FACULTAD DE FORMACIÓN DE PROFESORADO Y EDUCACIÓN

PROGRAMA DE DOCTORADO EN EDUCACIÓN

LINGÜÍSTICA COMPUTACIONAL APLICADA A LA INVESTIGACIÓN EDUCATIVA:

UN ENFOQUE MATEMÁTICO DE LA ENSEÑANZA DE VOCABULARIO EN LENGUA INGLESA PARA HISPANOHABLANTES

TESIS DOCTORAL

Presentada para obtener el Grado de Doctor por

Diana Cembreros Castaño

Directora de Tesis

Dra. María Dolores Ramírez Verdugo

Madrid, mayo 2014

Diana Cembreros Castaño **Dra. María Dolores Ramírez-Verdugo**

FACULTY OF TEACHER TRAINING AND EDUCATION

DOCTORAL PROGRAMME ON EDUCATION

COMPUTATIONAL LINGUISTICS APPLIED TO EDUCATIONAL RESEARCH:
A NEW APPROACH IN ENGLISH VOCABULARY TEACHING FOR SPANISH SPEAKERS

DISSERTATION THESIS

Presented in Fulfilment of the Requirements for the Degree
Doctor of Philosophy

Diana Cembreros Castaño

Dissertation supervisor

Dr. María Dolores Ramírez Verdugo

Madrid, mayo 2014

Diana Cembreros Castaño **Dr. María Dolores Ramírez-Verdugo**

To my parents,

who never failed to give me support.

AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento por su dedicación, seguimiento constante y paciencia a mi directora de Tesis, Mª Dolores Ramírez Verdugo, por haber sido a lo largo de esta aventura una fuente de inspiración académica y un gran apoyo, tanto a nivel personal como profesional.

Debo recordar también a Paloma Tejada Caller por su gran ayuda en uno de los momentos más difíciles de este proyecto y por haber tutelado mis primeros pasos en la investigación académica en el Máster en Lingüística Inglesa de la Universidad Complutense, que fue el germen de esta tesis doctoral.

Agradezco a Francisco Javier Murillo Torrecilla, director del Programa de Doctorado en Educación de la UAM, su inestimable labor como tutor a lo largo de todo el proceso. De igual manera, me gustaría reconocer a la directora del Departamento de Filologías y su Didáctica, Mª Victoria Sotomayor Sáez, quien tantas veces ha compartido conmigo de manera generosa su tiempo, conocimientos y experiencia.

Quiero agradecer a la Universidad Autónoma de Madrid y a la Universidad Camilo José Cela, que confiaron en mí como docente e investigadora y me inspiraron a seguir por este emocionante camino. Agradezco el apoyo de tantos compañeros, y quiero reconocer especialmente a Elena, por transmitirme cada día su optimismo y su pasión por la Universidad; a Marta, por todos sus valiosos consejos; a Alicia, a quien debo tanto; y a Pedro, por todo.

A mis padres, gracias por escucharme pacientemente hablar de *false cognates* y series armónicas, por las innumerables llamadas, por leer y releer, y por alentarme cada día. Toda mi gratitud para Ana, por hacerme reír en medio de mis peleas contra los corpus, por el tiempo que te han robado y, sobre todo, por creer en mí.

Quisiera reconocer a mi familia, amigos y, en general, a todos aquellos que me han acompañado en esta Tesis Doctoral y me han dado su apoyo y cariño.

Mi último agradecimiento es para George Kingsley Zipf por enunciar una ley tan apasionante.

ÍNDICE

LISTA DE TABLAS

LISTA DE FIGURAS

ABREVIATURAS, TERMINOLOGÍA Y SÍMBOLOS

Lista de abreviaturas

AWL: Academic Word List

BNC: British National Corpus.

COCA: Corpus of Contemporary American English.

EOI: Escuela Oficial de Idiomas

GSL: General Service List

L1: Lengua materna

L2: Lengua meta

PLH: Plan léxico adaptado a hispanohablantes.

NoPLH: Palabra no incluida en los objetivos de aprendizaje del PLH.

MCERL: Marco común europeo de referencia para las lenguas.

Trad. a.: Traducido por la autora.

VLT: Vocabulary Level Test

VST: Vocabulary Size Test

Sinónimos empleados

Salvo que se especifique lo contrario, se emplean indistintamente:

Léxico y vocabulario.	Lengua e idioma.
Palabra y unidad léxica.	Alumno, estudiante y aprendiz.
	Profesor y docente.
Aprendizaje y adquisición.	

Símbolos y operadores

$\forall$	Para todo
$\cup$	Unión de conjuntos
$\cap$	Intersección de conjuntos
$\in$	Pertenece a
$\subsetneq$	Contenido pero no igual a / subconjunto de
$\leq$	Menor o igual
$\geq$	Mayor o igual
$\sum$	Sumatorio
log (*n*)	Logaritmo de *n*
x :	*x* tal que
^	"y" (operador booleano)
$a \sim b$	*a* es proporcional a *b*
$a \rightarrow 1$	*a* tiende a 1
P *(a)*	Probabilidad de *a*
???	Palabra desconocida

INTRODUCCIÓN

El vocabulario es uno de los aspectos más importantes de cualquier sistema lingüístico y, sin embargo, ha sido desatendido sistemáticamente en la enseñanza de idiomas. Si bien es cierto que los primeros métodos incluían extensas listas de vocabulario que debía aprenderse de memoria, durante el último siglo las distintas recomendaciones pedagógicas han asumido que el vocabulario se aprende por mera exposición, fundamentalmente a través de la lectura. Desde esta perspectiva, la selección léxica en las programaciones docentes sería innecesaria.

A partir de los años 90 un grupo de investigadores, entre los que destacan Batia Laufer, Paul Meara, Paul Nation, Norbert Schmitt, James Coady y Michael McCarthy, publican una serie de estudios cuyos resultados cuestionan la validez de este enfoque (Schmitt, 2010). Sus trabajos reivindican la importancia de la selección del léxico prioritario así como su enseñanza explícita a través de distintas técnicas, especialmente en los niveles elementales. Surge entonces un renovado interés por aquello que ya defendían en los años 20 y 30 los lingüistas del llamado *Vocabulary Control Movement*, que fueron pioneros en la elaboración de un listado de las palabras más importantes para el aprendiz de inglés como L2. Este listado, llamado *General Service List* (West, 1953), ha cumplido 60 años y sigue siendo tremendamente influyente en la investigación sobre palabras de alta frecuencia con fines docentes.

La introducción de los avances tecnológicos ha permitido a los lingüistas estudiar el lenguaje humano mediante el procesamiento automático de enormes colecciones de textos que contienen millones de palabras. Nace así la lingüística computacional, una nueva rama de investigación que utiliza las herramientas informáticas sobre muestras de lenguaje real para detectar patrones de uso. Una de sus líneas de investigación en los últimas décadas ha sido, precisamente, elaborar listas de palabras frecuentes empleando modelos computacionales, una metodología que ofrece resultados mucho más rigurosos que la *General Service List*. Algunas de estas listas no estudian el lenguaje cotidiano, sino que están enfocadas únicamente al vocabulario técnico de un área concreta para utilizarlas como referencia para programaciones docentes de inglés profesional o con fines específicos. Así, se pueden encontrar listados de términos de alta frecuencia relativos a multitud de campos como la medicina, el derecho, el inglés de los negocios o las palabras recurrentes en el inglés académico. Es sorprendente, sin embargo, que no se puedan encontrar listas diseñadas especialmente para alumnos con una lengua materna determinada.

Un factor que puede explicar esto es el desprestigio del uso de la lengua materna como fuente de conocimiento transferible a la lengua meta. Esta visión, que parte de la teoría del *Análisis Contrastivo*, ha permanecido hasta hoy como dogma pedagógico, aunque ello implique desaprovechar el conocimiento potencial del alumno acerca del vocabulario que presenta similitudes en ambas lenguas. Aunque estructuralmente el inglés es una lengua de origen germánico, una proporción considerable de su léxico tiene una gran influencia latina. Estas palabras son especialmente representativas en el habla culta y en la mayor parte de registros científicos y académicos. La *Academic Word List* (Coxhead, 1998), la más influyente lista de inglés académico en la actualidad, contiene más de un 70% de palabras de raíz latina, muchas de las cuales pueden ser fácilmente deducibles por cualquier hablante de español que tenga unos conocimientos muy básicos de inglés, además de por hablantes de otras lenguas romances, como el portugués, el francés, el italiano o el rumano. Si además el alumno hispanohablante aprende las equivalencias de los afijos derivativos más comunes (-ly = -mente; *-less* = 'sin'; *-ence* = -encia; *-ize* = -izar, etc.) se estima que podría entender el significado y la categoría gramatical de unas 15.000 familias léxicas completas, incluso si las palabras están descontextualizadas (Montelongo, J. A., Hernández, A. C., & Herter, R. J., 2009).

A pesar de que la investigación ha demostrado que la influencia de la lengua materna puede ser un recurso facilitador de aprendizaje de segundas lenguas y no únicamente una fuente de errores, este conocimiento potencial aún no ha sido aprovechado en las listas de frecuencia ni en las programaciones léxicas de los libros de texto.

En definitiva, la investigación de las últimas dos décadas ofrece un marco teórico que favorece la selección léxica y la enseñanza explícita de vocabulario en niveles elementales. Sin embargo, las escasas aplicaciones docentes de estas teorías han sido demasiado generalistas y no han tenido en cuenta factores tan importantes como la influencia de la primera lengua del alumno. Si bien se pueden encontrar listados del vocabulario técnico de multitud de ámbitos, todos parecen estar enfocados a llegar a una meta concreta pero han pasado por alto algo tan importante como es el punto de partida, es decir, el conocimiento que ya tiene el alumno, aunque este sea únicamente potencial. Uno de los aspectos más importantes que se han descuidado son las palabras que el alumno es capaz de deducir debido a la semejanza formal y semántica entre lengua materna y lengua meta o, por el contrario, los *falsos amigos* con aparente parecido que en realidad suponen una trampa.

El propósito principal de este trabajo es diseñar un marco de trabajo para la enseñanza-aprendizaje de vocabulario optimizado para aprendices hispanohablantes de inglés como L2. El objetivo primario será elaborar un listado de vocabulario prioritario prestando especial atención a la influencia positiva y negativa del español como L1. Estadísticamente, este listado debe abarcar el número de términos necesario para poder leer textos auténticos[1] de manera autónoma. Una vez alcanzado este objetivo primario, se diseñará un sistema para determinar el vocabulario clave de planes de lectura extensa a través de modelos léxico-estadísticos.

Es importante remarcar que la elaboración de un listado de vocabulario no implica que su propósito final sea que los estudiantes memoricen una lista de palabras descontextualizadas. Lo que se pretende es fijar una serie de objetivos de aprendizaje

[1] En la didáctica de las lenguas, utilizamos las expresiones *textos auténticos* o *materiales auténticos* para referirnos a aquellos que están destinados a hablantes nativos; es decir, que no han sido adaptados ni escritos específicamente para ser utilizados como material de estudio de la L2.

prioritarios basada en la eficacia estadística, que se enseñará mediante las técnicas y metodologías que cada docente considere más adecuadas.

Se plantean las siguientes preguntas de investigación:

1. ¿Cuántas palabras de alta frecuencia es necesario conocer para poder leer textos auténticos en inglés sin que el vocabulario desconocido impida o afecte seriamente a la comprensión?
2. ¿Cuáles de esas palabras son deducibles por transferencia positiva del español y cuáles pueden inducir a error por transferencia negativa?
3. ¿Qué implicaciones pedagógicas derivan de estos datos y cómo pueden utilizarse para elaborar un plan léxico adaptado para hispanohablantes?

Asimismo, se pretende contestar a dos preguntas secundarias relativas a la aplicación directa de los resultados anteriores en la práctica docente:

4. ¿Cómo afecta la L1 de los hispanohablantes a los resultados de los tests estandarizados que evalúan el vocabulario receptivo en inglés? ¿Cómo se puede mejorar la fiabilidad de estas pruebas?
5. ¿Cómo se puede seleccionar el vocabulario más rentable de un programa de lectura extensa para que el número de objetivos de aprendizaje sea asumible?

El presente trabajo está dividido en cuatro partes. La primera parte establece la fundamentación teórica, revisa los estudios previos, describe el estado actual de la cuestión y plantea la motivación de esta tesis. La parte segunda presenta el enfoque, la validación estadística del diseño del estudio y de la metodología; y posteriormente, se presentan y discuten los resultados obtenidos. Por último, la parte tercera ofrece una propuesta de continuidad, resume las conclusiones más relevantes de esta tesis, así como sus aplicaciones pedagógicas, y presenta una perspectiva sobre el trabajo futuro.

PRIMERA PARTE

ANTECEDENTES Y ESTADO ACTUAL

I

1. LA ENSEÑANZA DEL VOCABULARIO

1.1. Antecedentes

El vocabulario es el corazón de todo sistema lingüístico, esta idea queda perfectamente reflejada en una conocida reflexión de Wilkins: «Without grammar very little can be conveyed, without vocabulary nothing can be conveyed»[2] (1972: 111). Cuando hablamos del vocabulario en el contexto del aprendizaje de una lengua extranjera, una de las citas más repetidas, generalmente atribuida a Stephen Krashen, es que cuando los alumnos viajan llevan consigo diccionarios, no libros de gramática.

Nadie parece dudar de que existe una relación entre cuántas palabras conoce una persona en una lengua extranjera (en adelante, L2) y su capacidad de comunicación en dicha lengua. El grado exacto de correlación entre la base léxica y la competencia comunicativa, así como las habilidades concretas en las que más influye el vocabulario, son preguntas que suscitan desde hace años un gran interés entre investigadores, profesores y psicólogos. Los distintos experimentos llevados a cabo no solo confirman la idea intuitiva de que un vocabulario amplio conlleva un mejor dominio del idioma, sino que sugieren que es posiblemente el factor más determinante, y el que parece

[2] Sin gramática se puede comunicar muy poco, sin vocabulario no se puede comunicar nada (trad. a.).

definir en mayor medida la competencia en todas las habilidades comunicativas (Meara, 1996; Schmitt, Jiang, & Grabe, 2011).

Concretamente, Laufer y Goldstein (2004) demostraron que conocer el significado de las palabras era el factor más influyente en las notas que obtenían los alumnos de inglés como L2 en los exámenes. Dos años después, se publicó un estudio (Alderson, 2006) que pone de manifiesto la relación entre la base léxica del alumno y su habilidad en comprensión oral, comprensión lectora y expresión escrita. Hay un alto grado de correlación entre todos ellos, destacando la clara dependencia entre lexicón mental y comprensión lectora. Sobre estos resultados, Schmitt (2010) apunta lo siguiente:

> Considering the multitude of the factors which could affect these scores (e.g. learner motivation, background knowledge, familiarity with test task),it is striking that a single factor, vocabulary knowledge, can account for such a large percentage of the variation. The relationship between vocabulary and writing is particularly strong, but even the individual skill subcomponents (e.g. inferencing) have strong relationships with vocabulary knowledge. (p. 4)

> Teniendo en cuenta la multitud de factores que podrían afectar a estas puntuaciones (por ejemplo, motivación del alumno, conocimientos previos, familiaridad con el tipo de prueba) es llamativo que sea un único factor, el conocimiento del vocabulario, el que pueda determinar un porcentaje tan grande de la variación. La relación entre el vocabulario y la escritura es particularmente sólida, pero incluso los subcomponentes de habilidades individuales (por ejemplo, la deducción) tienen una gran correlación con el conocimiento de vocabulario (trad. a.).

Como señalan Schmitt et al. (2011), el vocabulario obviamente no es el único factor que afecta a la comprensión lectora. El autor revisa los estudios de Droop y Verhoeven, (2003), Grabe (2009), van Gelderen et al. (2004) y Stoel, de Glopper y Hulstijn (2007) y agrupa el gran número de variables influyentes en tres grandes categorías: otros conocimientos lingüísticos (gramática, sintaxis), dificultad intrínseca

del discurso (longitud, tema, etc.) y factores relacionados con el lector en particular, como pueden ser su motivación y la capacidad de lectura comprensiva en su L1. El problema derivado de pretender clasificar *todos* los factores que afectan a la lectura es que el resultado pierde concreción; se podría argumentar que todas estas variables afectan a la lectura tanto como a cualquier otra habilidad lingüística. Incluso en habilidades productivas, como puede ser la expresión oral, parece obvio que los conocimientos lingüísticos, la dificultad del discurso y la capacidad del alumno para expresarse en su L1 afectarán a su competencia en esa habilidad.

En esta línea, hay otros estudios que sostienen que, al menos en la comprensión lectora, el vocabulario es un factor absolutamente determinante cuya influencia supera a la de todas las demás variables. En definitiva, de la investigación reciente podemos concluir que el vocabulario afecta en gran medida la competencia comunicativa de un aprendiz de L2 en todas las habilidades y que, concretamente, la comprensión escrita va a depender fundamentalmente de la base léxica en L2 del lector (Coady & Huckin, 1997; Huckin, Haynes, & Coady, 1995; Laufer, 2003; Pulido, 2009; Sökmen, 1997).

1.2. Conceptos clave del vocabulario

El Instituto Cervantes define el vocabulario como el conjunto de unidades léxicas de una lengua (1997a). Estas pueden ser simples (la idea intuitiva de *palabra*) o unidades compuestas con un único sentido, como pueden ser los *phrasal verbs* ingleses o incluso las fórmulas idiomáticas. En la lingüística aplicada a la enseñanza los términos *vocabulario* y *léxico* se utilizan de manera indistinta[3]. Sí es habitual en la investigación didáctica, en cambio, distinguir entre el *léxico global*, es decir, todas las voces existentes en la lengua meta, y *léxico individual*, también llamado *lexicón mental*, que se refiere al conjunto de palabras que conoce un aprendiz. En la didáctica de la L2, por tanto, nuestras acciones están dirigidas a mejorar el léxico individual del aprendiz para aproximarlo al léxico global de la lengua meta.

[3] En algunas disciplinas lingüísticas, *léxico* es una noción global que se refiere al conjunto de vocablos que forman un sistema lingüístico mientras que *vocabulario* son las palabras que un individuo utiliza en un momento concreto. Esta distinción tiene ciertos paralelismos con las dicotomías *langue* y *parole* de Sassure o *competence* y *performance* de Chomsky.

Al referirnos al léxico individual o lexicón mental de un aprendiz es necesario distinguir entre vocabulario receptivo y productivo, es decir, la capacidad que tiene un individuo para entender y para utilizar las palabras, respectivamente. Emplearemos también el término *vocabulario potencial* para referirnos a aquellas palabras que el alumno no conoce pero es capaz de deducir, ya sea por contexto o, fundamentalmente, por semejanza con su L1, como veremos en el punto 5 de este capítulo: *La influencia del español como L1 en el aprendizaje del inglés*.

Hay cierta controversia en cuanto al grado de dominio a partir del cual se puede considerar que el aprendiz *conoce* una palabra. En este sentido, Izquierdo (2003: 46) ofrece un exhaustivo listado de hasta 14 puntos relativos a las múltiples dimensiones del aprendizaje de una unidad léxica:

> a) reconocer la unidad léxica cuando se oye y saber pronunciarla,
>
> b) reconocer la forma escrita de la unidad léxica y ser capaz de escribirla,
>
> c) reconocer la morfología de la unidad léxica, es decir, los morfemas que la forman, relacionar dichas partes con su significado, así como ser capaz de formar la unidad léxica utilizando los morfemas correctos,
>
> d) reconocer las diferentes acepciones o significados y ser capaz de producir la unidad léxica para expresar su significado según el contexto,
>
> e) reconocer su categoría gramatical,
>
> f) conocer las estructuras sintácticas en las que puede aparecer y sus restricciones,
>
> g) reconocer y ser capaz de producir otras unidades con las que se relacione desde el punto de vista del significado (sinónimos, antónimos, cohipónimos, etc.),
>
> h) reconocer y ser capaz de producir las unidades léxicas con otras unidades con las cuales típicamente suele combinarse (“colocaciones”, relaciones sintagmáticas),
>
> i) conocer la adecuación pragmática de una unidad léxica a la situación o contexto comunicativo (según el lugar, el interlocutor, la intención, etc.),
>
> j) conocer su frecuencia de uso,

k) conocer a qué registro pertenece y utilizarla en una situación adecuada,

l) saber qué información cultural transmite para una comunidad lingüística,

ll) saber si pertenece a alguna expresión idiomática o institucionalizada,

m) reconocer y saber qué unidades están restringidas al discurso oral o escrito,

n) conocer sus equivalentes en otras lenguas

En este trabajo, sin embargo, utilizaremos la noción mucho más simplificada que se suele emplear en el ámbito de la didáctica de una lengua: diremos que el alumno *conoce* una palabra de forma receptiva si es capaz de entenderla con ciertas garantías, y de forma productiva si puede emplearla de manera eficaz en la comunicación, aunque no domine absolutamente todas las dimensiones del término. Este concepto va en línea con la puntualización que el Instituto Cervantes ofrece en su definición de lexicón mental:

> La existencia de una unidad léxica en el lexicón mental no supone necesariamente un conocimiento «completo» (fonología, ortografía, significado, construcción, etc.) ni «correcto» (el conocimiento individual puede diferir del valor de la unidad léxica en su comunidad hablante, o puede diferir de lo considerado normativo, puede incluso ser considerado erróneo por otros miembros de su comunidad).
>
> Diccionario de términos clave de ELE (Instituto Cervantes, 1997a)

Similarmente a lo que ocurre con *léxico* y *vocabulario*, en la didáctica de las lenguas los términos *palabra* y *unidad léxica* se suelen utilizar como sinónimos aunque técnicamente solo la segunda es el término riguroso para referirnos a la base mínima con valor semántico. En la presente tesis, sin embargo, nos centramos en las unidades léxicas simples, formadas por un único elemento, por lo que emplearemos los términos *unidades léxicas* y *palabras* de manera indistinta, salvo que se especifique lo contrario.

1.3. Los enfoques en la enseñanza del vocabulario

En el aprendizaje de una lengua extranjera, los errores de vocabulario son los más frecuentes y los más problemáticos para la comunicación. Sin embargo, en el último medio siglo tanto las metodologías más extendidas de enseñanza de idiomas como los libros de texto han minusvalorado la importancia del aprendizaje del vocabulario (Meara, 1980; O'Dell, 1997). Desde que se sustituyeron el Método Gramática-Traducción y el Método Directo por otros enfoques tales como el Método Audiolingual, el Enfoque Natural o las perspectivas humanistas, la enseñanza directa del vocabulario ha sido relegada o incluso desaconsejada (Gairns & Redman, 1986). Tampoco el Método Comunicativo ha favorecido, al menos en sus orígenes, la enseñanza explícita del vocabulario, optando en su lugar por un aprendizaje incidental basado fundamentalmente en estrategias de deducción de vocabulario por contexto (Cervero & Pichardo Castro, 2000).

Una de las cuestiones fundamentales en cuanto a la enseñanza del vocabulario es si esta debería ser implícita o explícita. El aprendizaje por deducción contextual, también llamado aprendizaje incidental, indirecto o implícito, sigue el paradigma del enfoque natural de Krashen y Terrell, que postula que el vocabulario se adquiere sin mayor esfuerzo a través de abundante exposición continuada. En el punto opuesto de esta línea de trabajo se encuentran los defensores del aprendizaje explícito, también llamado directo o sistemático. Este tipo de aprendizaje sigue una programación léxica controlada y requiere esfuerzo consciente del alumno mediante ejercicios y tareas diseñados específicamente para el aprendizaje de vocabulario. Estas metodologías suelen incluir también la enseñanza de estrategias de memorización (Huckin & Coady, 1999; Lee, 2003; Morin & Goebel Jr, 2001).

En la literatura actual es difícil encontrar un artículo sobre el aprendizaje explícito que no haga referencia a la paradoja del principiante —*Beginner's Paradox*— de Coady (1997) y la hipótesis del umbral mínimo —*Threshold Hypothesis*— de Laufer (1997). La paradoja de Coady plantea que en los niveles elementales hay un círculo vicioso: el aprendizaje de vocabulario a través la lectura difícilmente se podrá llevar a cabo hasta que el aprendiz no tenga la base léxica mínima necesaria para comprender el contexto. En el mismo sentido, Laufer defiende la existencia de un umbral o conocimiento mínimo de vocabulario a partir del cual es posible la comprensión de un texto auténtico.

Una de las soluciones que se proponen frecuentemente para abordar la paradoja del principiante es trabajar con libros simplificados destinados al aprendizaje de inglés como L2, llamados generalmente *graded readers* o lecturas graduadas. Estas pueden ser tanto obras adaptadas como libros originales con fines docentes que se caracterizan por una restricción de las estructuras gramaticales complejas y, fundamentalmente, por un control exhaustivo del vocabulario (Hafiz & Tudor, 1990). De hecho, la clasificación de los niveles de estos libros suele venir indicada en función del vocabulario necesario para su comprensión. Tras haber valorado sus ventajas e inconvenientes, algunos investigadores consideran que, si bien los *graded readers* no son materiales auténticos, permiten aprender vocabulario con palabras contextualizadas y pueden ser un medio de aprendizaje para poder llegar hasta los textos auténticos (Bertocchini, Costanzo, & Puren, 1998; Pérez Basanta, 1999).

Ya sea a través de lecturas graduadas o de textos auténticos, es evidente que la lectura en L2 favorece la adquisición de vocabulario y afianza el que ya se conoce. Algunos autores van incluso más lejos: en palabras de Krashen,

> Reading is good for you. The research supports a stronger conclusion, however. Reading is the only way, the only way we become good readers, develop a good writing style, an adequate vocabulary, advanced grammar, and the only way we become good spellers. (1993: 23)

> Leer es bueno para ti. Sin embargo, la investigación apoya una conclusión más sólida. Leer es el único camino, la única manera de llegar a ser buenos lectores, de desarrollar un buen estilo de escritura, un vocabulario adecuado, una gramática avanzada, y la única manera de llegar a tener buena ortografía (trad. a.).

Sin embargo, incluso con textos «fabricados» o simplificados hasta un nivel comprensible, la investigación ha demostrado que la adquisición indirecta no es el método más eficaz para la adquisición de vocabulario (Coady, 1997; Laufer, 2003). Si bien hay estudios (Nagy, Herman, & Anderson, 1985; Pulido, 2009) que indican, con

cifras variables, que el alumno suele enriquecer en mayor o menor grado su vocabulario mediante la lectura, Sökmen (1997) pone de manifiesto que el alumno que lee un texto en una lengua extranjera está más concentrado en comprender el contenido que en la retención del nuevo vocabulario, especialmente en el contexto de la lectura extensiva.

Al dar prioridad al mensaje, el alumno ignorará todas las palabras desconocidas que no interfieran excesivamente en la comprensión. Es más, incluso aquellas deducidas correctamente tienen muy pocas posibilidades de ser recordadas a medio plazo, a menos que haya un sistema de trabajo específico para la memorización y reciclaje de esas palabras (Pigada & Schmitt, 2006; Rott, 1999; Waring & Takaki, 2003).

La corriente que defiende el aprendizaje de vocabulario únicamente a través de la lectura actualmente es relativamente minoritaria, a pesar de haber estado apoyada en sus orígenes por lingüistas tan influyentes como Krashen. Su argumento principal es que ningún programa de instrucción explícita puede contener tanto vocabulario como un programa de lectura extensiva, llegando a comparar el aprendizaje explícito con intentar llenar una piscina con una cucharilla de té (Nagy et al., 1985). Si bien mediante la lectura el alumno estará expuesto a mucho más vocabulario que con un programa de instrucción explícita, hay que tener en cuenta que encontrar una palabra en un texto no implica necesariamente aprenderla; de hecho, el aprendizaje tras un único encuentro parece ser precisamente la excepción. En primer lugar, existe el riesgo de que el alumno le atribuya erróneamente un significado que no tiene y, en segundo lugar, estudios como el de Chun y Plass (1996) nos dan unas cifras desoladoras: «the research on learning words in context found only 5% to 15% probability that a given word would be learned at first exposure[4]» (p.184). Otro de los problemas fundamentales del aprendizaje por contexto es que no permite diseñar una planificación léxica porque no se puede controlar cuáles son las palabras que aprenderá el alumno, como señalan Day, Omura y Hiramatsu (1992), y Paribakht y Wesche (1997).

[4] La investigación sobre el aprendizaje contextual de vocabulario encontró una probabilidad de únicamente el 5% al 15% de que una palabra dada sea aprendida tras la primera exposición (trad. a.).

Actualmente la investigación apunta a que la adquisición por contexto y la instrucción explícita de vocabulario no tienen por qué ser posturas enfrentadas sino complementarias:

> All studies comparing incidental with intentional learning show that intentional learning is more efficient and effective. This should not be seen as a competition between incidental and intentional learning. Rather, a well balanced language programme should make good use of both types of learning. One without the other is inadequate (Waring & Nation, 2004: 20).

> Todos los estudios que comparan el aprendizaje incidental con el explícito muestran que el aprendizaje explícito es el más eficiente y eficaz. Esto no debe verse como una competición entre aprendizaje incidental y explícito. En su lugar, un programa bien equilibrado debería emplear ambos tipos de aprendizaje. El uno sin el otro es insuficiente (trad. a.).

Nadie parece poner en duda que la lectura debe favorecerse y que es muy beneficiosa tanto para la adquisición de nuevo vocabulario como para consolidar el que ya se conoce; sin embargo, en la práctica, la adquisición mediante la lectura es un camino lento y poco eficaz cuyos resultados de aprendizaje son relativamente impredecibles. Por el contrario, la instrucción explícita es más rápida y eficiente, además de que mantiene un control sobre las palabras de interés pero tiene la limitación de que no es posible enseñar de manera explícita todo el vocabulario existente debido al enorme número de unidades que contiene el léxico global. En este sentido, Nation apunta lo siguiente:

> The main problem with vocabulary teaching is that only a few words and a small part of what is required to know a word can be dealt with at any one time. This limitation also applies to incidental learning from listening or reading, but it is much easier to arrange for large amounts of independent listening and reading than it is to arrange for large amounts of teaching (Nation, 2006b: 329).

> El principal problema con la enseñanza de vocabulario es que en cada momento solo pueden tratarse unas pocas palabras y una pequeña parte de lo que se necesita saber sobre una palabra. Esta limitación se aplica también al aprendizaje incidental oral o escrito, pero es mucho más fácil organizar grandes cantidades de audición y lectura independiente que organizar grandes cantidades de enseñanza (trad. a.).

La estrategia más rentable parece ser utilizar la enseñanza explícita en los niveles más elementales y potenciar la adquisición por contexto cuando ya se ha alcanzado cierta base. En una primera fase, por tanto, debemos establecer cuidadosamente unos criterios de selección léxica, y encaminar nuestras acciones didácticas a la instrucción del vocabulario esencial hasta que el aprendiz alcance el umbral mínimo que le permitirá leer de manera autónoma. Cuanto antes consiga esta base, antes podrá desarrollar estrategias de deducción por contexto y enriquecer su vocabulario a través de la lectura extensa. Estas son las conclusiones que se derivan de los trabajos de los expertos en adquisición de léxico publicados en las últimas dos décadas (Coady & Huckin, 1997; Huckin, Haynes, & Coady, 1995; Huckin & Coady, 1999; Laufer, 2003; Laufer & Ravenhorst-Kalovski, 2010; O'Dell, 1997; Schmitt, 2010; Schmitt et al., 2011; Sökmen, 1997).

En resumen, para el aprendizaje del vocabulario básico es necesaria una instrucción explícita que debe estar enfocada principalmente a enseñar los rasgos semánticos principales y otros conocimientos básicos sobre las unidades léxicas. Cuando el alumno ha alcanzado una base mínima, se debe fomentar la lectura para la adquisición incidental, que servirá al alumno tanto para contextualizar y profundizar en las palabras aprendidas como para aprender vocabulario nuevo.

1.4. Criterios de selección del vocabulario

En 1917 Harold Palmer estudia la relación entre frecuencia y aprendizaje y señala que las palabras más frecuentes se aprenden antes. Esta conclusión ha sido apoyada por

las investigaciones de Hirsh y Nation (1992), Huckin y Coady (1999), Morin y Goebel (2001) y Laufer (2010), entre otros. Estas palabras frecuentes son también las más útiles, pues son las que permitirán al alumno entender y expresarse de manera más eficiente.

No obstante, en la mayoría de los libros de texto de inglés actuales no existe un programa léxico sistemático basado en la frecuencia o la rentabilidad, sino que suelen recopilar listas que agrupan una serie de unidades léxicas por campos semánticos o «centros de interés» tales como los medios de transporte o las profesiones. Posiblemente esto sea debido a que las tendencias metodológicas de los últimos años presumen que las palabras frecuentes se adquirirán de manera indirecta, principalmente a través de la lectura.

En general, los libros de texto de inglés como L2 suelen estar estructurados en torno a un programa gramatical. A pesar de que el Marco de referencia no establece directrices concretas en cuanto a los contenidos gramaticales, existe un esquema muy consolidado que parecen seguir la mayoría de los libros de texto, independientemente de su editorial. Un profesor de inglés experimentado que abre cualquier libro de texto de nivel B1-B2 no puede predecir cuál es el vocabulario clave que encontrará pero sabe que posiblemente los primeros temas incluyan los presentes simple y continuo, después se encuentren los tiempos pasados, el presente perfecto, *will* y *going to*, verbos modales, condicionales, pasivas, oraciones de relativo y, por último, *reported speech*. Este programa gramatical, que no está definido por el Marco de referencia sino más bien por la tradición, es la espina dorsal de las programaciones didácticas. En lo que difieren los distintos libros es en los conocimientos adicionales que aparecen como complemento al programa gramatical, tales como audiciones, lecturas, ejercicios de pronunciación y el vocabulario relativo al centro de interés que suele definir el título de cada tema.

Por tanto, se podría afirmar que la programación de gramática es universal mientras que los contenidos léxicos a partir del nivel intermedio pueden ser completamente diferentes en función del manual que se utilice [5] y, de hecho, al docente

[5] En los niveles muy elementales sí hay unos centros de interés comunes, tales como los colores, los meses de año o las partes del cuerpo. A medida que aumenta el nivel los contenidos varían.

le resultará difícil elegir cuál de ellos es el que tiene el programa léxico más adecuado porque es extremadamente inusual que una editorial justifique de manera transparente sus criterios de selección del vocabulario. Sin entrar en conjeturas sobre cuáles podrían ser sus fundamentos, se podría considerar que las palabras clave no suelen estar escogidas en función de su frecuencia, familiaridad o disponibilidad[6], sino porque pertenecen al campo semántico concreto que da título al tema y al que la editorial considera un centro de interés. Si bien hay algunos temas que parecen muy útiles para el aprendiz de un idioma extranjero, como pueden ser los relativos a los viajes o las visitas al médico, hay otros centros de interés que aparecen en distintos libros y que difícilmente se corresponden con una necesidad lingüística real en L2. Un ejemplo claro es el clásico tema acerca de los fenómenos naturales, entre cuyas palabras clave se suelen encontrar *earthquake* o *flood*[7] —'terremoto' e 'inundación'—, que son, para el hablante medio, términos muy alejados de su lenguaje cotidiano.

La tendencia en los libros de texto actuales es la inclusión de ejercicios para el vocabulario que consideran prioritario, es decir, apuestan por el aprendizaje explícito de algunos términos. Una de las peculiaridades es que las palabras que se trabajan en estos ejercicios suelen ser de baja frecuencia, por lo que podemos deducir que la filosofía que subyace es la visión tan afianzada de que el alumno aprenderá las palabras frecuentes por mera exposición y las de baja frecuencia mediante ejercicios de aprendizaje explícito. En la práctica, lo que ocurre suele ser muy distinto: muchos alumnos terminarán el curso dominando los términos *terremoto* e *inundación*, que rara vez van a necesitar, mientras que no habrán retenido la mayor parte del vocabulario útil porque no han hecho un esfuerzo consciente ni han seguido un plan de trabajo. Esto es consecuencia de la ineficacia de la adquisición por contexto frente al aprendizaje explícito, así como de la falta de control de las unidades aprendidas mediante adquisición incidental del que nos advierten Day et. al (1992) y Paribakht y Wesche (1997).

---

[6] *Familiaridad* y *disponibilidad* son dos criterios de selección de vocabulario relativos a las percepciones subjetivas sobre la frecuencia de un término y los contextos en los que se espera encontrar, respectivamente. Para más información ver Carter y McCarthy (1988) y López Morales (1978).

[7] Son palabras clave en el Tema 2 de English Alive 3, Ed. Oxford, (destinado a 3º de ESO) y el Tema 8 de Face2Face Intermediate, Ed. Cambridge (nivel B1-B2).

Por tanto, podemos concluir que en la planificación léxica de los libros de texto las editoriales combinan ambos enfoques, el aprendizaje incidental por exposición junto con ejercicios para el aprendizaje explícito. Sin embargo, la estrategia elegida es justo la contraria a la línea de trabajo que defiende la investigación reciente: los esfuerzos del alumno mediante aprendizaje explícito deberían estar dirigidos, en primer lugar, al vocabulario más frecuente, que será el más útil, universal y rentable para otros contextos.

2. LA DISTRIBUCIÓN LÉXICA EN INGLÉS

2.1. La frecuencia como criterio de selección del vocabulario

La frecuencia con la que se utiliza una unidad léxica es posiblemente el mejor criterio de selección para elaborar programaciones léxicas en cuanto a fiabilidad y eficacia. Hablamos de *fiabilidad* porque el estudio empírico a través de herramientas matemáticas tiene actualmente un grado de complejidad y precisión que es imposible de alcanzar en programaciones basadas en percepciones subjetivas como son la familiaridad o los centros de interés. La *eficacia* de la frecuencia como criterio de selección consiste en aprovechar la peculiar distribución del léxico, ya que las palabras no aparecen en el discurso de manera arbitraria, sino que hay cierto número de unidades que se repiten mucho más a menudo que las demás, en todo tipo de contextos y géneros.

Otra de las ventajas de la frecuencia a la hora de establecer niveles de prioridad léxica es que también es un indicador fiable para que el docente estime la probabilidad de que un alumno conozca una palabra. Schmitt ofrece la siguiente explicación:

> Frequent words are for the most part not inherently any easier than nonfrequent words, but, on average, they will be encountered more often, which means that they are more likely to be known than nonfrequent words. (Or more precisely, the most frequent meaning senses of these mostly-polysemous high-frequency words will be more likely to be known.) (Schmitt, 2010: 261)

> Las palabras frecuentes en su mayoría no son inherentemente más fáciles que las no frecuentes pero, en promedio, aparecerán más a menudo, lo que significa que es más

> probable que se sepan antes que las palabras no frecuentes. (O, más concretamente, las acepciones más frecuentes de las palabras polisémicas de alta frecuencia tienen más posibilidades de ser conocidas) (trad. a.).

En el ámbito de la Economía, una operación rentable es aquella que consigue los máximos beneficios con el mínimo coste. Trasladando esta idea al aprendizaje, podríamos considerar que este coste es el tiempo y esfuerzo que invierte el alumno, por lo que una programación léxica rentable debería tener como objetivo de aprendizaje aquellas palabras con mayor probabilidad de aparecer en cualquier contexto, que son las que permitirán al alumno maximizar los beneficios de su esfuerzo. Para determinar cuáles son estas palabras clave es importante tener en cuenta que la frecuencia de las palabras en inglés no es arbitraria sino que sigue una distribución muy específica: hay un número muy reducido de unidades que cubren la gran mayoría del lenguaje que se utiliza. La Figura I.1 muestra la distribución del léxico en el corpus SUBTLEXus, compuesto por subtítulos de películas estadounidenses (51 millones de palabras).

Una de las primeras publicaciones en la que se ofrece una sistematización matemática de este fenómeno es el trabajo del lingüista de Harvard George Kingsley Zipf (1949), que ha sido reproducido después en numerosos estudios y dio lugar a la llamada Ley de Zipf [8]. Se trata de una ley empírica que formula que la frecuencia de aparición de las palabras sigue una distribución que puede aproximarse por:

$$f_n \sim 1/n^a$$

donde *n* es el puesto que ocupa cada palabra en la lista ordenada, f_n es su frecuencia y el exponente *a* es próximo a 1. Esto significa que la palabra más frecuente se repite el doble que la segunda, el triple que la tercera y así sucesivamente.

[8] Manning y Schütze (1999) sostienen que el primero en advertir la relación entre frecuencia y rango de las palabras fue Estoup en 1916, pero que la publicación de Zipf fue quien la dio a conocer y por eso la ley lleva su nombre.

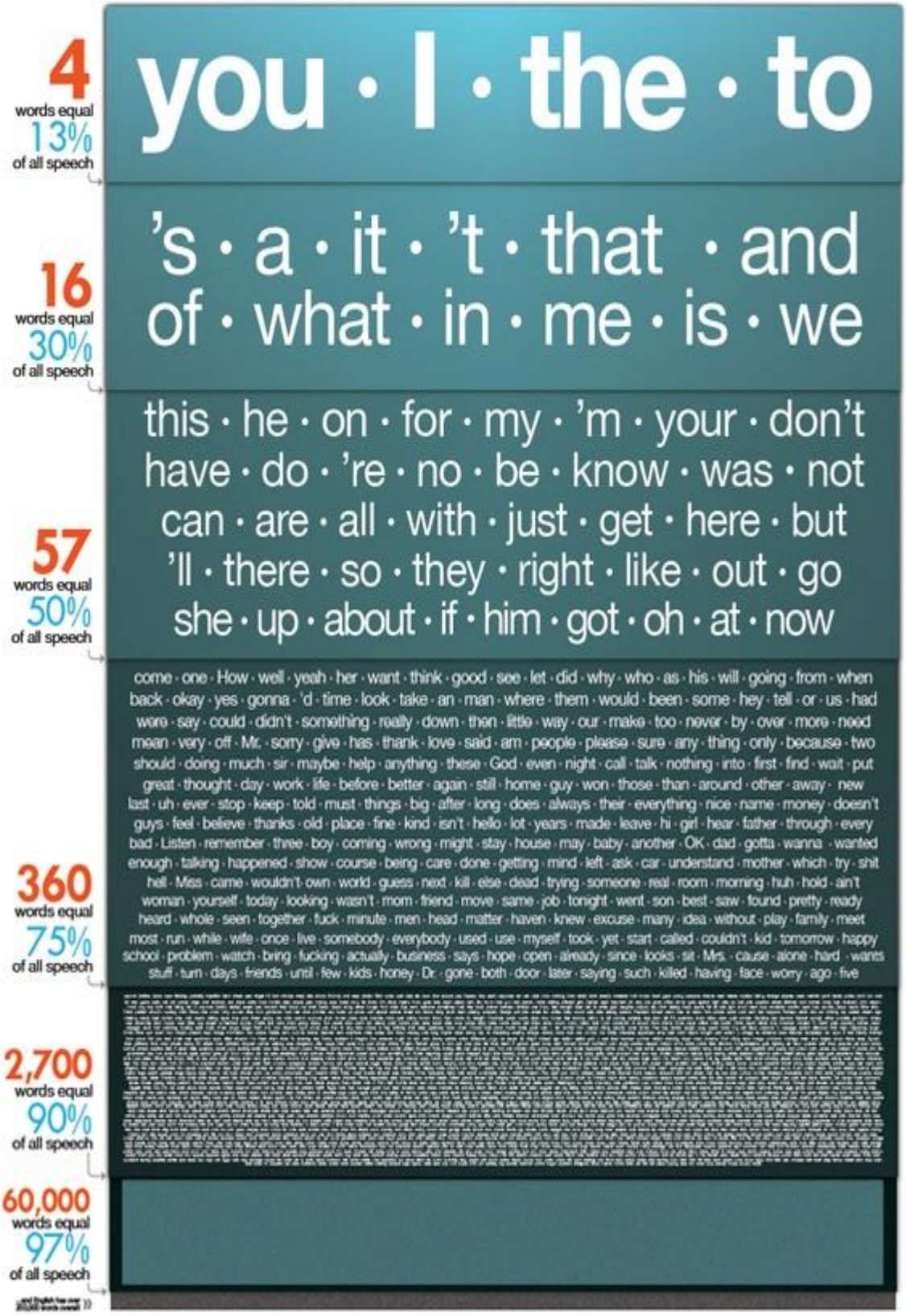


Figura I.1. Distribución léxica en el corpus SUBTLEXus. (http://resources.phrasemix.com/img/full/Word frequency.pdf)

Concretamente, Zipf determinó que si se examina un corpus extenso y se ordenan las palabras por orden decreciente de frecuencias se puede explorar la relación entre la frecuencia de una palabra, *f*, y su posición en la lista, *n*. A esta posición *n* la denominó rango. Zipf estableció que la palabra más frecuente del corpus (rango 1) aparece el doble que la siguiente (rango 2), el triple que la tercera (rango 3) y así sucesivamente.

Esto implica, en términos generales, que entre todo el léxico disponible hay una gran mayoría de unidades que rara vez aparecen, un conjunto menor de palabras de frecuencia media y unas pocas palabras que aparecen muy a menudo. Este último grupo contiene las palabras con carga semántica de uso más frecuente, así como la mayor parte de las unidades gramaticales (pronombres, determinantes, preposiciones, etc.). En definitiva, el vocabulario de alta frecuencia supone una proporción muy pequeña de todo el léxico disponible y, sin embargo, contiene la gran mayoría de las palabras que realmente se utilizan. Por esta razón, deberían ser el objetivo primordial de aprendizaje léxico.

2.2. Estimaciones subjetivas de la frecuencia

Las personas parecen poder intuir la frecuencia del vocabulario en su propia lengua. Un hablante de español sin nociones de lingüística es capaz de determinar que para un estudiante extranjero es más importante aprender la palabra *triste* antes que *afligido*, *desdichado* o *apesadumbrado*. Esta estimación subjetiva de la frecuencia léxica es una herramienta de la que se han servido tanto docentes como alumnos para discriminar entre el vocabulario «importante» frente a las palabras «raras» (o también referidas como «difíciles», aunque objetivamente su aprendizaje no sea más complicado que el del vocabulario frecuente).

Sin embargo, hay ciertas dudas sobre el grado de precisión con el que un hablante, y más concretamente un profesional experimentado de la enseñanza de L2, es capaz de determinar la frecuencia de las palabras en la lengua meta. Las primeras investigaciones en la década de los 70 apuntaban que la intuición de los hablantes nativos era fiable y acertada (Richards, 1976, citado por McCrostie, 2007: 54). Sin embargo, las réplicas posteriores de estos experimentos, como la llevada a cabo por Schmitt y Dunham (1999), revelan que había ciertas deficiencias en el diseño y los resultados previos, ya que las unidades que se presentaron a los participantes del estudio estaban situadas en

los rangos extremos, mientras que no se sabía qué ocurriría con las palabras de frecuencia media.

En 2007 McCrostie diseña un estudio sobre la intuición de frecuencia de amplio espectro en los hablantes nativos. Un grupo de participantes estaba formado por profesores de inglés nativos con cuatro años de experiencia y formación específica en lingüística, el otro grupo estaba compuesto por estudiantes universitarios de primer curso. Para una de las tareas se les presentó una lista de palabras y debían identificar las que pertenecían al rango más frecuente o al rango medio. El rango frecuente contenía palabras escogidas entre las 2.000 más utilizadas en inglés, y el medio comprendía los niveles 4.000 a 10.000. La elección de los rangos no es aleatoria, McCrostie estableció esa división porque la léxico-estadística ha estimado que con únicamente las 2.000 palabras más frecuentes se puede cubrir más del 80% de cualquier texto (Nation & Waring, 1997), por lo que es muy importante que un docente identifique correctamente cuáles son estas palabras prioritarias.

Las conclusiones del estudio, sin embargo, son desalentadoras:

> The results from this study indicate that the English teaching professionals' accuracy judgments do not seem to be significantly better than university undergraduates. Furthermore, both groups of native English speakers had difficulty judging the frequency of words in the middle frequency range. These results indicate the need for teachers to consult frequency lists rather than rely solely on their intuitions (McCrostie, 2007: 53).

> Los resultados de este estudio indican que la precisión de la estimación de los profesionales de la enseñanza del inglés no parece ser significativamente mejor que la de los universitarios. Es más, ambos grupos de hablantes nativos de inglés tuvieron dificultad para juzgar la frecuencia de palabras en el rango de frecuencia media. Estos resultados indican la necesidad de que los profesores consulten listas de frecuencia en lugar de depender exclusivamente de sus intuiciones (trad. a.).

A pesar de lo decepcionante de los resultados, Schmitt et al. (2011) señalan que no es recomendable emplear un criterio de selección que dependa únicamente de recuentos automáticos en corpus, ya que todos tienen sus limitaciones. Un ejemplo para ilustrar esto podría ser la palabra *nine*, que curiosamente tiene una presencia en corpus muy inferior que los ocho primeros números; sin embargo, en la práctica docente no tiene sentido hacer una programación léxica que no presente la secuencia completa de números.

En resumen, no es razonable pensar que un docente no está capacitado para determinar objetivos de aprendizaje basándose en su intuición y experiencia, o que la frecuencia en corpus debe ser un criterio exclusivo. Sin embargo, es muy importante que los profesores sean conscientes de sus propias limitaciones y que pueden servirse de la lingüística computacional para obtener información extremadamente útil para elaborar sus programaciones basadas en la rentabilidad léxica.

3. ¿CUÁNTAS PALABRAS DEBEN APRENDERSE?

Si asumimos que el vocabulario debe enseñarse de manera explícita hasta alcanzar la base que permita al alumno leer de manera autónoma, la pregunta imprescindible es dónde está situado el punto de inflexión, es decir, cuántas palabras se deben enseñar mediante instrucción explícita para llegar a ese nivel de comprensión.

3.1. Tamaño del vocabulario de un nativo

Para tener una referencia sobre el orden de magnitud del vocabulario de una L2, podemos tomar como primera referencia el tamaño de la base léxica de un hablante nativo. El problema, según Nation (1990), es que la mayoría de las investigaciones sobre este asunto han sido muy poco rigurosas en cuanto a su diseño metodológico, lo que ha dado lugar a resultados absolutamente dispares. En opinión de Schmitt (2010) uno de los primeros estudios fiables es el de Goulden, Nation y Read (1990), que estima un vocabulario de unas 17.000 familias léxicas en los estudiantes universitarios

neozelandeses, una cifra congruente, aunque ligeramente superior, a los resultados de un experimento similar diseñado por D'Anna, Zechmeister y Hall (1991).

Tanto Schmitt (2010) como Nation (2001) avalan también el diseño del experimento de Zechmeister, Chronis, Cull, D'Anna y Healy (1995) que estudia la diferencia entre el tamaño del léxico en diversos grupos de edad en Chicago. Este trabajo concluye que los alumnos del primer ciclo de Secundaria acertaron algo menos de 12.000 palabras, los estudiantes universitarios reconocieron unas 16.600, y los adultos jubilados, unas 21.000. Zechmeister et al. (1995) advierten de que el tipo de test de los alumnos de Secundaria era más fácil, por lo que «these results likely underestimate the difference between the younger and older adults in terms of general vocabulary knowledge[9]» (1995: 210). Schmitt (2010), en función de estos resultados y de sus propios experimentos sostiene que el tamaño real de un hablante nativo culto oscila entre 16.000 y 20.000 palabras.

3.2. Vocabulario necesario para la lectura en L2

Hacia la década de los 90 surge una importante corriente en la investigación que trata de responder a la pregunta de cuántas palabras necesita un alumno extranjero para leer en inglés. El punto de partida en ese momento eran las investigaciones previas que señalaban que para captar la idea general es necesario conocer al menos un 80% del vocabulario de un texto, lo que requiere saber aproximadamente 2.000 palabras (Milton, 2009). Sin embargo, una nueva corriente de investigaciones encabezada por Batia Laufer y Paul Nation pretendía averiguar cuántas palabras necesita realmente un alumno de L2, no para captar la idea general sino para entender textos auténticos con ciertas garantías.

Uno de los experimentos más influyentes fue el estudio *What percentage of text-lexis is essential for comprehension* de Laufer (1989) en el que los participantes debían subrayar las palabras que no conocían de un texto y posteriormente hacer un examen de

[9] Estos resultados probablemente infravaloran la diferencia entre los adultos jóvenes y mayores en cuanto al conocimiento del vocabulario general (trad.a.).

comprensión lectora sobre ese texto. Así, Laufer pudo calcular dónde estaba el porcentaje de cobertura que distinguía a los alumnos que aprobaban el examen frente a los que no. La conclusión de este estudio es que solamente aprobaban los que conocían al menos el 95% de las palabras.

A raíz de estos resultados Laufer elaboró su influyente *Hipótesis del umbral mínimo*, que postula que hay una base léxica a partir de la que el alumno es capaz de entender un texto dado (1992). Tomando como referencia el 95% de su estudio anterior y el análisis de corpus, concluyó que un alumno necesitaría conocer un mínimo de 3.000 palabras de alta frecuencia. Esta cifra tuvo una gran repercusión y ha permanecido durante años como sistema de referencia, avalado por estudios similares de expertos como Hirsh y Nation (1992).

A partir del año 2000 otros experimentos sugieren que la estimación de Laufer quizás era demasiado optimista. En el estudio de Hu y Nation (2000) los participantes necesitaron conocer un 98% de las palabras que aparecían en el texto, lo que requiere conocer más de 8.000 familias léxicas. Diez años después, la propia Laufer publica un artículo titulado *Lexical threshold revisited* en el que apoya la teoría de Nation según la cual el umbral óptimo es el 98%, que corresponde a unas 8.000 palabras, aunque sigue defendiendo la existencia de un umbral mínimo situado en el 95% (Laufer & Ravenhorst-Kalovski, 2010).

Esta nueva cifra de 8.000 palabras es abrumadora, hasta tal punto que, de ser cierta, prácticamente habría que desechar la idea de alcanzarla mediante instrucción explícita. Sin embargo, hay una razón que explica esta diferencia entre 2.000, 3.000 y 8.000 palabras a pesar de que todos los estudios parecen estar bien diseñados y seguir una metodología rigurosa. La clave de esta diferencia en los resultados se debe a una simple cuestión semántica en la pregunta de investigación: el umbral está situado en uno u otro punto dependiendo de lo que cada investigador ha considerado que significa *comprender un texto*. A mayor grado de exigencia en este concepto subjetivo, mayor será el vocabulario requerido. Si repasamos la metodología de los estudios citados, observamos que la referencia de Laufer (1989) era aprobar un examen de comprensión lectora, mientras que Nation en sus distintos experimentos (Hirsh & Nation, 1992; Nation & Waring, 1997; Nation, 2006a) considera que el punto de referencia es que el

alumno lea con fluidez, concretamente se insiste en que tenga la sensación de *lectura placentera*. Se puede concluir, por tanto, que los resultados no son contradictorios, sino que están midiendo conceptos distintos.

Por esta razón, la revisión reciente de Milton (2009) sigue dando validez tanto a los resultados iniciales sobre la existencia de un umbral muy básico situado en 2.000 palabras como a la estimación de Laufer acerca de las 3.000 palabras necesarias para aprobar exámenes. Concretamente, Milton sostiene que un nivel B1 se alcanza con 2.500 palabras y que aprobar el Cambridge First Certificate (B2) requiere 3.500. Por otra parte, Schmitt et al. (2011) realizaron un experimento sobre los umbrales del 95% y el 98% de Laufer y Nation y, aunque opinan que no hay ningún umbral a partir del cual el alumno suba su competencia de manera drástica, avalan la existencia de un punto mínimo de cobertura necesario para la lectura. Sus resultados lo sitúan en el 98% para leer textos académicos sin ninguna ayuda, mientras que consideran que el 95%, unas 3.000 palabras, puede ser un buen nivel para textos instructivos con apoyo del docente.

En resumen, los estudios demuestran que hay un umbral mínimo para la comprensión lectora que se alcanza mediante cierto número de palabras de alta frecuencia. Este umbral estará situado en un punto u otro en función del grado de comprensión que se quiera alcanzar. Con menos del 90% de cobertura no se garantiza siquiera entender la idea general. Para entender la mayor parte del texto, es necesario conocer al menos el 95% del vocabulario, unas 3.000 palabras. El umbral más exigente, que implica una lectura placentera sin ayuda externa, está situado en el 98%, es decir, al menos 8.000 palabras de alta frecuencia.

3.3. Unidades de medida del vocabulario

La investigación sobre el tamaño del vocabulario siempre requiere algún método de medida de unidades léxicas. El investigador, por tanto, debe valorar entre los distintos tipos de herramientas y técnicas de medida cuál es la más adecuada para describir con precisión la magnitud de los aspectos del vocabulario objeto de estudio.

Una de las cuestiones más importantes a la hora de diseñar una investigación léxica es decidir cuál será la unidad de medida utilizada. De hecho, uno de los problemas principales a la hora de evaluar los resultados de la investigación previa es que a menudo no ofrecen una indicación clara del sistema de referencia que han utilizado, sino que hablan únicamente de «palabras». Si un estudio concluye que el alumno debe aprender, digamos, 3.000 palabras, es muy importante saber si los términos *teach* y *teacher* se han contado como una unidad o dos.

Las herramientas de lingüística computacional nos permiten medir las palabras de un texto en tres unidades distintas: *tokens, types* y *groups*[10]. Si contamos *tokens* el resultado será el número de palabras que aparece en un texto, independientemente de que estén repetidas. Al contar *types* obtenemos el número de palabras distintas. Por último, *groups* es el número de familias léxicas distintas. Por ejemplo, si queremos contar el número de palabras que hay en la frase:

Sally worked last weekend, she works every weekend.

obtendremos 8 tokens, 7 types y 6 groups. Los 8 tokens son el número total de unidades, aunque *weekend* aparezca varias veces. Types son las palabras únicas, por lo que ignoramos la segunda ocurrencia de *weekend* pero contamos *worked /works* como dos palabras distintas. Por último, *groups* son las familias léxicas diferentes, por lo que consideramos que *woked/works* pertenecen a la misma categoría y las contamos una sola vez. Esto implica que cuando se cuentan palabras en un texto en *groups*, el recuento de ocurrencias de *work* será la suma de las ocurrencias de todas las palabras derivadas de ese lema: *work, working, worked, worker,* etc.

Para responder a la pregunta de cuántas palabras es necesario conocer se suele utilizar como referencia el número de familias léxicas o *groups*; es decir, se asume que el alumno que conoce la palabra *dance* es capaz de deducir el significado de sus derivadas formadas por afijos comunes, como *dancer* o *danced*. El problema, sin embargo, es establecer un límite en el grado de derivación. Por ejemplo, ¿conocer el

[10] Se pueden traducir como *casos, tipos* y *grupos o lemas*-, pero en la investigación lingüística publicada en español la tendencia es utilizar los términos ingleses, posiblemente porque las expresiones *tipos de palabras* y *grupos de palabras* pueden dar lugar a ambigüedad.

verbo *express* implica necesariamente que el alumno deducirá sus derivadas *expressionlessly* o *expressiveness*? Hay muy pocos estudios —quizás los únicos sean los de Paul Nation— que expliquen de forma transparente hasta qué grado de derivación están asumiendo que un alumno de L2 deducirá que dos palabras pertenecen a la misma familia[11].

En definitiva, para interpretar los datos sobre el porcentaje de texto que un alumno debe conocer es necesario tener en cuenta que se suelen utilizar dos unidades de medida diferentes en la presentación de resultados. En primer lugar, al hablar de una cobertura del 95% se hace referencia a *tokens*, es decir, el 95% de todas las palabras que aparecen en el texto, independientemente de que algunas estén repetidas. La razón es que una palabra desconocida que aparece varias veces será un reto para el alumno en cada una de sus ocurrencias. En segundo lugar, cuando se asocia un porcentaje de cobertura a una base léxica cuantificada, por ejemplo las 3.000 palabras más frecuentes, este número hace referencia a familias léxicas o *groups*, es decir, se cuenta como una unidad el grupo del lema principal y sus palabras derivadas. Suele ser opaco, sin embargo, cuál es el grado de derivación que admiten la mayor parte de los estudios publicados.

4. LA INVESTIGACIÓN DE CORPUS

Para poder hacer una generalización del número de palabras frecuentes que cubrirían un porcentaje específico de un texto cualquiera es necesario analizar grandes cantidades de textos de distintos géneros, esto es lo que se conoce como estudio de corpus.

4.1. Conceptos y evolución

En los últimos años, los lingüistas han explotado sistemas cada vez más sofisticados para compilar y procesar colecciones de texto cada vez mayores que sirvan

[11] El criterio de Nation para la elaboración de sus listas de alta frecuencia con respecto de los afijos admitidos se explicará en detalle en el punto 3 de la Parte Segunda.

como base para estudiar el lenguaje natural. Un corpus se suele definir como un conjunto estructurado de textos que tiene tres características: (a) está almacenado en formato electrónico para su proceso con un ordenador, (b) está compilado específicamente para el análisis lingüístico, y (c) trata de ser representativo, ya sea de todo un lenguaje o de una variedad específica.[12] (Llisterri, 2013; Manning & Schütze, 1999; Sinclair, 2005).

> A corpus constitutes an empirical basis not only for identifying the elements and structural patterns which make up the systems we use in a language, but also for mapping out our use of these systems. A corpus can be analysed and compared with other corpora or parts of corpora to study variation. Most importantly, it can be analysed distributionally to show how often particular phonological, lexical, grammatical, discoursal or pragmatic features occur, and also where they occur (Kennedy, 1998: 4).

> Un corpus constituye una base empírica para la identificación de los elementos y los patrones estructurales que componen los sistemas que utilizamos en un idioma, pero también para trazar el esquema de nuestro uso de estos sistemas. Un corpus se puede analizar y comparar con otros corpus o porciones de corpus para estudiar las variaciones. Lo más importante es que puede analizarse su distribución para mostrar la frecuencia con la que ocurren ciertas características fonológicas, léxicas, gramaticales, discursivas o pragmáticas, así como dónde se producen (trad. a.).

El estudio de corpus permite, fundamentalmente, descubrir patrones presentes en el lenguaje natural. Esto ha permitido abrir nuevas líneas en la investigación lingüística

[12] Kennedy (1998) apunta que esta es la noción actual de corpus pero históricamente esta definición no es del todo precisa: por un lado, los lingüistas ya trabajaban con corpus mucho antes de que existieran los ordenadores y, por otro lado, no todos los corpus que se utilizan para la investigación fueron compilados originalmente con ese fin. Para el caso específico de los corpus orales, Garrote (2010) añade una cuarta característica fundamental: que las muestras de habla sean auténticas. Es muy habitual que los corpus orales estén creados de manera artificial, lo que puede proporcionar datos no fiables.

con fines pedagógicos, así como en otras diversas áreas. Entre ellas, Lorenzo Salazar (2011: 9) cita las siguientes:

lexicología	lingüística histórica
semántica	traducción
pragmática	psicolingüística
análisis del discurso	antropología cultural
dialectología	psicología social
estudios sobre variación del lenguaje	lingüística forense
sociolingüística	lexicografía

Evidentemente, el análisis de corpus no es un simple proceso automático de recuento, sino una técnica de investigación que nos permite descubrir fenómenos lingüísticos. La habilidad más importante del lingüista computacional, según Kennedy (1998), no consiste en ser capaz de programar, ni siquiera en utilizar con soltura el software disponible, sino en saber plantear buenas preguntas sobre lingüística teórica que se puedan resolver mediante el estudio de grandes muestras de lenguaje real en forma de texto. Es decir, el objetivo de la investigación de corpus no consiste en proporcionar datos numéricos, sino en analizar estos resultados numéricos para descubrir patrones del lenguaje (Biber, Conrad, & Reppen, 1998, en Lorenzo Salazar, 2011).

De hecho, los análisis de corpus más influyentes para la selección léxica datan de mucho antes de que la tecnología permitiera procesar millones de palabras de forma automática. En la década de 1920 un grupo de profesores de inglés, entre los que destacan Michael West y Edward Thorndike, se planteó la posibilidad de llevar a cabo estudios de corpus con fines pedagógicos: su objetivo era maximizar los esfuerzos de aprendizaje de sus alumnos extranjeros elaborando una lista de términos de alta frecuencia. Su informe *Interim report on vocabulary selection*, publicado en 1936, resumió las conclusiones de las investigaciones del llamado *Vocabulary Control Movement*, que sostenía que para los alumnos suponía una gran ventaja aprender en primer lugar las palabras más frecuentes.

Dentro de esta nueva corriente educativa se elaboraron las primeras listas sistematizadas de palabras con fines docentes. Thorndike escribió tres libros de palabras de alta frecuencia, con el objetivo de ofrecer a los profesores información para discriminar entre las palabras frecuentes —que requieren énfasis en la instrucción— frente las que solo son necesarias para entender un contexto puntual. El primer libro de la serie, *The Teacher's Word Book* (1921) contiene las que, según el corpus de Thorndike, eran las 10.000 palabras más frecuentes en inglés estándar. Las palabras aparecían ordenadas alfabéticamente acompañadas por su número de ocurrencias y la frecuencia relativa en el corpus. Once años después se publicó el segundo libro, que duplicaba el número de palabras de su antecesor. El tercer libro de la serie —el más influyente de todos— salió a la luz en 1944 bajo el título *The Teacher's Word Book of 30,000 Words* (Thorndike & Lorge, 1944).

Sin embargo, el trabajo más significativo del *Vocabulary Control Movement* fue una pequeña lista de tan solo 2.000 palabras llamada *General Service List* (West, 1953). Su objetivo era compilar las palabras más frecuentes, que permitirían a un alumno poder comunicarse en inglés, aunque fuera de manera muy básica. En la lista, todas las palabras estaban anotadas con datos adicionales tales como su categoría gramatical o las distintas acepciones que podía tomar. La *General Service List* ha sido tremendamente influyente en la elaboración de programas de vocabulario básico hasta la actualidad. Es cierto que hoy se considera que algunas de las palabras de la lista necesitan ser reemplazadas, fundamentalmente debido a que West refleja el lenguaje de la primera mitad del siglo XX[13]. Sin embargo, los investigadores que han evaluado su eficacia con sofisticadas herramientas de lingüística computacional aún se sorprenden de la precisión y la validez estadística del listado que compilaron West y sus colegas de forma manual[14].

[13] Uno de los ejemplos más citados es *shilling*, 'chelín', una moneda británica que se acuñó por última vez en 1970. Por el contrario, faltan términos como *plastic*, *video* o *television*.

[14] Para hacernos a la idea de su alcance, el programa de análisis léxico on-line LexTutor, en el que están involucrados investigadores de la talla de Paul Nation o Tom Cobb, mantiene la General Service List entre sus principales listas de referencia, y la versión más reciente de AntWordProfiler, lanzada en 2013, la utiliza por defecto.

La aplicación de las tecnologías de la información a la lingüística de corpus facilitó enormemente la elaboración de listados de palabras de alta frecuencia. De entre los pioneros, los más citados probablemente sean *Computational analysis of present-day American English*, de Kučera & Francis (1967), y *Longman lexicon of contemporary English* (McArthur, 1981). Estos listados estaban basados en corpus que contenían aproximadamente un millón de palabras, que requerían varias horas para ser procesadas en aquella época. La evolución de los ordenadores ha hecho crecer de manera drástica el tamaño de los corpus que se pueden analizar de tal manera que actualmente los corpus de referencia oscilan entre los 100 millones del British National Corpus y los 450 millones que contiene el Corpus of American English[15].

Más recientemente, los lingüistas computacionales Kilgarriff y Grefenstette ofrecen la interesante propuesta de utilizar la web entera como corpus (2003). A pesar del atractivo inicial que ofrece un corpus enorme en constante actualización, el fuerte sesgo que tiene la web la invalida como fuente ya que hay palabras cuya frecuencia en Internet es exagerada (ej: *password, download, sex*).

En 2010 Google Inc. revoluciona el análisis léxico con una herramienta llamada Google Books N-gram Viewer que permite hacer consultas sobre una palabra (o una combinación de palabras) y devuelve una gráfica que muestra la evolución en la frecuencia de su uso a lo largo de la historia. Uno de los aspectos más relevantes de esta herramienta es el inmenso tamaño de su corpus: la colección de libros digitalizados de Google, que contiene más de 5,2 millones de obras completas y sigue aumentando. Tanto su base de datos como el corpus anotado están disponibles para su descarga y tratamiento estadístico. Dado que N-gram Viewer muestra la frecuencia por años, la mayor parte de las investigaciones que se sirven de esta herramienta se suelen centrar en el análisis de tendencias culturales y lingüísticas en la historia tales como la fecha de aparición de ciertos neologismos, los cambios de significado en una palabra a lo largo del tiempo (ej.: *gay*) o los términos que aparecen con mayor frecuencia en ciertas épocas.

[15] Hay algunos incluso más grandes, como el Bank of English de Collins que supera los 500 millones de palabras, aunque suelen contener casi exclusivamente lenguaje escrito.

Sin embargo, este valioso corpus ha abierto un nuevo horizonte en la investigación sobre las palabras de alta frecuencia y su evolución histórica, de tal manera que en los dos últimos años se han publicado interesantes investigaciones en esta línea. Sus conclusiones más relevantes para este estudio son que el vocabulario más frecuente permanece prácticamente invariable en los dos últimos siglos y que la ley empírica de Zipf, enunciada 60 años antes con recursos que hoy se consideran muy limitados, sigue teniendo la misma validez al analizar millones de libros Brysbaert, Buchmeier, Conrad, Jacobs, Bölte, & Böhl, 2011; Perc, 2012).

4.2. Corpus actuales de referencia

A pesar del potencial del corpus de Google para otras disciplinas lingüísticas, la preferencia para los estudios con fines docentes son los corpus de lenguaje contemporáneo, fundamentalmente aquellos que contienen distintos géneros y muestras tanto de inglés escrito como hablado. Esto es lo que se conoce como corpus equilibrado.

Posiblemente el corpus más popular de inglés británico sea el British National Corpus (en adelante, BNC), que se desarrolló en sus orígenes como un proyecto conjunto de distintas universidades y editoriales prestigiosas: Addison-Wesley Longman, Oxford University Press, Chambers Harrap, Oxford University Computing Services, Lancaster University y The British Library Research and Development Department. Sus datos ocupan aproximadamente 1,5 Gb, un tamaño que en sus inicios, 1991, requería un servidor de la universidad dedicado en exclusiva, pero que actualmente se procesa rápidamente con un PC doméstico. En cuanto a su composición, aproximadamente el 10% de sus muestras son transcripciones de habla espontánea —en contexto formal e informal—, y el 90% son obtenidas de textos de distintos géneros. Suman un total de aproximado de 100 millones de palabras.

La Tabla I.1 muestra la composición de la última versión del British National Corpus en la última versión de 2007 en formato XML anotado.

Tabla I.1. Composición del British National Corpus XML Edition.

CLASE DE TEXTO	TEXTOS	UNIDADES	% DEL BNC
Hablado – por grupos de edad	153	4.233.955	4,30
Hablado – por contextos	757	6.175.896	6,27
TOTAL HABLADO	**908**	**10.409.851**	**10,57**
Escrito - libros y periódicos	2.688	79.238.146	80,55
Escrito - para ser hablado	35	1.278.618	1,29
Escrito - miscelánea	421	7.437.168	7,56
TOTAL ESCRITO	**3.141**	**87.953.932**	**89,40**
TOTAL BNC	**4.049**	**98.363.783**	**99,97**

Adaptado de *BNC in numbers* (2009)

Para hacernos una idea del orden de magnitud, el propio BNC señala lo siguiente:

> To put these numbers into perspective, the average paperback book has about 250 pages per centimetre of thickness; assuming 400 words a page, we calculate that the whole corpus printed in small type on thin paper would take up about ten metres of shelf space. Reading the whole corpus aloud at a fairly rapid 150 words a minute, eight hours a day, 365 days a year, would take just over four years.
>
> The BNC in numbers (British National Corpus, 2009)

> Para poner estos números en perspectiva, el libro de bolsillo promedio tiene alrededor de 250 páginas por centímetro de espesor; asumiendo 400 palabras por página, calculamos que el corpus completo impreso en letra pequeña sobre papel delgado requeriría unos diez metros de espacio en una estantería. Leer todo el corpus

> en voz alta a velocidad rápida, 150 palabras por minuto, durante ocho horas al día, 365 días al año, llevaría unos cuatro años (trad. a.).

Por otra parte, la variedad de inglés americano contemporáneo está fielmente representada en el Corpus of Contemporary American English (en adelante, COCA), desarrollado por Mark Davis en colaboración con la Brigham Young University. Contiene 450 millones de palabras, con una proporción considerable de muestras de inglés hablado, y es uno de los pocos corpus en actualización constante. Sus fuentes y representación se detallan en la Tabla I.2.

Tabla I.2. Composición del Corpus of Contemporary American English.

GÉNERO	TAMAÑO	FUENTES
Discurso oral	~ 95 mill.	Transcripciones de conversación espontánea de más de 150 programas de TV y radio.
Ficción	~ 90 mill.	Relatos y obras de revistas literarias, revistas infantiles, libros editados a partir de 1990 y guiones cinematográficos.
Revistas	~ 95 mill.	Casi 100 revistas de temas como noticias, salud, decoración, mujer, finanzas, religión, deporte, etc.
Periódicos	~ 92 mill.	10 periódicos estadounidenses (USA Today, New York Times, etc.) Se incluyen distintas secciones como noticias locales, opinión o deportes.
Revistas académicas	~ 91 mill.	Casi 100 *journals* con revisión por pares. Se seleccionaron para cubrir el rango completo de la clasificación de la Biblioteca del Congreso de EEUU, es decir, un porcentaje de b (filosofía, psicología y religión), d (historia), k (educación), t (tecnología), etc.

Adaptado de: http://corpus.byu.edu/coca/help/texts_e.asp

Entre otros corpus relevantes queremos destacar los conocidos como Brown Corpus y Bank of English. El primero, cuyo nombre completo es *The Brown University Standard Corpus of Present-Day Edited American English* se publicó en 1964 por Francis y Kučera. Fue el corpus de referencia durante tres décadas hasta ser desplazado por el BNC. En comparación, hoy se considera reducido —solo 1 millón de palabras—

y ligeramente anticuado, pero su diseño marcó las directrices y la interfaz en las que se basan los corpus actuales. Por otra parte, John Sinclair, a quien se considera el investigador más influyente de la lingüística computacional, fue el inductor del ambicioso *Harper Collins COBUILD Bank of English*. Este corpus contiene más de 500 millones de palabras y sigue en expansión bajo el auspicio de la editorial Collins y la Universidad de Birmingham. La mayoría de sus muestras de lengua oral son transcripciones de películas o espacios televisivos bajo guión, apenas hay presencia de habla espontánea. A pesar de ser un corpus muy valioso, no es muy habitual en las publicaciones académicas debido a que su suscripción tiene un precio desorbitado. Se asume generalmente que es un proyecto diseñado principalmente para uso interno, es decir, lexicógrafos y creadores de materiales de la editorial Collins (Schmitt, 2010).

La razón por la que en los corpus grandes la proporción de lengua oral es relativamente baja es porque grabar, transcribir y codificar las muestras orales es mucho más exigente en tiempo y costes que procesar textos escritos. Existen corpus que contienen únicamente lenguaje oral, pero son mucho más reducidos. Uno de los más conocidos de inglés hablado es el *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE), que contiene 5 millones de palabras cuya codificación requirió ocho años de trabajo. Para ponerlo en perspectiva, los 90 millones de palabras procedentes de textos del BNC se compilaron en solo tres años.

Es frecuente que los corpus orales estén muy especializados en un tipo de habla muy concreto, ya sea un grupo demográfico, un área geográfica acotada o un contexto específico. Algunos ejemplos en inglés son el COLT, que contiene muestras del habla adolescente de Londres; el MICASE, grabaciones de discursos académicos en la Universidad de Michigan; o el WSC, de lengua oral de Nueva Zelanda. Hay otros con un grado de especialización incluso más alto, como el Basque Spoken Corpus, de muestras de habla espontánea en euskera o el CHIEDE, corpus de habla infantil espontánea del español (Garrote, 2010).

4.3. Las listas de frecuencia basadas en corpus

4.3.1. Listas de referencia de inglés general

Como ya hemos mencionado, las primeras listas de vocabulario clave con fines docentes se elaboraron sin medios tecnológicos, en lo que se conoce como «era pre-electrónica». Además de la *General Service List* de West[16], otra de las pioneras fue la llamada «lista de Ogden», publicada en 1935 en el libro *Basic English* de Charles K. Ogden. El término BASIC correspondía a las siglas del tipo de inglés utilizado: *British American Scientific International Commercial* (Ogden & Graham, 1930).

Los 850 términos de la lista se presentan en orden alfabético divididos en grupos semánticos. La primera categoría, *operations*, contiene 100 términos entre verbos, pronombres y partículas. La categoría *things* abarca 600 sustantivos y, por último, *qualities* tiene 150 adjetivos. A continuación se muestran algunas de las unidades pertenecientes a cada categoría de la lista de Ogden (1935) según la reedición de Brighman, (2005).

OPERATIONS – 100 WORDS

come, get, give, go, keep, let, make, put, seem, take, be, do, have, say, see

[…] quite, so, very, tomorrow, yesterday, north, south, east, west, please, yes

THINGS - 400 GENERAL WORDS

account, act, addition, adjustment, advertisement, agreement, air, amount,

[…] wind, wine, winter, woman, wood, wool, word, work, wound, writing , year

THINGS - 200 PICTURABLE WORDS

angle, ant, apple, arch, arm, army, baby, bag, ball, band, basin, basket, bath, bed,

[…] umbrella, wall, watch, wheel, whip, whistle, window, wing, wire, worm

QUALITIES - 100 GENERAL

able, acid, angry, automatic, beautiful, black, boiling, bright, broken, brown, cheap,

[…] thick, tight, tired, true, violent, waiting, warm, wet, wide, wise, yellow, young

[16] Howatt (citado por Schmitt, 2010) señala que aunque la GSL suele ir ligada al nombre de Michael West, el lingüista no merece el reconocimiento exclusivo. Esta lista fue la culminación de un proyecto en el que también tomaron parte Harold Palmer, Lawrence Faucett, Irving Lorge y Edward Thorndike.

QUALITIES - 50 OPPOSITES

awake, bad, bent, bitter, blue, certain, cold, complete, cruel, dark, dead, dear,

[…] short, shut, simple, slow, small, soft, solid, special, strange, thin, white, wrong

La lingüística de corpus moderna ha permitido elaborar listas de frecuencia con un enorme grado de precisión y fiabilidad, aunque Schmitt nos advierte de que «a frequency count is only as good as the corpus it is based upon[17]» (2010: 67). Entre las listas más prestigiosas se encuentran las cinco siguientes:

a) La versión actualizada de la *General Service List* adaptada en 1995 por Bauman y Culligan, que añade 284 palabras adicionales a las 2.000 de la lista original e incluye la frecuencia relativa de cada una de ellas en el Brown Corpus.
Disponible en: http://www.auburn.edu/~nunnath/engl6240/wlistgen.html

b) The Longman Defining Dictionary (LDOCE), compilada por Richard Kennaway y basada en diccionario Longman (1988). Contiene 2197 palabras, 10 prefijos y 39 sufijos.
Disponible en: www2.cmp.uea.ac.uk/~jrk/conlang.dir/LongmanVocab.html

c) La lista de frecuencia anotada de Adam Kilgarriff (1995). Replica sobre el BNC el proceso utilizado para crear el LDOCE de Longman. Contiene todas las palabras que tienen más de 800 ocurrencias en el BNC y se puede descargar en datos brutos o como lista lematizada.
Disponible en:http://www.kilgarriff.co.uk/bnc-readme.html

d) Word Frequencies in Written Spoken English (Leech, Rayson, & Wilson, 2001) . Su página web contiene la lista de frecuencia completa del BNC así como una gran cantidad de sublistas, tales como el corpus oral frente al escrito, categorías gramaticales específicas —solo verbos, nombres, o incluso determinantes o

[17] «Un recuento de frecuencia es solamente tan bueno como el corpus en el que esté basado» (trad.a.).

interjecciones— y géneros —escritura imaginativa frente a formal—, por nombrar algunas.
Disponible en: http://ucrel.lancs.ac.uk/bncfreq/flists.html

e) La lista de frecuencia de Paul Nation, creada con fines docentes sobre el BNC, el COCA y un corpus propio. Contiene 29 listas de 1.000 lemas acompañados de sus palabras derivadas. Esta lista es de especial relevancia para el presente trabajo y se describe con detalle en el punto 3 de la Parte segunda.
Disponible en: http://www.victoria.ac.nz/lals/about/staff/paul-nation

4.3.2. Listas de términos específicos

Una de las aplicaciones de las listas de frecuencia es emplearlas como datos de referencia para elaborar listas de vocabulario especializado. Una lista de frecuencia basada en un corpus de inglés general aporta datos sobre la esperanza matemática de las ocurrencias de cada palabra en un texto medio. Al compararla con un corpus de textos que giran en torno a un tema muy concreto podemos descubrir patrones que solamente ocurren en el corpus específico.

Hay dos procedimientos para elaborar listas de vocabulario sobre un área concreta. El más sencillo es seleccionar las de mayor frecuencia del corpus específico utilizando una *stop-list*, es decir, una lista de términos que el software debe ignorar en los recuentos. Lo más habitual es utilizar como *stop-list* las 2.000 palabras de mayor frecuencia del corpus primario. A partir de este nivel las palabras no se suelen repetir, por lo que las que tengan muchas ocurrencias serán previsiblemente términos clave en ese ámbito específico. El segundo procedimiento es algo más complejo y consiste en comparar el número de ocurrencias de cada palabra del corpus específico con su proporción esperada en un texto medio según el corpus primario. Esto nos permite analizar cuáles son las palabras que aparecen en un tema específico mucho más a menudo que en el corpus de inglés general sin el riesgo de eliminar términos de las 2.000 primeras palabras que puedan tener una acepción concreta en el inglés técnico del área.

Una de las listas de fines específicos más valoradas es la *Academic Word List* (en adelante, AWL) elaborada por Averil Coxhead en 1998 para su disertación de Máster y publicada dos años después en TESOL Quarterly. La AWL está elaborada siguiendo la primera técnica y utiliza como stop-list la *General Service List*. El resultado final son 570 palabras que suponen el 10% de los tokens de textos académicos pero solamente el 1,4% de un corpus de ficción de igual tamaño (Coxhead, 2000). En la última década, la mayoría de los libros de texto con fines académicos utilizan como referencia principal la AWL, que ha reemplazado por completo a su antecesora de 1984, la *University Word List* de Xue y Nation (Lessard-Clouston, 2012).

La Tabla I.3 muestra otras listas recientes elaboradas para la docencia del inglés técnico o para fines específicos. Se puedan encontrar fácilmente listas de vocabulario con fines docentes para casi cualquier disciplina, desde la medicina hasta la agricultura. Ante esta perspectiva resulta sorprendente que no existan listas especializadas para alumnos con una L1 determinada.

Tabla I.3. Listas de frecuencia de vocabulario específico con fines docentes

LISTA Y AUTOR	DESCRIPCIÓN
Business Word List (BWL1) (Konstantakis, 2007).	560 palabras que aparecen al menos 10 veces en el Business English Published Materials Corpus.
Science Word List (SWL) (Coxhead & Hirsh, 2007)	318 palabras que representan casi el 4% de un corpus de 1,5 millones de elementos de 14 disciplinas científicas, dividida en 6 sublistas. Está diseñada para estudiantes universitarios.
AgroCorpus List (Martínez, Beck, & Panza, 2009)	92 palabras de la AWL que tienen alta frecuencia en un corpus de artículos de investigación agrícola. Destacan las palabras que pueden ser de uso general pero tienen un sentido técnico, como *accumulation* o *region*.
Basic Engineering List (BEL) (Ward, 2009)	Aproximadamente 300 palabras de un corpus compuesto por los libros de texto más relevantes en cinco campos de la ingeniería. No utiliza como *stop-list* ni la GSL ni la AWL.
Newspaper Word List (NWL) (Chung, 2009)	588 palabras de alta frecuencia extraidas de un corpus de periódicos. Excluye nombres propios y las palabras de la GSL, salvo que tengan una gran presencia en el corpus.
Theological Word List (TWL) (Lessard-Clouston, 2010)	100 palabras obtenidas de conferencias académicas sobre teología. Dividida en dos sublistas de frecuencia.

Adaptado de Lessard-Clouston (2012)

Tomemos como ejemplo de ello la prestigiosa AWL de Coxhead. Tras un análisis contrastivo, vemos que entre sus 570 palabras hay 425 con raíz latina. Esto implica que tres de cada cuatro palabras de la AWL tienen muchas posibilidades de ser deducibles, incluso descontextualizadas, por el alumno que domine una lengua romance. En el caso concreto de un hispanohablante, la Tabla I.4 nos muestra que 16 de las 20 primeras palabras de la AWL serían fácilmente deducibles por cualquier hablante de español con conocimientos básicos de inglés.

En definitiva, la lingüística de corpus ha demostrado ser una herramienta extremadamente útil para elaborar listados de palabras clave con fines docentes, ya sean de inglés general o con fines específicos. Destaca, sin embargo, que la investigación en este ámbito haya pasado por alto un aspecto tan importante en el aprendizaje del vocabulario como la influencia de la L1.

Tabla I.4. Palabras deducibles entre las 20 primeras de la AWL

ORDEN	PALABRA	DEDUCIBLE
1	ABANDON	✓
2	ABSTRACT	✓
3	ACADEMY	✓
4	ACCESS	✓
5	ACCOMMODATE	✓
6	ACCOMPANY	✓
7	ACCUMULATE	✓
8	ACCURATE	
9	ACHIEVE	
10	ACKNOWLEDGE	
11	ACQUIRE	✓
12	ADAPT	✓
13	ADEQUATE	✓
14	ADJACENT	✓
15	ADJUST	✓
16	ADMINISTRATE	✓
17	ADULT	✓
18	ADVOCATE	
19	AFFECT	✓
20	AGGREGATE	✓

5. LA INFLUENCIA DEL ESPAÑOL COMO L1 EN EL APRENDIZAJE DEL INGLÉS

El idioma inglés tiene en todos sus registros un número considerable de palabras formadas a partir de raíces latinas que también están presentes en el español actual. El *Big Red Book of Spanish Cognates* contiene 14.000 raíces comunes en inglés y español, y el *Dictionary of Spanish Cognates* contempla 20.000 palabras similares seleccionadas en función de su frecuencia (Morán, 2011a; Thomas & Gaby, 2005). Estos números indican que el aprendiz hispanohablante de inglés cuenta desde el comienzo con una cantidad considerable de vocabulario potencial en L2. Estas unidades léxicas que presentan semejanzas formales y semánticas con su equivalente en L1 se denominan cognados.

5.1. Los cognados y su uso en la enseñanza de L2

La palabra *cognado* está formada por *co* + *gnatus*, es decir, 'nacidos juntos'. Se refiere a términos hermanos que provienen de una misma palabra y han seguido caminos distintos en su evolución. El estudio de los cognados es uno de los pilares básicos en lingüística histórica, ya que estos aportan indicios para reconstruir la ruta de aquellas palabras que evolucionaron de manera distinta a partir de un término común. Podemos citar como ejemplo el par *cadera* y *cátedra*, que derivan de mismo término latino *cathedra*[18]. Uno de los fenómenos más interesantes para el estudio sobre cognados es cuando aparecen en lenguas distintas, ya sea porque que han formado su léxico a partir de un ancestro común, o porque en algún momento han incorporado préstamos provenientes de otras lenguas.

El estudio de los cognados interlingüísticos es uno de los centros de interés del Análisis Contrastivo desarrollado en los años 50 a partir de las teorías de Lado basadas en el conductismo. Para Lado, la influencia de la L1 es siempre perjudicial para el

[18] *Cathedra* a su vez es un préstamo del griego *kathédra*, que significa 'silla'. En la Roma antigua el asiento de los profesores era una *cathedra* (butaca con respaldo) mientras que los pupilos se sentaban en un banco o taburete llamado *subsellium*. La evolución culta de la palabra es *cátedra*, que ha pasado de designar el lugar donde se sentaban los profesores para referirse al rango, *catedráticos*. Por otro lado, la evolución vulgar, *cadera*, se transformó por metonimia —por cercanía física—desde su concepto *silla* hasta su significado actual (Anders, 2011).

aprendizaje de una L2 y prácticamente todos los errores de los aprendices se deben a la influencia de los hábitos adquiridos por el aprendiz en su L1. El argumento principal es que los estímulos en L2 activan respuestas que se corresponden con viejos hábitos que inevitablemente serán una fuente de problemas. Los estudios del Análisis Contrastivo comparan las diferencias entre L1 y L2 con el objetivo principal de detectar y sistematizar errores (Buehler, 1995; Izquierdo, 2003).

En el caso concreto del vocabulario, el Análisis Contrastivo se centra fundamentalmente en analizar los errores a los que podrían inducir las palabras cuya aparente semejanza formal no se corresponde con una equivalencia semántica, es decir, los falsos cognados (ej: *carpet ≠ carpeta*). Esto es lo que se conoce como *transferencia negativa*. Sin embargo, la investigación ha puesto de manifiesto que también existe una transferencia positiva que puede ser de gran ayuda en el aprendizaje. El término neutro para referirse a ambos fenómenos es *influencia interlingüística*, que depende fundamentalmente del grado de hermanamiento que haya entre ambas lenguas (Carroll, 1992; Durán Escribano, 2004; Schmitt, 2010).

La autoridad de la escuela conductista hizo que durante muchos años se considerase que la influencia de la L1 era siempre negativa en el aprendizaje de la L2, por lo que potenciar los mecanismos de aprendizaje basados en técnicas de comparación interlingüística consciente ha sido una práctica docente denostada durante varias décadas. En cuanto a los factores que explican esto, la investigación sugiere que el método comunicativo no solo ha sido responsable del descrédito de la enseñanza explícita del vocabulario, sino también del rechazo a considerar los conocimientos en L1 como una condición ventajosa para el aprendizaje de la L2. Tal y como afirma Izquierdo:

> La importancia concedida a los objetivos de orden pragmático en el método comunicativo ha relegado a un segundo plano las consideraciones de orden metalingüístico, es decir, la dimensión lingüística comparativa (Dabène, 1995: 138). En la enseñanza de lenguas extranjeras se han marginado, pues, durante muchos años los conocimientos previos de la LM [lengua materna] del alumno y no se han utilizado como instrumento facilitador para el aprendizaje de la LE [lengua extranjera] (Izquierdo, 2003: 195).

Como consecuencia de ello, durante las últimas décadas se ha formado a los docentes en la idea de que se debe reprimir el uso de la L1 en la enseñanza de la L2. El vocabulario se debe aprender únicamente a través de recursos en los que no intervenga la L1 y se fomentará para ello el uso de diccionarios monolingües, apoyos visuales o mapas conceptuales. En una situación de aprendizaje en la que un alumno pregunte el significado de un término que ha encontrado en un texto, se espera que un «buen profesor» proporcione la explicación en inglés en lugar de su traducción. En general, cualquier uso de la L1 por parte del docente será considerado mala praxis o, cuanto menos, de una dejadez cuestionable. Sin embargo, el profesor que sigue todas estas directrices comprobará con frustración cómo, a pesar de sus esfuerzos, la tendencia natural del alumno será anotar la palabra inglesa junto con su significado en L1 en lugar de escribir un sinónimo o una explicación en la lengua meta. A medida que se avanza en el dominio de un idioma, la tendencia del aprendiz es utilizar fuentes directas en L2. En este sentido, Durán (2004) apunta que la estrategia natural de aprendizaje en las etapas elementales es recurrir a la lengua materna y transferir conocimientos; sin embargo, ha habido una tendencia represiva hacia el uso de la L1 que va en contra de los recursos naturales de aprendizaje.

Esta es la razón por la que son muy pocos los estudiantes que ven las palabras de su propio idioma como vehículo para el aprendizaje del inglés (Hancin-Bhatt & Nagy, 1994). A pesar de que es evidente que hay palabras que mantienen una gran semejanza formal y semántica, la investigación sugiere que, a menos que se enseñe de manera explícita, los alumnos no son conscientes del enorme potencial que tienen para ellos los cognados. Son diversos los estudios que demuestran que el entrenamiento para reconocer cognados utilizando pistas interlingüísticas es altamente beneficioso tanto para mejorar el proceso del léxico en L2 como para enriquecer el vocabulario y mejorar la capacidad de lectura en L2 (Albrechtsen, Haastrup, & Henriksen, 2008; Dressler, Carlo, Snow, August, & White, 2011; Haastrup, 1991).

En el caso del inglés y el español – y posiblemente, en todas las lenguas romances – la prevención de utilizar la L1 solo por recelo a la transferencia negativa de falsos cognados no tiene sentido desde el punto de vista estadístico (véase Figura I.2. y Tabla I.5). Observamos que en las distintas listas de vocabulario de alta frecuencia el porcentaje de cognados verdaderos es muy superior al de falsos cognados. Sus medianas

están situadas en 40,8% frente a 4,9%. Ante estas cifras, es necesario cuestionarse la eficacia de aplicar en la docencia los postulados de Lado, que se centran únicamente en una minoría problemática y desaprovechan el gran potencial de palabras cognadas que los alumnos pueden asimilar sin gran esfuerzo cognitivo, especialmente en algunos registros.

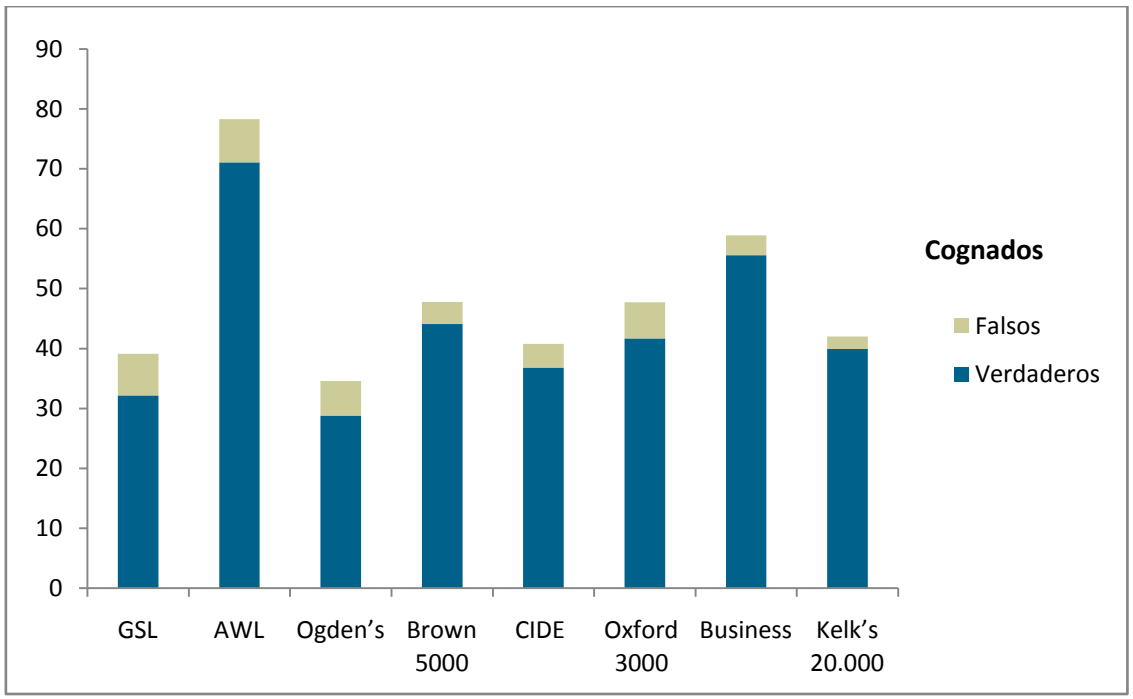


Figura I.2. Porcentaje de cognados verdaderos y falsos en listas de vocabulario relevantes.

Tabla I.5. Presencia de cognados en las listas de vocabulario relevantes.

	TOKENS	COGNADOS. VERDADEROS	%	COGNADOS. FALSOS	%
General Service List	2.284	736	32,2	158	6,9
Academic Word List	570	405	71,1	41	7,2
Ogden's Basic English Word List	850	245	28,8	49	5,8
Brown 5000 (Francis and Kucera, 1964)	5.000	2.207	44,1	184	3,7
CIDE Defining Vocabulary	3.732	1.375	36,8	148	4,0
Oxford 3000 Word List (Oxford UP)	3.457	1.441	41,7	207	6,0
Oxford UP Business and Finance Words	270	150	55,6	9	3,3
Kelk's UK English Word List (20.000)	20.833	8.316	39,9	431	2,1

Adaptado de Cognates.org

Un dato interesante que se aprecia en la Tabla I.5 es que en la lista de Ogden y en la *General Service List* hay una menor presencia de cognados que en el resto. Esto es debido a que son listas que contienen muy pocos elementos y todos de muy alta frecuencia, un rango en el que las palabras de raíz latina no reemplazaron a las de origen germánico. En cambio, en los registros cultos como la *Academic Word List* y la *Business and Finance Words* se ve claramente que la influencia del latín es mucho mayor. En este sentido, Meara (1996) señala que a pesar de que en el vocabulario básico no hay muchos cognados inglés-español, los hispanohablantes tienen una gran cantidad de vocabulario latente en el rango de las palabras de baja frecuencia. Esto implica que su habilidad para adquirir palabras nuevas aumenta drásticamente a medida que sube su nivel general de inglés.

Los postulados del Análisis Contrastivo dan un giro en los años 70 con la Teoría de la Interlengua (Selinker, 1972), «que supone un giro metodológico al estudiar y analizar tanto las producciones desviadas como las correctas, es decir, la producción total de los estudiantes, considerando que tanto unas como otras son relevantes en el proceso de aprendizaje» (Alexopoulou, 2005: 86). Según esta teoría, el aprendiz saca sus propias generalizaciones —en ocasiones excesivas— del sistema lingüístico de la L2 a partir de la L1. Mediante estos modelos es capaz de establecer sus propias hipótesis sobre palabras desconocidas si tienen cierta similitud. Eric Kellerman es uno de los primeros investigadores en aportar datos empíricos sobre esta teoría en *The empirical evidence for the influence of the L1 in interlanguage* (1984).

La investigación no deja lugar a dudas de que la L1 ejerce una influencia considerable en la L2, ya sea beneficiosa o perjudicial. Los estudios en psicolingüística demuestran que la L1 está activa durante el proceso léxico de la L2 tanto en alumnos principiantes como avanzados (Hall, 2002; Kroll & Sunderman, 2003). En el caso concreto del vocabulario, el reto es desarrollar técnicas docentes eficaces para abordar los problemas derivados de los falsos cognados y sacar partido de la ventaja que suponen los cognados verdaderos.

En este sentido, Granger sostiene que:

> Cognates are both an aid and a barrier to successful L2 vocabulary development. Teachers should therefore seek to find a happy medium between over-reliance on cognates and near-pathological mistrust of them, two attitudes which are equally detrimental to learners' vocabulary development (Granger, 1993, citado por Schmitt, 2010: 74).

> Los cognados son tanto una ayuda como una barrera para el desarrollo correcto del vocabulario en L2. Los profesores, por tanto, deben buscar un término medio entre la confianza excesiva en los cognados frente a una desconfianza casi patológica, dos actitudes que son igualmente perjudiciales para el desarrollo del vocabulario de los estudiantes.

Una de las dificultades que encontramos es que en la literatura no hay consenso en las características de un falso cognado. La acepción habitual es que dos palabras son falsos cognados si mantienen semejanza formal pero su significado es completamente distinto, ya sea porque los términos han seguido caminos de evolución muy diferentes o bien porque la aparente similitud entre el par se debe al mero azar, no a compartir un ancestro común (Morán, 2011a). Otra corriente más conservadora considera que, aunque comparta el significado principal, un cognado es falso si alguna de sus acepciones o de sus usos se corresponde con su equivalente en español (Prado, 2001).

Por ejemplo, *relative* puede ser un cognado falso o verdadero, dependiendo del contexto. La búsqueda del término *relative* en el British National Corpus nos ofrece estos dos ejemplos:

*Pope et al. (2006) reported the **relative** risk estimates of air pollution.* ('relativo')
*By Friday, **relatives** of those who perished were at the local cemetery*. ('familiar')

La primera escuela consideraría *relative* como cognado verdadero porque el sentido 'relativo' tiene muchas más ocurrencias que en su significado secundario, 'familiar'. Para otros, en cambio, *relative* sería un falso cognado porque tiene una

acepción secundaria que no se corresponde con 'relativo'. Evidentemente, para ciertas disciplinas lingüísticas, como puede ser el desarrollo de software para la traducción automática, es de gran relevancia contemplar todas las acepciones que puede tener una palabra. En la enseñanza, sin embargo, la polisemia se asume como un problema menor que, evidentemente, no es un fenómeno exclusivo de los cognados. La tendencia es simplemente presentar a los alumnos el significado más frecuente en los primeros niveles e introducir las acepciones secundarias de forma gradual. No es comprensible, por tanto, la razón por la cual los materiales docentes han tratado la polisemia en los cognados con una preocupación y celo excesivos, dando la impresión de que el alumno tiene que aprender absolutamente todos los sentidos de la palabra cognada, dominio que no se exige para las que no lo son. Sobre este asunto, Morán (2011a) señala lo siguiente:

> The word *good* is one of the first terms to be learnt; however, there is no analysis regarding the differences in meaning between *good morning*, *very good* and *this is for good*. Our claim is that cognates in general will always retain their core common meanings, although they sometimes have accessory or unrelated meanings and usages. A second, third or fourth additional meaning or usage of a word is not a reason to disregard its core meaning. WordNet 2.1, by Princeton University, 2005, gives 21 senses to the adjective *good*; 4 senses to its noun form; and 2 senses to its adverb form. Despite all this, *good* will basically always be the opposite of *bad*.

> La palabra *good* es uno de los primeros términos que se aprenden; sin embargo, no se analizan las diferencias entre el significado de *good morning*, *very good* y *this is for good* ['buenos días', 'muy bien' y 'esto es para siempre']. Nuestro argumento es que los cognados generalmente conservarán siempre un significado principal común, aunque a veces tengan otros usos o significados no relacionados. Un segundo, tercer o cuarto significado o un uso adicional de una palabra no es razón para hacer caso omiso de su significado principal. WordNet 2.1, de la Universidad de Princeton, 2005, ofrece 21 sentidos para el adjetivo *good*; 4 sentidos como sustantivo y 2 sentidos como adverbio. A pesar de todo esto, *good* básicamente siempre será lo contrario de *bad* (trad. a.).

5.2. Psicolingüística e identificación de cognados

Los estudios en psicolingüística han tratado de arrojar luz sobre el proceso cognitivo por el cual el cerebro asocia una palabra en L2 a otra similar en L1. Carroll (1992) explica el reconocimiento de cognados a través del modelo Cohort de Marslen-Wilson, según el cual el lexicón mental está organizado en entradas léxicas que llevan asociadas un elemento de reconocimiento que se activa ante ciertos estímulos. Las entradas léxicas se agrupan en barrios (*neighborhoods*) que contienen elementos formalmente similares. En la lectura, los aductos (*input*) estimulan distintas entradas de la L1 y mediante el contexto y otras pistas morfosintácticas el proceso reduce las entradas hasta seleccionar solamente una – o ninguna – adecuada. Entre los factores que activan el reconocimiento Carrol destaca que no tienen por qué ser semánticamente idénticos y que siempre hay algún tipo de semejanza formal.

El reconocimiento de cognados se atribuye a una similitud entre códigos semánticos, fonéticos y ortográficos, pero la investigación es unánime en la conclusión de que la semejanza ortográfica es el factor dominante (Dressler et al., 2011; Nagy, García, Durgunoğlu, & Hancin-Bhatt, 1993; Schwartz, Kroll, & Diaz, 2007).

> Holmes (1986) found that students differed in what they considered to be a cognate. He identified a "cline of cognate-ness", ranging from pairs most students classed as cognates, such as *progreso* – *progress* (which share extensive orthographic and complete semantic overlap) to words which few subjects considered cognates, such as mito – myth (little orthographic but complete semantic overlap) (Dressler et al., 2011: 244).

> Holmes (1986) reveló que los estudiantes discreparon en lo que consideraban un cognado. Identificó un "gradiente de *cognacidad*", que abarca desde los pares que la mayoría de los estudiantes clasificó como cognados, tales como *progreso* – *progress* (que presentan gran similitud ortográfica y una superposición semántica completa) hasta palabras que pocos participantes consideraron cognados, como mito – myth (poca similitud ortográfica pero superposición semántica completa) (trad. a.).

Montelongo, Hernández y Herter (2009) denominan «transparencia» al grado de similitud ortográfica entre cognados en una escala de 7 puntos en la que los cognados exactos, tales como *natural/*natural, tienen el grado más alto. Sus resultados sobre estimación subjetiva de la transparencia apuntan, como es de esperar, que a mayor divergencia ortográfica, mayor dificultad en el reconocimiento[19].

Diversos estudios en psicolingüística sugieren que reconocer cognados es una cualidad innata. El experimento de Méndez Pérez, Peña y Bedore (2010) reveló que los niños desde la etapa pre-escolar son capaces de reconocer cognados por transferencia de la L1 a la L2. El análisis se hizo mediante el ejercicio de imágenes del Test of Language Development-Primary:3 (TOLD-P:3). En cuanto a los adultos, el estudio de Friel y Kennison (2001) concluye que los adultos monolingües son capaces de reconocer y traducir correctamente el significado de palabras cognadas con su L1en una lengua desconocida.

Aunque parece ser una cualidad que se desarrolla sin esfuerzo en las primeras etapas del lenguaje, la formación explícita en estrategias de reconocimiento de cognados parece aumentar notablemente la habilidad del alumno. En el estudio de Bravo, Hiebert y Pearson (2007) los participantes subieron su puntuación en la identificación de cognados tras la instrucción explícita y transfirieron esta capacidad posteriormente a la comprensión lectora. Una de las técnicas más eficaces para el entrenamiento en el reconocimiento de cognados parece ser la reflexión metalingüística sobre los componentes morfológicos de las palabras en L1 y L2 (Kieffer & Lesaux, 2007).

5.3. La morfología en el reconocimiento de cognados

La «conciencia morfológica» (*morphological awareness*) es la habilidad para manipular morfemas y entender los componentes que forman las palabras complejas (Carlisle & Feldman, 1995). Esta conciencia morfológica parece transferirse de la L1 a la L2 de manera natural. Cuando el alumno comprende la equivalencia entre los afijos

[19] Otra de las conclusiones interesantes del estudio es que entre los cognados no evidentes la percepción sobre la transparencia es mayor cuando la terminación es distinta que cuando el comienzo es distinto (*yacht*-yate se percibe "más transparente" que *sugar*-azúcar).

derivativos, los conocimientos que tiene sobre el comportamiento y la formación de los adverbios, gerundios y adjetivos en L1 le permiten comprender el significado y los principios básicos del uso de palabras nuevas en la L2.

*probab**ly*** = probable**mente**	normal**ly** = normal**mente**
*abandon**ing*** = abandon**ando**	*consider**ing*** = consider**ando**
contin***uous*** = contin**uo**	*superfl**uous*** **=** superfl**uo**

Esta conciencia morfológica adquirida en la L1 parece mejorar considerablemente la comprensión lectora en la L2 (Ramírez, Chen, Geva, & Kiefer, 2010; Ramírez, Chen, & Pasquarella, 2013). El estudio de Ramírez et al. (2013) apunta a que la complejidad del sistema morfológico español puede ser una ventaja a la hora de procesar las estructuras en inglés:

> Spanish-speaking ELLs [English Language Learners] can use morphological skills developed in their first language to facilitate vocabulary development in their second language. Because Spanish has a more complex derivational system than English, Spanish speaking ELLs likely develop a heightened sensitivity to morphemes and morphological structures through exposure to Spanish. This sensitivity, in turn, enables them to analyze English words and acquire English vocabulary (Ramírez et al., 2013).

> Los aprendices de inglés hispanohablantes pueden usar habilidades morfológicas desarrolladas en su primer idioma para facilitar el desarrollo del vocabulario en su segunda lengua. Dado que el español tiene un sistema derivativo más complejo que el inglés, los aprendices hispanohablantes son más propensos a desarrollar una mayor sensibilidad a los morfemas y estructuras morfológicas a través de la exposición al español. Esta sensibilidad, a su vez, les permite analizar palabras inglesas y adquirir vocabulario (trad. a.).

De hecho, los estudios demuestran que los conocimientos sobre derivación ayudan a los alumnos a identificar palabras cognadas. Mediante el análisis morfológico, los alumnos parecen identificar mejor cuál es la raíz de la palabra y su categoría gramatical, especialmente si se trata de una palabra cognada (Hancin-Bhatt et al., 1994; Nagy, Berninger, & Abbott, 2006).

En 2005 se lanzó en California el programa CSI (*Cognate Strategy Instruction*) cuyo objetivo era enseñar estrategias de reconocimiento de cognados y procedimientos cognitivos para facilitar la transferencia positiva español-inglés. El CSI pretendía determinar si esta instrucción incrementaría el vocabulario y la capacidad para entender textos en inglés. Los participantes del estudio fueron 70 niños cuya L1 era el español dirigidos por 3 profesores.

La primera fase consistió en revisar los libros de ciencias naturales y sociales, entre los que se encontraron cientos de cognados verdaderos y solo unos pocos ejemplos de falsos cognados. La segunda fase del programa se fundamentaba en la enseñanza de técnicas para la identificación de cognados, con especial énfasis en la traducción de los afijos derivativos más frecuentes. Los profesores que participaban en el estudio remarcaron que el programa no benefició únicamente a los alumnos bilingües, sino que los niños monolingües también desarrollaron cierta conciencia morfológica que les permitía expandir su vocabulario a través del reconocimiento de raíces y afijos. Los resultados globales del CSI, según la evaluación de Lubliner y Grisham (2012), fueron excepcionales y en pruebas posteriores en el ámbito de la educación universitaria, los alumnos remarcaron la utilidad y versatilidad de la estrategia.

En los últimos años se observa que algunas editoriales de libros de texto de reconocido prestigio, tales como Longman, Cambridge y Oxford, empiezan a dar pequeños pasos hacia el reconocimiento de la influencia de la L1 en el aprendizaje, especialmente en los niveles elementales. Algunos libros incluyen un pequeño apéndice con el resumen gramatical en que el que, ocasionalmente, se aportan explicaciones sobre las diferencias gramaticales interlingüísticas L1/L2 que pueden dar lugar a error. En este sentido parece que la influencia de la escuela conductista sigue presente, ya que no aprovechan las similitudes, sino que advierten únicamente de los errores que pueden

producir las diferencias. A pesar de ello, el programa léxico no parece variar entre las ediciones internacionales.

A modo de conclusión, la investigación sugiere que tanto las programaciones didácticas como los programas de formación de profesores de inglés deberían incluir instrucción explícita sobre la importancia de los cognados, para así aprovechar la ventaja del conocimiento potencial que traen los alumnos. Son muchos los expertos en adquisición de L2 que defienden la importancia de los cognados a la hora de seleccionar los elementos para una programación léxica si los alumnos comparten una misma lengua materna (Folse, 2004; Koda, 2005; Laufer, 1990; O'Dell, 1997; Proctor, Carlo, August, & Snow, 2005; Ringbom, 2007; Schmitt, 2010).

La reflexión metalingüística y la toma de conciencia sobre las semejanzas y diferencias entre la L1 y la L2 se pueden englobar dentro de la competencia básica «aprender a aprender».

SEGUNDA PARTE

LA LISTA DE VOCABULARIO PRIORITARIO

II

1. ENFOQUE DEL ESTUDIO EMPÍRICO

El objetivo de este estudio es establecer unos objetivos de aprendizaje de vocabulario orientados a que el alumno pueda leer textos auténticos de manera autónoma lo antes posible. Para determinar cuáles son estos objetivos, en primer lugar se debe analizar cuánto vocabulario desconocido se puede tolerar en un texto sin que interfiera con la comprensión. A partir de este dato, podremos calcular el número mínimo de palabras de alta frecuencia que, estadísticamente, otorgarían a un alumno la base léxica necesaria para leer textos garantizando estar dentro de los umbrales de la comprensión. Por último, si se analiza la influencia de la L1 del aprendiz sobre estos resultados, es posible crear una lista de vocabulario prioritario específicamente diseñada para hispanohablantes.

1.1. Fases del estudio y punto de vista

El plan de trabajo del presente estudio se dividió en dos fases: elaboración de un plan léxico y análisis de su eficacia.

La primera fase consistió en un estudio empírico de corpus del que se obtuvo una lista del vocabulario mínimo necesario para que un aprendiz hispanohablante pueda leer de manera autónoma textos ingleses auténticos. El criterio primordial para la selección del vocabulario clave fue su frecuencia relativa en los corpus lingüísticos más representativos del inglés británico y americano: British National Corpus y Corpus of Contemporary American English.

En cuanto al procedimiento, en primer lugar definimos el porcentaje del corpus que se pretendía cubrir y, partiendo de ese dato, seleccionamos el conjunto mínimo de palabras que, estadísticamente, alcanza la cobertura deseada. Posteriormente eliminamos las unidades léxicas no rentables, tales como los cognados transparentes inglés-español y otras palabras deducibles, para así obtener la lista de vocabulario clave que, en teoría, permitiría a un aprendiz hispanohablante entender un texto al menos hasta el umbral definido. Por último, se optimizó el listado para su uso como material docente, estableciendo distintos niveles de dificultad que atienden a criterios tales como la frecuencia de las palabras, su rentabilidad o la probabilidad de que una palabra pueda inducir a error (falsos cognados).

En la segunda fase se estudió la validez de este listado de vocabulario analizando su presencia y cobertura en muestras de textos de tres géneros de interés para los alumnos de L2. En primer lugar, se pretendía determinar si la lista de vocabulario clave proporciona el vocabulario suficiente para que un hispanohablante supere las pruebas de comprensión lectora de exámenes correspondientes a los niveles B1 y B2 del Marco europeo de referencia para las lenguas. En segundo lugar se analizó la lista sobre distintas obras literarias, tanto versiones originales como ediciones simplificadas para estudiantes. También se estudiaron muestras de habla oral no espontáneas de un corpus creado ad hoc con subtítulos de películas y series y discursos políticos.

En cada muestra se analizó el porcentaje acumulado de texto que pertenece a estas categorías: (a) palabras clave, (b) cognados transparentes, (c) nombres propios, (d)

vocabulario que no pertenece a ninguna de estas categorías, es decir, palabras presumiblemente desconocidas. Basándonos en esos datos, se calculó la cobertura respecto del umbral mínimo de comprensión.

1.2. Validación estadística del diseño del estudio

1.2.1. Valores iniciales

En primer lugar, definimos *umbral de autonomía* como el porcentaje mínimo del vocabulario de un texto que queremos abarcar. Fijamos este umbral[20] en el 95% y ordenamos en una tabla todas las palabras del corpus según su frecuencia. Si vamos seleccionando desde la más frecuente hacia abajo, llegaremos a un punto en el que la suma de las ocurrencias de las *n* primeras palabras alcanza el 95% del total de ocurrencias del corpus.

Podemos aproximar el valor de *n* por la ley de Zipf: si la palabra más frecuente aparece *x* veces en el corpus, la segunda aparecerá $x/2$; la tercera $x/3$, y así sucesivamente.

Si tomamos las *n* primeras palabras, obtendremos la siguiente serie:

$$x + \frac{x}{2} + \frac{x}{3} + \ldots + \frac{x}{n} = x\underbrace{\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)}$$

Con infinitos términos, esta sería la serie armónica $\sum_{k=1}^{\infty} \frac{1}{k}$

[20] Ver punto 3.2 de la Primera parte (Vocabulario necesario para la lectura en L2, p.25)

Si consideramos el n-ésimo número armónico, —en nuestro caso, si tomamos las *n* primeras palabras— tenemos

$$H_n = \sum_{k=1}^{n} \frac{1}{k}$$

y posteriormente aplicaríamos el álgebra de series a este número armónico para hallar cuántas palabras es necesario escoger para cubrir el 95% del texto.

Sobre un corpus que contiene N palabras, llamamos f_1 a la frecuencia de la palabra que aparece mayor número de veces, f_2 a la segunda, y, en general, f_n a la frecuencia de la palabra que ocupa el lugar n-ésimo, tenemos

$$\sum f(n, N) = N$$

La ley de Zipf indica que $f_n = \frac{1}{n^a}$ con $a \to 1$. En términos de probabilidad, diríamos que

$$P_n = \frac{f_n\ (n, N)}{N}$$

P_n denota la probabilidad de que aparezca una palabra partiendo de su frecuencia relativa. Obviamente,

$$\sum_n P_n = 1$$

Al sumar las palabras, la distribución de la probabilidad sería

$$\zeta = 1 + \left(\frac{1}{2}\right)^a + \left(\frac{1}{3}\right)^a + \ ... \ + \left(\frac{1}{n}\right)^a$$

Si la primera aparece *x* veces,

$$x^a + \left(\frac{x}{2}\right)^a + \left(\frac{x}{3}\right)^a + \ ... \ + \left(\frac{x}{n}\right)^a = \quad x^a \underbrace{\left(1 + \left(\frac{1}{2}\right)^a + \ ... + \left(\frac{1}{n}\right)^a\right)}_{\zeta\ (a)}$$

entonces obtenemos la función zeta de Reamann. Como en la ley de Zipf consideramos $a \rightarrow 1$, la función de Reamann deviene en la ley armónica. Dado que la serie armónica no converge y que la suma de las probabilidades ≤ 1, esta no es, en rigor, una distribución de probabilidad; sin embargo, es una aproximación suficiente para el estudio.

1.2.2. Procedimiento y algoritmo

Elegimos un corpus representativo, llamado A, que está formado por una colección de textos. Ordenamos sus palabras por índice de frecuencia y las agrupamos en familias léxicas en la lista ordenada B, que no tiene elementos repetidos. De ahí se crean los conjuntos `Cognates` (C), `Frequency` (F) `y Keywords` (K) según el esquema de la Figura II.1.

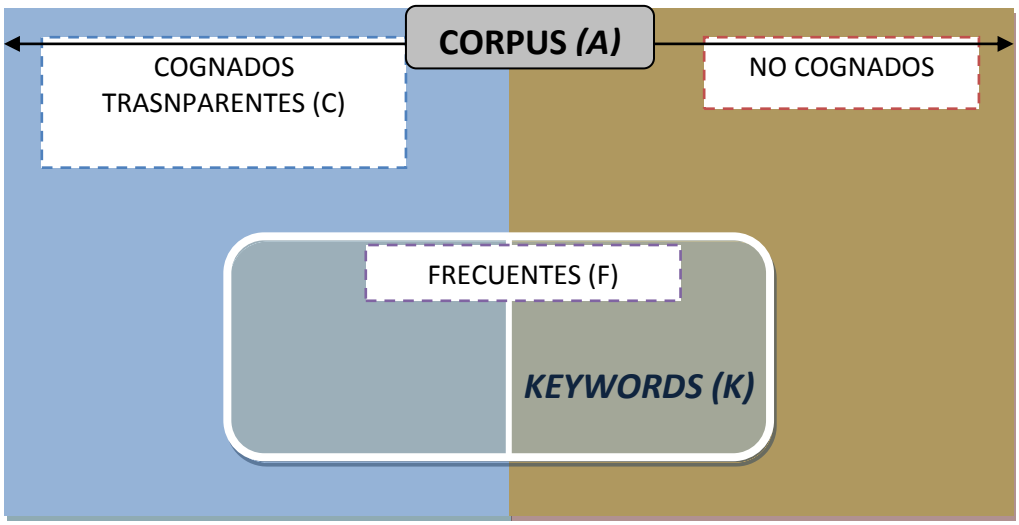


Figura II.1. Diagrama de los conjuntos Cognates, Frequency y Keywords

De la lista ordenada (B), por un lado seleccionamos los cognados transparentes inglés-español y se pasan a un archivo llamado `Cognates (C)`. Por otro lado, se escogen las *n* palabras más frecuentes del corpus y se pasan al archivo `Frequency (F)`. De ahí, las que no son cognadas se pasan a la lista `Keywords (K)`, que contiene el vocabulario prioritario.

Técnicamente, `Frequency` es una matriz [21] que contiene las *n* palabras más frecuentes de un corpus representativo hasta cubrir el 95% del total ocurrencias. El siguiente diagrama ilustra el proceso.

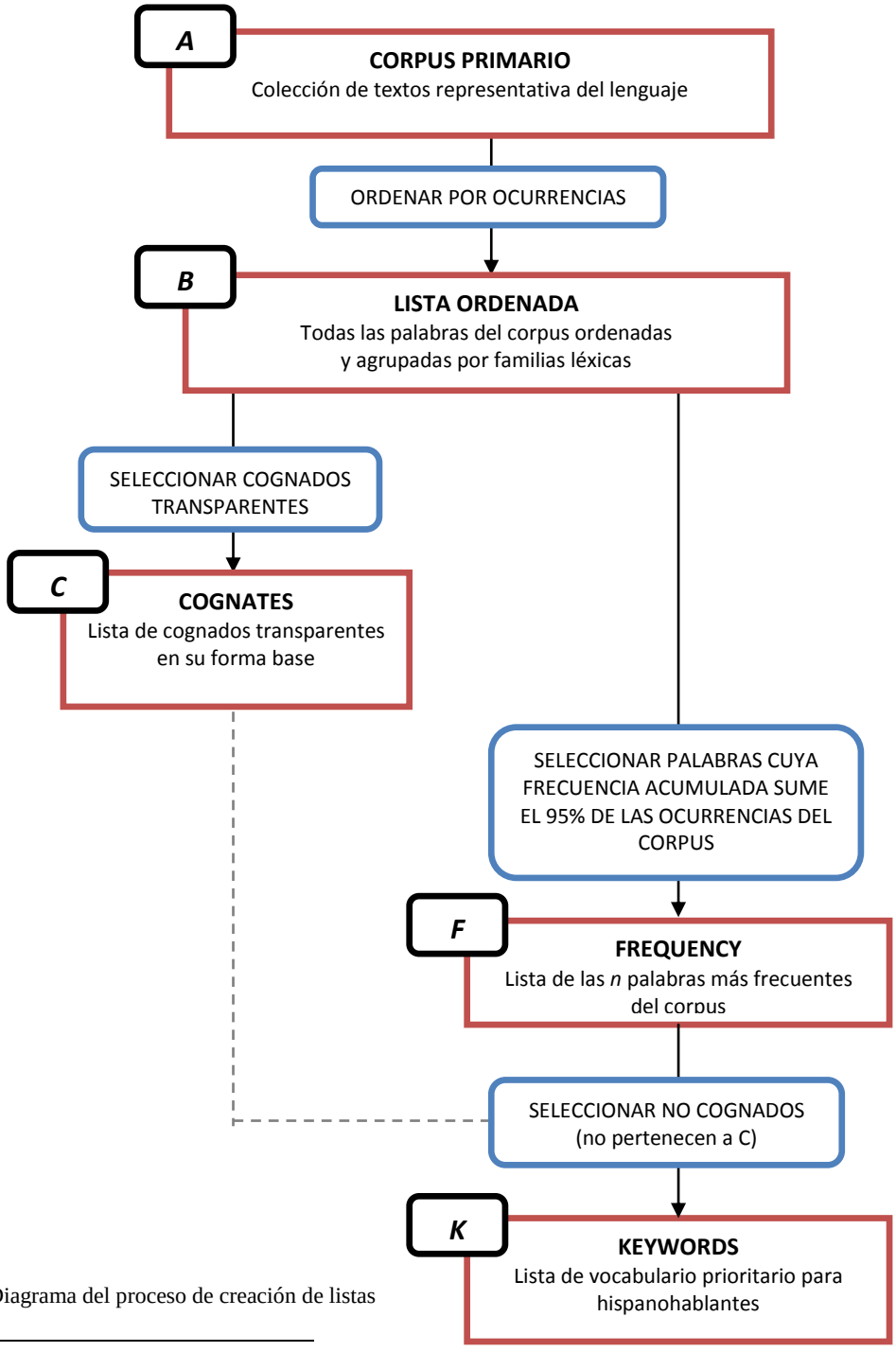


Figura II.2. Diagrama del proceso de creación de listas

[21] En programación, un vector o matriz es una estructura de datos que contiene una serie de elementos (valores o variables) almacenados de tal manera que cada uno de ellos tiene asignado una clave de posición. Desde el punto de vista lógico un vector es sencillamente un conjunto de elementos ordenados.

Ahora comprobaremos la aplicación en la enseñanza que tienen los elementos de `Keywords` (K) y `Frequency` (F) basándonos en su representatividad sobre el corpus (A).

Dado que el 95% de los elementos de A están representados en F,

$$\forall\, x \in A, \qquad P(x \in F) = 0{,}95$$

Es decir, la probabilidad de que una palabra cualquiera del corpus esté en la lista `Frequency` es P = .95

Si consideramos que el corpus es representativo del lenguaje escrito, se puede obtener el siguiente resultado:

Aplicación en la enseñanza 1:

Un alumno que aprende todas las palabras de `Frequency` puede entender al menos el 95% del vocabulario de un texto medio, alcanzando así la cobertura mínima para lectura autónoma.

A partir de diccionarios especializados creamos la matriz `Cognates`, que contiene los cognados transparentes inglés-español dentro de las 25.000 palabras más frecuentes del inglés. Para calcular cuántas palabras de `Frequency` son cognados transparentes se halla el número de los elementos que aparecen en ambas matrices, es decir:

$$Card(F \cap C)$$

Posteriormente creamos `Keywords,` una matriz que contiene todas las palabras de `Frequency` que no son cognados transparentes.

Si F = Frequency, C= Cognates, y K=Keywords

$$\forall\, x \in F,\ si\ x \notin C \Rightarrow x \in K$$

es decir, la matriz Keywords se define como $K = \{x : (x \in F) \wedge (x \notin C)\}$

por tanto, $Card(K) = Card(F) - Card(F \cap C)$

Aplicación en la enseñanza 2:

El aprendizaje de Frequency garantiza el 95% de cobertura. Dado que los cognados transparentes son deducibles, podemos eliminar estas palabras de la lista sin que esto afecte a la cobertura alcanzada por el alumno hispanohablante. El listado resultante, Keywords, también garantiza el umbral mínimo del 95%.

La matriz Cognates contiene un número considerable de cognados que no están entre las palabras más frecuentes en inglés. Dado que:

$$\forall\, x \in A, \qquad P(x \in F) = 0{,}95$$

$$F \subsetneq (K \cup C) \subsetneq A$$

Entonces, $P\big(x \in (K \cup C)\big) > 0{,}95$

Es decir, la probabilidad de que una palabra cualquiera del corpus representativo esté en Keywords o sea cognado evidente supera el 0,95.

Aplicación en la enseñanza 3:

Dado que el aprendizaje de la lista Keywords garantiza el 95% y además hay cognados transparentes fuera de Frequency, aprender Keywords implica una cobertura igual o mayor al 95% de un texto medio.

---

2. METODOLOGÍA PARA LA LISTA DE VOCABULARIO PRIORITARIO

2.1. Definición del punto de equilibrio

El primer paso para determinar las palabras necesarias para la lectura autónoma es definir un baremo a partir del cual consideramos que el aprendiz comprende un texto. Para establecer este valor evaluamos las conclusiones obtenidas de los estudios de Hirsh (1992), Huckin (1995), Laufer y Ravenhorst-Kalovski (2010) y Schmitt (2000), que apoyan la hipótesis del umbral mínimo situándolo en tres puntos diferentes según el nivel de comprensión requerido.

En este estudio definimos el concepto *punto de equilibrio* como el conjunto de palabras que alcanzaría el equivalente al *umbral mínimo* de Laufer en un corpus grande completo. Es importante remarcar que ambos conceptos tienen grandes diferencias en cuanto a su orden de magnitud: recordemos que dicho *umbral mínimo* establece proporciones en pequeños textos, mientras que el *punto de equilibrio* trabaja sobre corpus que contienen millones de palabras. Dado que ambas medidas son relativas al tamaño de su muestra, para un mismo valor porcentual la diferencia en números absolutos puede ser enorme.

Por esta razón, al situar el umbral mínimo de Laufer los investigadores se han podido mover con cierta libertad entre el 90% y el 98% de cobertura en función de criterios arbitrarios de exigencia. Sin embargo, cuando la muestra contiene millones de palabras, las consecuencias de una pequeña variación son considerables. Debido a que la distribución de frecuencia sigue una ley de potencias, a medida que aumentamos el umbral, el vocabulario necesario para cubrirlo sube de manera exponencial. Si a esto le sumamos que los valores que barajamos están al final de la curva, en el intervalo porcentual [90 – 98], una pequeña diferencia puede ser inasumible en números absolutos. Por ejemplo, en un texto de 1.000 palabras, pasar de una cobertura del 95% al 98% puede requerir únicamente unas 20 palabras, obstáculo que se puede solucionar fácilmente proporcionando un glosario al alumno. Si en lugar de un texto exploramos el British National Corpus completo, pasar del 95% al 98% requeriría aumentar la base léxica del alumno en unas 4.000 - 6.000 palabras.

Por tanto, debemos ser muy prudentes al fijar el porcentaje de corpus en el que estableceremos el punto de equilibrio, ya que este valor debe ofrecer ciertas garantías de comprensión y, simultáneamente, englobar un número absoluto de palabras asumible para ser alcanzado mediante instrucción explícita. De ahí que el término escogido sea «equilibrio» frente a «mínimo».

El punto de equilibrio para este estudio se fijó en el 95%, basándonos en la extrapolación para un corpus grande de los resultados empíricos sobre textos de los trabajos de Hirsh (1992), Huckin (1995), Laufer y Ravenhorst-Kalovski (2010) y Schmitt (2000). Según los estudios citados, en el primer umbral, situado al 90%, el alumno únicamente intuye la idea global, lo que es un alcance insuficiente para este estudio. El segundo umbral, situado en el 95%, es el primer valor donde el aprendiz parece poder leer de manera autónoma con ciertas garantías. Hay dos razones para descartar el tercer umbral, situado en un ambicioso 98%. La primera de ellas es que esta cobertura excede el objetivo de comprensión mínima del presente estudio. La segunda es que la carga léxica correspondiente al 98% de cobertura es tan extensa que difícilmente podría alcanzarse mediante enseñanza directa.

2.2. Elección del corpus primario

Nuestra matriz inicial se construyó sobre una versión revisada de términos de alta frecuencia elaborada por Paul Nation (2011). Las palabras de la lista de Nation están clasificadas en 29 niveles de frecuencia, cada uno de ellos contiene 1.000 familias léxicas ordenadas alfabéticamente. Los datos del primer y segundo nivel, que corresponden a las 2.000 palabras más frecuentes, provienen de un corpus equilibrado de inglés informal oral y escrito. El tercer nivel y sucesivos se obtienen de las palabras con mayor índice de frecuencia relativa en el BNC y el COCA al eliminar las que aparecen en los niveles anteriores.

Cada una de las 1.000 familias léxicas de un nivel está compuesta por una forma base (sin lexemas) y sus palabras derivadas. Por ejemplo, la familia léxica de la forma base *build* abarca los siguientes elementos: *build, builder, builders, building, buildings, builds, built, unbuilt, rebuild, rebuilding, rebuilt, prebuilt.*

Así, el primer nivel contiene, además de los 1.000 lemas, 5.348 palabras derivadas organizadas de manera jerárquica como se ilustra en la Tabla II.1. El criterio para formar las familias léxicas contiene hasta el nivel 6 de las reglas de afijos de Bauer y Nation (1993) , en la Tabla II.2.

Tabla II.1. Muestra de familias léxicas en el listado de Nation (2011).

FAMILIA #	FORMA BASE	DERIVADAS
248	**DRESS**	
		DRESSED
		DRESSES
		DRESSING
		UNDRESSED
		UNDRESS
		UNDRESSES
		UNDRESSING
249	**DRINK**	
		DRINKS
		DRINKER
		DRINKERS
		DRANK
		DRUNK
		UNDRUNK
		DRINKING
250	**DRIVE**	
		DRIVEN
		DRIVER
		DRIVERS
		DRIVES
		DRIVING
		DROVE

Tabla II.2.Reglas de afijos de Bauer y Nation (1993)

Nivel 1 Cada forma es una palabra diferente.
Nivel 2 Las palabras con inflexión regular son de la misma familia. Las categorías de inflexión son: plural, tercera persona singular del presente, pasado, gerundio, comparativo, superlativo y posesivo.
Nivel 3 -able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-, con usos restringidos.
Nivel 4 -al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in-, con usos restringidos.
Nivel 5 -age (leakage), -al (arrival), -ally (idiotically), -an (American), -ance (clearance), -ant (consultant), -ary (revolutionary), -atory (confirmatory), -dom (kingdom; officialdom), -eer (black marketeer), -en (wooden), -en (widen), -ence (emergence), -ent (absorbent), -ery (bakery; trickery), -ese (Japanese; officialese), -esque (picturesque), -ette (usherette; roomette), -hood (childhood), -i (Israeli), -ian (phonetician; Johnsonian), -ite (Paisleyite; also chemical meaning), -let (coverlet), -ling (duckling), -ly (leisurely), -most (topmost), -ory (contradictory), -ship (studentship), -ward (homeward), -ways (crossways), -wise (endwise; discussion-wise), anti- (anti-inflation), ante- (anteroom), arch- (archbishop), bi- (biplane), circum- (circumnavigate), counter- (counter-attack), en- (encage; enslave), ex- (ex-president), fore- (forename), hyper- (hyperactive), inter- (inter-African, interweave), mid- (mid-week), mis- (misfit), neo- (neo-colonialism), post- (post-date), pro- (pro-British), semi- (semi-automatic), sub- (subclassify; subterranean), un- (untie; unburden).
Nivel 6 -able, -ee, -ic, -ify, -ion, -ist, -ition, -ive, -th, -y, pre-, re-.

La razón que nos motivó a trabajar sobre esta lista en lugar de utilizar los datos brutos del BNC o el COCA es que Nation la diseñó especialmente para la enseñanza de inglés como L2 mediante un sistema que combinaba rigor estadístico y aplicación práctica. El autor había observado que en los primeros puestos de las listas de frecuencia basadas en el BNC faltaban ciertas palabras de uso muy común en inglés. Decidió, por tanto, que para obtener las primeras 2.000 palabras esenciales era necesario crear un corpus *ad hoc* de inglés cotidiano, con 10 millones de elementos de los que 6 millones provienen de inglés oral en sus variedades británica y americana. Para el tercer nivel y sucesivos tomó como referencia el BNC y el COCA tras excluir las palabras de los niveles 1 y 2.

Nation explica el proceso en *Information on the BNC-COCA word family lists* (2012):

> This unusual step of creating a special corpus for the first 2000 word families was followed because the previous lists made from the British National Corpus were so strongly influenced by the written formal nature of the corpus that they were not suitable lists for creating language courses or graded reader lists. […] Very common words in spoken English like alright, pardon, hello, dad, bye could then be included in the high frequency words. Other arbitrary adjustments included putting all the word forms of numbers (one, two, hundred) and weekdays in the 1st 1000, and the months of the year in the 2nd 1000, even though their frequency did not always justify this. The goal was to have a set of high frequency word lists that were suitable for teaching and course design (Nation, 2012: 1-2).

> Este paso inusual de crear un corpus especial para las primeras 2.000 familias de palabras, responde al hecho de que las listas anteriores del British National Corpus estaban tan influidas por la naturaleza formal y escrita del corpus que no eran adecuadas para crear cursos de idiomas ni listas para lecturas graduadas. […] Se añadieron a las palabras de alta frecuencia algunas palabras muy comunes en inglés hablado tales como *alright, pardon, hello, dad, bye* [vale, perdón, hola, papá, adiós]. Otros ajustes arbitrarios incluyeron los números (*one, two, hundred*) [uno, dos, cien] así como los días de la semana en la primera lista de 1000 palabras , y los meses del año en la segunda, aunque su frecuencia no lo justificara. El objetivo era tener una serie de listas de palabras de alta frecuencia adecuadas para la enseñanza y el diseño de cursos (trad. a.).

Originalmente, Nation elaboró esta base de datos para que pudiera ser utilizada con el software de análisis léxico RANGE (Nation, Coxhead, & Heatley, 2002), que fue el precursor de AntWordProfiler (Lawrence, 2013), el programa que empleamos para este estudio y que es capaz de procesar archivos con el mismo formato.

2.3. Umbral

Para calcular el número de palabras de la lista de Nation necesarias para cubrir el umbral de equilibrio fijado en el 95% del corpus tomamos como referencia los porcentajes acumulados de la lista lematizada de frecuencias del BNC elaborada por Kilgarriff (1995) y obtuvimos un valor próximo a 3.000 palabras. Estos resultados son congruentes con la ley de Zipf (1949) y las conclusiones de distintos estudios sobre cobertura de vocabulario de alta frecuencia en textos (Brysbaert et al., 2011; Hirsh & Nation, 1992; Laufer, 1989; Laufer, 1992; Laufer, 1997; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006a; Nation & Chung, 2009; Pulido, 2009).

Debido a que estos datos para hallar el número de palabras del punto de equilibrio se extrajeron únicamente del BNC mientras que la lista de Nation está basada en el BNC y el COCA y un corpus oral, valoramos que podía existir un margen de error si se replicaba el estudio tomando como referencia otro corpus basado en una variedad dialectal concreta. Sin embargo, cuando se trata de corpus equilibrados, las disparidades se aprecian principalmente en las unidades léxicas de baja frecuencia, mientras que las palabras de uso cotidiano no parecen variar. Concretamente, al comparar las 3.000 palabras de mayor frecuencia del BNC y el COCA pueden apreciarse pequeñas variaciones en cuanto al rango de algunas palabras, pero el conjunto de elementos es prácticamente igual (Newman, Baayen, & Rice, 2011), por lo que se puede asumir que el margen de error no pone en riesgo la asunción de que 3.000 palabras corresponden al punto de equilibrio estimado al 0,95 para cualquier corpus equilibrado.

Partiendo de este dato, construimos nuestra matriz inicial, llamada `Frequency`, replicando los tres primeros niveles de la lista de Nation, cuyo número de lemas y palabras derivadas se ilustra en la Tabla II.3:

Tabla II.3. Distribución de elementos en Frequency

NIVEL	LEMAS	DERIVADAS	TOTAL
NATION 1	1.000	5.857	6.857
NATION 2	1.000	5.370	6.370
NATION 3	1.000	4.880	5.880
TOTAL	3.000	16.107	19.107

2.4. Las palabras cognadas

En los 3.000 elementos de la matriz `Frequency` existen muchas palabras cuya traducción es evidente para cualquier hablante de español. La segunda fase de este estudio consistió en eliminar todas aquellas palabras que se pueden deducir a través de afijos comunes.

2.4.1. Definición de cognados transparentes

Llamamos cognados transparentes a aquellas palabras inglesas cuya similitud con su equivalente en español hace que sean deducibles para un hispanohablante que conozca los afijos derivativos más comunes en inglés. En primer lugar es necesario definir a partir de qué grado de similitud consideramos que un cognado será deducible. Evidentemente, los cognados exactos u homográficos (*motor, visible*) no presentarán ningún problema. Entre los cognados parciales también hay pares transparentes, (*plastic/*plástico; *attention/*atención), sin embargo, hay otros como *sugar*/azúcar, cuya evolución respecto del origen común ha dado lugar a variaciones ortográficas que dificultan su identificación.

Si bien la capacidad para reconocer cognados depende hasta cierto punto de la habilidad individual del alumno, la investigación reciente ha demostrado que hay ciertos patrones presentes en todos los cognados que son identificados correctamente por los participantes de los experimentos. Basándonos en estos resultados, determinamos el grado de similitud asignando a cada cognado una puntuación que se obtiene de la comparación de distintos factores en el par L1/L2. Utilizamos con los valores establecidos por Schwartz et al. (2007) y su fórmula basada en el algoritmo de Van Orden (1987), que tiene en cuenta los siguientes indicadores:

- Número de pares de letras adyacentes en el mismo orden
- Número de pares de letras adyacentes en orden inverso
- Número de letras comunes

- Media de la longitud de ambas palabras
- Ratio de la longitud de la palabra más corta respecto de la más larga
- Primera y/o última letra igual

Según la fórmula de Van Orden, el par ignition-ignición tiene una puntuación de similitud más alta que cube-cubo porque, entre otros factores, tiene la misma letra inicial y final. La longitud y la secuencia de letras compartidas explican por qué el par methodical- metódico es transparente mientras que guide-guía no lo es, a pesar de que tienen el mismo número de letras diferentes y que ninguno de los pares comparte la letra final.

En cuanto a las variedades dialectales, se hizo todo lo posible por tomar como referencia únicamente las palabras del español internacional estándar. En este sentido, no se clasificaron como cognados palabras como *car*, aunque sí lo sea en Perú, Colombia, México o Venezuela, donde *car* es equivalente a «carro». En el español de España o Argentina, en cambio, «carro» es un vehículo de tracción externa, y *car* es «coche», un par sin similitud ortográfica. En esta línea, tampoco se tuvieron en cuenta los préstamos ingleses que solamente se han incorporado al español cotidiano en una variedad dialectal concreta. Este fenómeno es muy habitual en México y otras regiones cercanas a EEUU.

2.4.2. Selección y exclusión de cognados

Una vez definido el tipo de cognado que se puede seleccionar, creamos la matriz `Cognates`, una lista de los cognados transparentes inglés-español que se encuentran dentro de las 25.000 palabras más frecuentes del inglés. A continuación se describe el complejo proceso de creación de la matriz `Cognates`.

En primer lugar, se extrajo un extenso listado de cognados mediante un archivo de órdenes (script) que consulta la base de datos de la aplicación web *The Cognate Highlighter* desarrollada por The Cognate Project (Morán, 2010). Este listado preliminar se completó con las entradas de diccionarios especializados tales como

Dictionary of cognates (Morán, 2011b), *NTC's dictionary of Spanish cognates: Thematically organized* (Nash, 1997) o *Instant Spanish: Vocabulary Builder* (Means, 2003). Posteriormente, se analizó la similitud ortográfica de cada par de cognados utilizando un algoritmo de la fórmula de Van Orden (1987) y se seleccionaron aquellos que cumplían con los valores de similitud ortográfica establecidos en Schwartz et al. (2007).

Observemos ahora cómo procesa el algoritmo de Van Orden[22] el par IGNITION-IGNICIÓN.

En primer lugar se analizan los indicadores de similitud ortográfica indicados anteriormente y se le asigna una puntuación a cada uno de ellos. En el ejemplo que nos ocupa, los valores de los indicadores son los siguientes:

PALABRA 1: IGNITION
PALABRA2: IGNICIÓN
LONGITUD DE LA PALABRA1 = 8
LONGITUD DE LA PALABRA2 = 8
PRIMERA LETRA COINCIDENTE = SÍ (I)
ÚLTIMA LETRA COINCIDENTE = SÍ (N)
LETRAS ÚNICAS EN LA PALABRA1 = I,G,N,T,O
LETRAS ÚNICAS EN LA PALABRA2 = I,G,N,C,O
PARES ADYACENTES EN LA PALABRA1 = "IG","GN","NI","IT","TI","IO","ON"
PARES ADYACENTES EN LA PALABRA2 = "IG","GN","NI","IC","CI","IO","ON"
PARES ADYACENTES EN ORDEN INVERSO = "NO","OI","IC","CI","IN","NG","GI"

[22].Para ilustrar todo el proceso del algoritmo con el ejemplo IGNITION-IGNICIÓN hemos utilizado la aplicación *The Weber (1970) and Van Orden (1987) algorithm for determining spelling similarity*, disponible en subjectpool.com/quest/reading/spelling_similarity.php

Una vez obtenidos estos valores, aplicamos la siguiente fórmula:

$$\Omega = 10 \cdot \frac{50A + 30V + 10C}{L} + 5R + 27P + 18U$$

donde L (longitud media de las palabras) = 8
P (primera letra común) = 1
C (número de letras comunes) = 4
U (última letra igual) = 1
A (Número de pares de letras adyacentes) = 5
R (Ratio de la longitud de la palabra más corta respecto de la más larga) = 1
V (Número de pares de letras adyacentes en orden inverso) = 0

Esto nos devuelve un grado de similitud Ω = 862,5. Este valor está dentro del intervalo establecido por Schwartz et al. (2007), por lo que la palabra *IGNITION* se considera un cognado transparente y se pasa a la matriz `Cognates`. Terminado este proceso, el archivo `Cognates` contenía una lista de palabras en su forma base ordenada por orden alfabético, como se ve en la Tabla II.4.

Posteriormente, mediante el programa AntWordProfiler comparamos los lemas de las 3.000 familias de Nation con la matriz `Cognates` y extrajimos los valores únicos. Esta orden nos devolvió una lista de todas las palabras de alta frecuencia que no son cognados transparentes hasta alcanzar el punto de equilibrio. La Tabla II.5 muestra los cuatro primeros lemas descartados de cada nivel junto con sus derivadas.

Tabla II.4. Muestra de la lista de cognados transparentes en su forma base.

…
ACCEDE
ACCELERATE
ACCENT
ACCEPT
ACCESSORY
ACCIDENT
ACCLIMATE
ACCOMPANY
ACCORDION
ACCREDITATION
ACCUMULATE
ACCUSE
ACETATE
ACETONE
ACETYLENE
ACHROMATIC
ACID
ACOUSTIC
AQUISITION
ACROBAT
ACRONYM
…

Tabla II.5. Palabras asignadas a la lista Cognates.

NIVEL 1	NIVEL 2	NIVEL 3
ABSOLUTE	**ABUSE**	**ABSENCE**
ABSOLUTELY	ABUSED	ABSENCES
ABSOLUTIST	ABUSIVE	**ACCELERATE**
ABSOLUTISTS	ABUSING	ACCELERATED
ACCEPT	ABUSES	ACCELERATES
ACCEPTABILITY	ABUSER	ACCELERATING
ACCEPTABLE	ABUSERS	ACCELERATOR
ACCEPTABLY	**ACCENT**	ACCELERATORS
UNACCEPTABLE	ACCENTED	ACCELERATION
ACCEPTANCE	UNACCENTED	ACCELERATIONS
ACCEPTED	ACCENTING	**ACCUSE**
ACCEPTING	ACCENTS	ACCUSING
ACCEPTS	**ACCESS**	ACCUSINGLY
UNACCEPTABLY	ACCESSED	ACCUSES
ACT	ACCESSES	ACCUSED
ACTED	ACCESSIBILITY	ACCUSATION
ACTING	ACCESSIBLE	ACCUSATIONS
ACTION	ACCESSING	ACCUSER
INACTION	INACCESSIBLE	ACCUSERS
ACTIONS	**ACCIDENT**	**ADAPT**
ACTIONABLE	ACCIDENTS	ADAPTED
ACTS	ACCIDENTAL	UNADAPTED
ACTOR	ACCIDENTALLY	ADAPTING
ACTORS		ADAPTS
ACTRESS		ADAPTATION
ACTRESSES		ADAPTATIONS
ACTIVE		ADAPTABLE
ACTIVELY		ADAPTER
ACTIVITIES		ADAPTERS
ACTIVITY		ADAPTOR
INACTIVE		ADAPTORS
ACTIVIST		ADAPTABILITY
ACTIVISTS		ADAPTABILITIES
ACTIVISM		ADAPTIVE

2.5. Matriz de palabras clave

Las palabras de `Frequency` que no eran cognados transparentes se pasaron a una matriz llamada `Keywords`, creada para contener el vocabulario que el alumno deberá aprender prioritariamente. `Keywords` está dividida en cuatro niveles de prioridad y su contenido se muestra por orden alfabético. Al eliminar los cognados, el número de elementos de `Keywords` es considerablemente más reducido que los 3.000 elementos de la lista `Frequency` y aún así, supuestamente, permite a un hispanohablante comprender al menos el mismo porcentaje de un texto.

La Tabla II.6 muestra los cuatro primeros lemas y sus palabras derivadas que permanecieron en cada nivel de `Keywords` tras excluir los cognados transparentes.

Tabla II.6. Primeros elementos de Keywords.

NIVEL 1	NIVEL 2	NIVEL 3
A	**ABOVE**	**ABBEY**
AN	**ACCORDING**	ABBEYS
ABLE	**ACCURATE**	**ABROAD**
ABLER	ACCURACY	**ACCOUNTANT**
ABLEST	ACCURACIES	ACCOUNTANTS
ABLY	INACCURACY	ACCOUNTANCY
UNABLE	INACCURACIES	**ACHE**
ABOUT	ACCURATELY	ACHES
ACCOUNT	INACCURATE	ACHED
ACCOUNTED	INACCURATELY	ACHING
ACCOUNTING	**ADVICE**	ACHY
ACCOUNTS	ADVICED	

2.6. Jerarquización de la lista

Con el fin de aplicar la lista de palabras clave a la creación de material docente, se jerarquizó `Keywords` en cuatro niveles de prioridad. Los criterios fueron frecuencia y dificultad específica para hispanohablantes. En primer lugar se creó el nivel 4, el de mayor dificultad para el alumno, compuesto únicamente por falsos cognados, es decir, por palabras que pueden inducir a error a un hispanohablante, tales como *advertise, actual o introduce* ('anunciar', 'real' y 'presentar', respectivamente). Posteriormente, se distribuyeron las palabras restantes en los tres primeros niveles según su frecuencia.

Una dificultad que encontramos fue que la lista de frecuencias de Nation está dividida en grupos de 1.000 familias, pero dentro de cada nivel las palabras están ordenadas alfabéticamente y no proporcionan datos sobre su frecuencia relativa. Como consecuencia de extraer los cognados y los falsos cognados, los tres niveles de `Keywords` quedaban muy descompensados en cuanto al número de elementos, por lo que dividirla equitativamente requería, en primer lugar, ordenar los elementos de `Keywords` según su frecuencia. Para obtener este dato se extrajeron los valores del rango de cada palabra que aparece en la lista lematizada del BNC de Kilgarriff (Kilgarriff, 1995). Hecho esto, las palabras se asignaron a los niveles 1, 2 y 3 según su frecuencia, siendo el primero el de las palabras más frecuentes.

3. METODOLOGÍA PARA EL ANÁLISIS DE TEXTOS

3.1. Géneros

Con el fin de verificar la eficacia del estudio así como la validez del punto de equilibrio establecido, analizamos con el programa AntWordProfiler textos de distintos géneros. En cada uno de ellos hallamos qué porcentaje de vocabulario cubre la matriz `Keywords`, cuántas palabras son deducibles y, finalmente, qué porcentaje de texto estaría formado por palabras presumiblemente desconocidas. Los textos analizados son exámenes de certificación de nivel L2, novelas originales y adaptadas y transcripciones de películas, series de TV y habla formal.

En primer lugar, se analizaron las pruebas de comprensión lectora de 24 exámenes de los niveles B1 y B2 del Marco europeo de referencia para las lenguas. Dada la clara orientación de aplicación docente de la presente tesis, el análisis de exámenes nos permitiría analizar la validez de `Keywords` como recurso sobre el que construir programas léxicos para los cursos de inglés destinados a preparar exámenes de certificación del nivel de inglés como L2 (ver organismos examinadores en la Tabla II.7).

Tabla II.7. Exámenes y equivalencias al MCERL

EXAMINADOR	NIVEL B1	NIVEL B2
Cambridge ESOL	Preliminary English Test (PET)	First Certificate in English (FCE)
Anglia	Intermediate	Advanced
Trinity Guildhall	Trinity Grades 5/6	Trinity Grades 7/9
Escuela Oficial de Idiomas	Nivel Intermedio	Nivel Avanzado

En cuanto a la narrativa, se analizaron las novelas *A Christmas Carol*, *The Adventures of Sherlock Holmes* y *Lady Chatterley's Lover*. Las dos primeras se compararon con sus respectivas adaptaciones para *graded readers* en tres niveles de dificultad.

Con el objetivo de comprobar los resultados del estudio en un registro coloquial se estudió la distribución del vocabulario en las películas *Shrek* y *The Goonies*, y en la serie de televisión *How I met your mother*. Si bien el objeto de estudio en esta tesis no es el lenguaje oral, hay que tener en cuenta que estas muestras no son, en realidad, habla espontánea. En otras palabras, analizamos guiones escritos que son recreaciones artificiales del discurso oral. Su relevancia para este análisis se fundamenta en que estas recreaciones tratan de replicar de manera fiel el registro coloquial en el habla espontánea. Como contraste, para analizar el habla no espontánea en un registro culto utilizamos un corpus de discursos políticos del presidente de EEUU Barack Obama.

3.2. Software

El programa utilizado fue AntWordProfiler (Lawrence, 2013). Es una herramienta de análisis lingüístico sobre perfiles de vocabulario que permite generar estadísticas sobre frecuencia relativa y absoluta de un corpus dado, así como la distribución de las palabras del corpus que aparecen en listados aportados por el usuario.

En primer lugar se debe introducir el corpus de textos que se desea analizar (*user files*). El programa permite el tratamiento de múltiples archivos, que se pueden analizar de manera independiente o conjunta (*batch process*). Los archivos deben proporcionarse en texto plano con formato .txt. Posteriormente, se deben introducir los archivos con las palabras cuya presencia queremos analizar en el corpus (*level list*). El programa puede procesar listados verticales de lemas con las palabras derivadas tabuladas, siguiendo el mismo formato definido para el programa RANGE (Nation et al., 2002).

El programa devuelve las palabras que pertenecen a cada *level list* así como las que no se encuentran en ninguno de ellos, además de las estadísticas de frecuencia del vocabulario del corpus. Si las *level lists* especifican lemas y familias, podremos obtener estadísticas en función de tres parámetros: *tokens, types* y *groups* (palabras totales, palabras únicas y familias léxicas, respectivamente). AntWordProfiler proporciona los siguientes datos para cada una de las *level lists*:

TOKEN	NÚMERO DE PALABRAS TOTALES
TOKEN %	PORCENTAJE DE PALABRAS TOTALES.
TOKEN%CUM	PORCENTAJE ACUMULADO DE PALABRAS TOTALES.
TYPE	NÚMERO DE PALABRAS ÚNICAS.
TYPE%	PORCENTAJE DE PALABRAS ÚNICAS.
TYPE%CUM	PORCENTAJE ACUMULADO DE PALABRAS ÚNICAS.
GROUP	NÚMERO DE FAMILIAS LÉXICAS.
GROUP%	PORCENTAJE DE FAMILIAS LÉXICAS.
GROUP%CUM	PORCENTAJE ACUMULADO DE FAMILIAS LÉXICAS.

3.3. Análisis de palabras prioritarias

Como *user files* se utilizaron versiones en texto plano de las muestras descritas en el punto 3.1 (en el caso de las películas utilizamos los archivos de subtítulos). Para el análisis de palabras prioritarias se utilizó como *level list* la matriz `Keywords`. Esto nos permitió conocer qué porcentaje de cada texto está cubierto por las familias léxicas del listado de palabras prioritarias.

Para determinar si el aprendizaje de estas palabras permite alcanzar el punto de equilibrio del 95%, es necesario conocer cuántas palabras pertenecen a `Keywords` y cuántas de las restantes son deducibles para hispanohablantes. Por tanto, es necesario utilizar una *level list* que contengan las palabras presumiblemente deducibles.

3.4. Análisis de cognados

Nuestro listado de cognados transparentes se compiló a través de diccionarios, por lo que contenía únicamente las formas base que contaban con su propia entrada en dichos diccionarios. Este dato era suficiente para determinar su presencia en las listas de Nation, que contienen las familias léxicas agrupadas y encabezadas por su forma base o lema. Sin embargo, para el análisis de textos originales cuyo vocabulario aparece en formas flexionadas, la lista de lemas de `Cognates` no podía aportar, originalmente, datos suficientes.

En este estudio asumimos que un alumno es capaz de razonar que *accidental* y *accidentally* pertenecen a la misma familia que *accident,* y que si deduce el significado de la forma base también podrá comprender las palabras derivadas mediante los sufijos más habituales en inglés. Sin embargo, para un programa informático este razonamiento no es un proceso tan sencillo. Para esta investigación fue preciso diseñar un sistema mediante el que el programa identificara que todas las palabras derivadas de una palabra cognada pertenecen a la misma familia léxica.

En lingüística computacional existen dos opciones para procesar las formas flexionadas del lenguaje natural. La primera se conoce como *stemming*, que es la más similar al proceso mental humano: identificar los prefijos y sufijos más comunes y

extraerlos para obtener la forma base de la palabra. Uno de los primeros investigadores en automatizar este proceso fue Martin Porter (1980), cuyo algoritmo consiste en una serie de reglas que se aplican de forma secuencial. Algunas de estas reglas son:

- SI LA PALABRA ACABA EN 'ED', ELIMINAR 'ED'.
- SI LA PALABRA ACABA EN 'ING', ELIMINAR 'ING'.
- SI LA PALABRA ACABA EN 'S', ELIMINAR 'S'.

A pesar de que el sistema de Porter es bastante eficaz y sigue siendo uno de los más utilizados hoy en día, no está exento de errores. Tomemos como ejemplo la palabra *ironic* (irónico): al quitar el sufijo, el programa considerará erróneamente que pertenece a la misma familia que *iron* (hierro). Otros ejemplos de categorizaciones equivocadas son los pares *several* - *severe* (varios - severo) o *animated* - *animal* (animado - animal).

La segunda opción para procesar formas flexionadas es la lematización, un proceso similar al *stemming* aunque mucho más complejo. La diferencia principal es que el *stemming* convierte todo el texto a las formas base, mientras que la lematización opera sobre las palabras en su forma flexionada tal y como aparecen en el texto mediante análisis morfológico y búsquedas en el diccionario. A pesar que da menos errores de clasificación que el *stemming*, la lematización es un proceso mucho más complicado y que tampoco está exento de problemas.

La opción que se escogió para este estudio fue la que garantizaba mayor rigor: trabajar sobre nuestros datos invirtiendo los procesos de *stemming* y lematización. En lugar de intentar obtener formas base a partir de formas flexionadas en textos, partimos de nuestra lista de cognados y construimos las familias léxicas completas de cada uno. De esta manera, el programa tendría los cognados en todas sus formas deducibles y esto le permitiría analizar cualquier texto sin alterarlo, con las palabras en el formato original. Para ello diseñamos un programa en Ruby[23] adaptado de un estudio previo (Cembreros, 2011) capaz de crear automáticamente palabras derivadas a partir de lemas siguiendo las normas de afijos de Bauer y Nation (1993) hasta el nivel 6, replicando así

[23] Ruby es un lenguaje de programación similar en algunos aspectos a Python y Perl que se distribuye bajo una licencia de software libre.

el sistema de Paul Nation. De las palabras resultantes, se seleccionaron únicamente aquellas que son palabras reales en inglés utilizando como referencia WordNet, la base de datos léxica de la Universidad de Princeton. Las consultas se hicieron mediante tratamiento por lotes a través de una interfaz de programación de aplicaciones (API) de WordNet. La Tabla II.8 es una muestra de las familias creadas correctamente para las palabras *accelerate, accent y accept*.

Tabla II.8. Familias léxicas derivadas de cognados transparentes.

ACCELERATE	ACCENT	ACCEPT
↓	↓	↓
ACCELERATED	ACCENTED	ACCEPTABILITY
ACCELERATEDLY	ACCENTS	ACCEPTABLE
ACCELERATES	ACCENTUATE	ACCEPTABLY
ACCELERATING	ACCENTUATED	ACCEPTANCE
ACCELERATION	ACCENTUATES	ACCEPTANCES
ACCELERATIONS	ACCENTUATING	ACCEPTATION
ACCELERATIVE	ACCENTUATION	ACCEPTED
ACCELERATOR	ACCENTUATIONS	ACCEPTING
ACCELERATORS		

3.5. Palabras invariables

En los textos traducidos hay un grupo de palabras que deben permanecer en su forma original, tales como los nombres propios y las voces extranjeras. Estas palabras no son objeto de estudio en el aprendizaje de la L2. Cuando el alumno identifica que una palabra es un nombre propio sabe que no necesita conocer su significado ya que, por definición, no tiene rasgos semánticos inherentes. Los nombres propios se pueden deducir fácilmente por contexto y por la peculiaridad de escribirse con mayúscula inicial. Por tanto, para el análisis de textos era necesario que el programa no considerase

a los nombres propios como palabras desconocidas, sino como términos invariables que no son objeto de aprendizaje.

Para ello creamos una base de datos de nombres de pila y apellidos comunes, topónimos y marcas comerciales globales con datos obtenidos del censo de EEUU, el Instituto de Estadística de la ONU y los informes anuales de reconocimiento de marcas globales de la agencia de estudios de mercado Millward Brown, respectivamente:

- United States Census Bureau:
  Frequently Occurring First Names from Census 1990.
  Surnames Occurring 100 or more times from Census 2000.
  Disponible en: http://www.census.gov/

- United Nations Statistics Division:
  City population by sex, city and city type.
  Disponible en: http://data.un.org

- Millward Brown:
  BrandZ Top 100. Most Valuable global Brands Report [de 2006 a 2013]
  Disponible en: http://www.millwardbrown.com

Este proceso dio lugar a dos problemas. El primero es que encontramos numerosos apellidos que también son palabras de uso común en inglés, tales como *White* o *Waters*. Esto ocurre también, aunque en menor medida, con los nombres de pila y las marcas. Para que el programa no identificara erróneamente palabras con contenido semántico como nombres propios, fue preciso eliminar de la lista aquellos nombres que también son palabras reales inglesas. Para llevarlo a cabo se cruzaron los datos con las listas del BNC. La Tabla II.9 muestra los cuatro primeros elementos eliminados de cada categoría.

Tabla II.9. Nombres propios con contenido semántico.

APELLIDOS	NOMBRES ♀	NOMBRES ♂	MARCAS
BROWN	ROSE	MARK	APPLE
YOUNG	QUEEN	WEST	TIMES
COOK	CHERRY	MILES	CATERPILLAR
BURNS	RUBY	GRANT	AMAZON

La segunda dificultad que encontramos es que este proceso eliminaba de la lista de nombres propios algunas palabras que en el pasado tuvieron contenido semántico, pero actualmente están en desuso y prácticamente sólo aparecen como nombre propio. Tomemos como ejemplo la palabra *tucker*. En el ranking de apellidos comunes en EEUU, *Tucker* ocupa una destacada posición 141, pero también es una palabra inglesa que designa a una pieza ornamental antigua similar a un cuello de encaje. Al consultar el término en el British National Corpus observamos que solamente en una de las 50 primeras ocurrencias mantiene su sentido original, es decir, en el 98% de sus ocurrencias, *tucker* aparece como nombre propio. En la Tabla II.10, que ilustra las primeras 10 muestras, se observa cómo la única en la que significa 'cuello de encaje' es la número 2.

Para solucionar estos posibles errores, se limitó la búsqueda de coincidencias en el BNC hasta las 6.000 palabras más frecuentes. La razón es puramente estadística: los niveles superiores contienen términos que rara vez aparecen en inglés mientras que la lista de nombres propios contiene únicamente elementos de alta frecuencia; por tanto, podemos asumir que si en un texto encontramos una palabra que puede pertenecer a ambas categorías, lo más probable es que sea un nombre propio. Al final de este proceso obtuvimos el archivo de palabras `Invariables`, formado inicialmente por un listado de nombres propios de alta frecuencia que no coincidieran con ninguna palabra hasta el rango 6.000 del British National Corpus. El proceso se resume en la Figura II.3.

Tabla II.10. Ocurrencias de 'Tucker' en el BNC

A9R 187 With the England captain, Richard Leman, in excellent form, Grinstead were too powerful in midfield for an Eastcote side without their skipper, Gary **Tucker**.
AHU 966 Instead he will, in best bib and **tucker**, be performing his last official act as the Masters champion.
AKM 1183 By PETER **TUCKER**
AT4 1806 'I'll come as far as **Tucker**'s, get some fags,' Nails said.
B1E 441 Moreover, **Tucker** (1987) has pointed out that the first period of large-scale deforestation in northern India occurred in the 1850s and 1860s as the British colonisation of India
BMF 1093 indeed the very first person to introduce me to canoeing was Marianne **Tucker** who subsequently represented England in the 1960 Olympics.
BML 319 Distinguished examples would include Nicholas **Tucker**, whose The child and the book (1981) is both theoretically sound and practically convincing
BMW 1673 If only she could lose her puppy fat and get her hair done at a proper salon instead of having it cut by Ivy **Tucker** who lived down the road and who did hairdressing for pin money.
BN1 15 Instant fame came to him, however, with the publication in 1891 of Die Anarchisten(The Anarchists), which was published in English that same year by Mackay's American friend Benjamin R. **Tucker**, in Boston.
BNK 351 The staff in those days included the formidable and devout figure of Susie **Tucker**, a great Norse scholar.

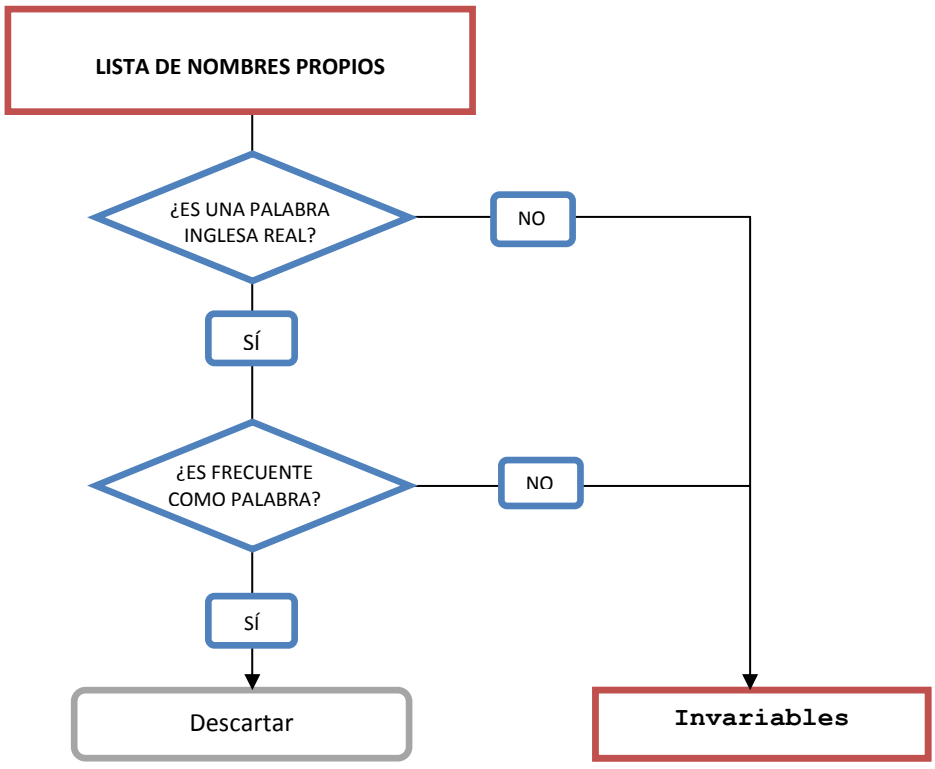


Figura II.3. Tratamiento de nombres propios en Invariables

La matriz `Invariables` contiene las palabras que un alumno no necesita aprender ya que permanecen en la lengua original en las traducciones comerciales de los textos. Es un conjunto de nombres propios, voces extranjeras, palabras inventadas y cadenas de caracteres sin contenido semántico[24]. `Invariables` se diseñó para que estas unidades léxicas no afectaran a las estadísticas y que no devolviesen resultados sesgados. En términos metodológicos, `Invariables` es un listado genérico de nombres propios de alta frecuencia así como otras expresiones que el software no debía computar como términos reales ingleses de baja frecuencia.

Durante el análisis, cada vez que se procesó un texto nuevo se añadieron a este listado los nuevos nombres propios que no estaban ya en el listado de `Invariables`. Para optimizar este procedimiento en textos extensos, programamos un código PHP[25] basado el algoritmo de Sosins (2009) que nos permitió obtener un listado automático de nombres potenciales de cada texto. Posteriormente se cotejaron a mano uno a uno los candidatos con el texto original para comprobar cuáles son, efectivamente, nombres propios. El programa, cuyo código se puede consultar en el Apéndice E, se sirve de una serie de indicadores para determinar si una palabra puede ser un nombre propio. Algunos de los aspectos que se analizan son los siguientes:

- Mayúscula inicial.
- Posición en el texto (ej.: que no sea la primera palabra de una frase).
- Palabras precedidas de una partícula de título (ej.: *Mr., Mrs., Ms., Dr., Sir, Miss*)
- Combinaciones de dos palabras que empiezan por mayúscula separadas por guiones o ciertas partículas (ej.: *of, the*).
- Palabras cuyas letras son todas mayúsculas (porque posiblemente sean siglas) y que estén aisladas (varias palabras en mayúsculas seguidas suelen ser el título del texto o de un capítulo).
- Nombres propios formados por más de una palabra.

[24] Un ejemplo son las cadenas de caracteres «C2», «E4», «B1» que se refieren a las casillas de un tablero de ajedrez en la película *Harry Potter y la piedra filosofal*, cuyo análisis veremos posteriormente.

[25] PHP es un lenguaje de programación de propósito general aunque es especialmente popular en el desarrollo de páginas web, ya que puede ser incrustado directamente en HTML. Es un lenguaje de código abierto, multiplataforma y que devuelve datos HTML dinámicos. Utiliza script del lado del servidor, igual que otros lenguajes tales como Java, Python o Ruby.

Los resultados incluyen como *candidatos* algunas palabras que el programa no tiene suficientes garantías para determinar si son o no nombres propios. Esto ocurre, por ejemplo, si una palabra aparece una única vez en el texto y al principio de una frase, lo que justificaría la mayúscula inicial.

4. RESULTADOS Y ANÁLISIS

4.1. La lista de vocabulario prioritario

El resultado de este proceso es un listado que contiene los 1.800 términos ingleses que, estadísticamente, son los más rentables para un aprendiz de inglés cuya lengua materna sea el español. La lista completa de los lemas ordenados alfabéticamente se incluye en el Apéndice A. Su jerarquización para su aplicación en la docencia se encuentra en el Apéndice B, que divide la lista en cuatro niveles atendiendo a su prioridad. Los niveles 1 a 3 están divididos por frecuencia, con la excepción de algunas series como los números o los meses del año, que se han agrupado en el primer nivel en el que aparezca un elemento de la serie. El cuarto nivel contiene los falsos cognados de `Keywords`, que se separan del resto de palabras porque requieren un enfoque docente especial. La configuración de los niveles es la siguiente:

Nivel 1:	691 palabras.
Nivel 2:	453 palabras.
Nivel 3:	486 palabras.
Nivel 4:	170 palabras.
Total:	1.800 palabras.

La Tabla II.11, cuyos datos están ilustrados en la Figura II.4, muestra la comparación entre Keywords y otras listas conocidas de frecuencia: la *General Service List* (GSL), de 2.000 palabras, y la *Academic Word List* (AWL), de 570 palabras. El 68% de las palabras de Keywords aparecen también en la GSL; en números absolutos,

hay 638 Keywords entre las primeras 1.000 de la GSL, y otras 547 en el segundo nivel. Con respecto a la AWL, como era previsible, el porcentaje de Keywords es muy bajo debido a la fuerte presencia de cognados deducibles en el lenguaje académico. Solamente el 5% de las Keywords aparecen también en la AWL, un total de 96 palabras. Por último, Keywords tiene 473 valores únicos, lo que implica que una de cada tres Keywords no pertenece ni a la AWL ni a la GSL.

Tabla II.11. Distribución de Keywords sobre GSL y AWL

NIVEL	TOKEN	TOKEN%	CUMTOKEN%
GSL 1000	684	38.00	38
GSL 2000	547	30.39	68.39
AWL 570	96	5.33	73.72
Otras	473	26.28	100
TOTAL:	1800		

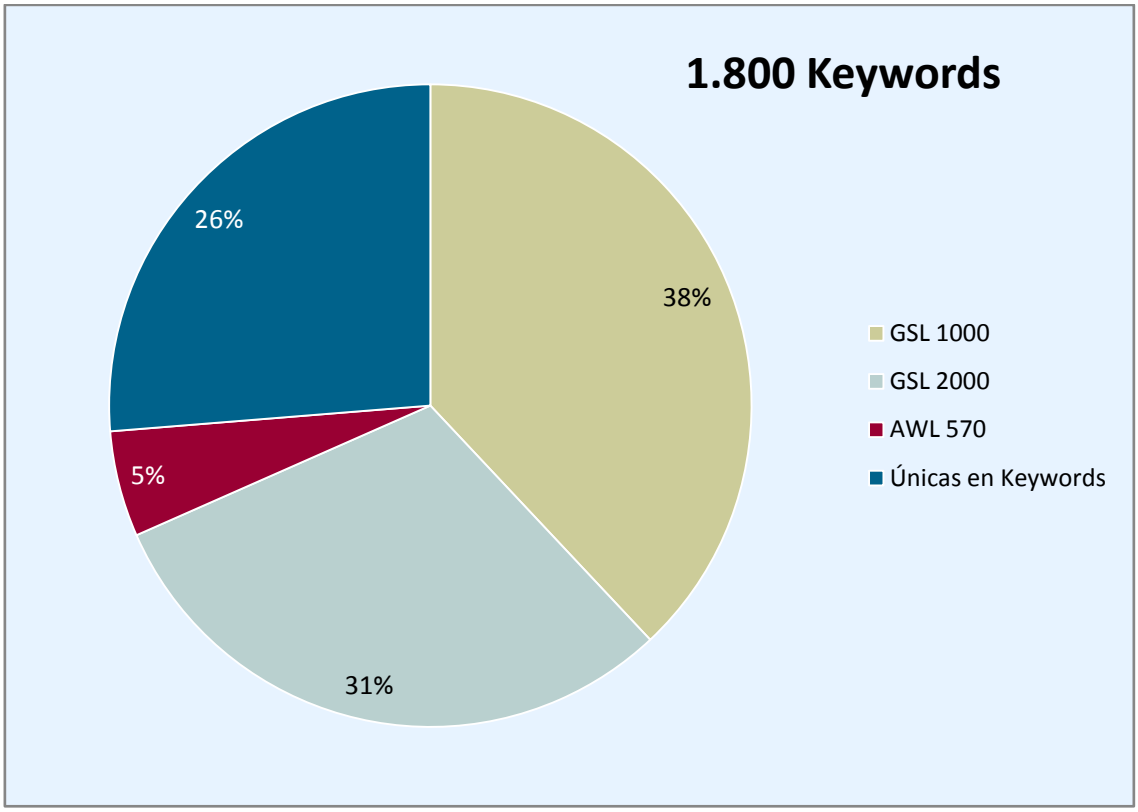


Figura II.4. Distribución de Keywords sobre la GSL y la AWL

Tomando la referencia inversa, de las 2.000 palabras de la GSL, el 61% son Keywords. En la AWL suponen el 17%, como se aprecia en la Figura II.5. En el Apéndice C se encuentran las Keywords que pertenecen a cada nivel de la GSL, a la AWL y los valores únicos.

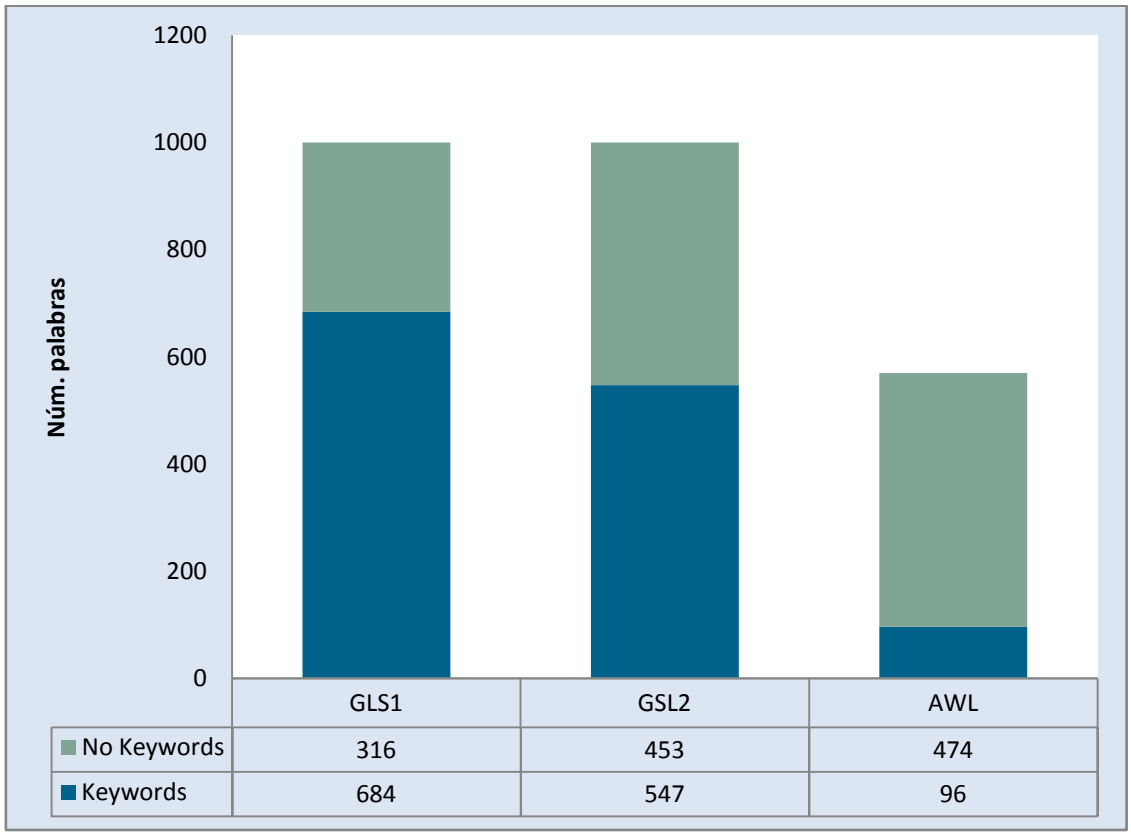


Figura II.5. Distrubución de la GSL y la AWL sobre Keywords

4.2. El plan léxico basado en la lista

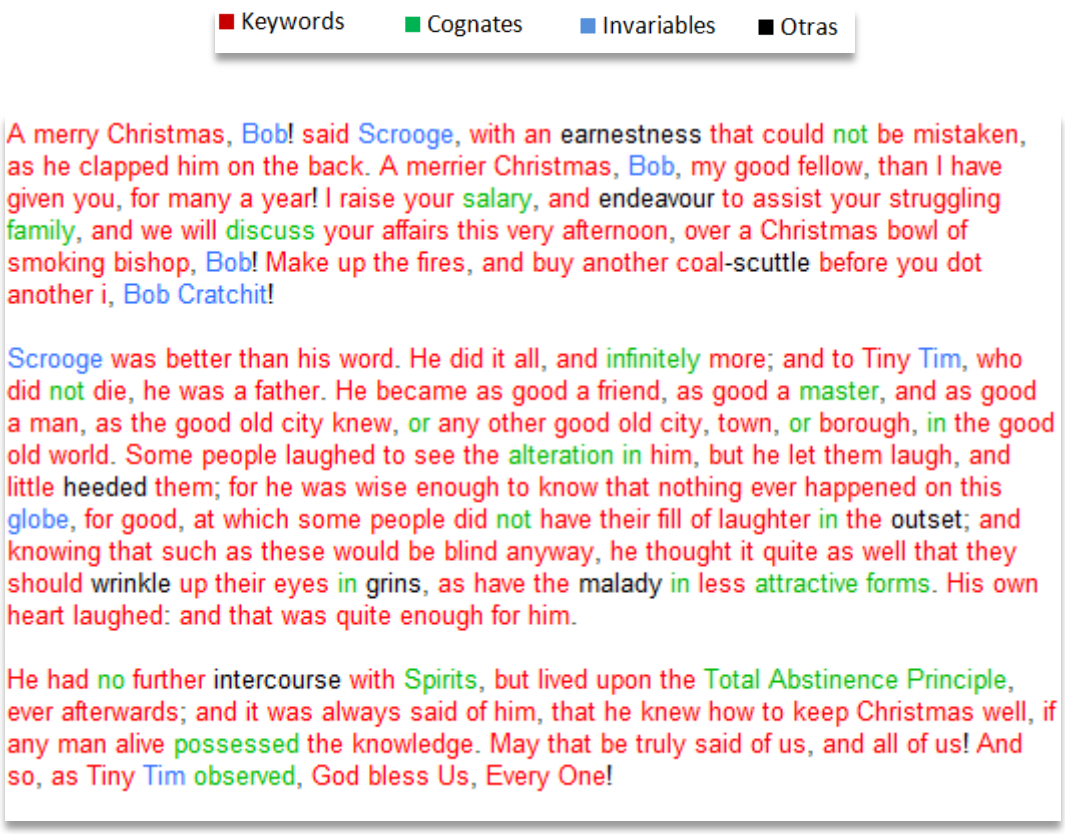
Esta sección estudia la eficacia de este listado como base sobre la que construir una planificación docente eficaz de inglés como L2 diseñada explícitamente para aprendices hispanohablantes. Utilizaremos la expresión *Programa Léxico Adaptado para Hispanohablantes* (en adelante, PLH) para referirnos a una programación que contiene las palabras de la lista de vocabulario prioritario Keywords y las reglas de derivación de Bauer y Nation (1993) que permiten reconocer e interpretar la categoría gramatical de los cognados. Nótese que el PLH no es una metodología, sino un objetivo

de aprendizaje que se puede alcanzar a través de distintas técnicas de instrucción explícita. A la hora de interpretar los resultados es importante remarcar que en este estudio asumimos que el alumno hispanohablante es capaz de deducir el significado de los cognados evidentes inglés-español y de identificar los nombres propios que encuentre en un texto dado. Al terminar el PLH, por tanto, el alumno debería reconocer todas las palabras englobadas en las categorías marcadas con el símbolo ✓ en la

Figura II.6. Únicamente deberían ser desconocidas las palabras no cognadas que sean de baja frecuencia.

Figura II.6. Esquema de categorías del PLH

Para poner estos datos en perspectiva, el siguiente texto es el final de la novela *A Christmas Carol* de Charles Dickens con un código de colores que corresponde a cada categoría del PLH. En principio, un alumno que termina el PLH entendería todo el vocabulario excepto las palabras en negro.



A continuación presentamos los resultados sobre la rentabilidad del programa PLH en la comprensión de distintos géneros de interés para el aprendiz de inglés como L2, tales como exámenes de certificación de nivel, obras literarias originales y adaptadas, películas, series de TV y discurso formal. Los resultados se presentan utilizando como referencia el segundo y tercer umbral descritos por Laufer y Nation. Valorando el alcance estimado para cada uno de ellos, utilizaremos los términos *umbral de autonomía* para referirnos a una cobertura que abarque el 95% de las palabras, y *umbral de garantía* para designar una cobertura mayor o igual al 98%.

4.3. Análisis de exámenes

La presente sección presenta la cobertura léxica que aporta PLH con las pruebas de comprensión escrita de 24 exámenes correspondientes al nivel «usuario independiente» fijado por el Marco común europeo de referencia para las lenguas del Consejo de Europa (en adelante, MCERL). El marco distingue entre tres amplios

bloques de competencia comunicativa, llamados A: usuario básico, B: usuario independiente y C: usuario competente. Cada bloque, a su vez, está dividido en dos niveles. En este trabajo nos centramos en la destreza «comprensión de lectura» de los niveles B1 y B2, para los que el MCERL especifica los descriptores que aparecen en la Tabla II.12.

Tabla II.12. Descriptores de competencia de lectura del MCERL

B1: umbral	Es capaz de comprender los puntos principales de textos claros y en lengua estándar si tratan sobre cuestiones que le son conocidas, ya sea en situaciones de trabajo, de estudio o de ocio.
B2: avanzado	Es capaz de entender las ideas principales de textos complejos que traten de temas tanto concretos como abstractos, incluso si son de carácter técnico siempre que estén dentro de su campo de especialización.

Fuente: Marco común europeo de referencia (Instituto Cervantes, 1997b)

4.3.1. Análisis de perfil léxico de los exámenes B1 y B2

a) Perfil del corpus B1.

El corpus de exámenes B1 contiene 4.450 palabras (tokens). La Tabla II.13 muestra su distribución a lo largo de las nueve primeras familias de 1.000 palabras frecuentes del BNC. En la tercera columna se observa que solamente el primer nivel aporta cuatro de cada cinco palabras que aparecen en los exámenes B1. A pesar de que todos los niveles contienen 1.000 familias léxicas, la presencia de tokens del segundo nivel es aproximadamente diez veces menor, y se observa que la progresión general es decreciente, confirmando el fenómeno descrito por la Ley de Zipf. Así, las 1.000 palabras del nivel 5 aparecen representadas únicamente por 14 tokens, el 0.31% del corpus. Los 149 nombres propios y otras palabras invariables suponen el 3,35% de los tokens. Asumiendo que se deducen los nombres propios, el umbral de autonomía (95%) se alcanzaría con las primeras 3.000 palabras del BNC. Para llegar al umbral de garantía (98%) serían necesarias 5.000 palabras. Las palabras que no pertenecen a los tres primeros niveles se consideran, generalmente, de baja frecuencia.

Tabla II.13. Distribución de frecuencia en el corpus de exámenes B1

NIVEL	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	149	3.35	3.35
BNC-1	3648	81.98	85.33
BNC-2	371	8.34	93.67
BNC-3	146	3.28	96.95
BNC-4	43	0.97	97.92
BNC-5	14	0.31	98.23
BNC-6	17	0.38	98.61
BNC-7	7	0.16	98.77
BNC-8	5	0.11	98.88
BNC-9	1	0.02	98.9
???*	49	1.1	100
TOTAL	4450		

* Palabras que no pertenecen a ninguna lista.

La Tabla II.14 muestra la forma base de algunas de las palabras del corpus que pertenecen al cuarto nivel y superiores. Nótese la presencia de cognados transparentes entre ellas.

Tabla II.14. Palabras de baja frecuencia en el corpus de exámenes B1

NIVEL	PALABRAS DE BAJA FRECUENCIA
BNC-4	EVOKE, HAUNT, SCRUTINY, LEISURE
BNC-5	BROCHURE, COMMEND, DUPLICATE, RENOWN
BNC-6	BROWSE, SLIPPERY, FANATIC, PROLIFERATE
BNC-7	GADGET, OVERLOAD, SKIP
BNC-8	BENEFACTOR, GOURMET, TYRE, OUTDATED
INVARIABLES	ADAM, KORD, TATTERBRIDGE, CALIFORNIA[26]

[26] En los textos originales, Kord es un nombre de pila (Kord Campbell) y Tatterbridge es el nombre de una ciudad.

Entre los elementos que no pertenecen a ninguna lista encontramos casi exclusivamente palabras compuestas, bien por dos lexemas (*countryside, headteacher, motorway, slideshow*) o por un lexema y afijos greco-latinos (*incoming, triplicated, parapsychology*). Si bien técnicamente son palabras de muy baja frecuencia en el BNC, algunas de ellas podrían ser fácilmente deducibles siempre que se conozca el significado de las partes que las componen. Por el contrario, otras palabras compuestas encontradas pueden inducir a error, como *workshop* (taller), o *countryside,* (campo), que el aprendiz podría interpretar como 'tienda de trabajo' y 'frontera' [27], respectivamente.

Perfil del corpus B2

El corpus de exámenes B2 contiene 5.529 tokens, repartidos entre los distintos rangos de frecuencia del BNC en la proporción que se muestra en la Tabla II.15. Igual que habíamos observado en los exámenes B1, la gran mayoría de los tokens del corpus B2 pertenecen al primer nivel y el resto están distribuidos en una progresión descendente. El primer nivel aporta casi el 80% del corpus, mientras que el nivel 5 únicamente representa el 0,61%. Los nombres propios suponen el 3,4%, una proporción muy similar al corpus B1.

Asumiendo que se conocen los nombres propios, el umbral de autonomía (95%) se alcanzaría con las primeras 3.000 palabras del BNC, igual que en el nivel B1. Por el contrario, para el umbral de garantía (98%) serían necesarias 6.000 palabras, 1.000 más que en el corpus B1. La Tabla II.16 muestra los lemas de algunas palabras de baja frecuencia del corpus B2.

[27] *Side* significa 'lado', *country* es una palabra polisémica que puede significar 'campo' o 'país'. Un alumno podría interpretar erróneamente que *countryside* es 'el lado del país', es decir, la frontera.

Tabla II.15. Distribución de frecuencia en el corpus de exámenes B2

LISTA	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	196	3.54	3.54
BNC-1	4387	79.35	82.89
BNC-2	473	8.55	91.44
BNC-3	210	3.8	95.24
BNC-4	94	1.7	96.94
BNC-5	34	0.61	97.55
BNC-6	44	0.8	98.35
BNC-7	23	0.42	98.77
BNC-8	7	0.13	98.9
BNC-9	16	0.29	99.19
???	45	0.81	100
TOTAL	5529		

Tabla II.16. Palabras de baja frecuencia en el corpus de exámenes B2

NIVEL	PALABRAS DE BAJA FRECUENCIA
BNC-4	NUTRITION, BUBBLE, MANUSCRIPT, CLAY
BNC-5	APPETITE, FLATTER, OFFSPRING, NURTURE
BNC-6	DUMMY, TURBINE, CAVITY, KNOB
BNC-7	STUNT, INTOXICATE, PELVIS, QUAINT
BNC-8	FIANCÉE, PERTURB, REVOLVER, TETHER
BNC-9	ABYSS, THERMOSTAT, INNOCUOUS, SCOURGE
NOMBRES	JOHN, MEDITERRANEAN, TITANIC, IPHONE

Paralelamente a lo observado en el corpus B1, el conjunto de tokens no pertenecientes a ninguna categoría está formado mayoritariamente por palabras compuestas, tales como *checkpoint, shipwrecked, tailbone, waterfall* y *widespread* [28]. También encontramos fuera de las listas del BNC vocabulario relativamente moderno como *blog, online, dvd* o *dj*. Sin embargo, a diferencia del corpus B1, en el nivel B2 sí aparecen términos ordinarios —no compuestos ni neologismos— cuya frecuencia es tan baja que no se encuentran entre las 9.000 primeras posiciones del BNC. Algunas de ellas son *imbibing*, *missives*, *monsoon*, *providence* y *trepidation*[29].

Con todo, la distribución del vocabulario en el corpus B2 se asemeja tanto a la del B1 que sus representaciones gráficas son prácticamente coincidentes, como se aprecia en la Figura II.7. La curva de frecuencia del corpus B1, representada por una línea sólida gris, se superpone en casi todo su recorrido a la curva de B2, negra y punteada. Las gráficas habituales que utilizan la escala lineal como la Figura II.7 no son muy útiles para comparar distribuciones de corpus porque las diferencias suelen resultar inapreciables. Debido a la Ley de Zipf, la distribución rango/frecuencia de cualquier texto dibuja una curva en la que el orden de magnitud de los primeros niveles siempre es enorme con respecto a los últimos, por lo que las diferencias al final del eje *x* son prácticamente imperceptibles. En el caso que nos ocupa, el primer nivel de los corpus B1 y B2 ronda los 4.000 tokens, mientras que en el último nivel solo tienen 1 y 16 tokens, respectivamente. En una escala sobre 4.000, una diferencia de 15 queda tan comprimida que resulta inapreciable, pero es un dato muy importante que puede determinar que un texto sea comprensible y el otro no.

Cuando los datos tienen órdenes de magnitud tan distintos, tomar logaritmos es una buena opción para reducirlos a un sistema más manejable. Toda distribución de Zipf es una función exponencial, por lo que la mejor manera de apreciar variaciones de forma visual es dibujar una gráfica de dispersión utilizando una escala semilogarítmica

[28] 'punto de control', 'naufragio', 'coxis', 'catarata' y 'extendido', respectivamente.

[29] Nótese que, excepto *imbibing*, las cuatro restantes son cognados, y al menos tres de ellas (*missives, providence* y *trepidation*) son deducibles. En los niveles de baja frecuencia se encuentra el registro de habla culta, donde se observa una gran presencia de palabras de raíz latina.

(eje *y*) o doble-logarítmica (ambos ejes)[30]. Esta escala divide el eje de ordenadas en subintervalos que no son iguales como en la escala lineal, sino que crecen de forma exponencial (1, 10, 100...). La Figura II.8 representa la gráfica en escala semilogarítmica de los mismos datos de la Figura II.7, pero ahora se aprecian claramente las diferencias que presentan el corpus B1 y B1, especialmente en el nivel 9.

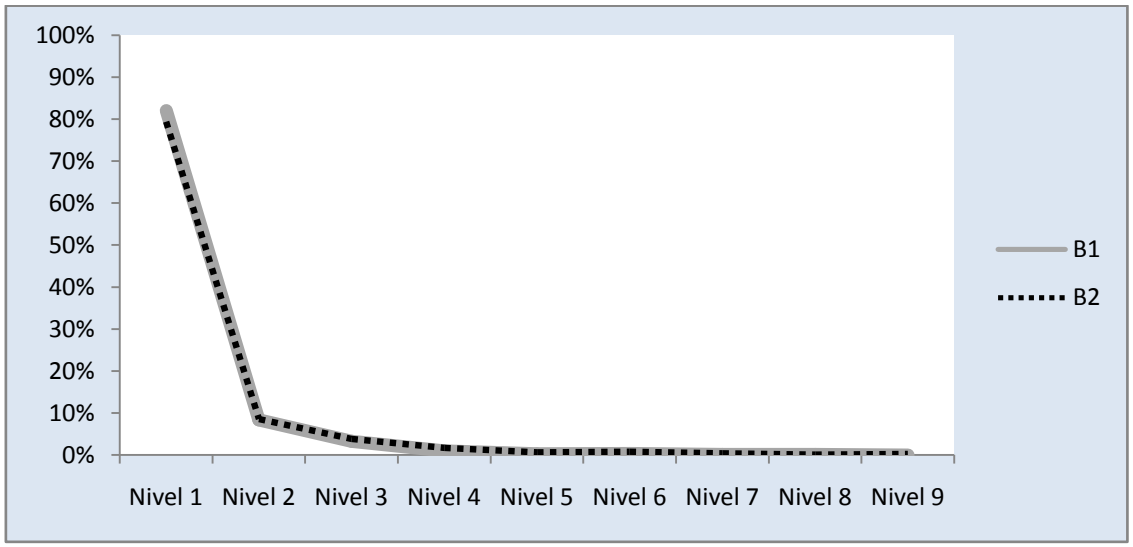


Figura II.7. Distribución de frecuencia en los corpus B1 y B2.

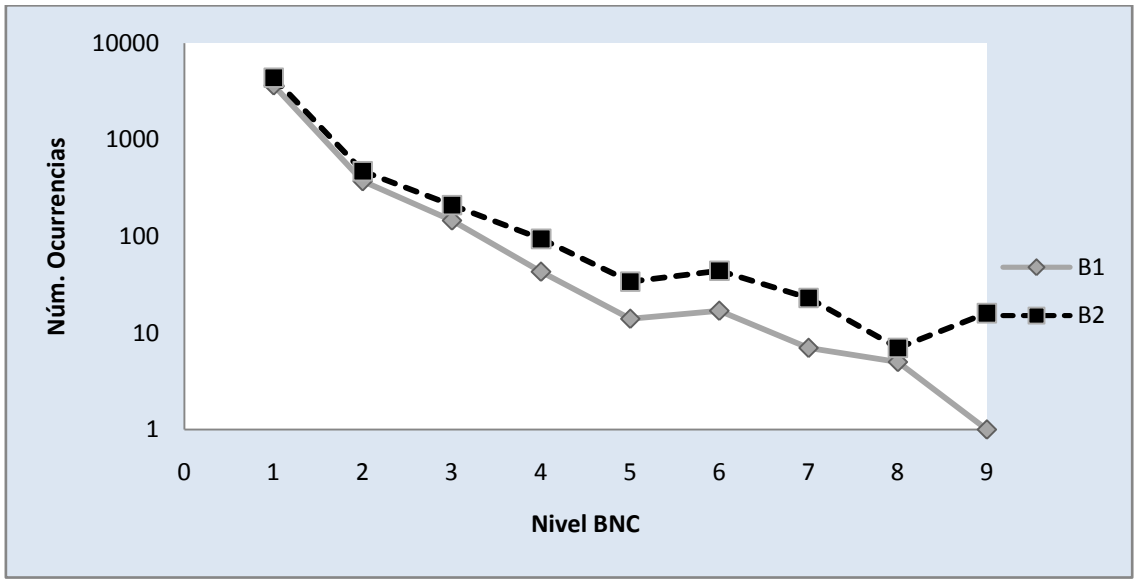


Figura II.8. Distribución de frecuencia en corpus B1 y B2 en escala semilogarítmica

[30] Salvo que se indique lo contrario, la base de todas las representaciones logarítmicas del presente trabajo es decimal.

4.3.2. Influencia de L1 y alcance del PLH

La Tabla II.17 y la Tabla II.18 muestran palabras de los exámenes identificadas como cognados deducibles en los exámenes B1 y B2, respectivamente. Aparecen en su forma flexionada tal y como se encuentran en los textos originales.

Tabla II.17. Palabras deducibles en el corpus de exámenes B1

ABSOLUTELY	CONSIDERATION	FUNDAMENTAL	OPPORTUNITIES	RESTAURANT
ACADEMY	CONTINUES	GALLERIES	OPTION	ROMANTIC
ACTION	CORRESPONDING	HELICOPTER	ORCHESTRA	SERIOUS
ACTIVITY	COURTESY	HISTORIC	ORGANISATION	SIGNIFICANT
ADDICTIVE	CREATIVITY	HOBBIES	PARAPSYCHOLOGY	SIMPLE
ADULTS	CYCLISTS	HOTEL	PART	SOCIAL
AFFECT	DEBATE	IDENTIFY	PERSONAL	SOUVENIRS
ALARM	DECIDED	ILLUSTRATION	PLANS	STIMULATION
APPRECIATE	DECISIONS	IMPOSSIBLE	POPULAR	STUDENTS
AREA	DIFFERENT	IMPULSE	PRIMITIVE	SUPERIOR
ARTISTS	DISTANCE	INDUSTRIAL	PROBABLY	TAXI
ATTRACTIVE	DISTRACTIONS	INFORMATION	PROBLEM	TECHNOLOGY
BENEFACTOR	DOCTOR	INTERNET	PRODUCTS	TENSIONS
CALORIES	EDUCATION	IRRELEVANT	PROGRAMME	TEXT
CAMERA	ELECTRONIC	LANGUAGES	PROVOKE	TEXTILES
CINEMA	EVENTS	LOCAL	PUBLIC	TRADITIONAL
COLOURS	EXAM	MARATHON	QUESTIONNAIRE	TRANSPORT
COMMUNICATION	EXCELLENT	MILLION	RADIO	TRIANGULAR
COMPARED	EXPERIENCE	MINUTES	REACTIONS	TUTORS
COMPLETING	FAMILY	MODERN	RECTANGULAR	VISIT
CONCERT	FANTASTIC	MOMENTS	RELAXATION	VISITORS
CONNECTED	FINAL	NECESSARILY	RESIDENTS	WIFI

Tabla II.18. Palabras deducibles en el corpus de exámenes B2

ACCIDENT	CALCULATIONS	FOSSIL	PARTICULAR	SYSTEM
ACTION	CANDID	HUMAN	PELVIS	TELEPHONE
ADDITIONALLY	CIRCULATES	IDEAL	PERSONALISED	TERRIBLE
ADJUST	CONDENSED	INSTRUCTOR	PORTIONS	TOMATOES
ADMIT	CONFESSED	INTERNET	POSITIONED	TRADITIONAL
AIR	CONTROL	LOCAL	PREDICTIONS	TRAFFIC
ALBUM	DECLARED	MANUAL	PROGRAM	TRANSPORT
ALCOHOL	DIFFERENT	MAXIMUM	PROTECTION	TROPICAL
AMBITION	DIGITAL	METHODS	PYRAMIDS	UNIQUE
ANIMAL	DIRECTOR	MILLIONS	RAPID	UNIVERSITY
APART	DISTRIBUTOR	MONSTERS	REGION	VAPOUR
APARTMENT	ECONOMIES	MUSICAL	ROMANTIC	VEHICLE
APPARENT	EDUCATION	NORMAL	ROUTINE	VIBRATION
APPROXIMATELY	EMISSIONS	OCCASIONAL	RUMOR	VIDEO
ARCHAEOLOGY	ENTHUSIASM	OPERATED	SCENE	VITAMINS
AREA	EXIST	OPINION	SIMPLE	WHISKY
BASED	EXTREMELY	OPTIONS	SUBMARINE	ZOOLOGY

El conjunto de palabras deducibles supone un porcentaje considerable de ambos corpus, como muestra la Tabla II.19. Los cognados transparentes suponen casi uno de cada cinco tokens del corpus B1 y se aprecia una proporción muy similar en el corpus B2, con 1.054 cognados que representan el 19,14% de los tokens.

Tabla II.19. Distribución de categorías del PLH en el corpus B1 y B2

	TOKEN		TOKEN%		CUMTOKEN%	
CORPUS	**B1**	**B2**	**B1**	**B2**	**B1**	**B2**
KEYWORDS	3387	4116	76.11	74.44	76.11	74.44
COGNATES	858	1058	19.28	19.14	95.39	93.58
INVARIABLES	149	196	3.35	3.54	98.74	97.12
???	56	159	1.26	2.88	100	100
TOTAL	4450	5529				

En el corpus B1 hay 3.387 palabras clasificadas como vocabulario clave por la lista Keywords, que suponen el 76,11% del corpus. En otras palabras, tres de cada cuatro palabras que el alumno encontrará en un examen es un término de Keywords. Junto con las palabras deducibles (cognados, nombres propios y palabras invariables), la cobertura del PLH alcanza el 98,74%, sobrepasando así tanto el umbral de autonomía como el de garantía.

El corpus B2 mantiene un porcentaje de cognados y nombres propios muy similar al B1. Sin embargo, la proporción de Keywords en B2 presenta una variación de -1,74% respecto del primero. Esta pequeña diferencia es la razón por la cual, si bien en ambos corpus se rebasa el umbral de autonomía, en el B2 no alcanza el umbral de garantía (98%) por -0,88 puntos porcentuales.

Sin embargo, al estudiar las palabras presuntamente desconocidas del corpus B2 encontramos que algunas de ellas (Tabla II.20) son potencialmente deducibles, especialmente si están contextualizadas. La primera columna presenta tres cognados evidentes que el software no detectó. La segunda columna muestra palabras compuestas por lemas pertenecientes a Keywords, lo que las haría deducibles, aunque plantean ciertas dudas los términos *checkpoint* y *courtroom*. La tercera columna contiene cognados no transparentes que, sin embargo, aún mantienen cierto grado de similitud con la L1. La cuarta columna contiene cuatro casos especiales. *Girlfriend* es una palabra que, técnicamente, es de muy baja frecuencia en el corpus y que, además, podría inducir a error si se traducen sus lexemas por separado; sin embargo, indudablemente es un término de alta familiaridad que se aprende en las etapas más elementales, por lo que se podría asumir que hay muchas posibilidades de que el alumno la conozca. Por otra parte, *grill, mailing y topping* son voces inglesas tan incorporadas al español que se podrían considerar en la misma categoría que los cognados evidentes.

Podemos asumir que el alumno es capaz de deducir algunas de estas palabras, fundamentalmente las de las columnas 1 y 4. Esto conllevaría un aumento de la cobertura del PLH en el corpus B2, que quedaría entre 0,75 y 0,39 puntos por debajo del umbral de garantía, lo que parece un intervalo asumible. La Figura II.9 ilustra el alcance de cada categoría del PLH en los corpus B1 y B2.

Tabla II.20. Palabras potencialmente deducibles del corpus B2

COGNADOS TRANSPARENTES	COMPUESTAS	COGNADOS NO TRANSPARENTES	CASOS ESPECIALES
ORBITERS	CHECKPOINT	CARRIAGE	GIRLFRIEND
PANIC	COURTROOM	DAILY	GRILL
VOLCANOES	DRUNKEN	FLIRTATIOUS	MAILING
	FINGERTIPS	FLIRTING	TOPPING
	FIREARMS	ISOLATED	
	HORSEBACK	MISADVENTURES	
	OVERCOOKING	ROBBER	
	SNOWFALLS	TECH	
	WIDESPREAD	TONNES	
		TONNESS	
		UNPRECEDENTED	

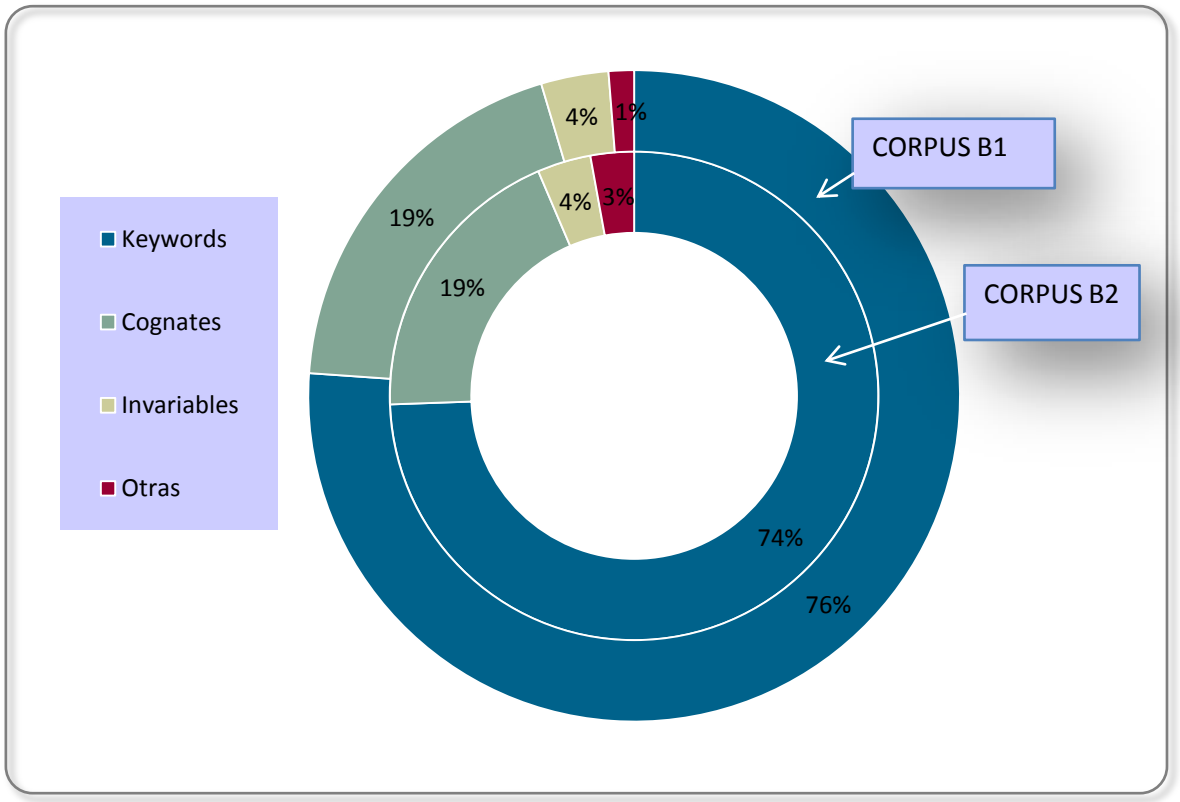


Figura II.9. PLH en los corpus B1 y B2

4.3.3. Eficacia del PLH frente a la lista BNC

En esta sección se muestra el alcance del PLH en los distintos exámenes y comparamos su rentabilidad frente a lo que supondría un programa basado en las listas de frecuencia que no tengan en cuenta la L1 del alumno. La Tabla II.21 toma como referencia la cobertura del PLH en los distintos exámenes y señala cuántos niveles del BNC serían necesarios para alcanzar los mismos resultados. La primera columna identifica los corpus, la segunda columna muestra la cobertura alcanzada por el PLH y la tercera es el nivel del BNC que aportaría al menos la misma cobertura.

Tabla II.21. Niveles del BNC necesarios para igualar al PLH

CORPUS	PLH %	NECESARIO BNC
B1 CAMBRIDGE	99.60	7000
B1 TRINITY	98.34	6000
B1 ANGLIA	99.69	6000
B1 EOI	97.48	5000
B2 CAMBRIDGE	99.24	4000
B2 TRINITY	98.89	3000
B2 ANGLIA	96.61	5000
B2 EOI	94.57	5000
B1 CORPUS	98.74	7000
B2 CORPUS	97.12	4000

Esta tabla nos permite comparar la rentabilidad del PLH como herramienta docente frente a un programa léxico basado en las listas originales. Tomemos como ejemplo el corpus completo de exámenes B1, hacia el final de la tabla. El PLH, con un objetivo de aprendizaje limitado a 1.800 palabras, cubre el 98,70% de las palabras del corpus. Para obtener ese porcentaje utilizando las listas del BNC serían necesarias 7.000 palabras. Esto implica que para un alumno hispanohablante el PLH tiene una eficacia 388% superior a las listas BNC.

Incluso en el caso de menor exigencia comparativa, el B2 de Trinity Guildhall, la diferencia son 1.800 palabras frente a 3.000, es decir, la rentabilidad docente del PLH es un 166% superior al BNC.

La Tabla II.22 muestra el alcance del PLH y el BNC respecto de los umbrales en los distintos exámenes. En las columnas PLH un número positivo señala por cuántos puntos porcentuales se ha superado el umbral; uno negativo, cuánto falta para alcanzarlo. Las columnas del BNC indican cuántos niveles de 1.000 familias del BNC son necesarios para alcanzar cada umbral.

Tabla II.22. Umbrales alcanzados por el PLH y el BNC en exámenes

	UMBRAL 95 DIF. PHL%	UMBRAL 98 DIF. PHL%	UMBRAL 95 NIVEL BNC	UMBRAL 98 NIVEL BNC
B1 CAMBRIDGE	4.6	1.6	2	3
B1 TRINITY	3.34	0.34	3	6
B1 ANGLIA	4.69	1.69	2	4
B1 EOI	2.48	-0.52	3	6
B2 CAMBRIDGE	4.24	1.24	2	3
B2 TRINITY	3.89	0.89	3	4
B2 ANGLIA	1.61	-1.39	4	7
B2 EOI	-0.43	-3.43	5	8

Se observa que el PLH supera el umbral de autonomía en todas las pruebas excepto en el B2 de la Escuela Oficial de Idiomas. Además, aunque excede los objetivos planteados para el PLH, se puede comprobar que también superaría el de garantía en 5 de los 8 corpus. Para interpretar este dato en su contexto es importante remarcar las grandes diferencias que hay entre exámenes que, aparentemente, pertenecen al mismo nivel del MCERL. Las pruebas de Cambridge ESOL con respecto a la Escuela Oficial de Idiomas presentan una variación de cobertura considerable, que va del 2,12% en el nivel B1 hasta el 4,67% en el nivel B2.

La razón de esta diferencia se observa claramente en su perfil léxico en la columna BNC: superar los umbrales en las pruebas de la Escuela Oficial de Idiomas requiere muchísimo más vocabulario del exigido por Cambridge ESOL. En el umbral de autonomía encontramos una diferencia de 1.000 palabras en B1, y de 3.000 en B2. Para el umbral de garantía la distancia es aún mayor, 3.000 palabras en B1 y 5.000 en B2. La Figura II.10 muestra las palabras no alcanzadas por el PLH e ilustra la diferencia en la exigencia léxica de los niveles supuestamente equivalentes.

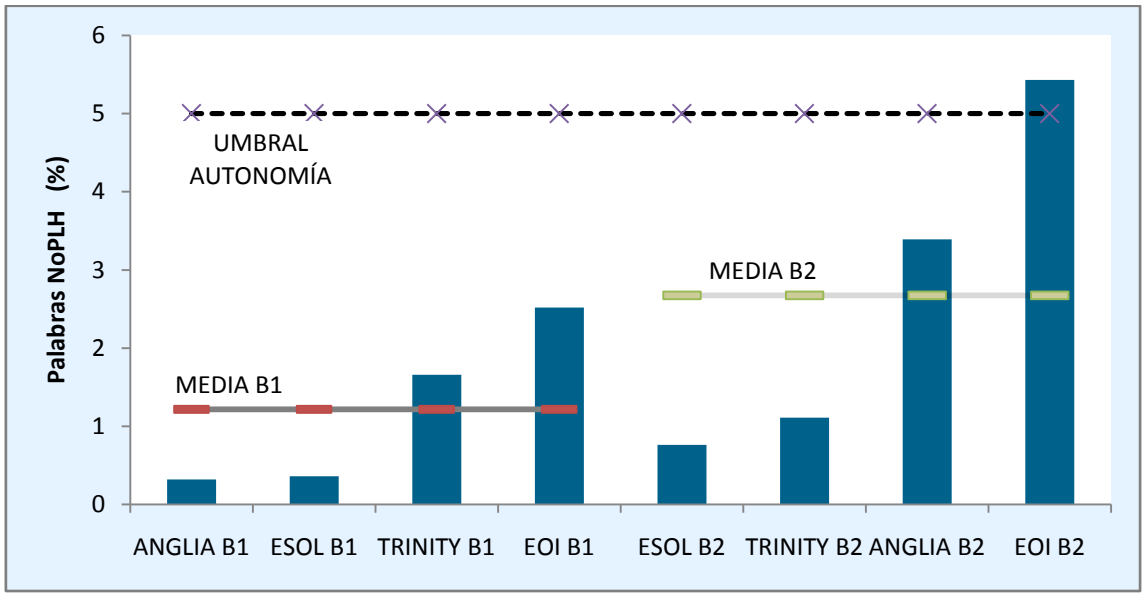


Figura II.10. Porcentaje de palabras NoPLH en los exámenes

Una de las razones que puede explicar esto es que el Consejo de Europa, responsable del MCERL, no valida los niveles de los organismos, sino que es cada escuela la que atribuye a sus propios cursos el nivel del MCERL al que consideran que corresponden. Sin entrar a valorar si la Escuela Oficial de Idiomas está correctamente baremada con respecto del Marco de referencia, es indudable es que el vocabulario que aparece en sus pruebas es mucho más exigente que el de los demás exámenes supuestamente equivalentes. Esto puede ser la razón por la cual el PLH presenta sus peores resultados en las pruebas de la Escuela Oficial de Idiomas y en el nivel B2 no consigue alcanzar el umbral de autonomía.

Por otra parte, las pruebas de Trinity Guildhall presentan una incoherencia en cuanto al perfil léxico. El PLH ofrece más cobertura en el nivel B2 que en el B1, que, en teoría, debería ser menos exigente. Este fenómeno se podría atribuir a una presencia inusual de cognados en las pruebas B2, pero la columna derecha de la Tabla II.22 revela que la explicación es sencillamente que hay más proporción de palabras de baja frecuencia en las pruebas B1. Concretamente, este nivel requiere 2.000 palabras más que el B2 para alcanzar el umbral de garantía.

Hay que ser cautelosos al interpretar este curioso dato, ya que no implica necesariamente que las pruebas de comprensión lectora de Trinity Guildhall B1 sean realmente más difíciles de superar que las B2; de hecho, muy probablemente no sea así. Además de que el vocabulario no es único factor del que depende la comprensión de un texto, hay otros elementos que se deberían valorar, tales como el vocabulario que aparece en las propias preguntas, no solamente en el texto. Sin embargo, todo parece indicar que las pruebas de Trinity Guildhall no controlan el vocabulario según su frecuencia sino que utilizan otros criterios.

En el otro extremo, Cambridge ESOL es el examinador que presenta mayor coherencia en el vocabulario exigido para cada nivel. Hay que remarcar que Cambridge ESOL tiene un proyecto llamado English Profile dedicado exclusivamente a establecer una correspondencia rigurosa y transparente entre sus propios niveles y los descriptores del MCERL. Tiene una sección específica para el análisis léxico, English Vocabulary Profile, que mide frecuencias en corpus de referencia con el fin de elaborar listados de vocabulario adecuado para cada uno de los distintos niveles del MCERL. La descripción del proyecto indica lo siguiente:

> The English Vocabulary Profile offers reliable information at word and sense level, based on extensive analysis of word frequency and learner use, using the Cambridge English Corpus (formerly known as the Cambridge International Corpus), the British National Corpus and the Cambridge Learner Corpus, together with other sources, including the Cambridge English Language Assessment vocabulary lists and classroom materials.
> (http://www.englishprofile.org/index.php/resources/wordlists)

> English Vocabulary Profile ofrece información fiable sobre el nivel léxico semántico basada en el análisis exhaustivo de frecuencia de las palabras y su uso por parte del aprendiz. Se utilizan el Cambridge English Corpus (antiguamente llamado Cambridge International Corpus), el British National Corpus y el Cambridge Learner Corpus junto con otras fuentes que incluyen las listas de vocabulario de Cambridge English Language Assessment y materiales docentes (trad. a.).

4.3.4. Conclusiones del análisis de exámenes

El análisis de los exámenes equivalentes al nivel «usuario independiente» del MCERL nos revela tres datos clave en cuanto a su perfil léxico y la eficacia del PLH como herramienta docente:

En primer lugar, los datos demuestran que la cantidad de vocabulario exigida para un nivel concreto del MCERL difiere considerablemente entre los distintos examinadores. En la Figura II.11, que muestra el porcentaje de palabras desconocidas según la institución, se aprecia claramente que la Escuela Oficial de Idiomas es, con mucho, el más difícil en ambos niveles. En el otro extremo, Cambridge ESOL es el organismo examinador que presenta el nivel de exigencia más adecuado a los descriptores del MCERL y mayor consistencia entre todas sus pruebas del mismo nivel.

En segundo lugar, concluimos que aprender los 1.800 términos de la lista Keywords proporciona al alumno hispanohablante la competencia léxica necesaria para alcanzar el umbral de autonomía en todas las pruebas de comprensión lectora tanto de B1 como de B2, excepto en las diseñadas por la Escuela Oficial de Idiomas. Es más, incluso se supera el objetivo del PLH, ya que también rebasa el umbral de garantía del 98% en la mayoría de las pruebas.

Por último, podemos afirmar que Keywords es un recurso mucho más eficiente y rentable que las listas de frecuencia genéricas cuando se trata de alumnos hispanohablantes. La cobertura que ofrece el PLH en los exámenes B1 y B2 es tan amplia que, para obtener el mismo resultado, el alumno necesitaría aprender entre 3.000

y 7.000 palabras del BNC, una carga léxica considerable al compararla con las 1.800 palabras clave del PLH.

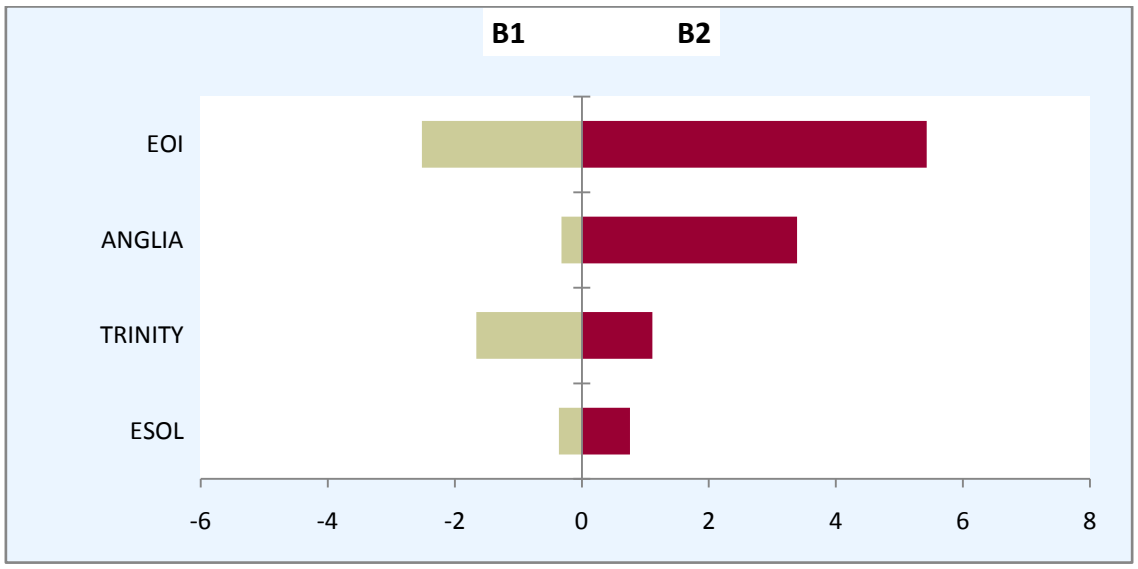


Figura II.11. Palabras fuera del PLH (%) en exámenes B1 y B2

4.4. Análisis de obras literarias originales y lecturas adaptadas

Esta sección presenta la cobertura del PLH en distintas obras literarias. Se escogieron dos textos para el público universal, *The Advertures of Sherlock Holmes*, de Sir Arthur Conan Doyle y *A Christmas Carol*, de Charles Dickens; así como una obra destinada exclusivamente a un público adulto, *Lady Chatterley's Lover*[31], de D.H. Lawrence. Los textos se obtuvieron del Proyecto Gutenberg[32]. Posteriormente, se presenta el análisis comparativo de las dos primeras obras junto con sus versiones adaptadas para estudiantes de inglés como L2. En conjunto, este corpus de obras literarias contiene un total de 653.700 palabras.

[31] *El amante de Lady Chatterley*, hoy considerado un clásico de la literatura, fue un escándalo tras su primera edición en Florencia en 1928. La primera edición sin censura no se publicó en Reino Unido hasta 1960, decisión que provocó que la editorial Penguin Books fuera llevada a juicio por obscenidad.

[32] El Proyecto Gutenberg es un repositorio on-line sin ánimo de lucro que alberga obras literarias y otros textos de dominio público. Contiene fundamentalmente clásicos digitalizados de la literatura universal cuyo copyright ha vencido. (http://www.gutenberg.org/)

4.4.1. Análisis léxico de The Adventures of Sherlock Holmes

La Tabla II.23 presenta el perfil léxico de la versión original de *The Adventures of Sherlock Holmes* tomando como referencia la frecuencia relativa en el BNC. Observamos que tiene un total de 105.249 tokens, y que el umbral de autonomía se alcanza con 4.000 palabras, asumiendo que se identifican los nombres propios.

Tabla II.23. Distribución léxica de Sherlock Holmes según el BNC

FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	2412	2.29	2.29
BNC-1	89298	84.84	87.13
BNC-2	5799	5.51	92.64
BNC-3	2264	2.15	94.79
BNC-4	1616	1.54	96.33
BNC-5	1030	0.98	97.31
BNC-6	654	0.62	97.93
BNC-7	401	0.38	98.31
BNC-8	340	0.32	98.63
BNC-9	231	0.22	98.85
BNC-10	153	0.15	99
???	1051	1	100
TOTAL	105249		

Una aplicación interesante del análisis léxico es que nos permite intuir el tema central de un texto en función de los patrones inusuales que presenta el vocabulario de baja frecuencia, siempre que se trate de un texto relativamente grande. Sabemos que en cualquier tipo de discurso la probabilidad indica que aproximadamente el 80% de sus tokens pertenecen al primer nivel del BNC y que, además, muchos de ellos se repiten varias veces. Sin embargo, cuando en un texto concreto hay ciertas palabras a partir del segundo nivel que presentan un alto número de ocurrencias y están relacionadas semánticamente, podemos intuir el escenario y el tema sobre el que gira la obra. En análisis del discurso estas se denominan palabras temáticas o *topic words*.

En la obra *The Adventures of Sherlock Holmes,* el primer nivel junto con los nombres propios supone el 87,13% de los tokens, por lo que el estudio de los patrones inusuales en el 12,87% restante nos permitirá deducir el género del texto. La Tabla II.24 muestra algunas de las palabras temáticas extraídas de entre los términos que deberían ser de baja frecuencia y, sin embargo, en el texto tienen un número de ocurrencias inusitadamente alto

Tabla II.24. Palabras temáticas en The Adventures of Sherlock Holmes

ADVENTURE	DETAIL	INVESTIGATE	RESULT
ARREST	DISAPPEAR	MURDER	SCENE
ATTENTION	EVIDENCE	MYSTERY	SILENCE
CONCLUSION	EXAMINE	OBSERVE	SOLVE
CONFESS	GENTLEMAN	PASSAGE	SUSPICION
CRIMINAL	INNOCENT	PIPE	THEORY
CURIOUS	INQUIRE	PRESUME	WHISPER
DEDUCE	INSPECT	REMARK	WITNESS

Encontramos vocabulario propio de libros de detectives, con palabras relativas a los crímenes (*murder investigate, confess, witness*[33]*...*) y a la investigación policiaca (*evidence, solve, inquire, inspect*[34]*...*). Hay también palabras relativas a su localización, como *gentleman*[35] y *Scotland Yard*, aunque esta última es realmente un nombre propio formado por dos tokens que se han procesado por separado. Una de las curiosidades de las palabras temáticas en Sherlock Holmes es el término *pipe*, posiblemente el elemento más característico de la iconografía del famoso detective. Apreciamos un patrón que es fundamental para que el PLH alcance una cobertura tan alta con un número reducido de palabras clave: entre las palabras temáticas de *The Adventures of Sherlock Holmes* observamos que 37 de las 40 palabras (todas excepto *murder, whisper* y *witness*) tienen

[33] Asesinato, investigar, confesar, testigo.
[34] Prueba, resolver, preguntar (formal), inspeccionar.
[35] 'Caballero', expresión frecuente en inglés británico.

raíz greco-latina, y la gran mayoría son deducibles para un alumno hispanohablante incluso si nunca las ha encontrado anteriormente.

La Tabla II.25 presenta el análisis de cobertura del PLH en la misma versión de *The Adventures of Sherlock Holmes*. Observamos que las Keywords representan el 81,8 del texto, hay un 12.16% de cognados deducibles, y los nombres propios y otras palabras invariables representan el 2,44%. Como resultado, el porcentaje acumulado del PLH alcanza el 96,4%, superando así el umbral de autonomía al 95%.

Tabla II.25. Porcentaje acumulado del PLH en Sherlock Holmes

FILE	TOKEN	TOKEN%	CUMTOKEN%
KEYWORDS	86090	81.80	81.8
COGNATES	12797	12.16	93.96
INVARIABLES	2570	2.44	96.4
???	3792	3.60	100
TOTAL	105249		

4.4.2. Análisis léxico de A Christmas Carol

A Christmas Carol [36] es una novela corta escrita por Charles Dickens en 1843. Desde su primera publicación nunca ha estado descatalogada y ha sido adaptada en numerosas ocasiones al cine, teatro, televisión y otros formatos (Douglas-Fairhurst, 2006: viii). La edición original contiene 28.792 tokens distribuidos en los rangos de frecuencia del BNC como se indica en la Tabla II.26. Asumiendo que se conocen los nombres propios, el umbral de autonomía se alcanzaría con 5.000 palabras, lo que implica que la lectura de esta obra requiere una base léxica mayor que *The Adventures of Sherlock Holmes*. En el mismo sentido, para alcanzar el umbral de garantía son necesarias 9.000 palabras, frente a las 7.000 requeridas en el texto anterior.

[36] *Cuento de Navidad*.

Tabla II.26. Distribución léxica de A Christmas Carol según el BNC

FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	627	2.18	2.18
BNC-1	23667	82.2	84.38
BNC-2	1779	6.18	90.56
BNC-3	572	1.99	92.55
BNC-4	515	1.79	94.34
BNC-5	414	1.44	95.78
BNC-6	285	0.99	96.77
BNC-7	174	0.60	97.37
BNC-8	102	0.35	97.72
BNC-9	92	0.32	98.04
BNC-10	69	0.24	98.28
???	496	1.72	100
TOTAL	28792		

Como es previsible, solo el primer nivel supone tres de cada cuatro palabras del texto. Mediante el análisis de patrones inusuales en los niveles de baja frecuencia del BNC podemos descubrir las palabras temáticas que definen el tema central. Entre las palabras con más de 10 ocurrencias se encuentran los términos listados en la Tabla II.27.

Tabla II.27. Palabras temáticas en A Christmas Carol

SPIRIT	FROST	BELL	CHAIN	PUDDING
GHOST	HOLLY	PHANTOM	CHEER	CANDLE
MERRY	SHADOW	CLERK	HEARTY	DELIGHT
NEPHEW	BLESS	SPECTRE	PLEASANT	GOOSE

Podemos distinguir cuatro grupos temáticos: palabras relativas a los fantasmas (*spirit*, *ghost*, *phantom*, *spectre*, *chain*), a la Navidad (*merry*, *holly*, *bell*, *frost*), a los sentimientos asociados a ella (*pleasant, cheer, hearty*) y a la comida (*goose*, *delight*, *pudding*). Esto nos permite hacer un mapa del escenario navideño en el que se

desarrolla el libro. Por último, aparecen términos muy específicos con una frecuencia altísima respecto de su esperanza matemática: *nephew* y *clerk*, que definen la relación de los protagonistas y su profesión.

Al contrario de lo que ocurría con *The Adventures of Sherlock Holmes*, en estas palabras temáticas la presencia de cognados transparentes no es demasiado alta, únicamente se pueden clasificar como tal los términos *spirit, spectre, pudding* y *delight*[37]. Si bien *phantom* parece ser un cognado de «fantasma», no mantiene una semejanza suficiente con su equivalente como para poder garantizar que es deducible.

La Tabla II.28 presenta la cobertura del PLH en la versión sin resumir de *A Christmas Carol*. Observamos que las Keywords suponen el 81,8% de los tokens, las palabras invariables son el 2,49%. y los cognados transparentes suponen una cobertura adicional del 11,05%. En total, el PLH abarca el 94,64% del texto, es decir, queda muy cerca del umbral de autonomía pero sin llegar a alcanzarlo.

Tabla II.28. Porcentaje acumulado del PLH en A Christmas Carol

FILE	TOKEN	TOKEN%	CUMTOKEN%
KEYWORDS	23351	81.10	81.1
COGNATES	3182	11.05	92.15
INVARIABLES	716	2.49	94.64
???	1543	5.36	100
TOTAL	28792		

Una de las posibles razones que podría explicar por qué el PLH no llega al 95% de cobertura es que la presencia de cognados es más baja de lo normal. Si lo comparamos con *The Adventures of Sherlock Holmes*, tanto las Keywords como los nombres propios mantienen una proporción muy similar en ambos textos; sin embargo,

[37] El análisis del BNC muestra los tokens agrupados por lemas; si bien *delight* no es un cognado evidente, la forma que aparece repetida en el texto es *delicious*, que sí lo es.

en el caso de los cognados, *A Christmas Carol* se queda a más de un punto de diferencia por debajo. El motivo de ello, como mencionamos anteriormente, podría ser que las palabras temáticas más repetidas de esta obra son relativas a un campo semántico muy específico, la Navidad, donde aparentemente el vocabulario de procedencia latina no ha reemplazado a las palabras de raíz germánica.

Sin embargo, la explicación más plausible de por qué el vocabulario de *A Christmas Carol* es tan «difícil» es, sencillamente, que su autor es Charles Dickens. El lexicógrafo Ben Zimmer en su artículo *How Dickens Helped Shape the Lexicon* (2012) destaca que el prolífico autor no solamente aportó su talento e imaginación a la literatura inglesa, sino que la cantidad de vocabulario nuevo que incorporó en sus obras lo ha convertido en el sexto autor más citado en el *Oxford English Dictionary*, tan solo por detrás de Shakespeare, Scott, Chaucer, Milton y Dryden. De las 9.218 citas de Dickens, 528 se corresponden con el primer registro escrito de una palabra y otras 1.586 son acepciones nuevas de voces existentes. Citando un artículo publicado por Eric Partridge en 1933, Zimmer destaca el cuidadoso esfuerzo de Dickens por transmitir la personalidad y procedencia de sus personajes mediante una fiel reproducción de la jerga de la época, fundamentalmente la clase trabajadora y el mundo de la delincuencia.

Por cuanto antecede, aunque estrictamente el PLH no alcance el umbral de autonomía, podemos considerar un éxito destacable conseguir un 94,64% de cobertura en un texto que incluye numerosas palabras de baja frecuencia, términos inventados y jerga de la clase obrera del siglo XIX.

4.4.3. Análisis léxico de Lady Chatterley's Lover

La novela *El amante de Lady Chatterley*, cuyos pasajes describen de manera explícita las apasionadas relaciones sexuales de sus personajes protagonistas, levantó una reacción tan escandalosa que la obra estuvo prohibida durante 32 años tanto en el Reino Unido como en EEUU. De hecho, hoy en día aún se discute sobre si se trata de una novela de carácter erótico o si se debería catalogar como mera pornografía (Hernández, 1997). El análisis de esta obra aporta a nuestro estudio datos sobre la

eficacia del PLH en textos que contienen numerosas palabras tabú que, en principio, no se encuentran entre las de más alta frecuencia en el Brisith National Corpus.

En el estudio *How Large a Vocabulary Is Needed For Reading and Listening?* Nation analiza el perfil léxico de *Lady Chatterly's Lover* con las listas del BNC, llegando a la conclusión de que son necesarias 4.000 palabras para alcanzar el umbral del 95% (Nation, 2006a). Según el citado estudio, la distribución léxica de *Lady Chatterley's Lover* es la que se muestra en la Tabla II.29.

Tabla II.29. Distribución léxica de *Lady Chatterley's Lover* según el BNC

NIVEL	COBERTURA SIN NOMBRES PROPIOS (%)	CON NOMBRES PROPIOS (%)
BNC-1	80.88	82.93
BNC-2	88.09	90.14
BNC-3	91.23	93.28
BNC-4	93.01	95.06
BNC-5	94.08	96.13
BNC-6	94.77	96.88
BNC-7	95.38	97.43
BNC-8	95.85	97.9
BNC-9	96.17	98.22
BNC-10	96.41	98.46
???	3.59	1.54

(Adaptado de Nation, 2006)

Los datos indican que el 80,88% de los tokens pertenecen al primer nivel y que el umbral de autonomía se alcanza, efectivamente, en el cuarto nivel del BNC. Según las palabras de Nation,

> With a vocabulary of 4,000 word-families and assuming that proper nouns are easily understood, 95.06% of the tokens would be familiar. This means that there would be 1 unknown word in about every 20 running words. With a vocabulary of 9,000 words plus proper nouns, 98.22% of the tokens would be familiar. (Nation, 2006a: 70).

> Con un vocabulario de 4.000 familias léxicas y asumiendo que los nombres propios se entienden fácilmente, el 95,06% de los tokens serían conocidos. Esto implica que habría una palabra desconocida por cada 20 palabras aproximadamente. Con un vocabulario de 9.000 palabras además de los nombres propios, el 98,22% de las palabras serían conocidas. (trad. a.)

En cuanto a las palabras de muy baja frecuencia que aparecen más de lo esperado en el texto, Nation destaca las voces '*ter*', *'mun'*, *'wi'*, *'yo'* e '*impudence'*, y señala que son términos que guardan relación con el argumento y que los cuatro primeros representan voces del dialecto de los personajes. En nuestro análisis de términos con frecuencia inusual encontramos también las palabras temáticas de la Tabla II.30.

Tabla II.30. Palabras temáticas de baja frecuencia en Lady Chatterly's Lover

TENDER	FLUSH	ADORE	INTENSE
PASSION	PLEASURE	SATISFACTION	SENSITIVE
BREAST	SECRET	BUTTOCK	FUCK
NAKED	AFFAIR	CONTACT	ARSE
PHYSICAL	PRIVATE	ANXIOUS	ERECT
INTIMATE	BITCH	PROSTITUTE	POSSESS
CONNECT	EMOTION	CURVE	LESBIAN
SENSUAL	PENIS	HUMILIATE	INTERCOURSE
DESIRE	CLING	MISTRESS	ASS

Valorando la publicación en su contexto histórico, no es de sorprender que el lector de 1928 se escandalizara con una obra en la que se encuentra con esa lista de términos, y concretamente con más de 100 ocurrencias de las palabras *sex, fuck, penis, arse* y *bitch*.

Como hemos visto anteriormente, la presencia de numerosos cognados entre las palabras temáticas parece ser un indicador preliminar del alcance del PLH en un texto

dado, ya que estas tienen numerosas ocurrencias. Observamos que aproximadamente la mitad de las que aparecen en la Tabla II.30 se podrían clasificar como cognados transparentes. La Tabla II.31 muestra que la predicción de la palabras temáticas/cognados parece ser acertada. El 82,2% de palabras pertenecen a Keywords, junto con los cognados y las palabras invariables, el PLH supone una cobertura del 95,28%, alcanzando el umbral de autonomía buscado.

Tabla II.31. Cobertura del PLH en Lady Chatterly's Lover

FILE	TOKEN	TOKEN%	CUMTOKEN%
KEYWORDS	97299	82.2	82.2
COGNATES	12306	10.4	92.6
INVARIABLES	3176	2.68	95.28
???	5590	4.72	100
TOTAL	118371		

4.4.4. Análisis de las versiones adaptadas para estudiantes de L2

Esta sección presenta el análisis de las novelas *The Adventures of Sherlock Holmes* y *A Christmas Carol* en las versiones adaptadas por el proyecto sin ánimo de lucro Mid-frequency Readers. Esta es una colección óptima para el análisis comparativo porque lo único que varía respecto de la obra original es el vocabulario. A diferencia de las adaptaciones hechas por editoriales comerciales, los Mid-frequency Readers no resumen el libro ni tampoco alteran las estructuras sintácticas. Otra ventaja de esta serie es que los niveles superiores alcanzan a un nivel de exigencia léxica imposible de encontrar en los *graded readers* del mercado. Nation ofrece las cifras concretas:

> Published series of graded readers end at around the 3,000 to 4,000 word family level, but learners need a vocabulary size of 8,000 to 9,000 word families to read novels and newspapers. The Mid-frequency Readers are designed to provide interesting, comprehensible reading to fill this gap of 6,000-7,000 words between the end of graded readers and the demands of unsimplified text.
> *About Mid-Frequency Readers* (Nation, s.f.: 1)

> Las colecciones de lecturas graduadas comerciales tienen alrededor de las 3.000 o 4.000 familias léxicas en los niveles más altos pero los aprendices necesitan un vocabulario de entre 8.000 y 9.000 palabras para leer novelas y periódicos. Los Mid-frequency Readers [lecturas graduadas de frecuencia media] están diseñados para proporcionar lecturas interesantes y comprensibles que llenen este vacío de 6.000 - 7.000 palabras entre el final de las lecturas graduadas y la exigencia de los textos sin simplificar (trad. a.).

La Tabla II.32 resume el criterio de control del vocabulario en los tres niveles de Mid-frequency Readers. Si tomamos como ejemplo el primer nivel, las pautas marcadas son las siguientes: las 4.000 primeras palabras del BNC se consideran conocidas; las que pertenecen al rango 5.000 se mantienen en el texto como objetivo del aprendizaje, y las de rango superior a 6.000 se reemplazan por un sinónimo más asequible. Excepcionalmente, una palabra de rango superior a 6.000 puede mantenerse si tiene muchas ocurrencias o no es posible modificarla porque es de especial relevancia para el texto.

Tabla II.32. Control de vocabulario en los niveles de Mid-frequency Readers

BASE LÉXICA	OBJETIVO DE APRENDIZAJE	REEMPLAZAR
4.000	5.000	> 6.000
6.000	7.000 - 8.000	>9.000
8.000	9.000 - 10.000	>11.000

Adaptado de *About Mid-Frequency Readers* (Nation, s.f.: 1)

El resultado de estos reemplazos se puede observar en la Tabla II.33. Se muestra la comparación las primeras líneas de la edición original de *The Adventures of Sherlock Holmes* frente la versión adaptada para el nivel 4.000. Las diferencias están resaltadas en sombra gris.

Tabla II.33. Diferencias entre versiones de Sherlock Holmes.

VERSIÓN ORIGINAL	VERSIÓN ADAPTADA NIVEL 4.000
To Sherlock Holmes she is always the woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex. It was not that he felt any emotion akin to love for Irene Adler.	To Sherlock Holmes she is always the woman. I have seldom heard him mention her under any other name. In his eyes she eclipses the whole of her sex. It was not that he felt any emotion like love for Irene Adler.
All emotions, and that one particularly, were abhorrent to his cold, precise but admirably balanced mind. He was, I take it, the most perfect reasoning and observing machine that the world has seen, but as a lover he would have placed himself in a false position.	All emotions, and that one particularly, were hateful to his cold, precise but admirably balanced mind. He was, I take it, the most perfect reasoning and observing machine that the world has seen, but as a lover he would have placed himself in a false position.
He never spoke of the softer passions, save with a gibe and a sneer. They were admirable things for the observer excellent for drawing the veil from men's motives and actions. But for the trained reasoner to admit such intrusions into his own delicate and finely adjusted temperament was to introduce a distracting factor which might throw a doubt upon all his mental results.	He never spoke of the softer passions, save with contempt. They were admirable things for the observer -- excellent for drawing the veil from men's motives and actions. But for the trained thinker to admit such things into his own delicate and finely adjusted temperament was to introduce a distracting factor which might throw a doubt upon all his mental results.

Podemos ver cómo habitualmente cambian las palabras difíciles por otras de mayor frecuencia en el BNC (*akin to* → *like; abhorrent* → *hateful*). En ocasiones, si el término que presenta dificultad no aporta mucho al texto o es redundante, directamente se elimina en la versión adaptada (*eclipses* <u>*and predominates*</u> → *eclipses*). También hay la posibilidad de que varios términos difíciles se transformen en una única palabra que abarque el significado de ambos (*a gibe and a sneer* → *contempt*).

En esta breve comparativa también se aprecian algunos de los problemas derivados de minusvalorar la importancia de la L1 del aprendiz a la hora de diseñar

material docente. Si bien la frecuencia de una palabra es un buen indicador de su dificultad previsible, el criterio de similitud debe prevalecer sobre la frecuencia. Tomemos el ejemplo de la última palabra modificada: *Intrusions* es una voz absolutamente transparente para un hispanohablante, pero en la versión adaptada se ha cambiado por una de mayor frecuencia, dando como resultado *things*. En este caso concreto, *things* es una palabra muy sencilla que conoce hasta el alumno del nivel más elemental, pero observemos otras palabras que en el texto que han sido alteradas siguiendo estrictamente el criterio de frecuencia:

AUTHORITATIVE → LOUD
SINISTER → EVIL
OPULENCE → WEALTH
INCISIVE → INSIGHTFUL

Estas modificaciones tienen para el alumno hispanohablante justo el efecto contrario del que se pretende conseguir: algunas alteraciones pueden hacer que ciertas frases del texto adaptado sean más difíciles que en la versión original.

Analicemos ahora el alcance del PLH para los tres niveles adaptados en comparación con las obras originales. La Tabla II.34 muestra que en *The Adventures of Sherlock Holmes* hay una descendencia progresiva en la cobertura a medida que se incluyen palabras de baja frecuencia.

Tabla II.34. PLH en versiones de Sherlock Holmes y Christmas Carol

VERSIÓN	SHERLOCK HOLMES (%)	CHRISTMAS CAROL (%)
MID FREQUENCY 4000	97.72	96.77
MID FREQUENCY 6000	97.21	95.89
MID FREQUENCY 8000	96.91	95.86
TEXTO ORIGINAL	96.40	94.64

Uno de los datos significativos que avalan la validez de los Mid-frequency Readers es que la distancia para pasar del primer al segundo nivel es la misma que la existente entre el último nivel y el texto original. Recordemos que las series comerciales de lecturas graduadas suelen terminar en torno a las 4.000 palabras, lo que deja un vacío

considerable entre ellos y los textos originales que los Mid-frequency Readers parecen ocupar de manera satisfactoria.

La Figura II.12 muestra el aumento en el porcentaje palabras que quedan fuera del PLH a medida que aumenta la dificultad del texto. Mientras Sherlock Holmes muestra una progresión regular, en *A Christmas Carol* tiene dos subidas abruptas interrumpidas por un valor casi constante entre las lecturas adaptadas a los niveles 6.000 y 8.000. El motivo de ello es que en el nivel 8.000 hay una mayor presencia de cognados que compensa el descenso de Keywords, por lo que el PLH tiene prácticamente el mismo porcentaje acumulado en ambos niveles. Posiblemente esto se deba a que en las modificaciones léxicas para adaptar el texto al nivel 6.000 se alteraron cognados transparentes que volvieron a aparecer en el nivel 8.000. En el texto original, sin embargo, el número de cognados no consigue neutralizar al de las palabras difíciles de baja frecuencia, cuya proporción sobrepasa en este texto el máximo fijado para umbral de autonomía.

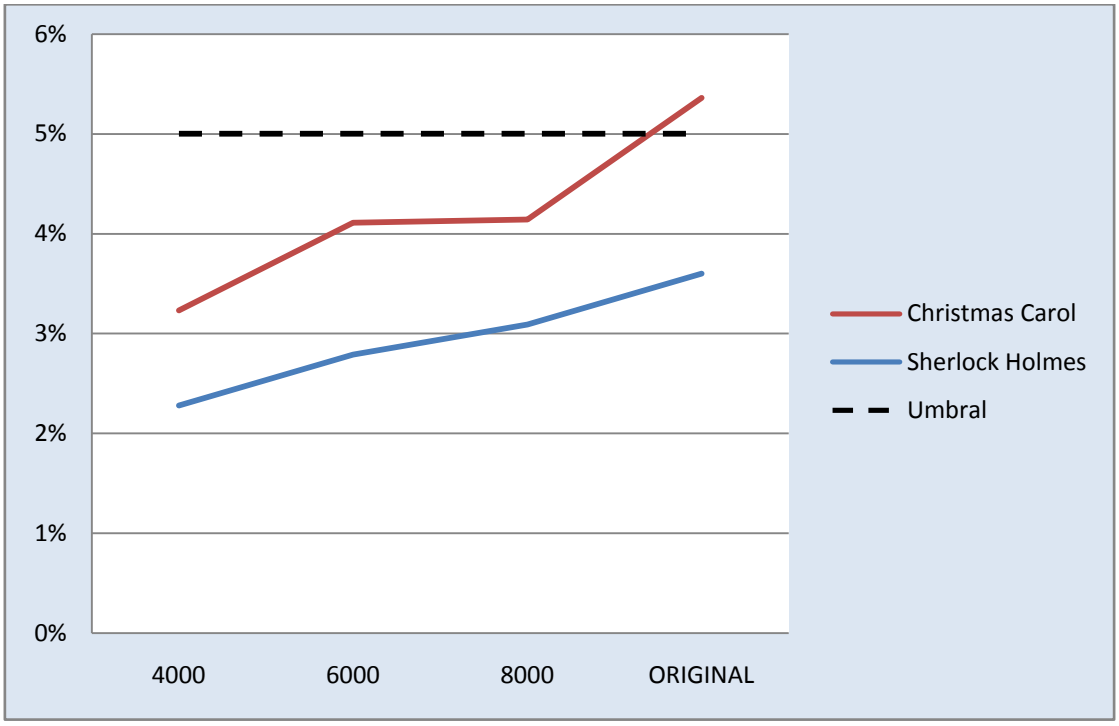


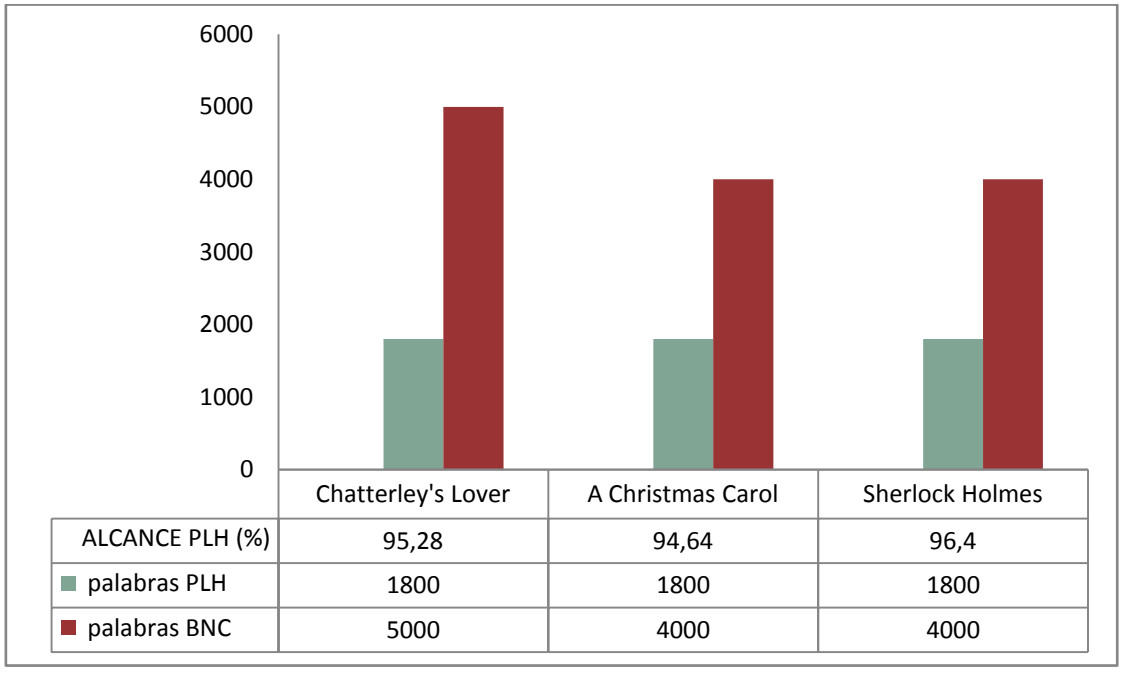
Figura II.12. Porcentaje de palabras que no cubre el PLH en las distintas versiones de Sherlock Holmes y A Christmas Carol

4.4.5. Conclusiones del análisis de obras literarias

El PLH alcanza el umbral de autonomía en dos de las tres obras originales, lo que es destacable ya que *Lady Chatterley's Lover* contiene múltiples palabras tabú y otros términos de baja frecuencia. En la obra de Dickens el PLH se queda muy cerca del umbral (- 0,46%) a pesar de que el texto presentaba un vocabulario exigente.

En las versiones de los mismos textos adaptadas para una base léxica de 4.000, 6.000 y 8.000 palabras también se consigue superar el umbral de autonomía. El nivel de esta serie es muy superior al de los *graded readers* del mercado, por lo que podemos deducir que probablemente en ellos se rebase el umbral ampliamente.Los datos indican que para alumnos hispanohablantes el PLH es más rentable como objetivo de aprendizaje que las listas de frecuencia genéricas del BNC.

La Figura II.13 resume los datos anteriores, señalando el alcance del PLH y el nivel del BNC que sería necesario para garantizar al menos la misma cobertura. Observamos para alcanzar el mismo resultado PLH, un programa basado en la frecuencia del BNC requeriría más del doble de objetivos de aprendizaje explícito.



	Chatterley's Lover	A Christmas Carol	Sherlock Holmes
ALCANCE PLH (%)	95,28	94,64	96,4
palabras PLH	1800	1800	1800
palabras BNC	5000	4000	4000

Figura II.13. Palabras del BNC necesarias para igualar al PLH en obras literarias

4.5. Análisis de habla no espontánea

Las Keywords que contiene el PLH tomaron como primera referencia las listas de Nation, en las que el BNC tiene una especial representatividad. Uno de los problemas más comunes de trabajar con este corpus es que muestra cierto sesgo por el que prevalece el lenguaje escrito, formal y británico. Se podría argumentar, por tanto, que el PLH ofrece buenos resultados al analizar textos de exámenes y obras literarias de autores británicos, pero su eficacia podría variar si nos encontramos ante otras situaciones comunicativas, tales como el lenguaje oral, un registro informal u otras variedades dialectales.

En esta sección, por tanto, estudiamos la eficacia del PLH en a) registros coloquiales, b) muestras de habla no espontánea, y c) inglés americano. El registro informal está representado en un corpus formado por subtítulos de películas. Para examinar el lenguaje oral se estudia un corpus de series de televisión dirigidas a un público joven. A continuación se analiza la transcripción de tres discursos del presidente de EEUU Barack Obama (2008; 2009; 2013), que representan un registro formal de inglés americano.

4.5.1. Registro informal

La comprensión oral es un área muy compleja en la que entran en juego muchos más factores que en el lenguaje escrito. Como apunta Nation, una película tiene la ventaja de proporcionar apoyo visual; sin embargo, también tiene la desventaja que caracteriza el lenguaje oral, y es que una vez oído desaparece (Nation, 2006a: 76). En ningún caso tenemos datos suficientes para hacer una correlación entre el vocabulario conocido y la probabilidad de que un alumno pueda entender una película de oído. Esta sección no pretende extrapolar la cobertura en lectura a la comprensión oral, la razón por la que se estudia este corpus es que los guiones de las películas nos ofrecen datos sobre un registro más coloquial.

El corpus está formado por los archivos de subtítulos de las películas *Harry Potter and the Philosopher's Stone*[38], *Shrek, The Goonies* y los 24 episodios que componen la primera temporada de la serie de la CBS *How I met your mother*[39]. En conjunto, suman un total de 91.689 tokens.

Harry Potter and the Philosopher's Stone es una película que tiene una gran riqueza léxica a pesar de que un nutrido grupo de su público objetivo son niños. Tiene especial interés para este estudio porque gira en torno a la magia, un tema muy específico y poco común en el lenguaje cotidiano, lo que implica que previsiblemente hay una gran presencia de palabras temáticas de baja frecuencia. *Shrek* es una película destinada a un público infantil que nos aporta datos sobre la eficacia del PLH en un registro muy familiar. La película de aventuras *The Goonies* nos proporciona datos sobre el habla adolescente. Por último, *How I met your mother* es una de las series de televisión más exitosas en la actualidad. Representa las conversaciones informales, incluyendo palabras vulgares y jerga, de un grupo de treintañeros que vive en Nueva York.

a) Análisis de Harry Potter and the Philosopher's Stone.

El universo de Harry Potter, igual que otras series de fantasía, tiene una gran cantidad de vocabulario propio. Dementores, mortífagos, *muggles*[40] y partidos de un peligrosísimo *quidditch*[41] son algunas de las muchas voces inventadas por JK Rowling que han pasado a ser parte del imaginario colectivo.

Una de las peculiaridades de las palabras inventadas para el universo Harry Potter es que la autora sigue un patrón lingüístico muy cuidado basado en raíces de apariencia germánica y latina [42]. Los términos que denominan a personas o cosas podrían ser voces reales inglesas, tales como *muggle* o *bludger*[43]; mientras que la mayor parte de las

[38] *Harry Potter y la piedra filosofal*.
[39] En España, *Cómo conocí a vuestra madre*.
[40] Persona incapaz de hacer magia
[41] Deporte que recuerda ligeramente al baloncesto al que se juega volando sobre escobas mágicas
[42] J.K. Rowling, autora de la serie de libros de Harry Potter, estudió Filología Clásica.
[43] Bola de hierro que se utiliza en el *quidditch*

palabras mágicas que activan los encantamientos tienen apariencia latina y están formadas a partir de raíces que a un angloparlante le recordarán a palabras conocidas.

Tabla II.35. Palabras inventadas para los encantamientos de Harry Potter

PALABRAS MÁGICAS	SIRVE PARA	RAIZ LATINA	PALABRA INGLESA
OCULUS REPARO	REPARAR (GAFAS)	OCULUS, OJO	OCULAR
		SUFIJO *RE + PARARE*, REPARAR	REPARE
LEVIOSA	LEVITAR	*LEVARE*, ELEVAR.	LEVITATE
PETRIFICUS TOTALUS	PETRIFICAR	*PETRA*, PIEDRA + *FACERE*, HACER	PETRIFIED
		TOTUS, TODO + SUFIJO *AL*	TOTAL
LUMUS	EMITIR LUZ	LUX, LUMEN-, LUZ	LUMINOSITY

Desde el punto de vista del análisis léxico, es importante tener en cuenta que una de las peculiaridades de las palabras inventadas es que no afectan a la comprensión porque en la propia obra el lector encontrará la definición de cada una de ellas. En el análisis de Harry Potter las voces inventadas que no fueron traducidas en las ediciones en español se asignaron a categoría de las palabras invariables[44].

La Tabla II.36 muestra la distribución léxica según su frecuencia relativa en el BNC de la primera de las películas de la saga, *Harry Potter and the Philosopher's Stone*. Observamos que las palabras invariables suponen el 6,31% de los tokens, una presencia notablemente más alta que la media de 2,5% presente en las obras literarias analizadas en la sección anterior. Es cierto que el vocabulario inventado de Harry Potter, incluido en la matriz de palabras invariables, contribuye ligeramente al aumento de su representatividad; sin embargo, realmente son los nombres propios los

[44] También se incluyó en esta categoría el término «snitch», que es una palabra real en inglés pero no está traducida porque en Harry Potter únicamente designa una pelota dorada con alas que deben atrapar los jugadores de *qudditch*.

responsables del elevado porcentaje de palabras invariables. En la película hay más de 40 personajes con nombre, además de ciudades, establecimientos y otros lugares con nombre propio. De hecho, solamente el token *Harry* aparece 71 veces. Para ponerlo en perspectiva, tiene casi tantas ocurrencias como *have* (77), la octava palabra más frecuente en inglés. Es más, el conjunto de los tokens *Harry, Hagrid, Snape, Hogwarts, Dumbledore* y *Ron* suponen una de cada 50 palabras de toda la película.

Los datos de la Tabla II.36 muestran que la película, destinada en parte al público infantil, requiere una base léxica menor que los clásicos de la literatura estudiados en la sección anterior. El umbral de autonomía se alcanza con únicamente 2.000 palabras del BNC, asumiendo que se deducen los nombres propios.

Tabla II.36. Distribución léxica de Harry Potter según su frecuencia relativa en el BNC

FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	530	6.31	6.31
BNC-1	7103	84.6	90.91
BNC-2	349	4.16	95.07
BNC-3	99	1.18	96.25
BNC-4	53	0.63	96.88
BNC-5	64	0.76	97.64
BNC-6	29	0.35	97.99
BNC-7	26	0.31	98.3
BNC-8	32	0.38	98.68
BNC-9	15	0.18	98.86
BNC-10	17	0.2	99.06
???	79	0.94	100
TOTAL	8396		

Entre los tokens de baja frecuencia encontramos palabras temáticas de la Tabla II.37. Están relacionadas tanto con el mundo mágico como con el académico, que nos permiten intuir el tema y el escenario, respectivamente.

Tabla II.37. Palabras temáticas en Harry Potter

PROFESSOR	CLOAK	ELIXIR	POTION	WEREWOLVES
DRAGON	COUNTERCURSE	JINXING	DORMITORIES	WITCH
BROOM	CURSE	TROLL	ROBE	WITCHCRAFT
BROOMSTICK	DUNGEON	OWL	TOAD	WIZARD
INCANTATION	GOBLINS	POTION	WAND	UNICORN

Algunas de estas palabras temáticas tales como *enchantment, incantation, troll, potion*, *elixir*, *dragon* y *unicorn* son cognados transparentes. Por otra parte, una curiosidad de estas palabras temáticas es *spell*, situada en el segundo nivel del BNC. Esta palabra puede inducir a error porque tiene dos acepciones, y casualmente las más habitual, 'deletrear', es un término de uso muy frecuente en las clases de lengua extranjera, por lo que el alumno pensará que la conoce. Sin embargo, en la Tabla II.38 se puede comprobar que de las nueve ocurrencias de *spell* en la película, solamente tiene el sentido 'deletrear' en la primera, mientras que en el resto toma su acepción menos frecuente, 'hechizo'.

Tabla II.38. Ocurrencias de «spell» en Harry Potter

Maybe if you wrote it down? —No, I can't spell it.
Are they real frogs? —It's a **spell**.
Fred gave me a **spell** to turn him yellow.
Standard Book of **Spells**, chapter seven.
I know a **spell** when I see one.
Cast a Christmas **spell.**
Spells, enchantments... Right.
Snape's already been here. He put a **spell** on the harp.

Encontrar una palabra polisémica suele ser una traba para la comprensión cuando el alumno conoce únicamente una de las acepciones, especialmente si la palabra resulta ser una *topic word* que presenta varias ocurrencias en el texto.Los datos de la Tabla II.39 presentan el porcentaje acumulado de los componentes del PLH, y observamos que cuatro de cada cinco palabras de la película son Keywords. De las restantes, la mitad son cognados evidentes. Suponiendo que se deducen los nombres propios, el PLH alcanza una cobertura del 97,48%, rebasando ampliamente el umbral de autonomía. Para lograr la misma cobertura con el BNC serían necesarias 5.000 palabras.

Tabla II.39. Porcentaje acumulado del PLH en Harry Potter

FILE	TOKEN	TOKEN%	CUMTOKEN%
KEYWORDS	6802	81.01	81.01
COGNATES	849	10.11	91.12
INVARIABLES	534	6.36	97.48
???	211	2.51	99.99
TOTAL	8396		

b) Análisis de Shrek

Shrek es una conocida película infantil estrenada en 2001 que ganó el Oscar a la mejor película de animación. Tiene 7.265 tokens distribuidos en progresión descendente, excepto el nivel 5, donde hay una representatividad más alta de lo esperado (véase la Tabla II.40). El alto porcentaje del nivel 5 se debe a que muchas palabras temáticas del género de aventuras se encuentran en este nivel, tales como *beast, helmet, noble* y *arrow*. En otros niveles de baja frecuencia encontramos *ogre, knight, quest, fairytale, rescue* o *monster*, vocabulario propio de los cuentos de hadas.

En Nation (2006a) se puede ver un análisis de Shrek con listas de frecuencias del BNC que ofrece unas cifras similares, aunque no coinciden completamente. Hay tres factores que explican las diferencias: (a) Nation utiliza el guión original de la película, mientras que en este estudio procesamos el archivo de subtítulos, (b) las listas del BNC

de este estudio son una versión más moderna, ya que Nation actualizó sus listas posteriormente a su artículo de 2006, y (c) nuestra lista de palabras invariables incluye expresiones como *um, oh, ah*, que en las listas de 2006 están en el nivel 3.

Tabla II.40. Distribución léxica de Shrek según el BNC

FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	171	2.35	2.35
BNC-1	6359	87.53	89.88
BNC-2	257	3.54	93.42
BNC-3	78	1.07	94.49
BNC-4	56	0.77	95.26
BNC-5	73	1	96.26
BNC-6	43	0.59	96.85
BNC-7	8	0.11	96.96
BNC-8	4	0.06	97.02
BNC-9	16	0.22	97.24
BNC-10	5	0.07	97.31
???	195	2.68	99.99
TOTAL	7265		

En cuanto a la eficacia del PLH en una película infantil, la Tabla II.41 muestra que el 83.33% de los tokens de *Shrek* son Keywords. Si asumimos que el alumno deduce los cognados evidentes y los nombres propios, el PLH supera cómodamente el umbral de autonomía.

Tabla II.41. Porcentaje acumulado del PLH en Shrek

FILE	TOKEN	TOKEN%	CUMTOKEN%
KEYWORDS	6054	83.33	83.33
COGNATES	819	11.27	94.6
INVARIABLES	173	2.38	96.98
???	219	3.01	99.99
TOTAL	7265		

c) Análisis de The Goonies

The Goonies es una película de aventuras basada en una historia de Steven Spielberg que se considera uno de los films de culto de los años 80. Narra las aventuras de un grupo de adolescentes en la búsqueda de un tesoro pirata perdido. La palabra que da título a la película es la que utilizan los protagonistas para referirse a su pandilla, que deriva del nombre del barrio en el que viven, *the Goon Docks.* Aunque ambos términos forman parte de un nombre propio, en el doblaje al castellano se refieren a ese lugar como «los muelles de Goon». Dado que el criterio para la lista de invariables es que la palabra no haya sido traducida en la versión en español, *goon* y *goonie* se han considerado nombres propios mientras que *dock* permanece en su lista original [45].

La película tiene 9.379 tokens distribuidos a lo largo de las familias del BNC como señala la Tabla II.42. El 85,12% de las palabras pertenecen al primer nivel y hay una gran presencia de palabras invariables.

Tabla II.42. Distribución léxica de The Goonies según el BNC

FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	492	5.25	5.25
BNC-1	7983	85.12	90.37
BNC-2	309	3.29	93.66
BNC-3	108	1.15	94.81
BNC-4	118	1.26	96.07
BNC-5	66	0.70	96.77
BNC-6	35	0.37	97.14
BNC-7	10	0.11	97.25
BNC-8	14	0.15	97.4
BNC-9	18	0.19	97.59
BNC-10	7	0.07	97.66
???	219	2.34	100
TOTAL	9379		

[45] Intencionadamente o no, *goonies* significa, en lenguaje coloquial, 'tontos'. El diccionario Merriam-Webster's también contempla la acepción 'albatros'.

Los 15 personajes que tienen nombre en *The Goonies* representan más del 90% de los tokens invariables, aunque la proporción global de esta lista en la película es muy inferior a la señalada en el film de Harry Potter, que tiene más de 40 personajes con nombre propio. Esto posiblemente se deba a que *The Goonies* es un guión original y *Harry Potter and the Philosopher's Stone* es la adaptación de una novela. Además, el nombre del protagonista, Mikey, tiene 42 ocurrencias frente a las 71 veces que mencionan a Harry.

The Goonies, al contrario que en los análisis previos, no tiene una correspondencia evidente entre términos de frecuencia inusitada y palabras temáticas. La Tabla II.43 muestra las palabras de los niveles altos que más se repiten en la película. Si bien todas son relevantes cuando se conoce previamente la historia, no encontramos grupos semánticos que identifiquen claramente la trama, el escenario o las características de los protagonistas.

Tabla II.43. Palabras de baja frecuencia recorrentes en The Goonies

ATTIC	CAVE	IDIOT	RABIES
BATHROOM	CHAIN	JERK	RESTAURANT
BIKE	CHASE	JEWELS	ROPE
BOOBY	CHOCOLATE	KIDDING	SCREW
BOOTY	CUSTOMER	MAMA	SHERIFF
BRAIN	DAMN	MAP	SHIT
BUCKET	DOCKS	PEPPERONI	STAIRS
BULLSHIT	ECHO	PIPE	TRANSLATE
BURY	FIREPLACE	PIRATE	TRAP
CALM	FUTURE	PLANK	TREASURE

Vemos pequeños grupos temáticos tales como partes de una casa (*attic, bathroom, fireplace*) o comida (*restaurant, chocolate, pepperoni*). Son muy importantes en la trama pero sería difícil intuir la historia o las características de los personajes únicamente a partir de esta lista de términos. De las 35 palabras, las únicas que son relativas al género de la película son *pirate, map, plank* y *jewels*. Aparte de eso, sí

podemos observar una serie de términos tales como *shit*, *damn*, *jerk* y *screw* —en la película es siempre parte de la expresión *screw up*— lo que nos da una pista del registro del habla coloquial de los personajes.

Una de las curiosidades de las palabras recurrentes es el par *booty / booby*. La expresión *booby traps* es un tipo de trampa explosiva que se detona al tocarla, concepto al que hacen referencia en repetidas ocasiones en la película. El personaje experto en explosivos es un adolescente chino que tiene dificultades con el idioma, por lo que las llama erróneamente *booty traps* a pesar de que sus amigos le corrigen cada vez que lo dice. Distintas variaciones del siguiente diálogo son un chiste recurrente en la película:

> Where you going?
>
> To set booty traps.
>
> You mean booby traps.
>
> That's what I said. Booby traps!
>
> (The Goonies, min. 00:53:21)

La Tabla II.44 muestra los porcentajes del PLH en *The Goonies*. Los datos revelan cuatro de cada cinco tokens son Keywords, el 11,91% cognados transparentes y el 5,11%, nombre propios y otras expresiones invariables. El resultado es que el porcentaje acumulado del PLH supera sin dificultades el umbral de autonomía (95%), quedándose a tan solo -0,37 puntos del umbral de garantía (98%).

Tabla II.44. Porcentaje acumulado del PLH en The Goonies

FILE	TOKEN	TOKEN%	CUMTOKEN%
KEYWORDS	7560	80.61	80.61
COGNATES	1117	11.91	92.52
INVARIABLES	479	5.11	97.63
???	223	2.38	100.01
TOTAL	9379		

d) Análisis de How I Met your mother

How I Met your Mother es una telecomedia americana estrenada en 2005 que sigue el día a día de un grupo de amigos que vive en Nueva York. El título hace referencia a que la narración utiliza un marco de referencia como recurso estilístico para engarzar las historias de cada episodio, otros ejemplos muy famosos que utilizan esta estructura son *Los cuentos de Canterbury* y *Las mil y una noches*. Cada capítulo comienza en el año 2030 con la voz en off del personaje principal que les relata a sus dos hijos la serie de eventos que condujeron a que conociera a su madre. Cada historia se desarrolla en el momento actual.

Para el análisis se escogieron los 24 episodios que componen la primera temporada de la serie. El corpus se elaboró a partir de los subtítulos extraídos de los DVD y tiene un total de 66.649 tokens distribuidos a lo largo de los niveles del BNC como muestra la Tabla II.45. El primer nivel contiene el 84,98% de los tokens, en línea con las cifras de las películas anteriores. Las palabras no traducibles representan el 5,42%. Solamente el nombre de los cinco protagonistas —*Ted, Robin, Marshall, Lily, Barney*— suponen una de cada 60 palabras de la serie.

Tabla II.45. Distribución léxica de How I Met your Mother

FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	3612	5.42	5.42
BNC-1	56640	84.98	90.4
BNC-2	2031	3.05	93.45
BNC-3	612	0.92	94.37
BNC-4	548	0.82	95.19
BNC-5	313	0.47	95.66
BNC-6	324	0.49	96.15
BNC-7	176	0.26	96.41
BNC-8	266	0.40	96.81
BNC-9	126	0.19	97.00
BNC-10	60	0.09	97.09
???	1941	2.91	100
TOTAL	66649		

Si observamos los patrones poco usuales en los niveles de baja frecuencia podemos encontrar las palabras temáticas listadas en la Tabla II.46. Identificamos palabras relativas a su localización urbana (*cab, metro*[46], *Liberty*[47]), referencias culturales estadounidenses (*Thanksgiving, prom, Halloween* acompañado de *costume y pumpkin*), y expresiones relativas al tema central de la serie (*romantic, girlfriend, boyfriend, fiancée, bride, jealous*). También aparece la profesión del protagonista, *architect.*

Lo más destacable de estas palabras temáticas es que a primera vista identifican claramente el registro de la serie. Por un lado encontramos palabras informales pero socialmente aceptables, como *awesome, dude, psyched, kidding* o *lame.* Además, hay multitud de términos vulgares y palabras tabú, tal es el caso de *damn, crap, slutty, hooker, pee, prostitute, ass, bitch* o *cock.*

Tabla II.46. Palabras de baja frecuencia recurrentes en How I Met your Mother

AWESOME	PUMPKIN	CUTE	ROMANTIC
DUDE	HALLOWEEN	METRO	ARCHITECT
DAMN	PROM	FIANCE	BITCH
GIRLFRIEND	COSTUME	BUDDY	BRIDE
BIRTHDAY	LEGENDARY	HOOKER	COCK
KIDDING	SLUTTY	AIRPORT	FAKE
CRAP	FIANCEE	PEE	KARAOKE
WEEKEND	LAME	PROSTITUTE	LESBIAN
CAB	PSYCHED	ANCHOR	LIBERTY
THANKSGIVING	BOYFRIEND	ASS	JEALOUS

[46] *Metro* no se refiere al medio de transporte -*subway*, en inglés americano-. En la serie *metro* se usa como una abreviatura de *Metropolitan*.

[47] *Statue of Liberty*, pero las palabras *statue* y *of* pertenecen a los niveles de alta frecuencia.

Para comprobar si el registro informal afecta negativamente a la cobertura del PLH, observamos los datos de su porcentaje acumulado en la Tabla II.47. La presencia de Keywords, un 78.21% es ligeramente menor que en las películas anteriores, pero se compensa con una mayor representatividad de los cognados.

Tabla II.47. Porcentaje acumulado del PLH en How I Met your Mother

FILE	TOKEN	TOKEN%	CUMTOKEN%
KEYWORDS	52129	78.21	78.21
COGNATES	8867	13.3	91.51
INVARIABLES	3612	5.42	96.93
???	2041	3.06	99.99
TOTAL	66649		

En principio, este dato iría contra la idea intuitiva de que es el registro culto el que contiene mayor número de palabras de raíz latina; sin embargo, los datos indican que en el lenguaje coloquial se utilizan frecuentemente términos como *totally, probably, moment, minute, problem* o *important*. Además, hay que tener en cuenta que no todos los cognados evidentes son consecuencia de la inclusión de palabras latinas en el inglés, sino que también el español ha incorporado a través de préstamos algunas palabras de uso habitual en inglés. El ejemplo más claro es *okay*, que en *How I met your Mother* aparece una media de 12 veces por episodio.

Los datos de la Tabla II.47 indican que el PLH cubre el 96,93% de todas las palabras que aparecen en la primera temporada de la serie *How I Met your Mother*, lo que implica que también ha superado el umbral de autonomía (95%) en un registro informal.

4.5.2. Registro formal

En esta sección se analiza la eficacia del PLH en lengua oral no espontánea en el registro formal del inglés estadounidense. Como referencia se utiliza un corpus formado por discursos inaugurales del presidente de EEUU Barack Obama y el primer discurso en que el utilizó públicamente el más famoso eslogan de su campaña, *Yes, we can*. Contiene un total de 5.578 tokens distribuidos a lo largo del BNC como se ve en la Tabla II.48.

Tabla II.48. Distribución léxica del corpus de discursos de Obama

FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	91	1.63	1.63
BNC-1	4433	79.47	81.1
BNC-2	501	8.98	90.08
BNC-3	312	5.59	95.67
BNC-4	86	1.54	97.21
BNC-5	50	0.9	98.11
BNC-6	33	0.59	98.7
BNC-7	26	0.47	99.17
BNC-8	7	0.13	99.3
BNC-9	3	0.05	99.35
BNC-10	3	0.05	99.4
???	33	0.59	99.99
TOTAL	5578		

La primera diferencia evidente respecto del registro informal es que Obama parece utilizar una menor proporción de palabras de alta frecuencia. El primer nivel aporta únicamente el 79,47% de los tokens, aproximadamente 5 puntos por debajo que la media de las películas y un porcentaje inferior incluso que el de las obras literarias. Esto podría ser un indicador de que el discurso de Obama tiene una gran complejidad léxica; sin embargo observamos que el segundo nivel ocurre el efecto contrario, con una proporción casi 5 puntos por encima de la media de las películas. Con estos datos, el discurso no debería ser especialmente complicado excepto para aprendices de los niveles más elementales. De hecho, el umbral de autonomía se supera en las 3.000 palabras, igual que en los exámenes B1.

En cuanto a las palabras invariables, el porcentaje de nombres propios que aparecen en un discurso político es obviamente muy inferior al de los diálogos de las películas. Casi todos ellos son topónimos de EEUU (*Hampshire, Carolina, Detroit, Iowa*), apenas hay nombres de personas (*Clinton, Biden*), y estos solamente aparecen una vez. La Tabla II.49 muestra las palabras de baja frecuencia que más se repiten en el corpus de discursos de Obama. Se puede reconocer el género a primera vista por la inmensa mayoría de palabras temáticas relacionadas con la política. Destaca también la altísima presencia de cognados transparentes entre las palabras del registro formal.

Tabla II.49. Palabras de baja frecuencia recurrentes en discursos de Obama

ALLIANCES	CREED	EVIDENT	MUTUAL	REPUBLICANS
AMBITIONS	CRISIS	FACTION	NUCLEAR	RESOLVE
CAMPAIGN	CYNICS	FAILURE	OATH	RESTORE
CAPACITY	DEBATES	FALSE	PATRIOTS	SEIZE
CELEBRATION	DECLARED	FASCISM	PIONEERS	SOURCE
CHARITY	DEFINE	FOREVER	PLEDGE	SUCCEED
CHARTER	DEMOCRACY	FOUNDING	POVERTY	TASK
CLIMATE	DEMOCRATS	HARNESS	PRECIOUS	THRIVES
COMMUNISM	DESTINY	IDEALS	PRINCIPLES	TOLERANCE
COMPEL	DIGNITY	IMMIGRANTS	PROCLAIM	TRANSITION
CONFIDENCE	DISTANT	INITIATIVE	PROSPERITY	TYRANNY
CONFLICT	DOCUMENTS	LASH	PURSUIT	ULTIMATELY
COOPERATION	ENDURING	LIBERTY	REFORM	VIRTUE
COURAGE	ERA	MAJORITY	REJECT	WEALTH

Para analizar cómo afectan estos factores al alcance del PLH podemos observar los datos de la Tabla II.50. La presencia de Keywords es relativamente baja, sin embargo, los cognados propios del habla culta representan el 17,7% de los tokens. Los resultados demuestran que a pesar del sesgo hacia el inglés británico escrito, el PLH

también supera el umbral de autonomía (95%) en un registro de habla culta estadounidense.

Tabla II.50. Porcentaje acumulado del PLH en los discursos de Obama

FILE	TOKEN	TOKEN%	CUMTOKEN%
KEYWORDS	4310	77.27	77.27
COGNATES	991	17.77	95.04
INVARIABLES	91	1.63	96.67
???	186	3.33	100
TOTAL	5578		

4.5.3. Conclusiones del análisis de habla no espontánea

La Tabla II.51 resume los datos relativos al alcance del PLH en el corpus de habla no espontánea. No resulta sorprendente que consiga la mayor cobertura en una película infantil, Shrek, y la menor en el discurso político de Obama. Sin embargo, y a pesar de ser dos géneros completamente distintos, la diferencia entre ambos es únicamente del 0,7%. Los datos demuestran que una baja presencia de Keywords no implica necesariamente un aumento de palabras desconocidas. En realidad, parecen equilibrarse con un aumento proporcional de cognados. En la Figura II.14 observamos que esto ocurre tanto en el registro formal como en el informal y posiblemente se debe, en el primer caso, a la influencia de los términos latinos en el habla culta y, en el segundo, a ciertas palabras cotidianas de alta frecuencia como *totally* o *really*.

Tabla II.51. Alcance del PLH en las muestras de habla no espontánea

	SHREK	HARRY	GOONIES	HOW I MET	OBAMA
KEYWORDS	83.33	81.01	80.61	78.21	77.27
COGNATES	11.27	10.11	11.91	13.3	17.77
INVARIABLES	2.38	6.36	5.11	5.42	1.63
???	3.01	2.51	2.38	3.06	3.33
TOTAL PLH	96.98	97.48	97.63	96.93	96.67

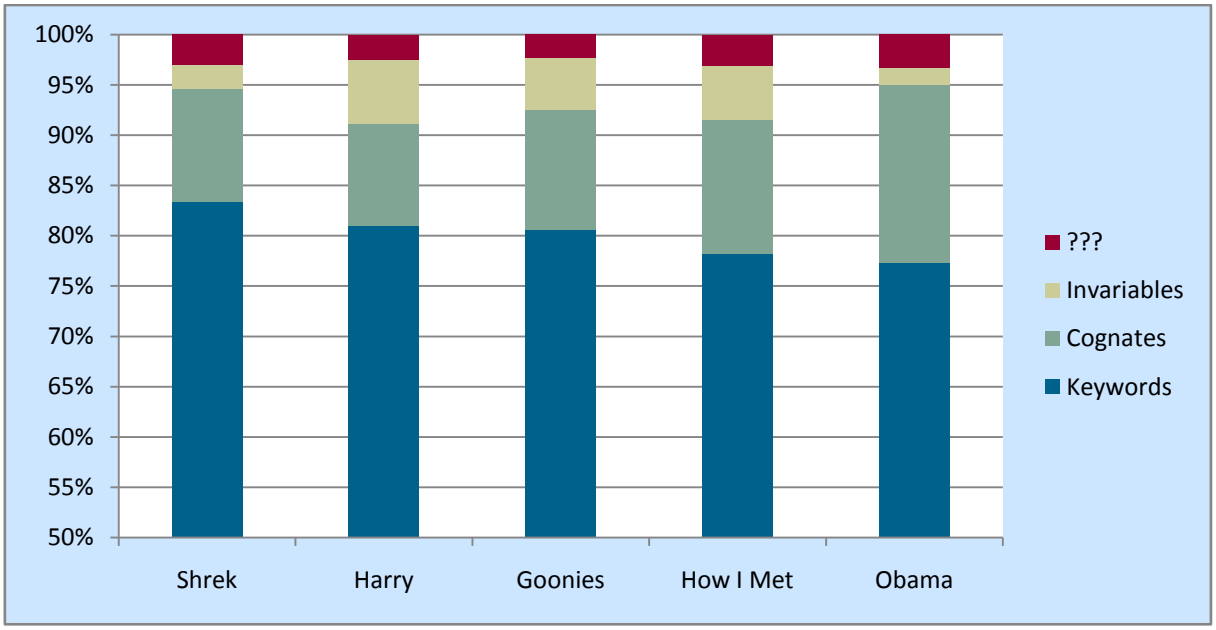


Figura II.14. Porcentaje acumulado del PLH en las muestras de habla no espontánea

La conclusión más destacable de este análisis es que en las muestras de distintos registros de habla espontánea y diferentes variedades dialectales el PLH ha alcanzado el umbral de autonomía al 95%. La consecuencia que deriva de ello es que el PLH parece ser un objetivo de aprendizaje más eficaz que las listas de frecuencia del BNC.

Tabla II.52. Comparativa de la cobertura del PLH frente al BNC en el corpus de habla no espontánea

	COBERTURA PLH (%)	EQUIVALENTE BNC	DIFERENCIA (Nº PAL)
OBAMA	96.67	4	2200
HARRY POTTER	97.48	5	3200
SHREK	96.98	8	6200
HOW I MET YOUR MOTHER	96.93	9	7200
GOONIES	97.63	10	8200

Por último, la Tabla II.52 compara el nivel del BNC necesario para alcanzar los mismos resultados que el PLH. Se observa que mientras el discurso de Obama y Harry Potter muestran unos valores similares a los que encontrábamos en la literatura, las muestras con registro coloquial presentan una gran diferencia en favor del PLH.

5. CONCLUSIONES GENERALES SOBRE LA PARTE PRIMERA

La aportación de esta parte de la investigación es el desarrollo de un Plan Léxico Adaptado para Hispanohablantes consistente en 1.800 palabras clave y una serie de afijos derivativos que permiten reconocer familias léxicas completas.

Los textos analizados componen un corpus de 359.626 tokens e incluye textos de las siguientes categorías:

- Materiales auténticos y adaptados.
- Variedades dialectales británica y americana.
- Lenguaje escrito y habla no espontánea.
- Registro culto e informal.
- Obras de los siglos XIX y XX y lenguaje actual.
- Obras dirigidas a un público infantil y exclusivamente adulto.

Si bien no es, estrictamente hablando, un corpus equilibrado, se puede considerar que es representativo de los géneros de mayor interés para las habilidades receptivas de los alumnos de L2: exámenes de certificación de nivel, obras literarias adaptadas y originales, películas, series de TV y discurso formal.

En cuanto a los resultados de los análisis de textos, el PLH ha superado el 95% de cobertura en prácticamente todas las muestras. Esto implica que ha alcanzado el denominado umbral de autonomía a partir del cual el vocabulario desconocido no plantea un obstáculo para la comprensión de la lectura. En muchas muestras se alcanzó incluso el 98% de cobertura, punto que define el umbral de garantía a partir del cual la lectura es fluida y placentera y se aprecian pequeños matices lingüísticos. Sobre el corpus global, la Figura II.15 muestra que el PLH ofrece una cobertura ligeramente superior al 96%, que engloba la suma de la representación de Keywords (81%), cognados (12,1%) y nombres propios y otras palabras invariables (3%). Esto implica que para un alumno que aprende los objetivos planteados por el PLH las palabras desconocidas supondrán únicamente un 3,9%.

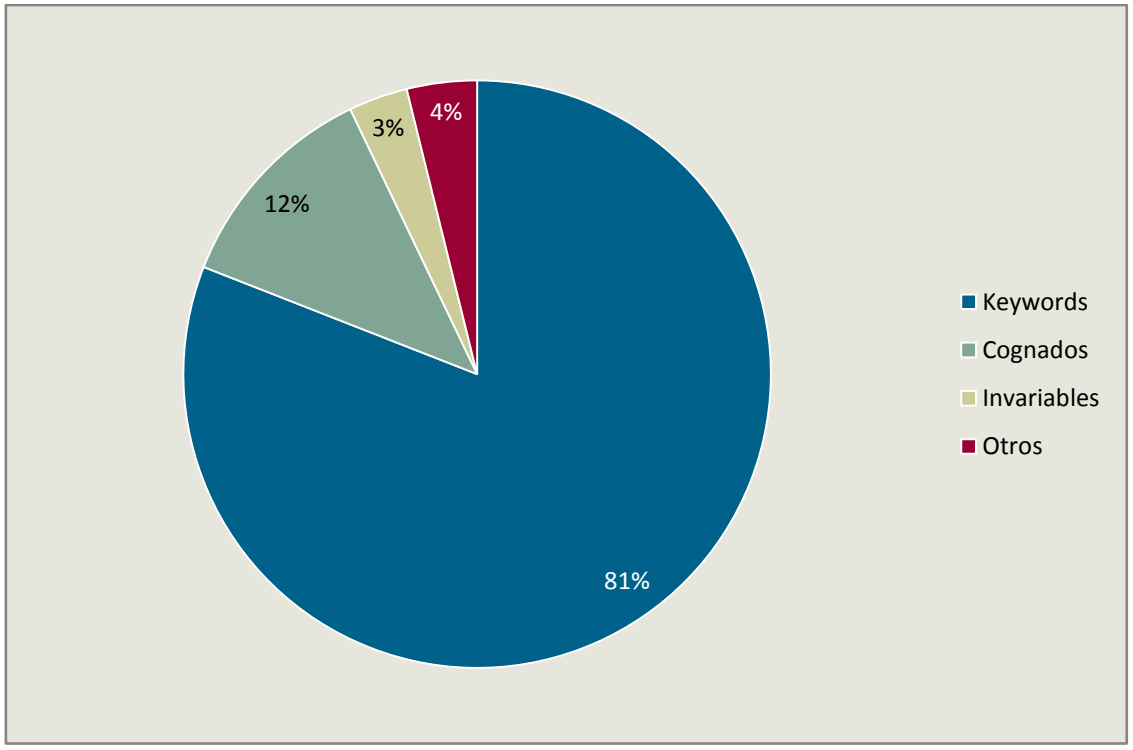


Figura II.15. Cobertura de las categorías del PLH sobre el corpus global

La innovación más relevante del PLH es que, hasta donde alcanza nuestro conocimiento, no hay ningún otro programa diseñado para instrucción explícita de vocabulario que valore el español L1 como facilitador de conocimiento potencial en inglés.

Los resultados abren un nuevo horizonte que amplía esa visión aún tan arraigada entre muchos profesores que consideran a la L1 únicamente como causa inevitable de errores, como consecuencia de la formación recibida basada en las teorías de Lado. En realidad, la diferencia fundamental entre otras listas de frecuencia y el PLH es precisamente que este aprovecha la influencia de la L1, y los resultados demuestran que, en lugar de ser causa de problemas, facilita de manera eficaz el aprendizaje.

Queremos destacar que para que un programa basado en las listas de Nation alcance la misma cobertura que el PLH sobre el corpus completo sería necesario un programa léxico de 6.000 palabras. Es decir, a través del PLH se llega al mismo punto dedicando menos de un tercio del tiempo y esfuerzo.

TERCERA PARTE

CONTINUIDAD TRAS EL PLH: SELECCIÓN LÉXICA EN PLANES DE LECTURA EXTENSA

III

1. LA EVALUACIÓN EN EL PLH

La evaluación es uno de los agentes más importantes en el sistema de enseñanza-aprendizaje. A menudo se asocia el término «evaluación» a un proceso que tiene lugar al final de un curso o una unidad para determinar el grado de éxito que han alcanzado docente y alumno sobre los objetivos de aprendizaje planteados. Sin pretender ser exhaustivos en los tipos de evaluación, resaltaremos únicamente que un buen sistema de evaluación debe transcurrir de manera continuada durante todo el proceso de enseñanza-aprendizaje y que una de sus funciones fundamentales es establecer un diagnóstico inicial sobre los conocimientos básicos del alumno así como sus herramientas, destrezas, hábitos y actitudes, entre otros.

En esta sección analizaremos un aspecto muy concreto de la evaluación: las pruebas a través de las que se puede analizar el grado de conocimiento del vocabulario receptivo. La integración del PLH en un curso requiere, en primer lugar, una prueba inicial de diagnóstico, que será fundamental para que el docente pueda ajustar los

contenidos pedagógicos a las características individuales del alumno. Durante la instrucción del PLH será necesario llevar a cabo distintas pruebas hasta verificar que el alumno ha alcanzado los objetivos planteados y, por tanto, debe pasar a la fase de lectura extensa.

El PLH puede ser la base sobre la que diseñar objetivos léxicos desde el nivel más elemental, pero también se puede incorporar fácilmente en la planificación didáctica de un curso con alumnos que no sean principiantes. Recordemos que el PLH no es una metodología, sino un conjunto de objetivos de aprendizaje que responden a un marco teórico sobre la enseñanza del vocabulario. En otras palabras, el PLH plantea una meta, pero no requiere una enseñanza lineal ni un sistema de trabajo específico que el alumno deba seguir paso a paso. Para adaptar el PLH a un curso dado, únicamente es necesario medir el punto de partida de los alumnos con el fin de adaptar los contenidos a sus necesidades. Los contenidos del PLH son muy específicos y están acotados al ámbito de las habilidades receptivas, por lo que no sirve cualquier prueba de nivel estándar. Es importante que el sistema de evaluación se ajuste a lo que debemos medir – alcance del vocabulario receptivo y reconocimiento de cognados –, y minimice dentro de lo posible las inferencias.

1.1. Tests validados y su aplicación en el PLH

Uno de los grandes retos a la hora de evaluar la base léxica de un alumno es escoger la prueba adecuada de entre las numerosas propuestas disponibles. John Read, sin duda la figura más influyente en la validación de pruebas de medida de vocabulario, nos advierte de que no debemos asumir que un test evalúa un aspecto concreto del vocabulario simplemente porque así lo haya etiquetado su autor (Read, 2000). En su artículo *A framework for second language vocabulary assessment* (Read & Chapelle, 2001) sugiere que antes de adoptar una prueba para medir el vocabulario en L2 es necesario plantearse previamente qué aspectos del vocabulario queremos medir y qué tipo prueba es la más adecuada para ofrecernos datos fiables sobre estos aspectos. La Tabla III.1 resume algunos de los puntos que se deben valorar según el marco de Read.

Tabla III.1. Variables en la adecuación de un test de vocabulario

ÁREA	EJEMPLOS
ASPECTOS	VOCABULARIO RECEPTIVO / PRODUCTIVO
	VOCABULARIO GENERAL / ESPECÍFICO
DISEÑO	PALABRAS AISLADAS / CONTEXTUALIZADAS
	TIPO TEST / RESPUESTA ABIERTA / RELLENAR HUECOS
	PUNTUACIÓN FIJA / SUBJETIVA
	CORRECCIÓN SOBRE PLANTILLA / CRITERIOS ABIERTOS
PROPÓSITO	INVESTIGACIÓN / DOCENCIA
	TEST DE DIAGNÓSTICO / TEST DE ALCANCE TRAS INSTRUCCIÓN
ALCANCE	VOCABULARIO BÁSICO / VOCABULARIO AVANZADO
	VOCABULARIO NATIVO / VOCABULARIO EN L2
VERIFICACIÓN	VALIDACIÓN POR EVIDENCIA / BASADO EN TEORÍA

Adaptado de Read & Chapelle, 2001

En el caso que nos ocupa, la prueba para evaluar el vocabulario de los alumnos que hayan seguido o vayan a seguir el PLH debe cumplir tres características fundamentales: (1) medir el vocabulario receptivo, (2) abarcar los rangos de alta y media frecuencia y, (3) estar diseñado para aprendices de L2. También se valorarán otros aspectos como que las palabras estén parcialmente contextualizadas o que esté enfocado a la docencia y no exclusivamente a la investigación[48].

A continuación discutiremos las tres pruebas más conocidas que cumplen estas características: *The Eurocentre Vocabulary Test*, *The Vocabulary Levels Test* y *The*

[48] En algunas pruebas diseñadas para la investigación se da por hecho la buena fe de los participantes. Por ejemplo, se puede pedir que en un texto dado subrayen las palabras que conocen y asumir la veracidad de las respuestas, como hizo Laufer en el experimento que dio lugar a la hipótesis del umbral mínimo (1989). Obviamente, este tipo de prueba no es fiable para certificar un nivel de inglés o para una prueba de acceso a una universidad.

Vocabulary Size Test. Han sido diseñadas por expertos de reconocido prestigio, validadas por marcos de verificación externos y ampliamente utilizadas en el ámbito educativo.

1.1.1. The Eurocentre Vocabulary Test

El *Eurocentres Test*, diseñado por Meara y Jones (1990), es actualmente uno de los test integrados en el proyecto DIALANG, un sistema on-line de evaluación del nivel de competencia lingüística sobre el Marco europeo de referencia en 14 idiomas[49].

La prueba contiene 5 niveles de 60 palabras cada uno, la referencia estadística para cada nivel viene determinada por las 10 primeras bandas de frecuencia de la lista de Thorndike y Lorge (1944). Tanto el listado de palabras como la versión on-line del «Yes/No Test» se pueden encontrar en la web del proyecto Lextutor de Tom Cobb. (http://www.lextutor.ca/tests/)

En la investigación léxica esta prueba es conocida como «The Yes/No Test» debido a su peculiar diseño, muy poco frecuente en pruebas de diagnóstico. Consiste en una serie de palabras con una casilla que el alumno únicamente debe marcar si conoce el significado de la palabra. Este test tiene un curioso sistema de control para evitar que el alumno mienta en las respuestas: se advierte a los participantes de que algunas de las palabras que aparecen no son términos reales en inglés. En la Tabla III.2 podemos encontrar *nonagrate, balfour, lannery, oxylate, degate* y *tooley*. Paul Meara las creó para el «Yes/No Test» bajo el nombre de *plausibles non-words*. Su mayor ventaja, tal y como destaca Read (2000), es que la simplicidad de la tarea permite procesar un gran número de palabras dentro del limitado tiempo de un test, por lo que el tamaño de la muestra permite alcanzar estimaciones muy fiables.

[49] http://www.lancaster.ac.uk/researchenterprise/dialang/about.htm

Tabla III.2. The Yes/No Test: primeras 30 palabras

1	☐	OBEY	11	☐	DOOR	21	☐	DEGATE
2	☐	THIRSTY	12	☐	GROW	22	☐	BATH
3	☐	NONAGRATE	13	☐	LANNERY	23	☐	BIRTH
4	☐	EXPECT	14	☐	RED	24	☐	GUMMER
5	☐	LARGE	15	☐	PLATE	25	☐	CHRISTIAN
6	☐	ACCIDENT	16	☐	HOLD	26	☐	SUCCEED
7	☐	COMMON	17	☐	LOVE	27	☐	CANTILEEN
8	☐	SHINE	18	☐	PULL	28	☐	WARM
9	☐	SADLY	19	☐	ENOUGH	29	☐	SONG
10	☐	BALFOUR	20	☐	OXYLATE	30	☐	TOOLEY

Sin embargo, este test tiene una dificultad intrínseca para alumnos hispanohablantes. Muchas de las pseudopalabras de control están fabricadas a partir de raíces latinas + afijos ingleses, de donde se obtienen presuntos cognados transparentes que en realidad no existen. Algunas de estas palabras artificiales son *condimented*, *descript*, *escrotal* y *fluctual*. Parece evidente es que si estas palabras fueran reales, su significado —e incluso su categoría gramatical— serían evidentes para un hispanohablante, lo que puede dar lugar a una falsa sensación de conocimiento. También la advertencia inicial de que hay palabras falsas puede tener el efecto contrario: dudar de la veracidad de otros cognados transparentes que sí son reales. El sistema de control, por tanto, añade un factor de estrés innecesario y los resultados se pueden ver afectados tanto por un exceso de confianza como por una cautela desmedida.

A modo de conclusión, utilizar este test para medir la base léxica del alumno hispanohablante puede dar resultados incorrectos, debido a una dificultad añadida en la que está involucrada su L1. Emitir un juicio negativo sobre palabras plausibles cognadas es una tarea que requiere unos conocimientos léxicos más profundos de los que, en principio, debería medir el test. Consideramos que el diseño de la prueba puede

ser válido siempre que se eliminen de entre las palabras artificiales aquellas que son cognadas, o se pasen por alto en el sistema de puntuación. Por otro lado, estos cognados artificiales sí podrían ser útiles para medir el grado de profundidad en los conocimientos del vocabulario de estudiantes muy avanzados, pero no en la fase de enseñanza explícita del PLH.

1.1.2. The Vocabulary Levels Test (VLT)

Se diseñó para medir el vocabulario receptivo en L2 en cinco rangos de frecuencia: 2.000; 3.000; 5.000; 10.000 y lenguaje académico. El diseño de la prueba original es de Paul Nation (1990) y su está validado por el marco de evaluación de Read (2000).

Nation publicó una actualización del VLT en 2001, y Schmitt, Schmitt y Clapham desarrollaron y verificaron dos nuevas versiones (2001). Uno de los cambios más destacables es que la prueba original tomaba el lenguaje académico de la University Word List (Xue & Nation, 1984) y en versiones posteriores esta se reemplazó por la *Academic Word List* (Coxhead, 2000). La versión de 1990 está disponible como test on-line en LexTutor [50] y las de Schmitt et al. (2001) se encuentran también en los apéndices de *Researching Vocabulary* (Schmitt, 2010).

En todos ellos el formato es el mismo: cada pregunta tiene tres definiciones y seis palabras; es decir, hay tres respuestas correctas y tres distractores. El alumno debe escribir el número de la palabra que corresponde a cada definición, como se muestra en la Tabla III.3. El test tiene cinco niveles y en cada uno de ellos hay seis preguntas, la tabla muestra las dos primeras preguntas del test. El problema principal es que si entre las palabras clave hay muchos cognados verdaderos, esto puede suponer una ventaja injusta para alumnos hispanohablantes, y un exceso de falsos cognados podría tener el efecto contrario. En ambos casos se obtendrían puntuaciones injustas que no se corresponden con el nivel real del alumno.

[50] http://www.lextutor.ca/tests/levels/recognition/2_10k

Tabla III.3. Dos preguntas del nivel 1 del test VLT de Nation (1990)

1. ADMINISTRATION		
2. ANGEL	_1_	MANAGING BUSINESS AND AFFAIRS
3. FRONT	_2_	SPIRIT WHO SERVES GOD
4. HERD	_4_	GROUP OF ANIMALS
5. MATE		
6. POND		
1. BENCH		
2. CHARITY	_6_	PART OF A COUNTRY
3. FORT	_2_	HELP TO THE POOR
4. JAR	_1_	LONG SEAT
5. MIRROR		
6. PROVINCE		

Adaptado de LexTutor (http://www.lextutor.ca/tests/levels/recognition/2_10k/test.html)

Analizamos a continuación la influencia de la L1 en una de las versiones más modernas del test (Schmitt et al., 2001). Contando tanto los distractores como las respuestas correctas, hay 60 palabras por nivel, 300 palabras en total. La siguiente tabla muestra la presencia de cognados en el test. Entre todas ellas únicamente encontramos 8 falsos cognados, una cifra dentro de los límites habituales. Por su parte, los cognados verdaderos representan el 36,67% del primer rango de frecuencia. Esta proporción es aún mayor en el tercer nivel del test, donde casi la mitad de las palabras son cognadas.

Tabla III.4. Porcentaje de cognados en el VLT de Schmitt et al. (2001)

NIVEL	2.000	3.000	5.000	10.000	ACADEMIC
COGNADOS (%)	36.67	46.67	48.33	40.00	78.33
FALSOS COGNADOS (%)	5.00	6.67	1.67	0.00	0.00
??? (%)	58.33	6.67	50.00	60.00	21.67

En la Figura III.1 observamos que en el lenguaje académico, como es previsible, los cognados suponen la gran mayoría de las palabras. Es más, en el VLT están incluso sobrerrepresentados. Si bien la AWL, fuente de donde se han extraído las palabras, ya tiene un 70% de cognados, en este rango del VLT suponen el 78%.

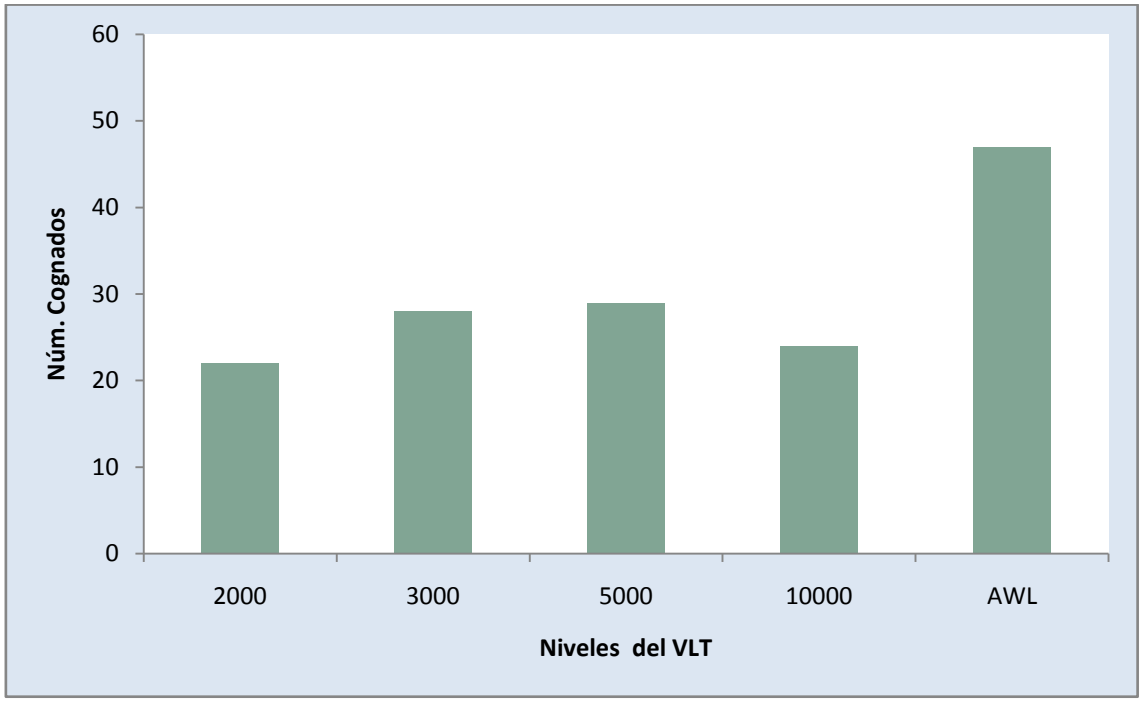


Figura III.1. Cognados en el VLT de Schmitt et al. (2001)

Estos datos nos indican que un alumno hispanohablante con una base mínima de inglés encontraría más asequible el rango de frecuencia 5.000 —nivel necesario para aprobar el Cambridge Proficiency of English— que el rango 2.000, que corresponde a una base léxica inferior a B1, según la estimación de Milton (2009).

De hecho, cuando Schmitt et al. (2001) exploraron los resultados de sus nuevas versiones del VLT, tuvieron en cuenta que los cognados podrían dar lugar a perfiles atípicos entre los hablantes de español y portugués. Tras examinar los datos obtenidos,

concluyeron que las puntuaciones de los hablantes de lenguas romances en los niveles 5.000 y 10.000 del VLT no pueden considerarse válidas:

> Speakers of Romance languages have a distinct advantage in this regard, for many English words can be guessed according to their similarity to Romance words. […] We had a large number of Spanish/ Portuguese examinees whom we could isolate in order to give an indication of this, so the first subset we looked at consisted of Romance speakers. […] Strictly speaking, the 5000 and 10 000 levels cannot be considered equivalent for the Romance speakers. (Schmitt et al., 2001: 77)

> Los hablantes de lenguas romances tienen una clara ventaja en este sentido, dado que muchas palabras en inglés pueden ser deducidas por su semejanza con palabras de las lenguas romances. [...] Examinamos a un gran número de españoles y portugueses a los que pudimos aislar para obtener señales de esto, así que el primer subconjunto que miramos fue el de hablantes de lenguas romances. […] Estrictamente hablando, los niveles de 5000 y 10.000 no pueden considerarse equivalentes para los hablantes de lenguas romances (trad.a.).

Nuestra conclusión es que el formato del VLT, verificado por Read (2000), es óptimo para medir el vocabulario exclusivamente receptivo, pero el docente PLH debe tener presente que el sistema de puntuación devolverá perfiles lingüísticos atípicos. Este test, sin embargo, puede resultar extremadamente útil para evaluar a alumnos PLH siempre que se puntúen por separado los cognados. Esto nos permitirá conocer la base léxica del alumno en palabras clave del PLH así como su habilidad para deducir cognados.

1.1.3. The Vocabulary Size Test (VST)

El VST mide el vocabulario receptivo, tanto de hablantes nativos como de aprendices de L2, en 14 niveles de frecuencia equivalentes a 1.000 palabras cada uno. Su diseño está avalado por Laufer y Nation (1999) y se ajusta al modelo Rasch de validación tests (Beglar, 2010). Una versión actualizada se puede encontrar en el

apéndice de *Researching Vocabulary*[51] (Schmitt, 2010). Por su parte, Nation ha publicado dos adaptaciones del VST ampliadas hasta el rango 20, así como versiones bilingües en coreano, japonés, chino, ruso y vietnamita[52]. En todas las versiones, cada nivel contiene 10 preguntas de respuesta múltiple, hay una palabra clave parcialmente contextualizada y cuatro definiciones de entre las que solo una es correcta, como se ve en la muestra de la Tabla III.5. Para evitar inferencias, las definiciones emplean siempre un vocabulario muy básico, como se puede ver en las preguntas del nivel 14, que corresponde al rango 14.000.

Tabla III.5. Diseño de las preguntas del VST

NIVEL 1	**NIVEL 14**
TIME: They have a lot of **time**.	MARSUPIAL: It is a **marsupial**.
a. money	a. an animal with hard feet
b. food	b. a plant that grows for several years
c. hours	c. a plant with flowers that turn to face the sun
d. friends	d. an animal with a pocket for babies

Analicemos ahora la influencia de la L1 en las puntuaciones previsibles para este test en la versión de Schmitt (2010). En la Tabla III.6 podemos ver las 140 palabras que componen los 14 niveles del test. Encontramos únicamente 2 falsos cognados —*figures* (en su acepción 'cifras') en el primer nivel, y *talons* en el 13º— que suponen únicamente el 1,42% de las palabras. Los cognados transparentes están marcados en letra negrita redonda y los cognados parciales, en negrita cursiva. Se han valorado estos últimos porque la respuesta múltiple ayuda mucho a su reconocimiento. Por ejemplo, la palabra *roubles* es un cognado parcial que si apareciera descontextualizada podría deducirse erróneamente como 'robles'. Sin embargo, al leer que una de las opciones es «*Russian money*» es muy probable que el alumno asocie «*roubles* = rublos».

[51] La versión de Schmitt está también disponible como test on-line en LexTutor: http://www.lextutor.ca/tests/levels/recognition/1_14k/

[52] Disponibles en la página web personal de Paul Nation (www.victoria.ac.nz/lals/about/staff/paul-nation)

Tabla III.6. Palabras de VST de Schmitt

RANGO	PALABRAS
1.000	**BASIS**; DRIVE; FIGURE; JUMP; **PERIOD**; POOR; SEE; SHOE; **STANDARDS**; TIME
2.000	***CIRCLE***; DRAWER; ***MAINTAIN***; **MICROPHONE**; ***NIL***; **PATIENCE**; **PRO** ; **PUB**; STONE; UPSET
3.000	DASHED; **DINOSAURS**; JUG; LONESOME; ***PAVED***; **RESTORED**; ROVING; SCRUBBING; ***SOLDIER***; STRAP
4.000	**ALLEGED**; **CANDID**; ***COMPOUND***; CRABS; **INPUT**; LATTER; **QUIZ**; **REMEDY**; TUMMY; **VOCABULARY**
5.000	**BACTERIUM; COMPOST; CUBE; DEFICIT; FRACTURE**; HAUNTED; **MINIATURE**; NUN; ***PEEL***; WEPT
6.000	**ACCESSORIES**; BUTLER; ***CAVALIER***; DEVIOUS; **MALIGN**; PREMIER; **STRANGLED**; **THESIS**; THRESHOLD; VEERED
7.000	**AZALEA**; **BLOC**; BRISTLE; **DEMOGRAPHY**; GIMMICK; **OLIVES**; QUILT; SHUDDERED; STEALTH; **YOGHURT**
8.000	**AUTHENTIC**; CABARET; **ECLIPSE**; **ERRATIC**; **KINDERGARTEN**; LOCUSTS; MARROW; MUMBLE; **NULL**; **PALETTE**
9.000	FENS; HALLMARK; LINTEL; **MONOLOGUE**; OCTOPUS; **PERTURBED**; **PURITAN**; **REGENT**; WEIR; WHIMS
10.000	AWE; CRANNY; CROWBAR; **EGALITARIAN**; LECTERN; **MYSTIQUE**; PEASANTRY; PIGTAIL; RUCK; UPBEAT
11.000	**APERITIF**; COUNTERCLAIM; **EMIR**; **EXCRETED**; HESSIAN; HUTCH; MUSSEL; PALLOR; **PUMA**; **YOGA**
12.000	***ALUM***; **CAFFEINE**; COVEN; HAZE; ***IMPALED***; **REFECTORY**; **REPTILE**; **SOLILOQUY**; SPLEEN; TRILL
13.000	***ATOLL***; BEAGLES; **COMMUNIQUÉ**; **DIDACTIC; JOVIAL; PLANKTON; *ROUBLES***; SKYLARK; TALONS**; *UBIQUITOUS***
14.000	ATOP; **AUGURED**; BAWDY; **CANONICAL**; **CORDILLERA**; **ERYTHROCYTE**; GAUCHE; **LIMPID**; **MARSUPIAL**; **THESAURUS**

Las palabras del VLT en la Tabla III.6 ilustran una peculiaridad lingüística: el número de caracteres es menor en las palabras frecuentes. En del primer rango del VLT tienen una media de 5 letras, que aumenta de manera proporcional hasta el rango 14.000, cuyas palabras tienen una media de 7,6 letras. Esto se podría atribuir a que en los rangos de baja frecuencia hay más palabras de raíz latina, que tienden a ser más largas que las germánicas. La explicación se encuentra, una vez más, en la Ley de Zipf: el número de caracteres de las palabras es inversamente proporcional a su frecuencia, y la proporción sigue una ley de potencias similar a la de rango / frecuencia (Manning & Schütze, 1999). Este fenómeno ocurre en la mayoría de las lenguas debido a la economía del esfuerzo. Lo que posiblemente ocurrió en el caso del inglés fue que las palabras latinas más frecuentes —y, por tanto, más cortas— no se incorporaron al léxico, porque esto implicaría reemplazar a palabras germánicas de uso muy común. Por esta razón, los préstamos latinos pertenecen, en su mayoría, a los rangos de frecuencia baja, que es donde suelen encontrarse las palabras largas en la mayoría de las lenguas.

La Figura III.2 muestra la proporción de cognados transparentes y parciales en el VST. Como esta prueba alcanza rangos mayores que el VLT, la fiabilidad es incluso menor. Observamos que en el último nivel se pueden deducir 7 de las 10 palabras. Esto implica que un hispanohablante con escasos conocimientos de inglés puede sacar una puntuación muy alta en el rango 14.000, que corresponde a una base léxica muy superior a la de un nativo de 12 años[53].

Por otro lado, esto solamente podría ocurrir si el alumno tiene un vocabulario muy rico en L1. El test requiere que se escoja la definición correcta, por lo que el estudiante tendría que saber el significado en L1 de palabras como *refectorio, soliloquio, canónico, eritrocito* y *tesauro*. Las palabras cognadas que son muy poco frecuentes en inglés tienden a serlo también en español.

[53] Ver punto 3.1 de la Primera parte: *Tamaño del vocabulario de un nativo* (p. 24)

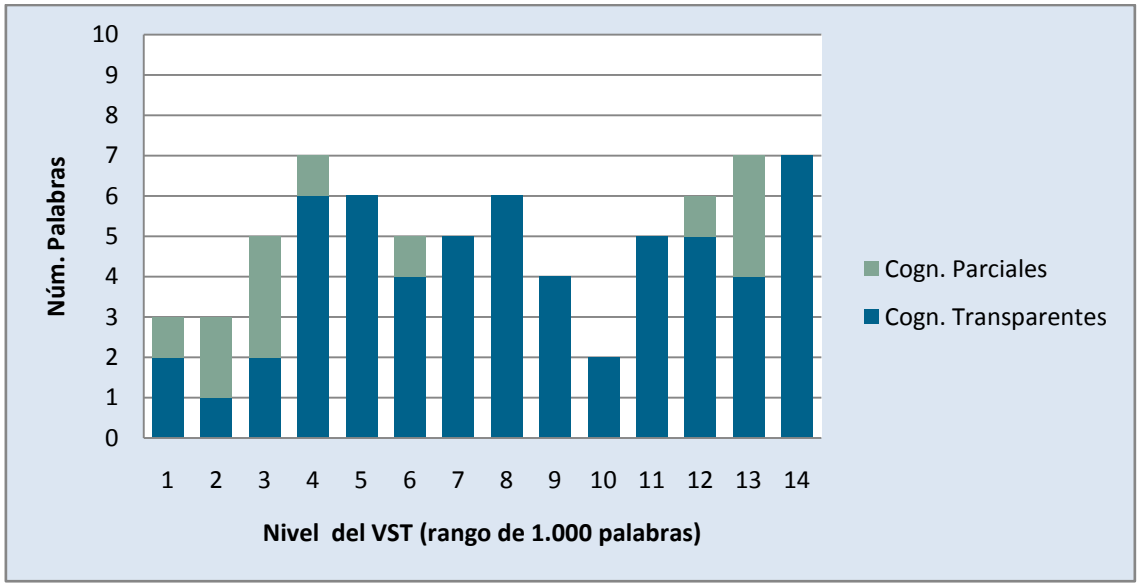


Figura III.2. Proporción de cognados en el VST

1.2. Conclusiones de las pruebas de evaluación

La evaluación del vocabulario receptivo requiere un tipo de prueba muy específico. Los sistemas de validación como el modelo de Read y Chapelle (2001) verifican que el diseño de la prueba es óptimo en cuanto a sus parámetros y sistema de puntuación, pero no valoran aspectos relacionados con las habilidades individuales de un alumno relativas a su L1. Por tanto, es muy importante tener en cuenta que la alta presencia de cognados puede invalidar los resultados, especialmente en los niveles altos.

De entre las tres pruebas analizadas, el «Yes/No Test» parece tener una dificultad añadida para el alumno hispanohablante, dado que muchas palabras artificiales están formadas a partir de raíces latinas. Este test puede ser interesante para medir conocimientos muy avanzados de vocabulario, pero no parece el más adecuado para evaluar a los alumnos PLH. Por el contrario, tanto el *Vocabulary Levels Test* como el *Vocabulary Size Test* parecen ser pruebas válidas para evaluar la base léxica de un alumno PLH, siempre que se tenga en cuenta la influencia de los cognados. Nuestra propuesta para evitar los falsos perfiles lingüísticos es eliminar las palabras cognadas de los test, o bien puntuarlas por separado. Esta última opción nos parece la más

interesante, ya que los resultados nos darán información tanto del vocabulario PLH como de la habilidad del alumno para reconocer cognados.

2. DESPUÉS DEL PLH: SELECCIÓN LÉXICA EN PLANES DE LECTURA

2.1. Fundamentos teóricos

El objetivo básico sobre el que se diseña el PLH es preparar a los alumnos para leer textos auténticos lo antes posible. Una vez terminada la instrucción del PLH, el siguiente reto es que los alumnos «aprendan a aprender» palabras nuevas a través de los textos que ya están preparados para leer. En un plan de lectura extensa el número de palabras desconocidas puede ser muy alto, por lo que es necesario que el docente sepa cómo seleccionar las más importantes con el fin de establecer objetivos de aprendizaje y elaborar materiales de apoyo léxico, tales como glosarios o ejercicios de vocabulario.

Como consecuencia del desprestigio que ha sufrido la enseñanza explícita del vocabulario, los cursos de formación de docentes de L2 no suelen incluir entrenamiento específico sobre las distintas estrategias que favorecen la adquisición de vocabulario a través de la lectura. Aparte de unas pocas técnicas básicas, como la deducción por contexto, el uso de diccionarios o, en el mejor de los casos, pequeños glosarios que suelen venir con los libros adaptados, el docente no tiene suficientes herramientas para enseñar a sus alumnos a sacar el máximo provecho posible del esfuerzo que invierten en la lectura. Una cuestión fundamental que muchos docentes no están preparados para responder es cómo se pueden determinar las unidades léxicas de mayor interés con las que se va a encontrar el alumno.

Es frecuente, por tanto, encontrar a los alumnos perdidos en el proceso del aprendizaje de vocabulario a través de textos, especialmente si son extensos y están programados para la lectura independiente fuera del aula. Las fórmulas que emplean para lidiar con el vocabulario desconocido suelen depender de sus preferencias personales más que de una estrategia fundamentada en propuestas pedagógicas basadas en la investigación. Muchos alumnos se centran únicamente en avanzar en la lectura, utilizando la deducción por contexto siempre que sea posible y el diccionario solo si un fragmento contiene tanto vocabulario desconocido que es imposible su comprensión. En el extremo opuesto encontramos alumnos muy interesados en enriquecer su vocabulario

y que para ello elaboran un glosario personal con el significado de todas las palabras desconocidas tras su consulta en el diccionario.

Redman (2011) denomina a esta última estrategia «la técnica de la pared de ladrillo», una analogía en la que las palabras desconocidas son pelotas lanzadas a un jugador de tenis que no debe dejar que alcancen una pared que se encuentra detrás de él. Esta técnica se podría resumir como:

«Aprende cada palabra que encuentres, sin importar lo rara o difícil que sea.»

Es habitual que en las lecturas que se hacen dentro del aula se aliente al alumno para que incluya en sus notas absolutamente todas las palabras desconocidas que encuentre en el texto y las repase posteriormente. En otras palabras, la técnica de la pared de ladrillo se fomenta en las lecturas de clase porque estas suelen ser fragmentos cortos incluidos en los libros de texto o seleccionados de otras fuentes por el profesor. El problema es que, a menos que se les indique lo contrario, los alumnos pueden asumir que esta es la técnica que deben emplear para todo tipo de lecturas. En un texto corto la pared de ladrillo no supone un problema pero en un libro de 300 páginas puede haber una gran cantidad de vocabulario desconocido, demasiado como para aprenderlo todo de una vez.

El objetivo de este capítulo es establecer un criterio para seleccionar el vocabulario fuera del PLH (en adelante, NoPLH) para un plan de lectura extenso, de tal manera que el número de objetivos de aprendizaje sea asumible. Se pretende aportar al docente información sobre cómo elaborar una planificación léxica eficaz para una colección de lecturas concretas, para que posteriormente pueda trabajar este vocabulario con la técnica que considere más conveniente.

En la investigación previa se pueden encontrar numerosas propuestas metodológicas que han demostrado su eficacia para mejorar la memorización y el proceso del vocabulario mediante la lectura extensa. Aunque el análisis de estas técnicas está fuera del ámbito de esta tesis, se puede encontrar información al respecto en Coady (1997), Krashen (1993), Laufer & Hulstijn (2001), Nagy et al. (1985), Oxford & Scarcella (1994), Pigada & Schmitt (2006) y Schmitt (2010). A modo de resumen, la investigación no deja dudas de que dos factores indispensables para la adquisición del

vocabulario mediante la lectura son el número de exposiciones de una palabra y su repaso a intervalos temporales (Huckin et al., 1995; Huckin & Coady, 1999; Laufer & Osimo, 1991; Laufer & Hulstijn, 2001; Milton, 2009; Schmitt, 1997; Schmitt, 2010).

2.2. Aplicación

Para ilustrar el proceso de selección léxica con un ejemplo concreto, en esta sección explicaremos cómo establecer los objetivos de aprendizaje para la versión adaptada de *The Adventures of Sherlock Holmes* para una base léxica de 8.000 palabras, el nivel más alto de Mid-frequency Readers. Para poner en perspectiva la extensión de esta obra, publicada únicamente como e-book, su versión impresa en formato de libro de bolsillo estándar tendría aproximadamente 300 páginas.Hemos escogido este libro porque es una colección de 12 relatos diferentes, lo que nos permite analizar los 12 capítulos como textos independientes que formaran parte de un plan de lectura global. De esta manera, el criterio de selección léxica que explicamos a continuación se puede replicar tanto con un libro como con cualquier colección de textos que no estén relacionados entre sí y formen parte del plan de lectura para un curso dado.

Partimos de la demostración estadística previa de que el PLH aporta al alumno el 95% de la cobertura del vocabulario de un texto medio. Si bien es un porcentaje muy alto, el número absoluto correspondiente al 5% de palabras presumiblemente desconocidas es realmente considerable en un texto largo. En el caso concreto de *The Adventures of Sherlock Holmes,* su extensión supera los 10.000 tokens, entre los que se encuentran 3.141 tokens NoPLH agrupados en 1.569 palabras distintas y 1.153 familias léxicas[54]. Es decir, el alumno PLH encontrará a lo largo del texto 1.153 palabras desconocidas diferentes, algunas de las cuales se repetirán varias veces, ya sea en su forma base o como formas derivadas mediante afijos.

Mediante la técnica de la pared de ladrillo, un alumno tendría que anotar y repasar las 1.153 palabras a medida que las fuera encontrando en la lectura. En la Tabla III.7 se

[54] En este trabajo utilizamos el PLH como referencia, pero el sistema que se describe en este capítulo se puede replicar para cursos con alumnos de procedencia heterogénea que conozcan al menos las 3.000 palabras inglesas más frecuentes. En este caso, en lugar de Keywords y Cognates se utilizarán como *stop-list* otras listas de frecuencia, como los 3 primeros niveles de Nation o el conjunto de la General Service List + *Academic Word List*, por citar algunos ejemplos conocidos.

pueden consultar el número total de tokens y las palabras presumiblemente desconocidas que contiene cada relato (columna «palabras NoPLH»). Uno de los datos más interesantes es la 4ª columna, «Nuevas en capítulo», que muestra cuántas de las palabras NoPLH aparecen por primera vez en cada uno de los relatos, es decir, el número de objetivos de aprendizaje por capítulo si utilizáramos la técnica de la pared de ladrillo. La última columna nos muestra el acumulado de las palabras que el alumno habría aprendido al final de cada capítulo.

Tabla III.7. Distribución de elementos no PLH por capítulos en Sherlock Holmes

CAPÍTULO	TOKENS	PALABRAS NO-PLH	NUEVAS EN CAPÍTULO	APRENDIDAS
CAP1	8622	178	178	178
CAP2	9240	202	167	345
CAP3	7071	141	86	431
CAP4	9723	191	128	559
CAP5	7364	169	88	647
CAP6	9323	233	109	756
CAP7	7919	168	60	816
CAP8	9925	225	84	900
CAP9	8381	205	83	983
CAP10	8173	146	53	1036
CAP11	9762	174	54	1090
CAP12	10027	190	63	1153
TOTAL	105530	-	1153	1153

Las repeticiones son un factor muy importante para la aplicación pedagógica de estos datos. Podemos observar que aunque el último capítulo contiene más palabras NoPLH que el primero (190 frente a 178), sin embargo tiene casi 3 veces menos objetivos de aprendizaje (63 frente a 178). Evidentemente, en el primer capítulo todas las palabras NoPLH que aparezcan son desconocidas pero en los capítulos sucesivos se cuenta con el conocimiento previo de las palabras que han aparecido anteriormente.

Estos datos ilustran claramente la imposibilidad de utilizar la técnica de la pared de ladrillo en lectura extensa: aprender más de 1.100 palabras de un libro de 300 páginas es, a todas luces, un objetivo inasumible. Un criterio para bajar la carga léxica atendiendo a la rentabilidad podría ser descartar aquellas palabras que solamente tienen una ocurrencia en el texto. Estas palabras se denominan «hápax legómenon».

2.3. Hápax legómenon

La expresión hápax legómenon, que en griego significa 'lo dicho una vez', designa a las palabras que solamente se encuentran una vez en un contexto dado, ya sea en un texto concreto, un corpus extenso o los registros escritos de toda una lengua. Uno de los ejemplos más conocidos es la palabra *golem*, que es un hápax legómenon de la Biblia. Es importante remarcar que un hápax legómenon es relativo a un contexto concreto, no significa que realmente se haya «dicho una única vez» o que sea el único registro histórico de la palabra. Estadísticamente los hápax legómenon son solo una pequeña parte de los tokens de cualquier texto o corpus, pero constituyen aproximadamente la mitad de sus tipos (palabras distintas). Esto es debido a la curva recíproca que presenta la distribución de las palabras por la ley de Zipf (Baayen, 2001; Manning & Schütze, 1999).

En lingüística computacional, cuando se procesan grandes cantidades de datos se tiende a descartar los hápax legómenon porque consumen demasiados recursos de memoria del ordenador, pero no afectan de manera significativa al modelo cualitativo (Manning & Schütze, 1999: 199). Aplicando esta premisa a la selección léxica con fines docentes, se podría considerar que el descarte de los hápax legómenon nos permite reducir a la mitad unos objetivos de aprendizaje excesivos para el alumno, sin que esto afecte de manera significativa a su capacidad de entender el texto. Por otro lado, la investigación ha demostrado que la probabilidad de retener palabras que aparecerán una única vez en el texto es realmente baja, por lo que mantener los hápax legómenon como objetivo de aprendizaje requeriría varios ejercicios complementarios de repaso y consolidación.

Cumpliendo lo previsto por la Ley de Zipf, prácticamente la mitad de las palabras NoPLH de *The Advenutres of Sherlock Holmes 8000* aparecen solamente una vez en toda la obra. La Figura III.3 muestra que las repeticiones de las palabras NoPLH forman

una curva de Zipf en su distribución de Rango/Frecuencia. Se observa cómo los hápax legómenon, representados en el extremo izquierdo del eje x, son el grupo más numeroso, y el número de palabras desciende proporcionalmente al aumentar las repeticiones. La dispersión de estos mismos datos en escala doble-logarítmica se puede ver en la Figura III.4.

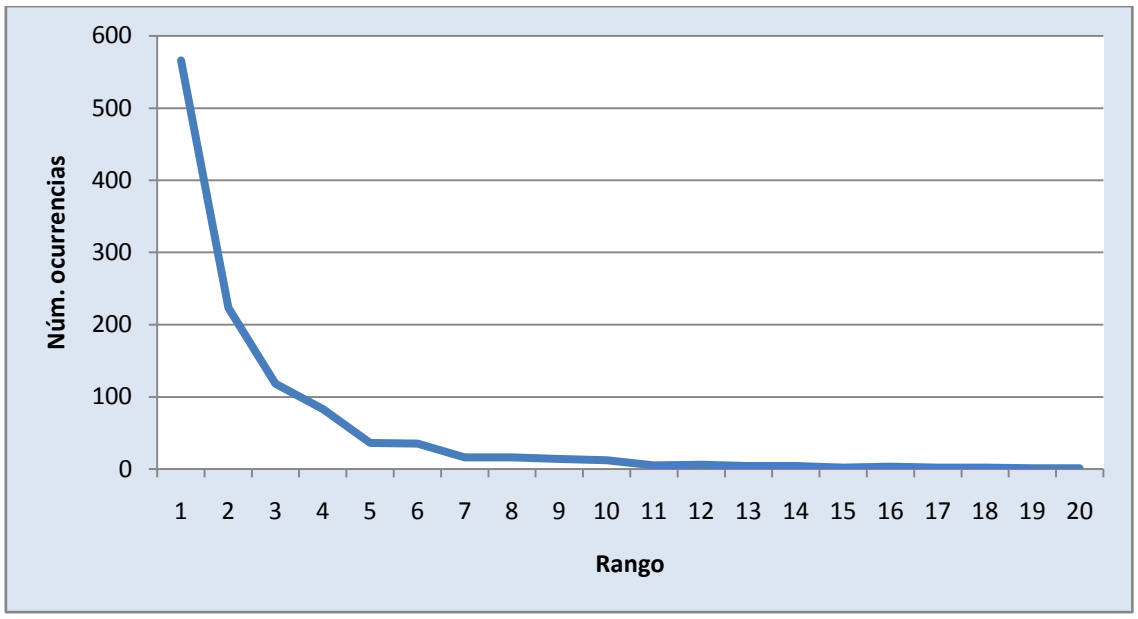


Figura III.3. Distribución de Zipf de Rango/Frecuencia en las palabras NoPLH en Sherlock Holmes

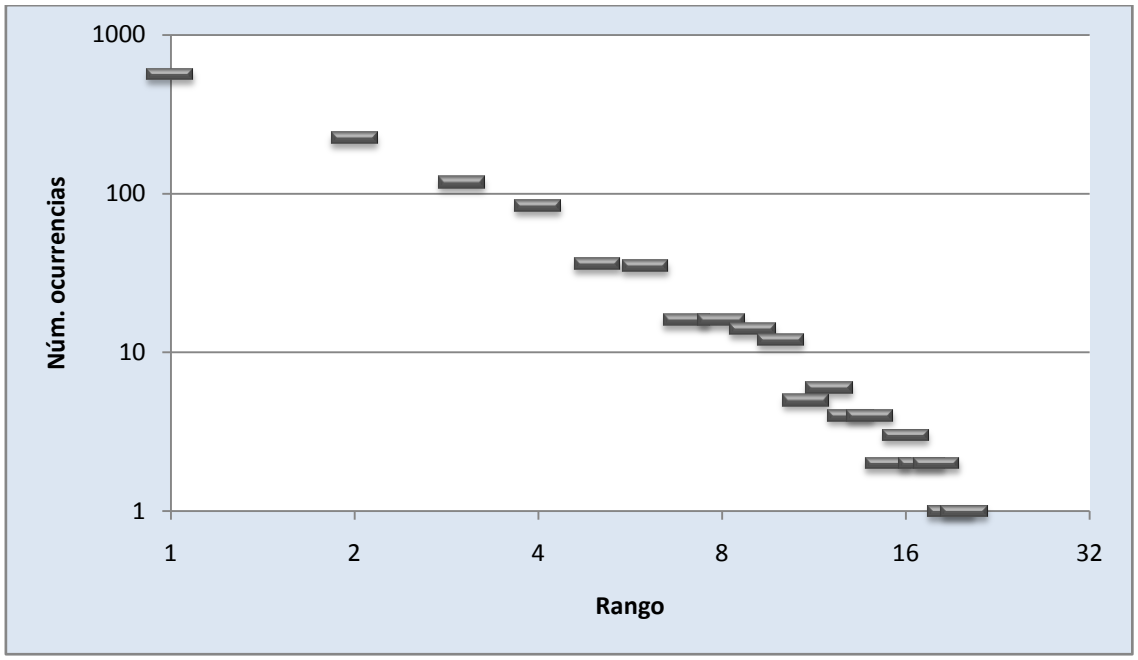


Figura III.4. Representación doble logarítmica de la dispersión de Rango/Frecuencia en las palabras NoPLH en Sherlock Holmes

En la Tabla III.8 las dos columnas a la izquierda muestran el número de repeticiones exactas, y las columnas de la derecha detallan cuántas palabras aparecen *al menos n* veces. En la primera columna observamos que de las 1.153 familias NoPLH, 556 son hápax legómenon. En la segunda columna podemos apreciar la demostración empírica de otra consecuencia de la Ley de Zipf: aproximadamente la mitad de las 1.153 palabras aparecen al menos 2 veces, un tercio aparecen al menos 3 veces, y así sucesivamente. Si estimamos que son necesarias, digamos 5 repeticiones, encontraremos que hay 163 palabras NoPLH que cumplen esta condición en el texto.

Tabla III.8. Hapax Legómenon y palabras repetidas en Sherlock Holmes

REPETICIONES	PALABRAS	REPETICIONES	PALABRAS
1	566	≥ 1	1153
2	223	≥ 2	597
3	118	≥ 3	364
4	83	≥ 4	246
5	36	≥ 5	163
6	35	≥ 6	127
7	16	≥ 7	92
8	16	≥ 8	76
9	14	≥ 9	60
10	12	≥ 10	46
11	5	≥ 11	34
12	6	≥ 12	29
13	4	≥ 13	23
14	4	≥ 14	19
15	2	≥ 15	15
16	3	≥ 16	13
17	2	≥ 17	10
18	2	≥ 18	8
19	0	≥ 19	6
20	1	≥ 20	6
>20	5	>20	5

2.4. Otros elementos de baja frecuencia

Una vez descartados los hápax legómenon, tendríamos como potenciales objetivos de aprendizaje las 597 palabras NoPLH que se repiten al menos una vez. De nuevo, podemos considerar que sigue siendo un número excesivo, por lo que habría que hacer una segunda criba.

Una opción sencilla podría ser repetir el proceso descartando esta vez las palabras que solamente aparecen solo dos, tres o incluso cuatro veces en el texto (*dis legomenon*, *tris legomenon* y *tetrakis legomenon*, respectivamente) hasta que el número de palabras resultante sea adecuado para nuestro propósito. En la columna derecha de la Tabla III.8 podemos comprobar que el número de objetivos de aprendizaje se reduciría a 364, a 246 y a 163 palabras, respectivamente.

Esto aligeraría la carga léxica del alumno y, simultáneamente, permitiría al docente planificar una metodología basada en el reciclaje a intervalos, ya que inevitablemente el alumno encontrará las palabras seleccionadas varias veces a lo largo de la lectura. Sin embargo, el problema es que un término que aparece con una frecuencia inusitada en un texto dado tiene altas probabilidades de ser una palabra temática, como sugiere el análisis de textos del capítulo anterior. Se podría argumentar, por tanto, que un criterio de selección basado en la frecuencia absoluta puede resultar muy útil para el alumno *durante* la lectura, pero es cuestionable su rentabilidad *después* de la lectura, ya que no valora la presencia de estas palabras en otros contextos.

Otra opción para maximizar la rentabilidad después de la lectura podría ser la frecuencia relativa, es decir, tomar como referencia el índice de las palabras en un corpus grande como el BNC. De esta manera, se seleccionarían las palabras con mayores probabilidades de aparecer en cualquier contexto. Sin embargo, el problema de pasar por alto las ocurrencias de las palabras en el propio texto es que no se puede garantizar que el alumno vaya a encontrar durante la lectura el número mínimo de exposiciones necesario para la memorización y consolidación.

El reto, por tanto, es trabajar sobre unos datos que, simultáneamente, tengan en cuenta la rentabilidad posterior a la vez que garanticen las exposiciones mínimas y el

reciclaje a intervalos en los textos seleccionados. Nuestra propuesta es utilizar como referencia el número de capítulos distintos en los que aparece una palabra. En primer lugar, dado que los textos son independientes, las palabras que aparecen en capítulos distintos tienen que tener necesariamente cierta flexibilidad contextual. En segundo lugar, el hecho de que sus ocurrencias estén separadas en el texto implica que el alumno encontrará múltiples repeticiones a intervalos, por lo que se optimizarán los procesos de reciclaje. Esto nos permite aprovechar el aprendizaje *durante* la lectura y maximizar las opciones para poder aplicar lo aprendido *después* de la lectura.

Veamos el contraste de los criterios de frecuencia absoluta y relativa en *The Adventures of Sherlock Holmes*. En la Figura III.5 podemos ver un diagrama que compara las palabras que aparecen más veces frente a las que aparecen en más capítulos.

.

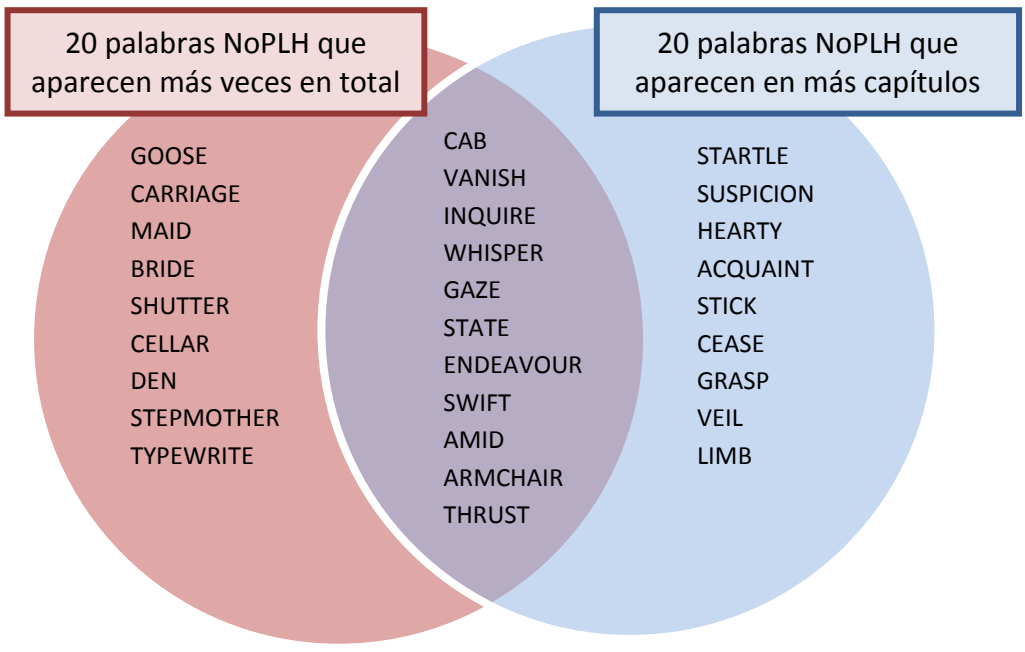


Figura III.5. Palabras no PLH con mayor frecuencia absoluta y relativa en Sherlock Holmes

A continuación, la Tabla III.9 muestra las 20 palabras NoPLH con mayor número de ocurrencias, mientras que la Tabla III.10 muestra las 20 palabras que aparecen en un mayor número de capítulos diferentes. Los datos numéricos que aportan ambas tablas son el número de ocurrencias totales, el número de capítulos distintos y cuántas ocurrencias presenta cada palabra en cada uno de los capítulos.Como es previsible, muchas de las palabras con más ocurrencias también aparecen en mayor número de capítulos diferentes, aunque esto no siempre es así, como ya hemos observado en la Figura III.5.

Un ejemplo interesante lo podemos observar en el término *goose*, la palabra de baja frecuencia que más veces aparece en todo el texto. Con sus 44 ocurrencias, a priori podría considerarse que *goose* es un término clave pero cuando analizamos la Tabla III.9 observamos que 43 de las 44 ocurrencias aparecen en un único capítulo. Esto implica que conocer su significado posiblemente es indispensable para entender el séptimo relato en concreto – es útil *durante* la lectura -, pero es ciertamente cuestionable la utilidad potencial de 'ganso' en el inglés cotidiano. Por el contrario, la última de la lista, *thrust*, tiene muchas menos ocurrencias que *goose* pero vemos que aparece en 7 de los 12 relatos, lo que hace que sea una de las 20 palabras que aparecen en más capítulos diferentes. Este dato parece indicar que *thrust*, 'empujar', es un concepto flexible que puede aparecer en muchos contextos distintos, por lo que al alumno probablemente le resultará útil tanto *durante* la lectura como *después* de haber finalizado el libro.

Casi con toda seguridad el alumno necesitará buscar en un diccionario la palabra *goose* para entender el relato en el que aparece 43 veces, pero son las palabras aplicables al inglés cotidiano las que aparecerán prioritariamente en esta selección léxica.

Tabla III.9. Las 20 palabras NoPLH con más ocurrencias en The Adventures of Sherlock Holmes 8000

		OCURRENCIAS	NÚM. CAP ≠	CAP. 1	CAP. 2	CAP. 3	CAP. 4	CAP. 5	CAP. 6	CAP. 7	CAP. 8	CAP. 9	CAP. 10	CAP. 12
1	GOOSE	44	2	0	1	0	0	0	0	43	0	0	0	0
2	CAB	39	9	8	5	5	3	0	10	2	1	3	1	0
3	CARRIAGE	30	5	11	1	0	7	0	0	0	0	9	1	0
4	MAID	25	6	3	0	0	3	1	0	2	0	2	7	0
5	STEPMOTHER	24	3	0	0	9	0	0	0	0	14	0	0	1
6	VANISH	24	10	1	1	2	2	0	1	1	1	2	6	1
7	INQUIRE	21	8	2	1	0	0	1	4	3	3	1	3	0
8	BRIDE	19	2	1	0	0	0	0	0	0	0	0	9	0
9	WHISPER	18	8	1	3	2	0	0	2	1	3	5	0	1
10	GAZE	16	9	1	2	2	0	0	1	1	3	1	1	3
11	SHUTTER	16	6	1	1	0	0	0	0	1	9	1	0	3
12	STATE	16	7	0	1	1	7	2	0	0	2	2	0	1
13	CELLAR	15	2	0	10	0	0	0	0	0	0	0	0	5
14	ENDEAVOUR	15	8	1	2	0	0	1	0	2	1	2	2	2
15	SWIFT	15	8	2	1	1	2	0	1	0	5	1	1	0
16	AMID	14	7	0	1	0	1	3	1	3	3	0	0	2
17	ARMCHAIR	14	7	4	2	2	0	0	2	2	1	1	0	0
18	DEN	14	3	0	0	0	0	0	12	0	1	1	0	0
19	TYPEWRITE	14	1	0	0	14	0	0	0	0	0	0	0	0
20	THRUST	14	7	0	2	0	0	1	3	1	2	0	2	1

Tabla III.10. Las 20 palabras NoPLH que aparecen en más capítulos de Sherlock Holmes 8000

		OCURRENCIAS	NÚM. CAP ≠	CAP. 1	CAP. 2	CAP. 3	CAP. 4	CAP. 5	CAP. 6	CAP. 7	CAP. 8	CAP. 9	CAP. 10	CAP. 12
1	VANISH	24	10	1	1	2	2	0	1	1	1	2	6	1
2	CAB	39	9	8	5	5	3	0	10	2	1	3	1	0
3	GAZE	16	9	1	2	2	0	0	1	1	3	1	1	3
4	INQUIRE	21	8	2	1	0	0	1	4	3	3	1	3	0
5	WHISPER	18	8	1	3	2	0	0	2	1	3	5	0	1
6	ENDEAVOUR	15	8	1	2	0	0	1	0	2	1	2	2	2
7	SWIFT	15	8	2	1	1	2	0	1	0	5	1	1	0
8	STARTLE	10	8	1	0	1	1	1	1	0	1	0	1	2
9	SUSPICION	10	8	1	0	2	1	1	1	1	2	0	0	1
10	STATE	16	7	0	1	1	7	2	0	0	2	2	0	1
11	AMID	14	7	0	1	0	1	3	1	3	3	0	0	2
12	ARMCHAIR	14	7	4	2	2	0	0	2	2	1	1	0	0
13	THRUST	14	7	0	2	0	0	1	3	1	2	0	2	1
14	HEARTY	13	7	2	0	0	0	0	1	2	1	4	1	1
15	ACQUAINT	12	7	1	1	0	0	0	1	1	2	4	1	0
16	STICK	12	7	2	3	1	2	0	0	1	0	1	0	2
17	CEASE	11	7	1	0	0	0	1	1	2	1	0	1	3
18	GRASP	11	7	2	0	1	1	2	0	0	0	2	1	1
19	VEIL	11	7	1	0	0	0	1	2	0	3	1	1	1
20	LIMB	10	7	1	1	0	2	0	1	0	2	0	1	1

2.5. Selección por flexibilidad contextual

La selección por frecuencia absoluta descarta progresivamente las palabras que aparecen con pocas repeticiones hasta llegar a un número asumible. La selección por flexibilidad contextual sigue un proceso muy similar, pero tomando como referencia el número de capítulos en los que aparece una palabra.

En *The Adventures of Sherlock Holmes* la dispersión de las palabras NoPLH según el número de capítulos en los que aparecen sigue una curva de Zipf, tal y como se aprecia en la Figura III.6. Podemos ver los datos en escala doble logarítmica la Figura III.7.

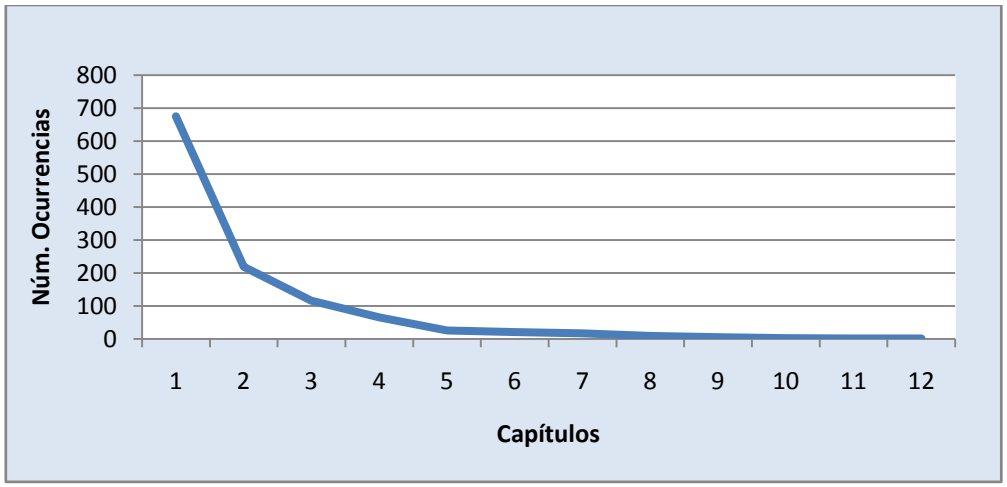


Figura III.6. Distribución de Zipf de Capítulos/Frecuencia en las palabras NoPLH en Sherlock Holmes

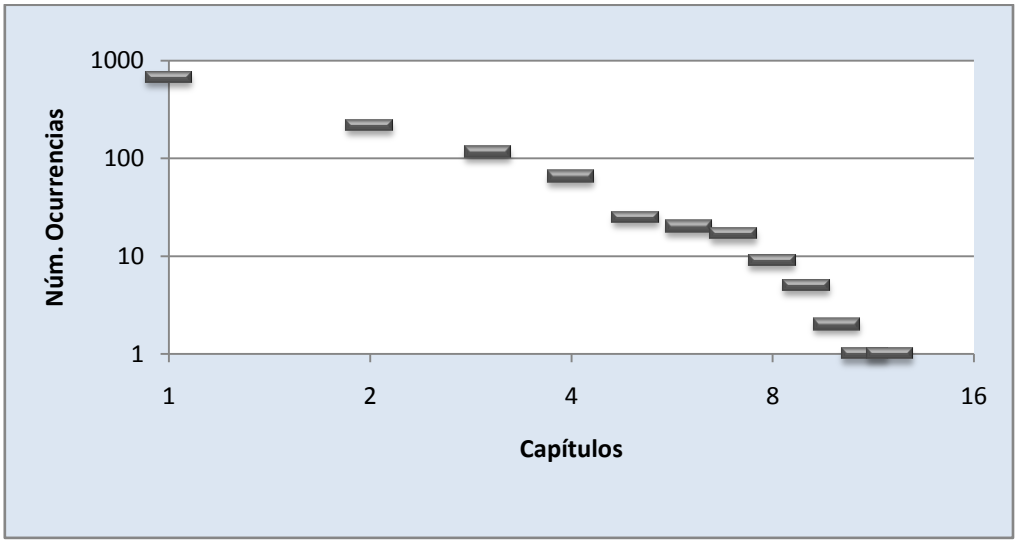


Figura III.7. Representación doble logarítmica de la dispersión Capítulo/Frecuencia en las palabras NoPLH en Sherlock Holmes

Esta distribución de Zipf implica que la carga léxica disminuirá de forma exponencial a medida que aumentemos la exigencia sobre el número de capítulos en los que debe aparecer una palabra para ser seleccionada. Los resultados numéricos se muestran en la Tabla III.11, las columnas a la izquierda señalan cuántas palabras aparecen en un número concreto de capítulos, las de la derecha muestran cuántas palabras aparecen en *al menos* n capítulos. Por ejemplo, hay 20 palabras que aparecen en exactamente 6 capítulos, y 53 palabras que aparecen en 6 o más capítulos.Nos fijaremos en la columna derecha para determinar el número de capítulos en los que debe aparecer una palabra para ser seleccionada. Si se descartan las que solamente aparecen en uno, se obtienen 478 objetivos. Esta cifra desciende a 259 y a 143 si se eliminan las palabras que aparecen en 2 y 3 capítulos, respectivamente. A partir de estos datos podremos escoger el grado de exigencia, que dependerá del la carga léxica que consideremos adecuada para el tipo de material que queremos elaborar. Si nuestro objetivo es proporcionar a los alumnos un glosario de apoyo a la lectura podemos incluir muchos más términos que si pretendemos preparar ejercicios cortos de repaso para el final de cada texto.

Tabla III.11. Número de capítulos en los que aparecen las palabras NoPLH en The Adventures of Sherlock Holmes

Nº CAPÍTULOS	PALABRAS	Nº CAPÍTULOS	PALABRAS
1	675	≥1	1153
2	219	≥2	478
3	116	≥3	259
4	65	≥4	143
5	25	≥5	78
6	20	≥6	53
7	17	≥7	33
8	9	≥8	16
9	5	≥9	7
10	2	≥10	2
11	0	≥11	0
12	0	≥12	0

Como referencia adicional, los datos están desglosados por capítulos en la Tabla III.12. La doble entrada nos permite conocer cuántas palabras aparecen en un capítulo concreto y en al menos *n* capítulos adicionales. Por ejemplo, la fila *Cap1* con la columna ≥*5* nos indica que hay 45 palabras en el Capítulo 1 que aparecen en al menos 5 capítulos diferentes. La primera columna numérica, por tanto, nos indica el total de palabras NoPLH que aparecen por capítulo.

Tabla III.12. Palabras NoPLH que aparecen en cada capítulo concreto y en al menos n capítulos adicionales.

	≥1	≥2	≥3	≥4	≥5	≥6	≥7	≥8	≥9	≥10	≥11	≥12
CAP1	178	140	102	69	45	36	26	13	7	2	0	0
CAP2	202	122	84	64	46	35	20	11	6	2	0	0
CAP3	141	109	78	52	33	24	16	10	5	2	0	0
CAP4	191	121	83	52	35	27	18	8	4	2	0	0
CAP5	169	122	81	55	37	27	19	9	3	0	0	0
CAP6	233	167	115	77	46	37	24	13	6	2	0	0
CAP7	168	123	95	66	43	36	24	12	6	2	0	0
CAP8	225	160	118	84	52	36	26	15	7	2	0	0
CAP9	205	135	97	68	46	34	22	12	6	2	0	0
CAP10	146	96	71	51	34	25	19	11	6	2	0	0
CAP11	174	125	86	58	43	29	21	12	5	1	0	0
CAP12	190	127	99	65	41	30	21	11	4	1	0	0

2.6. Objetivos, primera ocurrencia, reciclaje y aprendizaje acumulado

2.6.1. NoPLH completo

Como señalábamos previamente, dos aspectos fundamentales para la adquisición de vocabulario son el número de exposiciones y la repetición a intervalos. En sección anterior observamos el número de objetivos NoPLH presentes en cada capítulo, pero es importante conocer dónde aparecen por primera vez y cuántas repeticiones tienen. A medida que se avanza en la lectura, aumenta la probabilidad de que los objetivos NoPLH hayan aparecido en capítulos anteriores.

En general, el número de palabras nuevas por capítulo descenderá progresivamente mientras que el número de palabras recicladas aumentará de manera proporcional. La Figura III.8 ilustra esta evolución de las 1.153 palabras NoPLH que aparecen en *The Adventures of Sherlock Holmes*.

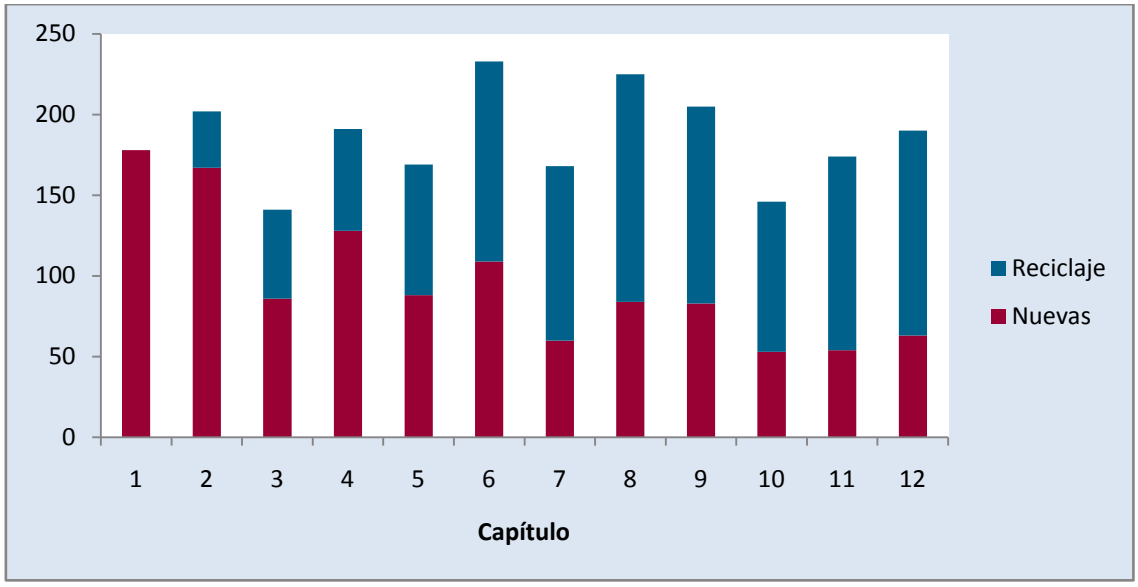


Figura III.8. Aumento progresivo del reciclaje de palabras NoPLH en cada capítulo de Sherlock Holmes

Los datos numéricos correspondientes se pueden consultar en la Tabla III.13. Vemos que en el primer relato hay 178 palabras NoPLH, obviamente todas ellas aparecen por primera vez y no puede haber reciclaje de capítulos anteriores. En la columna «Acumuladas» observamos que estas 178 palabras nuevas representan el 15% del objetivo global. El siguiente capítulo contiene 202 objetivos, de los que 167 aparecen por primera vez y 35 ya aparecieron en el capítulo anterior; es decir, el 83% de los objetivos son palabras nuevas y el 17% son reciclaje. En la última columna vemos que las palabras nuevas junto con las del capítulo anterior suponen el 30% del objetivo global. Previsiblemente, a medida que avanza la lectura, el porcentaje de palabras nuevas es menor y el reciclaje aumenta de manera proporcional. En el último capítulo, solamente uno de cada tres objetivos es nuevo, mientras que los dos tercios restantes se componen de palabras que han aparecido en capítulos anteriores.

Tabla III.13. Objetivos, reciclaje y aprendizaje acumulado de palabras NoPLH en The Adventures of Sherlock Holmes

	OBJETIVOS	1ª OCURRENCIA		RECICLAJE		ACUMULADAS	
CAP. 1	178	178	100%	0	0%	178	15%
CAP. 2	202	167	83%	35	17%	345	30%
CAP. 3	141	86	61%	55	39%	431	37%
CAP. 4	191	128	67%	63	33%	559	48%
CAP. 5	169	88	52%	81	48%	647	56%
CAP. 6	233	109	47%	124	53%	756	66%
CAP. 7	168	60	38%	108	64%	816	71%
CAP. 8	225	84	37%	141	63%	900	78%
CAP. 9	205	83	40%	122	60%	983	85%
CAP. 10	146	53	36%	93	64%	1036	90%
CAP. 11	174	54	31%	120	69%	1090	95%
CAP. 12	190	63	33%	127	67%	1153	100%
TOTAL		1153					

Estos datos describen la distribución de todas las palabras presuntamente desconocidas para un alumno. Como consecuencia de la ley de Zipf, la mitad de las palabras no se reciclarán en el texto. De las restantes, aproximadamente la mitad aparecerán solamente dos veces, y así sucesivamente.

2.6.2. NoPLH con repetición

Los hápax legómenon son palabras de baja rentabilidad durante y después de la lectura, por lo que podemos descartarlas como objetivos de aprendizaje prioritarios. Desde un punto de vista pedagógico, descartar los hápax legómenon optimiza los objetivos de aprendizaje porque nos centramos únicamente en las que tendrán repeticiones y reciclaje a intervalos. La Figura III.9 y la Tabla III.14 muestran los datos relativos a las 478 palabras que aparecen en al menos dos capítulos distintos, es decir, todos nuestros objetivos de aprendizaje tendrán al menos una oportunidad de reciclaje. Observamos que en el primer capítulo el alumno encontrará la primera ocurrencia de 140 palabras que se repetirán en capítulos sucesivos. Por el contrario, en último capítulo el 100% de los objetivos de aprendizaje habrán aparecido en capítulos anteriores.

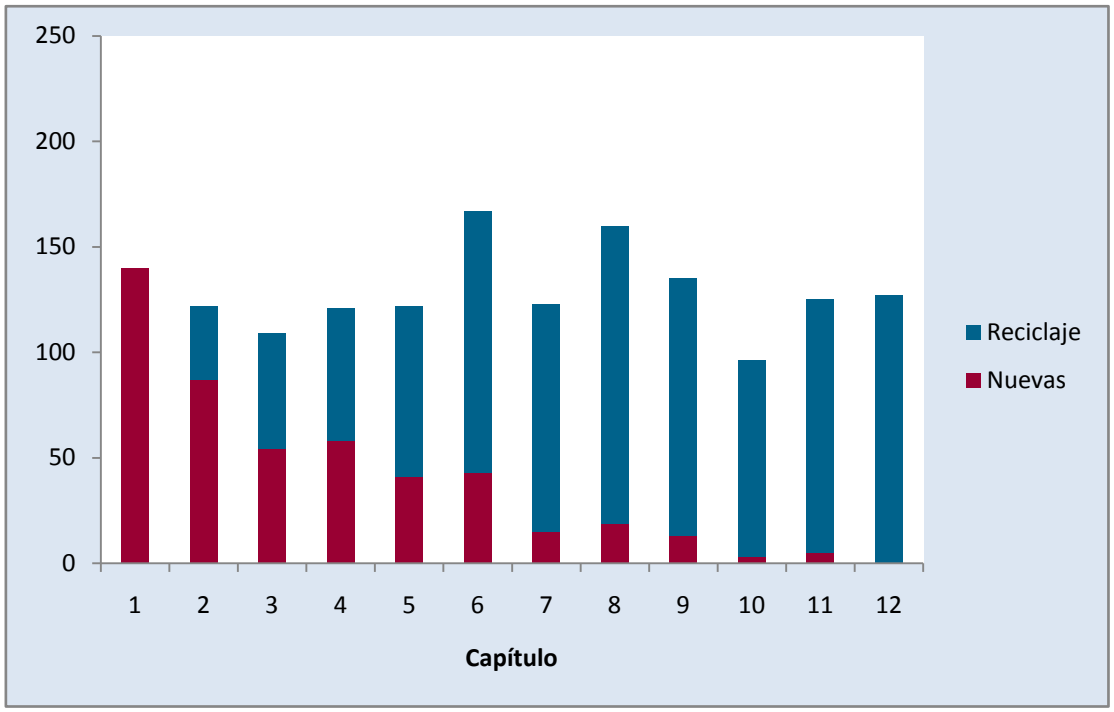


Figura III.9. Aumento progresivo del reciclaje de palabras NoPLH que aparecen en al menos 2 capítulos en Sherlock Holmes

Tabla III.14. Objetivos, reciclaje y aprendizaje acumulado de palabras NoPLH que aparecen en al menos 2 capítulos en Sherlock Holmes

	OBJETIVOS	1ª OCURRENCIA		RECICLAJE		ACUMULADAS	
CAP. 1	140	140	100%	0	0%	140	29%
CAP. 2	122	87	71%	35	29%	227	47%
CAP. 3	109	54	50%	55	50%	281	59%
CAP. 4	121	58	48%	63	52%	339	71%
CAP. 5	122	41	34%	81	66%	380	80%
CAP. 6	167	43	26%	124	74%	423	88%
CAP. 7	123	15	12%	108	88%	438	92%
CAP. 8	160	19	12%	141	88%	457	96%
CAP. 9	135	13	10%	122	90%	470	98%
CAP. 10	96	3	3%	93	97%	473	99%
CAP. 11	125	5	4%	120	96%	478	100%
CAP. 12	127	0	0%	127	100%	478	100%
TOTAL		478					

2.6.3. NoPLH que aparecen en al menos 3 capítulos

Si queremos reducir más el número de objetivos de aprendizaje, repetimos el proceso descartando las que aparecen en solamente dos capítulos. Esto nos da como resultado 259 palabras que aparecen en al menos tres capítulos diferentes.

En la Tabla III.15 se observa que ya en el segundo capítulo casi el 40% de sus objetivos de aprendizaje son reciclajes del capítulo anterior. Hacia la mitad del libro, las palabras recicladas suponen más del 90% de los objetivos y en los tres últimos, el 100% de los objetivos son reciclaje de palabras que han aparecido anteriormente. Todos los términos nuevos que encuentre el alumno en los últimos capítulos son hápax legómenon o palabras que solamente se repiten en dos textos. Llegados a este punto, todos los objetivos aparecen en al menos tres capítulos diferentes, lo que nos garantiza el reciclaje a intervalos de cada uno de ellos.

Tabla III.15. Objetivos, reciclaje y aprendizaje acumulado de palabras NoPLH que aparecen en al menos 3 capítulos en The Adventures of Sherlock Holmes

	OBJETIVOS	1ª OCURRENCIA		RECICLAJE		ACUMULADAS	
CAP. 1	102	102	100%	0	0%	102	21%
CAP. 2	84	51	61%	33	39%	153	32%
CAP. 3	78	29	37%	49	63%	182	38%
CAP. 4	83	29	35%	54	65%	211	44%
CAP. 5	81	14	17%	67	83%	225	47%
CAP. 6	115	21	18%	94	82%	246	51%
CAP. 7	95	7	7%	88	93%	253	53%
CAP. 8	118	4	3%	114	97%	257	54%
CAP. 9	97	2	2%	95	98%	259	54%
CAP. 10	71	0	0%	71	100%	259	54%
CAP. 11	86	0	0%	86	100%	259	54%
CAP. 12	99	0	0%	99	100%	259	54%
TOTAL		259					

Respecto del número de exposiciones necesario, la Tabla III.16 nos muestra la tasa de repetición a medida que se avanza en la lectura. Los datos son acumulativos y se refieren a ocurrencias tanto en el propio capítulo como en anteriores. Por ejemplo, en el capítulo 8 hay 12 palabras con >10 ocurrencias, esto significa que al final del octavo capítulo el alumno habrá encontrado 12 palabras que han aparecido más de 10 veces en

distintos puntos entre el primer y el octavo capítulo. Los datos revelan que la tasa de exposición es razonablemente buena: prácticamente un tercio de los objetivos aparecen por lo menos seis veces y casi el 40% de ellos tiene más de 10 ocurrencias.

Tabla III.16. Palabras que presentan *n* ocurrencias del vocabulario NoPLH que aparece en al menos 3 capítulos de Sherlock Holmes

	0 OCUR.	1-2 OCUR.	3-5 OCUR.	6-10 OCUR.	> 10 OCUR.
CAP. 1	157	94	6	1	1
CAP. 2	106	131	18	2	2
CAP. 3	77	134	40	6	2
CAP. 4	48	140	53	16	2
CAP. 5	34	135	69	18	3
CAP. 6	13	116	93	33	4
CAP. 7	6	103	104	40	6
CAP. 8	2	82	109	54	12
CAP. 9	0	61	122	58	18
CAP. 10	0	48	125	64	22
CAP. 11	0	26	133	77	23
CAP. 12	0	0	147	81	31

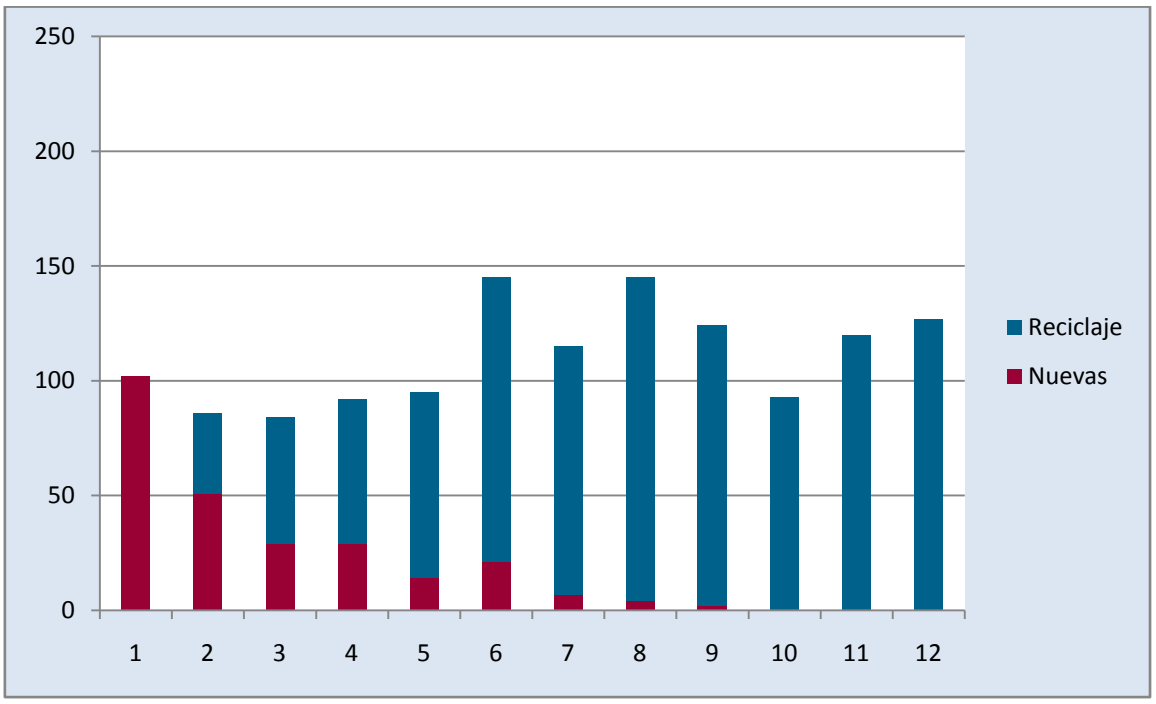


Figura III.10. Aumento progresivo del reciclaje de palabras NoPLH que aparecen en al menos 3 capítulos en Sherlock Holmes

2.7. Conclusiones del criterio de selección combinado

La adquisición de vocabulario por medio de la lectura extensa puede —y, en nuestra opinión, debe— tener una planificación de los objetivos prioritarios de aprendizaje. Incluso para una cobertura entre el 95% y el 98%, el número de palabras desconocidas en un plan extenso de lectura puede ser demasiado elevado como para pretender que el alumno procese y recuerde todas ellas. Es necesario, por tanto, establecer unos criterios de selección que valore tanto el apoyo durante la lectura como la rentabilidad que tendrán estas palabras posteriormente.

Una selección basada únicamente en la frecuencia en el texto puede dar lugar a unos objetivos de aprendizaje muy específicos, presentes únicamente en contextos muy restringidos. Por el contrario, utilizar como criterio la frecuencia relativa a un corpus grande no garantiza el número de exposiciones ni el repaso a intervalos durante la lectura. La idea propuesta en este trabajo es valorar el número de capítulos en los que aparecen las palabras e ir eliminando las de menor flexibilidad hasta llegar a un número de objetivos de aprendizaje adecuado para nuestro propósito.

Hemos escogido una obra que contiene distintos relatos para explicar cómo replicar el proceso con una colección de textos seleccionados como plan de lectura para un curso completo. Los datos relativos al análisis de *The Adventures of Sherlock Holmes* demuestran que este criterio permite rebajar considerablemente el número de objetivos de aprendizaje de 1.153 a 249, seleccionando aquellos que le serán útiles al alumno durante la lectura y, al mismo tiempo, le permitirán aplicar lo aprendido en situaciones comunicativas posteriores.

A medida que hemos reducido el número de objetivos se ha optimizado la tasa de repetición y la repetición a intervalos. Al eliminar las palabras que solamente aparecen en uno o dos capítulos, garantizamos que cada objetivo tenga un mínimo de 3 exposiciones, y casi un tercio de ellos supere las 6 exposiciones. En los primeros capítulos el alumno encontrará las primeras ocurrencias de las palabras y muy poco reciclaje. La tendencia varía progresivamente de tal manera que los últimos estarán enfocados principalmente al repaso y consolidación de los objetivos que han aparecido en capítulos anteriores.

Una vez hecha la selección, el docente puede crear materiales de apoyo tales como glosarios, *flashcards*, diccionarios de imágenes y ejercicios de pre y post-lectura de cada texto, por nombrar algunos. Debe tener en cuenta que los primeros capítulos son más exigentes para el alumno por una mayor presencia de vocabulario nuevo, mientras que los últimos estarán destinados a afianzar lo aprendido anteriormente. La Figura III.11 ilustra que la evolución en el reciclaje a medida que se avanza en el libro alcanza niveles óptimos, esta estrategia basada en exposiciones y repeticiones a intervalos está diseñada para consolidar los objetivos de aprendizaje planteados.

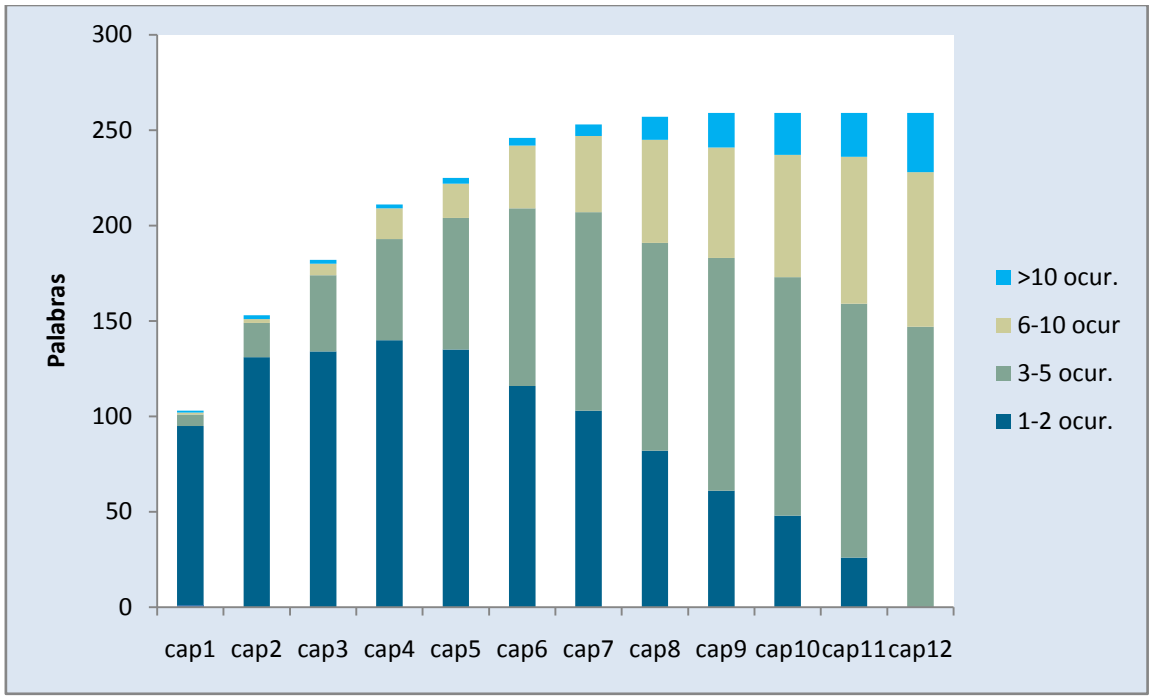


Figura III.11. Tasa de repetición y ocurrencias del vocabulario NoPLH que aparecen en al menos 3 capítulos de Sherlock Holmes

CUARTA PARTE

CONCLUSIONES Y PROSPECTIVA

IV

1. CONCLUSIONES

El objetivo de esta tesis era definir un marco de trabajo para la enseñanza-aprendizaje de vocabulario receptivo basado en la rentabilidad de las palabras y el aprovechamiento del conocimiento potencial que aporta la similitud léxica del español y el inglés.

El marco se compone de una primera fase de aprendizaje explícito del vocabulario prioritario y una segunda fase de aprendizaje incidental mediante la adquisición por contexto. Para la primera, se proporcionan los objetivos de aprendizaje y, para la segunda, un sistema de selección del vocabulario para cualquier plan de lectura extensa. El marco incluye, asimismo, una propuesta de pruebas de evaluación del alcance de la base léxica del alumno. A continuación se resumen las principales aportaciones de los resultados este trabajo tanto para la aplicación docente como para la investigación.

1.1. Principales aportaciones a la docencia

1.1.1. Lista de vocabulario prioritario

a) *Objeto de estudio.*

Uno de los objetivos específicos de esta tesis era determinar las palabras que deberían aprenderse prioritariamente para alcanzar la base léxica necesaria que permite leer textos auténticos.

b) *Metodología*.

Estas unidades léxicas se seleccionaron en función de su rango, frecuencia y similitud ortográfica con sus equivalentes en español. La selección se hizo mediante análisis de corpus, léxico-estadística y un sistema de clasificación de cognados en función de su similitud ortográfica. Posteriormente verificamos la validez de las palabras seleccionadas midiendo su alcance en distintos géneros de interés para el aprendiz de una lengua extranjera, tales como exámenes de certificación, novelas originales y adaptadas, subtítulos de películas y discurso formal.

c) *Hallazgos más significativos*.

Determinamos que para el alumno hispanohablante, 1.800 palabras inglesas y una serie de afijos derivativos son suficientes para alcanzar el 95% del vocabulario de un texto medio (umbral de autonomía). Únicamente hubo dos textos de gran complejidad léxica en los que no se alcanzó ese umbral, pero en ambos se superó el 94,5%. Sobre el corpus global se obtuvo una cobertura superior al 96%, alcanzando el 98% (umbral de garantía) en la mayoría de los exámenes de nivel. Partíamos de un dogma pedagógico muy extendido que considera que la L1 debe aislarse tanto en la enseñanza de L2 como en la planificación de objetivos de aprendizaje.

Los resultados de este estudio, sin embargo, han demostrado que valorar la transferencia positiva de la L1 permite elaborar una programación docente eficaz y asequible que aprovecha el conocimiento potencial. Esto reduce notablemente el tiempo y el esfuerzo del alumno respecto de otras programaciones basadas únicamente en la frecuencia de las palabras. La enseñanza explícita de estos

objetivos de aprendizaje es la manera más rápida de proporcionar a un alumno hispanohablante la base léxica necesaria para leer una amplia gama de textos auténticos en inglés sin que el vocabulario desconocido suponga un problema para la comprensión.

d) *Aplicación directa en la enseñanza.*

Este estudio aporta un listado de 1.800 palabras prioritarias dividido en cuatro niveles de dificultad que los docentes pueden utilizar tanto para elaborar una planificación léxica desde cero como para complementar una programación existente. Es importante remarcar que, si bien este marco apuesta por el aprendizaje directo del vocabulario prioritario, esto no implica que se recomiende proporcionar a los alumnos una lista de palabras descontextualizadas para que las memoricen. Al contrario, la lista pretende servir como base sobre la que elaborar actividades y materiales docentes encaminados a una meta —leer textos auténticos— que se puede alcanzar a través de distintas metodologías y técnicas de enseñanza de vocabulario.

1.1.2. Evaluación del vocabulario receptivo

a) *Objeto de estudio.*

La aplicación de la lista de vocabulario prioritario en la docencia requiere evaluar el alcance de las capacidades y competencias. Es necesario contar con un sistema fiable para medir el vocabulario del alumno tanto al inicio del curso para determinar los conocimientos previos, como a lo largo de todo el proceso de aprendizaje. El objetivo de la evaluación será diagnosticar el grado de éxito alcanzado con el fin de ajustar los contenidos pedagógicos a sus necesidades.

b) *Metodología.*

En primer lugar, estudiamos los distintos tests de reconocido prestigio y seleccionamos únicamente aquellos que ofrecían las garantías necesarias para evaluar la base léxica de un alumno hispanohablante que aprende —o va a

aprender— la lista de vocabulario prioritario. Se valoraron aspectos como la validación por marcos psicométricos, que el diseño de las preguntas fuera adecuado para fines docentes y que los test estuvieran enfocados al vocabulario receptivo en L2 minimizando otras inferencias. Finalmente seleccionamos tres únicas pruebas: *The Eurocentres Test*, *The Vocabulary Levels Test* y *The Vocabulary Size Test*. En cada uno de ellos estudiamos si el hispanohablante tendría ciertas ventajas analizando cuántas de las palabras escogidas para las preguntas podrían ser deducibles. También se valoraron las posibles desventajas derivadas de la transferencia negativa de la L1.

c) *Hallazgos más significativos.*

El resultado del análisis indicó que el español como L1 puede invalidar los resultados de las tres pruebas. *The Eurocentres Test* incluye como método de control una serie de palabras artificiales. Algunas de ellas están fabricadas sobre raíces latinas y son plausibles para el alumno hispanohablante, lo que le supone una desventaja considerable.

Ocurre el efecto contrario tanto en el VLT como en el VST: debido a la presencia de palabras de raíz latina en los registros cultos, un alumno hispanohablante obtendría unos resultados injustamente altos. Con una base mínima de inglés, podría obtener mayor puntuación en las palabras a partir del rango 5.000 que en las 1.000 más frecuentes. Concretamente, si se tiene una rica base léxica en L1, en el último rango del VST —que es superior al vocabulario de un nativo de 12 años— se podrían deducir por semejanza el 70% de las preguntas.

d) *Aplicación directa en la enseñanza.*

Si bien los resultados ofrecen un perfil atípico al alumno hispanohablante, el diseño y formato de los test es óptimo para medir vocabulario receptivo en L2. Por tanto, no se descarta su uso, pero se sugiere adaptarlos en función de la L1 del alumno. El docente podría modificar el *Eurocentres test* eliminando las palabras artificiales de raíz latina o bien reemplazarlas por otras pseudopalabras plausibles. Sin embargo, recomendamos en su lugar la adaptación del VLT o el VST. La opción más sencilla para adaptar estos test es puntuar de manera independiente los

cognados. Esto nos permite evaluar tanto el vocabulario general como la capacidad de reconocer cognados. Otra opción es reemplazar los cognados por palabras de su mismo rango de frecuencia si se quiere medir el alcance global, o bien utilizar el formato con los objetivos de aprendizaje de la lista si se desea hacer un seguimiento de la enseñanza explícita.

1.1.3. Selección léxica en planes de lectura

a) *Objeto de estudio.*

Tras aprender los objetivos de aprendizaje planteados, el alumno seguirá enriqueciendo su vocabulario en L2 por exposición, fundamentalmente a través de la lectura. El número de palabras desconocidas en un plan de lectura extensa es excesivo, por lo que es preciso seleccionar las unidades más importantes para incidir sobre ellas. El objetivo, por tanto, era establecer una sistematización matemática para determinar las palabras más relevantes en un plan de lectura extensa para un curso dado, con el fin de crear materiales de apoyo a la lectura.

b) *Metodología.*

Basándonos en el estudio de la función exponencial y logarítmica que determina la distribución del lenguaje, establecimos un sistema para reducir el número de objetivos de aprendizaje de manera progresiva hasta llegar a un número asumible. El sistema tenía en cuenta dos factores fundamentales: por un lado, un modelo probabilístico acerca de la rentabilidad de las palabras basado en los contextos distintos en los que puede aparecer; por otro lado, criterios psicolingüísticos y cognitivos sobre la necesidad de la revisión a intervalos y el número de repeticiones necesario para la memorización del vocabulario. Se analizó un caso práctico para ilustrar este proceso de selección léxica en un plan de lectura.

c) *Hallazgos más significativos.*

Utilizando esta sistematización sobre un libro de 300 páginas que contiene más de 1.000 unidades léxicas de baja frecuencia, pudimos seleccionar una serie

palabras que cumplían tres características: 1) el número de objetivos de aprendizaje era asumible, (2) aparecían en distintos contextos, de donde se infiere que el alumno puede aplicar posteriormente lo aprendido, y (3) seguían las recomendaciones pedagógicas sobre la revisión a intervalos y las repeticiones necesarias para la memorización.

d) *Aplicación directa en la enseñanza.*

El método propuesto para la selección de vocabulario en lectura extensa es muy riguroso, pero puede resultar extremadamente complejo llevarlo a cabo, pues no hay un programa informático que lo haga de forma automática. Para el docente no familiarizado con las herramientas de lingüística computacional este sistema no es,

de momento, una opción viable de aplicación directa. Sin embargo, los resultados ponen de manifiesto la importancia de formar al profesorado en técnicas de selección léxica con el fin de poder guiar a los alumnos para sacar el máximo provecho de la lectura. Por otro lado, el sistema puede sentar las bases sobre las que las editoriales comerciales creen material de apoyo a la lectura.

1.2. Principales aportaciones a la investigación

· La investigación previa ha elaborado listas de frecuencia que únicamente tienen en cuenta la meta. Esta investigación ofrece una lista de frecuencia que valora el punto de partida, es decir, el conocimiento potencial derivado de la L1.

· Los estudios previos demuestran que la transferencia positiva es facilitadora del aprendizaje, pero no plantean una aplicación práctica sólida. Este trabajo ofrece una demostración empírica de la eficacia de utilizar la transferencia positiva en la programación de contenidos docentes.

· Los estudios de corpus sobre la presencia de cognados se han limitado a una visión teórica. Esta investigación estudia los cognados desde el punto de vista pragmático, es decir, aquellos que serán deducibles por semejanza en los distintos rangos de frecuencia, así como sus implicaciones docentes.

· No hay, hasta donde sabemos, otras propuestas de sistematización matemática para la selección léxica en colecciones de textos no relacionados entre sí.

· El estudio aporta un análisis interdisciplinar que aplica las leyes matemáticas y el estudio de funciones logarítmicas y exponenciales a la enseñanza de un idioma.

2. LIMITACIONES DEL ESTUDIO

a) Relativas al enfoque

· Conocer el significado de la mayoría de las palabras de un texto, incluso de todas ellas, no garantiza al 100% la lectura comprensiva. Las conclusiones de este estudio pueden no ser aplicables a textos con numerosas referencias culturales o intrínsecamente difíciles como un poema o un manual técnico.

· El análisis se centra en unidades léxicas simples. Es necesario explorar el comportamiento de las unidades complejas formadas por varias palabras.

b) Relativas a la metodología

· Es un estudio empírico de corpus basado en léxico-estadística cuyas conclusiones son matemáticamente válidas sobre un corpus determinado. Sin embargo, su extrapolación a unos efectos esperados en la docencia se ha fundamentado en la investigación reciente, el presente estudio no aporta datos experimentales directos.

· El diseño está basado en modelos probabilísticos que se proyectan desde lo específico a lo general. El estudio de la probabilidad no es una ciencia exacta, nunca se puede tener una certeza del 100% sobre unos resultados basados en inferencia estadística.

3. TRABAJO FUTURO

Las dos líneas de investigación prioritarias para complementar las conclusiones de esta tesis son las siguientes:

· Este trabajo sienta unas bases empíricas, el siguiente paso es una investigación experimental que aporte evidencia a las conclusiones obtenidas en este estudio.

· Es necesario investigar cómo afectan a la comprensión lectora las formas léxicas complejas, locuciones y expresiones fijas en las que las palabras que las componen han perdido su autonomía semántica.

Si bien este estudio ha arrojado luz sobre algunas cuestiones relevantes, también ha abierto otros interrogantes. Otras líneas de investigación que surgen a raíz de los resultados de este trabajo son las siguientes:

· Estudio en profundidad de las distintas unidades de la lista de vocabulario prioritario: acepciones, colocaciones frecuentes, restricciones, traducción al español, etc.

· Diseño y validación de pruebas que midan el vocabulario receptivo en inglés adaptadas para hispanohablantes. Diseño de un test específico sobre la lista de vocabulario prioritario.

· Exploración de la validez de los resultados en otras lenguas romances como L1 y L2.

· Desarrollo de software capaz de procesar automáticamente el sistema de selección léxica en planes de lectura.

· Reconocimiento de cognados de alta frecuencia y técnicas docentes para potenciar y mejorar esta habilidad.

· Alcance del vocabulario productivo tras aprender la lista de palabras prioritarias, así como su proceso en comprensión oral.

· Presencia de vocabulario prioritario y cognados evidentes en los libros de texto de referencia utilizados en las distintas etapas educativas en España.

REFERENCIAS

Albrechtsen, D., Haastrup, K., & Henriksen, B. (2008). *Vocabulary and writing in a first and second language: Processes and development.* New York: Palgrave Macmillan.

Alderson, J. C. (2006). *Diagnosing foreign language proficiency: The interface between learning and assessment.* Londres: Continuum International Publishing Group.

Alexopoulou, A. (2005). La función de la interlengua en el aprendizaje de lenguas extranjeras. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas.*, *9*, 84-99.

Anders, V. (2011). Etimología de Cátedra. *Etimologías de Chile.* Recuperado de: http://etimologias.dechile.net/?ca.tedra.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International journal of Lexicography*, *6*(4), 253-279.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118.

Bertocchini, P., Costanzo, E., & Puren, C. (1998). *Se former en didactique des langues*. Paris: Ellipses.

Bravo, M. A., Hiebert, E. H., & Pearson, P. D. (2007). Tapping the Linguistic Resources of Spanish–English Bilinguals. The role of cognates in science. En R. K. Wagner, A. E. Muse & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension.* (pp. 140-156). New York: The Guilford Press.

British National Corpus. (2009). The BNC in numbers. *BNC website.* Recuperado de: http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=numbers.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie), 58*(5), 412-424.

Buehler, S. D. (1995). *Transfer in the interlanguage of native English speakers in first-year college Spanish.* Tesis de Máster, Rice University. doi:1911/13932

Carlisle, J. F., & Feldman, L. (1995). Morphological awareness and early reading achievement. *Morphological aspects of language processing.* (pp. 189-209)

Carroll, S. E. (1992). On cognates. *Second Language Research, 8*(2), 93-119.

Carter, R., McCarthy, M., & Channell, J. (1988). *Vocabulary and language teaching*. Londres: Longman.

Cembreros, D. (2011). *Predicting Readability for ESL Students: A Software Design/Method*. Saarbrücken: Lambert Academic Publications.

Cervero, M. J., & Pichardo Castro, F. (2000). *Aprender y enseñar vocabulario*. Madrid: Edelsa.

Coady, J. (1997). L2 vocabulary acquisition through extensive reading. En J. Coady, & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy*. (pp. 225-237). Cambridge: Cambridge University Press.

Coady, J., & Huckin, T. (1997). *Second language vocabulary acquisition: A rationale for pedagogy*. Cambridge: Cambridge University Press.

Coxhead, A. (1998). An academic word list. *English Language Institute Occasional Publications,18.* Victoria University of Wellington.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly, 34*(2), 213-238.

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue française de linguistique appliquée, 12*(2), 65-78.

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The modern language journal, 80*(2), 183-198.

Chung, M. (2009). The newspaper word list: A specialised vocabulary for reading newspapers. *JALT journal, 31*(2), 159-182.

Day, R. R., Omura, C., & Hiramatsu, M. (1992). Incidental EFL vocabulary learning and reading. *Reading in a foreign language, 7*, 541-541.

Douglas-Fairhurst, R. (Ed.). (2006). *A Christmas Carol and other Christmas Books.* . Oxford: Oxford University Press.

Dressler, C., Carlo, M. S., Snow, C., August, D., & White, C. (2011). Spanish-speaking students' use of cognate knowledge to infer the meaning of English words. *Bilingualism: Language and Cognition, 14*(2), 243-255.

Durán Escribano, P. (2004). Exploring cognition processes in second language acquisition: the case of cognates and false-friends in EST. *Ibérica, 7*, 87-106.

Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor, MI: University of Michigan Press.

Friel, B. M., & Kennison, S. M. (2001). Identifying German–English cognates, false cognates, and non-cognates: Methodological issues and descriptive norms. *Bilingualism: Language and Cognition, 4*(3), 249-274.

Gairns, R., & Redman, S. (1986). *Working with words: A guide to teaching and learning vocabulary*. Cambridge: Cambridge University Press.

Garrote, M. (2010). *Los corpus de habla infantil. Metodología y análisis.* Madrid: Servicio de publicaciones de la Universidad Autónoma de Madrid.

Haastrup, K. (1991). *Lexical inferencing procedures, or, talking about words: receptive procedures in foreign language learning with special reference to English*. Tübingen: Gunter Narr Verlag.

Hafiz, F. M., & Tudor, I. (1990). Graded readers as an input medium in L2 learning. *System, 18*(1), 31-42.

Hall, C. J. (2002). The automatic cognate form assumption: Evidence for the parasitic model of vocabulary development. *IRAL, 40*(2), 69-88.

Hancin-Bhatt, B., & Nagy, W. (1994). Lexical transfer and second language morphological development. *Applied Psycholinguistics, 15*, 289-310.

Hernández, L. P. (1997). A linguistic approach to the erotism of Lady Chatterly's lover. *Cuadernos de investigación filológica, 23*, 213-231.

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a foreign language, 8*(2), 689.

Hsueh-Chao, M. H., & Nation, I. S. P. (2000). Unknown Vocabulary Density and Reading Comprehension. *Reading in a foreign language, 13*(1), 403-430.

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language. *Studies in second language acquisition, 21*(2), 181-193.

Huckin, T., Haynes, M., & Coady, J. (1995). *Second Language Reading and Vocabulary Learning.* Norwood: Ablex Publishing Corporation.

Instituto Cervantes. (1997a). Diccionario de términos esenciales de ELE. *Centro virtual Cervantes,* recuperado de: *http://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/*

Instituto Cervantes. (1997b). Niveles comunes de referencia. *Marco común europeo de referencia*, recuperado de: http://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cap_03_01.htm.

Izquierdo, M. C. (2003). *La selección de léxico en la enseñanza del español como lengua extranjera. Su aplicación al nivel elemental en estudiantes francófonos*. Tesis doctoral, Universitat de València. doi:10803/9815

Kellerman, E. (1984). The empirical evidence for the influence of the L1 in interlanguage. *Interlanguage* (pp. 98-122). Edimburgo: Edinburgh University Press.

Kennedy, G. (1998). *An introduction to corpus linguistics*. Harlow: Addison Wesley Longman Ltd.

Kieffer, M. J., & Lesaux, N. K. (2007). Breaking down words to build meaning: Morphology, vocabulary, and reading comprehension in the urban classroom. *The reading teacher, 61*(2), 134-144.

Kilgarriff, A. (1995). Lemmatised list. *BNC database and word frequency lists*. Disponible en: http://www.kilgarriff.co.uk/BNClists/lemma.al.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics, 29*(3), 333-347.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach.* Cambridge: Cambridge University Press.

Konstantakis, N. (2007). Creating a business word for teaching Business English. *Elia: Estudios de lingüística inglesa aplicada,* (7), 79-102.

Krashen, S. D. (1993). *The power of reading: Insights from the research.* Englewood: Cliff Libraries Unlimited.

Kroll, J. F., & Sunderman, G. (2003). Cognitive processes in second language learners and bilinguals: The development of lexical and conceptual representations. En C. J. Doughty, & M. H. Long (Eds.), *The Handbook of Second Language Acquisition.* (pp. 104-129). Oxford: Blackwell Malden.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth: Dartmouth Publishing Group.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren, & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines.* (pp. 316-323). Clevedon: Multilingual Matters.

Laufer, B. (1990). Words you know: How they affect the words you learn. En J. Fisiak (Ed.), *Further insights into contrastive linguistics* (pp. 573-593). Amsterdam: Benjamins.

Laufer, B. (1992). How much lexis is necessary for reading comprehension**.**. En H. Bejoint, & P. Arnaud (Eds.), *Vocabulary and applied linguistics.* (pp. 16-323). Londres: Macillan.

Laufer, B. (1997). The lexical plight in second language reading: words you don't know, words you think you know and words you can't guess. En J. Coady, & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy*. (pp. 20-52). Cambridge: Cambridge University Press.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review/La revue canadienne des langues vivantes, 59*(4), 567-587.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied linguistics, 22*(1), 1-26.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language testing, 16*(1), 33-51.

Laufer, B., & Osimo, H. (1991). Facilitating Long-Term Retention of Vocabulary: The Second-Hand Cloze. *System, 19*(3)

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a foreign language, 22*(1), 15-30.

Lawrence, A. (2013). *AntWordProfiler (Version 1.4.0.1)*. [Software]. Tokyo: Waseda University. Disponible en: http://www.antlab.sci.waseda.ac.jp.

Lee, S. H. (2003). ESL learners' vocabulary use in writing and the effects of explicit vocabulary instruction. *System, 31*(4), 537-561.

Leech, G. N., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English.* Londres: Longman.

Lessard-Clouston, M. (2010). Theology lectures as lexical environments: A case study of technical vocabulary use. *Journal of English for Academic Purposes, 9*(4), 308-321.

Lessard-Clouston, M. (2012). Word lists for vocabulary learning. *The CATESOL Journal, 24*(1), 287-304.

López Morales, H. (1978). Frecuencia léxica, disponibilidad y programación curricular. *Boletín de la Academia Puertorriqueña de la Lengua Española, 6*(1), 73-86.

Lorenzo Salazar, D. (2011). *Lexical bundles in scientific English: A corpus-based study of native and non-native writing.* Tesis doctoral, Universitat de Barcelona. doi:10803/52083

Lubliner, S., & Grisham, D. L. (2012). Cognate Strategy Instruction. Providing powerful literacy tools to Spanish-speaking students. En J. Fingon, & S. Ulanoff (Eds.), *Learning from Culturally and Linguistically Diverse Classrooms: Using Inquiry to Inform Practice* (pp. 105-123). New York: Teachers College Press.

Llisterri, J. (2013). *Textos orales y corpus de lengua oral* Recuperado de: http://liceu.uab.cat/~joaquim/language_resources/spoken_res/Texto_oral.html.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* (6ª ed.). Boston: MIT Press.

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*(3), 183-198.

McArthur, T. (1981). *Longman lexicon of contemporary English*. Londres: Longman.

McCrostie, J. (2007). Investigating the accuracy of teachers' word frequency intuitions. *RELC journal, 38*(1), 53-66.

Means, T. (2003). *Instant Spanish: Vocabulary Builder*. New York: Hippocrene Books.

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching, 13*(3-4), 221-246. doi:10.1017/S0261444800008879

Meara, P. (1996). The dimensions of lexical competence. En G. Brown, K. Malmkjaer & J. Williams (Eds.), *Performance and Competence in Second Language Acquisition.* (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P., & Jones, G. (1990). *Eurocentres vocabulary size test.* Zurich: Eurocentres Learning Service.

Méndez Pérez, A., Peña, E. D., & Bedore, L. M. (2010). Cognates facilitate word recognition in young Spanish-English bilinguals' test performance. *Early childhood services (San Diego, California), 4*(1), 55.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Montelongo, J. A., Hernández, A. C., & Herter, R. J. (2009). *Transparency ratings for Spanish–English cognate words*. Recuperado de: http://works.bepress.com/jmontelo/3/.

Morán, R. (2010). *The cognate highlighter.* [Aplicación Web] Disponible en: http://www.cognates.org/h/highlight.html.

Morán, R. (2011a). *Cognate Linguistics* [Kindle Edition]. Recuperado de Amazon.com.

Morán, R. (2011b). *The Dictionary of Cognates.* [Kindle Edition].Recuperado de Amazon.com.

Morin, R., & Goebel Jr, J. (2001). Basic vocabulary instruction: Teaching strategies or teaching words? *Foreign Language Annals, 34*(1), 8-17.

Nagy, W., Berninger, V. W., & Abbott, R. D. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of educational psychology, 98*(1), 134.

Nagy, W., García, G. E., Durgunoğlu, A. Y., & Hancin-Bhatt, B. (1993). Spanish-English bilingual students' use of cognates in English reading. *Journal of Literacy Research, 25*(3), 241-259.

Nagy, W., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading research quarterly, 20*(2), 233-253. doi:10.2307/747758

Nash, R. (1997). *NTC's dictionary of Spanish cognates: Thematically organized*. Lincolnwood, IL: NTC Publishing Group.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2006a). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review/La revue canadienne des langues vivantes, 63*(1), 59-82.

Nation, I. S. P. (2006b). Teaching vocabulary. En P. Robertson, & R. Nunn (Eds.), *The Study of Second Language Acquisition in the Asian context.* (pp. 329-340). Seoul: Asian EFL Journal Press.

Nation, I. S. P. (2012). *Information on the BNC-COCA word family lists*. Recuperado de: http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf.

Nation, I. S. P. (s.f.). *About Mid-frequency Readers.* Recuperado de: http://www.victoria.ac.nz/lals/about/staff/publications/About-Mid-frequency-readers-2.pdf.

Nation, I. S. P., Coxhead, A., & Heatley, A. (2002). *RANGE and FREQUENCY*. [Software]. Disponible en: http://www.victoria.ac.nz/lals/about/staff/publications/BNC_COCA_25000.zip.

Nation, I. S. P., & Chung, T. (2009). Teaching and testing vocabulary. En M. H. Long, & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 543-559). Oxford: Wiley Blackwell. doi:10.1002/9781444315783.ch28

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. En N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). *Researching and Analyzing Vocabulary*. Boston: Heinle Cengage Learning.

Newman, J., Baayen, R. H., & Rice, S. (2011). *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Amsterdam: Rodopi.

O'Dell, F. (1997). Incorporating vocabulary into the syllabus. En N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy.* (pp. 258-278). Cambridge: Cambridge University Press.

Obama, B. H. (2008). *Transcription of Barack's Obama New Hampshire Primary Speech.* Recuperado de: http://www.nytimes.com/2008/01/08/us/politics/08text-obama.html.

Obama, B. H. (2009). *Inaugural speech.* Recuperado de: http://www.whitehouse.gov/blog/inaugural-address.

Obama, B. H. (2013). *Inaugural speech.* Recuperado de: http://www.whitehouse.gov/the-press-office/2013/01/21/inaugural-address-president-barack-obama.

Ogden, C. K., & Graham, E. (Eds.). (1930). *Basic English.* (ed. rev). Londres: Paul Kreber & Co. Ltd.

Oxford, R. L., & Scarcella, R. C. (1994). Second language vocabulary learning among adults: State of the art in vocabulary instruction. *System, 22*(2), 231-243.

Paribakht, T. S., & Wesche, M. B. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. En J. Coady, & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 174-200). Cambridge: Cambridge University Press.

Perc, M. (2012). Evolution of the most common English words and phrases over the centuries. *Journal of The Royal Society Interface, 9*(77), 3323-3328.

Pérez Basanta, C. (1999). La enseñanza del vocabulario desde una perspectiva lingüística y pedagógica. En M. S. Salaberri (Ed.), *Lingüística aplicada a la enseñanza de lenguas extranjeras* (pp. 262-307). Almería: Publicaciones de la Universidad de Almería.

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language, 18*(1), 1-28.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems, 14*(3), 130-137.

Prado, M. (2001). *Diccionario de falsos amigos. Inglés-español.* Madrid: Gredos.

Proctor, C. P., Carlo, M., August, D., & Snow, C. (2005). Native Spanish-Speaking Children Reading in English: Toward a Model of Comprehension. *Journal of educational psychology, 97*(2), 246.

Pulido, D. (2009). Vocabulary processing and acquisition through reading: Evidence for the rich getting richer. En Z. Han, & N. Anderson (Eds.), *Second language reading research and instruction: Crossing the boundaries.* (pp. 65-82). Ann Arbor, MI: University of Michigan Press.

Ramírez, G., Chen, X., Geva, E., & Kiefer, H. (2010). Morphological awareness in Spanish-speaking English language learners: Within and cross-language effects on word reading. *Reading and Writing, 23*(3-4), 337-358.

Ramírez, G., Chen, X., & Pasquarella, A. (2013). Cross-linguistic transfer of morphological awareness in Spanish-speaking English language learners: The facilitating effect of cognate knowledge. *Topics in Language Disorders, 33*(1), 73-92.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1-32.

Redman, C. (2011). The Brick Wall. *Zhtoolkit*. Recuperado de: http://www.zhtoolkit.com/posts/2011/12/hapax-legomena-vs-the-brick-wall/.

Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning*. Clevedon: Multilingual Matters.

Rott, S. (1999). The effect of exposure frequency on the intermediate language learner's incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition, 21*(04), 589-619.

Schmitt, N. (1997). Vocabulary learning strategies. En N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 199-227). Cambridge: Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Hampshire: Palgrave Macmillan.

Schmitt, N., & Dunham, B. (1999). Exploring native and non-native intuitions of word frequency. *Second Language Research, 15*(4), 389-411.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26-43.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language testing, 18*(1), 55-88.

Schwartz, A. I., Kroll, J. F., & Diaz, M. (2007). Reading words in Spanish and English: Mapping orthography to phonology in two languages. *Language and Cognitive Processes, 22*(1), 106-129.

Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching, 10*(1-4), 209-232.

Sinclair, J. (2005). Corpus and text-basic principles. En M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice.* (pp. 1-16). Oxford: Oxbow Books.

Sökmen, A. J. (1997). Current trends in teaching second language vocabulary. En N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 237-257). Cambridge: Cambridge University Press.

Sosins, A. (2009). *Proper nouns class.* [Software]. Disponible en: http://ar2rsawseen.users.phpclasses.org/package/6970-PHP-Extract-proper-nouns-from-texts.html.

Thomas, S., & Gaby, N. (2005). *The Big Red Book of Spanish Vocabulary: 30,000 Words Through Cognates, Roots, and Suffixes.* New York: McGraw-Hill.

Thorndike, E. L. (1921). *The teacher's word book.* New York: Teachers College, Columbia University. Recuperado de: http://archive.org/details/cu31924014451409.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's book of 30,000 words.* New York: Teachers college, Columbia University.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*(4), 442-458.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes, 28*(3), 170-182.

Waring, R., & Nation, I. S. P. (2004). Second language reading and incidental vocabulary learning. *Angles on the English speaking world, 4*, 97-110.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader. *Reading in a Foreign language, 15*(2), 130-163.

West, M. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Londres: Longman.

Wilkins, D. A. (1972). *Linguistics in language teaching*. Londres: Arnold.

Xue, G., & Nation, I. S. P. (1984). A University word list. *Language learning and communication, 3*(2), 215-229.

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., & Healy, N. A. (1995). Growth of a functionally important lexicon. *Journal of Literacy Research, 27*(2), 201-212.

Zimmer, B. (2012, 3 de febrero). "Not to Put Too Fine a Point Upon It": How Dickens Helped Shape the Lexicon. *Word Routes,* pp. 1. Recuperado de: http://www.visualthesaurus.com/cm/wordroutes/not-to-put-too-fine-a-point-upon-it-how-dickens-helped-shape-the-lexicon.

Zipf, G. (1949). *Human Behavior and Principle of Least Effort: an Introduction to Human Ecology*. Cambridge: Addison Wesley Press

APÉNDICES

APÉNDICE A. LISTA DE KEYWORDS PARA LA INVESTIGACIÓN

1. A
2. ABLE
3. ABOUT
4. ABOVE
5. ABROAD
6. ACCOMPLISH
7. ACCORDING
8. ACCOUNT
9. ACCURATE
10. ACHE
11. ACHIEVE
12. ACROSS
13. ACTUAL
14. ADD
15. ADDRESS
16. ADVERTISE
17. ADVICE
18. ADVOCATE
19. AERIAL
20. AFFAIR
21. AFFORD
22. AFRAID
23. AFTER
24. AFTERNOON
25. AGAIN
26. AGAINST
27. AGE
28. AGO
29. AGREE
30. AHEAD
31. AID
32. AIM
33. AIRCRAFT
34. AIRPORT
35. ALIVE
36. ALL
37. ALLEY
38. ALLOCATE
39. ALLOW
40. ALMIGHTY
41. ALMOST
42. ALONE
43. ALONG
44. ALONGSIDE
45. ALREADY
46. ALRIGHT
47. ALSO
48. ALTHOUGH
49. ALTOGETHER
50. ALWAYS
51. AMAZE
52. AMONG
53. AMOUNT
54. AMUSE
55. ANCIENT
56. AND
57. ANGER
58. ANGRY
59. ANNOY
60. ANOTHER
61. ANSWER
62. ANY
63. APOLOGY
64. APPAL
65. APPEAL
66. APPLE
67. APPLIANCE
68. APPLY
69. APPOINT
70. APPROACH
71. ARGUE
72. ARM
73. ARMY
74. AROUND
75. ARRANGE
76. ARREARS
77. ARROW
78. AS
79. ASHAMED
80. ASIDE
81. ASK
82. ASLEEP
83. ASSESS
84. ASSIST
85. ASSURE
86. ASTONISH
87. AT
88. ATTACH
89. ATTEMPT
90. ATTEND
91. AUNT
92. AUTUMN
93. AVAILABLE
94. AVERAGE
95. AVOID
96. AWAKE
97. AWARD
98. AWARE
99. AWAY
100. AWFUL
101. AWKWARD
102. BACK
103. BACKGROUND
104. BAD
105. BADGE
106. BAG
107. BAKE
108. BALL
109. BAN
110. BANG
111. BARE
112. BARGAIN
113. BARK
114. BARN
115. BASEMENT
116. BASH
117. BASKET
118. BATH
119. BATTLE
120. BE
121. BEACH
122. BEAM
123. BEAN
124. BEAR
125. BEAT
126. BEAUTY
127. BECAUSE
128. BECOME
129. BED
130. BEEF
131. BEER
132. BEFORE
133. BEG
134. BEGIN
135. BEHALF
136. BEHAVE
137. BEHAVIOUR
138. BEHIND

139. BELIEF
140. BELIEVE
141. BELL
142. BELONG
143. BELOW
144. BELT
145. BENCH
146. BEND
147. BENEATH
148. BESIDE
149. BEST
150. BET
151. BETWEEN
152. BEYOND
153. BID
154. BIG
155. BIKE
156. BILL
157. BIN
158. BIND
159. BIRD
160. BIRTH
161. BISCUIT
162. BISHOP
163. BIT
164. BITE
165. BITTER
166. BLACK
167. BLADE
168. BLAME
169. BLANK
170. BLANKET
171. BLESS
172. BLIMEY
173. BLIND
174. BLOKE
175. BLOOD
176. BLOOM
177. BLOW
178. BLUE
179. BOARD
180. BODY
181. BOG
182. BOIL
183. BOLD
184. BOLT
185. BOND
186. BONE
187. BOOK
188. BOOM
189. BOOST
190. BORING
191. BOROUGH
192. BORROW
193. BOSS
194. BOTH
195. BOTHER
196. BOTTOM
197. BOUNCE
198. BOUNDARY
199. BOW
200. BOWL
201. BOX
202. BOY
203. BOYFRIEND
204. BRACKET
205. BRAIN
206. BRAKE
207. BRANCH
208. BRAND
209. BRASS
210. BRAVE
211. BREAD
212. BREAK
213. BREAKDOWN
214. BREAKFAST
215. BREAST
216. BREATH
217. BREATHE
218. BREED
219. BRICK
220. BRIDGE
221. BRIEF
222. BRIGHT
223. BRING
224. BROAD
225. BROTHER
226. BROWN
227. BRUSH
228. BUCK
229. BUCKET
230. BUDGET
231. BUG
232. BUGGER
233. BUILD
234. BULK
235. BULL
236. BULLY
237. BUMP
238. BUNCH
239. BURDEN
240. BURGLE
241. BURN
242. BURST
243. BURY
244. BUSH
245. BUSINESS
246. BUST
247. BUSY
248. BUT
249. BUTCHER
250. BUTTER
251. BUY
252. BUZZ
253. BY
254. BYPASS
255. CALL
256. CAN
257. CANDLE
258. CANS
259. CAP
260. CAPABLE
261. CARD
262. CARDBOARD
263. CARE
264. CAROL
265. CARPET
266. CARROT
267. CARRY
268. CART
269. CASH
270. CAST
271. CASUALTY
272. CAT
273. CATCH
274. CATER
275. CEILING
276. CERTAIN
277. CHAIN
278. CHAIR
279. CHAIRMAN
280. CHALLENGE
281. CHAMBER
282. CHANCE
283. CHANGE
284. CHANNEL
285. CHAP
286. CHAPEL
287. CHAPTER
288. CHARGE
289. CHARITY
290. CHARM
291. CHART
292. CHASE
293. CHEAP
294. CHEAT
295. CHEEK
296. CHEER
297. CHEESE
298. CHEMIST
299. CHEST
300. CHEW
301. CHICKEN
302. CHIEF
303. CHILD
304. CHOICE
305. CHOKE
306. CHOOSE

307. CHOP
308. CHRISTMAS
309. CHUCK
310. CHUNK
311. CHURCH
312. CITIZEN
313. CITY
314. CLAIM
315. CLAP
316. CLASH
317. CLEAN
318. CLERK
319. CLEVER
320. CLIFF
321. CLIMB
322. CLOCK
323. CLOSE
324. CLOSES
325. CLOTH
326. CLOTHE
327. CLOUD
328. CLUE
329. CLUSTER
330. CLUTCH
331. COACH
332. COAL
333. COAT
334. COCK
335. COIN
336. COLD
337. COLLAR
338. COLLEGE
339. COME
340. COMPLAIN
341. COMPLIMENT
342. COMPREHENSIVE
343. CONCERN
344. CONFIDENCE
345. CONGRATULATE
346. CONTEST
347. CONVEY
348. COOK
349. COOL
350. COP
351. COPE
352. CORE
353. CORNER
354. COTTAGE
355. COTTON
356. COUGH
357. COULD
358. COUNCIL
359. COUNTER
360. COUNTRY
361. COUNTY
362. COUPLE
363. COUSIN
364. COW
365. CRACK
366. CRACKER
367. CRAFT
368. CRAMP
369. CRAP
370. CRASH
371. CRAWL
372. CRAZY
373. CREEP
374. CREW
375. CRIPPLE
376. CRISP
377. CROP
378. CROWD
379. CROWN
380. CRUSH
381. CRY
382. CUPBOARD
383. CURL
384. CUSHION
385. CUSTOMER
386. CUT
387. DAD
388. DAFT
389. DAMAGE
390. DAMN
391. DAMP
392. DANGER
393. DARE
394. DARK
395. DARLING
396. DASH
397. DATE
398. DAUGHTER
399. DEAD
400. DEAF
401. DEAL
402. DEAR
403. DEATH
404. DEBT
405. DEED
406. DEEP
407. DEFEAT
408. DEGREE
409. DELAY
410. DELIGHT
411. DELIVER
412. DENY
413. DEPUTY
414. DERBY
415. DESERVE
416. DESIGN
417. DESIRE
418. DESK
419. DESPITE
420. DETACH
421. DETAIL
422. DEVELOP
423. DEVIL
424. DIAL
425. DIE
426. DIG
427. DINE
428. DINNER
429. DIP
430. DIRECTION
431. DIRTY
432. DISABLED
433. DISAPPOINT
434. DISEASE
435. DISGRACE
436. DISGUISE
437. DISGUST
438. DISH
439. DISPLAY
440. DISPOSE
441. DISTRESS
442. DITCH
443. DIVE
444. DO
445. DOCK
446. DODGY
447. DOG
448. DOLE
449. DOOR
450. DOORSTEP
451. DOORWAY
452. DOT
453. DOUBT
454. DOWN
455. DRAFT
456. DRAG
457. DRAIN
458. DRAUGHT
459. DRAW
460. DRAWER
461. DREAD
462. DREADFUL
463. DREAM
464. DRESS
465. DRIFT
466. DRILL
467. DRINK
468. DRIP
469. DRIVE
470. DROP
471. DROWN
472. DRUM
473. DRY
474. DUCK

475. DUE
476. DULL
477. DUMMY
478. DUMP
479. DUST
480. DUTCH
481. DUTY
482. DYE
483. EACH
484. EAR
485. EARLY
486. EARN
487. EARTH
488. EASE
489. EAST
490. EASTER
491. EASTERN
492. EASY
493. EAT
494. EDGE
495. EGG
496. EIGHT
497. EITHER
498. ELBOW
499. ELDER
500. ELDEST
501. ELEVEN
502. ELSE
503. EMBARRASS
504. EMPTY
505. ENABLE
506. ENCOURAGE
507. END
508. ENGAGE
509. ENGINE
510. ENJOY
511. ENOUGH
512. ENQUIRE
513. ENSURE
514. ENTITLE
515. ENVELOPE
516. ENVIRONMENT
517. ESCORT
518. ESSAY
519. ESTATE
520. EVE
521. EVEN
522. EVENING
523. EVER
524. EVERY
525. EVIL
526. EVOLVE
527. EXCHANGE
528. EXCITE
529. EXHAUST
530. EXIT
531. EXPECT
532. EXPENSE
533. EXPLAIN
534. EXPOSE
535. EXPRESS
536. EYE
537. FACE
538. FACILITY
539. FACT
540. FACTORY
541. FADE
542. FAIL
543. FAILURE
544. FAINT
545. FAIR
546. FAITH
547. FALL
548. FANCY
549. FAR
550. FARE
551. FARM
552. FASHION
553. FAST
554. FAT
555. FATHER
556. FAULT
557. FEAR
558. FEATHER
559. FEATURE
560. FEE
561. FEED
562. FEEDBACK
563. FEEL
564. FELLOW
565. FEMALE
566. FENCE
567. FETCH
568. FEW
569. FIDDLE
570. FIELD
571. FIERCE
572. FIGHT
573. FILL
574. FIND
575. FINE
576. FINED
577. FINGER
578. FINISH
579. FIR
580. FIRE
581. FIRST
582. FISH
583. FIT
584. FIVE
585. FIX
586. FLAG
587. FLAME
588. FLAP
589. FLARE
590. FLAT
591. FLATTING
592. FLAVOUR
593. FLIGHT
594. FLOOD
595. FLOOR
596. FLOW
597. FLU
598. FLY
599. FOCUS
600. FOLD
601. FOLDER
602. FOLK
603. FOLLOW
604. FOND
605. FOOD
606. FOOL
607. FOOT
608. FOR
609. FORECAST
610. FOREIGN
611. FOREST
612. FOREVER
613. FORGET
614. FORGIVE
615. FORK
616. FORTH
617. FORTNIGHT
618. FORWARD
619. FOSTER
620. FOUR
621. FRAME
622. FRAMEWORK
623. FREE
624. FREEZE
625. FRESH 0
626. FRIDAY
627. FRIEND
628. FRIGHT
629. FRINGE
630. FROG
631. FROM
632. FROST
633. FRY
634. FUEL
635. FULL
636. FUME
637. FUN
638. FUR
639. FURNISH
640. FURNITURE
641. FURTHER
642. FUSS

643. GAIN
644. GAMBLE
645. GAME
646. GANG
647. GAP
648. GARDEN
649. GATE
650. GATHER
651. GEAR
652. GEE
653. GENTLE
654. GENTLEMAN
655. GET
656. GHOST
657. GIFT
658. GIRL
659. GIVE
660. GLAD
661. GLANCE
662. GLASS
663. GLAZE
664. GLOVE
665. GLOW
666. GLUE
667. GO
668. GOD
669. GOOD
670. GOODBYE
671. GOODS
672. GOSPEL
673. GRAB
674. GRAND
675. GRANT
676. GRASS
677. GRATEFUL
678. GREAT
679. GREED
680. GREEN
681. GREET
682. GREY
683. GRIEF
684. GRIND
685. GRIP
686. GROSS
687. GROUND
688. GROW
689. GUESS
690. GUEST
691. GUILTY
692. GUN
693. GUT
694. GUY
695. HAIR
696. HAIRDRESSER
697. HALF
698. HALL
699. HAM
700. HAMMER
701. HAND
702. HANDICAP
703. HANDLE
704. HANDY
705. HANG
706. HAPPEN
707. HAPPY
708. HARASS
709. HARD
710. HARDLY
711. HARM
712. HARVEST
713. HASSLE
714. HAT
715. HATE
716. HAVE
717. HAY
718. HAZARD
719. HE
720. HEAD
721. HEADMASTER
722. HEAL
723. HEALTH
724. HEAR
725. HEART
726. HEAT
727. HEAVEN
728. HEAVY
729. HECK
730. HEDGE
731. HEEL
732. HEIGHT
733. HELL
734. HELLO
735. HELP
736. HEN
737. HERE
738. HESITATE
739. HIDE
740. HIGH
741. HIGHLIGHT
742. HILL
743. HINT
744. HIP
745. HIRE
746. HIT
747. HOLD
748. HOLE
749. HOLIDAY
750. HOLY
751. HOME
752. HOMEWORK
753. HONEY
754. HOOD
755. HOOK
756. HOORAY
757. HOP
758. HOPE
759. HORSE
760. HOST
761. HOT
762. HOUSE
763. HOUSEHOLD
764. HOUSEWIFE
765. HOW
766. HOWEVER
767. HUGE
768. HUNDRED
769. HUNGER
770. HUNT
771. HURRY
772. HURT
773. HUSBAND
774. HUT
775. I
776. ICE
777. IDLE
778. IF
779. ILL
780. IMPROVE
781. INCH
782. INCOME
783. INCREASE
784. INDEED
785. INHERIT
786. INJURE
787. INN
788. INPUT
789. INSIDE
790. INSTANCE
791. INSTEAD
792. INSURE
793. INTENT
794. INTERVIEW
795. INTO
796. INTRODUCE
797. INVEST
798. INVOLVE
799. IRON
800. ISSUE
801. IT
802. JAIL
803. JAM
804. JANUARY
805. JAR
806. JAW
807. JEWEL
808. JOB
809. JOG
810. JOIN

811. JOINT
812. JOKE
813. JOLLY
814. JOURNALIST
815. JOURNEY
816. JOY
817. JUG
818. JUICE
819. JUMP
820. JUNCTION
821. JUST
822. KEEN
823. KEEP
824. KETTLE
825. KEY
826. KID
827. KIDDING
828. KIDNEY
829. KILL
830. KIND
831. KING
832. KISS
833. KIT
834. KITCHEN
835. KNEE
836. KNIFE
837. KNOCK
838. KNOT
839. KNOW
840. KNOWLEDGE
841. LABEL
842. LACK
843. LAD
844. LADDER
845. LADY
846. LAKE
847. LAMB
848. LAND
849. LANDLORD
850. LANDSCAPE
851. LANE
852. LAP
853. LARGE
854. LAST
855. LATE
856. LAUGH
857. LAUGHTER
858. LAUNCH
859. LAW
860. LAWN
861. LAY
862. LAYER
863. LAYOUT
864. LAZY
865. LEAD
866. LEAF
867. LEAFLET
868. LEAK
869. LEAN
870. LEAP
871. LEARN
872. LEAVE
873. LECTURE
874. LEFT
875. LEG
876. LEISURE
877. LEND
878. LENGTH
879. LESS
880. LET
881. LEVEL
882. LEVER
883. LIABLE
884. LIBRARY
885. LID
886. LIE
887. LIFE
888. LIFT
889. LIGHT
890. LIKE
891. LIKELY
892. LIKENESS
893. LINK
894. LIP
895. LISTEN
896. LITTER
897. LITTLE
898. LIVE
899. LOAD
900. LOAN
901. LOCATE
902. LOCK
903. LODGE
904. LOFT
905. LOG
906. LONE
907. LONG
908. LOO
909. LOOK
910. LOOP
911. LOOSE
912. LORD
913. LORRY
914. LOSE
915. LOSS
916. LOT
917. LOUD
918. LOUNGE
919. LOVE
920. LOW
921. LOYAL
922. LUCK
923. LUMP
924. LUNCHTIME
925. LUNG
926. LUXURY
927. MACHINE
928. MAD
929. MAGAZINE
930. MAIN
931. MAKE
932. MALE
933. MAN
934. MANAGE
935. MANOR
936. MANY
937. MAR
938. MARKET
939. MARRY
940. MARVEL
941. MATCH
942. MATE
943. MATH
944. MATTER
945. MAY
946. MEAL
947. MEAN
948. MEANTIME
949. MEANWHILE
950. MEASURE
951. MEAT
952. MEET
953. MELT
954. MEND
955. MERCHANT
956. MERGE
957. MERRY
958. MESS
959. MIDDLE
960. MIGHT
961. MILD
962. MILE
963. MILK
964. MILL
965. MIND
966. MIRACLE
967. MIRROR
968. MISS
969. MISTAKE
970. MIX
971. MIXTURE
972. MOAN
973. MOCK
974. MONDAY
975. MONEY
976. MONTH
977. MOOD
978. MOON

979. MOOR
980. MORE
981. MORNING
982. MORTGAGE
983. MOST
984. MOTHER
985. MOTION
986. MOTORBIKE
987. MOTORWAY
988. MOUNT
989. MOUSE
990. MOUTH
991. MOVIE
992. MRS
993. MUCK
994. MUD
995. MUG
996. MURDER
997. MUST
998. NAIL
999. NAIVE
1000. NAKED
1001. NAME
1002. NANNY
1003. NARROW
1004. NASTY
1005. NAUGHTY
1006. NAVY
1007. NAY
1008. NEAR
1009. NEAT
1010. NECK
1011. NEED
1012. NEEDLE
1013. NEGLECT
1014. NEIGHBOUR
1015. NEITHER
1016. NEST
1017. NETWORK
1018. NEVER
1019. NEVERTHELESS
1020. NEW
1021. NEWS
1022. NEWSPAPER
1023. NEXT
1024. NICE
1025. NICK
1026. NIECE
1027. NIGHT
1028. NIGHTMARE
1029. NIL
1030. NINE
1031. NIP
1032. NOD
1033. NOISE
1034. NONE
1035. NONSENSE
1036. NOR
1037. NORTHERN
1038. NOSE
1039. NOTICE
1040. NOUGHT
1041. NOW
1042. NOWT
1043. NUISANCE
1044. NURSE
1045. NURSERY
1046. NUT
1047. OAK
1048. ODD
1049. OF
1050. OFF
1051. OFTEN
1052. OIL
1053. OLD
1054. ONCE
1055. ONE
1056. ONLY
1057. ONWARDS
1058. OPEN
1059. ORANGE
1060. OTHER
1061. OTHERWISE
1062. OUGHT
1063. OUT
1064. OUTCOME
1065. OUTRAGE
1066. OVEN
1067. OVER
1068. OVERALL
1069. OVERWHELM
1070. OWE
1071. OWN
1072. PAD
1073. PADDY
1074. PAIN
1075. PAL
1076. PALE
1077. PAN
1078. PARADE
1079. PARCEL
1080. PARENT
1081. PARISH
1082. PARTNER
1083. PARTY
1084. PAT
1085. PATCH
1086. PATH
1087. PATTERN
1088. PAVE
1089. PAY
1090. PEACE
1091. PEAK
1092. PEER
1093. PEN
1094. PENCE
1095. PENCIL
1096. PENNY
1097. PEOPLE
1098. PERFORM
1099. PERHAPS
1100. PET
1101. PETROL
1102. PICK
1103. PICTURE
1104. PIE
1105. PIER
1106. PIG
1107. PIGEON
1108. PIN
1109. PINCH
1110. PINK
1111. PINT
1112. PIT
1113. PITCH
1114. PITY
1115. PLACE
1116. PLAIN
1117. PLASTER
1118. PLAY
1119. PLEASE
1120. PLEASURE
1121. PLENTY
1122. PLONK
1123. PLOT
1124. PLOUGH
1125. PLUG
1126. PLUMB
1127. PLUS
1128. POCKET
1129. POINT
1130. POISON
1131. POKE
1132. POLISH
1133. POLITE
1134. POLL
1135. POLLUTE
1136. POND
1137. POOL
1138. POOR
1139. PORTRAIT
1140. POSH
1141. POSTCARD
1142. POT
1143. POUND
1144. POUR
1145. POWDER
1146. PRACTITIONER

1147. PRAISE
1148. PRAY
1149. PREACH
1150. PREGNANT
1151. PRESSURE
1152. PRESUME
1153. PRETTY
1154. PREVENT
1155. PRIDE
1156. PRIEST
1157. PRINT
1158. PRIZE
1159. PRODUCE
1160. PROFIT
1161. PROOF
1162. PROUD
1163. PUB
1164. PULL
1165. PUMP
1166. PUNCH
1167. PUNISH
1168. PURCHASE
1169. PURPLE
1170. PURPOSE
1171. PURSE
1172. PUSH
1173. PUT
1174. PUZZLE
1175. QUEEN
1176. QUERY
1177. QUESTION
1178. QUEUE
1179. QUICK
1180. QUID
1181. QUIET
1182. QUITE
1183. QUOTE
1184. RABBIT
1185. RACK
1186. RAG
1187. RAIL
1188. RAIN
1189. RAISE
1190. RANDOM
1191. RANGE
1192. RAPE
1193. RATE
1194. RATHER
1195. RATTLE
1196. RAW
1197. REACH
1198. READ
1199. READY
1200. REALISE
1201. REAR
1202. RECALL
1203. RECEIPT
1204. RECESSION
1205. RECIPE
1206. RECKON
1207. RECORD
1208. RED
1209. REFUSE
1210. REGARD
1211. REGRET
1212. REHEARSE
1213. REJECT
1214. RELEASE
1215. RELIEF
1216. RELIEVE
1217. RELUCTANT
1218. RELY
1219. REMAIN
1220. REMARK
1221. REMEMBER
1222. REMIND
1223. REPLY
1224. REQUEST
1225. RESCUE
1226. RESEARCH
1227. RESEMBLE
1228. RESIGN
1229. RESORT
1230. RESOURCE
1231. RESUME
1232. RETAIL
1233. RETIRE
1234. REVOLT
1235. REWARD
1236. RIBBON
1237. RICE
1238. RICH
1239. RID
1240. RIDE
1241. RIGHT
1242. RING
1243. RIP
1244. RISE
1245. RISK
1246. RIVER
1247. ROAD
1248. ROCKET
1249. ROLL
1250. ROOF
1251. ROOM
1252. ROOT
1253. ROPE
1254. ROT
1255. ROTTEN
1256. ROUGH
1257. ROUND
1258. ROVE
1259. ROW
1260. ROYAL
1261. RUB
1262. RUBBER
1263. RUBBISH
1264. RULE
1265. RUN
1266. RUSH
1267. SAD
1268. SAFE
1269. SAIL
1270. SAKE
1271. SALE
1272. SAME
1273. SAMPLE
1274. SAND
1275. SATURDAY
1276. SAUSAGE
1277. SAVE
1278. SAY
1279. SCAN
1280. SCARE
1281. SCHEDULE
1282. SCHEME
1283. SCHOOL
1284. SCOPE
1285. SCORE
1286. SCOTLAND
1287. SCOUT
1288. SCRAP
1289. SCRAPE
1290. SCRATCH
1291. SCREAM
1292. SCREEN
1293. SCREW
1294. SCRIBBLE
1295. SCRIPT
1296. SCRUB
1297. SEA
1298. SEAL
1299. SEARCH
1300. SEASON
1301. SEAT
1302. SEE
1303. SEED
1304. SEEK
1305. SEEM
1306. SELDOM
1307. SELF
1308. SELL
1309. SEND
1310. SENIOR
1311. SENSE
1312. SENSIBLE
1313. SENSITIVE
1314. SET

1315. SETTLE
1316. SEVEN
1317. SEVERAL
1318. SEW
1319. SHADE
1320. SHADOW
1321. SHAKE
1322. SHALL
1323. SHAME
1324. SHAPE
1325. SHARE
1326. SHARP
1327. SHAVE
1328. SHE
1329. SHED
1330. SHEEP
1331. SHEER
1332. SHEET
1333. SHELF
1334. SHELL
1335. SHELTER
1336. SHIELD
1337. SHIFT
1338. SHINE
1339. SHIP
1340. SHIRT
1341. SHOE
1342. SHOOT
1343. SHOP
1344. SHORE
1345. SHORT
1346. SHOULD
1347. SHOULDER
1348. SHOUT
1349. SHOVE
1350. SHOVEL
1351. SHOW
1352. SHOWER
1353. SHUT
1354. SHY
1355. SICK
1356. SIDE
1357. SIGHT
1358. SIGN
1359. SIGNAL
1360. SIGNATURE
1361. SILK
1362. SILLY
1363. SILVER
1364. SIN
1365. SINCE
1366. SING
1367. SINGLE
1368. SINK
1369. SIR
1370. SISTER
1371. SIT
1372. SIX
1373. SIXPENCE
1374. SIZE
1375. SKETCH
1376. SKILL
1377. SKIN
1378. SKIP
1379. SKIRT
1380. SKY
1381. SLAP
1382. SLASH
1383. SLAVE
1384. SLEEP
1385. SLEEVE
1386. SLICE
1387. SLIDE
1388. SLIGHT
1389. SLIM
1390. SLIP
1391. SLOT
1392. SLOW
1393. SMACK
1394. SMALL
1395. SMART
1396. SMASH
1397. SMELL
1398. SMILE
1399. SMOKE
1400. SMOOTH
1401. SNAKE
1402. SNAP
1403. SNEAK
1404. SNIFF
1405. SNOW
1406. SO
1407. SOAK
1408. SOCK
1409. SOD
1410. SOFT
1411. SOIL
1412. SOLE
1413. SOLICITOR
1414. SOLVE
1415. SOME
1416. SOMEWHAT
1417. SON
1418. SONG
1419. SOON
1420. SORE
1421. SORRY
1422. SORT
1423. SOUND
1424. SOURCE
1425. SOUTH
1426. SOUTHERN
1427. SPARE
1428. SPARK
1429. SPEAK
1430. SPECTACLE
1431. SPEECH
1432. SPEED
1433. SPELL
1434. SPEND
1435. SPILL
1436. SPIN
1437. SPIT
1438. SPITE
1439. SPLASH
1440. SPLIT
1441. SPOIL
1442. SPOON
1443. SPOT
1444. SPREAD
1445. SPRING
1446. SQUARE
1447. SQUASH
1448. SQUEEZE
1449. STAB
1450. STACK
1451. STAFF
1452. STAGE
1453. STAIN
1454. STAIRS
1455. STAKE
1456. STALL
1457. STAND
1458. STAR
1459. STARE
1460. START
1461. STARVE
1462. STAY
1463. STEADY
1464. STEAL
1465. STEAM
1466. STEEL
1467. STEEP
1468. STEER
1469. STEP
1470. STICK
1471. STICKY
1472. STIFF
1473. STILL
1474. STINK
1475. STIR
1476. STITCH
1477. STONE
1478. STOP
1479. STORE
1480. STORM
1481. STRAIGHT
1482. STRAIGHTFORWARD

1483. STRAIN
1484. STRANGE
1485. STRAP
1486. STRAW
1487. STRAWBERRY
1488. STREAM
1489. STREET
1490. STRENGTH
1491. STRETCH
1492. STRIDE
1493. STRIKE
1494. STRING
1495. STRIP
1496. STRIPE
1497. STROKE
1498. STRONG
1499. STRUGGLE
1500. STUFF
1501. SUBTLE
1502. SUCCEED
1503. SUCH
1504. SUCK
1505. SUDDEN
1506. SUE
1507. SUIT
1508. SUMMER
1509. SUN
1510. SUNDAY
1511. SUPPER
1512. SUPPLY
1513. SURE
1514. SURFACE
1515. SURGEON
1516. SURGERY
1517. SURNAME
1518. SURROUND
1519. SURVEY
1520. SWALLOW
1521. SWAN
1522. SWAP
1523. SWEAR
1524. SWEAT
1525. SWEEP
1526. SWEET
1527. SWIM
1528. SWING
1529. SWITCH
1530. SWORD
1531. SYMPATHY
1532. TAB
1533. TABLE
1534. TACK
1535. TACKLE
1536. TAG
1537. TAIL
1538. TAKE
1539. TALE
1540. TALK
1541. TALL
1542. TAPE
1543. TARGET
1544. TASK
1545. TASTE
1546. TAX
1547. TEACH
1548. TEAM
1549. TEAR
1550. TEASE
1551. TEENAGE
1552. TELECOM
1553. TELL
1554. TEMPER
1555. TEN
1556. TEND
1557. TENDER
1558. TENT
1559. TERM
1560. TERRIFIC
1561. THAN
1562. THANK
1563. THE
1564. THEN
1565. THERE
1566. THEREFORE
1567. THEY
1568. THICK
1569. THIEF
1570. THIN
1571. THING
1572. THINK
1573. THIRTEEN
1574. THIRTY
1575. THIS
1576. THOROUGH
1577. THOUGH
1578. THOUSAND
1579. THREAT
1580. THREE
1581. THRILL
1582. THROAT
1583. THROUGH
1584. THROW
1585. THUMB
1586. THUNDER
1587. THURSDAY
1588. THUS
1589. TICK
1590. TIDE
1591. TIDY
1592. TIE
1593. TIGHT
1594. TILE
1595. TILLED
1596. TIMBER
1597. TIME
1598. TIMETABLE
1599. TIN
1600. TINY
1601. TIP
1602. TIRE
1603. TISSUE
1604. TO
1605. TOAST
1606. TODAY
1607. TOE
1608. TOGETHER
1609. TOILET
1610. TOKEN
1611. TOMORROW
1612. TONGUE
1613. TONIGHT
1614. TOO
1615. TOOL
1616. TOOTH
1617. TOP
1618. TORCH
1619. TOSS
1620. TOUCH
1621. TOUGH
1622. TOWARD
1623. TOWEL
1624. TOWER
1625. TOWN
1626. TOY
1627. TRACE
1628. TRACK
1629. TRADE
1630. TRAIL
1631. TRAIN
1632. TRANSLATE
1633. TRAVEL
1634. TREAD
1635. TREASURE
1636. TREAT
1637. TREE
1638. TREND
1639. TRIAL
1640. TRICK
1641. TRIGGER
1642. TRIP
1643. TROLLEY
1644. TROUBLE
1645. TROUSERS
1646. TRUCK
1647. TRUE
1648. TRUST
1649. TRY
1650. TUESDAY

1651. TUMBLE
1652. TUNE
1653. TURN
1654. TWELVE
1655. TWENTY
1656. TWIN
1657. TWIST
1658. TWO
1659. TYRE
1660. UGLY
1661. UN
1662. UNCLE
1663. UNDER
1664. UNDERGROUND
1665. UNDERLINE
1666. UNDERSTAND
1667. UNLESS
1668. UNTIL
1669. UP
1670. UPDATE
1671. UPON
1672. UPPER
1673. UPSET
1674. UPWARDS
1675. URGE
1676. UTILITY
1677. VACUUM
1678. VAT
1679. VERY
1680. VEST
1681. VET
1682. VICIOUS
1683. VIEW
1684. VILLAGE
1685. WAGE
1686. WAIT
1687. WAKE
1688. WALES
1689. WALK
1690. WALL
1691. WALLET
1692. WALLPAPER
1693. WANDER
1694. WANT
1695. WAR
1696. WARD
1697. WARDROBE
1698. WAREHOUSE
1699. WARM
1700. WARN
1701. WARRANT
1702. WASH
1703. WASTE
1704. WATCH
1705. WATER
1706. WAVE
1707. WAY
1708. WE
1709. WEAK
1710. WEALTH
1711. WEAPON
1712. WEAR
1713. WEATHER
1714. WED
1715. WEDNESDAY
1716. WEE
1717. WEED
1718. WEEK
1719. WEIGH
1720. WEIRD
1721. WELCOME
1722. WELFARE
1723. WELL
1724. WEST
1725. WESTERN
1726. WET
1727. WHACK
1728. WHAT
1729. WHATSOEVER
1730. WHEEL
1731. WHEELCHAIR
1732. WHEN
1733. WHERE
1734. WHEREABOUTS
1735. WHEREAS
1736. WHEREBY
1737. WHETHER
1738. WHICH
1739. WHILE
1740. WHIP
1741. WHISPER
1742. WHISTLE
1743. WHITE
1744. WHO
1745. WHOLE
1746. WHOOP
1747. WHY
1748. WICKED
1749. WIDE
1750. WIDOW
1751. WIFE
1752. WIG
1753. WILD
1754. WILL
1755. WIN
1756. WIND
1757. WINDOW
1758. WINE
1759. WING
1760. WINTER
1761. WIPE
1762. WIRE
1763. WISE
1764. WISH
1765. WITH
1766. WITHDRAW
1767. WITHIN
1768. WITHOUT
1769. WITNESS
1770. WOBBLE
1771. WOLF
1772. WOMAN
1773. WONDER
1774. WOOD
1775. WOOL
1776. WORD
1777. WORK
1778. WORKSHOP
1779. WORLD
1780. WORRY
1781. WORSE
1782. WORTH
1783. WORTHWHILE
1784. WOULD
1785. WOUND
1786. WRAP
1787. WRECK
1788. WRESTLE
1789. WRIST
1790. WRITE
1791. WRONG
1792. YEAR
1793. YELL
1794. YELLOW
1795. YES
1796. YESTERDAY
1797. YET
1798. YOU
1799. YOUNG
1800. YOUTH

APÉNDICE B. LISTA DE KEYWORDS PARA LA DOCENCIA

NIVEL 1 (Contiene 691 unidades léxicas)

A	AS	BIRTH	CATCH
ABOUT	ASHAMED	BIT	CERTAIN
ABOVE	ASK	BLACK	CHAIR
ACROSS	AT	BLOOD	CHANCE
ADD	AUTUMN	BLOW	CHANGE
ADDRESS	AWARE	BLUE	CHARGE
AFFORD	AWAY	BOARD	CHEAP
AFRAID	AWFUL	BONE	CHICKEN
AFTER	BACK	BOOK	CHILD
AFTERNOON	BAD	BORING	CHOICE
AGAIN	BAG	BOTH	CHRISTMAS
AGAINST	BATH	BOTHER	CHURCH
AGE	BE	BOX	CITY
AGREE	BEACH	BOY	CLEAN
AHEAD	BEAR	BREAD	CLIMB
ALLOW	BEAT	BREAK	CLOCK
ALMOST	BEAUTY	BREAKFAST	CLOSE
ALONE	BECAUSE	BREATH	CLOSED
ALONG	BECOME	BRIGHT	CLOTHES
ALREADY	BED	BRING	COAT
ALRIGHT	BEFORE	BROTHER	COLD
ALSO	BEGIN	BROWN	CONCERN
ALTHOUGH	BEHIND	BUILD	COOK
ALWAYS	BELIEVE	BURN	COOL
AMONG	BELOW	BUSH	CORNER
AMOUNT	BENEATH	BUSINESS	COULD
AND	BESIDE	BUSY	COUNTRY
ANGRY	BET	BUT	COUPLE
ANOTHER	BETWEEN	BUY	CRAZY
ANSWER	BEYOND	BY	CRY
ANY	BIG	CALL	CUT
AROUND	BILL	CARRY	DAD
ARRANGE	BIRD	CAT	DANGER

DARK
DARLING
DAUGHTER
DEAD
DEAL
DEAR
DEATH
DEEP
DEGREE
DIE
DIG
DIRTY
DO
DOG
DOOR
DOUBT
DOWN
DRAW
DREAM
DRESS
DRINK
DRIVE
DROP
DRY
EACH
EAR
EARLY
EARTH
EAST
EASY
EDGE
EGG
EIGHT
EITHER
ELSE
EMPTY
END
ENGINE
ENJOY
ENOUGH
EVEN
EVENING
EVER
EVERY
EXPRESS
EYE
FACE
FACT
FAIR
FALL
FAR
FARM
FAST
FAT
FATHER
FEAR
FEED
FEEL
FELLOW
FEW
FIELD
FIGHT
FILL
FIND
FINGER
FINISH
FIRE
FIRST
FISH
FIT
FIVE
FIX
FLAT
FLY
FOLLOW
FOOD
FOOT
FOREST
FORGET
FORWARD
FOUR
FREE
FREEZE
FRESH
FRIDAY
FRIGHT
FROM
FULL
FUN
FURTHER
GAME
GARDEN
GET
GIRL
GIVE
GLAD
GLANCE
GLASS
GO
GOD
GOOD
GOODBYE
GRASS
GREAT
GREEN
GREY
GROUND
GROW
GUESS
GUN
GUY
HAIR
HALF
HALL
HAND
HANDLE
HANG
HAPPEN
HAPPY
HARD
HAT
HATE
HAVE
HE
HEAD
HEALTH
HEAR
HEART
HEAT
HEAVY
HELL
HELLO
HELP
HERE
HIDE
HIGH
HILL
HIT
HOLD
HOLE
HOLIDAY
HOME
HOPE
HORSE
HOT
HOUSE
HOW
HOWEVER
HUGE
HUNDRED
HUNGER
HUNT
HURRY
HURT
HUSBAND
I
ICE
IF
INDEED
INSIDE
INSTEAD
INSURE
INTO
ISSUE
IT
JOB
JOIN
JOKE
JUMP
KEEP
KEY
KID
KILL
KIND
KING
KISS
KITCHEN
KNOCK
KNOW
LADY
LAKE
LAND
LAST
LATE
LAUGH
LAW
LAY
LAZY
LEAD
LEARN
LEAVE
LEFT
LEG
LESS
LET
LEVEL
LIFE
LIFT
LIGHT
LIKE
LIP
LISTEN
LITTLE
LIVE
LOAD
LOCK
LONG
LOOK
LORD
LOSE
LOT
LOUD
LOVE
LOW
LUCK
MACHINE
MAD
MAIN
MAKE
MANAGE
MANY
MARKET
MARRY
MAY

MEAL
MEAN
MEET
MESS
MIDDLE
MIGHT
MILE
MILK
MIND
MISS
MISTAKE
MONDAY
MONEY
MONTH
MORE
MORNING
MOST
MOTHER
MOUTH
MOVIE
MRS
MUST
NAME
NAUGHTY
NEAR
NEAT
NECK
NEED
NEIGHBOUR
NEVER
NEW
NEWS
NEXT
NICE
NIGHT
NINE
NOISE
NONE
NOSE
NOTHING
NOW
NURSE
ODD
OF
OFF
OFTEN
OIL
OLD
ONE
ONLY
OPEN
ORANGE
OTHER
OUGHT
OUT
OVER
OWN
PAY
PENNY
PEOPLE
PERHAPS
PICK
PICTURE
PLAY
PLEASE
PLENTY
PLUS
POOR
POT
POUND
PRETTY
PULL
PUSH
PUT
QUEEN
QUICK
RABBIT
RAIN
RAISE
RATE
RATHER
REACH
READ
READY
REMEMBER
REPLY
RICH
RID
RIDE
RIGHT
RING
RISE
RIVER
ROAD
ROLL
ROOM
ROUGH
ROUND
RUBBISH
RULE
RUN
SAD
SAFE
SAIL
SAME
SATURDAY
SAY
SCARE
SCHOOL
SEA
SEAT
SEE
SEEM
SELF
SELL
SEND
SENSE
SET
SETTLE
SEVEN
SHAKE
SHALL
SHAPE
SHARE
SHE
SHIP
SHIRT
SHOE
SHOOT
SHOP
SHORT
SHOULD
SHOULDER
SHOUT
SHOW
SHUT
SHY
SICK
SIDE
SIGHT
SIGN
SILLY
SINCE
SING
SINGLE
SIR
SISTER
SIT
SIX
SIZE
SKIN
SKY
SLEEP
SLIGHT
SLIP
SLOW
SMALL
SMELL
SMILE
SMOKE
SNOW
SO
SOFT
SOME
SONG
SOON
SORRY
SORT
SOUND
SOUTH
SPEAK
SPEND
SPOT
SPRING
SQUARE
STAGE
STAND
STAR
STARE
START
STEAL
STEP
STICK
STILL
STONE
STOP
STORE
STRAIGHT
STREET
STRIKE
STRONG
STUFF
SUCH
SUDDEN
SUIT
SUMMER
SUN
SUNDAY
SURE
SWEET
SWIM
TAIL
TAKE
TALK
TAPE
TASTE
TEACH
TEAM
TEAR
TELL
TEN
TEND
TERM
THAN
THANK
THE
THEN
THERE
THEY
THICK
THING
THINK
THIRTEEN
THIRTY

THIS	TROUBLE	WATCH	WIN
THOUGH	TRUE	WATER	WIND
THOUSAND	TRUST	WAVE	WINDOW
THREE	TRY	WAY	WINE
THROAT	TUESDAY	WE	WINTER
THROUGH	TURN	WEAR	WISH
THROW	TWELVE	WEATHER	WITH
THURSDAY	TWENTY	WED	WITHIN
TIE	TWO	WEDNESDAY	WITHOUT
TIGHT	UGLY	WEEK	WOMAN
TIME	UNCLE	WELL	WONDER
TO	UNDER	WEST	WOOD
TODAY	UNDERSTAND	WET	WORD
TOGETHER	UNLESS	WHAT	WORK
TOMORROW	UNTIL	WHEEL	WORLD
TONIGHT	UP	WHEN	WORRY
TOO	UPON	WHERE	WORSE
TOOTH	VERY	WHETHER	WORTH
TOP	VIEW	WHICH	WOULD
TOUCH	WAIT	WHILE	WRITE
TOWARD	WAKE	WHITE	WRONG
TOWN	WALK	WHO	YEAR
TRACK	WALL	WHOLE	YELLOW
TRAIN	WANT	WHY	YES
TRAVEL	WAR	WIDE	YESTERDAY
TREAT	WARM	WIFE	YET
TREE	WASH	WILD	YOU
TRIP	WASTE	WILL	YOUNG

NIVEL 2 (Contiene 453 unidades léxicas)

ACCORDING	BANG	BOOM	BURY
ACCOUNT	BARE	BOSS	BUTTER
AFFAIR	BASKET	BOUNCE	CANS
AID	BATTLE	BOW	CARROT
ALIVE	BEAN	BOWL	CASH
ALTOGETHER	BEER	BRAIN	CAST
AMUSE	BEG	BRAKE	CEILING
ANGER	BELL	BRANCH	CHAIN
ANNOY	BELONG	BRAND	CHALLENGE
APPEAL	BELT	BREAST	CHANNEL
APPLE	BEND	BREATHE	CHAPTER
APPROACH	BIKE	BREED	CHARM
ASIDE	BIN	BRICK	CHASE
ASLEEP	BIND	BRIDGE	CHEAT
ASSURE	BISCUIT	BRIEF	CHEEK
ATTACH	BITE	BROAD	CHEER
AVAILABLE	BITTER	BRUSH	CHEESE
AVERAGE	BLAME	BUCK	CHEST
AVOID	BLANKET	BUCKET	CHEW
AWAKE	BLESS	BUG	CHIEF
AWKWARD	BLOOM	BUMP	CHOP
BACKGROUND	BOIL	BUNCH	CITIZEN
BAKE	BOND	BURST	CLAIM

CLEVER
CLIFF
CLOTH
CLOUD
CLUE
COACH
COAL
COP
COPE
COTTAGE
COTTON
COUGH
COUNTER
COUNTY
COUSIN
COW
CRACK
CRASH
CRAWL
CREEP
CRISP
CROWD
CROWN
CURL
CUSTOMER
DAMAGE
DARE
DEBT
DENY
DESERVE
DESIRE
DESK
DETAIL
DEVELOP
DINE
DISAPPOINT
DISEASE
DISH
DIVE
DRAG
DRAWER
DRUM
DUCK
DUE
DUMP
DUST
DUTY
EARN
EASE
ELDER
ENCOURAGE
ENGAGE
ENVELOPE
ENVIRONMENT
EVIL
EXCHANGE
EXHAUST
EXPOSE
FAIL
FAINT
FAITH
FANCY
FASHION
FEATHER
FEATURE
FEMALE
FENCE
FETCH
FLAG
FLAME
FLIGHT
FLOOD
FLOW
FOLD
FOLK
FOOL
FOREIGN
FORGIVE
FORTH
FORTNIGHT
FRAME
FROG
FROST
FRY
FUR
FURNITURE
GAIN
GATE
GATHER
GEAR
GENTLEMAN
GHOST
GIFT
GRANT
GUEST
GUILTY
HANDY
HARM
HEAVEN
HEDGE
HEIGHT
HESITATE
HIRE
HONEY
HOOK
ILL
INCH
INCOME
INCREASE
INSTANCE
INTENT
INTERVIEW
IRON
JANUARY
JAW
JOURNEY
JOY
JUICE
KEEN
KNEE
KNIFE
KNOWLEDGE
LACK
LAMB
LANE
LAWN
LEAN
LEAP
LEND
LENGTH
LID
LIKELY
LOAN
LOCATE
LOG
LONE
LOOSE
LOSS
LUMP
MAGAZINE
MALE
MARVEL
MATCH
MEAT
MELT
MERRY
MILL
MIRROR
MIX
MOOD
MOON
MOUNT
MOUSE
MUD
NAIL
NANNY
NARROW
NASTY
NAVY
NEITHER
NEST
NEWSPAPER
NOR
NORTHERN
NUT
OAK
OTHERWISE
OVEN
OWE
PARTNER
PAT
PATCH
PATH
PEACE
PEN
PERFORM
PET
PIG
PIN
PINK
PITCH
PITY
PLAIN
PLEASURE
PLUG
POCKET
POISON
POLISH
POLITE
POLLUTE
POOL
POUR
PRAY
PREGNANT
PRESSURE
PRIDE
PRINT
PRODUCE
PROUD
PUB
PUMP
PUNCH
PUNISH
PURCHASE
PURPLE
PURPOSE
RECALL
REFUSE
REGARD
RELEASE
RELIEF
RELY
REMIND
RESEARCH
RICE
RIP
RISK
ROOF
ROOT
ROW
ROYAL
RUB
RUSH
SAKE

SAND
SAUSAGE
SCHEDULE
SCORE
SCRATCH
SCREAM
SCREEN
SCREW
SEAL
SEARCH
SEASON
SEED
SEEK
SEW
SHADE
SHADOW
SHAME
SHARP
SHAVE
SHED
SHEEP
SHEET
SHELF
SHELL
SHELTER
SHIFT
SHINE
SHORE
SHOVE
SHOWER
SIGNAL
SILVER
SINK
SKILL
SKIRT
SLAVE
SLIDE
SMART
SMASH
SMOOTH
SNAKE
SNAP
SOCK
SOIL
SOMEWHAT
SORE
SOUTHERN
SPARE
SPEECH
SPEED
SPELL
SPIN
SPLIT
SPOIL
SPREAD
STAFF
STAIRS
STARVE
STEADY
STEAM
STEEL
STIFF
STORM
STRAWBERRY
STREAM
STRENGTH
STRING
STRIP
STROKE
STRUGGLE
SUCK
SURFACE
SURROUND
SWALLOW
SWEAR
SWEEP
SWING
SWITCH
SWORD
TALE
TEASE
TEENAGE
TENT
THEREFORE
THIEF
THIN
THREAT
THUS
TIDE
TIN
TINY
TIP
TOAST
TOE
TOILET
TONGUE
ROT
TOOL
TOUGH
TOWEL
TOWER
TOY
TRACE
TRADE
TRIAL
TRICK
TRUCK
TUNE
TWIN
TWIST
UPPER
UPSET
VILLAGE
WAGE
WANDER
WARN
WEAK
WEAPON
WEED
WEIRD
WELCOME
WESTERN
WHEREAS
WHIP
WHISTLE
WICKED
WING
WIPE
WIRE
WISE
WITNESS
WOLF
WOOL
WOUND
WRAP
YELL

NIVEL 3 (Contiene 486 unidades léxicas)

ACCOMPLISH
ACCURATE
ACHE
ADVOCATE
AIM
AIRCRAFT
ALONGSIDE
AWARD
BAN
BARGAIN
BEAM
BEHAVE
BEHAVIOUR
BELIEF
BENCH
BID
BISHOP
BOOST
BOUNDARY
BUDGET
BURDEN
CHAIRMAN
CHAMBER
CHARITY
CHART
CLUSTER
COIN
CONVEY
CORE
CRAFT
CREW
CROP
CRUSH
DAMP
DEFEAT
DELAY
DEPUTY
DISABLED
DISPLAY
DISPOSE
DRAFT
DRAIN
DRIFT
DRILL
EASTERN
ENABLE
ENSURE
ENTITLE
ESSAY
EVOLVE
FADE
FAILURE
FIERCE
FLAVOUR
FOCUS
FORECAST
FOSTER
FRAMEWORK
FUEL
GAP
GLOW
GOODS
GRATEFUL
GREET
GRIP
GROSS
HAZARD
HEAL

HEEL
HIGHLIGHT
HINT
HIP
HOLY
HOST
HOUSEHOLD
INHERIT
INPUT
INVEST
JAIL
JOINT
LABEL
LANDSCAPE
LAUGHTER
LAUNCH
LAYER
LEAK
LIABLE
LIKENESS
LINK
LODGE
LOYAL
MEANWHILE
MERGE
MILD
MIXTURE
MORTGAGE
MOTION
NAKED
NEGLECT
NETWORK
NEVERTHELESS
NOD
OUTCOME
OUTRAGE
OVERALL
OVERWHELM
PAD
PARISH
PAVE
PEAK
PEER
PIT
PLOT
POLL
PORTRAIT
PRACTITIONER
PRAISE
PRIEST
PROFIT
PROOF
PUZZLE
RAIL
RANDOM
RAW
REAR
RECESSION
REGRET
REJECT
RELUCTANT
REQUEST
RESCUE
RESEMBLE
RESIGN
RESOURCE
RETAIL
REWARD
SAMPLE
SCAN
SCHEME
SCOPE
SHIELD
SILK
SLICE
SOLE
SOLVE
SOURCE
SPILL
SQUEEZE
STAIN
STAKE
STRAIN
SUBTLE
SURGERY
SURVEY
SWEAT
TACKLE
TASK
TENDER
THOROUGH
THRILL
TISSUE
TOSS
TRAIL
TREASURE
TREND
TRIGGER
UPDATE
URGE
UTILITY
WEALTH
WEIGH
WELFARE
WHISPER
WITHDRAW
YOUTH
AERIAL
AIRPORT
ALLEY
ALMIGHTY
APPAL
APPLIANCE
ARREARS
ASTONISH
BADGE
BARN
BASEMENT
BASH
BEHALF
BLADE
BLIMEY
BLOKE
BOG
BOLD
BOLT
BOROUGH
BOYFRIEND
BRACKET
BRASS
BREAKDOWN
BUGGER
BULK
BULL
BULLY
BURGLE
BUST
BUTCHER
BUZZ
BYPASS
CANDLE
CARDBOARD
CAROL
CATER
CHAP
CHAPEL
CHEMIST
CHUCK
CHUNK
CLAP
CLASH
CLERK
CLUTCH
COCK
CONGRATULATE
CRACKER
CRAMP
CRAP
CRIPPLE
CUPBOARD
CUSHION
DAFT
DASH
DEAF
DEED
DERBY
DETACH
DEVIL
DIAL
DIP
DISGUISE
DISTRESS
DITCH
DOCK
DODGY
DOLE
DOORSTEP
DOORWAY
DOT
DRAUGHT
DREAD
DREADFUL
DRIP
DROWN
DULL
DUMMY
DUTCH
DYE
EASTER
ELBOW
ELDEST
ENQUIRE
ESCORT
EVE
FARE
FEEDBACK
FIDDLE
FINED
FIR
FLAP
FLARE
FLATTING
FOLDER
FOND
FOREVER
FORK
FRINGE
FURNISH
FUSS
GAMBLE
GEE
GLAZE
GLOVE
GLUE
GOSPEL
GREED
GRIEF
GRIND
GUT
HAIRDRESSER
HAM
HAMMER
HANDICAP
HARASS

HARVEST
HASSLE
HEADMASTER
HECK
HEN
HOMEWORK
HOOD
HOORAY
HOP
HOUSEWIFE
HUT
IDLE
INN
JAR
JEWEL
JOG
JOLLY
JUG
JUNCTION
KETTLE
KIDDING
KIDNEY
KIT
KNOT
LAD
LADDER
LANDLORD
LAP
LAYOUT
LEAF
LEAFLET
LEISURE
LEVER
LITTER
LOFT
LOO
LOOP
LORRY
LOUNGE
LUNCHTIME
LUNG
MANOR
MAR
MEANTIME
MEND
MERCHANT
MIRACLE
MOAN
MOCK
MOOR
MOTORBIKE
MOTORWAY
MUCK
MUG
NAIVE
NAY
NEEDLE
NICK
NIECE
NIGHTMARE
NIL
NIP
NONSENSE
NOUGHT
NUISANCE
NURSERY
ONWARDS
PADDY
PAL
PENCIL
PIER
PINT
PLASTER
PLONK
PLOUGH
PLUMB
POKE
POND
POSH
POSTCARD
PREACH
PURSE
QUERY
QUEUE
QUID
RACK
RAG
RATTLE
RECEIPT
REHEARSE
REVOLT
RIBBON
ROCKET
ROVE
RUBBER
SCOTLAND
SCOUT
SCRAP
SCRAPE
SCRIBBLE
SCRIPT
SCRUB
SELDOM
SHEER
SHOVEL
SIXPENCE
SKETCH
SKIP
SLAP
SLASH
SLEEVE
SLIM
SLOT
SMACK
SNEAK
SNIFF
SOAK
SOD
SOLICITOR
SPARK
SPIT
ABROAD
SPITE
SPLASH
SPOON
SQUASH
STAB
STACK
STALL
STEEP
STEER
STICKY
STINK
STITCH
STRAIGHTFORWARD
STRAP
STRAW
STRIDE
STRIPE
SUE
SUPPER
SURGEON
SURNAME
SWAN
SWAP
TAB
TACK
TAG
TELECOM
TEMPER
THUMB
THUNDER
TICK
TIDY
TILE
TILLED
TIMBER
TIMETABLE
TOKEN
TORCH
TREAD
TROLLEY
TROUSERS
TUMBLE
TYRE
UN
UNDERGROUND
UNDERLINE
UPWARDS
VAT
VEST
VET
WALES
WALLET
WALLPAPER
WARD
WARDROBE
WAREHOUSE
WARRANT
WEE
WHACK
WHATSOEVER
WHEELCHAIR
WHEREABOUTS
WHEREBY
WHOOP
WIDOW
WIG
WOBBLE
WORKSHOP
WORTHWHILE
WRECK
WRESTLE
WRIST

NIVEL 4 (Contiene 170 falsos cognados)

ABLE
ACHIEVE
ACTUAL
ADVERTISE
ADVICE
AGO
ALL
ALLOCATE
ALSO
AMAZE
ANCIENT
APOLOGY
APPLY
APPOINT
ARGUE
ARM
ARMY
ARROW
ASSESS
ASSIST
ATTEMPT
ATTEND
AUNT
BALL
BARK
BEEF
BEST
BLANK
BLIND
BODY
BORROW
BOTTOM
BRAVE
CAN
CAP
CAPABLE
CARD
CARE
CARPET
CART
CASUALTY
CHOKE
CHOOSE
COLLAR
COLLEGE
COME
COMPLAIN
COMPLIMENT
COMPREHENSIVE
CONFIDENCE
CONTEST
COUNCIL
DAMN
DATE
DELIGHT
DELIVER
DESIGN
DESPITE
DINNER
DIRECTION
DISGRACE
DISGUST
EAT
ELEVEN
EMBARRASS
ESTATE
EXCITE
EXIT
EXPECT
EXPENSE
EXPLAIN
FACILITY
FACTORY
FAULT
FEE
FINE
FLOOR
FLU
FOR
FRIEND
FUME
GANG
GENTLE
GRAB
GRAND
HARDLY
HAY
IMPROVE
INJURE
INTRODUCE
INVOLVE
JAM
JOURNALIST
JUST
LARGE
LECTURE
LIBRARY
LIE
LUXURY
MAN
MATE
MATTER
MEASURE
MURDER
NOTICE
ONCE
PAIN
PALE
PAN
PARADE
PARCEL
PARENT
PARTY
PATTERN
PETROL
PIE
PIGEON
PINCH
PLACE
POINT
POWDER
PRESUME
PREVENT
PRIZE
QUESTION
QUIET
QUITE
QUOTE
RANGE
RAPE
REALISE
RECIPE
RECKON
RECORD
RED
RELIEVE
REMAIN
REMARK
RESORT
RESUME
RETIRE
ROPE
SALE
SAVE
SENIOR
SENSIBLE
SENSITIVE
SEVERAL
SIGNATURE
SIN
SON
SPECTACLE
STAY
STIR
STRANGE
STRETCH
SUCCEED
SUPPLY
SYMPATHY
TABLE
TALL
TARGET
TAX
TERRIFIC
TIRE
TO
TRANSLATE
TREE
VACUUM
VICIOUS

APÉNDICE C. KEYWORDS EN LA GSL Y LA AWL

684 KEYWORDS EN LAS PRIMERAS 1.000 GSL

A
ABLE
ABOUT
ABOVE
ACCORD
ACCOUNT
ACROSS
ACTUAL
ADD
ADDRESS
AFFAIR
AFTER
AGAIN
AGAINST
AGE
AGO
AGREE
ALL
ALLOW
ALMOST
ALONE
ALONG
ALREADY
ALSO
ALTHOUGH
ALWAYS
AMONG
AMOUNT
ANCIENT
AND
ANOTHER
ANSWER
ANY
APPLY
APPOINT
ARM
ARMY
AROUND
AS
ASK
AT
ATTEMPT
AVERAGE
AWAY
BACK
BAD
BALL
BATTLE
BE
BEAR
BEAUTY
BECAUSE
BECOME
BED
BEFORE
BEGIN
BEHIND
BELIEVE
BELONG
BELOW
BENEATH
BESIDE
BEST
BETWEEN
BEYOND
BIG
BILL
BIRD
BLACK
BLOOD
BLOW
BLUE
BOARD
BODY
BOOK
BOTH
BOX
BOY
BRANCH
BREAD
BREAK
BRIDGE
BRIGHT
BRING
BROAD
BROTHER
BUILD
BURN
BUSINESS
BUT
BUY
BY
CALL
CAN
CARE
CARRY
CATCH
CERTAIN
CHANCE
CHANGE
CHARGE
CHIEF
CHILD
CHOOSE
CHURCH
CITY
CLAIM
CLOSE
CLOUD
COAL
COIN
COLD
COLLEGE
COME
CONCERN
COTTON
COULD
COUNCIL
COUNTRY
CROWD
CROWN
CRY
CUT
DANGER
DARK
DATE
DAUGHTER
DEAD
DEAL
DEAR
DEEP
DEFEAT
DEGREE
DESIRE
DETAIL
DEVELOP
DIE
DIRECT
DO
DOG
DOOR
DOUBT
DOWN
DRAW
DREAM
DRESS
DRINK
DRIVE
DROP
DRY
DUE
DUTY
EACH
EAR
EARLY
EARTH
EAST
EASY
EAT
EGG
EIGHT
EITHER
ELEVEN
ELSE
END
ENJOY
ENOUGH
EVEN
EVENING
EVER
EVERY
EXCHANGE
EXPECT
EXPENSE
EXPLAIN
EXPRESS
EYE
FACE
FACT
FACTORY
FAIL
FAIR
FAITH
FALL
FAR
FARM
FAST
FATHER

FEAR
FEEL
FELLOW
FEW
FIELD
FIGHT
FILL
FIND
FINE
FINISH
FIRE
FIRST
FISH
FIT
FIVE
FIX
FLOOR
FLOW
FLY
FOLLOW
FOOD
FOR
FOREIGN
FOREST
FORGET
FORTH
FOUR
FREE
FRESH
FRIDAY
FRIEND
FROM
FULL
FURNISH
GAIN
GAME
GARDEN
GATE
GATHER
GENTLE
GET
GIFT
GIRL
GIVE
GLAD
GLASS
GO
GOD
GOOD
GREAT
GREEN
GROUND
GROW
HALF
HAND
HANG
HAPPEN
HAPPY
HARD
HARDLY
HAVE
HE
HEAD
HEAR
HEART
HEAT
HEAVEN
HEAVY
HELP
HERE
HIGH
HILL
HOLD
HOME
HOPE
HORSE
HOT
HOUSE
HOW
HOWEVER
HUNDRED
HUSBAND
I
IF
ILL
INCH
INCREASE
INDEED
INSTEAD
INTO
INTRODUCE
IRON
IT
JANUARY
JOIN
JOINT
JOY
JUST
KEEP
KILL
KIND
KING
KNOW
LACK
LADY
LAKE
LAND
LARGE
LAST
LATE
LAUGH
LAUGHTER
LAW
LAY
LEAD
LEARN
LEAVE
LEFT
LENGTH
LESS
LET
LEVEL
LIBRARY
LIE
LIFE
LIFT
LIGHT
LIKE
LIKELY
LIP
LISTEN
LITTLE
LIVE
LONG
LOOK
LORD
LOSE
LOSS
LOVE
LOW
MACHINE
MAIN
MAKE
MAN
MANY
MARKET
MARRY
MATTER
MAY
MEAN
MEASURE
MEET
MIDDLE
MIGHT
MILE
MILK
MIND
MISS
MONDAY
MONEY
MONTH
MOON
MORE
MORNING
MOST
MOTHER
MOUTH
MRS
MUST
NAME
NEAR
NEED
NEIGHBOUR
NEITHER
NEVER
NEW
NEWS
NEWSPAPER
NEXT
NIGHT
NINE
NONE
NOR
NORTH
NOTICE
NOW
OF
OFF
OFTEN
OIL
OLD
ONCE
ONE
ONLY
OPEN
OTHER
OTHERWISE
OUGHT
OUT
OVER
OWE
OWN
PARTY
PAY
PEACE
PEOPLE
PERHAPS
PICTURE
PLACE
PLAIN
PLAY
PLEASE
POINT
POOR
POUND
PRESSURE
PRETTY
PREVENT
PRODUCE
PROFIT
PROOF
PULL
PURPOSE
PUT

QUEEN
QUESTION
QUITE
RAISE
RATE
RATHER
REACH
READ
READY
REALISE
RECEIPT
RECORD
RED
REFUSE
REGARD
REMAIN
REMARK
REMEMBER
REPLY
RICH
RIDE
RIGHT
RING
RISE
RIVER
ROAD
ROLL
ROOM
ROUGH
ROUND
ROYAL
RULE
RUN
SAFE
SAIL
SALE
SAME
SATURDAY
SAVE
SAY
SCHOOL
SEA
SEASON
SEAT
SEE
SEEM
SELL
SEND
SENSE
SENSITIVE
SET
SETTLE
SEVEN
SEVERAL
SHADOW
SHAKE
SHALL
SHAPE
SHARE
SHE
SHINE
SHIP
SHOOT
SHORE
SHORT
SHOULD
SHOULDER
SHOW
SIDE
SIGHT
SIGN
SILVER
SINCE
SING
SINGLE
SIR
SISTER
SIT
SIX
SIZE
SKY
SLEEP
SMALL
SMILE
SNOW
SO
SOFT
SOME
SON
SOON
SORT
SOUND
SOUTH
SPEAK
SPEED
SPEND
SPITE
SPOT
SPREAD
SPRING
SQUARE
STAGE
STAND
STAR
START
STAY
STEEL
STEP
STILL
STONE
STOP
STORE
STRANGE
STREAM
STREET
STRENGTH
STRIKE
STRONG
STRUGGLE
SUCCEED
SUCH
SUMMER
SUN
SUNDAY
SUPPLY
SURE
SURFACE
SURROUND
SWEET
SWORD
TABLE
TAKE
TALK
TAX
TEACH
TEAR
TELL
TEN
TERM
THAN
THE
THEN
THERE
THEREFORE
THEY
THING
THINK
THIRTEEN
THIRTY
THIS
THOUGH
THOUSAND
THREE
THROUGH
THROW
THURSDAY
THUS
TIME
TO
TODAY
TOGETHER
TOO
TOP
TOUCH
TOWARD
TOWN
TRADE
TRAIN
TRAVEL
TREE
TROUBLE
TRUE
TRUST
TRY
TUESDAY
TURN
TWELVE
TWENTY
TWO
UNDER
UNDERSTAND
UNLESS
UNTIL
UP
UPON
VERY
VIEW
VILLAGE
WAGE
WAIT
WALK
WALL
WANT
WAR
WATCH
WATER
WAVE
WAY
WE
WEALTH
WEAR
WEDNESDAY
WEEK
WELCOME
WELL
WEST
WESTERN
WHAT
WHEN
WHERE
WHETHER
WHICH
WHILE
WHITE
WHO
WHOLE
WHY
WIDE
WIFE
WILD
WILL
WIN
WIND
WINDOW

WINTER
WISE
WISH
WITH
WITHIN
WITHOUT
WOMAN
WONDER
WOOD
WORD
WORK
WORLD
WORTH
WOULD
WOUND
WRITE
WRONG
YEAR
YES
YESTERDAY
YET
YOU
YOUNG
YOUTH

547 KEYWORDS EN LAS SEGUNDAS 1.000 GSL

ABROAD
ACHE
ADVERTISE
ADVICE
AFFORD
AFRAID
AFTERNOON
AHEAD
AIM
ALIVE
ALTOGETHER
AMUSE
ANGER
ANNOY
APOLOGY
APPLE
ARGUE
ARRANGE
ARROW
ASHAMED
ASIDE
ASLEEP
ASTONISH
ATTEND
AUNT
AUTUMN
AVOID
AWAKE
AWKWARD
BAG
BAKE
BARE
BARGAIN
BASKET
BATH
BEAM
BEAN
BEAT
BEG
BEHAVE
BEHAVIOUR
BELL
BELT
BEND
BIND
BIRTH
BIT
BITE
BITTER
BLADE
BLAME
BLESS
BLIND
BOIL
BOLD
BONE
BORROW
BOTTOM
BOUNDARY
BOW
BOWL
BRAIN
BRASS
BRAVE
BREAKFAST
BREATH
BREATHE
BRICK
BROWN
BRUSH
BUCKET
BUNCH
BURST
BURY
BUSH
BUSY
BUTTER
CAP
CARD
CART
CAT
CHAIN
CHAIR
CHARM
CHEAP
CHEAT
CHEER
CHEESE
CHEST
CHICKEN
CHRISTMAS
CLEAN
CLERK
CLEVER
CLIFF
CLIMB
CLOCK
CLOTH
COAT
COLLAR
COMPLAIN
CONFIDENCE
CONGRATULATE
COOK
COOL
CORNER
COTTAGE
COUGH
COUSIN
COW
CRACK
CRASH
CREEP
CROP
CRUSH
CUPBOARDS
CURL
CUSHION
CUSTOMER
DAMAGE
DAMP
DARE
DEAF
DEBT
DEED
DELAY
DELIGHT
DELIVER
DESERVE
DESK
DEVIL
DIG
DINNER
DIP
DIRT
DISAPPOINT
DISEASE
DISGUST
DISH
DITCH
DIVE
DOT
DRAG
DRAWER
DROWN
DRUM
DUCK
DULL
DUST
EARN
EASE
EDGE
ELDER
EMPTY
ENCOURAGE
ENGINE
ENVELOPE
EVIL
EXCITE
FADE
FAINT
FANCY
FASHION
FAT
FAULT
FEATHER
FEMALE
FENCE
FIERCE
FINGER
FLAG
FLAME
FLAT
FLAVOUR
FLOOD
FOLD
FOND
FOOL
FOOT
FORGIVE
FORK
FORWARD
FRAME
FREEZE

FRIGHT
FRY
FUN
FUR
GAP
GRAND
GRASS
GRATEFUL
GREED
GREET
GREY
GRIND
GUESS
GUEST
GUILTY
GUN
HAIR
HALL
HAMMER
HANDLE
HARM
HARVEST
HAT
HATE
HAY
HEAL
HEALTH
HESITATE
HIDE
HIRE
HIT
HOLE
HOLIDAY
HOLY
HOOK
HOST
HULLO
HUNGER
HUNT
HURRY
HURT
HUT
ICE
IDLE
IMPROVE
INN
INQUIRE
INSIDE
INSURE
INTEND
JAW
JEWEL
JOKE
JOURNEY
JUICE
JUMP
KEY
KISS
KITCHEN
KNEE
KNIFE
KNOCK
KNOT
LADDER
LAZY
LEAF
LEAN
LEG
LEND
LID
LOAD
LOAN
LOCK
LOG
LONE
LOOSE
LOT
LOUD
LOYAL
LUCK
LUMP
LUNG
MAD
MALE
MANAGE
MATCH
MEAL
MEANTIME
MEANWHILE
MEAT
MELT
MEND
MERCHANT
MERRY
MILD
MILL
MISTAKE
MIX
MOTION
MOUSE
MUD
MURDER
NAIL
NARROW
NEAT
NECK
NEEDLE
NEGLECT
NEST
NICE
NIECE
NOISE
NONSENSE
NOSE
NUISANCE
NURSE
NUT
ONWARDS
ORANGE
PAD
PAIN
PALE
PAN
PARCEL
PARENT
PATH
PATTERN
PEN
PENCIL
PENNY
PERFORM
PET
PICK
PIG
PIGEON
PIN
PINCH
PINK
PINT
PITY
PLASTER
PLENTY
PLOUGH
POCKET
POISON
POLISH
POLITE
POOL
POT
POUR
POWDER
PRAISE
PRAY
PREACH
PRIDE
PRIEST
PRINT
PRIZE
PROUD
PUMP
PUNISH
PURPLE
PUSH
PUZZLE
QUICK
QUIET
RABBIT
RAIL
RAIN
RAW
REGRET
RELIEVE
REMIND
REQUEST
RESCUE
RESIGN
RETIRE
REWARD
RIBBON
RICE
RID
RISK
ROOF
ROOT
ROPE
ROT
ROW
RUB
RUBBER
RUBBISH
RUSH
SAD
SAKE
SAMPLE
SAND
SCRAPE
SCRATCH
SCREEN
SCREW
SEARCH
SEED
SELDOM
SELF
SEW
SHADE
SHAME
SHARP
SHAVE
SHEEP
SHEET
SHELF
SHELL
SHELTER
SHIELD
SHIRT
SHOE
SHOP
SHOUT
SHOWER
SHUT
SICK
SIGNAL
SILK
SINK

SKILL
SKIN
SKIRT
SLAVE
SLIDE
SLIGHT
SLIP
SLOW
SMELL
SMOKE
SMOOTH
SNAKE
SOCK
SOIL
SOLVE
SORE
SORRY
SPARE
SPELL
SPILL
SPIN
SPIT
SPLIT
SPOIL
SPOON
STAFF
STAIN
STAIRS
STEADY
STEAL
STEAM
STEEP
STEER
STICK
STIFF
STIR
STORM
STRAIGHT
STRAP
STRAW
STRETCH
STRING
STRIP
STRIPE
STUFF
SUCK
SUDDEN
SUIT
SUPPER
SWALLOW
SWEAR
SWEAT
SWEEP
SWIM
SWING
SYMPATHY
TAIL
TALL
TASTE
TEMPER
TEND
TENDER
TENT
THANK
THICK
THIEF
THIN
THOROUGH
THREAT
THROAT
THUMB
THUNDER
TIDE
TIDY
TIE
TIGHT
TIN
TIP
TIRE
TOE
TOMORROW
TONGUE
TONIGHT
TOOL
TOOTH
TOUGH
TOWEL
TOWER
TOY
TRACK
TRANSLATE
TREASURE
TREAT
TRICK
TRIP
TUNE
TWIST
UGLY
UNCLE
UPPER
UPSET
UPWARDS
URGE
WAKE
WANDER
WARM
WARN
WASH
WASTE
WEAK
WEAPON
WEATHER
WEED
WEIGH
WET
WHEEL
WHIP
WHISPER
WHISTLE
WICKED
WIDOW
WINE
WING
WIPE
WIRE
WITNESS
WOOL
WORRY
WORSE
WRAP
WRECK
WRIST
YELLOW

96 KEYWORDS EN LA AWL

ACCURATE
ACHIEVE
ADVOCATE
AID
ALLOCATE
APPROACH
ASSESS
ASSIST
ASSURE
ATTACH
AVAILABLE
AWARE
BEHALF
BOND
BRIEF
BULK
CAPABLE
CHALLENGE
CHANNEL
CHAPTER
CHART
COMPREHENSIVE
CORE
COUPLE
DENY
DESIGN
DESPITE
DISPLAY
DISPOSE
DRAFT
ENABLE
ENSURE
ENVIRONMENT
ESTATE
EVOLVE
EXPOSE
FACILITATE
FEATURE
FEE
FOCUS
FRAMEWORK
GRANT
HIGHLIGHT
INCOME
INJURE
INPUT
INSTANCE
INVEST
INVOLVE
ISSUE
JOB
LABEL
LAYER
LECTURE
LINK
LOCATE
NETWORK
NEVERTHELESS
ODD
OUTCOME
OVERALL
PARTNER
PLUS
PRACTITIONER
PRESUME
PURCHASE
QUOTE
RANDOM

RANGE
REJECT
RELEASE
RELUCTANCE
RELY
RESEARCH
RESOURCE
SCHEDULE
SCHEME
SCOPE
SEEK
SHIFT
SOLE
SOMEWHAT
SOURCE
STRAIGHTFORWARD
SURVEY
TAPE
TARGET
TASK
TEAM
TRACE
TREND
TRIGGER
UTILISE
WELFARE
WHEREAS
WHEREBY

473 KEYWORDS QUE NO EN APARECEN EN LAS LISTAS GSL / AWL

ACCOMPLISH
AERIAL
AIRCRAFT
AIRPORT
ALLEY
ALMIGHTY
ALONGSIDE
ALRIGHT
AMAZE
APPAL
APPEAL
APPLIANCE
ARREARS
AWARD
AWFUL
BADGE
BAN
BANG
BARK
BARN
BASEMENT
BASH
BEACH
BEEF
BEER
BENCH
BET
BID
BIKE
BIN
BISCUIT
BISHOP
BLANK
BLANKET
BLIMEY
BLOKE
BLOOM
BOG
BOLT
BOOM
BOOST
BORING
BOROUGH
BOSS
BOTHER
BOUNCE
BOYFRIEND
BRACKET
BRAKE
BRAND
BREAKDOWN
BREAST
BREED
BUCK
BUDGET
BUG
BUGGER
BULL
BULLY
BUMP
BURDEN
BURGLE
BUST
BUTCHER
BUZZ
BYPASS
CANDLE
CARDBOARD
CAROL
CARPET
CARROT
CASH
CAST
CASUALTY
CATER
CEILING
CHAMBER
CHAP
CHAPEL
CHARITY
CHASE
CHEEK
CHEMIST
CHEW
CHOKE
CHOP
CHUCK
CHUNK
CLAP
CLASH
CLOTHE
CLUE
CLUSTER
CLUTCH
COACH
COCK
COMPLIMENT
CONTEST
CONVEY
COP
COPE
COUNTER
COUNTY
CRACKER
CRAFT
CRAMP
CRAP
CRAWL
CRAZY
CREW
CRIPPLE
CRISP
DAFT
DAMN
DARLING
DASH
DEPUTY
DERBY
DETACH
DIAL
DISABLED
DISGRACE
DISGUISE
DISTRESS
DOCK
DODGY
DOLE
DOORSTEP
DOORWAY
DRAIN
DRAUGHT
DREAD
DREADFUL
DRIFT
DRILL
DRIP
DUMMY
DUMP
DUTCH
DYE
EASTER
ELBOW
ELDEST
EMBARRASS
ENGAGE
ENTITLE
ESCORT
ESSAY
EVE
EXHAUST
EXIT
FARE
FEED
FEEDBACK
FETCH
FIDDLE
FINED
FIR
FLAP
FLARE
FLATTING
FLU
FOLDER
FOLK

FORECAST
FOREVER
FORTNIGHT
FOSTER
FRINGE
FROG
FROST
FUEL
FUME
FUSS
GAMBLE
GANG
GEAR
GEE
GHOST
GLANCE
GLAZE
GLOVE
GLOW
GLUE
GOODS
GOSPEL
GRAB
GRIEF
GRIP
GROSS
GUT
GUY
HAIRDRESSER
HAM
HANDICAP
HANDY
HARASS
HASSLE
HAZARD
HEADMASTER
HECK
HEDGE
HEEL
HEIGHT
HELL
HEN
HINT
HIP
HOMEWORK
HONEY
HOOD
HOORAY
HOP
HOUSEHOLD
HOUSEWIFE
HUGE
INHERIT
INTERVIEW
JAIL
JAM
JAR
JOG
JOLLY
JOURNALIST
JUG
JUNCTION
KEEN
KETTLE
KID
KIDDING
KIDNEY
KIT
LAD
LAMB
LANDLORD
LANDSCAPE
LANE
LAP
LAUNCH
LAWN
LAYOUT
LEAFLET
LEAK
LEAP
LEISURE
LEVER
LIABLE
LIKENESS
LITTER
LODGE
LOFT
LOO
LOOP
LORRY
LOUNGE
LUNCHTIME
LUXURY
MAGAZINE
MANOR
MAR
MARVEL
MATE
MATH
MERGE
MESS
MIRACLE
MIRROR
MOAN
MOCK
MOOD
MOOR
MORTGAGE
MOTORBIKE
MOTORWAY
MOUNT
MOVIE
MUCK
MUG
NAIVE
NAKED
NANNY
NASTY
NAUGHTY
NAVY
NAY
NICK
NIGHTMARE
NIL
NIP
NOD
NOUGHT
NOWT
NURSERY
OAK
OUTRAGE
OVEN
OVERWHELM
PADDY
PAL
PARADE
PARISH
PAT
PATCH
PAVE
PEAK
PEER
PENCE
PETROL
PIE
PIER
PIT
PITCH
PLONK
PLOT
PLUG
PLUMB
POKE
POLL
POLLUTE
POND
PORTRAIT
POSH
POSTCARD
PREGNANT
PUB
PUNCH
PURSE
QUERY
QUEUE
QUID
RACK
RAG
RAPE
RATTLE
REAR
RECALL
RECESSION
RECIPE
RECKON
REHEARSE
RESEMBLE
RESORT
RESUME
RETAIL
REVOLT
RIP
ROCKET
ROVE
SAUSAGE
SCAN
SCARE
SCORE
SCOTLAND
SCOUT
SCRAP
SCREAM
SCRIBBLE
SCRIPT
SCRUB
SEAL
SENIOR
SHED
SHEER
SHOVE
SHOVEL
SHY
SILLY
SIN
SIXPENCE
SKETCH
SKIP
SLAP
SLASH
SLEEVE
SLICE
SLIM
SLOT
SMACK
SMART
SMASH
SNAP
SNEAK
SNIFF
SOAK
SOD
SOLICITOR
SPARK
SPECTACLE

SPLASH
SQUASH
SQUEEZE
STAB
STACK
STAKE
STALL
STARE
STARVE
STICKY
STINK
STITCH
STRAIN
STRAWBERRY
STRIDE
SUBTLE
SUE
SURGEON
SURGERY
SURNAME
SWAN
SWAP
SWITCH
TAB
TACK
TACKLE
TAG
TALE
TEASE
TEENAGE
TELECOM
TERRIFIC
THRILL
TICK
TILE
TILLED
TIMBER
TIMETABLE
TINY
TISSUE
TOAST
TOILET
TOKEN
TORCH
TOSS
TRAIL
TREAD
TROLLEY
TROUSERS
TRUCK
TUMBLE
TWIN
TYRE
UN
UNDERGROUND
UNDERLINE
UPDATE
VACUUM
VAT
VEST
VET
VICIOUS
WALES
WALLET
WALLPAPER
WARD
WARDROBE
WAREHOUSE
WARRANT
WED
WEE
WEIRD
WHACK
WHATSOEVER
WHEELCHAIR
WHEREABOUTS
WHOOP
WIG
WITHDRAW
WOBBLE
WOLF
WORKSHOP
WORTHWHILE
WRESTLE
YELL

APÉNDICE D. ANÁLISIS DEL CORPUS DE EXÁMENES

ANÁLISIS BNC

ANÁLISIS PLH

ANGLIA – CORPUS B1

FILE	TOKEN	TOKEN%	CUMTOKEN%	FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	43	4.53	4.53	KEYWORDS	763	80.32	80.32
BNC-1	809	85.16	89.69	COGNATES	141	14.84	95.16
BNC-2	57	6.00	95.69	INVARIABLES	43	4.53	99.69
BNC-3	14	1.47	97.16	-	3	0.32	100.01
BNC-4	9	0.95	98.11				
BNC-5	1	0.11	98.22				
BNC-6	3	0.32	98.54				
-	14	1.47	100.01				

ESCUELA OFICIAL DE IDIOMAS – CORPUS B1

FILE	TOKEN	TOKEN%	CUMTOKEN%	FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	27	1.79	1.79	KEYWORDS	1084	71.88	71.88
BNC-1	1162	77.06	78.85	COGNATES	359	23.81	95.69
BNC-2	159	10.54	89.39	INVARIABLES	27	1.79	97.48
BNC-3	85	5.64	95.03	-	38	2.52	100.00
BNC-4	27	1.79	96.82				
BNC-5	8	0.53	97.35				
BNC-6	11	0.73	98.08				
BNC-7	7	0.46	98.54				
BNC-8	5	0.33	98.87				
BNC-9	1	0.07	98.94				
-	16	1.06	100				

CAMBRIDGE ESOL – CORPUS B1

FILE	TOKEN	TOKEN%	CUMTOKEN%	FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	56	4.03	4.03	KEYWORDS	1092	78.67	78.67
BNC-1	1192	85.88	89.91	COGNATES	235	16.93	95.60
BNC-2	108	7.78	97.69	INVARIABLES	56	4.03	99.63
BNC-3	15	1.08	98.77	-	5	0.36	99.99
BNC-4	6	0.43	99.2				
BNC-5	2	0.14	99.34				
BNC-6	1	0.07	99.41				
-	8	0.58	99.99				

TRINITY – CORPUS B1

FILE	TOKEN	TOKEN%	CUMTOKEN%	FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	23	3.81	3.81	KEYWORDS	448	74.17	74.17
BNC-1	485	80.30	84.11	COGNATES	123	20.36	94.53
BNC-2	47	7.78	91.89	INVARIABLES	23	3.81	98.34
BNC-3	32	5.30	97.19	-	10	1.66	100.00
BNC-4	1	0.17	97.36				
BNC-5	3	0.50	97.86				
BNC-6	2	0.33	98.19				
-	11	1.82	100.01				

ANÁLISIS BNC **ANÁLISIS PLH**

ANGLIA – CORPUS B2

FILE	TOKEN	TOKEN%	CUMTOKEN%	FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	47	4.30	4.30	KEYWORDS	804	73.63	73.63
BNC-1	837	76.65	80.95	COGNATES	204	18.68	92.31
BNC-2	94	8.61	89.56	INVARIABLES	47	4.30	96.61
BNC-3	40	3.66	93.22	-	37	3.39	100.00
BNC-4	22	2.01	95.23				
BNC-5	10	0.92	96.15				
BNC-6	18	1.65	97.80				
BNC-7	6	0.55	98.35				
BNC-8	1	0.09	98.44				
BNC-9	5	0.46	98.90				
-	12	1.10	100.00				

ESCUELA OFICIAL DE IDIOMAS – CORPUS B2

FILE	TOKEN	TOKEN%	CUMTOKEN%	FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	90	4.94	4.94	KEYWORDS	1218	66.81	66.81
BNC-1	1274	69.88	74.82	COGNATES	416	22.82	89.63
BNC-2	206	11.30	86.12	INVARIABLES	90	4.94	94.57
BNC-3	110	6.03	92.15	-	99	5.43	100.00
BNC-4	47	2.58	94.73				
BNC-5	21	1.15	95.88				
BNC-6	20	1.10	96.98				
BNC-7	14	0.77	97.75				
BNC-8	5	0.27	98.02				
BNC-9	10	0.55	98.57				
-	26	1.43	100.00				

CAMBRIDGE ESOL – CORPUS B2

FILE	TOKEN	TOKEN%	CUMTOKEN%	FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	42	2.45	2.45	KEYWORDS	1401	81.69	81.69
BNC-1	1526	88.98	91.43	COGNATES	259	15.10	96.79
BNC-2	101	5.89	97.32	INVARIABLES	42	2.45	99.24
BNC-3	22	1.28	98.60	-	13	0.76	100.00
BNC-4	11	0.64	99.24				
BNC-5	3	0.17	99.41				
BNC-6	3	0.17	99.58				
BNC-8	1	0.06	99.64				
BNC-9	1	0.06	99.70				
-	5	0.29	99.99				

TRINITY – CORPUS B2

FILE	TOKEN	TOKEN%	CUMTOKEN%	FILE	TOKEN	TOKEN%	CUMTOKEN%
INVARIABLES	17	1.89	1.89	KEYWORDS	693	77.09	77.09
BNC-1	750	83.43	85.32	COGNATES	179	19.91	97.00
BNC-2	72	8.01	93.33	INVARIABLES	17	1.89	98.89
BNC-3	38	4.23	97.56	-	10	1.11	100.00
BNC-4	14	1.56	99.12				
BNC-6	3	0.33	99.45				
BNC-7	3	0.33	99.78				

APÉNDICE E. CÓDIGO PHP PARA EXTRAER NOMBRES PROPIOS.

```
     <?php
/*************************************************************
* Extrae nombres propios de un texto mediante pistas sintácticas.
     * Contiene versiones modificadas de de un script de Arturs Sosins
     * bajo licencia BDS:
     * http://ar2rsawseen.users.phpclasses.org/browse/file/35277.html
     *************************************************************/
class nombres_propios
{
    //eliminar símbolos del texto
    private $simbolos = array('/','\\','\'','"',"'",',','. ','<','>','?',';','
:','[',']','{','}','|','=','+','-
','_',')','(','*','&','^','%','$','#','@','!','~','`','.','0','1','2','3','4',
'5','6','7','8','9');
    //Hallar nombres que tienen más de una palabra
    private $titulos = array(
        //Partícula al principio
        "principio" => array("the", "mr", "mrs", "ms", "dr", "mstr", "miss", "
sir"),          //en medio de dos nombres propios
        "medio" => array("of", "the", "and"),
        //llevan punto sin estar a final de frase
        "punto" => array("mr", "mrs", "ms", "dr")
    );
    private $stop_palabras = array();
    //omite símbolos que pueden separar frases
    private $omitir = array(" ", "\n", "\t", "\r", "\r\n");
    //puntuaciones que indican principio de frase y justifican las mayúsculas
    private $puntuacion = array(".", "?", "!", "'", '"');
    //incluye siglas
    private $siglas = false;
    //incluye candidatos a nombres propios que no se pueden garantizar, por
    //ejemplo si la palabra aparece solo al principio de una frase
    private $candidato = false;
    //combina nombres propios formados por más de una palabra
    private $multi_palabras = true;

    /*****************************
    * configuracion
    *****************************/
    public funcion set_conjunciones($arr, $type = "principio"){
        if(!is_array($arr))
        {
            $this->conjunciones[$type] = array(mb_strtolower($arr));
        }
        else
        {
            $nuevo = array();
            foreach($arr as $val)
            {
                $nuevo[] = mb_strtolower($val);
            }
            $this->conjunciones[$type] = $nuevo;
        }
    }

    public funcion set_simbolos($arr){
        if(!is_array($arr))
        {
            $this->simbolos = array(mb_strtolower($arr));
        }
        else
        {
            $nuevo = array();
```

```
            foreach($arr as $val)
            {
                $nuevo[] = mb_strtolower($val);
            }
            $this->simbolos = $nuevo;
        }
    }

    public funcion set_omitir($arr){
        if(!is_array($arr))
        {
            $this->omitir = array(mb_strtolower($arr));
        }
        else
        {
            $nuevo = array();
            foreach($arr as $val)
            {
                $nuevo[] = mb_strtolower($val);
            }
            $this->omitir = $nuevo;
        }
    }

    public funcion set_puntuacion($arr){
        if(!is_array($arr))
        {
            $this->puntuacion = array(mb_strtolower($arr));
        }
        else
        {
            $nuevo = array();
            foreach($arr as $val)
            {
                $nuevo[] = mb_strtolower($val);
            }
            $this->puntuacion = $nuevo;
        }
    }

    public funcion siglas($bool){
        if($bool)
        {
            $this->siglas = true;
        }
        else
        {
            $this->siglas = false;
        }
    }

    public funcion candidato($bool){
        if($bool)
        {
            $this->candidato = true;
        }
        else
        {
            $this->candidato = false;
        }
    }

    public funcion multi_palabras($bool){
        if($bool)
        {
            $this->multi_palabras = true;
        }
```

```
        else
        {
            $this->multi_palabras = false;
        }
    }
    public funcion stop_palabras($arr){
        if(!is_array($arr))
        {
            $this->stop_palabras = array(mb_strtolower($arr));
        }
        else
        {
            $nuevo = array();
            foreach($arr as $val)
            {
                $nuevo[] = mb_strtolower($val);
            }
            $this->stop_palabras = $nuevo;
        }
    }

    /*****************************
    * funciones private
    *****************************/
    private funcion limpiar_texto($texto){
        //cambiar simbolos no necesarios
        $texto = str_cambiar($this->simbolos, " ", $texto);
        //cambiar espacios multiples
        $texto = preg_cambiar('/\s+/', ' ', $texto);
        //dividir texto en palabras
        $keys = explode(" ", $texto);
        return $keys;
    }

    private funcion mb_ucfirst($str, $encoding = "UTF-
8", $lower_str_end = false){
        if (!funcion_existe('mb_ucfirst')) {
            $first_letter = mb_strtoupper(mb_substr($str, 0, 1, $encoding), $e
ncoding);
            $str_end = "";
            if ($lower_str_end) {
                $str_end = mb_strtolower(mb_substr($str, 1, mb_strlen($str, $e
ncoding), $encoding), $encoding);
            }
            else {
                $str_end = mb_substr($str, 1, mb_strlen($str, $encoding), $enc
oding);
            }
            $str = $first_letter . $str_end;
            return $str;
        }
        else
        {
            return mb_ucfirst($str);
        }
    }

    private funcion get_candidatos($palabras, $texto){
        $nombrepropio = array();
        foreach($palabras as $key)
        {
            if(trim($key) != "" && $key == $this-
>mb_ucfirst($key) && !in_array($key, $nombrepropio))
            {
                //siglas
                if($key == mb_strtoupper($key))
                {
```

```
                    if($this->siglas)
                    {
                        $nombrepropio[] = $key;
                    }
                }
                //comprobar mayúsculas
                else
                {
                    $stop = false;
                    if($this->estricto)
                    {
                        foreach($palabras as $val)
                        {
                            if(mb_strtolower($key) == mb_strtolower($val) && $
val != $this->mb_ucfirst($val))
                            {
                                $stop = true;
                                break;
                            }
                        }
                    }
                    //comprobar posición en la frase
                    if(!$stop)
                    {
                        $parts = explode($key, $texto);
                        $stop = false;
                        foreach($parts as $cnt => $part)
                        {
                            if(sizeof($parts) > $cnt + 1)
                            {
                                $cut = 1;
                                while(mb_strlen($part)-
$cut >= 0 && in_array(substr($part, mb_strlen($part)-$cut, 1), $this->omitir))
                                {
                                    $cut++;
                                }
                                if(mb_strlen($part)-
$cut > 0 && !in_array(substr($part, mb_strlen($part)-$cut, 1), $this-
>puntuacion))
                                {
                                    $nombrepropio[] = $key;
                                    $stop = true;
                                }
                                else if(!$this->estricto)
                                {
                                    $cut++;
                                    $str = "";
                                    while(mb_strlen($part)-
$cut >= 0 && (!in_array(substr($part, mb_strlen($part)-$cut, 1), $this-
>omitir) && !in_array(substr($part, mb_strlen($part)-$cut, 1), $this-
>puntuacion)))
                                    {
                                        $str = substr($part, mb_strlen($part)-
$cut, 1).$str;
                                        $cut++;
                                    }
                                    if(in_array(mb_strtolower($str), $this-
>conjunciones["punto"]))
                                    {
                                        $nombrepropio[] = $key;
                                        $stop = true;
                                    }
                                }
                            }
                            if($stop)
                            {
                                break;
```

```
                            }
                        }
                        if(!$stop && $this->candidato)
                        {
                            $nombrepropio[] = $key;
                        }
                    }
                }
            }
        }
        return $nombrepropio;
    }

    private funcion get_multi_palabras($nombrepropio, $palabras){
        $multi = array();
        for($i = 0; $i < sizeof($nombrepropio); $i++)
        {
            $palabra = "";
            for($j = 0; $j < sizeof($palabras); $j++)
            {
                if($nombrepropio[$i] == $palabras[$j])
                {
                    $palabra = $nombrepropio[$i];
                    //comprueba anterior al nombre propio
                    $principio = "";
                    for($k = $j-1; $k >= 0; $k--)
                    {
                        //añade titulos usados en el texto original
                        if(in_array(mb_strtolower($palabras[$k]), $this-
>conjunciones["principio"]))
                        {
                            if($principio != "")
                            {
                                $palabra = $palabras[$k]." ".$principio." ".$p
alabra;
                                $principio = "";
                            }
                            else
                            {
                                $palabra = $palabras[$k]." ".$palabra;
                            }
                        }
                        //comprueba si la conjunción esta en medio
                        else if(in_array(mb_strtolower($palabras[$k]), $this-
>conjunciones["medio"]))
                        {
                            $principio = $palabras[$k]." ".$principio;
                        }
                        //se añade si es otro nombre
                        else if(in_array($palabras[$k], $nombrepropio))
                        {
                            if($principio != "")
                            {
                                $palabra = $palabras[$k]." ".$principio." ".$p
alabra;
                                $principio = "";
                            }
                            else
                            {
                                $palabra = $palabras[$k]." ".$palabra;
                            }
                        }
                        else
                        {
                            break;
                        }
                    }
```

```
                    $end = "";

                    for($k = $j+1; $k < sizeof($palabras); $k++)
                    {
                        if(in_array(mb_strtolower($palabras[$k]), $this-
>conjunciones["medio"]))
                        {
                            $end = $palabras[$k]." ".$end;
                        }
                        else if(in_array($palabras[$k], $nombrepropio))
                        {
                            if($end != "")
                            {
                                $palabra = $palabra." ".$end." ".$palabras[$k]
;
                                $end = "";
                            }
                            else
                            {
                                $palabra = $palabra." ".$palabras[$k];
                            }
                        }
                        else
                        {
                            break;
                        }
                    }
                    if(!in_array($palabra, $multi))
                    {
                        $multi[] = $palabra;
                    }
                }
            }
        }
        return $multi;
    }

    /*****************************
    * Función principal
    *****************************/
    public funcion get($texto){
        $palabras = $this->limpiar_texto($texto);
        $nombrepropio = $this->get_candidatos($palabras, $texto);
        if(!empty($this->stop_palabras))
        {
            $nuevo = array();
            foreach($nombrepropio as $val)
            {
                if(!in_array(mb_strtolower($val), $this->stop_palabras))
                {
                    $nuevo[] = $val;
                }
            }
            $nombrepropio = $nuevo;
        }
        if($this->multi_palabras) ««
        {
            $nombrepropio = $this-
>get_multi_palabras($nombrepropio, $palabras);
        }
        return $nombrepropio;
    }
}
?>
```

APÉNDICE F. PALABRAS NO-PLH QUE APARECEN EN AL MENOS 4 CAPÍTULOS DE SHERLOCK HOLMES.

Primera aparición en el capítulo 1

ACCOMPLISH
ACQUAINT
ACUTE
AMIABLE
ARMCHAIR
AVERSE
AWAIT
BONNET
BRIDEGROOM
BRIM
BUTTONED
CAB
CARRIAGE
CEASE
CHIN
CHUCKLE
CLOAK
COMPELLED
CONCEAL
COUCH
DIM
DISTURB
DRUNKEN
EAGER
EARNEST
ENDEAVOUR
FASTEN
FATHOM
FISTS
FLUSH
FOREFINGER
FOREHEAD
FORGE
FORMER
GAZE
GLEAM
GLIMPSE
GRASP
GRATE
GRIM
HANDSOME
HEARTY
HENCE
HURL
INDEBTED
INQUIRE
LANDLADY
LATTER
LEATHER
LENS
LIMB
LIMP
LINED
MAID
MOIST
MUMBLE
OMINOUS
OVERPOWERING
PACE
PEERING
PERCH
PITY
PLUNGE
PREPOSTEROUS
PROCEEDINGS
PURSUE
QUARREL
QUEST
SAVAGE
SCATTERED
SCENT
SECURE
SEIZE
SHRIEK
SHRUG
SHUTTER
SLIT
SNAKE
SNEER
SPARKLE
STAGGER
STARTLE
STICK
SUMMON
SUSPICION
SWIFT
TANGLE
TINTED
TRIFLE
TRIM
TWEED
VANISH
VEIL
VILE
WEAVE
WHENCE
WHISKER
WHISPER
WIT
YAWN
AMID

Primera aparición en el capítulo 2

ASCERTAIN
BIZARRE
BLAND
BLAZE
CLATTER
CLUMP
COMMONPLACE
CUNNING
DANGLING
DECEIVE
DESPAIR
ENTER
FLASH
FROCK
GAPE
GLOOM
GOLDEN
GROAN
HOLLOW
INK
LANTERN
LAUGHTER
MANTELPIECE
OVERCOAT
PEEP
PIERCE
PROTRUDE
PUFF
REST
ROAR
RUEFUL
SKIRT
STATE
STOUT
SUBDUE
THRUST
TRADESMAN
TUG
USE
WAISTCOAT
WAX
WEARY
WINDING
WRING
WRINKLE
WRITHE
YELL

Primera aparición en el capítulo 3

ABSTRACTED
BROW
BUNDLE
CLANG
DWELL
EXACTED
FATE
FIRM
FORBID
GLARE
HANDKERCHIEF
HOMELY
HUDDLED
LOOM
MOISTURE
PLEDGE
POINT
SHILLING
SIMPLE
SOB
STEPMOTHER
UTTERLY
VENTURE
WINK
WREATH

Primera aparición en el capítulo 4

AROUSED
BEARD
BENEATH
CAMP
CURSE
DECEASED
DEPOSE
DROOPING
FOUL
GARMENT
GRIN
HEIR
HIDEOUS
HIGHROAD
ONTO
RUMMAGED
SHATTERED
SNARL
STROLL
SUBMIT
TUT
USHER
UTTER
WIDESPREAD
YARDS

Primera aparición en el capítulo 5

ATTAIN
CHILL
CLAD
CRUMPLED
FULFIL
HASTE
HOWL
MERCY
PASS
RESEMBLE
SCRAWL
TAP
VAIN

Primera aparición en el capítulo 6

BARRED
DEN
ERRAND
FRANTIC
GOWN
HORRID
LEDGER
MIDST
MUTTER
PILLOW
RUG
SMEAR
SOOTHE
STOOP
TWINKLE
WHITEWASH

Primera aparición en el capítulo 7

BRUTE
GASPED
NOISE
SLIPPER
SLOPE

Primera aparición en el capítulo 8

BRISK
CHEERFUL

Primera aparición en el capítulo 9

LEST
OBEY

APÉNDICE G. INVARIABLES EN SHERLOCK HOLMES.

ABBOTS
ABERDEEN
ADLER
AFFAIRE
AFGHAN
AFGHANISTAN
AGRA
ALBERT
ALDERSGATE
ALDERSHOT
ALEXANDER
ALICE
ALICIA
ALOYSIUS
AMERICA
AMERICAN
AMERICANS
AMOY
ANDERSON
ANDOVER
ANSTRUTHER
APACHE
APACHES
ARABIAN
ARCHERY
ARCHIE
ARIGHT
ARIZONA
ARMITAGE
ARMOUR
ARNSWORTH
ARTHUR
ATKINSON
ATLANTIC
ATTICA
AUCKLAND
AUGUSTINE
AUSTRALIA
AUSTRALIAN
AUSTRALIANS
B
BACHELOR
BACKWATER
BALLARAT
BALMORAL
BALZAC
BARTON
BAXTER
BECHER
BEECHES
BENGAL
BERKSHIRE
BERMUDA
BERYL
BIRCHMOOR
BLANCHE
BLOOMSBURY
BOB
BOHEMIA
BOHEMIAN
BOONE
BORDEAUX
BOSCOMBE
BOSWELL
BRADSHAW
BRADSTREET
BRECKINRIDGE
BRIM
BRIONY
BRISTOL
BRITAIN
BRITANNICA
BRITISH
BRIXTON
BULLET
BURNWELL
C
CAL
CALCUTTA
CALHOUN
CALIFORNIA
CALIFORNIAN
CAMBERWELL
CARBUNCLE
CARLO
CARLSBAD
CAROLINAS
CARR
CASSEL
CATHERINE
CEDARS
CHARING
CHARLES
CHESTERFIELD
CHINA
CHINESE
CHRIST
CHUBB
CLAIR
CLARA
CLARK
CLAY
CLOTILDE
COBB
COBURG
COEUR
CONSTABLE
CORNWALL
CORONER
CORONET
COUNTESS
COVENT
COVENTRY
CRANE
CREWE
CROWDER
CUSACK
CUVIER
D
DAILY
DANE
DARLINGTON
DE
DEVONSHIRE
DONNA
DORAN
DUCHESS
DUNCAN
DUNDAS
DUNDEE
E
EDGEWARE
EDWARD
EG
EGLONITZ
EGLOW
EGRIA
ELEY
ELIAS
ELISE
EMBANKMENT
ENDELL
ENGLAND
ENGLISH
ESQ
ETHEREGE
ETON
EUROPE
EUROPEAN
EUSTACE
EYFORD
EZEKIAH
F
FAIRBANK
FAREHAM
FARINTOSH
FARRINGTON
FELSTEIN
FENCHURCH
FERGUSON
FINNS
FLAUBERT
FLEET
FLORIDA
FLUFFY
FORDHAM
FOWLER
FRANCE
FRANCIS
FRANCISCO
FRANCO
FRANK
FREEBODY
FRENCH
FRENCHMAN
FRESNO
FRISCO
FRITZ
G
GAZETTEER
GEORGE
GEORGIA
GERMAN
GERMANS

GESELLSCHAFT
GLADSTONE
GLOBE
GODFREY
GOODGE
GOODWINS
GORDON
GOTTSREICH
GRAVESEND
GREENWICH
GRICE
GRIMESBY
GROOM
GROSVENOR
GUINEA
GUSTAVE
H
HAFIZ
HAGUE
HALIFAX
HAMPSHIRE
HANKEY
HANOVER
HARDY
HARE
HARLEY
HARRIS
HARROW
HATHERLEY
HATTY
HAYLING
HEBREW
HELEN
HENRY
HERCULES
HEREFORD
HEREFORDSHIRE
HIGHNESS
HOLBORN
HOLLAND
HOLMES
HONORIA
HOPKINS
HORACE
HORNER
HORSHAM
HOSMER
HUDSON
HUGH
HYDE
IGNOTUM
II
III
INDIA
INDIAN
INDIANS
INNER
INST
IRENE
IRISH
ISA
ITALIAN
IV
IX
J
JABEZ
JACK
JACKSON
JAMES
JANE
JEM
JEREMIAH
JOHN
JONES
JOSEPH
JOVE
JULIA
JURY
K
KATE
KENSINGTON
KENT
KILBURN
KRAMM
KU
L
LA
LANCASTER
LANGHAM
LASCAR
LEADENHALL
LEATHERHEAD
LEBANON
LEE
LESTRADE
LLOYD
LONDON
LONDONERS
LOTHMAN
LOUISIANA
LUCY
LYON
LYSANDER
M
MADAME
MAGGIE
MAGNIFICO
MALAY
MALL
MARBANK
MARSEILLES
MARY
MATHESON
MAUDSLEY
MAURITIUS
MCCARTHY
MCCARTHYS
MCCAULEY
MCFARLANE
MCQUIRE
MELBOURNE
MENDICANT
MENINGEN
MEREDITH
MERRYWEATHER
MEXICO
MIDDLESEX
MILLAR
MONICA
MONTAGUE
MONTANA
MORAN
MORCAR
MOROCCO
MORRIS
MORTIMER
MOULTON
MUFF
MUNICH
MUNRO
N
NED
NEVILLE
NEZ
NORTHUMBERLAN
D
NORTON
NOVA
O
OAKSHOTT
OCCURRENCE
ODESSA
OMNE
OPENSHAW
ORMSTEIN
OXFORD
OXFORDSHIRE
P
PA
PACIFIC
PADDINGTON
PALL
PALMER
PANCRAS
PARADOL
PARAMORE
PARIS
PARR
PATERSONS
PAUL
PEARL
PENNSYLVANIA
PENTONVILLE
PERCY
PERSIAN
PETER
PETERSFIELD
PETERSON
PETRARCH
PEW
PHILADELPHIA
PINCE
PIPS
PLANTAGENET
PONDICHERRY
PORTSDOWN
PRAGUE
PRENDERGAST
PRITCHARD
PROOSIA
PRUSSIAN
QUINCEY
R
REGENCY
REGENT
RIEN
ROBERT
ROBINSON
ROCKIES
ROSS
ROTTERDAM
ROYLOTT
ROYLOTTS
RUCASTLE
RUCASTLES
RUSSELL
RY
RYDER
SALLY
SAN
SAXE
SAXON
SCALA
SCOTIA
SCOTT
SEVERN
SHERLOCK
SHOLTO
SHOLTOS
SIGISMOND
SIMON
SOPHY
SOUTHAMPTON
SOUTHERTON

SPAULDING
SPECKLED
SPENCE
STARK
STEVENSON
STOKE
STONER
STOPER
STRAND
STREATHAM
STROUD
SURREY
SUSSEX
SUTHERLAND
SWAIN
SWANDAM
SWINDON
T
TANKERVILLE
TENNESSEE
TEXAS
THAMES
THOREAU
THREADNEEDLE
TOLLER
TOLLERS
TOTTENHAM
TOUT
TRAFALGAR
TREPOFF
TRINCOMALEE
TUDOR
TURKISH
TURNER
U
UFFA
ULSTER
V
VENNER
VERE
VI
VICTORIA
VII
VIII
VINCENT
VON
WALLENSTEIN
WALSALL
WALSINGHAM
WARBURTON
WARSAW
WATERLOO
WATSON
WELLINGTON
WESTAWAY
WESTBURY
WESTHOUSE
WESTPHAIL
WHARF
WHEAL
WHITNEY
WHITTINGTON
WIGHT
WIGMORE
WILHELM
WILLIAM
WILLOWS
WILSON
WILTON
WIMPOLE
WINCHESTER
WINDIBANK
WINDIGATE
WOODCOCK
X
XI
XII
YARD
ZEALAND