ± 1.47 %)
3	Nº de palabras/frase	17,81
4	Nº de palabras contenido	87 ± 18 (55.80 % ± 3.38 %)
5	Nª de palabras función	41 ± 11 (26.15 % ± 3.56 %)
6	Porcentaje de palabras distintas	62.06 % ± 5.54 %
7	Nº de locuciones	8.6 ± 2.6 (5.67 % ± 2.03 %)
8	Nº de Sintagmas Nominales	14.61 ± 4.95 (9.22 % ± 2.05 %)
9	Promedio de determinantes en el SN	1.39 ± 0.20

Tabla 5.4: Nueve índices calculados en los textos de penpal_2.

5.2.3. Índice de estructuras anidadas o criterio de máximo nivel

Como decíamos más arriba, todos los índices, más o menos genéricos, expuestos anteriormente nos dan información cuantitativa de las características de un texto pero no dan un nivel. Por tanto, proponemos otro índice que conjugue los resultados obtenidos a partir del fichero de estructuras.

Proponemos un índice que aglutine varios de los criterios que conforman los índices anteriores a nivel sintagmático ya que tales índices se consideraron para diseñar el fichero de estructuras. En realidad, este índice se configura con dos criterios, el de posición y el de pertenencia. Sirve para calificar el nivel de una categoría gramatical o *PoS*, según su posición en el texto y según su pertenencia a una o varias de las estructuras niveladas.

Esto es, como ya hemos expresado más arriba, se ha creado una colección de estructuras sintácticas y sintagmáticas a las que hemos otorgado un determinado nivel sintáctico, desde A1 a C2. Esta colección de estructuras se localizan en el texto procesado siempre que existan. Al procesar el texto, nos encontraremos que determinadas partes del texto, determinados *PoS*, pertenecen a varias estructuras, y que estas estructuras pueden tener niveles distintos. Denominaremos a estas estructuras "estructuras anidadas o superpuestas". Este hecho nos permite generar un criterio para recalificar los *PoS* del texto. Así pues, definimos el nivel sintáctico de un *PoS* como el nivel de la estructura sintáctica

Nº	Concepto	21 textos de penpal_1
1	Tokens	164 ± 41
2	Nº de frases	8.7 1 ± 3.04 (5.32 % ± 1.37 %)
3	Nº de palabras/frase	18,85
4	Nº de palabras contenido	91 ± 23 (55.65 % ± 3.97 %)
5	Nª de palabras función	43 ± 12 (26.08 % ± 3.30 %)
6	Porcentaje de palabras distintas	60.77 % ± 4.92 %
7	Nº de locuciones	7.7 ± 2.8 (4.71 % ± 1.13 %)
8	Nº de Sintagmas Nominales	14.19 ± 3.42 (8.78 % ± 1.68 %)
9	Promedio de determinantes en el SN	1.33 ± 0.28

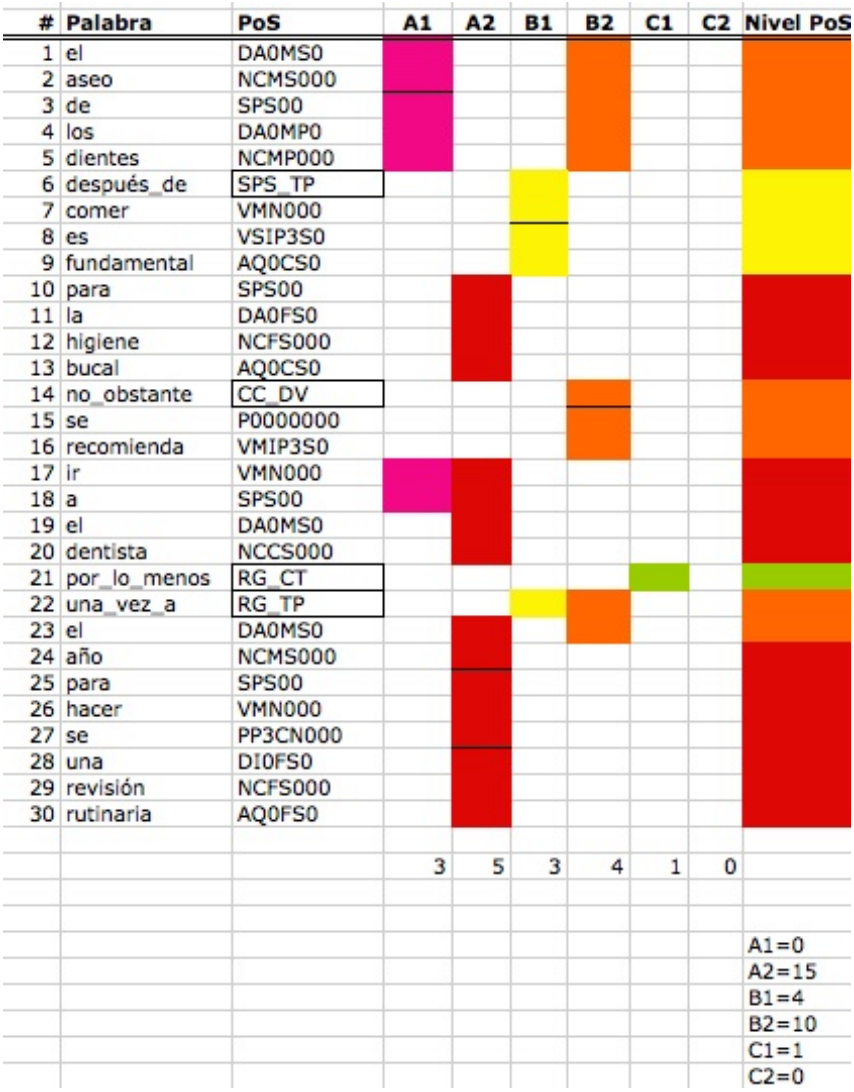
Tabla 5.5: Nueve índices calculados en los textos de penpal_1.

de máximo nivel al que pertenece. Al procedimiento utilizado para calcular el nivel de referencia de un texto, lo denominamos *criterio de máximo nivel*.

En la figura 5.1 se ilustra este criterio de manera gráfica con el texto: *El aseo de los dientes después de comer es fundamental para la higiene bucal. No obstante se recomienda ir al dentista por lo menos una vez al año para hacerse una revisisón rutinaria.* Se lematiza el texto con los módulos correspondientes y se etiquetan los sintagmas con Sintactor aplicando el criterio de máximo nivel.

La imagen 5.1 ejemplifica el proceso automatizado de nivelación de las estructuras y recalificación de los *PoS* anidados o superpuestos de nuevas estructuras en un texto. Explicamos a continuación cómo es el proceso. En un primer paso, una vez etiquetado cada vocablo del texto con su respectivo *PoS* o categoría gramatical, los *PoS* se organizan por estructuras sintagmáticas con sus niveles. Por ejemplo, en la figura 5.1, los siguientes conjuntos de lemas consecutivos (1,2), (3,4,5) y (17,18,19) conforman 3 estructuras cuyos *PoS* tienen un nivel A1. Con respecto a los demás *PoS* y su distribución, obtenemos dos tipos de datos. Por un lado, el número de estructuras niveladas identificadas en el texto conforman 3 estructuras de A1 (en rosa), 5 de A2 (en rojo), 3 de B1 (en amarillo), 4 de B2 (en color naranja), 1 de C1 (en verde claro) y 0 de C2. Por otro lado, continuando con el ejemplo de la figura 5.1, hacemos el recuento del número de *PoS* por niveles, después de aplicar el criterio de máximo nivel. De manera que obtenemos 0 estructuras de A1, 15 de A2, 4 de B1, 10 de B2, 1 de C1 y 0 de C2. En un segundo paso, calculamos dos parámetros normalizados con estos datos. El primero se halla calculando el cociente entre el número de estructuras de un nivel por el número total de éstas en el texto (16). Esto es, el cociente de las estructuras sería para el nivel A1= 3/16, para el nivel A2=5/16, para el nivel B1= 3/16, etc. El segundo parámetro se calcula hallando el cociente entre el número de *PoS* de un nivel por el número total de *PoS*. Es decir, el cociente para los *PoS* sería para el nivel A1=0/30, para el nivel A2=15/30, para el nivel B1=4/30 y así sucesivamente.

A su vez, estos dos parámetros cuantitativos nos permiten obtener un valor único que es el que se identifica con un nivel de referencia concreto, es decir, el que califica el texto.



#	Palabra	PoS	A1	A2	B1	B2	C1	C2	Nivel PoS
1	el	DA0MS0							
2	aseo	NCMS000							
3	de	SPS00							
4	los	DA0MP0							
5	dientes	NCMP000							
6	después_de	SPS_TP							
7	comer	VMN000							
8	es	VSIP3S0							
9	fundamental	AQ0CS0							
10	para	SPS00							
11	la	DA0FS0							
12	higiene	NCFS000							
13	bucal	AQ0CS0							
14	no_obstante	CC_DV							
15	se	P0000000							
16	recomienda	VMIP3S0							
17	ir	VMN000							
18	a	SPS00							
19	el	DA0MS0							
20	dentista	NCCS000							
21	por_lo_menos	RG_CT							
22	una_vez_a	RG_TP							
23	el	DA0MS0							
24	año	NCMS000							
25	para	SPS00							
26	hacer	VMN000							
27	se	PP3CN000							
28	una	DI0FS0							
29	revisión	NCFS000							
30	rutinaria	AQ0FS0							
			3	5	3	4	1	0	
									A1=0
									A2=15
									B1=4
									B2=10
									C1=1
									C2=0

Figura 5.1: Representación de estructuras sintácticas. Al aplicar el criterio de máximo nivel, tenemos los siguientes valores correspondientes al número de estructuras ordenadas por niveles: 0, 15, 4, 10, 1, 0. Contabilizando el número de estructuras por niveles, tenemos los valores: 3, 5, 3, 4, 1, 0.

5.3. Herramienta de análisis sintáctico: Sintactor

Una vez expuesto el índice de calificación explicado más arriba, vamos a procesar un texto. Por ello, se ha diseñado un fichero con una plantilla de estructuras y la herramienta "Sintactor" con vistas a computar la distribución sintáctica y nivelar cualquier texto. Primero, identificamos las estructuras sintácticas niveladas a partir de los criterios del *PCIC* y, luego, calculamos los parámetros cuantitativos de la distribución sintáctica del discurso. Por último, obtenemos el nivel de referencia del texto. Además, el proceso de adjudicación de un nivel a las estructuras sintácticas, cuando se computa con el fichero de estructuras oracionales y sintagmáticas, más los criterios aplicados de nivelación, permitirá otorgar un nivel a cualquier texto en castellano.

Sintactor se conforma a partir de un conjunto de módulos que permiten procesar

en distintas etapas las estructuras sintácticas que tiene cualquier texto. En una primera fase, se procesan las estructuras sintácticas, ya niveladas previamente en el fichero de estructuras. Este fichero recoge 2.586 estructuras diferentes, con sus respectivos niveles, (ver tabla 5.1) que se han generado para esta investigación. No obstante, este fichero es susceptible de implementación y revisiones futuras. Como ya hemos dicho más arriba, dicho fichero se configura mediante diferentes combinaciones de *PoS*, cada conjunto de *PoS* se identifica con un ejemplo, se describen las categorías gramaticales que conforman la estructura y se adjunta el nivel asignado a la estructura sintáctica. Este nivel se adjunta según los criterios de complejidad sintáctica postulados o según la asignación de nivel que otorga el *PCIC* en las distintas etapas de aprendizaje. Mostramos un fragmento del fichero en la sección A.6.1 del apéndice.

A su vez, otro módulo aplica el criterio de máximo nivel. El criterio de máximo nivel permite renivelar las estructuras atendiendo a los *PoS* o categorías gramaticales que aparecen en distintas estructuras con niveles diferentes. El análisis que obtenemos con esta herramienta, después de procesar un texto con las estructuras sintácticas niveladas y después de renivelarlo con el módulo que aplica el criterio del máximo nivel, es el que aparece visualizado en la figura 5.2. Esta figura representa visualmente todo el proceso que realiza "Sintactor" para identificar las estructuras y sus niveles. El color rojo lo identificamos con el nivel A, el naranja con el B y el verde con el C. La distinción dentro de los niveles de 1 y 2 se diferencia en la figura por el menor (1) y el mayor (2) tamaño de la letra. De manera que en el procesamiento del texto, al igual que en esta figura, están representados todos los niveles según el nivel de la estructura renivelada del texto.

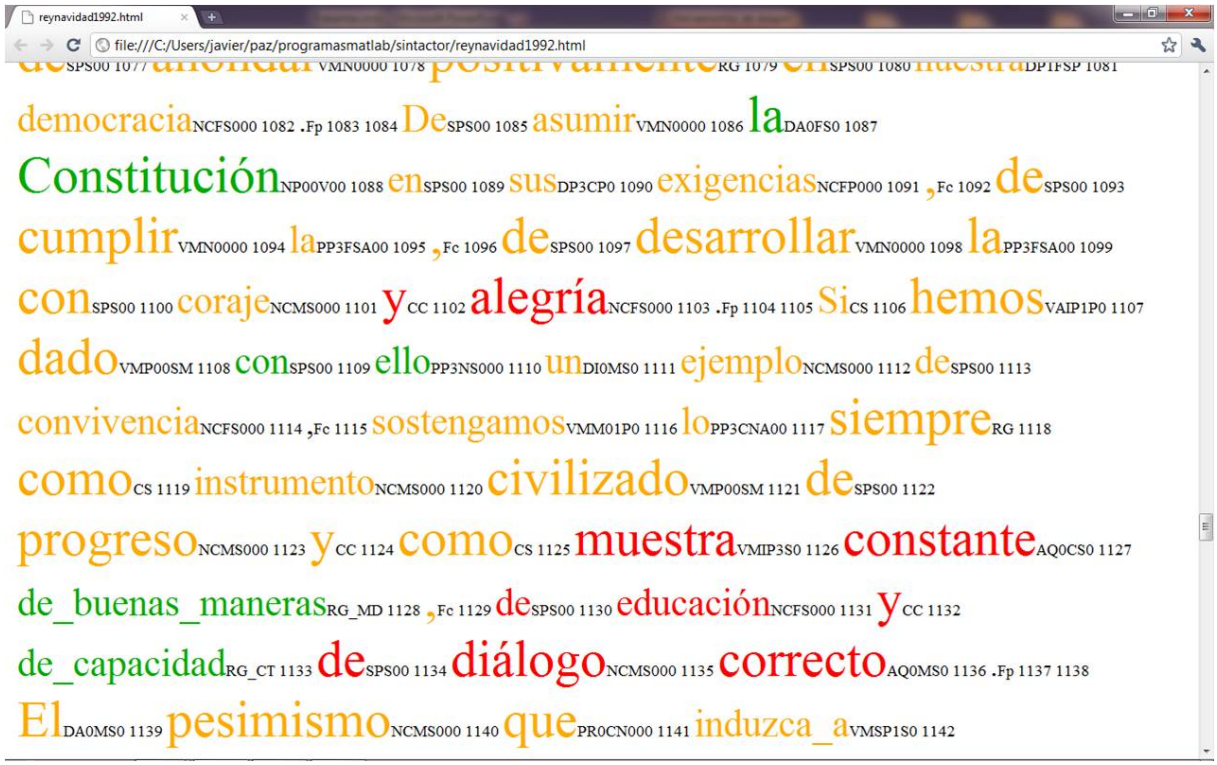


Figura 5.2: Fragmento de estructuras sintácticas niveladas correspondientes al discurso navideño de 1992.

En la figura 5.2 ya se han aplicado dos criterios, el criterio de nivel de referencia y el criterio de máximo nivel, descritos en el apartado 5.2.3, ya que son los criterios que se han utilizado para la nivelación y recuento de las estructuras y para la renivelación de los *PoS* del texto, como veremos a continuación. No obstante, es posible que algún *PoS* de algún texto permanezca sin valoración sintáctica si no ha sido posible incluirlo en ninguna de las estructuras sintácticas contenidas en el fichero de estructuras.

5.4. Método de nivelación sintáctica del texto

A pesar de la cantidad de índices computables a nivel sintáctico que apuntamos anteriormente y recordamos ahora como los 60 índices de Coh-Metrix registrados en la tabla 2.1, o los de nuestra adaptación a Coh-Metrix, en las tablas 3.13, 3.14, 3.15, 3.17, etc., y los índices primarios y secundarios de Hunt y Véliz con la especificación de algunos de estos índices expuestos en el apartado 5.2, nosotros simplificamos el proceso. La computación de tantos índices nos aportarían una serie de números o datos para caracterizar y perfilar un texto. Por su sencillez, optamos por el criterio de máximo nivel para obtener el nivel de referencia de un texto.

5.4.1. Método de la máxima diferencia positiva

Definitivamente, nos hallamos ante el problema básico de esta tesis doctoral: ¿cómo podemos establecer un único valor del nivel sintáctico de un texto? Este problema ya lo resolvimos en el caso del análisis léxico aplicando una variación de la ley de Zipf de distribución de frecuencias de palabras. Veíamos en el análisis léxico que la fiabilidad en la determinación del nivel léxico de un texto era mayor cuanto más largo era un texto. Este mismo hecho se va a poner de manifiesto ahora, con el inconveniente de que las frecuencias de aparición de las estructuras sintácticas van a ser menores que las frecuencias de aparición de lemas. Ello es debido a que las estructuras sintácticas agrupan varios lemas en una sola unidad sintáctica, disminuyendo de esta manera la frecuencia de aparición de una estructura para un número determinado de palabras.

No obstante, una vez niveladas las estructuras, otro módulo de Sintactor es el que nos permitirá establecer el nivel de referencia de la sintaxis de un texto. Puesto que en un texto van a existir siempre estructuras de varios niveles, proponemos que el texto sea nivelado en función de la distribución en niveles de las estructuras identificadas en el texto. En nuestra propuesta, la distribución hallada para un *corpus* de referencia servirá para establecer los niveles. En nuestro caso, se utilizará una vez más el *corpus* de los discursos navideños del Rey como texto modelo de referencia del español para la nivelación sintáctica.

Independientemente del *corpus* de referencia, el procedimiento de nivelación sintáctica de cualquier texto comienza igualmente identificando automáticamente las estructuras sintácticas. Una vez realizada la identificación de estructuras, calculado el número de estructuras en cada uno de los seis niveles y aplicado el criterio del máximo nivel, es posible proporcionar un nivel sintáctico a cada uno de los *PoS* del texto. Este último procedimiento produce una distribución de los *PoS* del texto. De esta manera, obtenemos dos

distribuciones porcentuales. una para cada uno de los dos criterios explicados anteriormente (número de estructuras identificadas según niveles y número de *PoS* según niveles sintácticos después de aplicar el criterio de máximo nivel). Estos dos valores numéricos para cada uno de los niveles se pueden expresar gráficamente, identificando cada nivel con el porcentaje obtenido en cada uno de los dos criterios. De esta manera se obtienen 6 puntos, uno para cada nivel, que representan la distribución de niveles sintácticos de un determinado texto.

Es importante recordar que ambos parámetros para cada nivel no son independientes ya que ambos están ligados a la identificación de estructuras en el texto. Sin embargo, sí que ofrecen información complementaria que puede ser utilizada, por ejemplo, para analizar el grado de anidamiento de las estructuras sintácticas del texto. Además, según se verá a continuación, han permitido la definición de un criterio automático para la nivelación sintáctica de un texto.

En un primer momento, con el fin de validar la utilización de la ley de Zipf, hemos procesado estas frecuencias de aparición de estructuras sintácticas siguiendo el mismo esquema de trabajo y el mismo procedimiento que en el caso del análisis léxico. Lamentablemente, los resultados no han sido tan determinantes ni tan claros como en el análisis léxico. Por ello, hemos abandonado este camino y hemos propuesto otro basado en la ubicación de las distribuciones de aparición de estructuras según sus niveles.

En este punto, nos parece adecuado aclarar que la estructura sintáctica de un texto puede depender de diversos factores (Cervera *et al.*, 2006, 285) tales como la función, la audiencia, el tema del mismo (las estructuras sintácticas de un texto personal o narrativo puede diferir sustancialmente de las usadas para un texto referencial o argumentativo) o de otros factores sociológicos, literarios o jurídicos "pero no afectan de manera esencial a la organización de la información en el texto" (Cervera *et al.*, 2006; Attali y Burstein, 2006; Attali *et al.*, 2010, 338). Sin embargo, consideramos de importancia definir y diseñar un *corpus* de textos de referencia modelo o tipo con el que se comparen los textos que se van a evaluar sintácticamente.

En nuestro caso, hemos elegido como *corpus* de referencia los textos de los discursos navideños del Rey Juan Carlos I, por tener una sintaxis modelo y servir de "gold standard corpus" (Zaanen *et al.*, 2004). Se trata de 36 discursos bien construidos que contienen una amplia variedad de estructuras sintácticas. Al procesar cada discurso navideño hemos obtenido una representación gráfica de seis "nubes de puntos" asociados a cada uno de los niveles de referencia. Esto es, cada discurso se disecciona por estructuras de nivel representadas por puntos que, a su vez, se reagrupan en una región (nube) con un nivel determinado como se aprecia en la figura 5.3. Este hecho nos ha indicado una gran homogeneidad sintáctica en todos los discursos navideños ya que contienen propociones similares de estructuras en cada uno de los niveles de referencia, excepto en los niveles C1 y C2 por ser las estructuras registradas con menor número en nuestro fichero de estructuras. Como proponemos que el *corpus* navideño tiene un nivel sintáctico C1, lo hemos tomado como texto nivelado de referencia para nivelar otros textos de menor o igual nivel que el de referencia.

Una vez observado el comportamiento de agrupación que se refleja en la figura 5.3, hemos definido para estos discursos dos parámetros: un valor medio (simbolizado por un

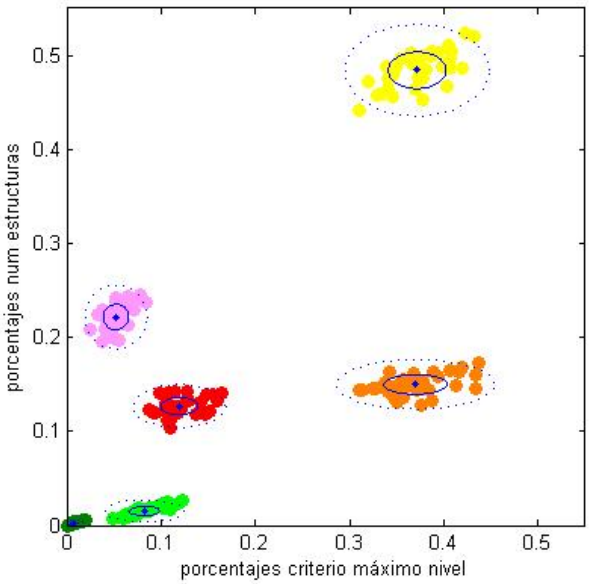


Figura 5.3: Distribución de estructuras por niveles de referencia en los discursos navideños del Rey Juan Carlos.

punto central) y un rango de incertidumbre (simbolizado por una elipse). Si el punto medio expresa la localización promedio de las distribuciones estadísticas de las estructuras sintácticas para cada nivel, el área de incertidumbre expresa la dispersión de los valores obtenidos para cada uno de los discursos individuales. Ambos parámetros serán claves a la hora de definir el nivel sintáctico de un texto en comparación, en este caso, con el *corpus* de referencia de los discursos navideños. Se señala, además, un umbral de calificación (simbolizado por un círculo de puntos discontinuos) para decidir si la calificación es fiable o no, dependiendo de la ubicación del texto calificado con respecto a ese umbral de calificación.

A partir de aquí, contando con el criterio de nivelación y el *corpus* de referencia, empezamos a nivelar sintácticamente un determinado texto. Como ejemplo, valga la carta de Penpal_2_16 escrita por un estudiante español nativo, cuyo texto original y fragmentos procesados documentamos en las figuras A.6, A.7 y A.8 en el apartado A.7 del apéndice. Primero, se procesa el texto para identificar sus estructuras y obtener las dos distribuciones estadísticas correspondientes a la aplicación de los dos criterios de reparto en niveles: el recuento de estructuras sintácticas por niveles (independiente de la anidación de estructuras sintácticas) y el recuento de los niveles a partir del *PoS*, aplicando el criterio de máximo nivel (que depende de la anidación de las estructuras sintácticas).

Los valores calculados de las distribuciones se comparan ahora con los valores medios y el nivel de incertidumbre obtenidos al procesar el *corpus* de referencia. En este punto, es importante subrayar que, cuando un determinado nivel se aleja en una dirección del valor de referencia (por ejemplo, aumentando su proporción), otro u otros niveles se han de alejar en dirección contraria (por ejemplo, disminuyendo su proporción), ya que entre todos han de sumar el 100 % de las estructuras o *PoS* nivelados. Por tanto, el aumento de la proporción de un nivel será un indicador del nivel predominante en el texto, siempre

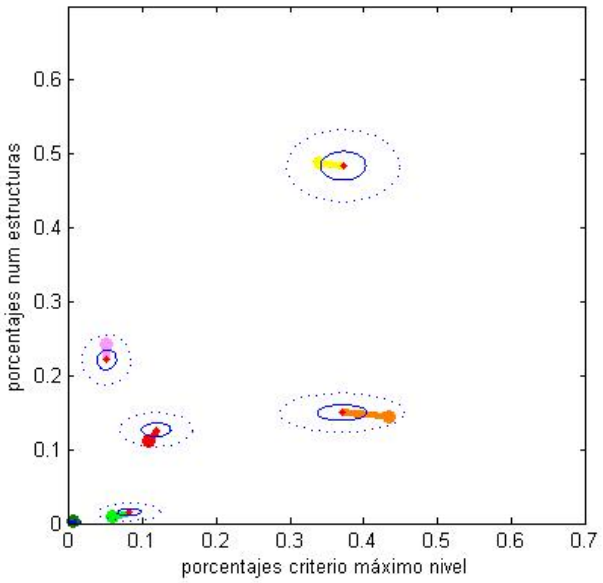


Figura 5.4: Localización de las estructuras sintácticas del discurso navideño del Rey de 1986 respecto al *corpus*.

respecto a las proporciones observadas para el texto de referencia. Esto nos indica que el comportamiento que debemos identificar como calificador es el de aquel punto-vector que nos sitúe a la derecha y hacia arriba en el cuadrante (0^{o}-90^{o}) con respecto a la posición del punto de ese *nivel* en el corpus de referencia, como se observa en las figuras 5.6 y 5.7.

Una simple aplicación de este criterio se fijaría únicamente en la distancia entre los dos puntos, el punto medio de referencia y el punto de incertidumbre de un nivel determinado. Es más, este procedimiento difícilmente podría identificar textos de niveles superiores C1 o C2 ya que su proporción es muy baja y, por tanto, las distancias implicadas también muy pequeñas. Para compensar este hecho hemos decidido incluir un procedimiento de normalización del valor de la distancia.

Esta distancia se va a medir en unidades del rango de incertidumbre, que es distinto para cada nivel. Por ello, se han expresado en la gráfica unas áreas-rangos de incertidumbre alrededor de los puntos asignados a cada nivel. El tamaño de las áreas-rangos tiene que ver con la incertidumbre en el valor de la media para cada nivel, como se aprecia en las figuras 5.4 y 5.5. Estas regiones de incertidumbre se han definido mediante el cálculo de "la desviación cuadrática media" en cada uno de los niveles obtenidos en el conjunto de textos que componen el *corpus* navideño.

Mediante este procedimiento se obtiene una serie de vectores-nivel, cuya dirección indica la posición de los valores para cada uno de los niveles de un determinado texto y cuyo tamaño señala cuánto nos separamos de la posición correspondiente al valor medio obtenido con el *corpus* de referencia como se observa en las figuras 5.6 y 5.7.

Utilizando estos vectores, definimos el nivel sintáctico del texto como aquel nivel correspondiente al vector de mayor longitud que se sitúa en el primer cuadrante de la representación gráfica. Es decir, aquel vector que apunta hacia la derecha y hacia arriba

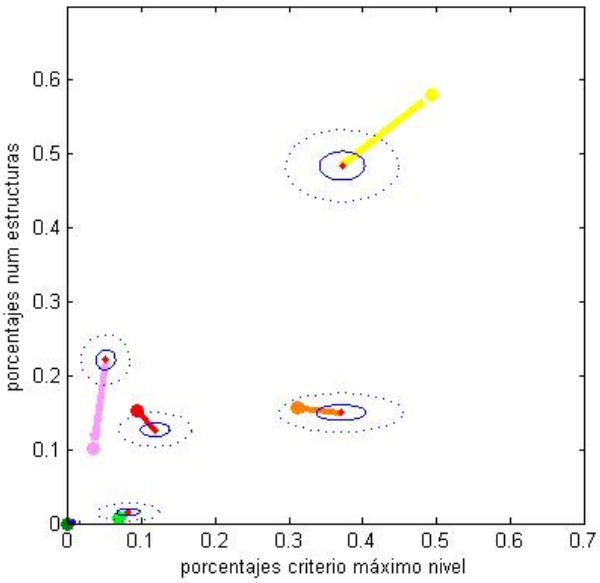


Figura 5.5: Localización de las estructuras sintácticas de la carta de "Penpal_2_16".

y que tiene la mayor longitud. Tal como se observa en la figura 5.6, el nivel sintáctico de referencia de este texto procesado sería de un B1 (vector amarillo).

En casos en los que aparece más de un vector en el primer cuadrante, un análisis más detallado de la posición de los vectores para cada uno de los niveles nos permitiría afinar más la determinación del nivel, aumentando o disminuyendo el valor del nivel sintáctico obtenido con el criterio anterior.

Llegados a este punto, cabe plantearse la siguiente situación: ¿qué nivel tendría un texto que estuviese muy cercano a los valores medios obtenidos para el *corpus*? Este caso será de especial relevancia en función de la calidad del *corpus* considerado. En el caso que nos ocupa, nuestro *corpus* de referencia, los discursos navideños del Rey, podría calificarse sintácticamente como de un nivel C1. Por esta razón, un texto que coincidiese con el valor medio del *corpus* no debería tener un nivel distinto al C1.

Para resolver esta situación hemos definido un nuevo parámetro que actúa como umbral para definir el nivel. Este parámetro se basa en el valor de la longitud de "los vectores diferencia normalizados" definidos previamente. Cuando estos vectores tienen un tamaño inferior al umbral, consideramos que no es posible realizar la calificación sintáctica y rehusamos proporcionar un número de forma automática. Es más, la superación de este valor umbral puede considerarse como un parámetro de fiabilidad del procedimiento de calificación. No obstante, a pesar de que esta situación no permite calificar con fiabilidad el nivel sintáctico, nos aventuramos a aplicar el mismo procedimiento para calificar un discurso. En el caso mostrado en la figura 5.7, el nivel del texto sería C2 ya que el vector correspondiente es el de mayor longitud en el interior del primer cuadrante.

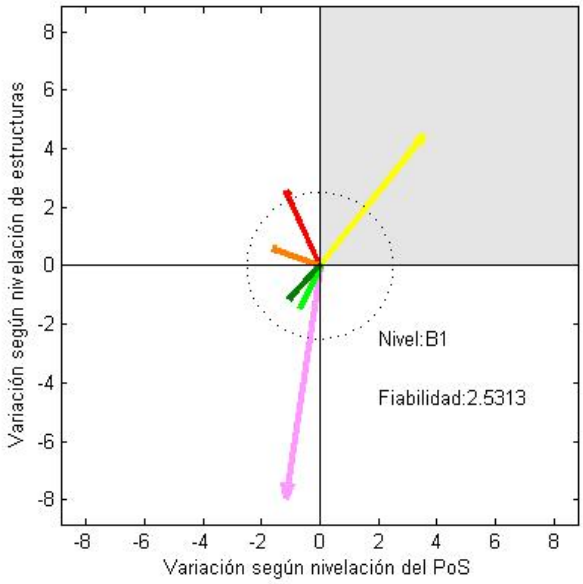


Figura 5.6: Vectores de diferencia normalizados de la carta de "Penpal_2_16". El cuadrante de interés está sombreado en gris. El círculo indica el umbral de calificación. Los colores de los vectores están codificados por niveles como en gráficas anteriores.

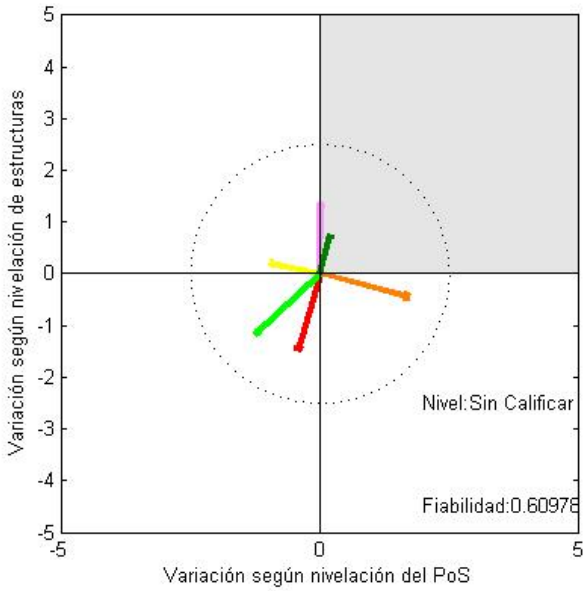


Figura 5.7: Vectores de diferencia normalizados del discurso navideño de 1986.

Capítulo 6

Análisis semántico

6.1. Evaluación semántica

Los métodos que presentamos para el análisis semántico de los textos se diferencian porque unos son más precisos y otros más genéricos. En cada uno de los apartados de este capítulo se muestran estos métodos concebidos para la extracción temática de textos dentro de la práctica conocida como minería o extracción de datos.

6.1.1. Análisis Semántico Latente

Mediante este método se podrá detectar si el contenido del texto analizado está más o menos "próximo" semánticamente al contenido esperado. Este método, explicado en el apartado 3.2.3, tiene una amplia aplicabilidad en otras áreas como la música, la imagen, las comunicaciones, la medicina, etc. ya que permite realizar estudios estadísticos o comparativos de datos. También en lingüística es muy productivo. El investigador que lo ha utilizado profusamente en estudios de lingüística computacional ha sido el Dr. Landauer (Landauer *et al.*, 1998a,b, 2000, 2003, 2004, 2009) con brillantes resultados y aplicaciones. Otros grupos aplican este método matemático del LSA también para calcular el contenido semántico de textos y agruparlos en *clusters* (Ziempekis y Gallopoulos, 2006) con el fin de clasificar textos por el contenido (Olmos *et al.*, 2009), detectar correlaciones entre los textos o diferencias en el contenido como se hace con su homólogo Latent Semantic Indexing (LSI) (Kontostathis y Pottenger, 2006).

En nuestro caso, este método forma parte del módulo Semantor, herramienta desarrollada para aplicar este modelo estadístico (ver apartado 6.3).

6.1.2. esWordNet

Para el análisis semántico, el glosario esWordnet 1.6 es útil porque se puede detectar cierto grado de coherencia léxica en el texto. Este grado lo vamos a identificar al procesar cada lema y rastrear, si existe, alguna relación semántica entre lemas. Es decir, una de las

características del glosario esWordnet es que muchos lemas tienen entre sí relaciones de hiperonimia, meronimia y sinonimia. Por ello, si en el texto se dan alguna de estas relaciones, podremos detectarlo. Además, creemos que es un método complementario al Análisis Semántico Latente porque nos informa de las relaciones semánticas que se establecen entre los lemas.

Este método que aplicamos a nivel léxico se basa en las características implícitas de este diccionario. Según las característica descritas en el apartado 3.1.5 permite detectar relaciones taxonómicas. Sin embargo, la falta de disponibilidad de un algoritmo que establezca de forma rápida las relaciones entre lemas y frases utilizando esWordnet hace poco rentable este método ya que se precisa mucho tiempo para hallar la similitud o relación entre los vocablos de un texto. Esto supone que, aunque los resultados obtenidos son significativos, no hayamos procesado apenas textos con este método. Además, la inexistencia en esWordnet 1.6 de relaciones, vocablos y categorías que sí se han implementado en otras lenguas y glosarios, como la inclusión de adverbios, es un inconveniente que impide establecer mayor número de relaciones en los textos analizados. Es en 2009, en el contexto de la investigación sobre la traducción y desarrollo de sucesivos proyectos de investigación como KYOTO, KNOW, KNOW2 del grupo IXA, en el que intervienen la Universidad del País Vasco y la Universidad Politécnica de Cataluña, cuando se ha desarrollado un algoritmo que permite establecer, de forma eficiente y precisa, las relaciones semánticas entre palabras atendiendo al esquema general de Wordnet (Agirre y Soroa, 2009; Agirre *et al.*, 2009b,a, 2010). De hecho, FreeLing ha incorporado esta aplicación en la versión FeeLing 2.1 (Padró *et al.*, 2010b) y versión 2.2. Dicho algoritmo no lo aplicamos en este trabajo porque hemos trabajado con versiones más antiguas, FreeLing 1.5 y esWordnet 1.6.

No obstante, presentamos nuestra propuesta para hallar el grado de parentesco entre los lemas de un texto. Por una parte, en la figura 6.1, se observan las frecuencias de aparición de seis grados de parentesco para el discurso navideño del Rey de 1979. Además, se ha añadido una estimación de carácter exponencial que permite deducir que el considerar más grados de parentesco resulta mucho más costoso computacionalmente.

Por otra parte, en las figuras 6.2 y 6.3, representamos las relaciones que establece esWordnet en un texto. Los lemas están unidos con líneas azules que intersectan en el punto rojo al nivel de parentesco identificado. Las líneas son más gruesas conforme el parentesco es más cercano. Con el fin de clarificar la representación, en la figura 6.2, sólo mostramos un fragmento del discurso navideño de 1979, esto es, treinta lemas nominales, verbales y adjetivales consecutivos. A modo de detalle, para mejorar la visualización, en la figura 6.3, seleccionamos y representamos las relaciones de parentesco de un conjunto de 15 lemas consecutivos extraídos del discurso navideño del Rey de 1979.

Las relaciones se establecen en distintos grados. Por ejemplo, a partir de la figura 6.3 de un fragmento del discurso navideño del Rey, se halla una relación de sinonimia (preocupación-inquietud) en grado 1, de meronimia (participar-compartir) en grado 2, de hiponimia (hablar-participar) en grado 4, de cierta antonimia con diferente categoría gramatical (mismo-complicación) en grado 5, etc. Para las relaciones de grado 6 resulta más complicado especificar el tipo de relación. Sin embargo, si se creara el algoritmo correspondiente, se podría definir el tipo de relación taxonómica entre los lemas también

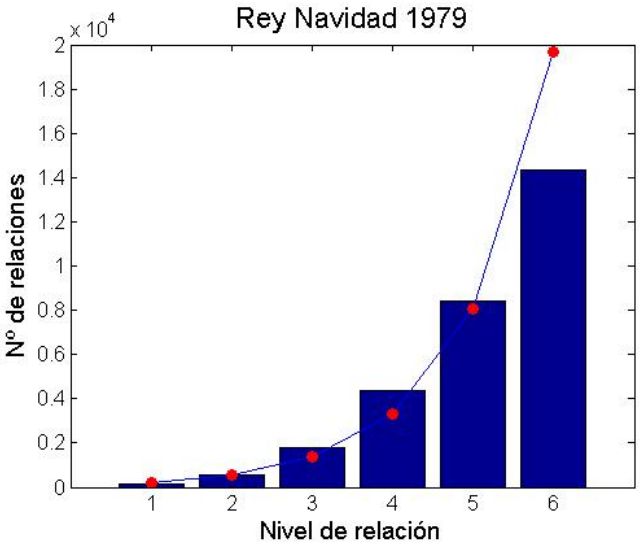


Figura 6.1: Histograma de seis grados de parentesco para el discurso navideño del Rey de 1979. La línea que une los puntos rojos expresa un ajuste exponencial de los valores del histograma en función del grado de parentesco.

de forma automática, ya que está especificada en la base de datos de esWordnet (ver tablas 3.10 y 3.11), y convertirse dicho tipo de relación en otro índice computable.

6.1.3. Campos semánticos del *PCIC*

Un vez más, el "Índice de nociones generales y nociones específicas" del *PCIC* es útil porque organiza el vocabulario en campos semánticos catalogando el léxico según pertenezca a una noción general o a una noción específica. Mediante un índice numérico que acompaña a cada vocablo, se ubica el léxico en un campo semántico determinado. Estos campos semánticos son de utilidad como referencia semántico-contextual para identificar qué tipo de nociones se expresan en un texto. Un listado somero de los campos, subcampos y subsubcampos semánticos que especifica el *PCIC* se exponen en el apéndice, apartado A.5.

Para procesar el "Índice de nociones generales y nociones específicas" del *PCIC*, se ha creado un módulo computable que recoge las siglas NG (Nociones Generales) y NE (Nociones Específicas), siglas que acompañan a los vocablos y van seguidas de las cifras numéricas que identifican los campos semánticos. Además, este módulo de identificación de los campos semánticos implementa una aplicación más de la herramienta Semantor. De esta manera, si el módulo del LSA y el del área temática de la Dra. Fuensanta López en Semantor nos informa de que un texto está o no fuera de tema, el módulo del campo semántico nos informará de qué trata el texto.

A continuación, en las figuras 6.4, 6.5, 6.6, 6.7 y 6.8, se muestran gráficamente los campos semánticos en los que se distribuyen los lemas extraídos de los textos analizados, según la clasificación del "Índice de nociones generales y nociones específicas" del *PCIC*. La figura 6.4 recoge 8.662 lemas del "Índice" del *PCIC* distribuidos en aquellos campos semánticos que han especificado los expertos del Instituto Cervantes. En la figura

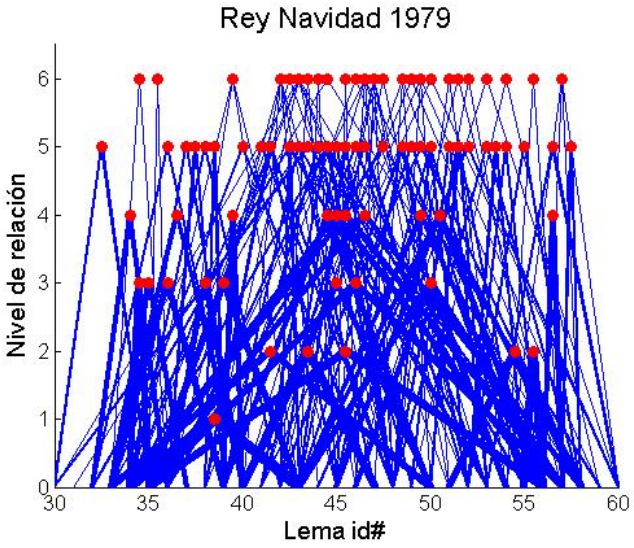


Figura 6.2: Número de relaciones genealógicas en el discurso de 1979.

6.4 se muestran incluso lemas repetidos cuando tienen o pertenecen a más de un campo semántico. Es decir, existen lemas que se repiten bien por tener niveles diferentes o bien por pertenecer a campos semánticos diferentes. Sirva de ejemplo el lema "bajo". "Bajo", desambiguado, tiene dos categorías gramaticales: preposición y nombre, y pertenece a tres campos semánticos distintos: NG A1 2.6.3; NG C1 3.3 y NE C2 18.2, es decir, al campo semántico de nociones cuantitativas, espaciales y artístico-musicales, respectivamente. Con nuestro método de procesamiento, al tener el vocablo tres campos semánticos, lo computamos como tres lemas distintos. Matizamos que, a pesar de que hemos etiquetado morfológicamente los vocablos del "Índice de nociones generales y nociones específicas" del *PCIC* para desambiguar, el procesamiento mejoraría si se sometiese el "Índice" electrónico del *PCIC* a una revisión manual detallada.

Un ejemplo de procesamiento de texto con este módulo del "Índice de nociones generales y nociones específicas" del *PCIC* es el discurso del Rey en la Navidad de 1992 que muestra la figura 6.5.

Además, para obtener una buena representación de algunos campos semánticos en un texto utilizando el "Índice de nociones generales y nociones específicas" del *PCIC*, hemos creado una *stoplist*. Con esta *stoplist*, obtendremos unas frecuencias más reales y evitaremos el "ruido" que producen las repeticiones de estos lemas en distintos campos cuando procesemos los textos. Esta *stoplist* la conforman todas las preposiciones simples (a, ante, bajo, etc.), el determinante "el" y los verbos auxiliares "ser, estar y haber". Las figuras que se presentan a continuación muestran un estudio de los campos semánticos en el discurso del Rey de la Navidad de 1992 donde se aplican sucesivamente los siguientes cambios: eliminación de los lemas repetidos, aplicación de la *stoplist* y la ponderación de un lema que pertenezca a uno u otro campo semántico.

Podemos observar que en la figura 6.5 se computan los 558 lemas que tiene el texto, sin aplicar la *stoplist* y con la repetición de lemas. A continuación, de los valores de cada campo, expresamos entre paréntesis el promedio de repetición de los lemas después de aplicar la *stoplist*.

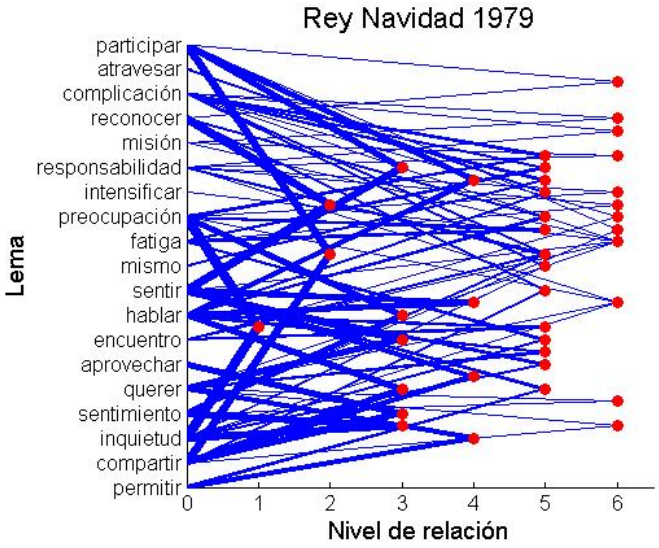


Figura 6.3: Relación genealógica o de coherencia de un fragmento del discurso de navidad de 1979.

En la figura 6.6 aparece representado el mismo discurso con todos los lemas sin incluir los repetidos.

Continuando con un proceso de refinamiento, en la figura 6.7, se han eliminado los lemas de la *stoplist* y, una vez más, se han eliminado las repeticiones para obtener unos resultados semánticos distintivos, ya que lo que buscamos es saber de qué trata el texto.

Siguiendo con este proceso de mejora, otro de los aspectos que se ha considerado en el análisis semántico es la repetición de un lema cuando pertenece a diferentes campos semánticos, tal y como hemos visto más arriba con el ejemplo del vocablo "bajo". Tal repetición representa la ubicación de un mismo lema, desambiguado o no, en distintos campos semánticos. Esto supone que los lemas repetidos se distribuyen en los campos sin afinar el campo semántico al que se refiere el texto. Para ello, hemos aplicado el "criterio de ponderación semántica" al módulo que asigna el campo semántico a un vocablo. Esto es, la asignación de los lemas a un campo semántico se aplica utilizando un factor de ponderación que es directamente proporcional a la distribución de otros lemas a campos semánticos en una unidad lingüística (frase) e inversamente proporcional al número de campos semánticos asignables a un determinado lema.

Este criterio se va a aplicar únicamente dentro de una frase de manera que, a aquel lema etiquetado con varios campos semánticos en el "Índice" del *PCIC*, se le asignará aquel campo al que se refiera otro lema incluido en un campo semántico coincidente con el de otro lema incluido en dicha frase. Es decir, con este criterio afinamos la asignación semántica de los lemas en el co-texto de la frase. La figura 6.8 muestra el resultado de aplicar a este texto el "criterio de ponderación semántica", lo cual nos permite configurar un parámetro que marque el grado de proximidad a la temática deseada y, por tanto, la contextualización y coherencia del texto analizado. Como se observa en la figura 6.8, los campos que aumentan en el discurso del Rey, tras la ponderación del léxico, son los esperados: gobierno, política y sociedad. Efectivamente, los discursos del Rey tratan de cuestiones de gobierno, política y sociedad (marrón) en el territorio español y fuera de él

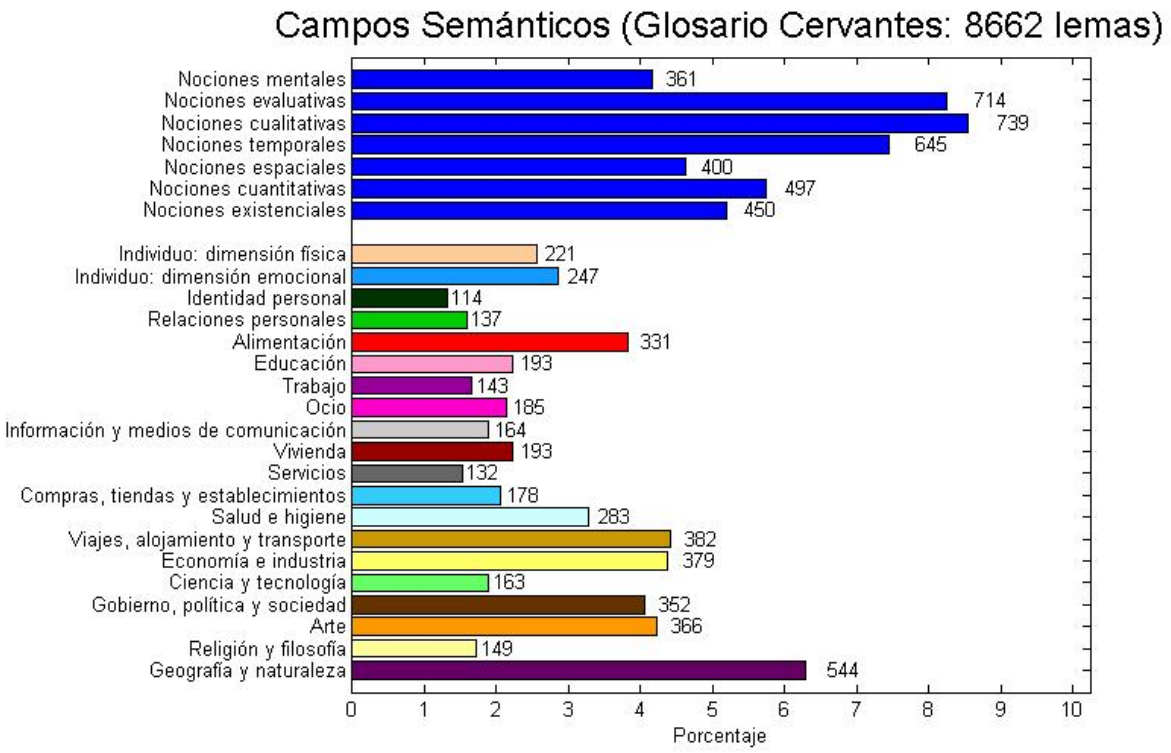


Figura 6.4: Campos semánticos del "Índice" del *PCIC*.

(morado). Muchos de los discursos navideños del Rey tienen una parte afectiva, personal y de interés humano (azul turquesa y verde oscuro). También destacan su preocupación e interés por el empleo (malva) y el ocio (fucsia), la vivienda (granate), los servicios a los ciudados (gris oscuro) y la educación (rosa claro).

Cuando leemos el discurso de Navidad de 1992, comprobamos que es el año de la Exposición Universal de Sevilla y los Juegos Olímpicos de Barcelona, la II cumbre Hispanoamericana y la necesidad de una Europa que apoye a Sudamérica, la Conmemoración del V Centenario del descubrimiento de América y la capitalidad europea de la cultura de Madrid. También se recuerda la Conferencia de Paz para Oriente Medio, la guerra en Europa (de Bosnia) y las víctimas y temores del terrorismo. Aún más, el Rey alude a la xenofobia e insta a ser solidarios con los extranjeros que han decido vivir con nosotros y a ser conscientes de nuestro patrimonio natural después del desastre [del Prestige] en Galicia.

En resumen y en relación a los datos extraídos automáticamente y una vez leído el discurso, anticipábamos que el discurso trataba de asuntos de gobierno, política y sociedad (marrón), nuestra actividad cultural y actitud con el resto de países y con la naturaleza (morado). Como la mayoría de españoles sabe, los discursos navideños del Rey tratan sobre varios temas que el Rey repasa dotándolos de cierta sensibilidad, compromiso personal, solidaridad, incluso espiritualidad (azul turquesa) y humanidad (verde).

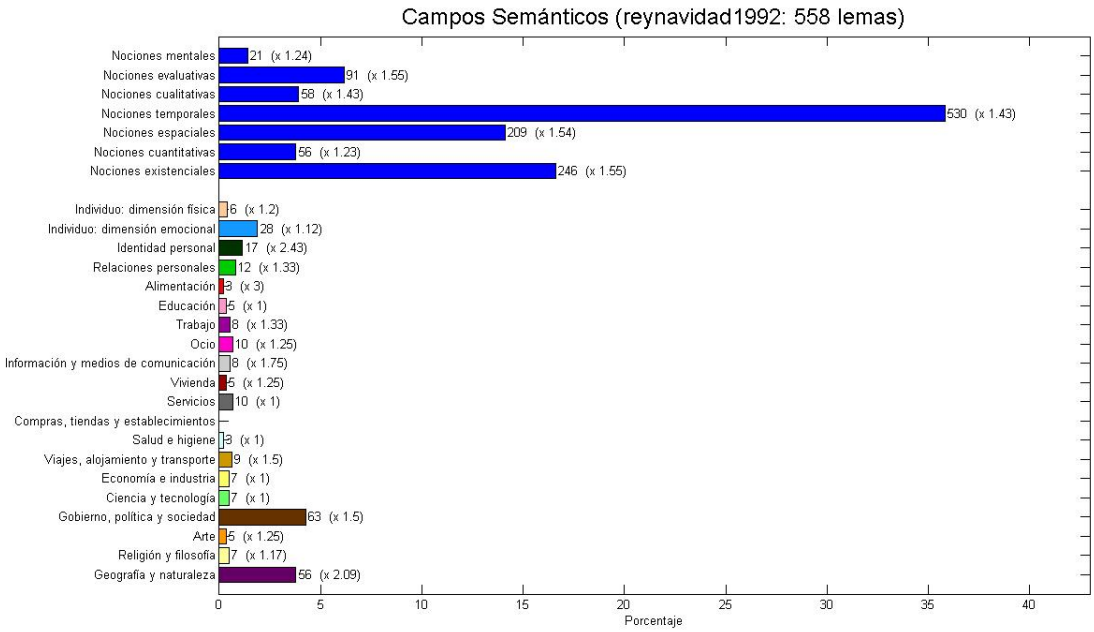


Figura 6.5: Distribución de campos semánticos del discurso navideño del Rey de 1992 con todos los lemas, incluidos los repetidos.

6.1.4. Áreas temáticas de la Dra. Fuensanta López

Uno de los criterios básicos para diagnosticar semánticamente un texto es poder identificar que el texto trata del tema que esperamos. Para ello, a partir de los datos del glosario de la Dra. Fuensanta López, se ha desarrollado un módulo de identificación semántica con el que se identifica el área temática a la que pertenece el texto objeto de análisis.

Por una parte, y en primer lugar, mostramos la figura 6.9 donde se representan todos los lemas del glosario de la Dra. Fuensanta López distribuidos por áreas temáticas. Recordamos que este glosario sólo indexa lemas nominales y verbales, y que la distribución de lemas por áreas es desigual. Es menor el número de lemas en la sección de Matemáticas y mayor en el área de Lengua y Literatura. El punto rojo que aparece en la figura 6.9 corresponde a todo el glosario de la Dra. Fuensanta López y muestra el sesgo de este glosario. Sin embargo, el punto rojo en las figuras 6.10, 6.11, 6.12 y 6.13 es la ubicación del texto analizado en su área correspondiente. Por ello, para ajustar los textos mejor a su área temática, eliminando este sesgo, se ha aplicado el "criterio de ponderación de área" ya que las áreas constan de lemas que comparten los cuatro sectores. En definitiva, este criterio es un factor de corrección inversamente proporcional a la distribución de áreas del glosario de la Dra. Fuensanta López. Es decir, cuanto más lemas existen en una determinada área menos peso van a tener los lemas de esa área.

En segundo lugar, se presentan cuatro figuras más concretas (6.10, 6.11, 6.12 y 6.13) que muestran una selección de cuatro textos distribuidos en su cuadrante según la temática correspondiente a cada una de las cuatro áreas temáticas generales a las que pertenecen los lemas de dichos textos. Estas áreas son "matemáticas", "ciencias naturales", "ciencias sociales" y "lengua y literatura". En todas las gráficas que se muestran a continuación, el

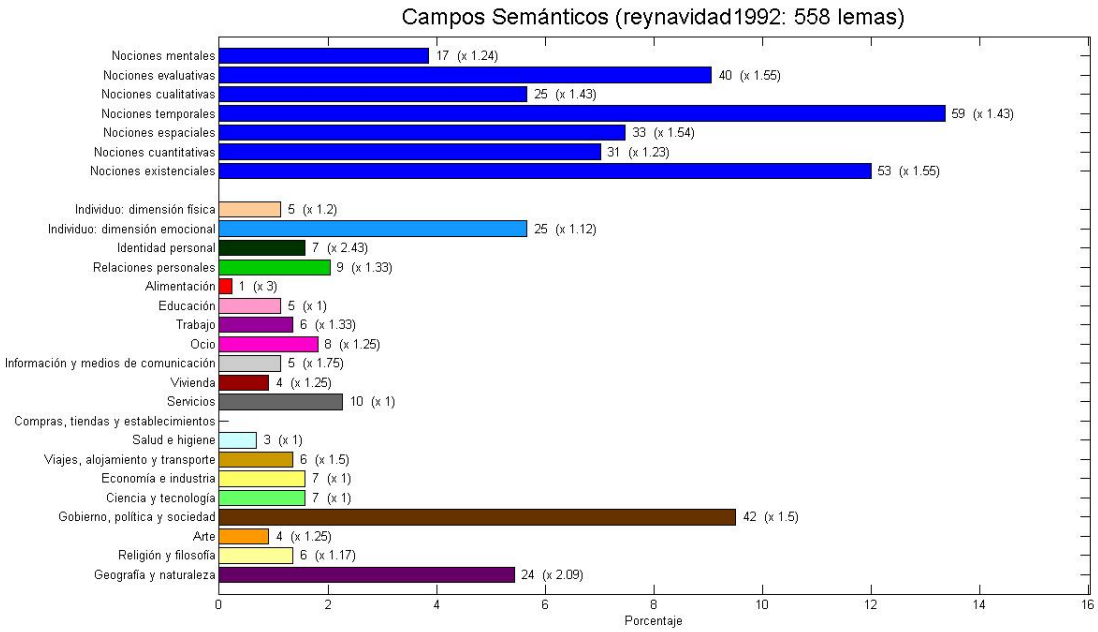


Figura 6.6: Distribución de campos semánticos del discurso navideño del Rey de 1992 con todos los lemas, sin incluir los repetidos. Los números entre paréntesis indican el factor medio de repetición de cada lema en cada uno de los campos semánticos

punto rojo es el texto analizado y ubicado automáticamente en el área temática correspondiente.

Consideramos que este "criterio de ponderación de área" funciona muy bien con los textos procesados. La elección de textos se ha hecho en función de las cuatro áreas temáticas que pueden distinguirse y de un texto extenso sin imágenes. El texto de matemáticas trata sobre vectores, el de Ciencias Naturales sobre el paisaje, el de Lengua y Literatura sobre los géneros literarios, y el de Ciencias Sociales sobre la Revolución Francesa. Todos ellos se han extraído del banco de materiales didácticos del Instituto de Tecnologías Educativas mientras que el de Ciencias Sociales procede de la Wikipedia.

Por otra parte, como aplicación del método a un discurso conocido, sometemos a este método al discurso del Rey de 1992. Como se observa en la figura 6.14, y como es de esperar, el discurso se ubica entre el cuadrante de Lengua y Literatura y de Ciencias Sociales.

6.2. Definición de índices semánticos

Si resultaba ardua la tarea de hallar un valor numérico o índice cuantitativo que mostrase el nivel léxico y sintáctico de un texto, no lo es menos para el nivel semántico. A continuación, proponemos índices medibles susceptibles de marcar el perfil o proximidad del contenido semántico esperado en un texto:

- Índice de coherencia textual. Se calcula hallando la correlación de las frecuencias de

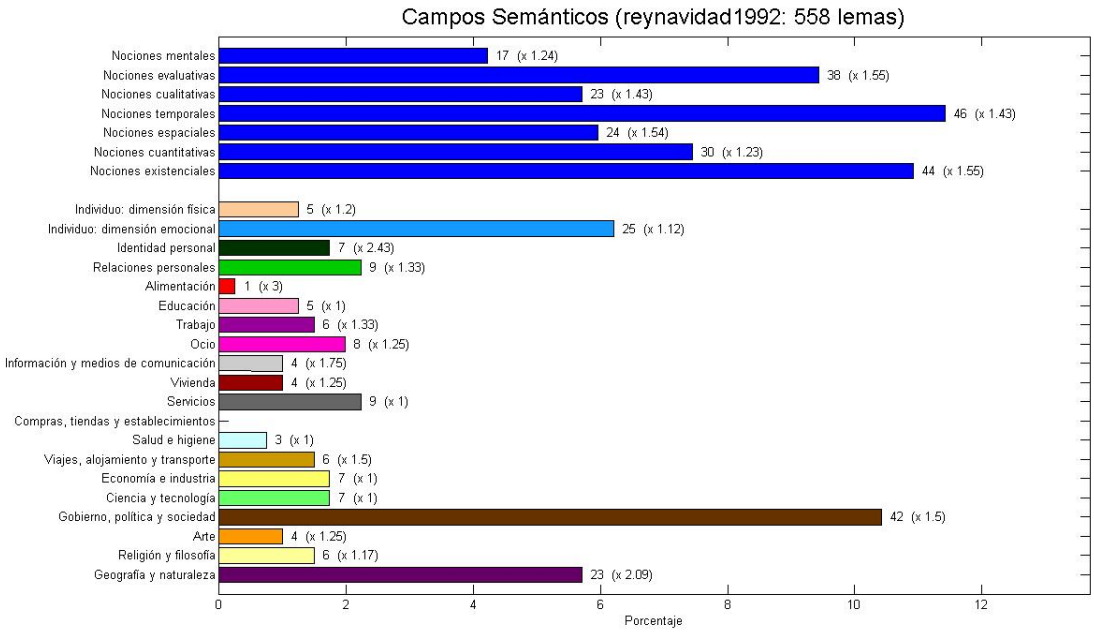


Figura 6.7: Distribución de campos semánticos del discurso navideño del Rey de 1992 con todos los lemas sin incluir los repetidos y aplicando la *stoplist*.

aparición de los lemas-contenido del texto.

- Índice de similitud del texto analizado con el *corpus* de referencia o texto sintético. Se obtiene un mapa de identificación semántica susceptible de comparación y medida cuando se somete un texto al Análisis de Componentes Principales o Análisis Latente Semántico.

- Índice de relación genealógica entre los lemas contenido del texto. Se puede obtener el grado de relación existente entre dos nombres, adjetivos o verbos dentro de la unidad de análisis (frase, párrafo, texto, etc.). Este índice se basa en el grado de relación entre los lemas al establecerse conexiones de sinonimia, antonimia, hiperonimia, meronimia, etc. que se especifican en el glosario de esWordnet, en su archivo de *relations*.

- Índice de coherencia semántica basado en los campos semánticos del "Índice general de nociones generales y específicas" del *PCIC* y del glosario de áreas temáticas de la Dr. Fuensanta López.

6.3. Herramienta de análisis semántico: Semantor

Esta herramienta, al igual que Lexicator y Sintactor, es un conjunto de módulos algorítmicos que compilan los criterios y los índices definidos en los apartados anteriores para indicarnos el grado de contenido temático esperado en un texto. ¿Cómo indicará Semantor ese grado de contenido? Podrá identificar un texto con un grado semántico bajo,

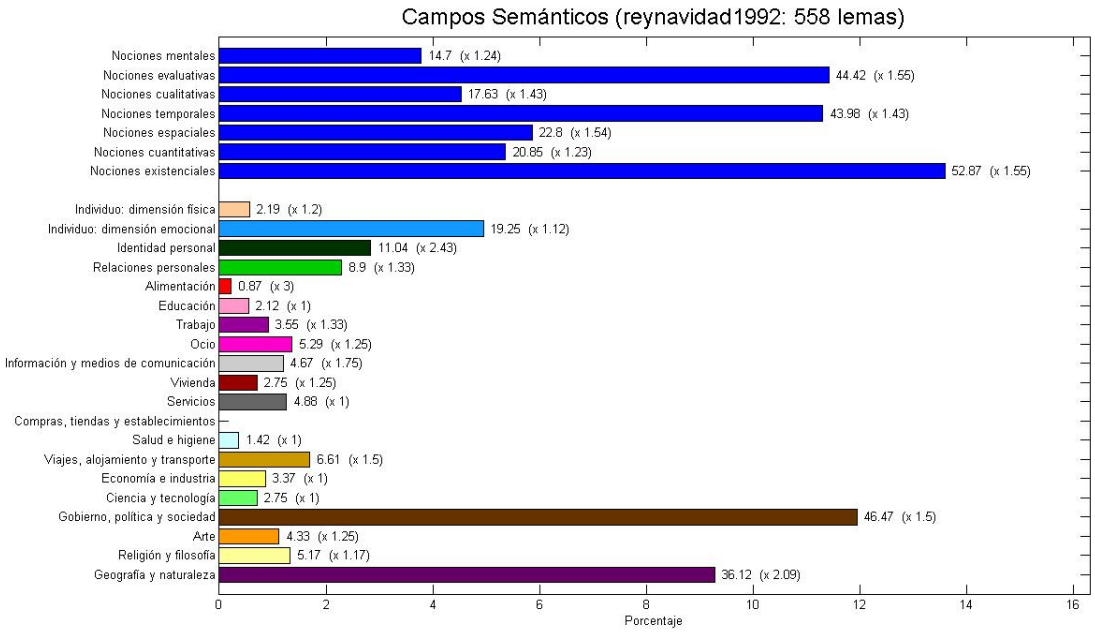


Figura 6.8: Distribución de campos semánticos del discurso navideño del Rey de 1992, aplicando la *stoplist*, no incluyendo los lemas repetidos y ponderando la asignación de campo semántico.

medio, alto o fuera de tema en función de la correlación de dicho texto con un *corpus* de referencia previamente nivelado. Es decir, un texto sometido a análisis obtendrá un grado de calificación según la mayor o menor aproximación a un texto modelo de referencia.

6.4. Método de evaluación semántica del texto

Los descriptores de nivel o grado de contenido semántico de un texto no están en función del número de vocablos de un texto ni de la suma obtenida por los distintos métodos propuestos sino en función de un buen *corpus* de referencia, cuando se aplica el método del LSA y el de los campos semánticos. También la evaluación semántica se basa en un glosario de referencia bien configurado, en nuestro caso el esWordnet, para hallar cierta relación de coherencia, y en el glosario de la Dra. Fuensanta López, para determinar las áreas temáticas. Por ello, en nuestro estudio, cualquier parámetro semántico se define en función del grado de aproximación semántica de un texto a un *corpus* de referencia o a una base de datos específica.

6.4.1. Métodos de correlación textual con *corpus* de referencia

6.4.1.1. Método del LSA

Empezamos con el método más utilizado en la actualidad en el análisis del contenido semántico: el Análisis Semántico Latente (LSA) o Análisis de Componentes Principales

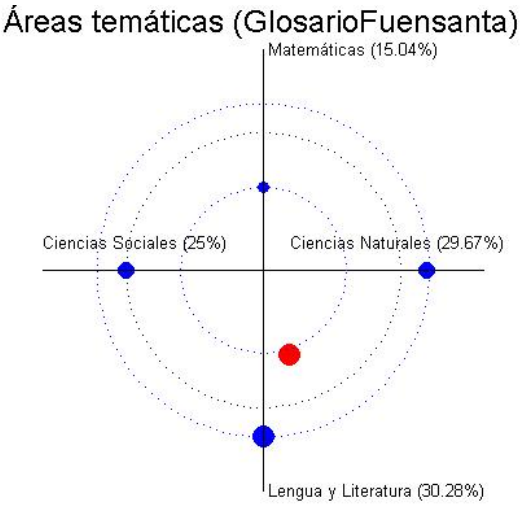


Figura 6.9: Áreas temáticas del glosario de la Dra. Fuensanta López sin ponderar.

(PCA). El método del LSA nos marcará el "índice de similitud" semántica entre textos. Indicará aquellas zonas de máxima, media, mínima o nula correlación semántica entre textos comparados entre sí. Observamos la figura 6.15. En ella se representa la matriz de varianza-covarianza correspondiente a todos los discursos navideños del Rey. Esa matriz es el punto de partida para el análisis con el método del LSA. Los resultados se expresan como sigue: en color rojo oscuro, se marca la máxima coincidencia semántica entre los textos; en color naranja, la aproximación media entre discursos diferentes; en color amarillo, la correlación baja; y en color azul, la coincidencia nula. El cálculo del "índice de similitud" será la media entre todos los grados calculados distinguiendo en los grados alto, medio, bajo sus correspondientes niveles A1-A2, B1-B2, C1-C2, respectivamente, o bien calculando las distancias al centro de los *clusters*.

Cuando aplicamos el análisis semántico latente a todos los discursos navideños, éstos se reorganizan en *clusters* más o menos compactos, dependiendo del mayor o menor grado de relación temática entre ellos. En la figura 6.16, se reagrupan los discursos por similitud temática. Los trazos discontinuos marcan la línea cronológica de los discursos. Los puntos representan la ubicación de cada uno de los 36 discursos con respecto a los demás. La distancia entre los puntos es significativa. Cuanto más próximos están los puntos entre sí, más cercana es su temática. Se observa la proximidad de la mayoría de los discursos por décadas. Por ejemplo, están muy próximos, por un lado, los de 1975 y 1976; los de 1988 y 1986; y los de 1977, 1978, 1980, 1981, 1982, 1983, 1984, 1985, 1987, 1989. Por otro lado, se reagrupan los de los años noventa por una parte y, por otra, los del dos mil. El más diferente a todos ellos es el de 1979. La figura 6.16 ya fue presentada en el apartado 3.2.3 donde se explicaban los fundamentos del método LSA.

Como se observa en las figuras 6.15 y 6.16, el análisis de los textos con el método del LSA permite diferenciar muy bien la realidad temática del texto; además, la distinción visual de la organización temática de los discursos, en *clusters*, es clara y distintiva.

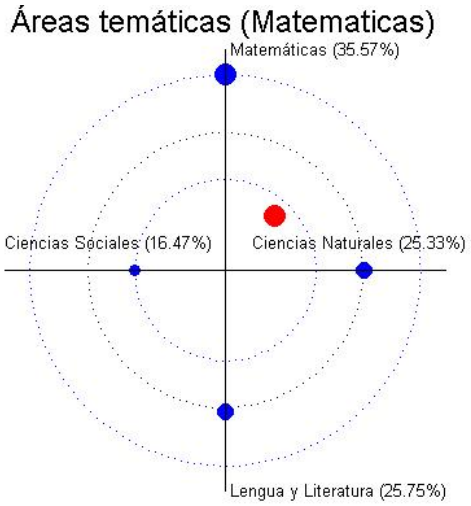


Figura 6.10: Área temática para un texto de Matemáticas según el glosario de la Dra. Fuensanta López. El círculo representa la localización del texto. El punto azul grande indica el área predominante. El punto azul pequeño indica el área menos presente. Se han trazado dos círculos discontinuos para indicar esta posición. El círculo intermedio corresponde a una distribución igual para todas las áreas temáticas.

6.4.1.2. Método de los campos semánticos

Este método de análisis semántico es un híbrido, ya que se computa la relación de los campos semánticos de un texto con los de un *corpus* de referencia. Este método lo aplicamos y detallamos en el análisis de los exámenes de DELE intermedio en el apartado 7.2.4.1. Anticipamos que esta metodología consiste, primero, en compilar textos con unas características comunes que puedan conformar un *corpus* de referencia; luego, en hallar los campos semánticos ponderados del *corpus* de referencia y de cada uno de los exámenes que analizamos; después, en comparar estos campos con los de un *corpus* de referencia, *corpus* representativo de los campos que queremos identificar. Este método se diferencia del LSA en que comparamos los exámenes, previo análisis de los campos semánticos, con unos textos de referencia.

6.4.2. Métodos de correlación textual con bases de datos

6.4.2.1. Método del parentesco

El método del parentesco se asocia a la base de conocimiento del esWordnet. Éste método nos permite hallar el "índice de relación de coherencia o de genealogía". Dicho índice nos marca un grado de relación alta, media, baja o nula entre los vocablos contenido del propio texto y entre un texto de referencia propio del nivel y temática que queremos medir. Como hemos expuesto más arriba, dado el coste computacional para procesar los textos, no lo aplicamos a los exámenes sino a un discurso del Rey, a modo de ejemplo, en el apartado 6.1.2.

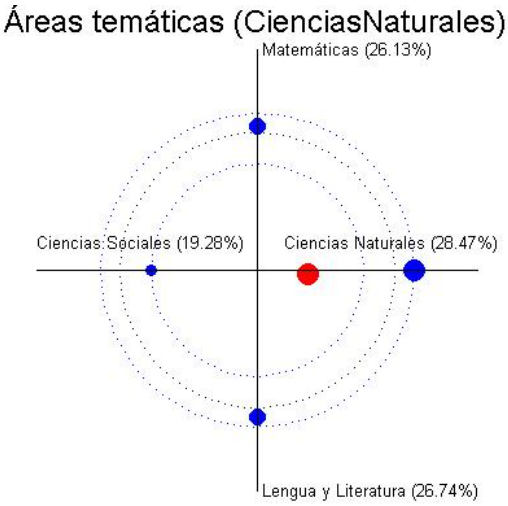


Figura 6.11: Áreas temáticas para un texto de Ciencias Naturales según el glosario de la Dra. Fuensanta López.

6.4.2.2. Método de las áreas semánticas

El método de las áreas semánticas se realiza con el glosario de la Dra. Fuensanta López. Éste nos puede proporcionar igualmente un "índice de pertenencia" alto, medio, bajo en ciertas áreas y se presta muy útil para discriminar tipos de texto con cierta especialización, como hemos desarrollado en el apartado 6.1.4. Sin embargo, debido a la generalidad del método, tampoco lo aplicamos a los exámenes.

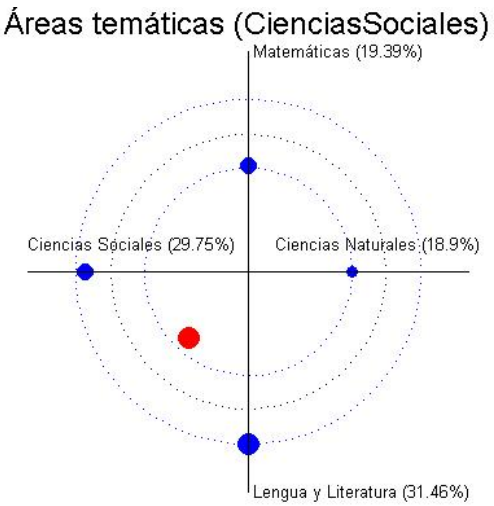


Figura 6.12: Áreas temáticas para un texto de Ciencias Sociales según el glosario de la Dra. Fuensanta López.

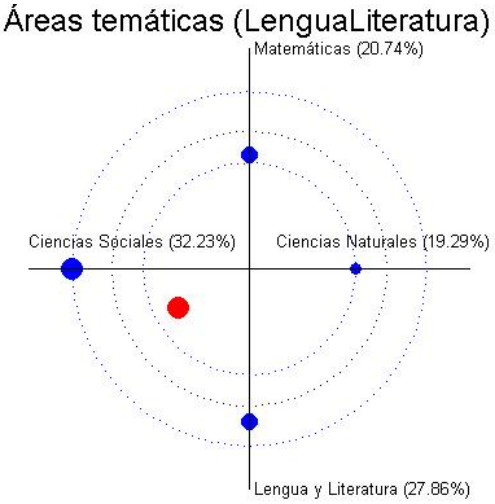


Figura 6.13: Áreas temáticas para un texto de Lengua y Literatura según el glosario de la Dra. Fuensanta López.

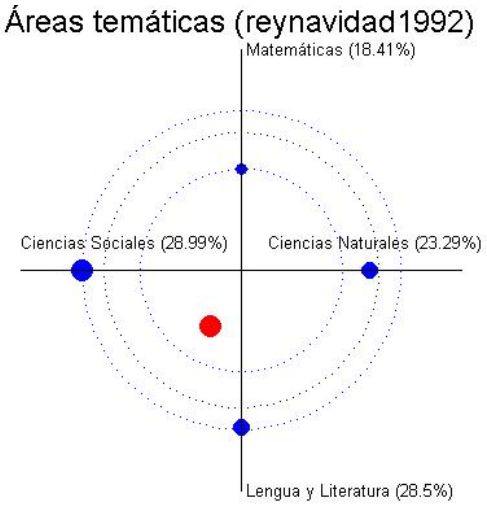


Figura 6.14: Areas temáticas para el discurso del Rey de 1992.

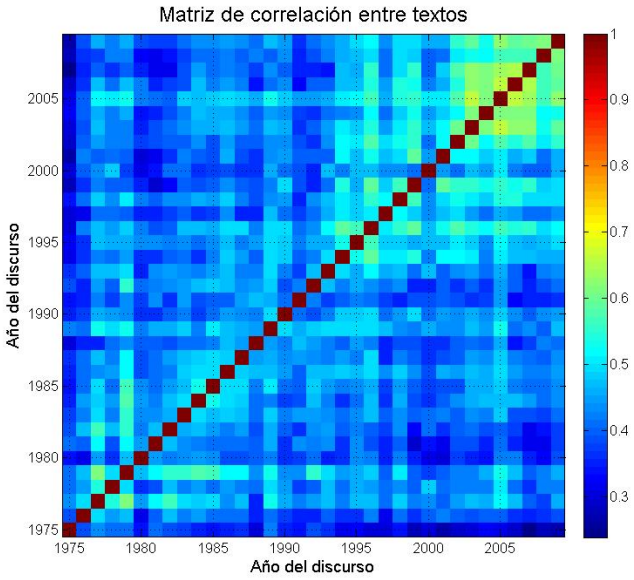


Figura 6.15: Matriz de correlación entre los discursos navideños del Rey.

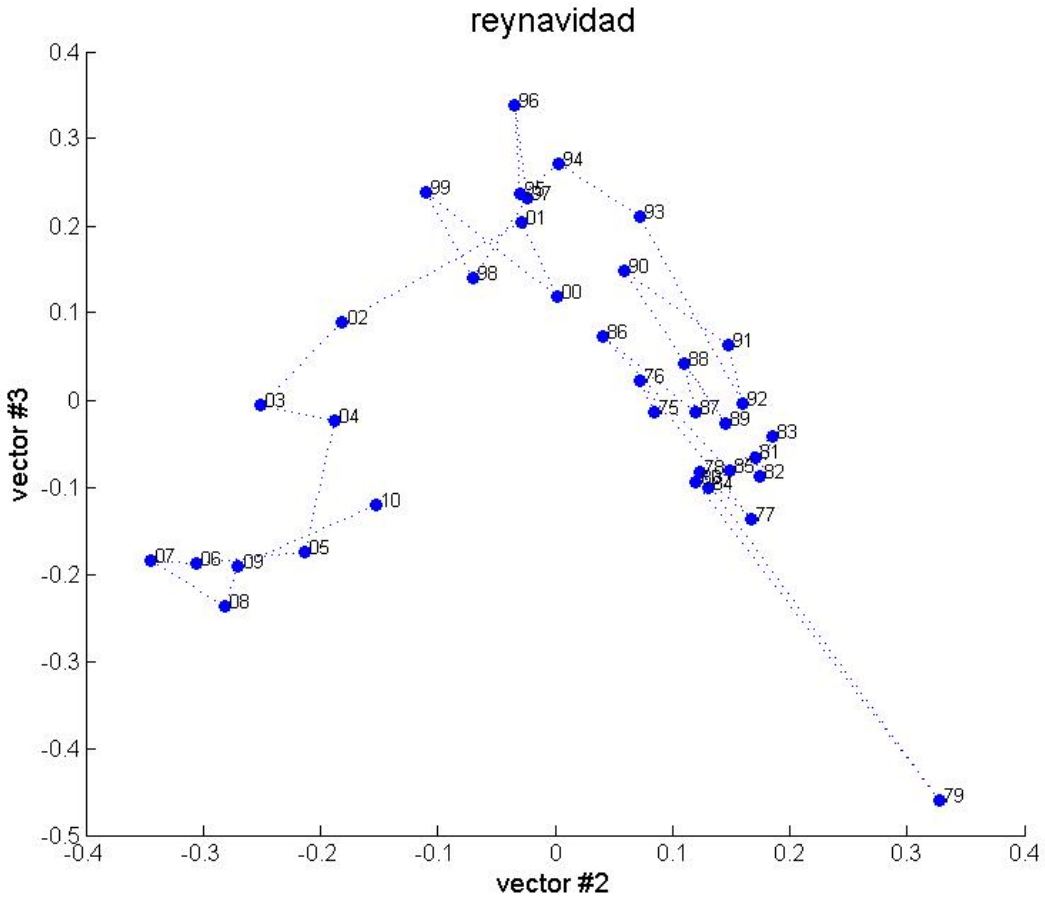


Figura 6.16: Representación de los vectores #2 y #3 obtenidos a mediante la LSA para los discursos navideños del Rey. Los números indican el año del discurso las líneas los unen cronológicamente. Se puede observa una agrupación de los discursos en *clusters*.

Capítulo 7

Aplicación del método, resultados y discusión

7.1. Descripción de los textos analizados

Los textos utilizados para probar los métodos y analizar su contenido a lo largo de este trabajo son de distintos autores. Desde el punto de vista de la autoría y naturaleza, se clasifican en dos grandes grupos: Los discursos del Rey y los escritos por estudiantes, nativos unos y extranjeros otros. Desde el punto de vista de la aplicación al estudio, en unas ocasiones unos escritos son útiles como textos de referencia; en otras, como textos a los que se analiza para comprobar su nivel de aprendizaje; y en otras tantas, como textos de prueba y demostración de las funcionalidades de las herramientas automáticas. Por último, desde la perspectiva del nivel de referencia en el aprendizaje de una lengua, distinguimos textos estándar de nivel alto y de nivel medio. En los siguientes apartados describimos en detalle tipo, origen, número y utilidad de cada grupo de textos.

7.1.1. Textos de referencia

Dentro de los textos de referencia, distinguimos dos clases de textos: los textos de nivel alto y los de nivel medio, ambos utilizados como textos de prueba y referencia. Cuando se alude a cada grupo de textos como un conjunto homogéneo, según su función, lo denominamos *corpus* de referencia o *corpus* de análisis.

7.1.1.1. Textos de nivel alto

Consideramos textos estándar de alto nivel a los discursos navideños del Rey. Estos discursos resultan válidos como *corpus* de referencia porque están escritos en un castellano estándar y supeditados a una temática y una estructura similar a lo largo del tiempo. Contamos con los discursos pronunciados por el Rey D. Juan Carlos I entre los años 1975 -2010. El año del discurso es el código de identificación para su procesamiento y estudio. Estos discursos se han extraído de la página web de la Casa Real correspondiente a los

discursos pronunciados por D. Juan Carlos. A su vez, cada discurso se ha convertido en un fichero "txt" de texto plano. En nuestros archivos nombramos cada discurso con su año correspondiente, por ejemplo: `reynavidad1992.txt`.

7.1.1.2. Textos de nivel medio

Disponemos de un conjunto de textos escritos por nativos adultos españoles con una competencia media en la expresión escrita. Independientemente de la destreza de estos alumnos en la escritura, ha sido útil haber podido contar con estos textos. Se han convertido para este estudio en *corpus* de referencia para probar y calibrar la eficacia de las herramientas a nivel léxico, sintáctico y semántico. Dentro de este grupo que conforma una serie de 144 textos, hay dos subgrupos. Un grupo tiene un nivel ligeramente más bajo (el 1) y el otro, apenas más alto (el 2). Sin preparación previa, cada grupo ha redactado textos relacionados con tres de las pruebas escritas propuestas en los exámenes del DELE intermedio del 23 y 24 de mayo de 2008.

Los textos solicitados a los alumnos adultos españoles han sido similares a los del DELE. No obstante, estos estudiantes no han sido entrenados para realizar ningún tipo de texto. Simplemente se les ha pedido que siguiesen las pautas que indicaba el ejercicio correspondiente a la sección de escritura de los exámenes del DELE. En tres ocasiones se les propuso escribir una redacción similar a la que se plantea en los mencionados exámenes del DELE. En una sesión, redactan una carta destinada a un *penpal* para practicar el idioma con una persona nativa (44 textos); en otra, escriben una reclamación a una compañía aérea por el extravío de su equipaje (33 textos); y en una tercera, describen un lugar especial para ellos (37 textos).

Estos textos se han escrito con unos fines muy claros: por un lado, con vistas a esta investigación, recopilar varios textos con la misma temática y estructura para convertirlos en textos de referencia para el análisis semántico de los exámenes del DELE Intermedio; y por otro, *a posteriori* de la realización de la prueba, formar a los alumnos, ya que el objetivo didáctico era que observasen los convencionalismos de los distintos tipos de texto, que practicasen las reglas ortográficas y, tras corregir los errores, que reflexionasen sobre los diferentes tipos de discursos, sus peculiaridades y los errores cometidos.

Todos los textos de estos alumnos nativos adultos se han escrito a mano en clase y, posteriormente, se han convertido a formato electrónico. Señalamos que se han corregido en la transcripción tildes y faltas ortográficas para que las herramientas automáticas procesaran el máximo de vocablos. No obstante, el número de palabras solicitadas en estos ejercicios oscilan entre 120-200. En nuestro registro se identifican estos textos por tema y grupo (1 o 2), seguido de un dígito que identifica a cada alumno. Por ejemplo, `Lugar_especial_2_1` se refiere a la descripción del lugar especial realizada en el grupo 2 por el alumno 1.

Se ha pedido autorización a estos alumnos para utilizar sus escritos, procesarlos y, luego, poder comparar los textos entre sí y con los exámenes de los candidatos al DELE intermedio, siempre manteniendo el anonimato de todos ellos. Se puede ver un modelo de texto solicitado, encabezado con la petición de consentimiento para el uso de estos escritos, en el apartado A.6 del apéndice.

7.1.2. Textos de candidatos al DELE: Características formales de los exámenes

Todo nuestro estudio se basa precisamente en el análisis textual y la interpretación de los resultados de un cierto número de exámenes del nivel intermedio y avanzado. Los exámenes analizados fueron realizados el 23 y 24 de mayo de 2008 por aprendices aspirantes a la obtención del diploma intermedio o superior. Dentro de la variedad temática y tipológica de los exámenes que disponemos, no contamos con muchos textos monotemáticos sino que una parte de ellos tratan sobre unos temas y otra parte sobre otros. Hay varios exámenes con diferentes temas y éstos no coinciden por pertenecer a una u otra convocatoria de exámenes o a una u otra selección de un tipo de texto, aunque sí coinciden los niveles de los test. Tenemos, por un lado, los exámenes de nivel intermedio y, por otra, los de nivel superior. En consecuencia, este hecho implica que no tengamos muchos textos similares. Destacamos también que, lógicamente, los exámenes de DELE intermedio y superior se diferencian entre sí, en función del nivel, en cuanto al tipo de texto, la longitud y la temática del mismo.

A continuación, describimos detalladamente las características de cada nivel y la de los textos-sujetos que disponemos.

7.1.2.1. Nivel intermedio

Un test de DELE de nivel intermedio propone la elección de un tipo de texto, entre dos opciones, en cada uno de las dos partes del test, y la elaboración de dos textos entre 150-200 palabras, cada uno con una temática y estructura determinada.

Test del 23 de mayo de 2008. Nivel intermedio

Este test consta de las partes que describimos en los siguientes párrafos (Cervantes, 2008a, Expresión escrita: 14-15).

En una primera parte, el alumno tiene que escribir una carta personal cuyo contenido debe desarrollar tras eligir un tema entre dos propuestos.

La primera opción propone un texto de estructura epistolar donde el aprendiz exponga su interés por practicar castellano. La propuesta del test detalla explícitamente que "La escuela de lenguas donde usted estudia le ha proporcionado la dirección de un estudiante interesado en escribir cartas en español para poder practicar". De forma que el aprendiz debe "presentarse; presentar su satisfacción e interés por esta actividad, exponer los motivos por los que estudia español, despedirse e invitarle a que conteste" (14).

La segunda opción pide al candidato que redacte una carta de reclamación para recuperar su equipaje ya que "La aerolínea con la que ha viajado ha perdido su maleta y después de tres días todavía no se la han mandado", y que siga una estructura temática con el fin de "presentarse; explicar el problema; describir el contenido de la maleta; solicitar una solución a su problema" (14).

En una segunda parte, el candidato debe escribir una redacción, de nuevo, eligiendo entre dos temas:

En la primera opción se le pregunta al candidato: "¿Hay algún lugar que para usted sea especial?" Además, se le pide que describa "dónde está ese lugar; cuándo estuvo allí por primera vez; cómo es el lugar y por qué es especial; por qué le gustaría volver" (15).

En la segunda opción se le propone al candidato que escriba como "Con frecuencia podemos escuchar que la primera impresión que tenemos de alguien es la que permanece". En el escrito, el candidato debe "exponer argumentos a favor o en contra; dar ejemplos que justifiquen su opinión; hablar de su experiencia personal; elaborar una breve conclusión" (15).

Test del 24 de mayo de 2008. Nivel intermedio

Igualmente, este test presenta sus propuestas en las partes que describimos en los siguientes párrafos (Cervantes, 2008b, Expresión escrita: 14-15) .

En una primera parte, el alumno tiene que escribir una carta personal cuyo contenido debe desarrollar tras elegir un tema entre dos propuestos.

La primera opción propone también una estructura epistolar del texto donde el aprendiz invite a una persona a un viaje. La propuesta del test explícita: "Usted acaba de ganar un viaje para dos personas. Escríbale una carta a una amiga a la que hace tiempo que no ve, en la que deberá: saludarla; comunicarle la noticia; describir el viaje y explicarle cuándo y dónde es; invitarla a que vaya con usted" (14).

La segunda opción pide al candidato que redacte una carta de reclamación para indemnizar daños y perjuicios en la que se apunta que "Usted ha realizado un viaje en tren por motivos de trabajo y el tren ha llegado a su destino con dos horas de retraso. Por tal motivo usted no ha podido asistir a una cita muy importante. Escríbale una carta a la compañía ferroviaria en la que deberá: saludar e identificarse; explicar los motivos de la carta; exponer los daños causados; reclamar una indemnización" (14). Esta opción no es elegida por ninguno de los sujetos que estudiamos.

En una segunda parte, el candidato debe escribir una redacción, de nuevo, eligiendo entre dos temas:

En la primera opción se le pide al candidato que redacte que "Cuando éramos pequeños todos soñábamos con tener una profesión ideal de mayores. Escriba una redacción en la que cuente: qué profesión era; por qué la admiraba tanto; si ha cambiado de opinión; si se ha cumplido su sueño o no y por qué" (15).

En la segunda opción se le propone al candidato que escriba como "Algunas personas prefieren pasar su tiempo libre con amigos en sus casas y otras prefieren reunirse con ellos en lugares públicos. Elabore un escrito en el que deberá: expresar su opinión a favor o en contra; dar ejemplos que la justifiquen; hablar de su experiencia personal; elaborar una breve conclusión" (15).

7.1.2.2. Nivel superior

El test del nivel superior sigue una estructura similar a la del nivel intermedio. Igualmente, se pide que los textos tengan una extensión de 150 a 200 palabras. Sin embargo,

mientras que en la primera parte del test se pide redactar una carta formal sobre uno de los dos temas propuestos, en la segunda parte se propone una redacción sobre un tema que se puede elegir entre tres opciones.

Test del 23 de mayo de 2008

Este test consta de las siguientes partes (Cervantes, 2008c, Expresión escrita: 14-15) :

En la primera parte del test se presenta la elección de uno entre dos temas de la siguiente manera:

Opción 1: "Usted acaba de llegar a un país en el que va a pasar varios años por cuestiones de trabajo y se ha enterado de que hay un centro cultural en el que se reúnen personas de su país de origen. Escriba una carta a dicho centro en el que deberá: presentarse y explicar su situación; solicitar información sobre las actividades que realizan; expresar su deseo de unirse al centro; ofrecerse para ayudar en lo que pueda" (14).

Opción 2: "Usted es un profesor que ha decidido ir con sus alumnos a visitar una importante exposición guiada en otra ciudad. Sin embargo, cuando llega no aparece el guía y debe realizar la visita sin su ayuda. Escriba una carta de reclamación a los organizadores de la exposición en la que deberá: presentarse y explicar el motivo de la misma; contar detalladamente su experiencia; pedir una explicación y una compensación por lo ocurrido" (14).

En la segunda parte del test se debe elegir una entre las tres opciones siguientes:

Opción 1: "Seguramente usted ha pensado en algún trabajo que, por diversos motivos, no estaría capacitado para hacer. Escriba un texto en el que cuente: de qué trabajo u oficio se trata; por qué ha llegado a esta conclusión; cómo actuaría en el caso de tener que realizarlo" (15).

Opción 2: "Probablemente ha habido alguna ocasión en la que habría deseado ser invisible. Redacte un texto en el que deberá: comentar cuándo tuvo ese deseo y qué lo provocó; cómo se desarrolló toda la situación; cuál fue el desenlace y cómo lo recuerda hoy día" (15).

Opción 3: "*Experiencia es el nombre que damos a nuestras equivocaciones*. Exprese su opinión respecto a esta frase indicando: las razones por las que está de acuerdo o no; algún ejemplo que apoye a su tesis; una conclusión a su argumentación". (15)

Test del 24 de mayo de 2008

Este test consta de las siguientes partes (Cervantes, 2008d, Expresión escrita: 14-15) :

En la primera parte del test se presenta la elección de uno entre dos temas con formato de carta-reclamación:

Opción 1: "Usted pidió información a una compañía de comunicaciones sobre sus ofertas para Internet y telefonía. Aunque les ha expresado su desinterés, ellos siguen insistiendo con diversas proposiciones por medios telefónicos y postales. Escriba una carta a dicha compañía. En ella deberá: presentarse y contar lo que ha pasado; describir su cansancio ante la situación; pedir que no se le moleste más" (14).

Opción 2: "Debajo de su casa han abierto un restaurante que tiene una insonorización deficiente, lo que les crea muchas molestias a usted y a sus vecinos. Usted ha intentado hablar con los propietarios para solucionar esta situación pero no le han hecho ningún caso. Escriba una carta al Ayuntamiento de su ciudad en la que deberá: explicar su caso; contar qué molestias le ocasiona; describir algún momento u ocasión en particular; pedir una solución a este problema" (14).

En la segunda parte del test se elige una entre las tres opciones siguientes:

Opción 1: "Probablemente haya un libro cuya lectura le ha dejado una huella especial. Escriba una redacción en la que deberá: comentar cuándo lo leyó; describir la historia y sus sensaciones durante la lectura; recomendarlo argumentando el porqué" (15).

Opción 2: "Probablemente durante su infancia o adolescencia usted era coleccionista de algún tipo de objetos. Escriba un texto en el que deberá: explicar qué coleccionaba; por qué comenzó dicha colección y cuánto tiempo la siguió; comentar qué ocurrió con esa colección" (15).

Opción 3: "*La mayor riqueza del hombre no es una cuantiosa fortuna sino un buen carácter*. Elabore un escrito en el que: exprese su opinión a favor o en contra; dé algún ejemplo que justifique su opinión; elabore una breve conclusión sobre el tema" (15).

7.1.2.3. Tipología, temática y número de los exámenes estudiados

Nivel intermedio

Dentro del DELE de nivel intermedio contamos con 40 textos de 20 candidatos, 2 por candidato. Se caracterizan porque unos pertenecen al género epistolar con la función de invitar o reclamar; otros, al género narrativo, descriptivo y de opinión. Como no todos los textos tratan sobre los mismos temas, hacemos una relación de cómo se distribuyen los temas por partes, convocatorias y opciones.

Primera parte del test:

- 23 de mayo
  - Opción 1: Carta de *penpal*: 8 textos
  - Opción 2: Carta-reclamación de una maleta: 7 textos
- 24 de mayo
  - Opción 1: Carta de invitación a un amigo para un viaje ganado: 5 textos
  - Opción 2: Carta-reclamación por la ausencia de un guía: 0 textos

Segunda parte del test:

- 23 de mayo
  - Opción 1: Redacción de un lugar especial: 12 textos
  - Opción 2: Opinión sobre la primera impresión que se tiene de alguien: 3 textos
- 24 de mayo
  - Opción 1: Redacción sobre una profesión favorita: 4 textos
  - Opción 2: Opinar sobre quedar con los amigos en casa o en lugares públicos: 1 texto

Nivel superior

Dentro del DELE de nivel superior también contamos con 40 textos de 20 candidatos. Se caracterizan porque unos pertenecen al género epistolar con la función de solicitar información o reclamar; otros, al género narrativo o descriptivo; y otros tantos, al de opinión y argumentación. Como no todos los exámenes tratan sobre los mismos temas, volvemos a hacer una relación de cómo se distribuyen por partes, convocatorias y opciones.

Primera parte del test:

- 23 de mayo
  - Opción 1: Carta solicitando información sobre un centro cultural: 6 textos
  - Opción 2: Carta-reclamación por ausencia de un guía contratado: 6 textos
- 24 de mayo
  - Opción 1: Carta-reclamación a una compañía de telefonía: 5 textos
  - Opción 2: Carta-reclamación por ruido de un bar-restaurante: 3 textos

Segunda parte del test:

- 23 de mayo
  - Opción 1: Redacción sobre un trabajo que el aprendiz no se vea capacitado para realizar: 5 textos
  - Opción 2: Redacción sobre haber deseado ser invisible alguna vez: 0 textos
  - Opción 3: Texto de opinión sobre nuestra experiencia fundamentada en los errores: 6 textos

- 24 de mayo de 2008
  - Opción 1: Texto de opinión sobre un libro: 2 textos
  - Opción 2: Redacción sobre una colección de objetos: 3 textos
  - Opción 3: Texto de opinión sobre el buen carácter como la mayor riqueza del hombre: 4 textos

7.1.2.4. Nueva propuesta del DELE superior: mejor procesamiento y análisis

Extensión y estructura

Respecto a los modelos de examen de los niveles que nos ocupa esta investigación, B2 y C1, el nuevo C1 es el que mejor se puede adaptar, y con más garantía de un buen análisis, al procesamiento y evaluación automática. Sólo en el recién estrenado DELE superior en noviembre de 2010 de nivel de referencia C1 se observan novedades en la prueba escrita respecto a los anteriores modelos de examen de este nivel. Mientras que el modelo anterior de B2 (Cervantes, 2009a) y C2 (Cervantes, 2009b) permanecen semejantes en cuanto a extensión del texto y estructura, el modelo de C1 ha variado. Por un lado, esta nueva prueba de C1 aumenta la extensión del texto de 150-200 a 220-250 palabras, respecto a las anteriores ediciones de los niveles intermedio y superior; y por otro lado, tanto en la tarea de la primera parte del examen como en la segunda, la "Prueba de expresión e interacción escritas" está diseñada de tal manera que "ayudará a acotar" (Cervantes, 2010, 22) el contenido y la estructura, no sólo al aprendiz sino también al evaluador.

A continuación citamos literalmente parte de la nueva propuesta de examen escrito a la que se enfrenta el candidato. En la primera parte de la prueba "el candidato debe captar los puntos principales de un texto oral y elaborar un texto que contenga una valoración u opinión personal. La audición tendrá una duración aproximada de 4 minutos y se escuchará dos veces. La tarea consiste en redactar un texto argumentativo o expositivo de entre 220 y 250 palabras en el que el candidato exponga, defienda o rebata de manera clara, detallada y bien estructurada las principales ideas del texto de entrada, y en el que respete las convenciones y rasgos del género que se le haya solicitado. El candidato cuenta con información sobre el tipo de texto que debe elaborar y se basa para su resumen o argumentación en una conferencia, presentación o discurso relacionado con los ámbitos público, profesional o académico" (Cervantes, 2010, 19-20).

La segunda parte, aparentemente similar a las ediciones anteriores, resalta el aspecto estructural y convencional de cualquiera de las opciones en las que "el candidato debe redactar textos argumentativos o epistolares con el fin de persuadir, argumentar, valorar u opinar sobre algo. La tarea consiste en redactar un texto formal de entre 220 y 250 palabras en el que se expongan argumentos de manera clara, detallada y bien estructurada, y se respeten las convenciones y los rasgos del género especificado. El candidato debe elegir una entre dos opciones del siguiente tipo:
• Opción a: Redactar una reseña, un informe o un artículo de revista.
• Opción b: Redactar una carta de reclamación, de solicitud de una beca o de recomendación" (Cervantes, 2010, 21).

Temas

En la cuestión temática del nivel C1, sin embargo, se apunta que "el candidato cuenta con información sobre el contenido del texto de salida, que estará relacionado con los ámbitos público, académico o profesional, y con un estímulo escrito, consistente en un anuncio de prensa, en instrucciones o en un breve resumen de un artículo de opinión, que le ayudará a acotar y contextualizar su propio texto" (Cervantes, 2010, 22).

Estos cambios producidos en el nivel C1 son de gran interés ya que las herramientas de procesado automático se pueden ajustar mejor y calibrar con mayor precisión cuando los entornos están muy acotados. Es decir, por un lado, el hecho de escuchar un tema y redactar sobre él y, por otro, el ceñirse a un contexto en el que el candidato dispone de material de referencia sobre el que debe escribir, facilitan la focalización del contenido de la producción escrita. Sin duda, esta fórmula es la ideal para futuros exámenes de DELE en formato electrónico, "DELE-e", y su posterior procesamiento automático como ya se propone en el New TOEFL (Cumming *et al.*, 2006).

En conclusión, como postulamos en nuestros modelos de medida, un texto más extenso y bien acotado temáticamente proporciona mayor fiabilidad en el diagnóstico del nivel de referencia.

7.1.3. Texto de prueba

Se ha elegido un texto muy específico para comprobrar los resultados lingüísticos que proporcionan los módulos de Lexicator y Sintactor. Nos ha parecido que las características del capítulo 68 de *Rayuela* redactado en gíglico eran óptimas (Cortázar, 1986). La mayoría de sus vocablos, inventados, están correctamente estructurados a nivel morfo-sintáctico con las normas del castellano. Por ello, hemos analizado un texto que pudiera ser distinto en su concepción y que las herramientas automáticas dieran cuenta en su análisis de cierta anomalía. El estudio de este texto está desarrollado en el apartado 7.3.4.

7.2. Resultados de evaluación de los exámenes de DELE

Al igual que la fiabilidad en el diagnóstico del nivel del léxico en un texto pueda depender en principio de la extensión de dicho texto, también en el análisis sintáctico hay más posibilidades de calcular mejor el número de habilidades adquiridas por el aprendiz y de obtener mayor fiabilidad en el análisis cuanto más extenso sea el texto que se analiza. En nuestro estudio contamos con textos de exámenes de DELE de nivel intermedio (B2) y superior (C1) de una extensión que oscila entre 150 y 200 vocablos. Aunque son textos breves, vamos a computar todos los vocablos, incorrectos o repetidos, y vamos a diagnosticar un nivel léxico, sintáctico y semántico al texto.

7.2.1. Calificación por un experto

Una vez realizado el análisis de los exámenes, vamos a comparar las calificaciones de los textos de la prueba escrita de los exámenes de DELE entre las de los correctores humanos y la de un ordenador. Primeramente, presentamos las calificaciones de los textos evaluados por dos expertos del Instituto Cervantes y, posteriormente, las comprobaremos con los niveles de los textos obtenidos de forma automática en este estudio.

Como disponemos de las calificaciones de los textos escritos por los candidatos para la obtención del diploma intermedio y superior del Instituto Cervantes, por un lado, vamos a reproducir las partes calificadas por los dos expertos (A y B): la gramática (G) y el vocabulario (V). Estas secciones se corresponden con las partes estudiadas en este trabajo, la sintaxis y el léxico, respectivamente. Además, estas calificaciones se expresan en la tabla 7.1, donde se registran las del nivel intermedio; y en la tabla 7.2, las del nivel superior.

En primer lugar, exponemos en tablas los resultados totales emitidos por los expertos para el nivel intermedio en la tabla 7.1. Se marcan con un asterisco aquellos aprendices que han superado la prueba escrita, según el baremos de los correctores expertos.

Código	Texto1A	Texto1B	Texto2A	Texto2B	Total Gram.	Total Voc.
I_1	G:3 | V: 4	G:4 | V:4	G:3 | V:4	G:3 | V:4	3,25 *	4 *
I_2	G:3 | V: 4	G:3 | V:3	G:3 | V:3	G:3 | V:3	3 *	3,25 *
I_3	G:3 | V: 4	G:3 | V:3	G:3 | V:3	G:3 | V:4	3 *	3,5 *
I_4	G:2 | V: 3	G:3 | V:2	G:2 | V:2	G:3 | V:3	2,5	2,5
I_5	G:3 | V: 3	G:4 | V:3	G:3 | V:2	G:3 | V:3	3,25 *	2,75
I_6	G:3 | V: 4	G:3 | V:4	G:3 | V:3	G:4 | V:3	3,25 *	3,5 *
I_7	G:3 | V: 4	G:3 | V:3	G:3 | V:4	G:3 | V:4	3 *	3,75 *
I_8	G:3 | V: 3	G:3 | V:3	G:2 | V:3	G:3 | V:3	2,75	3 *
I_9	G:2 | V: 3	G:3 | V:3	G:2 | V:3	G:3 | V:3	2,5	3 *
I_10	G:2 | V: 3	G:3 | V:3	G:3 | V:3	G:3 | V:3	2,75	3 *
I_11	G:2 | V: 3	G:3 | V:3	G:2 | V:2	G:3 | V:3	2,5	2,75
I_12	G:3 | V: 3	G:3 | V:4	G:3 | V:3	G:3 | V:3	3 *	3,25 *
I_13	G:3 | V: 3	G:2 | V:2	G:3 | V:2	G:2 | V:3	2,5	2,5
I_14	G:4 | V: 3	G:3 | V:4	G:4 | V:4	G:4 | V:4	3,75 *	4 *
I_15	G:4 | V: 4	G:4 | V:4	G:4 | V:4	G:4 | V:4	4 *	4 *
I_16	G:3 | V: 3	G:3 | V:3	G:2 | V:2	G:3 | V:3	2,75	2,75
I_17	G:2 | V: 3	G:2 | V:3	G:2 | V:3	G:2 | V:4	2	3,25
I_18	G:4 | V: 4	G:4 | V:4	G:4 | V:4	G:4 | V:4	4 *	4 *
I_19	G:2 | V: 3	G:1 | V:3	G:2 | V:2	G:1 | V:3	1,5	2,75
I_20	G:3 | V: 3	G:3 | V:3	G:3 | V: 3	G:3 | V:3	3 *	3 *

Tabla 7.1: Calificaciones en gramática y vocabulario de los dos expertos para nivel intermedio. El asterisco indica la superación del umbral en cada uno de los aspectos: gramática y vocabulario.

Como puede apreciarse, en la tabla 7.1 se muestran las calificaciones de la prueba escrita de 20 candidatos al diploma de nivel intermedio. Esto es, 40 textos correspondientes

a 2 textos (Texto-1 y Texto-2) por aprendiz. Dos correctores (A y B) califican un mismo texto en una escala de 1 a 4. Para hallar el total se calcula la media de los dos correctores de ambos textos en los dos apartados, la gramática y el vocabulario. Señalamos que, conforme a la calificación del Instituto Cervantes, el umbral de apto para cada apartado está en el valor 2,80 que corresponde al 70 % de la máxima calificación de 4,00. En las columnas de la calificación total se han destacado con un asterisco (*) aquellos candidatos que superan este umbral en cada una de las áreas.

A su vez, en la tabla 7.2, se muestra igualmente el mismo esquema, en este caso, para el nivel superior.

Código	**Texto1A**	**Texto1B**	**Texto2A**	**Texto2B**	**Total Gram.**	**Total Voc.**
S_1	G:3 | V:2	G:3 | V:3	G:2 | V:2	G:2 | V:3	2,5	2,5
S_2	G:2 | V:2	G:1 | V:3	G:2 | V:2	G:1 | V:2	1,5	2,25
S_3	G:2 | V:3	G:2 | V:2	G:2 | V:2	G:2 | V:2	2	2,25
S_4	G:3 | V:3	G:3 | V:3	G:3 | V:2	G:3 | V:3	3 *	2,75
S_5	G:3 | V:2	G:2 | V:2	G:2 | V:2	G:2 | V:3	2,25	2,25
S_6	G:2 | V:2	G:2 | V:2	G:2 | V:2	G:2 | V:2	2	2
S_7	G:4 | V:3	G:4 | V:4	G:3 | V:4	G:3 | V:4	3,5 *	3,75 *
S_8	G:3 | V:4	G:4 | V:4	G:3 | V:4	G:4 | V:4	3,5	4
S_9	G:3 | V:2	G:4 | V:3	G:2 | V:2	G:2 | V:2	2,75	2,25
S_10	G:4 | V:3	G:4 | V:4	G:4 | V:4	G:4 | V:4	4 *	3,75*
S_11	G:3 | V:3	G:3 | V:3	G:2 | V:2	G:3 | V:2	2,75	2,5
S_12	G:3 | V:3	G:3 | V:4	G:4 | V:3	G:3 | V:4	3,25 *	3,5 *
S_13	G:3 | V:3	G:4 | V:4	G:3 | V:4	G:4 | V:4	3,5 *	3,75 *
S_14	G:2 | V:2	G:1 | V:3	G:2 | V:2	G:2 | V:2	1,75	2,25
S_15	G:3 | V:3	G:4 | V:3	G:2 | V:3	G:4 | V:3	3,25 *	3*
S_16	G:3 | V:3	G:3 | V:4	G:3 | V:3	G:4 | V:4	3,25 *	3,5 *
S_17	G:3 | V:4	G:4 | V:4	G:3 | V:3	G:4 | V:4	3,5 *	3,75 *
S_18	G:3 | V:3	G:3 | V:4	G:2 | V:2	G:3 | V:2	2,75	2,75
S_19	G:2 | V:2	G:3 | V:3	G:2 | V:2	G:2 | V:2	2,25	2,25
S_20	G:3 | V:3	G:2 | V:4	G:3 | V:3	G:3 | V:3	2,75	3,25*

Tabla 7.2: Calificaciones en gramática y vocabulario de los dos expertos para el nivel superior.

Una vez calculadas las medias de los dos correctores para cada examen-candidato, se presentan los valores en las columnas del "Total Gramática" y "Total Vocabulario". Aquellas calificaciones marcadas con un asterisco en las dos últimas columnas señalan que se han superado los dos apartados, según el baremo del Instituto Cervantes.

En los siguientes apartados presentamos los niveles léxicos obtenidos de forma automática en los distintos exámenes de DELE: intermedio y superior.

7.2.2. Nivelación léxica automática

Para los dos niveles, intermedio y superior, se han aplicado los mismos métodos de análisis léxico. Con los criterios del *PCIC*, del glosario de multi-vocablos y de la combinación de diccionarios, se ha obtenido el porcentaje del vocabulario correspondiente a cada glosario en el conjunto de exámenes referidos al Texto-1 y Texto-2 que vemos en la figura 7.1 para el nivel intermedio y en la figura 7.4 para el nivel superior. Concretamente, según el criterio de nivel de referencia del Instituto Cervantes, calculamos qué porcentaje del Texto-1 y Texto-2 pertenece a cada nivel de referencia. Estos porcentajes por niveles se ven en la figura 7.2 para el nivel intermedio y en la figura 7.5 para el nivel superior. Por último, mediante el criterio de distribución de frecuencias y niveles, hallamos el nivel léxico de cada examen como se ve en las figuras 7.3 y 7.6 para el nivel intermedio y superior respectivamente. Hay que considerar que, en el procesamiento de estos exámenes, computamos sólo los lemas de todos aquellos vocablos bien escritos.

7.2.2.1. Exámenes del DELE intermedio

Del estudio léxico de los exámenes de DELE intermedio, destacamos la distribución de porcentajes del léxico en los distintos glosarios utilizados para el estudio. Respecto al total del léxico procesado en los exámenes, se obtiene un 95,98 % del vocabulario procesado en el Texto-1 y a un 97,75 % en el Texto-2 (ver figura 7.1). Estos datos indican que se procesa un alto porcentaje del léxico de los textos; que este porcentaje de lemas identificados se corresponden con vocablos bien escritos en los exámenes; y que los resultados obtenidos por la combinación de diccionarios permite identificar en torno a un 30 % más vocablos de los que califica el Instituto Cervantes.

Todos los valores del léxico del Texto-1 y del Texto-2 por glosario se pueden observar en la figura 7.1. Los resultados totales y los datos numéricos de la distribución por diccionarios de cada examen referidos al Texto-1 y Texto-2 se hallan en las tablas 7.9, A.1 y A.4.

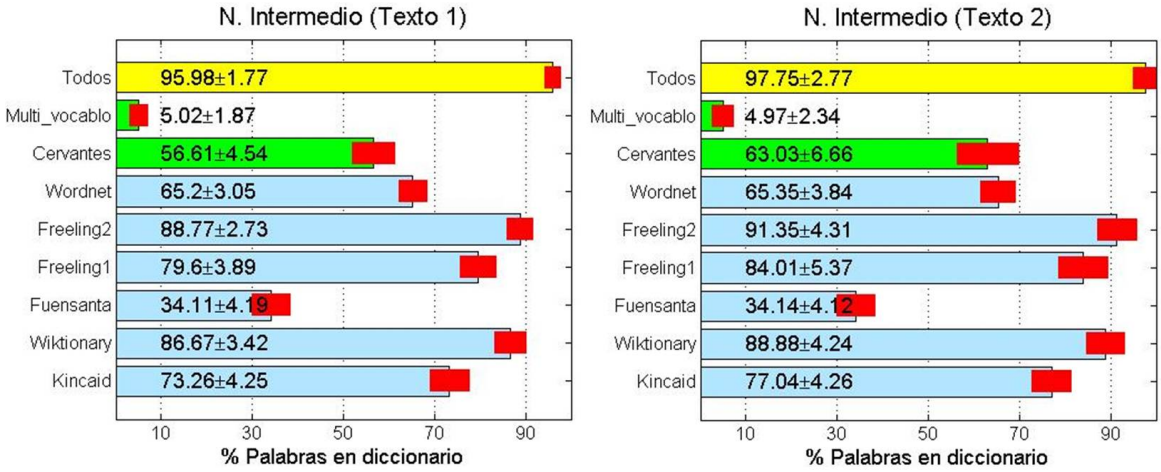


Figura 7.1: Identificación del léxico de los exámenes del DELE intermedio en todos los diccionarios. La banda roja expresa la dispersión en valores entre los exámenes considerados.

Respecto a la distribución de vocablos en niveles de referencia, observamos los por-

centajes que proporciona la combinación de glosarios junto con el criterio del glosario del Instituto Cervantes en la figura 7.2. De los resultados de este análisis léxico, destacamos la similitud de porcentajes por niveles tanto en los exámenes del Texto-1 como del Texto-2. No obstante se califican menos lemas en el Texto-1 (en torno al 80 %) que en el Texto-2 (en torno al 85 %). El método de la combinación de glosarios y el glosario de multi-vocablos nos ha permitido nivelar aproximadamente un 30 % más del léxico respecto al que nivelaría el "Índice" del Instituto Cervantes solo. Los valores numéricos están recogidos en las tablas A.2 y A.5 del apéndice.

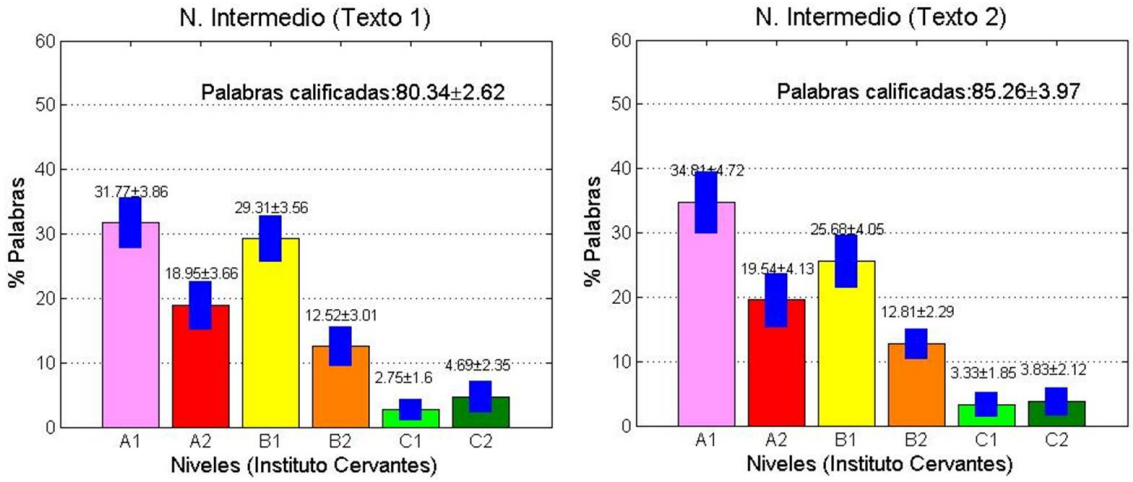


Figura 7.2: Identificación del léxico de los exámenes de DELE intermedio por niveles de referencia del Instituto Cervantes. Las bandas azules indican la dispersión en los datos de cada uno de los exámenes.

En la figura 7.3 se calcula el nivel de referencia del texto desde el punto de vista léxico. Se halla su nivel mediante el criterio de distribución de frecuencias y niveles. Los valores de cada examen se representan con un cuadrado azul para indicar el nivel de referencia de cada examen y con un punto azul se señala la tendencia a otro nivel. La mayor o menor extensión de la línea azul que une el punto con el cuadrado es la que indica la mayor o menor tendencia a otro nivel. Además, cada examen se dispone dentro de una sección de gamas de color rojo y verde. Cuanto más fuerte es el rojo de la sección menos fiable es el resultado y cuanto más verde, más válido. Cuando calculamos los niveles de referencia para cada uno de los exámenes y el índice de fiabilidad, comprobamos que los exámenes del Texto-1 se califican en su mayoría con un nivel B1 con tendencia a B2 aunque con fiabilidad negativa para los exámenes # 2 y 13; en cambio, con nivel B1 tendiendo a B2 y fiabilidad positiva son los exámenes # 3, 4, 8, 10, 17 y 18. El resto tiene valores definidos con otras tendencias, pero en su mayoría son de un nivel B1 y tienen fiabilidad positiva. Por el contrario, los exámenes del Texto-2 tienen mayoritariamente un nivel B2 con tendencia a B1, aunque la fiabilidad es negativa para los exámenes #1, 4, 6, 7, 10, 12, 13, 14 y 20.

Al observar esta figura 7.3, deducimos que el nivel de referencia más elevado es el de los exámenes del Texto 2. Para su comprobación numérica, los valores de cada examen correspondiente al Texto-1 y al Texto-2 se detallan en la tabla A.3 y A.6 del apéndice.

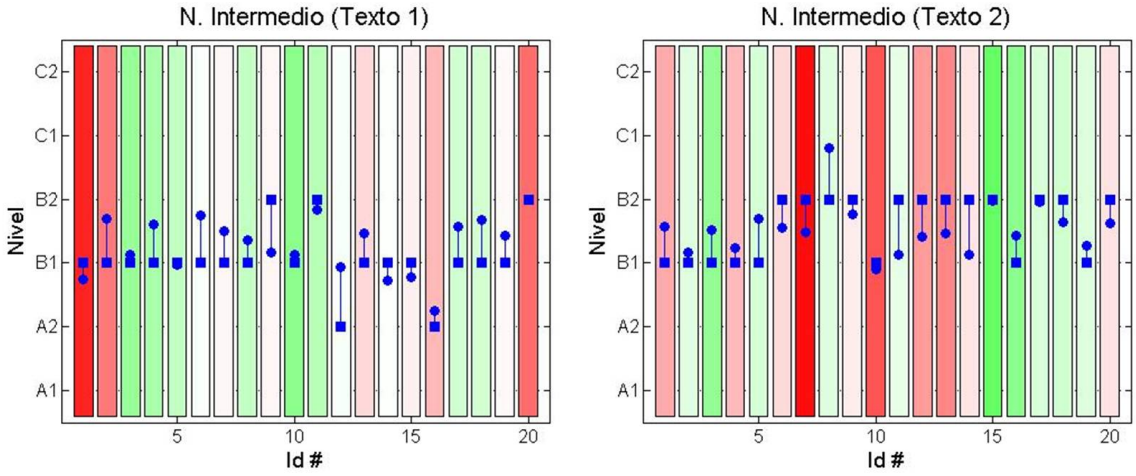


Figura 7.3: Nivel léxico de cada examen del DELE Intermedio.

7.2.2.2. Exámenes del DELE superior

El análisis léxico de los exámenes de DELE superior, mediante la distribución de porcentajes del léxico en los distintos glosarios, nos muestra un punto más del total del léxico procesado en los exámenes. Se obtiene un 96,44 % del vocabulario procesado en el Texto-1 a un 99,72 % en el Texto-2. Igualmente, la combinación nos permite procesar un mayor porcentaje del léxico de los exámenes, alcanzando a identificar de nuevo en torno a un 30 % más vocablos que los que podríamos procesar con el "Índice" del Instituto Cervantes.

Todos los valores de cada glosario del Texto-1 y del Texto-2 se pueden observar en la figura 7.4. Los datos numéricos de cada examen referidos al Texto-1 y Texto-2 se hallan en las tablas A.15 y A.18 del Apéndice.

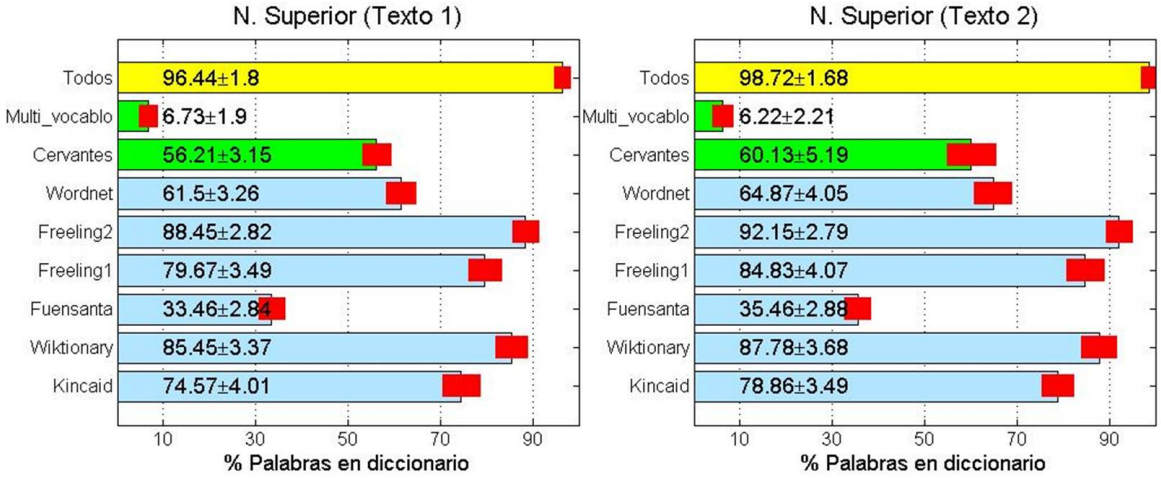


Figura 7.4: Identificación del léxico de los exámenes del DELE superior en todos los diccionarios.

Según el criterio de niveles de referencia para el léxico del Instituto Cervantes, en la figura 7.5 tenemos la distribución de niveles de los lemas para todos los exámenes. Obser-

vamos que los porcentajes que proporciona la combinación de diccionarios, el glosario de multivocablos y el glosario del Instituto Cervantes son muy similares en los exámenes del Texto-1 como del Texto-2. Concretamente, se llega a calificar entre un 82 % y un 85 % de los lemas del Texto-1 y Texto-2. Los valores numéricos están recogidos en las tablas A.16 y A.19 del apéndice.

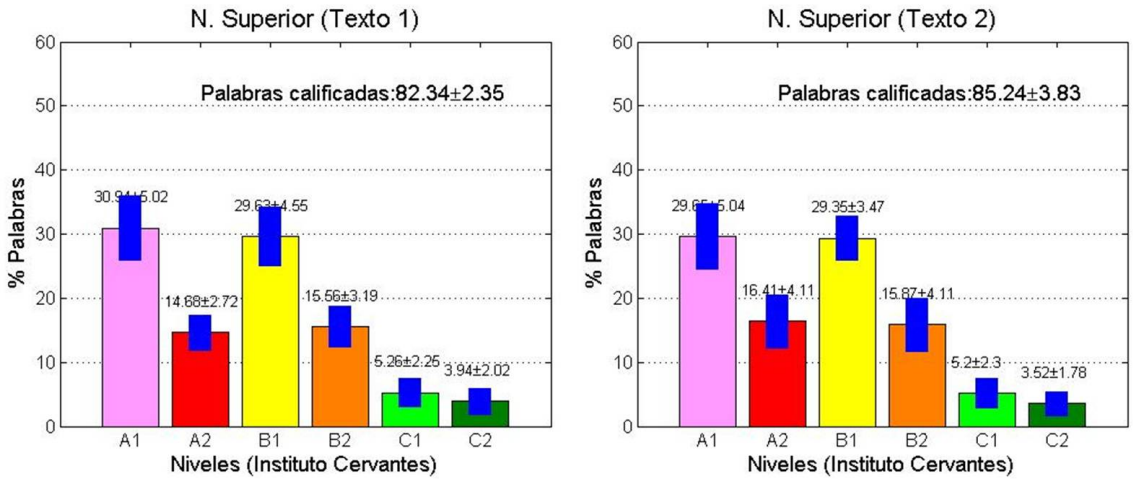


Figura 7.5: Identificación del léxico de los exámenes de DELE superior por niveles de referencia del Instituto Cervantes.

Una vez más, ahora para el nivel superior, en la figura 7.6 se calcula el nivel de referencia del texto desde el punto de vista léxico. Se halla su nivel mediante el criterio de distribución de frecuencias y niveles. Los valores de cada examen se representan con un cuadrado azul para indicar el nivel de referencia de cada examen y con un punto azul se marca la tendencia a otro nivel. La mayor o menor extensión de la línea azul que une el punto con el cuadrado es la que indica la mayor o menor tendencia a otro nivel. También, cada examen se dispone dentro de una sección de gamas de color rojo y verde. Cuanto más fuerte es el rojo de la sección menos fiable es el resultado y cuanto más verde, más fiable. Cuando calculamos los niveles de referencia para cada uno de los exámenes y el índice de fiabilidad, comprobamos que los exámenes del Texto-1 se califican, una parte (#2, 9, 10, 12) con un nivel B2 pero con tendencia a B1, mientras el #18 es un B2 propiamente con tendencia a C1 y tiene una fiabilidad positiva. Por otra parte, los exámenes #1, 3, 11, 15, 17 y 19, con un nivel B1, tienen una tendencia a B2. Mientras que los #5, 7 y 16 tienen un nivel A2 con tendencia a B1. Por otro lado, el grado de fiabilidad es bajo para los exámenes del Texto-1 (#1, 3, 4, 5, 6, 9, 10, 13, 14 y 16).

Por el contrario, los exámenes del Texto-2 tiene una parte de exámenes (#5, 9, 10, 12, 13, 14 y 16) con un nivel B2 con cierta tendencia a B1; otra parte de los exámenes (#2, 3, 6, 7, 8, 11, 15, 17, 18, 19, 20) destaca con un nivel B1, pero con mayor tendencia a B2; mientras que se da, incluso, un nivel A1 en el examen #1 y un nivel A2 en el #4. No obstante, la fiabilidad de los resultados del Texto-2, en general, es mayor.

Si observamos la figura 7.6, deducimos que el nivel de referencia también se eleva mínimamente en los exámenes del Texto-2. Para su comprobación numérica, los valores de cada examen correspondiente al Texto-1 y al Texto-2 se detallan en las tablas A.17 y A.20 del apéndice.

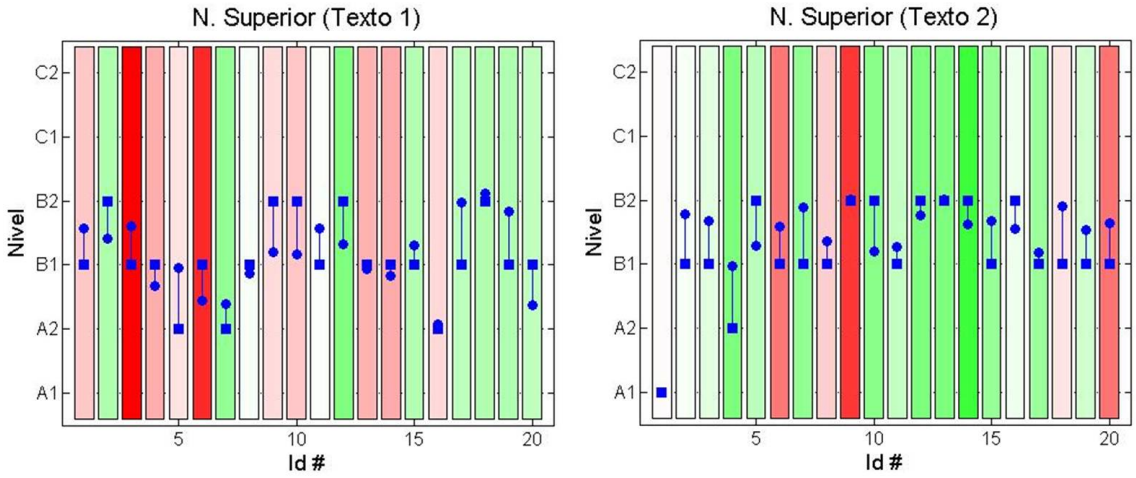


Figura 7.6: Nivel léxico de cada examen del DELE superior.

7.2.3. Nivelación sintáctica automática

Para nivelar los exámenes del Texto-1 y Texto-2, se procesan mediante el método de la "máxima diferencia positiva", anteriormente desarrollado en el apartado 5.4.1. Para aplicar este método hemos recurrido a otro *corpus* de referencia. En esta ocasión, el *corpus* está conformado por los discursos navideños del Rey. Además, este *corpus* está marcado con estructuras sintácticas niveladas después de procesarlo con el fichero de estructuras. Como hemos considerado que el *corpus* del Rey, *a priori*, es un texto modelo de nivel sintáctico C1, lo tomamos como referencia para medir cuantitativamente el nivel sintáctico de los exámenes (Zaanen *et al.*, 2004). De manera que, para hallar el nivel sintáctico de un examen, calcularemos la distancia normalizada entre los parámetros sintácticos del examen y la del *corpus*. Los parámetros sintácticos, definidos con anterioridad en el apartado 5.4.1 y aplicados, son el número de estructuras identificadas en un texto y la distribución en niveles de los *PoS*.

7.2.3.1. Exámenes del DELE intermedio

Para el grupo de exámenes identificado como Texto-1, se observan los resultados en la figura 7.7. En la parte izquierda de la figura se representa la distribución de todos los exámenes intermedios por niveles. Cada punto de color representa un nivel (rosa y rojo para A1 y A2; amarillo y naranja para B1 y B2; y verde claro y verde oscuro para C1 y C2, respectivamente). En la figura de la izquierda se representan todos los niveles que aparecen en cada examen. Para calificar, sin embargo, nos centramos en el cuadrante superior derecho, ampliado en la derecha de esta misma figura 7.7. En este cuadrante, nos encontramos que los exámenes #4, 7, 13 y 14 tienen un nivel de referencia sintáctico de A2 (rojo); los exámenes #2, 3, 5, 8, 9, 10, 11, 12, 15, 16, 17, 18, 19 y 20 tienen un nivel B1 (amarillo); y el examen #1 es el único con nivel B2 propiamente. Destacamos que el examen # 6 queda sin calificar porque sus estructuras sintagmáticas están muy próximas a las de los discursos navideños. Los datos numéricos de este análisis están recogidos en las tablas A.7, A.8 y A.9 del apéndice.

El posicionamiento de los puntos en uno u otro arco (0-1, 1-2, 2-3, 3-4, 4-5) del cuadrante superior derecho indica la fiabilidad del cálculo del nivel. Cuanto más alejado está el examen del origen, más fiable es su calificación.

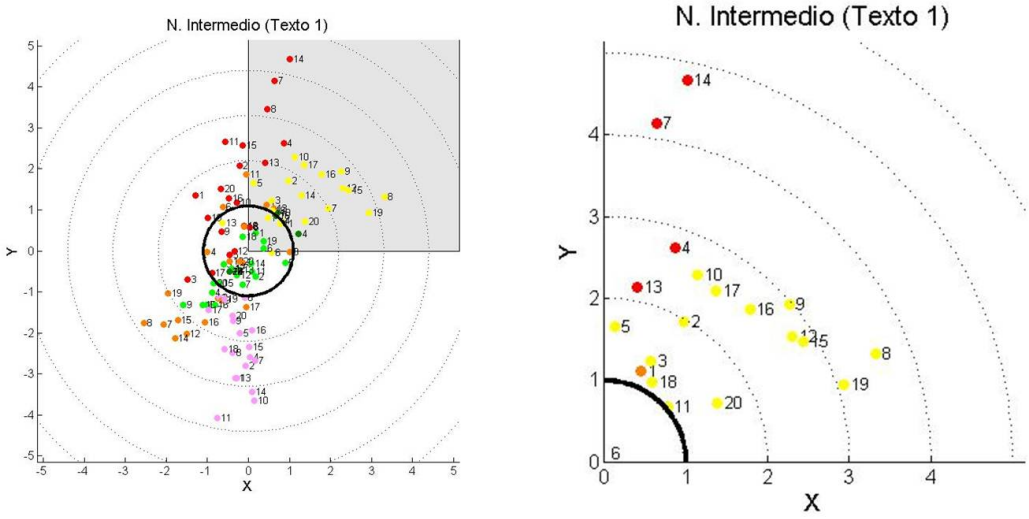


Figura 7.7: Nivelación sintáctica de cada texto de la opción 1ª de los exámenes del DELE intermedio. Cada texto está asociado a 6 puntos de color (uno para cada nivel) con un número identificador a su lado. El círculo de trazo continuo indica el umbral de calificación. Los círculos de trazo discontinua están a distancias múltiplos del umbral. La figura de la derecha representa el primer cuadrante de la figura de la izquierda.

Para el grupo de exámenes identificado como Texto-2, se observan los resultados en la figura 7.8. En la parte izquierda de la figura se representa la distribución de todos los exámenes por niveles. Igualmente que en el análisis anterior, cada punto de color representa un nivel (rosa y rojo para A1 y A2; amarillo y naranja para B1 y B2; y verde claro y verde oscuro para C1 y C2, respectivamente). En la figura de la izquierda se representan todos los niveles que aparecen en cada examen. Para dar un nivel a cada examen, sin embargo, nos centramos en el cuadrante superior derecho, ampliado en la derecha de esta misma figura 7.8. En este cuadrante, nos encontramos que el examen #20 tiene un nivel de referencia sintáctico de A2 (rojo); los exámenes #1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 15, 18 y 19 tienen un nivel B1 (amarillo); los # 6 y 17 tienen un B2 (naranja); y, por último, el examen 14 es el único con nivel C1 (verde claro). De nuevo, podemos observar que los exámenes # 7 y 16 quedan sin calificar, lo que significa que estos exámenes están muy próximos a las estructuras sintagmáticas de los discursos navideños. Los datos numéricos de este análisis están recogidos en las tablas A.10, A.11 y A.12.

El posicionamiento de los puntos en uno u otro sector (0-1, 1-2, 2-3, 3-4, 4-5) del cuadrante superior derecho indica la fiabilidad del cálculo del nivel. Cuanto más alejado está el examen del origen, más fiable es su calificación.

7.2.3.2. Exámenes del DELE superior

El mismo procedimiento se sigue para el estudio de los exámenes de DELE de nivel superior. Para el grupo de exámenes identificado con Texto-1, se observan los resultados

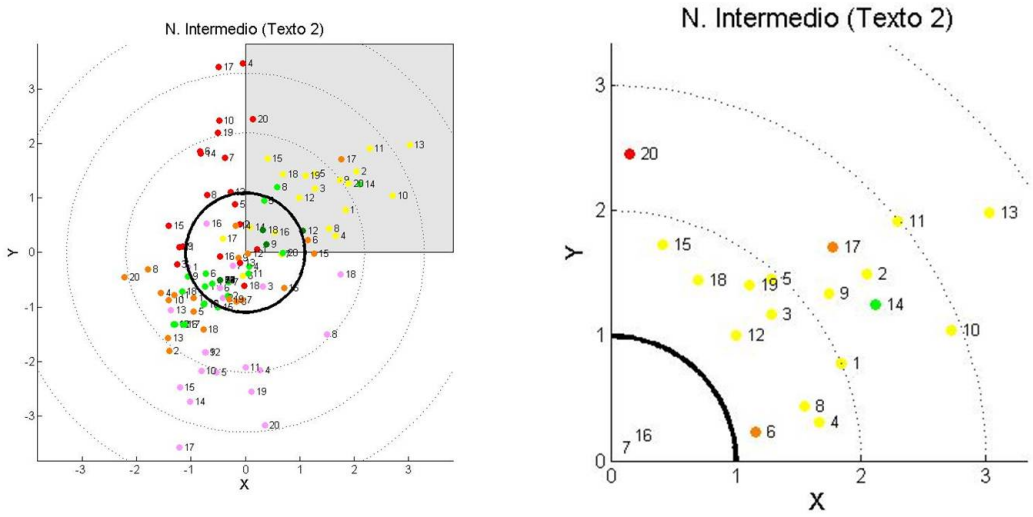


Figura 7.8: Nivelación sintáctica de cada texto de la opción 2ª de los exámenes del DELE intermedio.

en la figura 7.9. En la parte izquierda de la figura se representa la distribución de todos los exámenes por niveles. Igualmente que en los análisis anteriores, cada punto de color representa un nivel (rosa y rojo para A1 y A2; amarillo y naranja para B1 y B2; y verde claro y verde oscuro para C1 y C2, respectivamente). En la figura de la izquierda se representan todos los niveles que aparecen en cada examen. Para dar un nivel a cada examen, sin embargo, nos centramos en el cuadrante superior derecho, ampliado en la derecha de esta misma figura 7.9. En este cuadrante, nos encontramos que los exámenes #8 y 9 tienen un nivel de referencia sintáctico de A2 (rojo); los #1, 3, 6, 10, 17, 18 y 20 tienen un nivel B1 (amarillo); los # 4, 7, 11, 14, 15, 16 y 19 tienen un B2 (naranja); y son únicos en el nivel C1 (verde claro) el #12, y en el C2 (verde oscuro) el #13. También aquí, podemos observar que los exámenes # 2 y 5 quedan sin calificar. Es decir, estos exámenes están muy próximos a las estructuras sintagmáticas de los discursos navideños. Los datos numéricos de este análisis están recogidos en las tablas A.21, A.22 y A.23.

El posicionamiento de los puntos en uno u otro arco (0-1, 1-2, 2-3, 3-4, 4-5) del cuadrante superior derecho indica la fiabilidad del cálculo del nivel. Cuanto más alejado está el examen del origen, más fiable es su calificación.

Finalmente, para el grupo de exámenes identificado como Texto-2, se observan los resultados en la figura 7.10. En la parte izquierda de la figura se representa la distribución de todos los exámenes por niveles. Igualmente que en todos los análisis anteriores, cada punto de color representa un nivel: rosa y rojo para A1 y A2; amarillo y naranja para B1 y B2; y verde claro y verde oscuro para C1 y C2, respectivamente. En la figura de la izquierda se representan todos los niveles que aparecen en cada examen. Para dar un nivel a cada examen, sin embargo, nos centramos en el cuadrante superior derecho, ampliado en la derecha de la misma figura 7.10. En este cuadrante, nos encontramos que el examen #9 tiene un nivel de referencia sintáctico de A2 (rojo); los exámenes #3, 4, 6, 7, 8, 10, 12, 13, 14, 16, 17 y 20 tienen un nivel B1 (amarillo); los # 5 y 18 tienen un B2 (naranja); y con el nivel C1 (verde claro) el #1, 2 y 11. Una vez más, podemos observar que los exámenes # 15 y 19 quedan sin calificar. Los datos numéricos de este análisis están recogidos en las

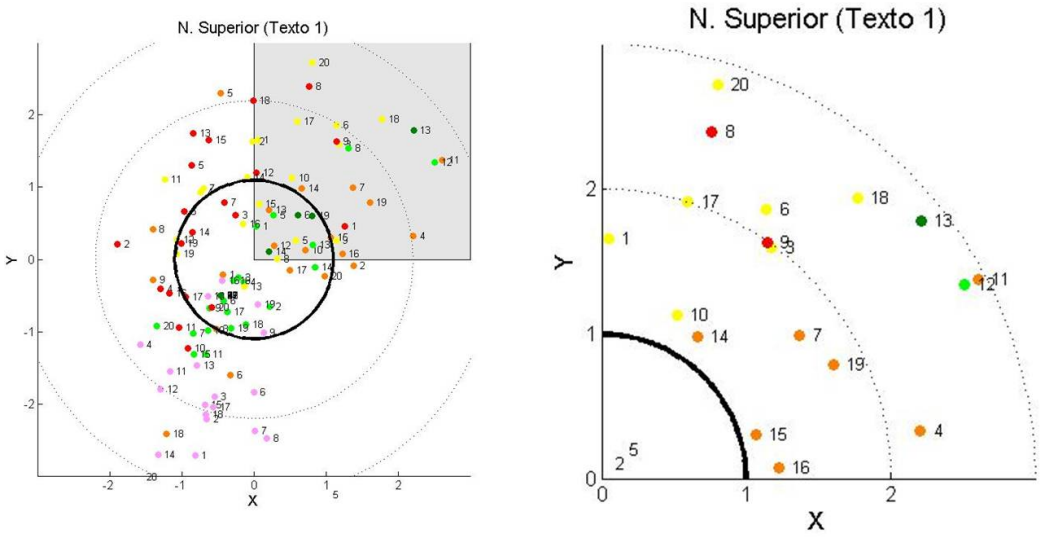


Figura 7.9: Nivelación sintáctica de cada texto de la opción 1ª de los exámenes del DELE superior.

tablas A.24, A.25, y A.26.

El posicionamiento de los puntos en uno u otro arco (0-1, 1-2, 2-3, 3-4, 4-5) del cuadrante superior derecho indica la fiabilidad del cálculo del nivel. Cuanto más alejado está el examen del origen, más fiable es su calificación.

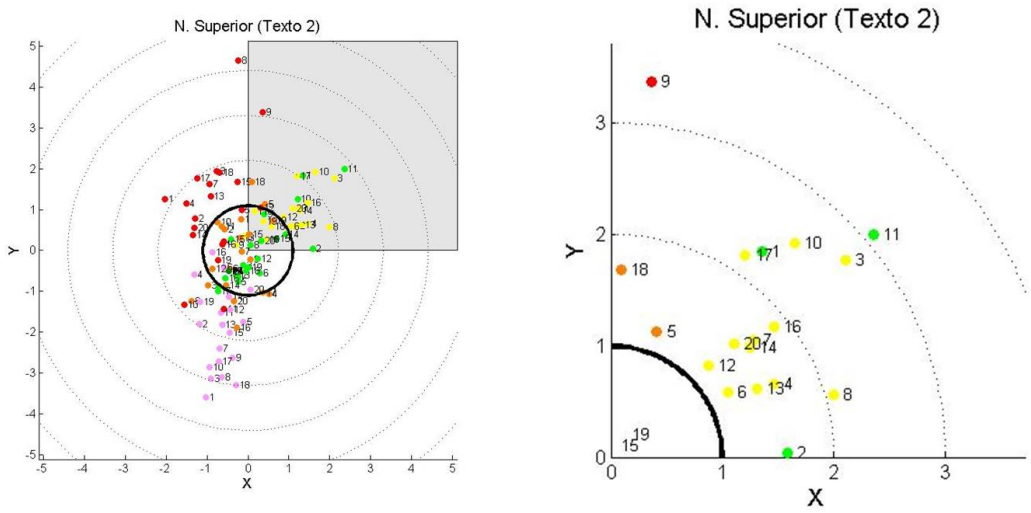


Figura 7.10: Nivelación sintáctica de cada texto de la opción 2ª de los exámenes del DELE superior.

7.2.4. Identificación semántica automática

7.2.4.1. Método de los campos semánticos

Esta metodología se aplica a los exámenes de DELE intermedio para hallar los campos semánticos ponderados en cada uno de los exámenes. Después se comparan estos campos

con los de un *corpus* de referencia que representa un modelo situacional o de registro similar a los exámenes. Lo ideal sería disponer de un *corpus* referencial, diseñado a modo de *rubric* u hoja de especificaciones, que permitiese la identificación *ad hoc* de los textos que se van a evaluar con un modelo lingüístico, sistémico-funcional (Halliday, 1985) o diatípico (Conrad y Biber, 2009) esperable en cada tipo de texto que se vaya a diagnosticar. En nuestro caso, el *corpus* referencial, son los 114 textos escritos por los estudiantes adultos, nativos españoles. Estos textos, escritos también con un tipo de formato similar, tratan sobre tres temas idénticos a los requeridos en el examen del 23 de mayo de 2008: la reclamación del equipaje extraviado por una compañía aérea, la carta a un *penpal* y la descripción de un lugar especial. Además, en estos textos españoles también se han hallado los campos semánticos ponderados, según la configuración establecida por el *PCIC* para poder comparar resultados.

Respecto a los exámenes del DELE intermedio, se analizan los 40 textos conjuntamente. Sin embargo, 20 textos pertenecen a la primera parte del examen y otros 20 textos a la segunda.

En la figura 7.11, presentamos los valores de las correlaciones que se establecen entre los 40 exámenes de los candidatos y los 114 textos de los españoles. La diagonal, en blanco, es la comparación de cada texto consigo mismo, así que vamos a observar el área de la sección derecha limitada por la diagonal en el cuadro. En realidad, las dos mitades de la figura, a ambos lados de la diagonal, son iguales. La sección que presentamos en color amarillo pertenece a los exámenes de DELE, y las de colores azul, verde y rojo representan las áreas de los textos españoles. Los valores, en gama de grises, muestran el grado de correlación de los exámenes intermedios de DELE con los textos escritos por los estudiantes españoles adultos, después de haber sido hallados los campos semánticos de todos ellos. En esta figura se aprecian tres agrupamientos por temas pertenecientes a los textos de los españoles en los sectores más claros adyacentes a la diagonal. Los elementos de los cuadros de colores indican los valores de correlación entre los exámenes de DELE y los textos de los españoles, limitando los cuadros de colores cada uno de los tres temas de los textos. Respecto a los valores de correlación representados en esta figura, cuanto más claro es el gris, más alta es la correlación. En los exámenes de DELE se alcanza un rango de correlación entre 0,35 y 0,74 (ver tablas 7.3 y 7.4).

Otra representación más simplificada y explícita de la figura anterior es la 7.12. En esta figura se especifica automáticamente qué exámenes del Texto-1 (textos de la 1ª parte del DELE de la convocatoria del 23 y 24 de mayo de 2008) y del Texto-2 (textos de la 2ª parte del DELE de la convocatoria del 23 y 24 de mayo de 2008) tratan sobre una de las temáticas del *corpus*. De esta manera se distingue qué exámenes representan qué opción dentro de cada parte del examen. Es decir, sabemos qué textos se refieren a la carta de *penpal*, a la reclamación (1ª y 2ª opción de la primera parte del examen) o al lugar especial (1ª opción de la segunda parte), y qué exámenes contienen otras opciones o se encuadran en alguna de las especificadas por su proximidad de contenido. En esta figura 7.12 se identifican aquellos exámenes cuyos campos semánticos están más próximos entre sí en relación a los del *corpus* de referencia. Por un lado, se simboliza con el signo "+" al examen que se identifica claramente con el *corpus* de referencia; por otro, el mayor o menor grosor del punto azul representa el índice de fiabilidad. Este grado de fiabilidad viene dado por la mayor o menor correlación de un examen con la media de los distintos

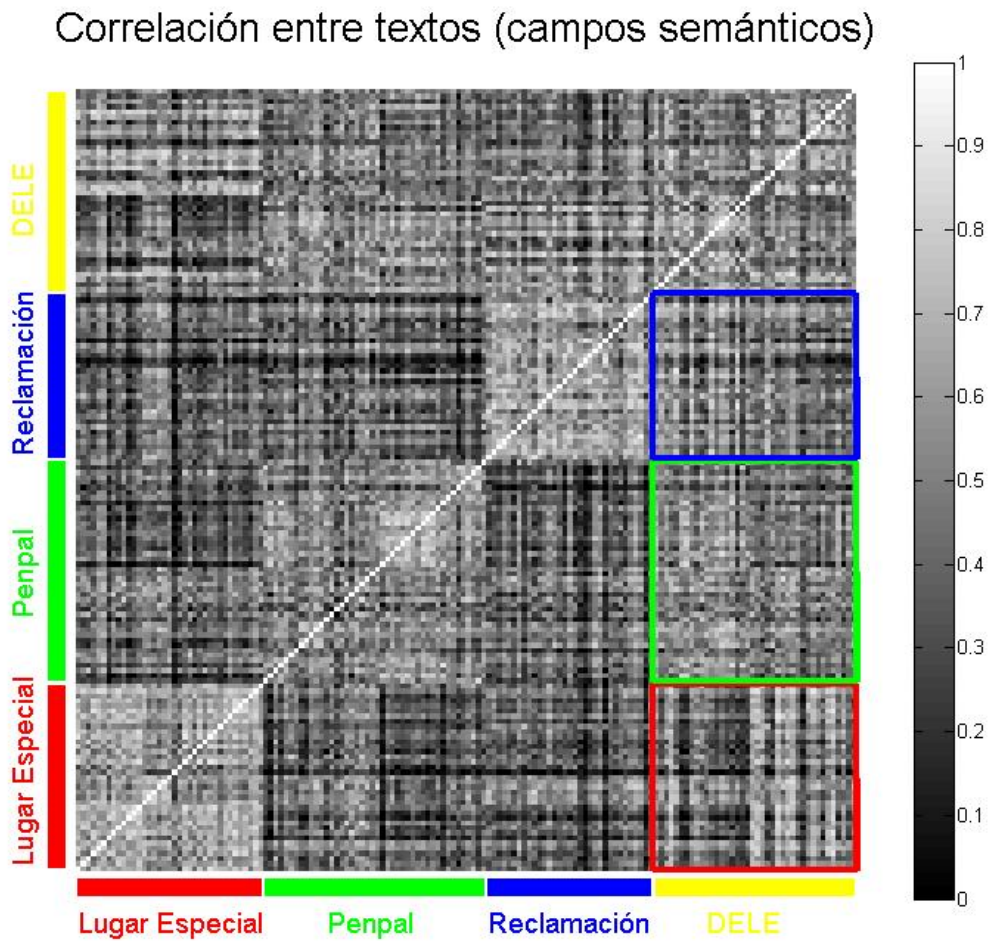


Figura 7.11: Correlación de los exámenes de DELE intermedio con los campos semánticos del *PCIC*.

campos semánticos.

Aunque vamos a presentar los valores automáticos junto a los de los calificadores expertos, los valores automáticos no son idénticos a los de los correctores humanos. Es decir, el criterio de contenido semántico, identificado automáticamente, difiere conceptualmente del de adecuación del texto, calificado por los expertos. Recordamos que la adecuación del texto incluye no sólo un contenido ajustado al tema, sino la función del mensaje y a quién y cómo se trasmite. Sin embargo, nuestro criterio semántico sólo identifica la temática del contenido esperada a través del procesamiento de su léxico como criterio más próximo al de los correctores. No obstante, consideramos que la identificación semántica con la adecuación es lo más próximo que tenemos y nos parece la más acertada.

Cuando analizamos los datos del campo semántico en la figura 7.11 y su fiabilidad en la figura 7.12, observamos que los valores están muy próximos a los de los correctores. En las tablas 7.3 y 7.4 se representan estos valores de manera aún más explícita en forma de comparativa. Sabemos que las notas de los dos correctores no pueden ser mayores del

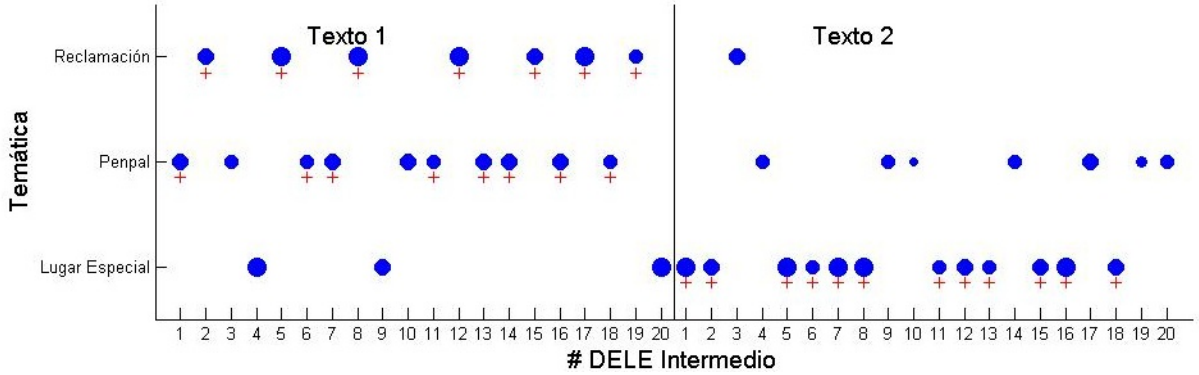


Figura 7.12: Índice de proximidad del contenido de los campos semánticos y la fiabilidad de los exámenes de DELE intermedio.

valor 3. Como hay dos valores, uno de cada corrector, se ha hecho la media entre las dos y se ha normalizado, no pudiendo ser su valor total superior a 1 en nuestra comparativa. Respecto a los exámenes, tanto para el Texto-1 como para el Texto-2, nuestros "Índices de correlación" automáticos oscilan entre 0,35 (la nota más baja) y 0,74 (la nota más alta). Ambos extremos se hallan en el Texto-2 (Tabla 7.4).

TEXTO 1	ADECUACIÓN		Nota media (máx. 3)	CAMPOS SEMÁNTICOS		CONTENIDO
Candidato #	Corrector 1	Corrector 2	Normalizada	Ind. Corr.	Calculados	Real
1	2	2	0,67	0,63	Penpal	Penpal
2	1	1	0,33	0,62	Reclamación	Reclamación maleta
3	2	2	0,67	0,53	*Penpal*	Invitación viaje
4	2	3	0,83	0,72	*Lugar Especial*	Invitación viaje
5	2	2	0,67	0,67	Reclamación	Reclamación maleta
6	2	2	0,67	0,56	Penpal	Penpal
7	3	3	1,00	0,59	Penpal	Penpal
8	1	3	0,67	0,72	Reclamación	Reclamación maleta
9	3	3	1,00	0,57	*Lugar Especial*	Invitación viaje
10	3	3	1,00	0,58	*Penpal*	Invitación viaje
11	2	2	0,67	0,53	Penpal	Penpal
12	3	3	1,00	0,68	Reclamación	Reclamación maleta
13	3	3	1,00	0,58	Penpal	Penpal
14	3	3	1,00	0,63	Penpal	Penpal
15	3	3	1,00	0,60	Reclamación	Reclamación maleta
16	2	3	0,83	0,62	Penpal	Penpal
17	1	2	0,50	0,69	Reclamación	Reclamación maleta
18	3	3	1,00	0,54	Penpal	Penpal
19	3	3	1,00	0,56	Reclamación	Reclamación maleta
20	3	3	1,00	0,68	*Lugar Especial*	Invitación viaje

Tabla 7.3: Índice de correlación semántica de los exámenes de DELE del Texto-1.

Respecto al Texto-1, en la tabla 7.3, observamos que los índices de correlación de todos los exámenes de los candidatos está por encima de 0,52, incluso el del candidato 2 que tiene la nota inferior dada por los correctores. No podemos establecer un umbral de apto de forma automática sino simplemente presentar los valores de los índices de los correctores humanos y del automático e interpretarlos. Por ello, observamos que nuestros valores marcan en su mayoría una correlación media-alta del examen analizado respecto al *corpus* de referencia.

Se aprecia, además, que los exámenes de los candidatos # 3, 4, 9, 10 y 20, que optaron

por la carta de invitación a un amigo a un viaje, se han identificado con el contenido de "*penpal*" o "lugar especial". Como todos los textos son cartas a un amigo en las que se habla de lugares de procedencia o de destinos de viaje, este hecho justifica la proximidad temática de estos exámenes 3, 4, 9, 10 y 20.

TEXTO 2	ADECUACIÓN		Nota media (máx. 3)	CAMPOS	SEMÁNTICOS	CONTENIDO
Candidato #	Corrector 1	Corrector 2	Normalizada	Ind. Corr.	Calculados	Real
1	3	3	1,00	0,69	Lugar Especial	Lugar Especial
2	2	3	0,83	0,62	Lugar Especial	Lugar Especial
3	3	3	1,00	0,63	*Reclamación*	Profesión favorita
4	3	3	1,00	0,52	*Penpal*	Profesión favorita
5	3	3	1,00	0,73	Lugar Especial	Lugar Especial
6	3	3	1,00	0,53	Lugar Especial	Lugar Especial
7	3	3	1,00	0,67	Lugar Especial	Lugar Especial
8	3	3	1,00	0,70	Lugar Especial	Lugar Especial
9	3	2	0,83	0,52	*Penpal*	Amistades
10	3	3	1,00	0,35	*Penpal*	Profesión favorita
11	3	3	1,00	0,52	Lugar Especial	Lugar Especial
12	3	3	1,00	0,64	Lugar Especial	Lugar Especial
13	3	3	1,00	0,53	Lugar Especial	Lugar Especial
14	3	3	1,00	0,48	*Penpal*	1ª impresión
15	3	3	1,00	0,62	Lugar Especial	Lugar Especial
16	3	3	1,00	0,74	Lugar Especial	Lugar Especial
17	1	0	0,17	0,57	*Penpal*	1ª impresión
18	3	3	1,00	0,58	Lugar Especial	Lugar Especial
19	3	3	1,00	0,47	*Penpal*	1ª impresión
20	3	3	1,00	0,51	*Penpal*	Profesión favorita

Tabla 7.4: Índice de correlación semántica de los exámenes de DELE del Texto-2.

En cuanto al Texto-2, los datos figuran en la tabla 7.4. Concretamente, destacamos que el índice de correlación más bajo, 0,35, es el del candidato 10. Sin embargo, este examen obtiene la nota más alta por el corrector humano (1,00). La explicación la hallamos en que el texto del candidato 10 trata de una "profesión favorita" y no hay *corpus* de referencia específico con ese contenido. Temáticamente, el asunto más próximo al que se puede correlar es al de "*penpal*" ya que en ese tipo de carta uno suele informar a su receptor sobre el tipo de actividad que realiza en su vida cotidiana.

Igualmente, en general, en esta tabla 7.4 apreciamos diferencias de los campos semánticos asignados a los exámenes de los candidatos 3, 4, 9, 10, 14, 17, 19 y 20. La primera razón es que los exámenes se corresponden con los del día 24 de mayo y se proponen temas diferentes. Segunda, no tenemos *corpus* de referencia para los exámenes relativos al Texto-2. Entre estos exámenes encontramos una redacción sobre una profesión favorita, un escrito de opinión sobre la primera impresión que tenemos de alguien, y un ensayo sobre las preferencias y hábitos de socialización con los amigos. Por ello, al identificarse cada examen con un *corpus* de referencia, la mayoría se asocia al de "*penpal*", excepto el candidato 3. Esto es, si los otros temas se relacionan con el de "*penpal*", es porque a un *penpal* se le cuentan temas muy diversos. Pero llama la atención que el examen del candidato 3 con el tema de "profesión favorita" se correle con el de "reclamación". Revisamos manualmente a qué se debe este resultado y comprobamos que la "profesión favorita" del candidato 3 es la de ser piloto y todo lo relacionado con la aviación. En consecuencia, el Texto-2 del candidato 3 se asocia al *corpus* de referencia de "reclamación" porque está temáticamente relacionado. Recordemos que en la "reclamación" se solicita a la compañía área que gestione el extravío de una maleta después de un viaje en avión.

En conclusión, este método da muy buenos resultados siempre que se tengan buenos *corpora* de referencia. Si los temas y los destinatarios de los textos están bien delimitados, los resultados serán óptimos.

7.2.4.2. Método de la LSA

El siguiente método para identificar el contenido semántico de los exámenes es el *Latent Semantic Analysis* (LSA). Se ha explicado y ejemplificado anteriormente en el apartado 3.2.3 cómo se aplica este método matemático al estudio de los textos. De nuevo, aplicamos este método sólo a los exámenes de DELE intermedio ya que carecemos de un *corpus* para analizar los exámenes de nivel superior. Una vez más, puesto que este método funciona mejor cuanto mayor es el conjunto de textos que se analizan, se toman conjuntamente los exámenes del DELE intermedio y el *corpus* de referencia redactado por el grupo de adultos nativos, compuesto por la carta de penpal (44 textos), la reclamación de la maleta (33 textos) y el lugar especial (37 textos).

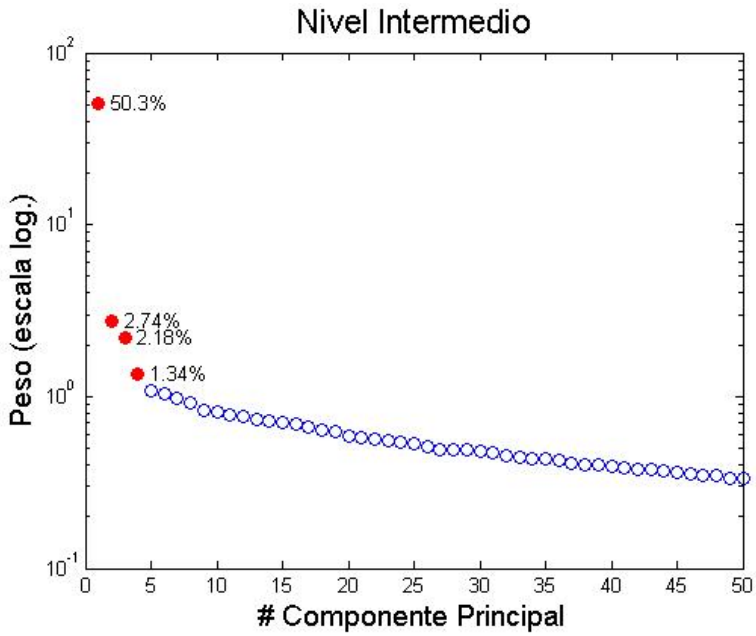


Figura 7.13: Pesos de los autovalores o vectores en los exámenes de DELE intermedio.

Después de aplicarse el método del LSA, obtenemos cuatro vectores. En la figura 7.13 estos cuatro vectores muestran el porcentaje de lemas comunes de todos los exámenes en todos los textos analizados: los de los candidatos y los de los alumnos españoles. De nuevo, igual que como veíamos en los discursos del Rey, los vectores que nos van a ser válidos para visualizar e interpretar mejor el contenido semántico son el vector #2 y #3 (Landauer *et al.*, 2004). Debemos ratificar que los valores de los cuatro vectores obtenidos en la figura 7.13, tanto antes como después de aplicar la *stoplist* (ver tabla 7.5), no afectan a la representación del contenido semántico. Este hecho ya pudo comprobarse en el análisis semántico de los discursos del Rey, tal y como se mostró en la tabla 3.22.

Dado que los vectores propios más relevantes para la interpretación de los resultados son el v#1, v#2 y v#3, son éstos los que representamos en la figura 7.14. Cada color

Componente	Discursos completos	Discursos aplicando la *stoplist*
#1	50,3 %	29,14 %
#2	2,74 %	4,05 %
#3	2.18 %	3,31 %
#4	1,34 %	1,88 %

Tabla 7.5: Porcentajes o pesos de los componentes principales #1 a #4 para los exámenes de DELE intermedio antes y después de aplicar la *stoplist*.

de las circunferencias representa un vector. La circunferencia azul representa el vector v#1 donde se concentra el mayor número de lemas del conjunto de textos sintéticos. La circunferencia verde simboliza el vector v#2 y la roja el v#3. Al calcular la distancia entre los dos vectores v#2 y v#3, la diferencia resultante permite discernir qué examen se corresponde con qué contenido semántico y en qué medida.

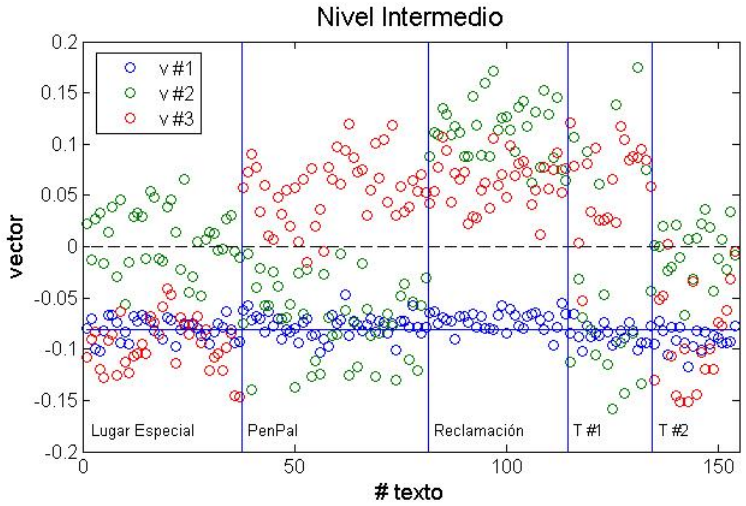


Figura 7.14: Vectores propios de los exámenes del DELE intermedio y de los textos de los nativos españoles.

El cálculo de estos vectores v#2 y v#3 aplicado a todos estos textos permite distinguir dónde se sitúan los exámenes en todo el conjunto de textos. En la figura 7.15 se visualiza la ubicación de cada examen en el conjunto de textos procesado y su mayor o menor identificación con este conjunto. Esta organización de los exámenes por grupos o *clusters* es el resultado final que proporciona la aplicación del análisis de los componentes principales a un gran conjunto de textos para hallar su contenido semántico.

Como en el método anterior, se han procesado sólo los exámenes de nivel intermedio. Estos exámenes son una muestra de la eficacia de este modelo de identificación semántica de un texto. Se han procesado sólo los exámenes de nivel intermedio porque, para que funcione el método, se precisa de una gran cantidad de texto de referencia. En este caso, como referencia, sólo contamos con el *corpus* de tres tipos de textos escritos por el grupo de nativos españoles. Como se observa en la figura 7.15, los textos de los españoles se simbolizan con las nubes de puntos de colores: las 44 cartas de penpal se representan con puntos azules; las 33 reclamaciones de la maleta, con puntos verdes; y los 37 textos sobre

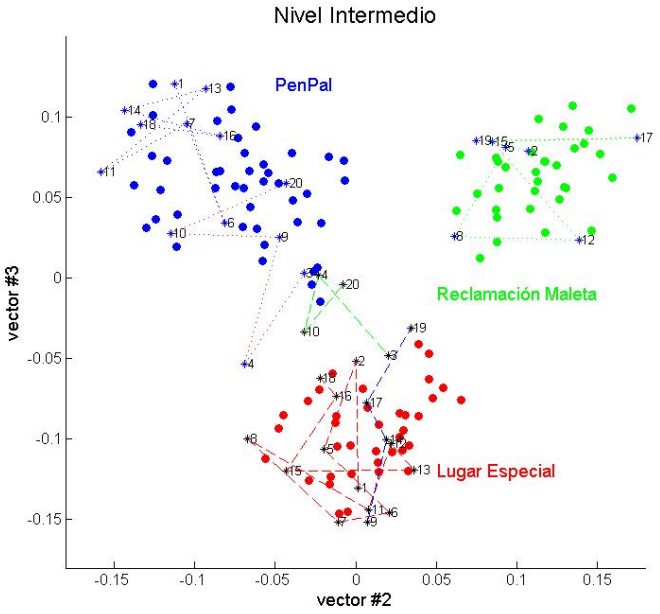


Figura 7.15: Mapa temático de todos los exámenes de nivel intermedio.

un lugar especial, con puntos rojos. En torno a estos puntos de referencia, se agrupan todos los exámenes de DELE intermedio, representados por asteriscos numerados. Los exámenes del DELE intermedio suponen en nuestro caso un total de 40 cuarenta textos. Todo el conjunto de exámenes se identifican bien dentro del Texto-1 o bien dentro del Texto-2. El Texto-1 se representa por líneas de puntos y el Texto-2 por líneas discontinuas que unen los asteriscos numerados. Cada uno de esos números identifica los exámenes de los candidatos para cada una de las dos opciones. Por ello se repite cada número dos veces.

Los exámenes correspondientes al Texto-1 son la carta de "*penpal*" con 8 textos y los de la "reclamación maleta" con 7 textos; los pertenecientes al Texto-2 son los del "lugar especial" con 12 exámenes. En torno a estos temas de referencia es donde se distribuyen los exámenes. Por eso, cuando los exámenes no se asocian a uno de los grupos de contenido anteriores, es porque existe un cambio de tema en algunos exámenes. Entonces, este cambio se aprecia cuando se generan subconjuntos de temas diferentes o alejamientos del grupo de referencia.

En la figura 7.15, el cambio temático se observa en los exámenes del Texto-1 cuando tratan temas de otras opciones como una carta de invitación a un viaje (5 textos: # 3, 4, 9, 10 y 20). También, hay un cambio temático cuando el Texto-2 trata de una redacción sobre una profesión favorita (4 textos: # 3, 4, 10, 20) o sobre las amistades (1 texto: #9) o trata de un escrito de opinión acerca de la primera impresión sobre alguien (4 textos: #14, 17 y 19).

Es decir, lo que representan estos textos apartados es que, cuando se separan del conjunto de referencia, tratan sobre un tema diferente. Por ello, hay exámenes, como los 5 exámenes anteriores del Texto-1, que se posicionan en áreas límites (# 9, 10 y 20) o alejadas (# 3, 4). Otros, como los del Texto-2, se ubican en áreas diferentes (#4, 10, 20) o alejados y dentro de otro grupo temático (#3) e, incluso, inmersos en otra área temática

distinta de la propia como son los exámenes #14, 17 y 19.

Comprobamos manualmente esos alejamientos y, efectivamente, los temas de esos exámenes no se corresponden con los textos del *corpus* de referencia. En el Texto-1, los exámenes # 3, 4, 9, 10 y 20 están a caballo entre dos nubes, aunque todos tratan sobre "invitación a un viaje", el # 3, 4 tienden al "lugar especial" mientras los candidatos # 9, 10 y 20 se quedan alejados dentro del área de "*penpal*".

De forma similar se comportan los exámenes del Texto-2 : #3, 4, 10, 20. Como su temática es "profesión favorita", los # 4, 10, 20 se reagrupan entre ellos, pero el #3 tiende hacia "lugar favorito" y la "reclamación maleta". Revisamos este examen #3 y leemos que la profesión favorita del candidato es la de ser piloto y hacer actividades relacionadas con aviones. También los exámenes #14, 17 y 19 son de temática diferente a la del *corpus* de referencia. Éstos tratan sobre la "primera impresión" que tenemos de alguien. Sin embargo, están bien integrados en "penpal". La razón es que estos exámenes describen tipos de personas, caracteres y comportamientos de personas que luego se hacen sus amigas, lo cual es un tema propio en un texto de *penpal*.

En definitiva, desde una perspectiva visual, deducimos que en aquellos casos en los que se produce un desvío del tema o un alejamiento de la "nube", se debería revisar el examen por un experto para comprobar por qué un texto difiere del resto, cuando se espera que todos traten sobre un mismo tema.

Desde una perspectiva objetivamente cuantitativa, también disponemos de unos valores que indican la mayor o menor aproximación de un examen a un tema. Los valores que hallamos con el LSA son la distancia de cada examen al valor medio calculado de cada *cluster* temático. En la figura 7.16 se representan estas distancias de correlación entre un examen y su correspondiente *cluster*. Los valores numéricos de las distancias se pueden ver en las tablas A.13 y A.14 del apéndice. Se considera que la distancia menor que se obtiene del LSA es la que identifica mejor un examen a su *cluster*. Los círculos de colores azul, rojo y verde simbolizan la media de los textos de los nativos españoles que hacen de *corpus* de referencia. La elipse que rodea cada círculo puede considerarse como el umbral de identificación semántica debido al *corpus* y aplicable a los textos analizados. Cada circunferencia de color azul (*penpal*), roja (lugar especial) o verde (reclamación) representa cada examen, y estas circunferencias se sitúan en función de aquellos grupos con los que mejor se correlan.

No obstante, tanto para el método de LSA como para el método de los campos semánticos, la disponibilidad de muchos textos que traten sobre el mismo tema es un requisito válido porque todos los textos funcionan como *corpus* de referencia. Este hecho nos ha permitido disponer de un *corpus* específico para identificar parte de la variedad de exámenes con temática diferente. Por ello, no sólo no hemos podido correlar con precisión algunos exámenes sino que incluso no hemos podido identificar los exámenes de DELE superior por carecer de un *corpus* de referencia para este nivel. En realidad, el buen funcionamiento de nuestro *corpus* muestra la inexistencia de todos los temas que concurren en los exámenes. Por ello, algunos exámenes se correlan con *clusters* que no les corresponden, aunque se parecen, porque no tenemos un *corpus* completo. Un ejemplo de la compatibilidad de los dos métodos son los resultados que vemos en las tablas 7.6 y 7.7. En ambas tablas se contrastan los resultados automáticos con el real. Tras esta corrección

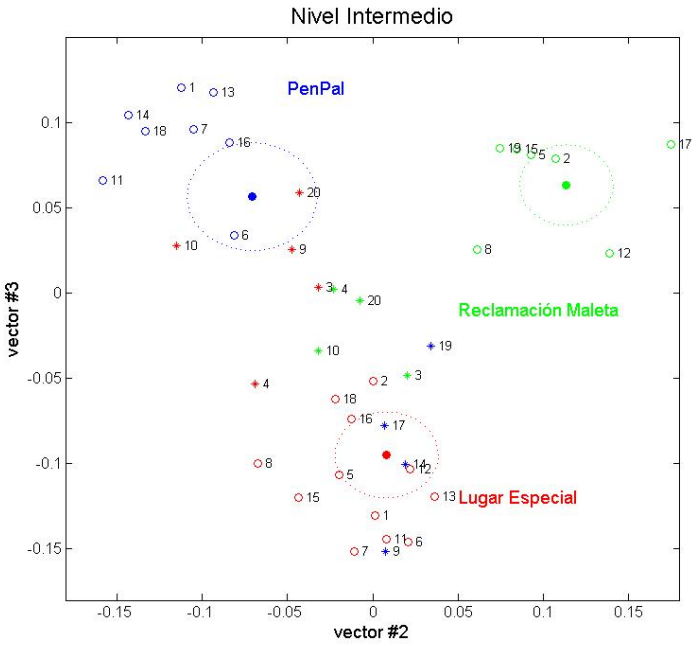


Figura 7.16: Representación de las distancias de los exámenes a sus correspondientes medias de los *clusters*.

del conjunto de exámenes del Texto-1 y Texto-2, utilizamos tres símbolos ("OK", "-", " / ") para identificar los resultados: "OK", representa la conformidad; "-", una cierta similitud temática y "/ " una temática diferente entre los exámenes.

En conclusión, consideramos que los dos métodos expuestos, el de los campos semánticos y el del LSA, son compatibles ya que uno refuerza y ratifica al otro. Además, nuestro *corpus* de referencia es un modelo *ad hoc* y nuestros valores no son generalizables. Sin embargo, es factible establecer unas distancias válidas entre el *corpus* y cada examen para determinar su correlación (ver tablas A.13 y A.14). Como ya hemos insistido en este capítulo, el *corpus* debería estar compuesto por muchos más textos, configurado sin errores gramaticales y diseñado específicamente para identificar con fiabilidad un cierto grado de contenido temático en cada examen. Creemos que la utilización de un gran macrotexto monotemático para el análisis de microtextos no funcionaría tan bien como la de un gran *corpus* de numerosos textos sobre un mismo tema, estructura y función. Es decir, es fundamental configurar un *corpus* compuesto con un gran número de textos, destinados a un receptor concreto y que traten sobre un tema bien acotado, como hemos pretendido con nuestro *corpus*.

TEXTO-1	CORRECCIÓN			RESULTADOS
Candidato #	Campo semántico	LSA	Real	OK / -
1	Penpal	Penpal	Penpal	OK
2	Reclamación	Reclamación	Reclamación	OK
3	Penpal	Penpal	Invitación viaje	-
4	Lugar Especial	Lugar especial	Invitación viaje	-
5	Reclamación	Reclamación	Reclamación	OK
6	Penpal	Penpal	Penpal	OK
7	Penpal	Penpal	Penpal	OK
8	Reclamación	Reclamación	Reclamación	OK
9	Lugar especial	Penpal	Invitación viaje	-
10	Penpal	Penpal	Invitación viaje	-
11	Penpal	Penpal	Penpal	OK
12	Reclamación	Reclamación	Reclamación	OK
13	Penpal	Penpal	Penpal	OK
14	Penpal	Penpal	Penpal	OK
15	Reclamación	Reclamación	Reclamación	OK
16	Penpal	Penpal	Penpal	OK
17	Reclamación	Reclamación	Reclamación	OK
18	Penpal	Penpal	Penpal	OK
19	Reclamación	Reclamación	Reclamación	OK
20	Lugar especial	Penpal	Invitación viaje	-

Tabla 7.6: Comparativa de las correcciones automáticas y manuales de los exámenes de DELE intermedio. Texto-1.

7.3. Discusión de los resultados

La comparativa entre los resultados obtenidos por un corrector experto y un corrector humano se presenta en las tablas 7.8 y 7.9. En estas tablas se reproducen los resultados comparados: en color verde con el signo "=", se simbolizan los valores similares de ambos correctores, el experto y el automático; en color rojo con el signo "-", aquellos resultados que el corrector experto califica con un valor más bajo que el automático; y en color azul con el signo "+", aquellos que el corrector experto califica más alto que el automático.

Es importante resaltar que los valores numéricos correspondientes al corrector experto no son equivalentes a los valores del calificador automático. Aunque los dígitos coinciden aparentemente, su significado difiere. El calificador humano puntúa entre 1 y 4 tanto el vocabulario como la gramática. Sólo se aprueban estas categorías cuando se supera el umbral, el 70 % de la puntuación máxima. Es decir, un 2,80 es el mínimo para obtener la calificación de apto.

Sin embargo, los valores automáticos que, van desde 1 hasta 6, representan los niveles de referencia desde el A1 al C2, correspondiéndose los valores con los detallados en la tabla 7.10.

Es preciso comentar ciertos resultados relativos al léxico y a la sintaxis de los exámenes del nivel intermedio. Como puede observarse en la tabla 7.8 del nivel intermedio y en la tabla 7.9 del nivel superior, el calificador experto ha valorado de forma distinta que el

TEXTO-2		CORRECCIÓN		RESULTADOS
Candidato #	**Campo Semántico**	**LSA**	**Real**	OK / -
1	Lugar especial	Lugar especial	Lugar especial	OK
2	Lugar especial	Lugar especial	Lugar especial	OK
3	Reclamación	Lugar especial	Profesión favorita	-
4	Penpal	Penpal	Profesión favorita	/
5	Lugar especial	Lugar especial	Lugar especial	OK
6	Lugar especial	Lugar especial	Lugar especial	OK
7	Lugar especial	Lugar especial	Lugar especial	OK
8	Lugar especial	Lugar especial	Lugar especial	OK
9	Penpal	Lugar especial	Amistades	-
10	Penpal	Lugar especial	Profesión favorita	-
11	Lugar especial	Lugar especial	Lugar especial	OK
12	Lugar especial	Lugar especial	Lugar especial	OK
13	Lugar especial	Lugar especial	Lugar especial	OK
14	Penpal	Lugar especial	1ª impresión	-
15	Lugar especial	Lugar especial	Lugar especial	OK
16	Lugar especial	Lugar especial	Lugar especial	OK
17	Penpal	Lugar especial	1ª impresión	-
18	Lugar especial	Lugar especial	Lugar especial	OK
19	Penpal	Lugar especial	1ª impresión	-
20	Penpal	Penpal	Profesión favorita	/

Tabla 7.7: Comparativa de las correcciones automáticas y manuales de los exámenes de DELE intermedio. Texto-2.

calificador automático. En el nivel intermedio el calificador experto valora más bajo que el automático, mientras que en el nivel superior ocurre lo contrario. Por tanto, consideramos que algunos resultados automáticos podrían parecer sesgados tanto para el léxico como para la sintaxis. Sin embargo podemos concluir que el evaluador automático califica siempre con los mismos criterios, mientras que el corrector humano está expuesto a otros factores.

Respecto al léxico, aunque calificamos según los niveles de referencia propuestos por el *Plan Curricular de Instituto Cervantes* (*PCIC*), es importante recordar que el glosario actual del *PCIC* califica entre un 50-60 % del léxico de un texto y que se obtiene más de un 80 % de vocablos nivelados en la mayoría de los textos gracias al criterio de nivelación por combinación de diccionarios. Ese aumento de un 30 % de vocablos nivelados se obtiene con el criterio de combinación de todos los glosarios que exponíamos en el apartado 4.1.4, incluido el glosario de multi-vocablos nivelados. Consideramos que, aunque la combinación de glosarios para nivelar es bastante satisfactoria, la nivelación léxica automática agrupa muchos más vocablos en el nivel B1. Por tanto, en la nivelación resultante, es lógico que la calificación combinada tienda a la baja. Este fenómeno de infranivelación de vocablos de B2 como B1 o de supranivelación de C2 para vocablos de C1 lo observamos y apuntamos con anterioridad en el apartado 4.1.4 y lo ilustramos en su momento con la figura 4.4.

Respecto a la sintaxis, por ejemplo, apreciamos que la nivelación automática es más baja que la de los expertos. Apuntamos dos posibles razones. Una razón plausible es

Intermedio

	Léxico					Sintaxis				
# Candidato	Experto		Automático			Experto		Automático		
1	Apto	4	3	B1	=	Apto	3,25	3,5	B1-B2	=
2	Apto	3,25	3	B1	=	Apto	3	3	B1	=
3	Apto	3,5	3	B1	=	Apto	3	3	B1	=
4	No apto	2,5	3	B1	-	No apto	2,5	2,5	A2-B1	=
5	No apto	2,75	3	B1	-	Apto	3,25	3	B1	=
6	Apto	3,5	3,5	B1-B2	=	Apto	3,25	2	A2	+
7	Apto	3,75	3,5	B1-B2	=	Apto	3	1	A1	+
8	Apto	3	3,5	B1-B2	=	No apto	2,75	3	B1	-
9	Apto	3	4	B2	=	No apto	2,5	3	B1	-
10	Apto	3	3	B1	=	No apto	2,75	3	B1	-
11	No apto	2,75	4	B2	-	No apto	2,5	3	B1	-
12	Apto	3,25	3	B1	=	Apto	3	3	B1	=
13	No apto	2,5	3,5	B1-B2	-	No apto	2,5	2,5	A2-B1	=
14	Apto	4	3,5	B1-B2	=	Apto	3,75	3,5	B1-B2	=
15	Apto	4	3,5	B1-B2	=	Apto	4	3	B1	=
16	No apto	2,75	2,5	A2-B1	=	No apto	2,75	1,5	A1-A2	=
17	Apto	3,25	3,5	B1-B2	=	No apto	2	3,5	B1-B2	-
18	Apto	4	3,5	B1-B2	=	Apto	4	3	B1	=
19	No apto	2,75	3	B1	-	No apto	1,5	3	B1	-
20	Apto	3	4	B2	=	Apto	3	2,5	A2-B1	+

Tabla 7.8: Comparativa de los resultados totales, léxicos y sintácticos, de los exámenes del DELE intermedio.

que gran parte de las estructuras sintácticas listadas se corresponden con los niveles de A2 y B1, y aunque existen un número importante de B2, existen muy pocas estructuras de C1 y C2; otra razón de nivelación con tendencia a la baja es que hay textos cuya calificación automática da valor "0". Esta valoración automática de "0" es una valoración cautelar de la fiabilidad del programa ya que partimos del postulado de que los discursos navideños del Rey son sintácticamente de un nivel C1. Así que, para aquellos exámenes cuyas estructuras sintagmáticas tienen gran proximidad a este *corpus* de referencia, se ha programado Sintactor para dar un valor "0". Por tanto, estos exámenes podrían calificarse como de C1, como el propio *corpus*.

7.3.1. Nivel intermedio

7.3.1.1. Léxico y sintaxis

Los valores detallados y la nota media dados por los correctores expertos al vocabulario y la gramática se observan en la tabla 7.1, mientras que los valores automáticos detallados de los exámenes intermedios para el léxico se especifican en las tablas A.2 y A.5, y los de la sintaxis en las tablas A.7 y A.10. Estas cuatro últimas tablas están ubicadas en el apéndice. Sin embargo, las medias de los correctores expertos y del calificador automático comparados se pueden ver en la tabla 7.8 de este apartado.

Superior

	Léxico					Sintaxis				
# Candidato	Experto		Automático			Experto		Automático		
1	No apto	2,5	2	A2	=	No apto	2,5	4	B2	-
2	No apto	2,25	3,5	B1-B2	=	No apto	1,5	2,5	A2-B1	=
3	No apto	2,25	3	B1	=	No apto	2	3	B1	=
4	No apto	2,75	2,5	A2-B1	=	Apto	3	3,5	B1-B2	+
5	No apto	2,25	3	B1	=	No apto	2,25	2	A2	=
6	No apto	2	3	B1	=	No apto	2	3	B1	=
7	Apto	3,75	2,5	A2-B1	+	Apto	3,5	3,5	B1-B2	+
8	Apto	4	3	B1	+	Apto	3,5	2,5	A2-B1	+
9	No apto	2,25	4	B2	-	No apto	2,75	2	A2	=
10	Apto	3,75	4	B2	=	Apto	4	3	B1	+
11	No apto	2,5	3	B1	=	No apto	2,75	4,5	B2-C1	-
12	Apto	3,5	4	B2	=	Apto	3,25	4	B2	=
13	Apto	3,75	3,5	B1-B2	+	Apto	3,5	4,5	B2-C1	=
14	No apto	2,25	3,5	B1-B2	=	No apto	1,75	3,5	B1-B2	=
15	Apto	3	3	B1	+	Apto	3,25	2	A2	+
16	Apto	3,5	3	B1	+	Apto	3,25	3,5	B1-B2	+
17	Apto	3,75	3	B1	+	Apto	3,5	3	B1	+
18	No apto	2,75	3,5	B1-B2	=	No apto	2,75	3,5	B1-B2	=
19	No apto	2,25	3	B1	=	No apto	2,25	2	A2	=
20	Apto	3,25	3	B1	+	No apto	2,75	3	B1	=

Tabla 7.9: Comparativa de los resultados totales, léxicos y sintácticos, de los exámenes del DELE superior.

En los exámenes intermedios, al comparar los valores otorgados por los correctores expertos con los resultados automáticos, observamos que la calificación léxica de los expertos no coincide en algunos casos porque la calificación automática del léxico es algo superior a la que otorga el experto. Con la simbología de signos y color utilizada en la tabla 7.8, mostramos en la comparación que, cuando el candidato es "apto" por el experto y tiene un B1 por el automático, el resultado es similar (=); cuando el candidato es "apto" y con el automático obtiene B1 o B1-B2 consideramos que la calificación del experto supera la automática (+); y cuando el experto califica con un "no apto" y con el automático el candidato obtiene un A2-B1, se considera equiparada la calificación (=). Como la comparación del resultado automático se hace tomando como referencia la nota del calificador humano, en general, podemos afirmar que la calificación automática coincide más veces que discrepa. En las discrepancias, el corrector automático califica más alto el léxico que los expertos. Entre los niveles léxicos, destacamos con el nivel más bajo el examen # 16; con nivel B1, los # 1, 2, 3, 4, 5, 10, 12 y 19; con B1-B2, los #6, 7, 8, 13, 14, 15, 17 y 18; y con B2, sólo los #9, 11 y 20.

Respecto a la calificación gramatical, la calificación del experto califica más alto los exámenes de los candidato #6, 7 y 20 que el corrector automático. Sin embargo, el automático da más valor a los exámenes # 8, 9, 10 y 11 por una pequeña diferencia, mientras que se discrepa bastante con los exámenes # 17 y 20. Igualmente, consideramos que, en general, los resultados comparados entre el automático y el experto se equiparan más que se diferencian.

Valor	Nivel
1	A1
1,5	A1-A2
2	A2
2,5	A2-B1
3	B1
3,5	B1-B2
4	B2
4,5	B2-C1
5	C1
5,5	C1-C2
6	C2

Tabla 7.10: Equivalencia entre valores numéricos y sus correspondientes calificaciones en los 6 niveles estándar.

Un ejemplo de suspenso muy claro y coincidente por los dos tipos de correctores es el del candidato #16 con un "no apto" tanto para el léxico como para la sintaxis.

7.3.1.2. Nivel semántico

En este nivel, como ya hemos anticipado, no comparamos la calificación de los correctores expertos con el resultado automático. Nosotros aportamos la identificación del contenido semántico con un *corpus* de referencia. Esto es, identificamos cuantitativamente la proximidad de un examen al *corpus* de referencia por su contenido semántico. Además, el criterio de adecuación que aplica el experto corrector no se basa en el criterio que articula el método del LSA, aunque en ambos criterios la concepción semántica es fundamental. En la tabla 7.11 se observa que diferenciamos los exámenes en Texto-1 y Texto-2 porque no tiene significado hallar una calificación media para cada opción. Tampoco es adecuado identificar el "OK" del corrector automático con el "apto" del experto. Por otro lado, suponemos que la nota umbral de "apto" corresponde a 2,1, que es el 70 % del valor máximo 3 en este apartado. Respecto al valor numérico del corrector automático, el mejor valor numérico es la menor distancia del examen al contenido semántico específico dentro del *corpus* de referencia. La menor distancia es la seleccionada entre el conjunto de valores de los tres *clusters* del *corpus* de referencia. Estos valores se registran en las tablas A.13 y A.14 .

7.3.2. Nivel superior

7.3.2.1. Léxico y sintaxis

Los valores detallados y la nota media del vocabulario y la gramática dados por los correctores expertos se observan en la tabla 7.2. A su vez, los valores automáticos de los exámenes intermedios para el léxico se especifican en las tablas A.17 y A.20; y los de la

DELE Intermedio

Adecuación		**Contenido automático**	
Texto-1	Texto-2	Texto-1	Texto-2
2	3	OK (2,31)	OK (1,44)
1	2,5	OK (0,70)	OK (1,73)
2	3	_ (1,97)	_ (1,90)
2,5	3	_ (3,04)	/ (2,13)
2	3	OK (1,07)	OK (1,03)
2	3	OK (0,77)	OK (2,06)
3	3	OK (1,54)	OK (2,35)
2	3	OK (2,49)	OK (2,50)
3	2,5	_ (1,16)	_ (2,26)
3	3	_ (1,47)	_ (2,77)
2	3	OK (2,32)	OK (1,96)
3	3	OK (1,93)	OK (0,56)
3	3	OK (2,02)	OK (1,34)
3	3	OK (2,44)	_ (0,42)
3	3	OK (1,38)	OK (1,97)
2,5	3	OK (1,07)	OK (1,08)
1,5	0,5	OK (2,45)	_ (0,68)
3	3	OK (2,06)	OK (1,63)
3	3	OK (1,69)	_ (2,67)
3	3	_ (0,73)	/ (2,54)

Tabla 7.11: Registro de calificaciones del DELE intermedio en el nivel semántico.

sintaxis, en las tablas A.21 y A.24. La cuatro últimas tablas están ubicadas en el apéndice. Además, las notas medias de los correctores expertos y del calificador automático, comparadas, se pueden ver en la tabla 7.9 de este apartado.

En este nivel observamos que los valores automáticos son más bajos en estos exámenes, tanto en el nivel léxico como sintáctico. De ahí que la nota de los calificadores expertos, cuando discrepa, es porque su nota es más alta (+) que la del calificador automático. Consideramos que la causa de este desajuste es la nivelación léxica que, al estar algo sesgada a la baja la nivelación por la combinación de diccionarios, el sistema automático tiende a marcar más léxico con nivel B1.

En relación a los resultados del léxico, como mostramos en la tabla 7.9, insistimos que ahora no distinguimos entre el Texto-1 y Texto-2 sino que comparamos las notas medias de ambos tipos de correctores para cada candidato. Como en el caso de los exámenes intermedios, los resultados comparados de los dos tipos de correctores, humanos y automáticos, son equiparables, y es mayor la similitud entre los resultados que la diferencia. Aunque el nivel léxico de los exámenes del DELE superior no asciende mucho respecto al intermedio, sí que observamos que hay más candidatos en el nivel superior que tienen niveles bajos de A2, A2-B1 (# 1, 4, 7) y de B1 (#3, 5, 6, 8, 11, 15, 16, 17 y 19), igual número de candidatos con nivel B2 (# 9, 10, 12) y menos que estén a caballo entre los dos niveles de B1-B2 (#2, 13, 14, 18). Esta poca variabilidad en el ascenso de nivel del intermedio al superior parece ratificar la afirmación de Laufer & Paribakht y Nation de que hay una diferencia entre el vocabulario que se conoce y el vocabulario que se plasma

en la realidad, ya que es difícil la aparición de vocablos poco frecuentes (Nation, 2001, 182), (Gass y Selinker, 2008, 452).

Con respecto a la sintaxis, también observamos que la calificación automática es más baja que la de los expertos. Como explicábamos arriba, la mayor cantidad de estructuras listadas con niveles más bajos parece sesgar los resultados a la baja. Sin embargo, destacamos al candidato #11 para mostrar que la diferencia de calificación entre el corrector humano y el automático es significativa. El calificador automático le otorgaría un "apto" pero para el experto humano, cuya valor mínimo para ser "apto" es 2,80, el #11 no pasa la prueba al obtener un 2,75.

No obstante, concluimos que la mayoría de los resultados léxicos y sintácticos coinciden con el juicio del experto humano; que la mayoría de los candidatos del nivel superior utiliza un vocabulario más frecuente y menos diferenciador de nivel. Además, resaltamos que el sistema automático mide todos los exámenes por igual, incluso con el sesgo sistemático que apuntábamos arriba.

7.3.2.2. Nivel semántico

Como ya adelantábamos, no podemos hacer la comparativa semántica para los exámenes de DELE superior porque no tenemos un *corpus* de referencia para el análisis semántico (ver tabla 7.12).

DELE Superior			
Adecuación		**Contenido automático**	
Texto-1	Texto-2	Texto-1	Texto-2
1	3		
3	3		
3	3		
1	1,5		
1	2,5		
2,5	3		
2,5	3		
3	3		
1	3		
3	3		
1	2,5		
2,5	3		
3	3		
2,5	3		
3	3		
3	3		
3	3		
1	3		
1	3		
3	3		

No hay corpus de referencia

Tabla 7.12: Registro de calificaciones del DELE superior en el nivel semántico.

7.3.3. Resumen de resultados

En las siguientes tablas se muestra un resumen de los resultados obtenidos en los distintos niveles de lengua de los exámenes de DELE.

Consideramos que estos resultados validan la hipótesis de esta tesis doctoral al otorgar una calificación del nivel de referencia estándar empleando métodos automáticos para el análisis de los textos. A la vez, las herramientas y entornos desarrollados en este trabajo son susceptibles de mejora para aumentar la eficacia y fiabilidad de esta evaluación automática.

Con todo ello, no se pretende la sustitución del evaluador humano sino tan sólo proporcionar un método alternativo y automatizable que pueda servir para la autoevaluación de los aprendices.

7.3.3.1. DELE Intermedio

Para este nivel se presentan los resultados en la tabla 7.13. Se resumen los valores del nivel léxico y sintáctico obtenido a partir de las herramientas automáticas y se comparan con las calificaciones dadas por los evaluadores expertos ("vocabulario" y "gramática"). También, esta misma tabla contiene los resultados obtenidos en el área semántica y se cotejan con los índices de calificación de los expertos. Para ello se ha tomado la categoría de "adecuación" por ser la más cercana al "contenido semántico".

INTERMEDIO	**Vocabulario**	**Léxico**		**Gramática**	**Sintaxis**		**Adecuación**		**Contenido automático**	
Candidato #	Expertos	Automático		Expertos	Automático		Texto-1	Texto-2	Texto-1	Texto-2
1	4 *	3	B1	3,25 *	3,5	B1-B2	2	3	OK (2,31)	OK (1,44)
2	3,25 *	3	B1	3 *	3	B1	1	2,5	OK (0,70)	OK (1,73)
3	3,5 *	3	B1	3 *	3	B1	2	3	_ (1,97)	_ (1,90)
4	2,5	3	B1	2,5	2,5	A2-B1	2,5	3	_ (3,04)	/ (2,13)
5	2,75	3	B1	3,25 *	3	B1	2	3	OK (1,07)	OK (1,03)
6	3,5 *	3,5	B1-B2	3,25 *	2	A2	2	3	OK (0,77)	OK (2,06)
7	3,75 *	3,5	B1-B2	3 *	1	A1	3	3	OK (1,54)	OK (2,35)
8	3 *	3,5	B1-B2	2,75	3	B1	2	3	OK (2,49)	OK (2,50)
9	3 *	4	B2	2,5	3	B1	3	2,5	_ (1,16)	_ (2,26)
10	3 *	3	B1	2,75	3	B1	3	3	_ (1,47)	_ (2,77)
11	2,75	4	B2	2,5	3	B1	2	3	OK (2,32)	OK (1,96)
12	3,25 *	3	B1	3 *	3	B1	3	3	OK (1,93)	OK (0,56)
13	2,5	3,5	B1-B2	2,5	2,5	A2-B1	3	3	OK (2,02)	OK (1,34)
14	4 *	3,5	B1-B2	3,75 *	3,5	B1-B2	3	3	OK (2,44)	_ (0,42)
15	4 *	3,5	B1-B2	4 *	3	B1	3	3	OK (1,38)	OK (1,97)
16	2,75	2,5	A2-B1	2,75	1,5	A1-A2	2,5	3	OK (1,07)	OK (1,08)
17	3,25	3,5	B1-B2	2	3,5	B1-B2	1,5	0,5	OK (2,45)	_ (0,68)
18	4 *	3,5	B1-B2	4 *	3	B1	3	3	OK (2,06)	OK (1,63)
19	2,75	3	B1	1,5	3	B1	3	3	OK (1,69)	_ (2,67)
20	3 *	4	B2	3 *	2,5	A2-B1	3	3	_ (0,73)	/ (2,54)

Tabla 7.13: Tabla conjunta de los resultados del DELE intermedio.

7.3.3.2. DELE Superior

Al igual que en el caso anterior, en la tabla 7.14 se incluyen los valores para el vocabulario y el léxico, y los correspondientes a la gramática y la sintaxis. Sin embargo, no

podemos cotejar los valores semánticos de los expertos con ninguna predicción automática ya que carecemos de un corpus de referencia para el análisis semántico de este nivel superior.

SUPERIOR	Vocabulario	Léxico		Gramática	Sintaxis		Adecuación		Contenido automático	
Candidato #	Expertos	Automático		Expertos	Automático		Texto-1	Texto-2	Texto-1	Texto-2
1	2,5	2	A2	2,5	4	B2	1	3		
2	2,25	3,5	B1-B2	1,5	2,5	A2-B1	3	3		
3	2,25	3	B1	2	3	B1	3	3		
4	2,75	2,5	A2-B1	3 *	3,5	B1-B2	1	1,5		
5	2,25	3	B1	2,25	2	A2	1	2,5		
6	2	3	B1	2	3	B1	2,5	3		
7	3,75 *	2,5	A2-B1	3,5 *	3,5	B1-B2	2,5	3		
8	4	3	B1	3,5	2,5	A2-B1	3	3		
9	2,25	4	B2	2,75	2	A2	1	3		
10	3,75*	4	B2	4 *	3	B1	3	3		
11	2,5	3	B1	2,75	4,5	B2-C1	1	2,5		
12	3,5 *	4	B2	3,25 *	4	B2	2,5	3		
13	3,75 *	3,5	B1-B2	3,5 *	4,5	B2-C1	3	3		
14	2,25	3,5	B1-B2	1,75	3,5	B1-B2	2,5	3		
15	3*	3	B1	3,25 *	2	A2	3	3		
16	3,5 *	3	B1	3,25 *	3,5	B1-B2	3	3		
17	3,75 *	3	B1	3,5 *	3	B1	3	3		
18	2,75	3,5	B1-B2	2,75	3,5	B1-B2	1	3		
19	2,25	3	B1	2,25	2	A2	1	3		
20	3,25*	3	B1	2,75	3	B1	3	3		

No hay corpus de referencia

Tabla 7.14: Tabla conjunta de los resultados del DELE superior.

7.3.4. Cortázar: la prueba de Evaluator

A favor de la eficacia de Evaluator, podemos enumerar algunas alarmas de diagnóstico de nivel que se produzcan cuando el nivel de un texto sea muy alto o muy bajo en algún nivel gramatical. Es decir, al ser un proceso automatizado, se pueden detectar anomalías en un texto cuando en éste:

- Se identifican muchos lemas igual que los vocablos.
- Se detectan muchos lemas de niveles muy altos o muy bajos.
- Se nivela menos de un 80 % de los lemas porque no están "marcados" los niveles de sus lemas o sus lemas no se hallan en los glosarios.

Un buen ejemplo de texto-alarma es un fragmento del capítulo 68 en gíglico de *Rayuela* de Julio Cortázar (Cortázar, 1986). Indiscutiblemente, es un fragmento de calidad. Sin embargo, después de procesarlo con FreeLing y con Lexicator, obtenemos gran variedad de lemas no listados en los glosarios. En la figura 7.17 observamos muy pocos lemas del texto en gíglico en los glosarios (73,33 %). Este hecho no lo atribuimos a errores tipográficos o léxicos, como sería el caso para un aprendiz, sino a la creatividad literaria de Cortázar. Al aplicar el método de nivelación del léxico mediante la combinación de diccionarios, el glosario del Cervantes califica el 37,78 %, la identificación léxica del Cervantes junto con

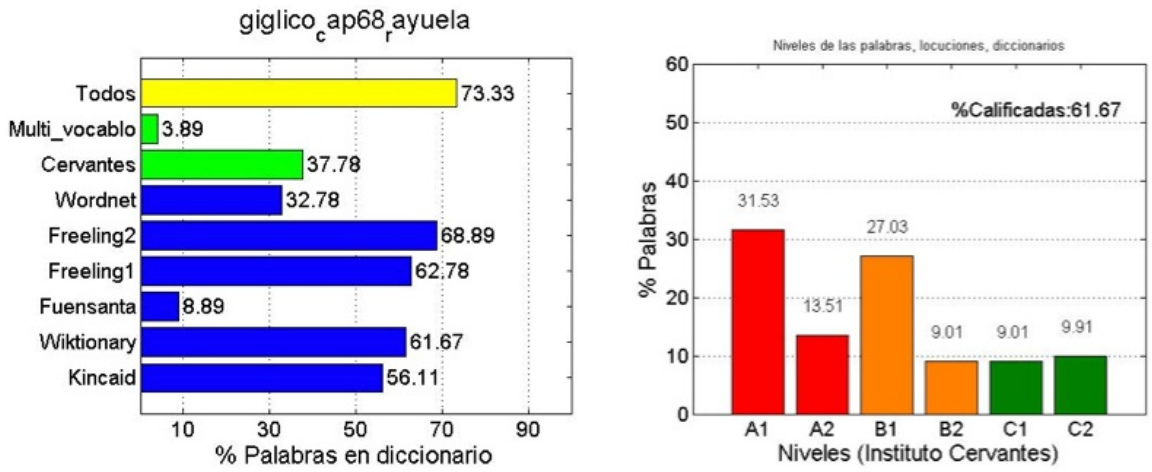


Figura 7.17: Porcentajes de léxico procesado y nivelado con los diferentes diccionarios.

el glosario de multivocablos es del 41,67 % y la combinación de todos los glosarios con el criterio de *PCIC* califica el 61,67 %. Todo esto apunta a que el léxico es muy específico.

Más aún, los datos anteriores no cuadran con el nivel léxico del texto que nos proporciona el método de las áreas porque, tras el cálculo, apunta a un nivel B2 con tendencia a B1 (-0,37) y con una fiabilidad aceptable de 0,35. También sorprenden los resultados del método Kincaid o método K-2000 cuando nos revela un porcentaje de sólo un 56.11 %. Estos datos nos alertan de que estamos ante un texto muy especial ya que quedan un 26.67 % de lemas sin identificar.

Respecto a la calificación sintáctica, los resultados son más asombrosos. La calificación automática, después de hallar todos los niveles de calificación sintáctica, da al mismo tiempo el nivel A2 y C1 en el cuadrante derecho superior de la figura 7.18 (la de la izquierda). Sin embargo, nuestro criterio sintáctico para la elección de nivel en los resultados automáticos optaría por el nivel A2 que se indica en la misma figura 7.18 (la de la derecha). Estos resultados son muy llamativos.

En resumen, cuando un texto sea diferente de lo esperado, cada módulo de Evaluator nos va a proporcionar datos contradictorios entre sí. Aunque en este caso del capítulo de *Rayuela*, dadas las características del famoso texto, lo esperado es que todas las alarmas señaladas arriba se cumplan y se requiera la revisión de un humano para decidir la excepcionalidad de un caso de particularidades similares.

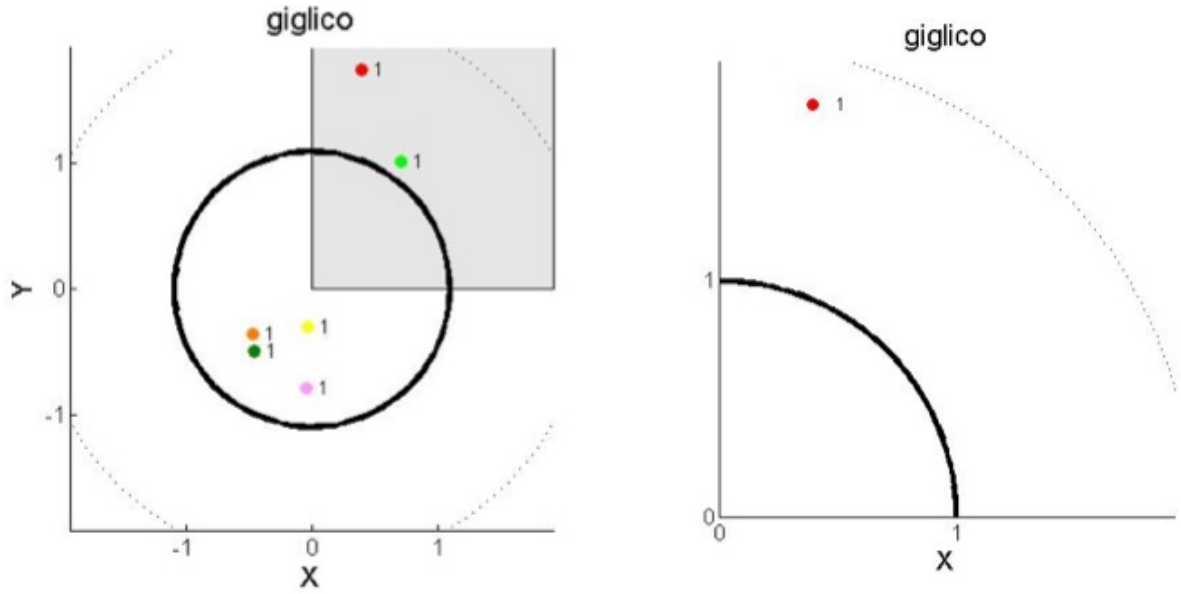


Figura 7.18: Nivelación sintáctica del texto gíglico de Cortazar.

Capítulo 8

Conclusiones

Esta memoria de tesis doctoral ha conseguido validar la hipótesis planteada y ofrece un conjunto de métodos integrables en una única herramienta de evaluación para la calificación automática de cualquier texto. Concretamente, esta herramienta se ha diseñado para diagnosticar textos escritos por aprendices de español como lengua extranjera. Todas estas contribuciones son originales y han sido adecuadamente justificadas y probadas en textos de distintos niveles de conocimiento y de diferentes autores, reconocidos y anónimos, españoles y extranjeros.

Las principales contribuciones realizadas son:

1. Se han desarrollado algoritmos y herramientas en un entorno automatizado de análisis que califican y marcan el valor de un texto a nivel léxico y sintáctico, e identifican la validez del contenido semántico. Se han tenido en cuenta las directrices especificadas en el *Plan Curricular del Instituto Cervantes* en relación con las áreas léxicas, sintácticas y semánticas estudiadas.

    - En el área léxica:
        - Se ha ampliado el número de locuciones del fichero de FreeLing. A las locuciones ya listadas y a las nuevas se les ha marcado con categorías gramaticales y con contenido funcional en los *PoS*. Las locuciones se han calificado con niveles estandarizados atendiendo a criterios de complejidad sintáctica y a diversos parámetros cuantitiativos, teniendo como referencia la clasificación y calificación propuesta por el *Plan Curricular del Instituto Cervantes*.
        - Con el fin de ampliar la identificación y nivelación de vocablos y lemas de un texto, se han propuesto criterios de calificación en función de su pertenencia a un conjunto de diccionarios combinados. Estos criterios están basados en las características de los diccionarios, en la fiabilidad otorgada a cada uno de los mismos y en su tamaño.
        - Se han propuesto parámetros o indicadores de nivel léxico de un texto en función del comportamiento de los vocablos calificados por niveles y ordenados por frecuencias de aparición. Este procedimiento no requiere la

utilización de *corpora* de referencia. Junto con el nivel otorgado se obtienen parámetros de tendencia y de fiabilidad de la calificación.

- En el área sintáctica:
  - Se han construido estructuras sintácticas y se han calificado éstas por niveles de acuerdo a criterios de complejidad sintáctica o a la asignación del *Plan Curricular del Instituto Cervantes*.
  - Se ha establecido un criterio de calificación sintáctica de una categoría o *PoS* en función de su pertenencia a estructuras sintácticas de mayor o menor nivel dentro de un esquema de estructuras anidadas. Al combinar la distribución de elementos por niveles y la distribución de estructuras identificadas y niveladas, es posible generar criterios objetivos para la calificación sintáctica en comparación con un *corpus* de referencia previamente analizado con el fichero de estructuras niveladas por el mismo método.
- En el área semántica:
  - Se identifica semánticamente un texto o un conjunto de textos al ubicarlos dentro de un área temática, basándonos en las áreas de conocimiento definidas por la Dra. Fuensanta López en su tesis.
  - Por otro lado, se puede establecer la relación genealógica entre vocablos al medirse y analizarse cada vocablo en relación a los demás vocablos del texto mediante la utilización del diccionario esWordNet.
  - Los vocablos se han agrupado en función de los campos semánticos definidos por el *Plan Curricular del Instituto Cervantes* y codificados en su glosario-inventario. Para ello, se han utilizando criterios de repetición y de ponderación en función de los campos semánticos del entorno. Los campos semánticos de cada texto se correlan con los de un *corpus* específico para identificar la temática del texto.
  - Finalmente, se han empleado herramientas de estadística multivariante interpretando los resultados en el ámbito de la lingüística. Las herramientas del análisis semántico latente, adaptadas a textos de pocas palabras, han permitido la identificación temática de textos respecto a un *corpus* de referencia.

2. Se han comparado los resultados de la calificación automática de un conjunto de 80 textos de aprendices de español como segunda lengua con los resultados de la calificación asignada por dos correctores especializados del Instituto Cervantes. Los resultados obtenidos en el nivel léxico y sintáctico son satisfactorios. Por otro lado, desde un punto de vista semántico, se han podido identificar y correlar la temática de textos escritos por aprendices de español con textos similares escritos por hablantes nativos españoles.

3. El sistema desarrollado se halla en un nivel de computación. Los algoritmos y funciones que conforman los módulos analizadores son practicables y las bases de datos reutilizables, actualizables e implementables. El sistema es capaz de diferenciar textos de nivel A2, B1, B2 y C1 a nivel léxico y sintáctico.

El sistema de nivelación, creado y bautizado como Evaluator, se concibe como un recurso de medida de la competencia escrita para aprendices de español. Esta aplicación alcanzará su plenitud cuando se mantenga, se siga implementando y se haga visible en internet con una interfaz sencilla. La retroalimentación que pudiera proporcionar a los usuarios que quisieran nivelar sus textos y conocer sus errores será el fruto del trabajo conjunto de un grupo interdisciplinar de expertos.

Apéndice A

Datos y tablas

A.1. Exámenes de DELE de nivel intermedio: Tablas de valores numéricos de nivelación

A.1.1. Nivel léxico

A.1.1.1. Texto-1

La tabla A.1 muestra la distribución de lemas en todos los diccionarios del análisis léxico del Texto-1 del nivel intermedio.

En la tabla A.2 se observa la distribución de lemas por niveles de referencia del análisis léxico del Texto-1 del nivel intermedio.

La tabla A.3 resume los niveles de referencia y fiabilidad del análisis léxico del Texto-1 del nivel intermedio.

A.1.1.2. Texto-2

La tabla A.4 muestra la distribución de lemas en los diccionarios del análisis léxico del Texto-2 del nivel intermedio.

En la tabla A.5 se observa la distribución de lemas por niveles de referencia del análisis léxico del Texto-2 del nivel intermedio.

La tabla A.6 resume los niveles de referencia y fiabilidad del análisis léxico del Texto-2 del nivel intermedio.

Candidato #	Kincaid	Wiktionary	Fuensanta	Frlng 1	Frlng 2	esWordnet	Cervantes	Multivocablos	Total	# Glosados
1	76,87	92,52	38,78	85,03	93,20	64,63	63,27	4,08	98,64	147
2	70,00	84,55	30,00	77,27	87,27	67,27	57,27	3,64	92,73	110
3	81,82	91,48	37,50	84,66	94,32	66,48	66,48	2,84	98,30	176
4	67,47	84,34	32,13	78,71	87,15	68,67	53,82	4,02	93,98	249
5	69,17	81,20	27,07	73,68	82,71	61,65	51,88	9,77	96,24	133
6	70,59	82,35	33,53	79,41	87,65	61,18	52,35	5,88	95,29	170
7	69,38	87,50	32,50	76,88	89,38	62,50	53,75	5,63	96,25	160
8	70,44	81,13	30,19	76,10	84,91	62,26	54,09	7,55	92,45	159
9	76,30	88,89	37,78	80,74	90,37	62,22	56,30	3,70	96,30	135
10	73,83	88,59	33,56	75,84	88,59	71,14	52,35	3,36	95,30	149
11	74,16	88,20	34,83	82,02	88,20	66,29	53,93	5,62	96,63	178
12	75,94	86,10	29,41	82,35	87,70	62,57	58,82	5,88	97,33	187
13	79,88	92,68	42,68	85,98	91,46	68,29	64,02	2,44	96,95	164
14	66,28	88,37	37,21	74,42	86,63	66,86	55,23	6,40	96,51	172
15	68,12	81,16	28,99	71,74	86,23	60,87	47,83	5,07	94,20	138
16	75,34	86,99	39,73	81,51	91,78	66,44	60,27	6,16	98,63	146
17	71,23	86,32	29,72	81,13	89,62	67,45	58,02	2,83	95,75	212
18	76,80	88,66	39,69	80,93	88,14	64,43	55,15	7,73	97,94	194
19	78,42	86,84	32,63	84,21	91,58	62,63	61,58	3,16	96,32	190
20	73,28	85,50	34,35	79,39	88,55	70,23	55,73	4,58	93,89	131

Tabla A.1: Distribución de lemas en todos los diccionarios. Nivel intermedio: Texto-1.

Candidato #	A1	A2	B1	B2	C1	C2	# Calificados
1	39,66	14,66	28,45	12,07	4,31	0,86	116
2	32,56	13,95	31,40	15,12	2,33	4,65	86
3	36,05	21,77	26,53	8,16	2,72	4,76	147
4	28,50	19,50	29,50	14,00	3,50	5,00	200
5	29,63	13,89	30,56	14,81	3,70	7,41	108
6	26,09	21,01	26,09	14,49	7,97	4,35	138
7	32,52	21,95	28,46	12,20	1,63	3,25	123
8	31,54	16,15	28,46	15,38	1,54	6,92	130
9	30,91	21,82	34,55	9,09	0,91	2,73	110
10	25,00	25,86	35,34	8,62	0,86	4,31	116
11	27,86	22,86	36,43	7,86	2,14	2,86	140
12	36,31	16,56	26,75	10,83	3,18	6,37	157
13	38,17	20,61	28,24	8,40	1,53	3,05	131
14	31,01	20,16	26,36	14,73	3,88	3,88	129
15	28,04	14,02	26,17	18,69	0,93	12,15	107
16	29,27	20,33	30,08	12,20	4,07	4,07	123
17	30,17	12,29	34,08	13,97	2,79	6,70	179
18	32,03	19,61	26,80	16,99	1,96	2,61	153
19	35,26	23,08	21,79	11,54	3,21	5,13	156
20	34,91	18,87	30,19	11,32	1,89	2,83	106

Tabla A.2: Distribución de lemas por niveles de referencia del análisis léxico. Nivel Intermedio. Texto-1.

# Candidato	Nivel	Tendencia	Nivel+Tendencia	Fiabilidad
1	3	-0,25	2,75	-0,30
2	3	0,69	3,69	-0,18
3	3	0,14	3,14	0,14
4	3	0,61	3,61	0,12
5	3	-0,03	2,97	0,08
6	3	0,75	3,75	0,01
7	3	0,50	3,50	-0,01
8	3	0,35	3,35	0,08
9	4	-0,83	3,17	-0,01
10	3	0,12	3,12	0,16
11	4	-0,17	3,83	0,10
12	2	0,93	2,93	0,01
13	3	0,47	3,47	-0,05
14	3	-0,27	2,73	0,00
15	3	-0,22	2,78	-0,02
16	2	0,25	2,25	-0,10
17	3	0,57	3,57	0,06
18	3	0,68	3,68	0,06
19	3	0,43	3,43	-0,01
20	4	-0,02	3,98	-0,20

Tabla A.3: Tabla de niveles de referencia y fiabilidad del análisis léxico. Nivel intermedio. Texto-1.

Candidato #	Kincaid	Wiktionary	Fuensanta	Frlng 1	Frlng 2	esWordnet	Cervantes	Multivocablos	Total	# Glosados
1	79,89	91,62	37,99	91,06	97,21	68,16	69,27	2,79	100,00	179
2	72,31	90,77	29,23	75,38	86,15	69,23	56,92	6,15	98,46	130
3	71,17	90,18	37,42	85,89	95,09	66,87	66,26	3,68	98,77	163
4	78,57	90,08	38,49	82,94	89,68	68,25	63,89	3,97	96,43	252
5	70,75	82,31	29,25	76,19	85,03	62,59	53,06	7,48	97,28	147
6	74,07	85,71	28,04	83,60	88,36	63,49	61,38	5,29	97,35	189
7	81,71	95,12	36,59	92,07	97,56	68,90	73,17	1,83	99,39	164
8	79,37	88,36	38,62	85,19	91,53	65,08	65,08	5,82	97,88	189
9	83,56	94,52	40,41	89,04	96,23	64,73	71,23	3,42	99,66	292
10	82,18	97,03	40,59	92,57	98,02	72,77	76,24	1,49	99,50	202
11	78,95	85,53	30,92	82,24	87,50	63,16	53,95	7,89	97,37	152
12	74,89	88,99	33,92	85,90	92,07	61,67	61,67	3,52	96,48	227
13	83,12	93,51	37,01	92,21	95,45	72,08	62,34	2,60	98,05	154
14	77,17	86,30	30,59	80,37	90,41	58,45	52,05	9,59	100,00	219
15	76,65	86,23	32,34	80,24	88,02	62,87	65,87	6,59	96,41	167
16	73,18	87,15	31,84	84,92	90,50	64,80	71,51	6,70	99,44	179
17	81,18	89,41	32,94	84,12	95,29	63,53	60,00	4,71	100,00	170
18	68,98	78,70	26,39	74,07	82,87	59,26	57,87	2,31	87,04	216
19	78,61	87,70	35,29	82,35	90,91	62,03	59,89	4,28	96,26	187
20	74,42	88,37	34,88	79,84	89,15	68,99	58,91	9,30	99,22	129

Tabla A.4: Distribución de lemas en todos los diccionarios. Nivel intermedio: Texto-2.

Candidato #	A1	A2	B1	B2	C1	C2	# Calificados
1	36,02	18,01	26,71	13,66	3,11	2,48	161
2	40,20	17,65	23,53	13,73	0,98	3,92	102
3	35,00	16,43	24,29	13,57	5,71	5,00	140
4	36,49	16,59	28,91	13,74	2,37	1,90	211
5	36,29	9,68	28,23	13,71	4,03	8,06	124
6	34,15	15,85	27,44	14,02	3,05	5,49	164
7	39,86	20,27	22,30	14,19	2,70	0,68	148
8	33,74	19,02	22,09	17,79	3,68	3,68	163
9	32,95	25,58	25,58	12,79	1,94	1,16	258
10	42,20	29,48	16,76	9,25	1,73	0,58	173
11	23,81	23,02	32,54	14,29	3,17	3,17	126
12	34,04	21,81	27,13	10,11	4,26	2,66	188
13	32,14	17,86	32,86	13,57	1,43	2,14	140
14	29,21	18,54	31,46	7,30	7,87	5,62	178
15	43,36	18,88	19,58	12,59	3,50	2,10	143
16	33,13	16,27	27,11	13,25	4,82	5,42	166
17	30,28	21,13	26,06	9,15	7,04	6,34	142
18	40,24	17,75	21,30	12,43	0,59	7,69	169
19	34,44	23,84	23,84	11,92	1,99	3,97	151
20	28,57	23,21	25,89	15,18	2,68	4,46	112

Tabla A.5: Distribución de lemas por niveles de referencia del análisis léxico. Nivel Intermedio. Texto-2.

Candidato #	Nivel	Tendencia	Nivel+Tendencia	Fiabilidad
1	3	0,57	3,57	-0,12
2	3	0,16	3,16	0,04
3	3	0,51	3,51	0,14
4	3	0,24	3,24	-0,09
5	3	0,69	3,69	0,05
6	4	-0,44	3,56	-0,05
7	4	-0,51	3,49	-0,33
8	4	0,80	4,80	0,04
9	4	-0,24	3,76	-0,02
10	3	-0,09	2,91	-0,23
11	4	-0,87	3,13	0,03
12	4	-0,58	3,42	-0,13
13	4	-0,54	3,46	-0,16
14	4	-0,86	3,14	-0,04
15	4	-0,02	3,98	0,21
16	3	0,43	3,43	0,16
17	4	-0,05	3,95	0,05
18	4	-0,37	3,63	0,06
19	3	0,27	3,27	0,05
20	4	-0,38	3,62	-0,04

Tabla A.6: Tabla de niveles de referencia y fiabilidad del análisis léxico. Nivel intermedio. Texto-2.

Candidato #	Nivel	Fiabilidad	Mod A1	Ang A1	Mod A2	Ang A2	Mod B1	Ang B1	Mod B2	Ang B2	Mod C1	Ang C1	Mod C2	Ang C2
1	4	1,20	3,12	-95,84	1,87	133,44	0,93	59,67	1,20	68,18	0,47	67,54	0,68	-132,84
2	3	1,97	2,80	-91,42	2,09	95,92	1,97	60,19	1,38	-122,86	0,65	-75,11	0,68	-132,84
3	3	1,35	1,34	-122,17	1,63	-154,92	1,35	64,93	1,02	-1,85	0,69	-151,09	0,68	-132,84
4	2	2,76	2,59	-89,13	2,76	71,78	1,09	42,15	1,00	-178,21	1,35	-131,27	1,30	18,82
5	3	1,66	2,02	-96,08	0,46	-167,88	1,66	85,75	0,31	-115,44	0,94	-18,49	0,68	-132,84
6	0	0,57	1,14	-94,60	0,57	85,68	0,57	-4,37	1,24	120,27	0,38	10,80	0,68	-132,84
7	2	4,19	2,68	-86,62	4,19	81,25	2,20	27,95	2,73	-139,16	0,84	-100,06	0,68	-132,84
8	3	3,57	2,51	-98,97	3,49	82,47	3,57	21,64	3,10	-145,38	1,67	-128,04	0,68	-132,84
9	3	2,97	1,76	-102,45	0,81	144,77	2,97	40,40	1,38	-118,16	2,06	-140,28	0,68	-132,84
10	3	2,55	3,66	-87,82	1,20	103,49	2,55	63,51	0,52	-151,14	1,73	-130,38	1,16	52,77
11	3	1,03	4,15	-100,55	2,72	101,96	1,03	40,92	1,86	91,44	0,52	-79,44	0,68	-132,84
12	3	2,76	0,60	-135,00	0,34	-178,89	2,76	33,68	2,52	-126,77	0,66	-115,49	0,68	-132,84
13	2	2,18	3,11	-95,12	2,18	79,26	0,93	132,75	1,20	59,38	0,52	-114,18	0,68	-132,84
14	2	4,78	3,45	-88,43	4,78	77,80	1,87	46,23	2,78	-130,14	0,31	-75,51	0,68	-132,84
15	3	2,85	2,33	-89,36	2,58	93,10	2,85	31,10	2,40	-135,62	1,05	-131,98	0,68	-132,84
16	3	2,58	1,94	-87,28	1,36	110,44	2,58	46,28	2,04	-121,32	1,55	-121,77	1,08	51,73
17	3	2,49	1,74	-123,93	1,04	-148,63	2,49	56,78	1,36	-92,26	0,59	-133,80	0,68	-132,84
18	3	1,14	2,47	-103,56	0,61	99,37	1,14	59,28	0,60	100,08	0,38	110,69	0,68	-132,84
19	3	3,08	1,31	-116,74	1,27	140,68	3,08	17,65	2,21	-152,25	0,44	34,06	0,68	-132,84
20	3	1,56	1,63	-103,74	1,65	114,06	1,56	27,29	0,33	-130,65	1,15	-137,09	0,68	-132,84

Tabla A.7: Valores-vectores de nivel sintáctico. Nivel intermedio. Texto-1.

A.1.2. Nivel sintáctico

A.1.2.1. Texto-1

La tabla A.7 muestra la nivelación sintáctica del Texto-1 de cada candidato con su correspondiente "Nivel" (1, 2, 3, 4, 5, 6) equivalente a un nivel de referencia (A1, A2, B1, B2, C1, C2) respectivamente; a continuación, sigue el índice de "Fiabilidad" del nivel indicado. Los demás valores expresados son la orientación (Ang) y el tamaño (Mod) del vector indicador de nivel. Ambos valores, "Ángulo" y "Módulo", marcan el nivel en un cuadrante equivalente al nivel del texto, como se mostraba en la figura 7.7.

La tabla A.8 representa el histograma de estructuras sintácticas niveladas del Texto-1 del DELE intermedio.

En la tabla A.9 aparece el número de etiquetas de las categorías gramaticales niveladas o *Part of Speech* (*PoS*), según el criterio de máximo nivel. Esta figura registra también una serie de elementos no computados (Items NC) que incluyen mayormente signos de puntuación y algún *PoS* solo, que queda sin calificar al estar excluido o no pertenecer a ninguna de las estructuras registradas en el fichero de estructuras.

A.1.2.2. Texto-2

La tabla A.10 muestra la nivelación sintáctica del Texto-1 de cada candidato con su "Nivel" (1, 2, 3, 4, 5, 6) equivalente a un nivel de referencia (A1, A2, B1, B2, C1, C2) respectivamente; a continuación, sigue el índice de fiabilidad del Nivel indicado. Los demás valores expresados son la orientación (Ang) y el tamaño (Mod) del vector indicador de nivel. Ambos valores, "Ángulo" y "Módulo", marcan el nivel en un cuadrante equivalente al nivel del texto, como se mostraba en la figura 7.8.

Candidato #	A1	A2	B1	B2	C1	C2	Total
1	23	31	102	35	4	0	195
2	16	22	71	15	1	0	125
3	48	29	142	39	3	0	261
4	38	53	146	42	1	1	281
5	26	21	95	24	2	0	168
6	34	26	89	33	3	0	185
7	23	39	93	18	1	0	174
8	20	30	79	15	0	0	144
9	25	21	88	18	0	0	152
10	18	28	108	26	0	1	181
11	18	40	109	42	2	0	211
12	49	30	132	23	2	0	236
13	24	36	105	36	2	0	203
14	18	40	93	16	2	0	169
15	23	30	89	17	1	0	160
16	30	30	110	20	0	1	191
17	50	33	169	33	3	0	288
18	29	29	109	34	4	0	205
19	40	32	116	27	4	0	219
20	27	26	83	23	1	0	160

Tabla A.8: Representa el histograma de estructuras sintácticas niveladas del Texto-1 del DELE intermedio.

Candidato #	Items NC	A1	A2	B1	B2	C1	C2	Total
1	34	6	8	59	59	13	0	179
2	33	5	11	45	31	9	0	134
3	47	5	8	73	80	10	0	223
4	77	13	40	106	70	11	5	322
5	19	6	13	51	48	16	0	153
6	45	8	20	68	52	16	0	209
7	31	9	24	82	31	12	0	189
8	47	6	22	96	24	6	0	201
9	47	5	11	69	40	2	0	174
10	57	8	15	65	47	5	2	199
11	45	5	17	80	68	16	0	231
12	28	7	19	101	45	13	0	213
13	42	7	23	53	69	12	0	206
14	44	9	28	77	36	14	0	208
15	29	7	15	74	30	7	0	162
16	39	8	14	74	41	7	2	185
17	41	4	15	95	73	13	0	241
18	50	6	21	76	66	14	0	233
19	37	6	13	110	38	18	0	222
20	42	5	11	61	45	6	0	170

Tabla A.9: Número de *PoS* o categorías gramaticales computadas con el criterio de máximo nivel y distribuidos por niveles de referencia. Nivel intermedio. Texto-1.

Candidato #	Nivel	Fiabilidad	Mod A1	Ang A1	Mod A2	Ang A2	Mod B1	Ang B1	Mod B2	Ang B2	Mod C1	Ang C1	Mod C2	Ang C2
1	3	2,00	1,09	-165,74	0,22	13,79	2,00	22,98	1,26	-138,87	0,97	-139,58	0,68	-132,84
2	3	2,53	0,93	-116,47	0,53	101,46	2,53	36,19	2,28	-127,89	0,86	-111,82	0,68	-132,84
3	3	1,74	0,71	-62,45	1,28	-170,38	1,74	42,43	0,92	-99,94	0,84	-137,30	0,68	-132,84
4	3	1,70	2,18	-82,80	3,48	90,78	1,70	10,40	1,73	-154,54	0,26	-74,32	0,68	-132,84
5	3	1,94	2,26	-103,48	0,91	101,88	1,94	48,55	1,45	-131,13	1,01	69,92	0,68	-132,84
6	4	1,18	0,81	-125,92	2,04	114,13	0,43	-95,87	1,18	11,36	0,82	-152,07	0,68	-132,84
7	0	1,18	0,34	-133,90	1,79	102,28	0,66	-2,50	0,88	-93,68	0,62	-119,75	0,68	-132,84
8	3	1,61	2,13	-44,94	1,27	123,76	1,61	15,75	1,82	-170,19	1,34	63,83	0,68	-132,84
9	3	2,19	1,98	-111,87	1,15	174,68	2,19	37,53	0,17	-141,55	1,15	-157,08	0,42	21,70
10	3	2,91	2,32	-110,52	2,48	101,22	2,91	21,00	1,66	-148,01	1,21	-128,93	0,68	-132,84
11	3	2,99	2,11	-89,86	1,22	175,07	2,99	39,88	1,53	-149,08	0,40	-81,88	0,68	-132,84
12	3	1,41	1,98	-111,81	1,15	103,84	1,41	45,20	0,05	-21,32	1,85	-134,65	1,14	20,57
13	3	3,62	1,73	-142,33	0,22	-116,75	3,62	33,24	2,12	-132,21	1,86	-134,97	0,68	-132,84
14	5	2,46	2,93	-110,34	2,00	114,26	0,48	79,51	0,53	109,37	2,46	30,60	0,68	-132,84
15	3	1,78	2,76	-115,92	1,49	160,92	1,78	76,68	1,27	-1,12	1,13	-116,65	0,68	-132,84
16	0	0,66	0,89	143,66	0,48	-170,75	0,66	34,72	0,97	-42,59	1,75	-131,16	0,68	-132,84
17	4	2,46	3,79	-108,71	3,45	98,11	0,49	148,73	2,46	44,03	1,72	-130,08	0,68	-132,84
18	3	1,60	1,81	-12,78	0,61	-91,40	1,60	64,35	1,62	-118,55	1,37	-148,56	0,53	51,68
19	3	1,79	2,56	-87,42	2,26	102,89	1,79	51,77	0,91	-109,10	1,22	-128,82	0,68	-132,84
20	2	2,45	3,20	-83,59	2,45	86,67	2,29	33,65	2,28	-168,57	0,69	-0,04	0,68	-132,84

Tabla A.10: Valores-vectores de nivel sintáctico. Nivel intermedio. Texto-2.

La tabla A.11 representa el histograma de estructuras sintácticas niveladas del Texto-2 del DELE intermedio.

La tabla A.12 aparece el número de etiquetas de las categorías gramaticales niveladas o *Part of Speech* (*PoS*), según el criterio de máximo nivel. Esta tabla registra también una serie de elementos no computados (Items NC) que incluyen mayormente signos de puntuación y algún *PoS* que queda, solo, sin calificar, al estar excluido o no pertenecer a ninguna de las estructuras registradas en el fichero de estructuras.

A.1.3. Nivel semántico

A.1.3.1. Texto-1

En la tabla A.13 se listan todas las distancias de los textos en relación a la media de cada *cluster*. El valor menor es indicador de mayor proximidad semántica de un examen a un *cluster*.

A.1.3.2. Texto-2

En la tabla A.14 se listan todas las distancias de los textos en relación a la media de cada *cluster*. El valor menor es indicador de mayor proximidad semántica de un examen a un *cluster*.

Candidato #	A1	A2	B1	B2	C1	C2	Total
1	53	32	130	32	2	0	249
2	32	23	92	17	1	0	165
3	46	28	124	29	2	0	229
4	48	67	160	42	4	0	321
5	28	28	105	23	5	0	189
6	55	47	127	43	3	0	275
7	47	37	106	28	2	0	220
8	35	31	103	29	6	0	204
9	63	51	216	58	4	1	393
10	34	42	122	29	1	0	228
11	28	24	107	24	2	0	185
12	46	44	153	43	0	1	287
13	41	27	128	24	0	0	220
14	39	51	152	49	9	0	300
15	38	38	156	41	1	0	274
16	63	33	132	35	0	0	263
17	25	51	122	48	0	0	246
18	59	32	157	32	2	1	283
19	32	42	130	30	1	0	235
20	15	24	71	18	2	0	130

Tabla A.11: Número de estructuras sintácticas por niveles de referencia. Texto-2.

Candidato #	Items NC	A1	A2	B1	B2	C1	C2	Total
1	33	3	23	90	51	9	0	209
2	26	5	15	69	33	9	0	157
3	31	10	9	75	57	9	0	191
4	57	15	29	123	59	21	0	304
5	20	5	16	68	42	14	0	165
6	42	7	15	71	90	10	0	235
7	20	7	16	67	58	11	0	179
8	65	18	15	87	39	19	0	243
9	52	8	18	146	104	11	3	342
10	50	5	19	115	50	10	0	249
11	28	8	9	84	40	13	0	182
12	41	6	23	97	81	6	4	258
13	36	1	17	89	37	4	0	184
14	33	4	17	82	77	37	0	250
15	28	2	8	66	78	10	0	192
16	37	5	17	73	76	6	0	214
17	26	2	16	57	87	6	0	194
18	42	23	25	89	64	7	2	252
19	36	10	17	82	62	9	0	216
20	44	8	16	65	23	14	0	170

Tabla A.12: Número de PoS o categorías gramaticales computadas con el criterio de máximo nivel y distribuidos por niveles de referencia. Nivel intermedio. Texto-2.

TEXTO-1	MEDIA DE LAS DISTANCIAS			CONTENIDO		RESULTADO
Candidato #	Lugar especial	Penpal	Reclamación	Automático	Real	OK /
1	9,44	2,31	8,59	Penpal	Penpal	OK
2	7,64	4,73	0,70	Reclamación	Reclamación	OK
3	4,12	1,97	5,90	Penpal	Invitación viaje	/
4	3,04	3,48	8,31	Lugar especial	Invitación viaje	/
5	7,54	4,37	1,07	Reclamación	Reclamación	OK
6	5,91	0,77	7,22	Penpal	Penpal	OK
7	8,45	1,54	8,10	Penpal	Penpal	OK
8	5,11	3,60	2,49	Reclamación	Reclamación	OK
9	5,12	1,16	6,09	Penpal	Invitación viaje	/
10	6,35	1,47	8,48	Penpal	Invitación viaje	/
11	8,43	2,32	9,92	Penpal	Penpal	OK
12	6,38	5,60	1,93	Reclamación	Reclamación	OK
13	9,08	2,02	7,89	Penpal	Penpal	OK
14	9,36	2,44	9,54	Penpal	Penpal	OK
15	7,56	4,18	1,38	Reclamación	Reclamación	OK
16	7,90	1,07	7,29	Penpal	Penpal	OK
17	9,08	6,52	2,45	Reclamación	Reclamación	OK
18	8,89	2,06	9,12	Penpal	Penpal	OK
19	7,47	3,92	1,69	Reclamación	Reclamación	OK
20	6,33	0,73	5,72	Penpal	Invitación viaje	/

Tabla A.13: Cálculo de las distancias de los exámenes a las medias de los *clusters* para el Texto-1.

TEXTO-2	MEDIA DE LAS DISTANCIAS			CONTENIDO		RESULTADO
Candidato #	Lugar especial	Penpal	Reclamación	Automático	Real	OK /
1	1,44	6,22	9,22	Lugar especial	Lugar especial	OK
2	1,73	3,90	6,42	Lugar especial	Lugar especial	OK
3	1,90	4,08	5,84	Lugar especial	Profesión favorita	OK
4	3,99	2,13	5,63	Penpal	Profesión favorita	/
5	1,03	5,34	8,72	Lugar especial	Lugar especial	OK
6	2,06	6,84	9,53	Lugar especial	Lugar especial	OK
7	2,35	6,78	10,22	Lugar especial	Lugar especial	OK
8	2,50	4,95	9,58	Lugar especial	Lugar especial	OK
9	2,26	6,91	9,95	Lugar especial	Amistades	/
10	2,77	3,03	6,74	Lugar especial	Profesión favorita	/
11	1,96	6,68	9,64	Lugar especial	Lugar especial	OK
12	0,56	5,61	7,84	Lugar especial	Lugar especial	OK
13	1,34	6,23	8,27	Lugar especial	Lugar especial	OK
14	0,42	5,50	7,77	Lugar especial	1ª impresión	/
15	1,97	5,63	9,67	Lugar especial	Lugar especial	OK
16	1,08	4,39	7,42	Lugar especial	Lugar especial	OK
17	0,68	4,71	7,16	Lugar especial	1ª impresión	/
18	1,63	3,98	7,29	Lugar especial	Lugar especial	OK
19	2,67	3,91	4,96	Lugar especial	1ª impresión	/
20	3,64	2,54	5,28	Penpal	Profesión favorita	/

Tabla A.14: Cálculo de las distancias de los exámenes a las medias de los *clusters* para el Texto-2.

Candidato #	Kincaid	Wiktionary	Fuensanta	Frlng 1	Frlng 2	esWordnet	Cervantes	Multivocablos	Total	# Glosados
1	77,6	91,2	38,4	83,2	92,8	66,4	53,6	6,4	99,2	125
2	69,94	82,66	31,21	73,99	84,97	62,43	52,60	12,14	98,84	173
3	75,96	91,26	28,96	81,97	91,26	64,48	55,19	7,10	98,91	183
4	80,85	87,23	29,79	81,91	88,30	58,51	60,64	7,45	96,81	94
5	73,55	83,47	34,71	74,38	85,12	61,98	51,24	7,44	93,39	121
6	73,03	87,50	30,92	79,61	90,79	62,50	57,24	5,26	96,05	152
7	78,46	87,18	36,92	82,05	88,72	62,05	56,41	6,15	97,44	195
8	65,96	79,43	29,79	73,05	82,98	56,03	53,90	7,09	92,20	141
9	72,48	81,65	35,78	77,98	86,24	65,14	53,21	7,34	96,33	109
10	79,04	82,04	32,34	83,23	86,83	59,88	53,89	6,59	94,01	167
11	69,93	87,41	37,76	80,42	88,81	55,94	56,64	7,69	97,20	143
12	79,46	88,39	34,82	84,82	92,86	66,07	60,71	4,46	98,21	112
13	68,55	78,62	29,56	74,21	83,65	59,12	57,86	10,06	96,23	159
14	78,23	87,90	36,29	82,66	90,73	60,48	57,26	6,05	97,18	248
15	75,14	85,55	32,37	78,61	89,02	61,85	53,76	4,62	94,80	173
16	73,97	85,39	34,70	83,56	92,24	60,27	63,01	3,20	96,80	219
17	72,13	87,43	33,88	76,50	89,62	62,84	54,10	6,56	97,27	183
18	78,45	85,08	35,91	79,56	87,29	66,30	54,70	7,73	96,13	181
19	71,54	83,08	33,85	80,77	87,69	55,38	61,54	6,15	96,92	130
20	77,07	86,62	31,21	80,89	89,17	62,42	56,69	5,10	94,90	157

Tabla A.15: Distribución de lemas por diccionarios. Nivel superior: Texto-1. En esta tabla se observa cómo se distribuyen los lemas del Texto-1 en los distintos diccionarios.

A.2. Exámenes de DELE de nivel superior: Tablas de valores numéricos de nivelación

A.2.1. Nivel léxico

A.2.1.1. Texto-1

En la tabla A.15 se observa cómo se distribuyen los lemas del Texto-1 en los distintos diccionarios. En la tabla A.16 se muestra la distribución de lemas por niveles de referencia del Texto-1 del nivel superior. La tabla A.17 resume los niveles de referencia y fiabilidad del Texto-1 del nivel superior.

A.2.1.2. Texto-2

En la tabla A.18 se observa cómo se distribuyen los lemas del Texto-2 en los distintos diccionarios.

En la tabla A.19 se muestra la distribución de lemas por niveles de referencia del Texto-2 del nivel superior.

La tabla A.20 resume los niveles de referencia y fiabilidad del Texto-2 del nivel superior.

Candidato #	A1	A2	B1	B2	C1	C2	# Calificados
1	29,52	13,33	39,05	14,29	1,90	1,90	105
2	25,17	14,97	23,81	23,13	6,80	6,12	147
3	25,83	15,89	33,77	20,53	3,31	0,66	151
4	41,77	10,13	27,85	11,39	5,06	3,80	79
5	26,04	18,75	37,50	12,50	2,08	3,13	96
6	38,98	16,10	22,03	13,56	8,47	0,85	118
7	29,94	12,10	33,12	19,75	3,18	1,91	157
8	23,42	19,82	27,03	14,41	8,11	7,21	111
9	25,27	17,58	29,67	15,38	5,49	6,59	91
10	34,75	11,35	31,21	12,77	7,80	2,13	141
11	32,20	16,10	27,97	14,41	7,63	1,69	118
12	30,93	15,46	27,84	15,46	6,19	4,12	97
13	31,58	12,03	21,80	18,80	9,02	6,77	133
14	39,70	12,56	28,64	13,07	4,02	2,01	199
15	34,53	12,23	31,65	10,07	5,04	6,47	139
16	32,97	12,97	27,03	16,76	4,86	5,41	185
17	29,53	11,41	33,56	17,45	4,03	4,03	149
18	25,66	16,45	30,92	16,45	5,92	4,61	152
19	28,70	18,52	25,00	17,59	5,56	4,63	108
20	32,28	15,75	33,07	13,39	0,79	4,72	127

Tabla A.16: Distribución de lemas por niveles de referencia. Nivel superior. Texto-1.

Candidato #	Nivel	Tendencia	Nivel+Tendencia	Fiabilidad
1	3	0,56	3,56	-0,08
2	4	-0,58	3,42	0,11
3	3	0,61	3,61	-0,39
4	3	-0,33	2,67	-0,11
5	2	0,95	2,95	-0,04
6	3	-0,56	2,44	-0,29
7	2	0,39	2,39	0,15
8	3	-0,14	2,86	0,01
9	4	-0,80	3,20	-0,05
10	4	-0,83	3,17	-0,08
11	3	0,57	3,57	0,01
12	4	-0,68	3,32	0,17
13	3	-0,07	2,93	-0,10
14	3	-0,17	2,83	-0,11
15	3	0,30	3,30	0,10
16	2	0,08	2,08	-0,05
17	3	0,98	3,98	0,09
18	4	0,12	4,12	0,10
19	3	0,83	3,83	0,11
20	3	-0,62	2,38	0,08

Tabla A.17: Niveles de referencia y fiabilidad. Nivel superior. Texto-1.

Candidato #	Kincaid	Wiktionary	Fuensanta	Frlng 1	Frlng 2	esWordnet	Cervantes	Multivocablos	Total	# Glosados
1	83,43	92,57	40,00	90,29	97,71	69,71	59,43	2,29	100,00	175
2	74,24	87,34	34,50	83,41	90,39	65,94	65,07	9,17	100,00	229
3	77,90	89,86	39,49	83,33	91,30	65,22	57,25	7,25	99,28	276
4	83,33	93,86	35,09	87,72	94,74	61,40	68,42	2,63	98,25	114
5	78,40	90,00	33,60	86,00	92,80	63,20	58,40	5,20	98,40	250
6	83,20	92,21	39,75	90,57	95,08	72,13	71,31	4,92	100,00	244
7	76,15	87,61	40,83	86,70	94,04	71,56	58,26	5,96	100,00	218
8	74,81	85,19	31,11	82,22	91,11	57,04	61,48	7,41	98,52	135
9	80,00	89,03	34,84	86,45	90,97	67,10	49,68	8,39	99,35	155
10	80,82	85,62	36,99	82,19	89,73	63,70	50,68	10,27	100,00	146
11	77,10	81,68	34,35	80,92	87,79	61,07	57,25	9,16	96,95	131
12	79,79	82,45	35,11	84,04	89,36	62,23	61,17	8,51	98,40	188
13	72,59	83,78	34,36	80,31	93,05	61,39	59,46	6,18	99,61	259
14	79,90	86,60	35,05	82,47	92,78	63,92	63,92	7,22	100,00	194
15	78,64	91,36	35,91	86,36	95,00	64,09	64,09	5,00	100,00	220
16	78,61	87,86	34,68	88,44	95,38	64,74	64,16	2,89	100,00	173
17	76,40	85,39	29,78	76,97	90,45	65,73	57,87	6,18	97,19	178
18	77,92	86,58	31,60	80,95	87,88	61,47	54,55	4,76	93,94	231
19	76,39	82,64	34,72	82,64	88,19	62,50	57,64	6,94	95,14	144
20	87,50	94,05	37,50	94,64	95,24	73,21	62,50	4,17	99,40	168

Tabla A.18: Distribución de lemas por diccionarios. Nivel superior. Texto-2.

Candidato #	A1	A2	B1	B2	C1	C2	# Calificados
1	27,21	24,49	29,25	12,24	4,08	2,72	147
2	29,27	10,73	25,85	21,46	7,32	5,37	205
3	25,76	18,34	30,57	15,72	5,68	3,93	229
4	42,71	11,46	28,125	12,5	2,08	3,13	96
5	27,80	20,49	28,78	12,68	7,80	2,44	205
6	38,60	21,05	24,12	13,16	0,88	2,19	228
7	23,78	17,30	30,27	15,68	7,57	5,41	185
8	32,41	22,22	25,93	9,26	5,56	4,63	108
9	27,27	12,88	34,09	17,42	7,58	0,76	132
10	22,58	12,90	35,48	16,94	8,06	4,03	124
11	23,64	10,91	35,45	21,82	7,27	0,91	110
12	29,17	19,64	27,38	9,52	8,93	5,36	168
13	34,39	9,50	31,67	14,48	4,52	5,43	221
14	26,83	16,46	26,22	25,00	3,05	2,44	164
15	35,98	17,99	26,46	12,70	4,23	2,65	189
16	31,82	16,88	23,38	17,53	3,90	6,49	154
17	27,03	12,84	29,73	20,27	4,05	6,08	148
18	27,68	16,95	32,20	18,64	2,82	1,69	177
19	27,05	15,57	28,69	18,03	6,56	4,10	122
20	32,03	19,61	33,33	12,42	1,96	0,65	153

Tabla A.19: Distribución de lemas por niveles de referencia. Nivel superior. Texto-2.

Candidato #	Nivel	Tendencia	Nivel+Tendencia	Fiabilidad
1	1	0	1	0,00
2	3	0,79	3,79	0,01
3	3	0,67	3,67	0,04
4	2	0,97	2,97	0,18
5	4	-0,72	3,28	0,09
6	3	0,59	3,59	-0,19
7	3	0,88	3,88	0,15
8	3	0,36	3,36	-0,07
9	4	0,02	4,02	-0,27
10	4	-0,80	3,20	0,16
11	3	0,27	3,27	0,08
12	4	-0,24	3,76	0,18
13	4	0,02	4,02	0,18
14	4	-0,37	3,63	0,26
15	3	0,68	3,68	0,14
16	4	-0,44	3,56	0,02
17	3	0,19	3,19	0,13
18	3	0,90	3,90	-0,04
19	3	0,54	3,54	0,06
20	3	0,64	3,64	-0,19

Tabla A.20: Niveles de referencia y fiabilidad para el análisis léxico. Nivel superior. Texto-2.

A.2.2. Nivel sintáctico

A.2.2.1. Texto-1

La tabla A.21 muestra la nivelación sintáctica del Texto-1 de cada candidato con su Nivel (1, 2, 3, 4, 5, 6) equivalente a un nivel de referencia (A1, A2, B1, B2, C1, C2) respectivamente, a continuación sigue el índice de fiabilidad del Nivel indicado. Los demás valores expresados son la orientación (Ang) y el tamaño (Mod) del vector indicador de nivel, valores indicadores de un nivel en un cuadrante para marcar el nivel del texto como se mostraba en la figura 7.9.

La tablaA.22 representa el histograma de estructuras sintácticas niveladas del Texto-1 del DELE superior.

En la tabla A.23 se registra el número de etiquetas de las categorías gramaticales niveladas o *part of speech* (*PoS*) según el criterio de máximo nivel. Esta figura lista también una serie de elementos no computados (Items NC) que incluyen mayormente signos de puntuación y algún *PoS* solo, que queda sin calificar al estar excluido o no pertenecer a ninguna de las estructuras registradas en el fichero de estructuras.

A.2.2.2. Texto-2

La tabla A.24 muestra la nivelación sintáctica del Texto-1 de cada candidato con su Nivel (1, 2, 3, 4, 5, 6) equivalente a un nivel de referencia (A1, A2, B1, B2, C1, C2) respectivamente, a continuación sigue el índice de fiabilidad del Nivel indicado. Los demás valores expresados son la orientación (Ang) y el tamaño (Mod) del vector indicador de

Candidato #	Nivel	Fiabilidad	Mod A1	Ang A1	Mod A2	Ang A2	Mod B1	Ang B1	Mod B2	Ang B2	Mod C1	Ang C1	Mod C2	Ang C2
1	3	1,66	2,83	-106,76	1,34	19,85	1,66	88,29	0,48	-154,24	0,45	86,21	0,68	-132,84
2	0	1,66	2,30	-106,51	1,91	173,54	1,63	90,72	1,38	-3,85	0,69	-72,03	0,68	-132,84
3	3	1,98	1,98	-106,20	0,67	112,85	1,98	53,59	1,10	-118,37	0,33	-131,14	0,68	-132,84
4	4	2,23	1,97	-143,24	1,36	-162,43	1,19	128,42	2,23	8,40	0,34	-114,76	0,68	-132,84
5	0	0,67	3,41	-72,84	1,57	123,44	0,64	24,31	2,35	101,44	0,67	66,52	0,68	-132,84
6	3	2,19	1,84	-90,16	1,18	145,51	2,19	58,48	1,64	-101,58	0,71	-126,07	0,86	45,54
7	4	1,69	2,37	-89,77	0,89	117,63	1,21	125,55	1,69	35,89	1,33	-129,33	0,68	-132,84
8	2	2,51	2,48	-85,93	2,51	72,32	0,32	0,30	1,46	163,61	2,02	49,63	0,68	-132,84
9	2	1,99	1,02	-82,62	1,99	54,87	1,17	12,91	1,43	-168,54	0,92	-132,69	0,68	-132,84
10	3	1,25	0,82	-141,51	1,54	-126,63	1,25	65,14	0,73	9,86	1,17	-123,23	0,68	-132,84
11	4	2,94	1,94	-127,02	1,40	-137,79	1,67	138,12	2,94	27,77	1,48	-116,96	0,68	-132,84
12	5	2,84	2,22	-125,68	1,20	88,39	1,11	165,63	0,33	34,76	2,84	28,14	0,68	-132,84
13	6	2,84	1,67	-118,29	1,94	115,72	0,40	-109,24	0,71	72,88	0,84	13,64	2,84	38,91
14	4	1,19	3,01	-116,27	0,93	155,98	1,14	94,53	1,19	56,00	0,85	-7,54	0,23	27,03
15	4	1,11	2,13	-108,72	1,77	110,72	0,77	84,86	1,11	16,02	1,56	-122,44	0,68	-132,84
16	4	1,22	0,53	-145,98	1,26	-158,41	0,51	107,15	1,22	3,49	0,41	-134,13	0,68	-132,84
17	3	2,01	2,12	-105,37	1,08	-150,92	2,01	72,67	0,52	-16,87	0,81	-117,07	0,68	-132,84
18	3	2,63	2,26	-107,32	2,20	90,39	2,63	47,62	2,70	-116,65	0,91	-97,08	0,68	-132,84
19	4	1,79	0,63	-85,39	1,04	167,63	1,07	175,76	1,79	26,19	1,00	-108,57	1,00	36,95
20	3	2,84	3,40	-117,65	0,89	-131,19	2,84	73,59	1,01	-13,37	1,63	-145,65	0,68	-132,84

Tabla A.21: Valores-vectores de nivel sintáctico. Nivel superior. Texto-1.

Candidato #	A1	A2	B1	B2	C1	C2	Total
1	19	20	82	21	3	0	145
2	38	34	145	38	2	0	257
3	38	34	135	30	3	0	240
4	31	20	90	27	2	0	170
5	15	21	66	28	3	0	133
6	37	33	133	25	2	1	231
7	42	43	157	52	1	0	295
8	25	33	87	29	6	0	180
9	25	22	66	19	1	0	133
10	52	25	137	39	1	0	254
11	42	26	133	46	0	0	247
12	26	25	80	25	5	0	161
13	39	38	105	38	4	2	226
14	56	58	230	75	6	1	426
15	40	43	135	41	0	0	259
16	71	39	170	51	4	0	335
17	44	33	166	42	2	0	287
18	31	37	120	18	1	0	207
19	47	31	114	40	1	1	234
20	26	24	133	31	1	0	215

Tabla A.22: Número de estructuras sintácticas por niveles de referencia. Texto-1.

Candidato #	Items NC	A1	A2	B1	B2	C1	C2	Total
1	30	3	22	45	40	10	0	150
2	39	5	4	61	80	15	0	204
3	42	6	19	82	58	13	0	220
4	13	0	5	29	51	7	0	105
5	35	10	9	49	39	11	0	153
6	32	8	11	71	53	10	2	187
7	34	10	19	61	93	9	0	226
8	38	8	22	55	35	19	0	177
9	23	6	19	49	27	6	0	130
10	24	5	12	67	70	9	0	187
11	22	2	10	41	87	8	0	170
12	29	1	13	31	42	20	0	136
13	28	4	12	56	60	18	5	183
14	56	2	19	90	105	29	2	303
15	29	5	15	64	78	8	0	199
16	47	8	13	77	101	15	0	261
17	31	6	13	75	74	12	0	211
18	43	5	20	85	45	13	0	211
19	25	7	9	38	66	9	2	156
20	31	0	14	67	70	4	0	186

Tabla A.23: Número de *PoS* o categorías gramaticales computadas con el criterio de máximo nivel y distribuidos por niveles de referencia. Nivel superior. Texto-1.

nivel, valores indicadores de un nivel en un cuadrante para marcar el nivel del texto como se mostraba en la figura 7.10.

La tabla A.25 representa el histograma de estructuras sintácticas niveladas del Texto-2 del DELE superior.

En la tabla A.26 se registra el número de etiquetas de las categorías gramaticales niveladas o *part of speech* (*PoS*) según el criterio de máximo nivel. Esta figura lista también una serie de elementos no computados (Items NC) que incluyen mayormente signos de puntuación y algún *PoS* solo, que queda sin calificar al estar excluido o no pertenecer a ninguna de las estructuras registradas en el fichero de estructuras.

A.3. Discursos navideños del Rey

A.3.1. Nivel léxico

La tabla A.27 registra, para cada discurso, el porcentaje de lemas hallados en cada diccionario, el "Total" de lemas procesados y el número de lemas (repetidos y no repetidos) que conforman el discurso.

En la tabla A.28 aparece, para cada discurso, la distribución de lemas por niveles de referencia y el número total de lemas calificados.

La fiabilidad de los niveles de referencia calculados para cada discurso se observa en

Candidato #	Nivel	Fiabilidad	Mod A1	Ang A1	Mod A2	Ang A2	Mod B1	Ang B1	Mod B2	Ang B2	Mod C1	Ang C1	Mod C2	Ang C2
1	5	2,28	3,74	-105,95	2,38	148,10	1,66	38,52	0,79	102,29	2,28	53,70	0,68	-132,84
2	5	1,58	2,16	-123,41	1,50	148,77	1,34	26,85	0,77	137,22	1,58	1,49	0,68	-132,84
3	3	2,75	3,26	-106,10	2,09	111,55	2,75	39,99	1,29	-139,06	0,38	-106,77	0,68	-132,84
4	3	1,60	1,44	-155,81	1,90	142,59	1,60	24,23	1,18	-64,43	1,11	-129,23	0,68	-132,84
5	4	1,20	1,75	-93,78	1,00	98,93	0,35	105,79	1,20	69,89	0,80	-107,63	0,68	-132,84
6	3	1,20	0,71	-144,12	0,63	159,10	1,20	28,98	0,63	-135,02	0,62	-61,82	0,68	-132,84
7	3	1,66	2,49	-106,20	1,88	120,00	1,66	39,36	0,15	-169,76	0,68	-109,65	0,68	-132,84
8	3	2,08	3,17	-101,67	4,65	92,96	2,08	15,75	1,85	-138,22	0,14	64,13	0,68	-132,84
9	2	3,39	2,65	-98,14	3,39	83,83	0,32	156,40	0,24	-73,92	0,97	65,92	0,68	-132,84
10	3	2,53	3,00	-108,26	2,03	-139,50	2,53	49,27	1,02	136,42	1,75	46,07	0,68	-132,84
11	5	3,08	1,65	-113,66	1,54	-112,39	0,96	80,39	0,87	137,37	3,08	40,28	0,68	-132,84
12	3	1,20	1,52	-106,30	1,08	73,62	1,20	42,99	0,97	-152,49	0,31	-41,24	0,68	-132,84
13	3	1,45	1,92	-109,15	1,61	124,45	1,45	25,21	0,36	120,93	0,67	-115,47	0,68	-132,84
14	3	1,59	1,22	-112,20	1,39	164,43	1,59	38,22	1,01	-122,21	1,00	23,28	0,68	-132,84
15	0	0,75	2,07	-102,42	1,70	98,60	0,52	32,05	0,39	85,47	0,49	146,57	0,75	23,17
16	3	1,87	0,86	-176,60	0,64	167,22	1,87	38,52	1,91	-98,06	0,88	-129,32	0,68	-132,84
17	3	2,17	2,80	-104,63	2,16	125,20	2,17	56,30	1,08	-70,34	1,24	-126,32	0,68	-132,84
18	4	1,68	3,31	-94,97	2,03	110,01	0,83	45,72	1,68	86,90	0,50	-94,29	0,68	-132,84
19	0	0,95	1,71	-132,75	0,77	-162,03	0,80	61,95	0,95	49,94	0,41	-87,73	0,68	-132,84
20	3	1,50	0,95	-86,12	1,42	156,99	1,50	42,77	1,29	-105,11	0,40	35,23	0,68	-132,84

Tabla A.24: Valores-vectores de nivel sintáctico. Nivel superior. Texto-2.

# Candidato	A1	A2	B1	B2	C1	C2	Total
1	22	34	116	37	8	0	217
2	61	55	194	62	6	0	378
3	42	62	205	46	4	0	359
4	38	29	97	23	1	0	188
5	50	46	153	55	2	0	306
6	69	44	170	46	3	0	332
7	36	42	136	38	2	0	254
8	21	42	91	21	3	0	178
9	26	40	95	28	5	0	194
10	21	16	96	28	5	0	166
11	31	17	96	30	7	0	181
12	40	35	121	32	3	0	231
13	58	57	185	57	3	0	360
14	46	34	133	32	5	0	250
15	50	54	161	52	6	1	324
16	59	35	145	27	2	0	268
17	35	45	153	33	1	0	267
18	35	54	161	61	3	0	314
19	34	23	98	32	2	0	189
20	42	31	118	26	4	0	221

Tabla A.25: Representa el histograma de estructuras sintácticas niveladas del Texto-2 del DELE superior.

#Candidato	Items NC	A1	A2	B1	B2	C1	C2	Total
1	35	3	3	78	59	23	0	201
2	28	3	13	109	76	35	0	264
3	57	6	22	144	78	21	0	328
4	14	1	5	55	47	6	0	128
5	43	12	28	91	101	18	0	293
6	52	8	22	110	81	23	0	296
7	53	6	15	97	74	15	0	260
8	32	4	14	68	33	11	0	162
9	35	6	21	53	57	15	0	187
10	22	3	6	71	44	19	0	165
11	27	4	12	51	42	24	0	160
12	50	7	25	81	55	17	0	235
13	58	8	19	121	91	18	0	315
14	50	7	10	89	62	23	0	241
15	47	8	23	87	80	14	3	262
16	34	4	15	82	59	10	0	204
17	30	5	10	81	70	9	0	205
18	38	10	20	98	89	19	0	274
19	29	2	12	58	61	12	0	174
20	28	9	9	76	57	16	0	195

Tabla A.26: Número de *PoS* o categorías gramaticales computadas con el criterio de máximo nivel y distribuidos por niveles de referencia. Nivel superior. Texto-2.

la tabla A.29, además de la tendencia hacia un nivel superior o inferior al Nivel 4 (B2).

A.3.2. Nivel sintáctico

Con la tabla A.30 se muestran todos los valores de los vectores que apuntan hacia cada nivel de referencia. Cada vector tiene dos parámetros: el módulo (Mod), o distancia del vector al centro, y el ángulo (Ang), u orientación del vector en uno u otro cuadrante. El cuadrante que marca el nivel sintáctico es el que se sitúa entre 0^{o} y 90^{o}. Por ejemplo, el discurso 1975, #1, tiene un valor sintáctico de B1. El "Nivel: 0" indica que el texto no se califica porque queda dentro del umbral de referencia. Además, en esos casos de "Nivel 0" o no calificado, la fiabilidad está por debajo de cero.

En la figura A.1 se visualizan, como puntos con color de nivel, los valores de los vectores en longitud (Módulo) y orientación (Ángulo) que reproducimos en la tabla A.30. La mayoría de los valores aparecen dentro del círculo que define el umbral ya que estos discursos son los que utilizan como referencia para la nivelación sintáctica. Recordamos que aquellos vectores de otros textos procesados y que se ubiquen dentro del primer cuadrante (0-90^{o}) son los que indicarán el nivel de referencia de dicho texto para la sintaxis.

En la tabla A.31 se lista el número de estructuras sintácticas de cada discurso por niveles de referencia.

La reagrupación de los *PoS* de la tabla A.31, cuando se aplica el criterio de máximo nivel, permite renivelar el total de las estructuras como se observa en la tabla A.32.

Discurso-año	Kincaid	Wiktionary	Fuensanta	Frlng 1	Frlng 2	esWordnet	Cervantes	Multivocablos	Total	# Glosados
1975	75,94	88,87	35,19	89,05	95,69	66,61	57,27	3,41	99,46	557
1976	71,41	86,65	33,88	86,78	93,32	64,61	57,56	5,67	99,50	794
1977	76,68	87,47	31,51	87,21	94,34	62,40	59,44	4,44	99,83	1149
1978	75,40	85,26	33,13	87,92	95,12	65,99	60,30	3,82	99,47	1126
1979	72,23	84,70	31,31	86,53	93,31	64,01	56,19	5,16	99,91	2287
1980	73,67	84,52	31,16	87,44	94,57	65,53	57,19	4,32	100,00	995
1981	73,23	86,20	32,88	88,41	95,15	66,81	58,27	3,90	100,00	949
1982	72,41	85,03	32,09	85,67	94,33	63,42	56,58	5,03	99,79	935
1983	70,94	83,99	30,66	85,99	92,55	63,33	55,72	5,76	99,84	1249
1984	73,25	86,43	33,04	86,14	93,61	63,13	56,54	4,62	99,90	1017
1985	71,16	84,94	31,71	86,06	93,86	63,98	58,09	4,62	100,00	1255
1986	72,52	83,51	33,31	89,68	94,67	64,86	59,95	4,08	99,33	1201
1987	72,58	84,98	33,19	86,51	93,36	64,96	58,87	4,57	99,67	919
1988	73,30	84,21	33,30	86,03	93,49	65,07	58,85	4,50	99,52	1045
1989	77,00	87,15	32,04	88,82	93,74	63,05	58,13	4,68	99,60	1261
1990	75,51	86,17	33,64	88,44	94,86	66,06	60,47	3,85	99,55	1323
1991	73,81	84,76	32,04	88,55	93,57	66,06	59,64	4,53	99,34	1214
1992	72,17	83,35	31,51	85,65	92,29	64,53	55,80	5,41	99,42	1387
1993	75,24	85,16	33,24	87,40	93,77	64,95	59,09	4,85	99,57	1381
1994	73,86	83,88	35,04	87,88	93,62	66,29	54,87	5,33	99,65	1427
1995	73,41	83,43	31,66	86,64	93,69	65,06	57,95	4,70	99,44	1617
1996	74,48	85,49	34,46	87,90	93,74	67,68	56,33	4,54	99,52	1454
1997	74,23	84,69	33,70	88,34	93,18	66,61	56,86	5,15	99,44	1261
1998	73,62	83,59	34,51	86,35	91,95	65,57	59,59	5,90	99,54	1304
1999	74,94	84,30	34,97	88,58	93,34	67,88	56,86	4,60	99,52	1261
2000	74,29	85,99	31,84	86,67	93,33	67,21	57,28	5,17	99,32	735
2001	72,67	83,97	32,60	88,36	93,63	66,52	58,62	5,05	99,56	911
2002	73,95	81,99	33,30	86,54	91,52	68,01	56,12	6,29	99,39	1144
2003	71,74	83,16	32,64	87,17	92,36	65,58	56,90	5,19	98,59	1348
2004	71,78	82,24	31,41	86,10	91,37	67,84	55,86	5,03	97,99	1194
2005	73,22	83,46	33,61	87,25	92,79	70,18	58,95	5,39	99,62	1318
2006	73,54	84,16	32,66	88,42	93,05	67,71	56,13	4,93	99,25	1338
2007	71,38	81,17	31,48	85,86	91,91	66,49	56,02	5,64	99,12	1471
2008	72,37	82,40	31,65	85,74	92,70	67,26	56,68	5,73	99,45	1466
2009	71,73	81,03	31,40	84,90	91,74	66,67	56,40	6,62	99,48	1344
2010	73,35	83,33	35,53	86,63	92,22	70,76	56,99	5,89	99,30	1002

Tabla A.27: Distribución de lemas por diccionarios en los discursos navideños del Rey.

Discurso-año	A1	A2	B1	B2	C1	C2	# Calificados
1975	31,03	11,53	28,93	15,30	9,22	3,98	477
1976	31,40	11,05	28,49	14,83	9,45	4,80	688
1977	34,32	12,42	28,00	11,30	8,86	5,09	982
1978	33,47	12,04	27,04	12,14	9,90	5,41	980
1979	29,61	10,34	28,14	15,47	11,05	5,38	1972
1980	29,33	11,20	29,10	15,94	9,35	5,08	866
1981	32,53	11,25	26,48	13,42	10,04	6,29	827
1982	29,59	11,61	28,09	14,98	10,61	5,12	801
1983	27,72	12,08	28,75	14,61	11,70	5,15	1068
1984	29,50	11,73	28,34	13,59	10,69	6,16	861
1985	28,56	10,72	30,96	13,49	10,54	5,73	1082
1986	32,30	9,98	26,27	14,88	11,39	5,18	1062
1987	27,89	11,18	30,10	14,86	10,07	5,90	814
1988	29,42	11,42	28,87	12,29	11,09	6,92	911
1989	31,26	11,92	30,34	14,39	8,89	3,21	1091
1990	31,45	10,79	30,35	12,56	9,36	5,48	1186
1991	30,37	9,53	28,97	16,07	10,37	4,67	1070
1992	28,56	10,74	30,47	13,41	11,49	5,33	1201
1993	30,07	13,01	28,75	13,92	9,69	4,56	1207
1994	25,57	11,81	31,55	16,83	9,63	4,61	1236
1995	31,39	10,42	26,81	15,41	9,78	6,19	1421
1996	26,83	12,38	30,95	15,16	9,92	4,76	1260
1997	29,43	9,63	29,52	16,20	10,35	4,86	1111
1998	28,83	12,46	29,27	14,29	9,76	5,40	1148
1999	29,36	10,48	30,62	14,81	9,58	5,15	1107
2000	33,18	9,43	29,25	15,41	7,86	4,87	636
2001	31,49	10,46	29,15	14,39	8,86	5,66	813
2002	28,37	9,23	29,56	15,58	11,11	6,15	1008
2003	29,58	8,44	29,92	16,11	11,00	4,94	1173
2004	29,37	7,41	30,42	15,40	11,79	5,61	1052
2005	29,41	8,64	30,08	15,34	11,61	4,92	1180
2006	27,88	8,90	29,72	16,96	11,75	4,79	1191
2007	27,56	8,39	30,28	14,83	12,97	5,98	1288
2008	26,59	8,92	31,90	16,06	10,15	6,38	1301
2009	26,35	9,14	29,18	17,29	10,97	7,07	1203
2010	25,77	11,45	29,07	15,75	11,78	6,17	908

Tabla A.28: Distribución de lemas por niveles. Aparece, para cada discurso, la distribución de lemas por niveles de referencia y el número total de lemas calificados.

Discurso-año	Nivel	Tendencia	Nivel+Tendencia	Fiabilidad
1975	4	0,23	4,23	0,08
1976	4	-0,69	3,31	0,15
1977	4	-0,92	3,08	0,21
1978	4	-0,65	3,35	0,19
1979	4	0,26	4,26	0,19
1980	4	0,24	4,24	0,20
1981	4	-0,35	3,65	0,20
1982	4	-0,14	3,86	0,11
1983	4	-0,25	3,75	0,19
1984	4	-0,44	3,56	0,14
1985	4	0,08	4,08	0,18
1986	4	0,54	4,54	0,16
1987	4	-0,22	3,78	0,18
1988	4	-0,53	3,47	0,15
1989	4	-0,44	3,56	0,22
1990	4	-0,37	3,63	0,13
1991	4	0,23	4,23	0,22
1992	4	-0,10	3,90	0,14
1993	4	-0,12	3,88	0,21
1994	4	-0,05	3,95	0,15
1995	4	0,09	4,09	0,19
1996	4	-0,24	3,76	0,21
1997	4	0,33	4,33	0,11
1998	4	-0,54	3,46	0,14
1999	4	0,09	4,09	0,14
2000	4	-0,77	3,23	0,18
2001	4	-0,83	3,17	0,22
2002	4	0,09	4,09	0,24
2003	4	-0,40	3,60	0,24
2004	4	0,13	4,13	0,18
2005	4	-0,23	3,77	0,22
2006	4	-0,07	3,93	0,17
2007	4	0,07	4,07	0,12
2008	4	0,11	4,11	0,20
2009	4	0,48	4,48	0,18
2010	4	-0,47	3,53	0,17

Tabla A.29: Representación de la fiabilidad de los niveles de referencia calculados para cada discurso se observa en esta tabla, además de la tendencia hacia un nivel superior o inferior al Nivel 4 (B2).

Discurso-año	Discurso #	Nivel	Fiabilidad	Mod A1	Ang A1	Mod A2	Ang A2	Mod B1	Ang B1	Mod B2	Ang B2	Mod C1	Ang C1	Mod C2	Ang C2
1975	1	3	1,05	0,94	-120,26	0,73	126,26	1,05	49,75	0,88	-85,19	0,51	-150,72	0,24	-132,49
1976	2	0	0,71	0,28	94,25	0,74	-111,87	0,71	51,60	0,75	-106,90	0,22	-117,43	0,28	-9,77
1977	3	0	0,54	0,16	-118,19	0,26	-139,34	0,54	29,86	0,08	123,45	0,64	-129,87	0,26	-133,06
1978	4	0	0,24	0,44	-33,68	0,38	-142,96	0,18	110,44	0,18	55,73	0,10	140,77	0,24	35,09
1979	5	0	0,37	0,35	-150,58	0,12	-154,45	0,37	16,78	0,17	153,19	0,09	26,25	0,26	-133,55
1980	6	0	0,54	0,35	-141,05	0,16	-84,91	0,33	75,77	0,28	-121,07	0,54	30,42	0,68	-132,84
1981	7	5	1,40	0,60	-99,42	0,61	117,40	0,35	-164,58	0,09	29,69	1,40	46,15	0,43	-132,44
1982	8	0	0,64	0,25	28,64	0,36	-151,47	0,64	6,91	0,56	-163,56	0,12	-25,91	0,27	75,31
1983	9	3	1,09	0,74	-82,67	0,65	-150,93	1,09	42,42	0,40	-169,95	0,10	-47,89	0,68	-132,84
1984	10	0	0,72	0,68	-93,92	0,07	-135,53	0,40	102,91	0,72	51,47	0,92	-126,80	0,44	-132,83
1985	11	0	0,65	0,54	-133,07	0,32	156,71	0,65	41,84	0,27	-83,57	0,25	-128,39	0,68	-132,84
1986	12	0	0,35	0,61	90,24	0,70	-107,37	0,44	165,73	0,81	-15,07	0,77	-136,89	0,35	73,89
1987	13	0	0,47	0,28	-169,16	0,19	-168,64	0,47	9,21	0,23	-171,79	0,05	-62,18	0,11	43,49
1988	14	0	0,78	0,42	-125,10	0,30	-173,87	0,48	64,11	0,56	-105,65	0,78	43,93	0,45	-132,40
1989	15	0	0,89	0,94	-154,48	0,61	-147,01	0,30	168,55	0,89	47,73	0,68	10,58	0,22	-152,06
1990	16	0	0,13	0,49	155,76	0,67	-166,22	0,13	7,74	0,52	-8,28	0,05	-125,88	0,50	-132,31
1991	17	0	0,87	0,29	175,03	0,29	178,73	0,44	-157,35	0,87	27,02	0,19	-137,01	0,48	-132,44
1992	18	0	0,50	0,51	-134,74	0,23	169,99	0,50	52,61	0,19	-52,05	0,51	-122,01	0,19	46,21
1993	19	0	0,76	0,21	177,92	1,00	-101,47	0,35	132,00	0,76	41,93	0,48	-131,00	0,13	82,00
1994	20	0	0,37	0,27	-54,13	0,25	166,61	0,14	-82,07	0,43	92,70	0,37	42,30	0,51	-133,27
1995	21	0	0,69	0,49	92,46	0,67	100,93	0,52	-36,72	0,63	-114,43	0,34	-126,97	0,69	47,79
1996	22	0	0,40	0,22	-121,71	0,08	-67,67	0,15	75,92	0,40	44,07	0,60	-139,38	0,68	-132,84
1997	23	0	0,40	0,49	169,97	0,50	-46,17	0,27	99,51	0,28	-50,05	0,27	-164,17	0,30	-131,83
1998	24	4	1,18	0,17	43,73	0,40	-113,71	0,71	-160,37	1,18	46,32	0,44	-120,58	0,50	-132,66
1999	25	0	0,55	0,55	71,76	0,33	-98,69	0,37	59,76	0,69	-72,33	1,05	-139,59	0,26	-120,56
2000	26	5	1,05	0,36	-136,30	0,86	37,12	0,19	-62,31	0,71	-161,61	1,05	57,43	0,02	144,34
2001	27	0	0,15	0,61	172,39	0,64	-33,77	0,10	60,82	0,15	75,70	0,35	-149,34	0,14	-135,62
2002	28	0	0,67	0,30	-36,97	0,67	78,63	0,67	-134,39	0,53	64,07	0,27	69,14	0,25	-134,72
2003	29	0	0,74	0,58	65,08	0,74	40,91	0,65	-121,69	0,45	-137,92	0,46	48,30	0,73	56,68
2004	30	0	0,86	0,86	35,48	0,80	15,97	0,40	-91,13	0,76	-159,40	0,28	119,34	0,67	55,19
2005	31	6	1,58	0,63	18,65	1,05	33,28	0,55	-136,34	0,53	-157,73	0,17	-169,66	1,58	41,89
2006	32	6	1,22	1,06	42,31	0,68	-21,37	1,17	-133,76	0,60	122,32	0,79	53,09	1,22	42,68
2007	33	6	1,10	0,56	27,80	0,32	54,98	0,71	-135,66	0,38	-179,65	1,04	37,42	1,10	36,31
2008	34	1	1,07	1,07	26,39	0,46	18,52	0,76	-135,82	0,21	131,96	0,17	127,14	0,44	42,79
2009	35	0	0,74	0,74	59,28	0,32	63,97	0,62	-82,54	0,39	171,93	0,08	-28,67	0,52	48,94
2010	36	0	0,87	0,33	24,93	0,87	43,39	0,68	-115,75	0,47	123,04	0,21	46,41	0,32	-159,11

Tabla A.30: Valores-vectores de niveles sintácticos de los discursos.

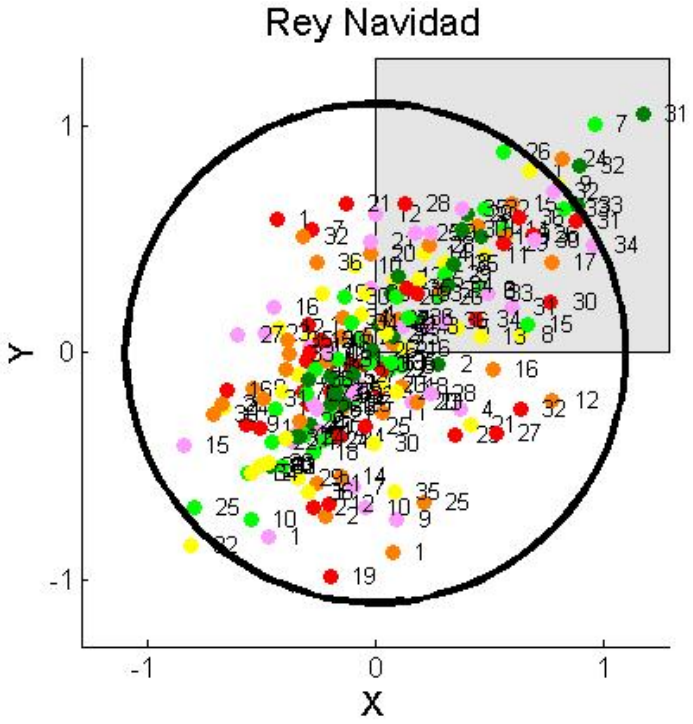


Figura A.1: Primer cuadrante superior derecho: Umbral de nivel de referencia de los discursos navideños.

Discurso-año	Discurso #	A1	A2	B1	B2	C1	C2	Total
1975	1	156	113	419	102	10	1	801
1976	2	264	127	583	150	15	2	1141
1977	3	361	205	826	253	16	2	1663
1978	4	345	197	794	249	26	4	1615
1979	5	708	413	1601	499	52	4	3277
1980	6	323	186	752	217	28	0	1506
1981	7	283	196	670	212	38	1	1400
1982	8	306	167	660	198	20	4	1355
1983	9	352	214	928	265	26	0	1785
1984	10	292	185	737	242	10	1	1467
1985	11	383	239	926	263	24	0	1835
1986	12	447	206	901	267	17	6	1844
1987	13	296	170	655	201	20	3	1345
1988	14	329	198	790	213	34	1	1565
1989	15	408	234	952	328	33	3	1958
1990	16	461	249	976	299	30	1	2016
1991	17	402	230	857	290	25	1	1805
1992	18	427	261	1024	298	21	5	2036
1993	19	456	214	1019	336	23	5	2053
1994	20	422	253	937	318	36	1	1967
1995	21	546	327	1072	310	28	9	2292
1996	22	458	267	1042	335	23	0	2125
1997	23	418	221	923	269	27	2	1860
1998	24	434	228	906	332	21	1	1922
1999	25	450	225	938	250	14	2	1879
2000	26	216	141	480	146	26	2	1011
2001	27	287	152	623	197	17	2	1278
2002	28	342	226	728	258	29	2	1585
2003	29	444	257	845	264	36	8	1854
2004	30	353	196	686	212	27	6	1480
2005	31	380	234	772	241	25	10	1662
2006	32	431	213	774	287	40	9	1754
2007	33	465	269	925	303	46	9	2017
2008	34	462	254	888	300	33	6	1943
2009	35	423	233	788	264	26	6	1740
2010	36	303	189	606	215	23	2	1338

Tabla A.31: Histograma de estructuras sintácticas. Se lista el número de estructuras sintácticas de cada discurso por niveles de referencia.

Discurso-año	Discurso #	Items NC	A1	A2	B1	B2	C1	C2	Total
1975	1	86	20	54	233	207	35	2	637
1976	2	133	40	83	317	275	61	7	916
1977	3	215	57	127	471	423	75	4	1372
1978	4	183	72	118	413	426	89	9	1310
1979	5	327	96	263	917	822	197	8	2630
1980	6	122	43	121	379	359	102	0	1126
1981	7	126	46	100	327	356	116	2	1073
1982	8	164	55	97	392	303	81	6	1098
1983	9	233	68	113	536	417	105	0	1472
1984	10	188	51	119	370	413	60	2	1203
1985	11	184	50	132	513	468	95	0	1442
1986	12	192	61	129	400	513	69	8	1372
1987	13	144	39	101	372	321	76	6	1059
1988	14	123	46	110	407	375	111	2	1174
1989	15	170	30	117	434	522	137	4	1414
1990	16	185	49	115	505	547	107	2	1510
1991	17	173	51	127	411	525	92	2	1381
1992	18	229	56	152	552	531	99	10	1629
1993	19	207	62	151	486	574	95	8	1583
1994	20	209	81	153	530	523	133	2	1631
1995	21	235	82	182	647	558	118	18	1840
1996	22	193	70	177	546	575	92	0	1653
1997	23	185	45	172	461	482	89	4	1438
1998	24	182	73	146	419	573	95	2	1490
1999	25	185	72	147	483	485	61	5	1438
2000	26	109	32	114	279	231	78	4	847
2001	27	123	29	133	340	338	63	4	1030
2002	28	193	68	144	382	443	98	4	1332
2003	29	211	80	198	462	457	127	14	1549
2004	30	222	88	187	437	365	90	12	1401
2005	31	240	94	216	449	432	99	26	1556
2006	32	219	104	204	416	460	137	22	1562
2007	33	245	100	190	490	497	172	24	1718
2008	34	242	123	210	488	529	115	14	1721
2009	35	269	86	170	506	452	114	13	1610
2010	36	179	62	153	352	351	89	2	1188

Tabla A.32: Histograma de *PoS* nivelados.

Lectura-I #	Kincaid	Wiktionary	Fuensanta	Frlng 1	Frlng 2	esWordnet	Cervantes	Multivocablos	Total	# Glosados
1	71,82	86,96	35,91	80,00	91,25	63,26	59,23	5,69	99,30	1565
2	69,34	84,49	31,30	79,87	88,27	63,07	55,12	4,99	95,84	1083
3	78,10	88,29	33,96	85,19	89,93	63,82	62,35	5,56	97,83	1708
4	75,75	84,92	33,50	83,33	90,08	65,00	60,33	4,92	95,75	1200
5	70,54	88,26	32,55	83,88	93,28	65,85	64,57	4,91	98,61	937
6	72,20	88,26	31,67	83,79	92,20	61,97	61,44	5,98	98,56	1320
7	76,55	90,19	33,75	87,20	94,43	67,92	65,07	4,11	98,96	1437
8	68,35	82,97	27,67	81,30	90,22	66,47	56,90	3,28	94,95	1861
9	69,93	86,07	32,49	80,96	91,98	65,95	56,48	6,40	98,87	1859
10	71,80	84,41	32,72	82,96	93,36	66,41	66,70	3,75	97,59	1039
11	74,00	91,00	33,89	85,67	94,44	66,00	61,78	4,11	99,33	900

Lectura-A #	Kincaid	Wiktionary	Fuensanta	Frlng 1	Frlng 2	esWordnet	Cervantes	Multivocablos	Total	# Glosados
1	70,93	84,32	29,32	78,27	89,06	63,69	55,79	6,34	96,35	2002
2	69,31	79,76	35,24	86,34	94,60	71,33	59,87	2,87	98,48	593
3	72,45	86,99	34,19	81,75	88,73	64,98	58,25	4,02	94,63	2290
4	70,69	80,80	32,07	80,69	87,24	61,49	61,03	4,71	93,22	870
5	74,95	88,36	33,06	82,66	92,29	64,06	53,53	6,05	98,99	1984
6	71,51	85,15	31,23	81,37	89,36	62,15	56,51	5,24	95,35	3205
7	67,42	78,46	31,57	82,73	92,52	64,99	58,40	2,55	96,20	1685
8	75,04	87,20	34,34	82,27	90,77	65,07	54,96	5,57	97,36	2047
9	67,61	77,15	25,51	82,00	89,05	64,63	57,59	3,76	94,84	639
10	73,08	86,46	32,54	82,50	90,03	65,80	57,07	5,75	97,49	2348
11	72,49	83,96	32,58	82,97	89,80	63,03	55,64	5,98	97,34	5685
12	71,50	84,89	35,26	84,77	91,52	66,71	58,97	4,79	97,30	814

Tabla A.33: Distribución de lemas por diccionarios de las lecturas de nivel intermedio (I) y avanzado (A).

A.4. "Lecturas paso a paso" del Centro Virtual Cervantes (CVC)

Denominamos indistintamente a estas lecturas en su conjunto, según su nivel: "CVC 1" o "Lectura I" para el nivel intermedio, y "CVC 2" o "Lectura A" para las lecturas de nivel avanzado.

A.4.1. Nivel léxico

En la tabla A.33 se lista el número de lemas correspondientes a cada diccionario. En la primera parte de la tabla, se representan las lecturas del nivel intermedio (Lectura-I) y, en la segunda, las lecturas del nivel avanzado (Lectura-A). En la figura A.2 se visualiza esta tabla A.33.

En la tabla A.34 se lista la distribución de lemas por niveles de aprendizaje, según el criterio del Instituto Cervantes, en cada nivel de lectura, intermedio (I) y avanzado (A).

Se visualiza esta tabla A.34 en la figura A.3. En color azul se representa el rango de error de cada lectura respecto a la media.

En la tabla A.35 se lista el nivel de referencia de cada lectura y la tendencia hacia un nivel u otro, además de la fiabilidad. La figura A.4 representa la visualización de la tabla anterior, A.35, que trata sobre el nivel léxico de las lecturas según los niveles de

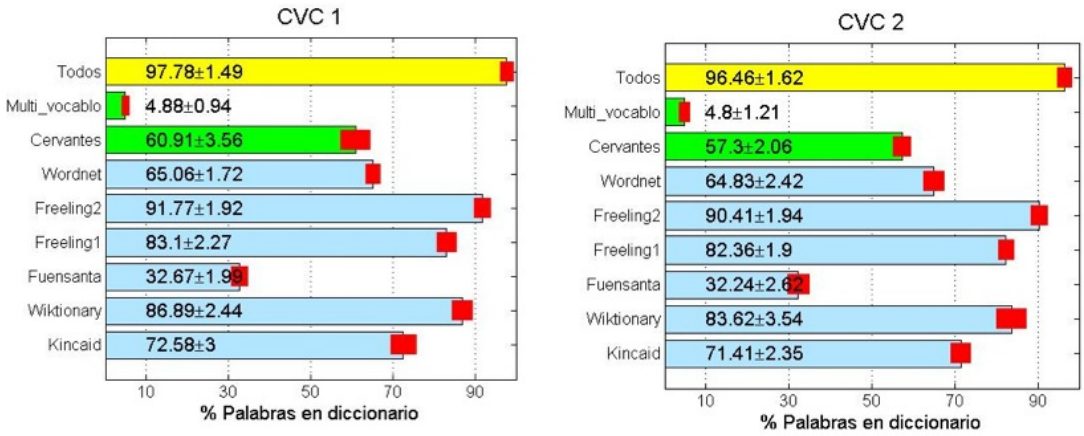


Figura A.2: Representación visual de la distribución de lemas de las lecturas en cada diccionario. En esta representación gráfica de la tabla A.33 también se observa el rango de error (en rojo) de los lemas de cada lectura pertenecientes a cada diccionario.

Lectura-I #	A1	A2	B1	B2	C1	C2	# Calificados
1	28,40	19,49	27,04	13,60	4,98	6,50	1324
2	34,86	18,51	27,54	11,43	2,86	4,80	875
3	35,39	16,34	31,12	9,83	3,66	3,66	1475
4	33,00	22,89	28,44	10,01	2,68	2,97	1009
5	34,70	16,06	23,33	15,12	5,39	5,39	853
6	31,16	19,19	26,94	12,76	4,40	5,55	1136
7	36,72	16,23	28,05	11,19	4,02	3,78	1269
8	35,22	16,88	25,80	10,13	4,14	7,83	1570
9	30,70	14,87	26,24	16,94	5,62	5,62	1547
10	37,37	13,06	23,98	12,53	5,89	7,17	934
11	32,10	16,62	29,28	14,19	3,84	3,96	782

Lectura-A #	A1	A2	B1	B2	C1	C2	# Calificados
1	31,95	19,66	27,21	11,20	4,32	5,66	1643
2	30,35	11,36	24,95	13,59	10,43	9,31	537
3	35,67	18,10	27,86	11,91	3,39	3,07	1856
4	33,61	16,46	23,32	13,31	7,27	6,04	729
5	31,63	15,81	31,12	12,78	4,81	3,86	1581
6	34,24	20,19	27,06	12,15	2,68	3,69	2576
7	29,58	11,12	25,26	14,82	9,54	9,68	1457
8	33,80	18,29	28,89	10,90	4,00	4,12	1651
9	33,99	10,25	26,26	10,61	10,25	8,63	556
10	32,67	18,19	27,85	12,45	3,86	4,98	1968
11	28,11	16,95	29,30	14,26	5,62	5,75	4642
12	32,06	10,88	29,38	13,56	8,76	5,37	708

Tabla A.34: Distribución de lemas por niveles de las lecturas de nivel intermedio (I) y avanzado (A).

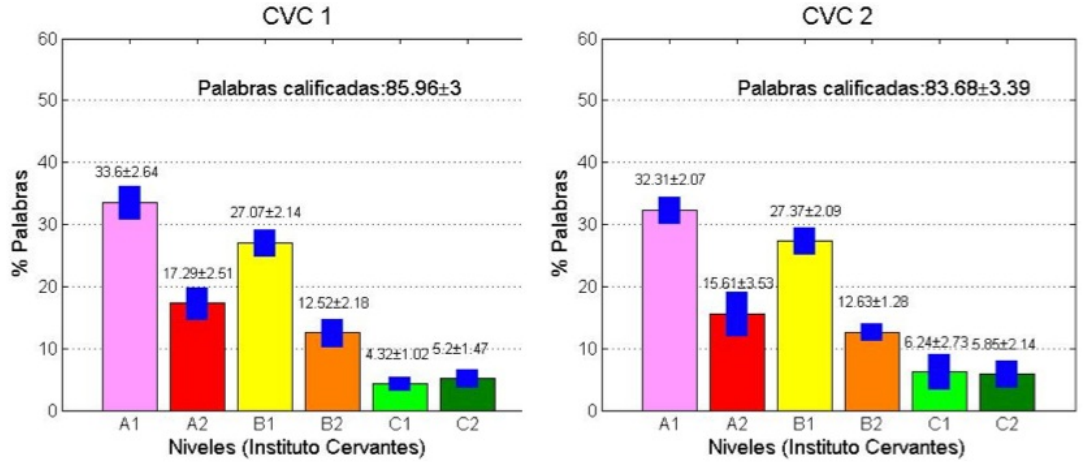


Figura A.3: Visualización de los lemas por niveles de las lecturas del Centro Virtual Cervantes (CVC) de nivel intermedio (CVC 1) y avanzado (CVC 2).

Lectura-I #	Nivel	Tendencia	Nivel+Tendencia	Fiabilidad
1	4	-0,96	3,04	0,15
2	4	-0,83	3,17	0,07
3	3	0,98	3,98	0,10
4	3	0,72	3,72	0,23
5	4	-0,58	3,42	0,15
6	4	-0,32	3,68	0,08
7	4	-0,74	3,26	0,16
8	4	-0,24	3,76	0,12
9	2	0,93	2,93	0,11
10	4	-0,44	3,56	0,13
11	3	0,51	3,51	0,18

Lectura-A #	Nivel	Tendencia	Nivel+Tendencia	Fiabilidad
	4	-0,61	3,39	0,13
2	4	0,58	4,58	0,14
3	4	-0,91	3,09	0,16
4	3	0,95	3,95	0,11
5	4	-0,50	3,50	0,04
6	3	0,80	3,80	0,08
7	4	0,58	4,58	0,14
8	4	-0,81	3,19	0,12
9	4	-0,49	3,51	0,16
10	4	-0,71	3,29	0,13
11	4	0,19	4,19	0,19
12	4	-0,80	3,20	0,15

Tabla A.35: Fiabilidad del nivel del léxico de las lecturas de nivel intermedio (Lectura I) y avanzado (Lectura A).

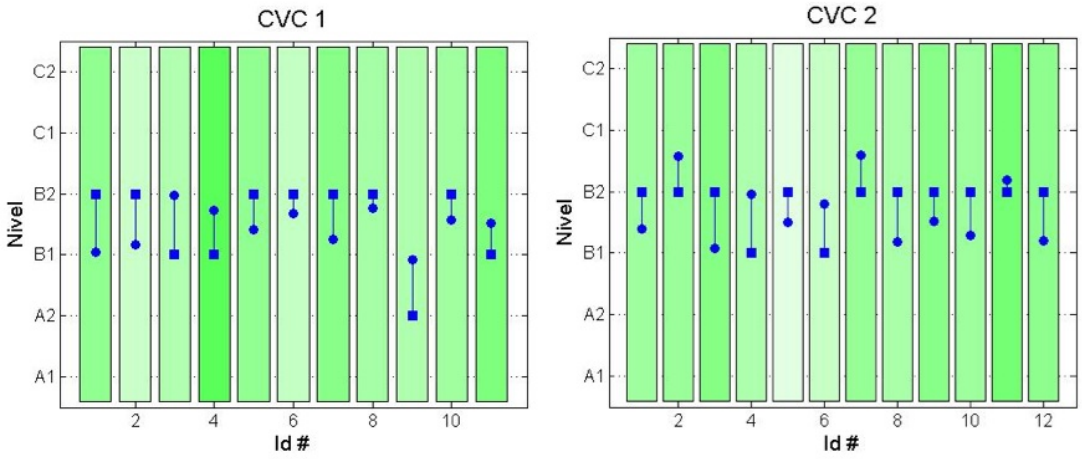


Figura A.4: Nivel léxico de las lecturas del Centro Virtual Cervantes (CVC) de nivel intermedio (CVC 1) y avanzado (CVC 2).

referencia.

A.4.2. Nivel sintáctico

Con la tabla A.36 se muestran todos los valores de los vectores que apuntan hacia cada un nivel de referencia. Cada vector tiene dos parámetros: el módulo (Mod), o distancia del vector al centro, y el ángulo (Ang), u orientación del vector en uno u otro cuadrante. El cuadrante que marca el nivel sintáctico es el que se sitúa entre 0^{o} y 90^{o}. Por ejemplo, la Lectura-I #1, tiene un valor sintáctico de B1. En otras lecturas, el "Nivel: 0" indica que dicha lectura no se califica porque queda dentro del umbral de referencia. En esos casos de "Nivel 0" o no calificado, la fiabilidad está por debajo de 1.

En la figura A.36 se visualizan, como puntos con color de nivel, los valores de los vectores en longitud (Módulo) y orientación (Ángulo) que reproducimos en la tabla A.5. La mayoría de los valores aparecen fuera del círculo que define el umbral para las lecturas de nivel intermedio (CVC 1). El nivel sintáctico de las lecturas avanzadas (CVC 2) se sitúa dentro y fuera del umbral.

Lectura-I #	Nivel	Fiabilidad	Mod A1	Ang A1	Mod A2	Ang A2	Mod B1	Ang B1	Mod B2	Ang B2	Mod C1	Ang C1	Mod C2	Ang C2
1	3	1,08	2,13	-107,15	1,16	117,19	1,08	43,20	0,51	100,95	0,43	-99,52	0,29	-117,48
2	0	1,00	1,53	-101,63	1,00	59,39	0,99	61,95	0,95	-144,67	0,19	-29,68	0,86	44,86
3	3	1,26	1,44	-109,39	0,70	143,01	1,26	39,68	0,41	-169,12	0,24	-69,25	0,05	69,20
4	2	1,36	1,67	-82,41	1,36	83,59	1,30	44,55	1,15	-139,90	0,19	-161,99	0,46	-134,50
5	3	1,91	0,60	-158,02	1,66	-141,78	1,91	35,44	0,48	-94,27	0,98	-128,94	0,68	-132,84
6	3	1,01	1,40	-112,80	0,82	128,17	1,01	47,18	0,14	-134,35	0,36	-71,88	0,68	-132,84
7	3	1,38	0,53	-104,99	0,88	162,89	1,38	40,39	0,87	-90,26	0,98	-131,03	0,68	-132,84
8	3	1,38	0,56	0,94	0,85	63,39	1,38	24,84	2,07	-130,72	0,78	-130,73	0,35	59,88
9	3	1,64	2,44	-100,33	1,54	126,84	1,64	38,27	0,44	147,05	0,32	-97,45	0,54	-133,35
10	3	1,17	0,59	-164,62	1,31	-138,87	1,17	55,75	0,81	-49,30	0,62	-140,30	0,68	-132,84
11	3	1,64	1,33	-122,42	1,56	-141,96	1,64	50,97	0,36	-123,23	0,63	34,44	0,12	30,22

Lectura-A #	Nivel	Fiabilidad	Mod A1	Ang A1	Mod A2	Ang A2	Mod B1	Ang B1	Mod B2	Ang B2	Mod C1	Ang C1	Mod C2	Ang C2
1	0	0,79	2,16	-78,81	2,40	93,79	0,79	8,84	0,94	162,53	0,38	36,24	0,38	-132,29
2	0	0,79	0,93	98,18	0,70	-46,41	0,45	-159,65	0,71	-1,76	1,23	-143,57	0,68	-132,84
3	3	1,21	1,45	-87,42	0,56	72,69	1,21	55,52	0,72	-149,95	0,25	-124,83	0,56	-133,51
4	0	0,64	0,29	-60,80	0,64	15,98	0,32	48,72	0,35	-157,33	0,48	-160,12	0,68	-132,84
5	3	2,02	3,10	-103,85	1,51	151,57	2,02	53,58	0,12	105,50	0,15	37,05	0,31	-133,36
6	2	1,73	1,83	-75,48	1,73	69,21	1,34	49,12	1,64	-142,20	0,19	-103,29	0,20	-153,14
7	0	0,85	0,85	80,50	0,48	-47,16	0,60	13,01	1,13	-123,00	0,26	-165,14	0,37	-135,03
8	3	1,71	2,94	-103,11	1,39	93,76	1,71	55,63	0,65	-168,96	0,33	-50,83	0,18	-141,36
9	0	0,36	1,21	112,80	0,36	4,64	0,82	-18,58	0,99	-120,43	0,44	-158,19	0,29	-133,76
10	3	1,16	1,62	-91,68	0,90	74,68	1,16	55,12	0,86	-151,33	0,18	-41,85	0,34	-137,59
11	0	0,99	2,26	-104,97	0,72	125,12	0,99	56,58	0,65	108,82	0,54	31,11	0,38	-130,09
12	3	1,15	1,25	-138,57	0,91	-140,32	1,15	61,38	0,56	-17,54	0,51	-146,03	0,39	-131,47

Tabla A.36: Valores-vectores de nivel sintáctico de las lecturas de nivel intermedio (Lectura I) y avanzado (Lectura A).

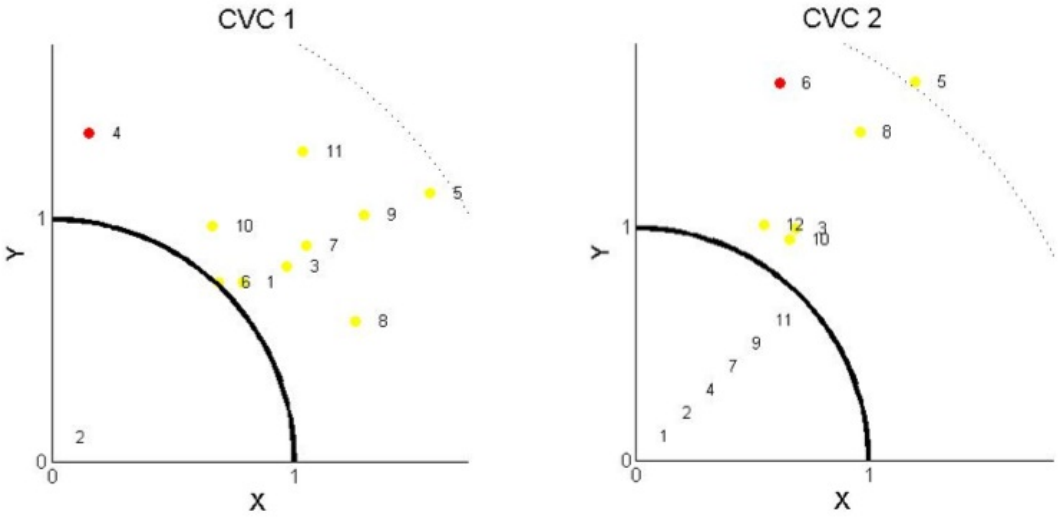


Figura A.5: Visualización del nivel sintáctico de las lecturas del Centro Virtual Cervantes (CVC) de nivel intermedio (CVC 1) y avanzado (CVC 2).

Lectura-I #	A1	A2	B1	B2	C1	C2	Total
1	326	321	1102	346	22	2	2119
2	240	206	736	190	20	6	1398
3	414	322	1228	348	30	5	2347
4	249	238	791	196	22	1	1497
5	299	144	751	192	9	0	1395
6	330	263	961	273	21	0	1848
7	424	276	1092	264	14	0	2070
8	499	326	1150	245	19	7	2246
9	351	387	1323	388	29	1	2479
10	342	169	839	212	17	0	1579
11	255	145	756	197	27	3	1383

Lectura -A #	A1	A2	B1	B2	C1	C2	Total
1	351	427	1139	367	42	2	2328
2	221	101	416	131	6	0	875
3	493	398	1516	401	37	1	2846
4	237	146	550	163	15	0	1111
5	345	411	1609	437	47	3	2852
6	615	625	2018	469	50	6	3783
7	548	261	1074	275	32	2	2192
8	334	424	1468	390	33	4	2653
9	231	114	419	114	12	1	891
10	478	421	1513	398	40	3	2853
11	1084	1028	3818	1213	136	6	7285
12	241	141	662	181	15	1	1241

Tabla A.37: Histograma de estructuras por niveles.

Lectura-I #	NC	A1	A2	B1	B2	C1	C2	Total
1	590	47	140	649	544	119	6	2095
2	402	45	157	439	329	96	14	1482
3	513	62	157	764	577	147	10	2230
4	438	69	149	516	346	87	2	1607
5	227	31	50	453	339	52	0	1152
6	409	45	125	561	479	115	0	1734
7	472	66	108	629	515	77	0	1867
8	683	127	252	845	464	110	14	2495
9	578	69	135	864	624	148	2	2420
10	220	34	72	434	426	64	0	1250
11	260	25	51	395	310	91	6	1138

Lectura-A #	NC	A1	A2	B1	B2	C1	C2	Total
1	686	129	220	847	580	187	5	2654
2	99	28	85	200	253	24	0	689
3	1001	119	283	934	701	168	2	3208
4	203	49	131	336	297	55	0	1071
5	331	55	106	928	736	175	6	2337
6	1569	214	483	1402	836	257	11	4772
7	336	94	227	693	530	119	4	2003
8	809	61	235	907	644	185	8	2849
9	132	23	87	272	207	41	2	764
10	931	117	307	982	714	204	6	3261
11	1430	183	554	2312	1972	569	15	7035
12	161	17	68	333	334	52	2	967

Tabla A.38: Histograma de *PoS* nivelados.

En la tabla A.37 se refleja el histograma de las estructuras sintácticas por niveles de las lecturas de nivel intermedio (I) y avanzado (A).

La reagrupación de los *PoS* se observa en la tabla A.38. Después de aplicar el criterio de máximo nivel, permite renivelar el total de las estructuras de las lecturas de nivel intermedio (I) y avanzado (A) que aparecen en la tabla A.37.

A.5. Campos semánticos del *Plan Curricular del Instituto Cervantes* (*PCIC*)

A.5.1. Nociones generales según el *PCIC*

Se representan con barras en azul oscuro en las gráficas.

1. Nociones existenciales

1.1. existencia, inexistencia

1.2. presencia, ausencia

1.3. disponibilidad, no disponibilidad

1.4. cualidad general 1.5. acontecimiento

1.6. certeza, incertidumbre

1.7. realidad, ficción

1.8. necesidad, contingencia, obligación

1.9. generalidad, especificidad.

2. Nociones cuantitativas

2.1. cantidad numérica

2.2. cantidad relativa

2.3. aumento, disminución

2.4. proporción

2.5. grado

2.6. medidas

2.6.1. medidas generales

2.6.2. talla 2.6.3. tamaño

2.6.4. distancia

2.6.5. velocidad

2.6.6. peso

2.6.7. superficie

2.6.8. volumen, capacidad

2.6.9. temperatura

2.6.10. presión.

3. Nociones espaciales

3.1. localización

3.2. posición absoluta

3.3. posición relativa

3.4. distancia

3.5. movimiento, estabilidad

3.6. orientación, dirección

3.7. orden

3.8. origen.

4. Nociones temporales

4.1. referencias generales

4.2. localización en el tiempo

4.2.1. presente

4.2.2. pasado

4.2.3. futuro

4.3. aspectos de desarrollo

4.3.1. simultaneidad

4.3.2. anterioridad

4.3.3. posterioridad

4.3.4. puntualidad

4.3.5. anticipación

4.3.6. retraso

4.3.7. inicio

4.3.8. finalización

4.3.9. continuación

4.3.10. repetición

4.3.11. duración, transcurso

4.3.12. frecuencia

4.3.13. cambio

4.3.14. permanencia

4.3.15. singularidad.

5. Nociones cualitativas

5.1. formas y figuras

5.2. dimensiones

5.3. consistencia, resistencia

5.4. textura

5.5. acabado

5.6. humedad, sequedad

5.7. materia

5.8. visibilidad, visión

5.9. audibilidad, audición

5.10. sabor

5.11. olor

5.12. color

5.13. edad, vejez

5.14. accesibilidad

5.15. limpieza.

6. Nociones evaluativas

6.1. evaluación general

6.2. valor, precio

6.3. atractivo

6.4. calidad

6.5. aceptabilidad

6.6. adecuación

6.7. conformidad

6.8. corrección

6.9. precisión, claridad

6.10. interés

6.11. éxito, logro

6.12. utilidad, uso

6.13. capacidad, competencia

6.13. importancia

6.15. normalidad

6.16. facilidad.

7. Nociones mentales

7.1 reflexión, conocimiento

7.2. expresión verbal.

A.5.2. Nociones específicas según el *PCIC*

Se representan en distintos colores en la barras de las gráficas.

1. Individuo: dimensión física (color carne)

1.1. partes del cuerpo

1.2. característica físicas

1.3. acciones y posiciones que se realizan con el cuerpo

1.4. ciclo de la vida y reproducción

2. Individuo: dimensión perceptiva y anímica (turquesa oscuro)

2.1. carácter y personalidad

2.2. sentimientos y estados de ánimo

2.3. sensaciones y percepciones físicas

2.4. estados mentales

2.5. modales y comportamiento

2.6. valores personales

2.7. suerte

3. Identidad personal (verde oscuro)

3.1. datos personales

3.1.1. nombre

3.1.2. dirección

3.1.3.

3.1.4. lugar y fecha de nacimiento

3.1.5.

3.1.6. edad

3.1.7. sexo

3.1.8. estado civil

3.1.9. profesión

3.2. documentación

3.3. objetos personales

4. Relaciones personales (verde claro)

4.1. relaciones familiares

4.2. relaciones sociales

4.3. celebraciones y actos familiares, sociales y religiosos

4.4. actitudes y formas de comportarse

5. Alimentación (rojo)

5.1 dieta y nutrición

5.2 bebida

5.3. alimentos

5.4. recetas

5.5. platos

5.6. utensilios de cocina y mesa

5.7. restaurante

6. Educación (rosa)

6.1. centros e instituciones educativas

6.2. profesorado y alumnado

6.3. sistema educativo

6.4. aprendizaje y enseñanza

6.5. exámenes y calificaciones

6.6. estudios y titulaciones

6.7.

6.8. material educativo y mobiliario de aula

7. Trabajo (morado)

7.1. profesiones y cargos

7.2.

7.3. actividad laboral

7.4. desempleo y búsqueda de trabajo

7.5. derechos y obligaciones laborales

7.6. características de un trabajador

8. Ocio (fucsia)

8.1. tiempo libre y entretenimiento

8.2. espectáculos y exposiciones

8.3. deportes

8.4. juegos

9. Información y medios de comunicación (gris claro)

9.1. información y comunicación

9.2. correspondencia escrita

9.3. teléfono

9.4. prensa escrita

9.5. televisión y radio

9.6. internet

10. Vivienda (rojo ladrillo)

10.1. acciones relaciones con la vivienda

10.1.1. construcción

10.1.2. compra y alquiler

10.1.3. ocupación

10.2. características de la vivienda

10.2.1. tipos

10.2.2. partes

10.2.3. personas

10.2.4. condiciones

10.3.

10.3.1. limpieza de la casa

10.3.2. decoración de la casa

11. Servicios (gris oscuro)

11.1. servicio postal

11.2. servicios de transporte

11.3. servicios financieros

11.4. servicios sanitarios

11.5. servicios educativos

11.6. servicios de protección y seguridad

11.7. servicios sociales

11.8. servicios de abastecimiento público

12. Compras, tiendas y establecimientos (turquesa claro)

12.1. lugares, personas y actividades

12.2. ropa, calzado y complementos

12.3. alimentación

12.4. pagos

13. Salud e higiene (azul claro)

13.1. salud y enfermedad

13.2. heridas y traumatismos

13.3. síntomas

13.4. centros de asistencia sanitaria

13.5. medicina y medicamentos

13.6. higiene

13.7. estética

14. Viajes, alojamiento y transporte (oro)

14.1. viajes

14.1.1. objetos y documentos relacionados con los viajes

14.1.2. tipos de viajes

14.2. alojamiento

14.3. sistema de transporte

14.3.1. red de transportes

14.3.2. tipos de transportes

14.3.3. la conducción.

15. Economía e industria (amarillo)

15.1. finanzas y bolsa

15.1.1. economía y dinero

15.1.2. mercado financiero

15.1.3. organismos e instituciones financieros y mercantiles

15.2. renta

15.3. comercio

15.3.1. transacciones comerciales y mercados

15.3.2. comercio exterior

15.3.3. publicidad, mercadotecnia

15.4. entidades y empresas

15.4.1. tipos de empresa y organización

15.4.2. situación de la empresa

15.5. industria y energía

15.5.1. construcción, industria pesada y ligera

15.5.2. sector agropecuario

15.5.3. pesca.

16. Ciencia y tecnología (verde fosforito)

16.1. cuestiones generales

16.2. biología

16.3. matemáticas

16.4. informática y nuevas tecnologías

16.5. física y química.

17. Gobierno, política y sociedad (marrón oscuro)

17.1. sociedad

17.1.1. vida en comunidad

17.1.2. conducta social

17.2. política y gobierno

17.2.1. instituciones políticas y órganos de gobierno

17.3. ley y justicia

17.4. ejército.

18. Actividades artísticas (naranja)

18.1. disciplinas y cualidades artísticas

18.2. música y danza

18.3. arquitectura escultura y pintura

18.4. literatura

18.5. fotografía

18.6. cine y teatro.

19. Religión y filosofía (amarillo claro)

19.1. religión

19.2. filosofía.

20. Geografía y naturaleza (morado)

20.1. universo y espacio

20.2. geografía

20.2.1. geografía física, humana y política

20.2.2. paisaje y accidentes geográficos

20.3. espacios urbanos y rústicos

20.3.1. ciudad

20.3.2. campo

20.4. clima y tiempo atmosférico

20.5. fauna

20.6. flora

20.7. problemas medioambientales y desastres naturales.

A.6. Fragmentos de bases de datos

A.6.1. Fichero de "complejidad_sintaxis.txt"

CC y conjunción-coordinante A2

CC* sino_que locución-conjuntiva B2

CC CS ni-que conjunciónC-conjunciónS B2

CS CC porque-ni conjunciónS-conjunciónC B2

CC P* ni-otro conjunción-pronombre B1

CC D* N* ni-el-calor conjunción-determinante-nombre B1

CC D* P* ni-ningún-otro conjunción-determinante-pronombre B2

CC D* D* NC* ni-ninguna-otra-cosa conjunción-determinante-determinante-nombre B2

CC RG y_así conjunción-coordinante B1

CC* RG sino_que_también locución-conjuntiva-adverbio C1

CC* RG* sino_que-más_aún locución-conjuntiva-adverbio C1

CC Fc y-, conjunción-coma B1

CC* Fc sin_embargo-, locución-conjuntiva-coma B2

CC Fc CC* Fc y-,sin_embargo-, conjunción-coma-conjunción-coma B2

SPS00 NC* CC NC* en-guerra-y-paz preposición-nombre-conjunción-nombre A2

SPS00 NC* CC SPS00 NC* en-guerra-y-en-paz preposición-nombre-conjunción-preposición-nombre A2

SPS00 D* NC* CC D* NC* en-la-guerra-y-la-paz preposición-determinante-nombre-conjunción-determinante-nombre B1

SPS00 D* NC* CC SPS000 D* NC* en-la-guerra-y-en-la-paz preposición-determinante-nombre-conjunción-determinante-nombre B1

A.6.2. Fichero de multivocablos "locuciones_es_SPSXX_pfg.dat"

buena_cantidad_de buena_cantidad_de DI0000_Partitivo I

<buen>_<persona>buena_persona $2:NC I B1

<buen>_<vista>buena_vista $2:NC I B2

<buen>_<manera>buenas_maneras $2:NC I B2

<buen>_<modal>buenos_modales $2:NC I

<buen>_<papel>buen_papel $2:NC I

buen_saque buen_saque NCMS000 I C2

bueno_, bueno RG_XP I

bueno_,_sí bueno_sí RG_AF I

<buzón>_de_<correo>buzón_de_correo $1:NC I

<buzón>_de_correo_electrónico buzón_de_correo_electrónico $1:NC I

<buzón>_de_voz buzón_de_voz $1:NC I B1

cabe_concluir_que cabe_concluir_que CS_CN I

cabello_de_ángel cabello_de_ángel NCMS000 I

<caber>_la_posibilidad caber_la_posibilidad $1:VM I C1

<caber>_duda caber_duda $1:VM I

<caber>_duda_alguna caber_duda_alguna $1:VM I

<caber>_la_menor_duda caber_la_menor_duda $1:VM I

<caber>_la_mínima_duda caber_la_mínima_duda $1:VM I

<caber>_ninguna_duda caber_ninguna_duda $1:VM I

cabeza_de cabeza_de NCF0000_Partitivo I

<cabeza>_de_familia cabeza_de_familia $1:NC I C2

<cabina>_de_ducha cabina_de_ducha $1:NC I B2

cada_año cada_año RG_TP I B1

cada_cosa_a_su_tiempo cada_cosa_a_su_tiempo RG_MD I C2

cada_cosa_por_su_lado cada_cosa_por_su_lado RG_MD I

cada_dos_por_tres cada_dos_por_tres RG_TP I C2

cada_día cada_día RG_TP I B1

A.7. Ejemplo de texto de los alumnos españoles nativos

En la figura A.6 se muestra la copia del texto Penpal_2_16, a modo de modelo, de los textos en formato papel utilizados para esta investigación y escritos por los alumnos que estudian en la Universidad Popular de Alcobendas para la obtención del Graduado en Educación Secundaria Obligatoria. Con la figura A.7 mostramos el mismo texto en formato plano para ser procesado. Finalmente, la figura A.8 es un ejemplo de qué tipo de texto procesado obtenemos después de procesarlo con FreeLing.

A.8. Ejemplo del glosario de frecuencias del Dr. Padró

Mostramos un fragmento del listado de frecuencias del Dr. Padró en su versión original en la figura A.9; otro fragmento del listado de frecuencias del Dr. Padró en su versión adaptada con lemas nivelados se observa en la figura A.10.

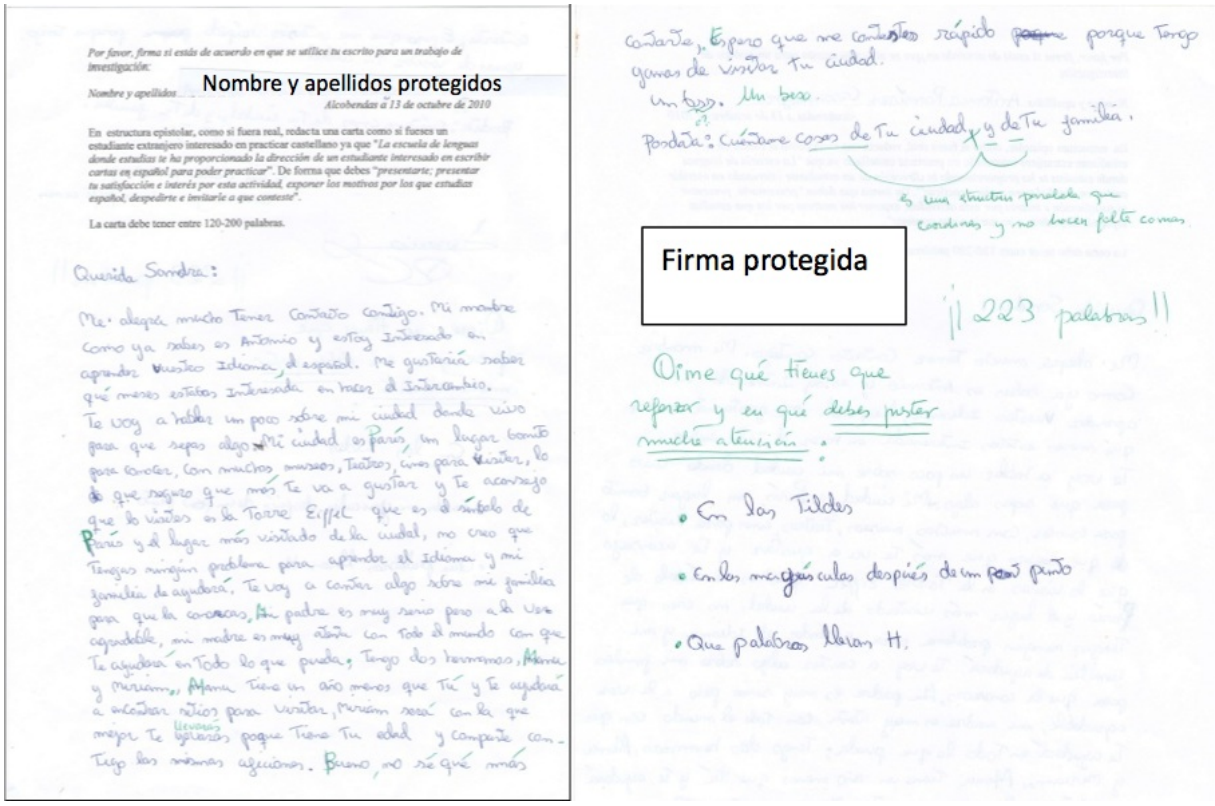


Figura A.6: Texto completo escaneado de la carta de Penpal_2_16 con nombre y firma protegidos.

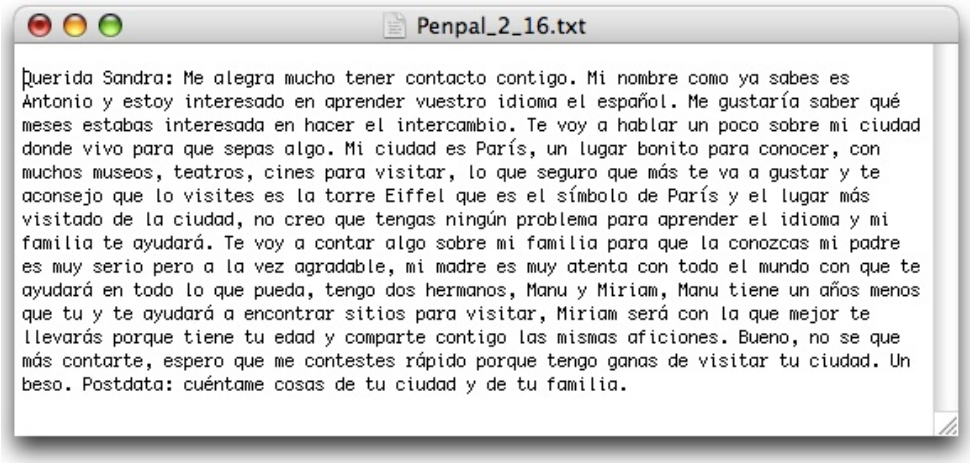


Figura A.7: Ejemplo de texto digitalizado del original.

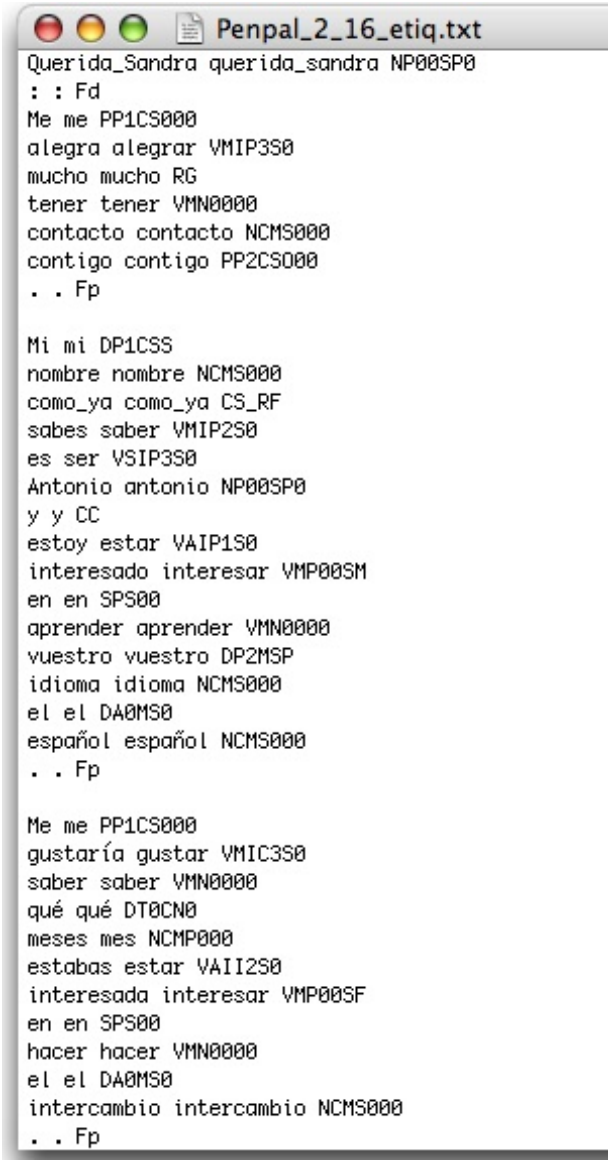


Figura A.8: Fragmento de la carta de Penpal_2_16 después de ser lematizado y etiquetado con los PoS con FreeLing.

```
561485 el
358626 ,
318021 de
220696 .
163550 que
140135 y
125374 a
123185 en
120002 uno
 87666 ser
 86872 se
 56543 su
 56265 no
 53299 haber
 48477 con
 47579 "
 45233 por
 30539 --
 30339 para
 27259 lo
 26643 como
 24616 este
 24279 más
 23136 me
 22651 le
 21589 estar
 21521 todo
 20141 poder
 20086 tener
 19488 pero
 18876 o
 18724 :
 18414 hacer
 16918 decir
 16644 ese
 15045 otro
 14835 -
 13690 ?
 13564 ¿
 13472 si
 13070 él
 12200 sin
 11727 mi
 11703 ir
 11544 ;
 10510 cuando
 10383 )
 10217 (
 10133 ver
 10027 ya
  9525 sobre
  9106 2
  8817 dar
```

Figura A.9: Fragmento del listado de frecuencias y lemas del Dr. Padró en su versión original.

Lema	Frecuencia	Nivel
el	561485	1,00
de	318021	0,00
y	140135	1,00
a	125374	0,00
en	123185	0,00
uno	120002	3,00
ser	87666	1,00
se	86872	1,67
no	56265	3,00
haber	53299	1,50
con	48477	3,00
por	45233	0,00
para	30339	1,00
este	24616	1,67
más	24279	1,00
estar	21589	0,00
todo	21521	3,00
poder	20141	0,00
tener	20086	3,00
o	18876	0,00
hacer	18414	0,00
decir	16918	1,00
ese	16644	1,00
otro	15045	0,00
si	13472	0,00
sin	12200	2,00
mi	11727	3,00
ir	11703	3,00
ver	10133	1,67
ya	10027	0,00
sobre	9525	3,00

Figura A.10: Fragmento del listado de frecuencias y lemas del Dr. Padró en su versión adaptada.

```
#### Splitter options
SplitterFile="/usr/local/share/FreeLing/es/splitter.dat"

#### Morfo options
SuffixAnalysis=yes
MultiwordsDetection=yes
NumbersDetection=yes
PunctuationDetection=yes
DatesDetection=yes
QuantitiesDetection=yes
DictionarySearch=yes
ProbabilityAssignment=yes
NERecognition=yes
DecimalPoint=","
ThousandPoint="."
#LocutionsFile=/usr/local/share/FreeLing/es/locucions.dat
#LocutionsFile=/users/paz/documents/freelingpaz/dicclocesp/locuciones_es_pfg.dat
LocutionsFile=/users/paz/documents/freelingpaz/dicclocesp/locuciones_es_SPSXX_pfg.dat
QuantitiesFile=/usr/local/share/FreeLing/es/quantities.dat
#SuffixFile=/usr/local/share/FreeLing/es/sufixos.dat
SuffixFile=/users/paz/documents/freelingpaz/dicclocesp/sufixos_version2.dat
#SuffixFile=/users/paz/documents/freelingpaz/dicclocesp/sufixos_es_pfg.dat
ProbabilityFile=/usr/local/share/FreeLing/es/probabilitats.dat
DictionaryFile=/usr/local/share/FreeLing/es/maco.db
NPDataFile=/usr/local/share/FreeLing/es/np.dat
PunctuationFile=/usr/local/share/FreeLing/common/punct.dat
ProbabilityThreshold=0.001

## NEC options
NECclassification=yes
NECFilePrefix=/usr/local/share/FreeLing/es/nec/nec

## Sense annotation options (none,all,mfs)
SenseAnnotation=mfs
SenseFile=/usr/local/share/FreeLing/es/senses.db

#### Tagger options (relax,hmm)
Tagger=relax
TaggerHMMFile=/usr/local/share/FreeLing/es/tagger.dat
#TaggerRelaxFile=/usr/local/share/FreeLing/es/constr_gram.dat
TaggerRelaxFile=/users/paz/documents/freelingpaz/dicclocesp/constr_gram_es_pfg.dat
TaggerRelaxMaxIter=500
TaggerRelaxScaleFactor=670.0
TaggerRelaxEpsilon=0.001
TaggerRetokenize=yes

#### Parser options
GrammarFile=/usr/local/share/FreeLing/es/grammar-dep.dat
#GrammarFile=/users/paz/documents/freelingpaz/dicclocesp/grammar-dep_es_pfg.dat

#### Dependence Parser options
HeuristicsFile=/usr/local/share/FreeLing/es/dependences.dat
```

Figura A.11: Fragmento del archivo de gestión de FreeLing 1.5 en su versión adaptada.

A.9. Archivo personalizado para FreeLing 1.5

En la figura A.11 se muestra un fragmento del archivo de configuración que gestiona el funcionamiento de FreeLing 1.5. De esta manera se utilizan archivos personalizados para las locuciones y las estructuras gramaticales. Estos archivos son adaptaciones de los archivos originales de FreeLing 1.5.

Bibliografía

Aarts, J., De Haan, P. y Nelleke, O. (Editores) 1992. *English Language Corpora: Design, Analysis and Exploitation (Language and Computers). Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi.

Adamic, L. A. y Huberman, B. A., 2002. Zipf law and the Internet. *Glottometrics*, 3:143–150. URL `www.hpl.hp.com/research/idl/papers/ranking/adamicglottometrics.pdf`. To honor K.C. Zipf.

Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K., Pociello, E. y Quintian, M., 2006. Improving the Basque WordNet by corpus annotation. En *Proceedings of the first International WordNet Conference*, 22–26.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M. y Soroa, A., 2009a. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. En *Proceedings of NAACL-HLT*.

Agirre, E., Cuadros, M., Rigau, G. y Soroa, A., 2010. Exploring Knowledge Bases for Similarity. En *Proceedings of LREC*.

Agirre, E., López de Lacalle, O. y Soroa, A., 2009b. Knowledge-based WSD and specific domains: performing over supervised WSD. En *Proceedings of IJCAI*.

Agirre, E. y Soroa, A., 2009. Personalizing PageRank for Word Sense Disambiguation. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Alda, J. y Ferrero, P., 2007a. *Algunas cuestiones de ciencia. Libro Homenaje al profesor Manuel Quintanilla*, capítulo Análisis computacional de textos aplicados a una muestra de publicaciones en óptica, 655–667. Prensas Universitarias de Zaragoza.

Alda, J. y Ferrero, P., 2007b. Aplicación de herramientas de lingüística computacional en foros virtuales. En Crespo, A. F.-V., Cesteros, A. F.-P. y Granizo, J. M. (Editores) *Innovación en el Campus virtual: metodologías y herramientas*, 128–138. III Jornada Campus virtual UCM, Universidad Complutense de Madrid.

Alda, J. y Ferrero, P., 2009. Medida de la eficacia del campus virtual. En *Buenas prácticas e indicios de calidad*. V Jornada Campus virtual UCM, Universidad Complutense de Madrid.

Alderson, J. C., 2005. *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. Continuum. URL http://books.google.es/books?id=Y3x-sr7g3aIC&printsec=frontcover&dq=Diagnosing+Foreign+Language+Proficiency:+The+Interface+between+Learning+and+Assessment&source=bl&ots=jZCbvk7Dmm&sig=CKSnMZMAs0egYU8fwHrWfsrj_Zw&hl=es&ei=tyEsTd6DJ8nB4gaD84yNCw&sa=X&oi=book_result&ct=result&resnum=2&ved=0CCMQ6AEwAQ#v=twopage&q&f=false.

Alegria, I., Díaz-de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K., Forcada, M. L., Ortiz-Rojas, S. y Padró, L., 2005. An open architecture for transfer-based machine translation between Spanish and Basque. En *Proceedings of the X Machine Translation Summit workshop OSMaTran: Open-Source Machine Translation X*.

Alonso, L., Castellón, I. y Padró, L., 2002a. Lexicón computacional de marcadores del discurso. *Procesamiento del Lenguaje Natural*, 29:239–246.

Alonso, L., Castellón, I. y Padró, L., 2002b. X-Tractor: A Tool for Discourse Marker Extraction. En *Proceedings of LREC Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*.

Alonso, L., Castellón, I., Padró, L. y Gibert, K., 2002c. Clustering Discourse Markers. En *Proceedings of Congrés Catalá d'Intelligència Artificial, CCIA*.

Alonso, L., Castellón, I., Padró, L. y Gibert, K., 2002d. Discourse Marker characterisation via clustering: extrapolation from supervised to unsupervised corpora. *Procesamiento del Lenguaje Natural*, 29:223–230.

Alonso, L., Castellón, I., Padró, L. y Gibert, K., 2002e. An Empirical Approach to Discourse Markers by Clustering. En *Cinquè Congrés Català d'Intel·ligència Artificial, CCIA'02*.

Association, G. W., 1999. Eurowordnet. URL http://www.illc.uva.nl/EuroWordNet/.

Atienza-Cerezo, E., 1992. *Propuesta de evaluación del texto escrito en Enseñanza Secundaria. Propuesta de evaluación normativa y criterial para el texto expositivo académico*. Tesis Doctoral, Departament de Didàctica de la Llengua I la Literatura de la Universitat de Barcelona.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L. y Padró, M., 2006. Freeling 1.3: Syntatic and Semantic Service in an Open-Source NLP Library. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), ELRA. Genoa, Italy*. URL http://www.lsi.upc.edu/~nlp/freeling.

Atserias, J., Climent, S., Farreres, J., Rigau, G. y Rodríguez, H., 1997. Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. En *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP'97)*. URL http://nlp.lsi.upc.edu/papers-grup/papers/atserias97.pdf.

Attali, Y., Bridgeman, B. y Trapani, C., 2010. Performance of a Generic Approach in Automated Essay Scoring. *Journal of Technology, Learning, and Assessment*, 10 (3). URL `http://escholarship.bc.edu/jtla/vol10/3/`.

Attali, Y. y Burstein, J., 2006. Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment*, 4 (3). URL `http://escholarship.bc.edu/jtla/vol4/3/`.

Baayen, R., 2008. *Analyzing Linguistic Data: A practical Introduction to Statistics using R*. Cambridge University Press.

Baayen, R. H., Wurm, H. L. y Aycock, J., 2007. Lexical dynamics for low-frequency complex words. A regression study across tasks and modalities. *The Mental Lexicon*, 2.3. URL `http://www.ualberta.ca/~baayen/publications.BaayenWurmAycock.pdf`.

Baker, M., 1993. *Text and Technology: In Honour of John Sinclair*, capítulo Corpus Linguistics and Translation Studies. Implications and Applications, 233–250. John Benjamins Publishing Company.

Bakhtin, M., 1998. *Estética de la creación verbal*. Editorial Siglo XXI, 8ª edición.

Bennett, R. E., 2004. Moving the Field Forward: Some Thoughts on Validity and Automated Scoring. Research Memorandum RM-04-01, Educational Text Service (E.T.S), Research Publications Office Mail Stop 7-R ETS Princeton, NJ 08541. URL `http://www.ets.org/Media/Research/pdf/RM-04-01.pdf`.

Biber, D., 1988. *Variation across speech and writing*. Cambridge University Press.

Biber, D., Conrad, S. y Reppen, R., 1998. *Corpus linguistics. Investigating language structure and use*. Cambridge University Press, 1ª edición.

Bloor, T. y Bloor, M., 2004. *The Functional Analysis of English*. Oxford University Press, 2º edición.

Bordón Martínez, T., 2004. *Vademécum para la formación de profesores. Enseñar español como segunda lengua(L2)/lenguas extranjeras(LE)*, capítulo La evaluación de la expresión oral y de la comprensión auditiva, 983. SGEL.

Bordón Martínez, T., 2006. *La evaluación de la lengua en el marco de E/L2: Bases y procedimientos*. Arco-Libros: Cuadernos de didáctica del español/LE, 1ª edición.

Bornkessel, I., Schlesewsky, M., Comrie, B. y Friederici, A. D., 2006. *Semantic Role Universals and Argument Linking: Theoretical, Typological, and Psycholinguistic Perspectives*, capítulo Introduction, 1–13. Berlin: Mouton de Gruyter.

Bosque, I., 1989. *Las categorías gramaticales. Relaciones y diferencias*. Síntesis.

Bosque, I., 2004. *Redes. Diccionario combinatorio del español contemporáneo*. Ediciones SM.

Bosque, I. y Demonte, V., 1999. *Gramática descriptiva de la lengua española*, tomo 3 de *4ª parte: Entre la oración y el discurso*. Espasa Calpe: Colección Lebrija y Bello.

Burstein, J., Chodorow, M. y Leacock, C., 2004. Automated Essay Evaluation: The Criterion Online Writing Service. *AI Magazine*, 25(3):27–36. URL http://www.aaai.org/ojs/index.php/aimagazine/article/view/1774/1672.

Calfee, R., 2000. To grade or not to grade. *IEEE Intelligent Systems*, 15(5):35–37. URL http://www.knowledge-technologies.com/papers/IEEEdebate2000.pdf.

Carpenter, W. J., 2005. *Refiguring Prose Style. Possibilities for Writing Pedagogy*, capítulo Rethinking Stylistic Analysis in the Writing Class, 181–197. Utah University Press.

Carreras, X., Chao, I., Padró, L. y Padró, M., 2004. FreeLing: An Open-Source Suite of Language Analyzers. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. URL http://www.lsi.upc.edu/~nlp/papers/carreras04.pdf.

Cassany, D., 1993. *La cocina de la escritura*. Anagrama: Colección Argumentos.

Cerrolaza-Gili, s., 2005. *Diccionario práctico de gramática. 800 fichas de uso correcto del español*. Grupo Didascalia, Edelsa.

Cervantes, I., 2006. *Plan curricular del Instituto Cervantes. Niveles de referencia para el español. A1, A2, B1, B2, C1, C2*. Biblioteca Nueva. Madrid.

Cervantes, I., 2008a. Diploma de español. Nivel intermedio B2. Prueba 1: Comprensión lectora. Prueba 2: Expresión escrita. URL http://diplomas.cervantes.es/docs/ficheros/200805300001_7_12.pdf.

Cervantes, I., 2008b. Diploma de español. Nivel intermedio B2. Prueba 1: Comprensión lectora. Prueba 2: Expresión escrita. URL http://diplomas.cervantes.es/docs/ficheros/200805300001_7_16.pdf.

Cervantes, I., 2008c. Diploma de español. Nivel Superior C1. Prueba 1: Comprensión de lectura. Prueba 2: Expresión escrita. URL http://diplomas.cervantes.es/docs/ficheros/200805300001_7_20.pdf.

Cervantes, I., 2008d. Diploma de español. Nivel superior C1. Prueba 1: Comprensión de lectura. Prueba 2: Expresión escrita. URL http://diplomas.cervantes.es/docs/ficheros/200805300001_7_24.pdf.

Cervantes, I., 2009a. Diplomas de Español como Lengua Extranjera (DELE). Nivel Intermedio B2. URL http//diplomas.cervantes.es/docs/ficheros/200906180002_7_12.pdf.

Cervantes, I., 2009b. Diplomas de Español como Lengua Extranjera (DELE). Nivel Superior C2. URL http//diplomas.cervantes.es/docs/ficheros/200906180002_7_20.pdf.

Cervantes, I., 2010. Diplomas de Español como Lengua Extranjeras (DELE). Nivel C1. Explicación y ejemplo de examen. URL `http://diplomas.cervantes.es/docs/ficheros/20100729004_7_0.pdf`.

Cervera, A., Hernández, G., Pichardo, C. y Sánchez, J., 2006. *Saber escribir*. Editorial Aguilar.

Chall, J. S. y Dalee, E., 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Inc.

Checa García, I. y Lozano, C., 2002. Los índices de madurez sintáctica de Hunt a la luz de las distintas corrientes generativistas. XVII Encuentro de la Asociación de Jóvenes Lingüistas (AJL). Alicante.

Chomsky, N., 1989. *El conocimiento del lenguaje: su naturaleza, origen y uso*. Alianza Editorial.

Climent Roca, S., 2000. Individuación e información Parte-Todo. Representación para el procesamiento computacional del lenguaje. *Estudios de Lingüística del Español*, 8. URL `http://elies.rediris.es/elies8`.

Collins, A. y Loftus, E., 1975. A spreading activation theory of semantic memory. *Psychological Review*, 82:407–428.

Connolly, J. H. y Dik, S. C. (Editores) 1989. *Functional grammar and the computer*. Foris Publications.

Conrad, D. y Biber, S., 2009. *Register, Genre and Style*. Cambridge University Press.

Cortázar, J., 1986. *Rayuela*. Cátedra, 3ª edición.

Coseriu, E., 1977. *Principios de semántica estructural*. Gredos.

Crossley, S. A., Greenfield, J. y McNamara, D. S., 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42:475–493.

Crossley, S. A. y McNamara, D. S., 2009. Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, 18:119–135. URL `http://csep.psyc.memphis.edu/pdf/crossley_jslw.pdf`.

Crossley, S. A., McNamara, D. S., Weston, J. y McLain Sullivan, S. T., 2010. The development of writing proficiency as a function of grade level: A linguistic analysis. Informe técnico, University of Memphis. URL `http://w-pal.memphis.edu/main/pdf/The_development_of_writing_Proficiency_tech_report.pdf`.

Cruse, D., 1986. *Lexical Semantics*. Cambridge University Press.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U. y James, M., 2006. Analysis of Discourse Features and Verification of Scoring Levels for Independent and Integrated Prototype Written Tasks for the New TOEFL. Informe técnico, Educational Testing Services (ETS). URL `http://www.ets.org/Media/Research/pdf/RR-05-13.pdf`.

da Cunha, I., SanJuan, E., Torres-Moreno, J.-M., Lloberes, M. y Castellón, I., 2010a. Discourse Segmentation for Spanish based on Shallow Parsing. *Lecture Notes in Computer Science*, 6347:13–23. URL http://springerlink.com/content/r402013247484361/.

da Cunha, I., SanJuan, E., Torres-Moreno, J.-M., Lloberes, M. y Castellón, I., 2010b. DiSeg: Un segmentador discursivo automático para el español. *Procesamiento del Lenguaje Natural*, 45:145–152. URL http://www.sepln.org/ojs/ojs2.2/index.php/pln/article/view/776.

da Cunha, I. y Torres-Moreno, J. M., 2010c. Automatic Discourse Segmentation: Review and Perspectives. En *Proceedings of the International Workshop on African Human LanguagesTechnologies*. URL http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/fich_art/paper_iria.pdf.

Daudé, J., 2005. *Enlace de jerarquías usando el etiquetado por relajación*. Tesis Doctoral, Universidad Politécnica de Cataluña.

Daudé, J., Padró, L. y Rigau, G., 2001. A Complete WN1.5 to WN1.6 Mapping. En *WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. NAACL, Pittsburg, PA, USA.

Debowski, L., 2002. Zipf's law against the text size: A half-rational model. *Glottometrics*, 4:49–60.

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Harshman, R. A., Landauer, T. K., Lochbaum, K. E. y Streeter, L. A., 1989. Computer information retrieval using latent semantic structure. URL http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearchbool.html&r=1&f=G&l=50&co1=AND&d=PTXT&s1=4839853.PN.&OS=PN/4839853&RS=PN/4839853.

Díez Orzas, P. L., 1999. La relación de meronimia en los sustantivos del léxico español: Contribución a la semántica computacional. *Estudios de lingüística española*, 2. URL http://elies.rediris.es/elies2.

Dik, C. S., 1980. *Studies in Functional Grammar*. London: Academic Press.

Dik, C. S., 1997. *The Theory of Functional Grammar*, tomo 2 de *Functional Grammar Series 20/21*. Mouton de Gruyter. URL http://books.google.es/books?id=qeMLE_5uvHcC&printsec=frontcover&dq=The+Theory+of+Functional+Grammar&source=bl&ots=yJtZM7wl68&sig=5GZ574xvhTDr5FuGAklaswU_3mU&hl=es&ei=T1FBTaeLKMe5hAef6-30AQ&sa=X&oi=book_result&ct=result&resnum=4&ved=0CDsQ6AEwAw#v=onepage&q&f=false.

Dufty, D. F., McNamara, D., Louwerse, M., Cai, Z. y Graesser, A. C., 2004. Automated Evaluation of Aspects of Document Quality. En *Proceedings of the 22nd Annual International Conference on Design of Communication: The Engineering of Quality Documentation*. Special Interest Group on Design of Documentation (SIGDOC). URL http://csep.psyc.memphis.edu/pdf/p142-dufty.pdf.

Dumais, S. T., 1993. LSI meets TREC: A status report. En Harman, D. (Editor) *The First Text REtrieval Conference (TREC1). Special Publication 500-207*, 137–152. National Institute of Standards and Technology.

Dunteman, G. H., 1989. *Principal Component Analysis.* Número 69 en Quantitative Applications in the Social Sciences. Sage Publications, Inc.

Enbar, N., 1999. This is e-rater. It'll be scoring your essay today. URL `http://www.businessweek.com/bwdaily/dnflash/jan1999/nf90121d.htm`. Retrieved January, 2010.

España-Bonet, C., Vila, M., Martí, M. A. y Rodríguez, H., 2009. CoCo, a Web Interface for Corpora Compilation. *Procesamiento del Lenguaje Natural*, 367–368.

de Europa, C., 2005. *Reference Level Descriptions for National and Regional Languages (RDL), Draft Guide for the production of RDL (Version 2).* Estrasburgo, Language Policy Division.

Faber, P. y Mairal, R., 1998. *The Structure of the Lexicon in Functional Grammar*, capítulo Methodological Criteria for the Elaboration of a Functional Lexicon-Based Grammar of the Semantic Domain of Cognitive Verbs, 3–24. Studies in Language Companion. Amsterdam / Philadelphia: John Benjamins.

Fernández-Montraveta, A., Vázquez, G. y Fellbaum, C., 2008. *Text Resources and Lexical Knowledge. Selected Papers from the 9th Conference on Natural Language Processing. KONVENS*, capítulo The Spanish Version of WordNet 3.0, 175–182. Berlin / New York: Mouton de Gruyter. URL `http://grial.uab.es/archivos/contribution1.pdf`.

Fernández-Pampillón, A., 2010. *La construcción de tesauros académicos. Un modelo general y método inductivo con aplicación al e-learning.* Tesis Doctoral, Universidad Complutense de Madrid.

Fernández-Ramírez, S., 2009. Archivo gramatical de la lengua española. Categorías gramaticales. Centro Virtual Cervantes. URL `http://cvc.cervantes.es/obref/agle/categorias_gramaticales/`.

Ferrero, A., Alda, J., Campos, J., López-Alonso, J. M. y Pons, A., 2007. Principal Component Analysis of the Photo-Response Non Uniformity of a Matrix Detecto. *Applied Optics*, 46:9–17.

Ferrero, P., 2011. *Entornos virtuales y modelos de aprendizaje: Espacios para la aplicación de herramientas y métodos de análisis lingüístico-computacionales.* Proyecto Fin de Carrera, Universidad Autónoma de Madrid.

Ferrero, P. y Alda, J., 2005. La tutorización virtual y la expresión de las emociones. En Crespo, A. F.-V., Cesteros, A. F.-P. y Granizo, J. M. (Editores) *Cómo integrar investigación y docencia en el CV-UCM*, 129–133.

Fischer, K. (Editor) 2006. *Approaches to Discourse Particles.* Studies in Pragmatics 1. Oxford: Elsevier.

Flower, L. y Hayes, R. J., 1981. A Cognitive Process of Writing. *College Composition and Communication*, 32(4):365–387.

Foltz, P., Gilliam, S. y Kendall, S., 2000. Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments*, 8 (2):111–129. URL http://www.knowledge-technologies.com/SupContBas2000.pdf.

Gass, S. M. y Selinker, L. (Editores) 2008. *Second Language Acquisition. An Introductory Course*. New York: Routledge, 3rd edición.

Goodfellow, R., Jones, G. y Lamy, M.-N., 2002. Assessing Learners Writing using Lexical Frequency. *ReCALL*, 14 (1):133–145. URL http://iet-staff.open.ac.uk/r.goodfellow/ReCall01%282%29.htm.

Gorman, S., 2010. NAEP 2011 Writing Computer-Based Assessment (CBA) at grades 8 and 12. Brochure (3 hojas) by National Center for Education Statistics (NCES).

Graesser, A., 2008. Advances in text comprehension: Commentary and final perspective. *Applied Cognitive Psychology*, 22:425–429. URL http://www.autotutor.org/publications/pubs.htm.

Graesser, A. C., McNamara, D. S. y Louwerse, M. M., 2010. *Handbook of Reading Research*, tomo IV, capítulo Methods of automated text analysis. Mahwah, NJ: Routledge / Erlbaum. URL http://sites.google.com/site/graesserart/files/Methods-of-automated-text-analysis.pdf?attredirects=0.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. y Cai, Z., 2004. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, and Computers*, 36:193–202. URL http://cohmetrix.memphis.edu/CohMetrixWeb2/HelpFile2.htm.

Graesser, A. C. y Petschonek, S., 2005. *Advancing health outcomes research methods and clinical applications*, capítulo Automated Systems that Analyze Text and Discourse: QUAID, Coh-Metrix, and AutoTutor. International Society for Quality of Life Research. Degnon Associates. URL http://www.memphis.edu/psychology/graesser/publications/documents/QOLgraesser040305.doc.

Griffiths, T. L., Lucas, C., Williams, J. J. y Kalish, M. L., 2009. Modeling human function learning with Gaussian processes. *Advances in Neural Information Processing Systems 21*, 21. URL http://cocosci.berkeley.edu/tom/papers/funclearn1.pdf.

Griffiths, T. L. y Steyvers, M., 2002. A Probabilistic Approach to Semantic Representation. En *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 381–386. URL http://cocosci.berkeley.edu/tom/papers/semrep.pdf.

Grimes, D. y Warschauer, M., 2010. Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *Journal of Technology, Learning, and Assessment*, 8 (6). URL http://www.jtla.org.

Group, P. K. T., 2007. Evidence for Reliability, Validity, and Learning Effectiveness. Informe técnico, Pearson Education, Inc. URL `http://www.knowledge-technologies.com/papers/WTLReliabilityValidity-082007.pdf`.

Ha, L. Q., Sicilia-García, E. I., Ming, J. y Smit, F. J., 2002. Extension of Zipf's Law to Words and Phrases. En *Proceedings of the 19th International Conference on Computational Linguistics*, tomo 1, 1–6. Taipei, Taiwan.

Hall, C., Lewis, G., McCarthy, P., Lee, D. y McNamara, D., 2006. Using Coh-Metrix to Assess Differences between American and English/Welsh. En Sun, R. y Miyake, N. (Editores) *Proceedings of the 28th Annual Conference of the Cognitive Science Scociety*, 2498.

Halliday, M. A., 1985. *An Introduction to Functional Grammar*. London: Edward Arnold.

Halliday, M. A. K., 1970. *Functional Sentence Perspective*, capítulo Language structure and language function, 11–37. Mouton.

Halliday, M. A. K., 1974. *System of Linguistic Description*, capítulo The Place of Functional Sentence Perspective, 43–53. Mouton.

Heinzle, J. y Haynes, J.-D., 2009. Multivariate functional connectivity between fine-grained cortical activation patterns. En Neuroscience, B. (Editor) *Eighteenth Annual Computational Neuroscience Meeting: CNS*2009, Berlin, Germany*, tomo 10(1), 76.

Hirschman, L., Breck, E., Light, M., Burger, J. y Ferro, L., 2000. Automated Grading of Short Answer Tests. *IEEE Intelligent Systems*, 31-35:15(5). URL `http://www.knowledge-technologies.com/papers/IEEEdebate2000.pdf`.

Hunt, K. W., 1965. Gramatical Structures Written at Three Grade Levels. *Research Report, National Council of Teachers of English (NCTE)*.

Hunt, K. W., 1977. *Evaluating Writing: Discribing, Measuring, Judging*, capítulo Early Blooming and Late Blooming Syntactic Structures, 99–104. National Council of Teachers of English (NCTE).

Jackendoff, R. S., 1990. *Semantic Structures (Current Studies in Linguistics)*. The MIT Press. URL `http://www.amazon.com/Semantic-Structures-Current-Studies-Linguistics/dp/026260020X#reader_026260020X`.

Jackendoff, R. S., 1992. *Languages of the mind. Essays on Mental Representation*. The MIT Press. URL `http://www.amazon.com/Languages-Mind-Essays-Mental-Representation/dp/0262600242#reader_0262600242`.

Jenvald, J., Morin, M. y Kincaid, P., 2001. A Framework for Web-based Dissemination of Models and Lessons Learned from Emergency-Response, Exercises, and Operations. *International Journal of Emergency Management (IJEM)*, 1, No.1:82–94.

Jorge-Botana, G., León, J., Olmos, R. y Hassan-Montero, Y., 2010. Visualizing polysemy using LSA and the predication algorithm. *Journal of the American Society for Information Science and Technology*, 61, nº 8. URL http://www.elsemantico.com/Documentos/visualizing.pdf.

Jorge-Botana, G., Olmos, R. y A, L. J., 2007. Análisis de la Semántica Latente (LSA) y estimación automática de las intenciones del usuario en diálogos de telefonía (call routing). *Revista FAZ*, 1:53–66.

Kintsch, W., 2002. The Potential of Latent Semantic Analysis for Machine Grading of Clinical Case Summaries. *Journal of Biomedical Informatics*, 35(1):3–7.

Klare, G. R., 1974-1975. Assessing Readability. *Reading Research Quarterly*, 10:62–102.

Konchady, M., 2006. *Text Mining Application Programming*. Charles River Media.

Kontostathis, A. y Pottenger, W., 2006. A framework for understanding LSI performance. *Information Processing and Management*, 42 (1):56–73.

Krippendorff, K., 2003. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, 2ª edición.

Krippendorff, K. y Bock, A. (Editores) 2009. *The Content Analysis Reader*. Sage Publications. URL http://www.amazon.com/Content-Analysis-Reader-Klaus-Krippendorff/dp/1412949661#reader_1412949661.

Kukich, K., 2000. Beyond Automated Essay Scoring. *IEEE Intelligent Systems*, 15(5):22–27. URL http://www.knowledge-technologies.com/papers/IEEEdebate2000.pdf.

Landauer, T., Lochbaum, K. E. y Dooley, S., 2009. A new formative assessment technology for reading and writing. *Theory Into Practice*, 48:44–52. URL http://www.knowledge-technologies.com/papers/ab-TIP48.shtml.

Landauer, T. K. y Dumais, S. T., 1997. A solution to Plato's problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

Landauer, T. K., Foltz, P. W. y Laham, D., 1998a. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284. URL http://www.knowledge-technologies.com/papers/IntroLSA1998.pdf.

Landauer, T. K., Laham, D. y Derr, M., 2004. From Paragraph to Graph: Latent Semantic Analysis for Information Visualization. En *Proceedings of the National Academy of Science*, tomo 101, 5214–5219. URL http://www.knowledge-technologies.com/papers/pnas.0400341101.pdf.

Landauer, T. K., Laham, D. y Foltz, P. W., 1998b. *Advances in Neural Information Processing Systems*, capítulo Learning Human-like Knowledge by Singular Value Decomposition: A progress report, 45–51. 10. The MIT Press. URL http://www.knowledge-technologies.com/papers/nips1998.pdf.

Landauer, T. K., Laham, D. y Foltz, P. W., 2000. The Intelligent Essay Assessor. *IEEE Intelligent Systems*, 15(5):27–31. URL `http://www.knowledge-technologies.com/papers/IEEEdebate2000.pdf`.

Landauer, T. K., Lahm, D. y Flotz, P., 2003. Automatic Essay Assessment. *Assessments in Education*, 10, Nº3:295–308. URL `http://www.pearsonkt.com/papers/ab-CAIE_10_3_04lores.shtml`.

Landauer, T. K., S., M. D., Dennis, S. y Kintsch, W. (Editores) 2007. *Handbook of Latent Semantic Analysis*. University of Colorado. Institute of Cognitive Science.

Laufer, B. y Nation, P., 1995. Vocabulary size and use: Lexical richness in L2 written productions. *Applied Linguistics*, 16(3):307–322.

Li, W., 1992. Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

Lin, D. y Pantel, P., 2001a. DIRT@SBT@ Discovery of inference rules from text. En *Proceedings of KDD-2001. The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 323–328. Association for Computing Machinery (ACM). Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD).

Lin, D. y Pantel, P., 2001b. Induction of semantic classes from natural language text. En *Proceedings of KDD-2001. The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 317–322. Association for Computing Machinery (ACM). Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD).

Louwerse, M. M., 2001. *Multidisciplinary Approaches to Discourse*, capítulo Context in Causal and Diagnostic Readings: Cognitive Evidence from Eye Tracking, 11–26. Amsterdam-Muenster,Uitgaven Stichting Neerlandistiek VU, Nodus.

Louwerse, M. M., 2004. Un modelo conciso de cohesión en el texto y coherencia en la comprensión. *Revista Signos*, 37, 56:41–58.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S. y Graesser, A. C., 2004. Variation in Language and Cohesion across Written and Spoken Registers. En K.D. Forbus, T. R. E., D. Gentner (Editor) *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, 843–848. Lawrence Erlbaum Publishers, Mahwah, NJ.

López Alonso, C., 2006. El texto electrónico en el aprendizaje de las lenguas. *Estudios de Lingüística del Español*, 23. URL `http://elies.rediris.es/elies23/lopez.htm`.

López-Alonso, J. M. y Alda, J., 2002. Bad Pixel Identification by Means of the Principal Component Analysis. *Optical Engineering*, 41:2152–2157.

López-Alonso, J. M. y Alda, J., 2004. Operational Parametrization of the 1/f Noise of a Sequence of Frames by Means of the Principal Components Analysis in Focal Plane Arrays. *Optical Engineering*, 42:257–265.

López-Alonso, J. M., Alda, J. y Bernabéu, E., 2002. Principal Component Characterization of Noise for Infrared Images. *Applied Optics*, 41:320–331.

López-Alonso, J. M., Rico-García, J. M. y Alda, J., 2004. Photonic Crystal Characterization by FDTD and Principal Component Analysis. *Optics Express*, 12:2176–2186.

López Martín, F., 1999. *El vocabulario básico de orientación didáctica.* Tesis Doctoral, Facultad de Letras. Departamento de Lengua Española y Lingüística General de Murcia.

Lucas Cuesta, J. M., Fernández Martínez, F., Ferreiros López, J., López Ludeña, V. y San Segundo Hernández, R., 2010. Clustering of Syntactic and Discursive Information for the Dynamic Adaptation of Language Models. *Procesamiento de Lenguaje Natural*, 45:175–182. URL `http://www.sepln.org/ojs/ojs-2.2/index.php/pln/article/view/790`.

Lyons, J., 1980. *Semántica.* Teide.

Martín Úriz, A., Whittaker, R., Chaudron, C., Barrio Luis, M., Hidalgo Downing, L., Blanco Paetsch, S. y Ordóñez de Celis, L., 2005. *La composición como comunicación: una experiencia en las aulas de lengua inglesa en bachillerato.* Ediciones de la Universidad Autónoma de Madrid, 1ª edición.

Martín Úriz, A. M., Hidalgo, L. y Whittaker, R., 2001. Desarrollo y complejidad de la frase nominal en composiciones de estudiantes. En *XIX Congreso de AESLA.*

Martín Zorraquino, M. A., 1998. *Los marcadores del discurso teoría y análisis.* Arco-Libros.

Martyniuk, W. (Editor) 2010. *Aligning Test with the CEFR. Reflections on using the Council of Europe draft Manual.* Cambridge University Press.

McCarthy, P. M., 2005. *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical diversity (MTLD).* Tesis Doctoral, University of Memphis. URL `http://gateway.proquest.com/openurl%3furl_ver=Z39.88-2004%26res_dat=xri:pqdiss%26rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation%26rft_dat=xri:pqdiss:3199485`.

McNamara, D. S., Louwerse, M. M., Cai, Z. y Graesser, A., 2006. Coh-Metrix 2.0. URL `http://cohmetrix.memphis.edu/CohMetrixDemo/demo.htm`.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M. y Graesser, A. C., 2010. Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes*, 47:292–330.

McNamara, L. M. M. G. C. A., S. Danielle, 2002-2005. Coh-Metrix: Automated Cohesion and Coherence Scores to Predict Text Readability and Facilitate Comprehension. Project funded by the office of educational research and improvement reading program, University of Memphis.

de Miguel, E. (Editor) 2009. *Panorama de la lexicología.* Editorial Ariel.

Miller, G., 1962. Some Psychological Studies of Grammar. *American Psychologist*, 17(11):748–762.

Miller, G., 1986. Dictionaries in the Mind. *Language and Cognitive Processes*, 1(30):171–185.

Miller, G., 1990. Wordnet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4).

Miller, G., 1991. Semantic Networks of English. *Cognition*, 41:197–229.

Miller, G. A. y McKean, K. O., 1964. A Chronometric Study of Some Relations between Sentences. *Quarterly Journal of Experimental Psychology*, 16:297–308.

Mislevy, R. J., Behrens, J. T., Bennett, R. E., Demark, S. F., Frezzo, D. C., Levy, R., Robinson, D. H., Rutstein, D. W., Shute, V. J., Stanley, K. y Winters, F. I., 2010. On the Roles of External Knowledge Representations in Assessment Design. *Journal of Technology, Learning, and Assessment*, 8(2). URL `http://www.jtla.org`.

Moliner, A. M., 2007. *Diccionario del uso del español*, tomo 2. Gredos, 3ª edición.

Olloqui de Montenegro, L., 1991. *La enseñanza de español como lengua materna*, capítulo La investigación de la madurez sintáctica y la enseñanza de la lengua materna, 113–132. Universidad de Puerto Rico. URL `http://books.google.es/books?id=7pdU-ht5N-4C&printsec=frontcover#v=onepage&q&f=false`.

Moreno Ortiz, A., 1998. El lexicón en la lexicografía computacional: adquisición y representación de la información léxica. *Alfinge: Revista de Filología*, 10:249–272.

Moreno Ortiz, A., 2000. Diseño e implementación de un lexicón computacional para lexicografía y traducción automática. *Estudios de Lingüística del Español*, 9. URL `http://elies.rediris.es/elies9/index.htm`.

Morris, L. y Cobb, T., 2004. Analysis of TESL and TEFL trainees grammatical and lexical knowledge. *System*, 35:75–87.

Nation, I. S. P., 1993. Using dictionaries to estimate vocabulary size: essential, but rarely followed, procedures. *Language Testing*, 10(1):27–40.

Nation, I. S. P., 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, P. y Kyongho, H., 1995. Where Would General Service Vocabulary Stop and Special Purposes Vocabulary Begin? *System 1995*, 23(1):35–41. URL `www.victoria.ac.nz/lals/staff/publications/paul-nation/1995-Hwang-Special-purposes.pdf`.

O'Donnell, M., Murcia, S., García, R., Molina, C., Rollinson, P., MacDonald, P., Stuart, K. y Boquera, M., 2009. Exploring the proficiency of English learners: The TREACLE project. En *Proceedings of the Fifth Corpus Linguistics Conference*. University of Liverpool.

Olmos, R., León, J., Escudero, I. y Jorge-Botana, G., 2009. Análisis del tamaño y especificidad de los corpus en la evaluación de resúmenes mediante el LSA. Un análisis comparativo entre LSA y jueces expertos. 42:69. URL http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342009000100004&lng=en&nrm=iso&ignore=.html.

Oxford, U., 2010. Oxford TM 3000 Text Checker. URL http://www.oup.com/elt/catalogue/teachersites/oald7/oxford_3000/oxford_3000_profiler?cc=gb.

Padró, L., 2006. *FreeLing User Manual 1.5.*

Padró, L., 2009. *FreeLing User Manual 2.1.*

Padró, L., 2011a. Página web del Dr. Padró. URL http://www.lsi.upc.es/~padro/index.php?page=nlp.

Padró, L., 2011b. Several frequency counts of word forms, lemmas, and PoS from a 5.5 milion word Spanish corpus of unrestricted text. URL http://www.lsi.upc.es/~padro/index.php?page=nlp.

Padró, L., Collado, M., Reese, S., Lloberes, M. y Castellón, I., 2010a. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. En *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10)*. La Valletta, Malta. URL http://www.lsi.upc.edu/~nlp/papers/padro10b.pdf.

Padró, L., Reese, S., Agirre, E. y Soroa, A., 2010b. *Principles, Construction, and Application of Multilingual Wordnets: Proceedings of the 5th Global Wordnet Conference*, capítulo Semantic Services in FreeLing 2.1: WordNet and UKB, 99–105. Narosa. URL http://www.lsi.upc.edu/~nlp/papers/padro10a.pdf.

Page, E. B., 1966. Grading Essays by Computer: Progress Report in Notes. En *Testing Problems*, 87–100.

Page, E. B., 1994. Computer Grading of Student Prose, Using Modern Concepts and Software. *Journal of Experimental Education*, 62(2):127–142.

Page, E. B., Poggio, J. P. y Keith, T. Z., 1997. Computer Analysis of Student Essays: Finding Trait Differences in the Student Profile. *Annual Meeting of the American Educational Research Association*, 8. URL http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED411316.

Paredes Chavarría, E. A., 2006. *Prontuario de lectura, lingüística, redacción, comunicación oral y nociones de literatura*. Editorial Limusa, 2ª edición. URL http://books.google.es/books?id=eKLCl2GkMr4C&pg=PA161.

Peramos Soler, N., Leontaridi, E. y Ruiz Morales, M., 2010. Las unidades fraseológicas del español: su enseñanza y adquisición en la clase de ELE. URL http://www.ispania.gr/arthra/ispanika/1518-las-unidades-fraseologicas-del-espanol-su-ensenanza-y-adquisicion-en-la-clase-de

Pfau, R., 2000. *Features and Categories in Language Production*. Tesis Doctoral, Johann Wolfgang Goethe Universität. URL `http://dingo.sbs.arizona.edu/~hharley/courses/535/PfauThesis.pdf`.

Pons Bordería, S., 2006. *Approaches to Discourse Particles*, capítulo A Functional Approach for the Study of Discourse Markers, 77–99. Elsevier.

Pérez Hernández, C., 1994. *Corpus-based Bilingual Lexicography: The Use of Computerized Corpora for the Identification of Translation Equivalents between English and Spanish*. Tesis Doctoral, University of Exeter.

Pérez Marín, D. R., 2007. *Adaptive Computer Assisted Assessment of Free-Text Students Answers: An Approach to Automatically Generate Students Conceptual Models*. Proyecto Fin de Carrera, Escuela Politécnica Superior. Universidad Autónoma de Madrid. URL `http://www.ii.uam.es/esp/alumnos/egresados/web/d_perez.html`.

Rodríguez Fonseca, L., 1999. *Estudios de lingüística hispánica: homenaje a María Vaquero*, capítulo ¿Qué nos dicen y qué no os dicen los índices de madurez sintáctica?, 523–535. Universidad de Puerto Rico. URL `http://books.google.es/books?id=O7uh8WbWj1oC&pg=PA523&lpg=PA523&dq=unidad-T+hunt&source=bl&ots=RRNqZ6zt0Q&sig=y8FLcH_3eN2yLMAIPZfz809nQxs&hl=es&ei=DNirTeKKOMKi8QPh4pG5Ag&sa=X&oi=book_result&ct=result&resnum=2&ved=0CBoQ6AEwAQ#v=onepage&q=unidad-T%20hunt&f=false`.

Romero Trillo, J., 2007. Adaptive Management in Discourse: the case of involvement discourse markers in English and Spanish conversations. *Catalan Journal of Linguistics*, 6:81–94. URL `http://www.raco.cat/index.php/catalanjournal/article/view/74211`.

Ruiz Gurillo, L., 1997. Relaciones categoriales de las locuciones adverbiales. *Contextos*, 15(29-30):19–31.

Sanborn, A. N., Griffiths, T. L. y Navarro, D. J., 2011. Rational Approximations to Rational Models: Alternative Algorithms for Category Learning. *Psychological Review*, in press. URL `http://cocosci.berkeley.edu/tom/papers/approx-pdf`.

Sigley, R., 1997. Text Categories and Where You Can Stick Them: A Crude Formality Index. *International Journal of Corpus Linguistics*, 2(2):199–237.

Simón Granda, J., 1992. Evaluación automática de la dificultad de lectura: hacia un índice de legibilidad flexible y fiable. En Losada Durán, J. R. y Mansilla García, M. (Editores) *Actas del VIII Congreso de la Asociación Española de Lingüística Aplicada (AESLA)*, 643–56. Universidad de Vigo.

Smith III, M., 2009. The Reading-Writing Connection. URL `http://www.lexile.com/m/resources/materials/Reading-Writing_Connection.pdf`. President of MetaMetrics, Inc.

Sánchez, A., Sarmiento, R., Cantos, P. y José, S., 1995. *Cumbre. Corpus lingüístico del español contemporáneo: Fundamentos, metodología y aplicaciones de los corpus lingüísticos*. SGEL.

Sánchez Lobato, J., 2006. *Saber escribir*. Editorial Aguilar.

Sánchez Lobato, J. y Santos Gargallo, I., 2004. *Vademécum para la formación de profesores. Enseñar español como segunda lengua(L2)/lenguas extranjeras(LE)*. Sociedad General Española de Librería (SGEL).

Streeter, L., Psotka, J., Laham, D. y MacCuish, D., 2002. The credible grading machine: Automated essay scoring in the DoD. En *The Interservice Industry Training, Simulation & Education Conference (I/ITSEC)*. URL `http://www.knowledge-technologies.com/papers/CredGrading2002.pdf`.

Thomas, M., Jaffe, G. J., Kincaid, P. J. y Stees, Y., 1992. Learning to Use Simplified English: A Preliminary Study. *Technical Communication*, 69–73.

Tonta, Y. y Darvish, H. R., 2010. Diffusion of Latent Semantic Analysis as a Research Tool: A Social Network Analysis Approach. *Journal of Informetrics*, 4(2):166–174. URL `http://hacettepe.academia.edu/Ya%C5%9FarTonta/Papers/197666/Diffusion_of_Latent_Semantic_Analysis_as_a_Research_Tool_A_Social_Network_Analysis_Approach`.

Torres González, A. N., 1993. *Madurez sintáctica en estudiantes no universitarios de la zona metropolitana de Tenerife*. Tesis Doctoral, Universidad de La Laguna. Tenerife.

Venegas, R., 2006. La similitud léxico-semántica en artículos de investigación científica en español: una aproximación desde el análisis semántico latente. *Revista Signos*, 39(60):75–106. URL `http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342006000100004&lng=es&nrm=iso`.

Venegas, R., 2009. Toward a Method for Assessing Summaries in Spanish Using LSA. En *Proceedings of the Twenty-Second International FLAIRS Conference*, 310–311.

Véliz, M., 1988. Evaluación de la madurez sintáctica en el discurso escrito. *Revista de lingüística teórica y aplicada (RLA)*, 26:105–140.

Véliz, M., 1999. Complejidad sintáctica y modo del discurso. *Estudios filológicos*, 34:181–192. URL `http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0071-17131999003400013&lng=es&nrm=iso`.

Véliz, M., 2004. Procesamiento de estructuras sintácticas complejas en adultos mayores y adultos jóvenes. *Estudios filológicos*, 39:65–81. URL `http://www.scielo.cl/scielo.php?pid=S0071-17132004003900004&script=sci_arttext`.

Voorhees, E. M., Harman, D. K., Buckley, C., Robertson, S., Callan, J., Dumais, S. T., Belkin, N. J., Hawking, D. y et alii, 2005. *TREC. Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. The MIT Press. URL `http://trec.nist.gov`.

Vossen, P., 1994. A Functional Approach to the Grammatical and Conceptual Individuation of First-order Nouns. En *Proceedings of the Functional Grammar Conference. York*.

Wolfe-Quintero, K., Inagaki, S. y Kim, H.-Y., 1998. *Second language development in writing: measure of fluency, accuracy and complexity*. Second Language Teaching and Curriculum Center. University of HawaiŽi.

Zaanen, M. v., Robert, A., Atwell, E., Woodhouse, L. y Jt, L., 2004. *Proceedings of the Workshop: The Amazing Utility of Parallel and Comparable Corpora*, capítulo A Multilingual Parallel Parsed Corpus as Gold Standard for Grammatical Inference Evaluation, 58–61.

Ziempekis, D. y Gallopoulos, E., 2006. *Grouping Multidimensional Data: Recent Advances in Clustering*, capítulo Tmg: A Matlab Toolbox for Generating Term-Document Matrices from Text Collections, 187–210. Springer.

Zipf, G., 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA.: Harvard University Press.

Zipf, G. K., 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.

Índice de figuras

Índice de tablas

Este texto se ha compuesto utilizando el sistema de procesado de textos científicos LaTeXen su versión pdfLaTeX. Los editores utilizados han sido LyX y WindEdt 5.4. La mayor parte de las figuras han sido generadas mediante el entorno de cálculo MatLab.

La actual versión se terminó de componer e imprimir
en Alcobendas el sábado, 30 de abril de 2011.