FACULTAD DE FILOSOFÍA Y LETRAS
DEPARTAMENTO DE FILOLOGÍAS INGLESA Y ALEMANA

Tesis Doctoral

"Recopilación, desarrollo pedagógico y evaluación de un banco de colocaciones frecuentes de la lengua inglesa a través de la lingüística de corpus y computacional"

La Directora de la tesis, La Doctoranda,

Dra. Carmen Pérez Basanta María Moreno Jaén

Granada, 2009

Editor: Editorial de la Universidad de Granada
Autor: María Moreno Jaén
D.L.: GR 2294-2009
ISBN: 978-84-692-3097-8

AGRADECIMIENTOS

Todo aquél que conoce a la Dra. Carmen Pérez Basanta, directora de esta tesis, me comprenderá perfectamente si digo que resulta verdaderamente imposible agradecer en unas pocas líneas todo lo que ha hecho por mí. ¿Cómo agradecer su desvelo constante de varios años, sus correcciones a las 6 de la mañana, su guía, su ayuda, su generosidad, su respaldo incondicional tanto personal como académico? Gracias, Carmen, por tu dedicación, por transmitirme tu pasión y enseñarme a disfrutar de la investigación y de la docencia, por haber hecho que cada día de los últimos cuatro años dé gracias por poder trabajar a tu lado y aprender de ti. Gracias, de corazón, por todo.

Mi más sincero agradecimiento a todos los miembros del proyecto ADELEX, de los que tanto he aprendido, y muy especialmente a la Dra. Mª Teresa López-Mezquita, cuya labor investigadora ha sido siempre un ejemplo para mí y ha supuesto la base sobre la que se fundamenta esta tesis. Gracias a todos por vuestro apoyo constante y por todo lo que me habéis enseñado estos años. Esta tesis no habría sido posible sin todo lo que me habéis aportado cada uno de vosotros.

Gracias al equipo ECPC de la Universitat Jaume I, especialmente a su directora, la Profesora María Calzada y a Mª Jesús Blasco, Noemí Marín y José Manuel Martínez. Todos los viajes y proyectos que hemos compartido forman también parte de esta tesis.

A los profesores y becarios del Departamento de Filología Inglesa y Alemana, por potenciar mi vocación por los estudios ingleses, por vuestro apoyo y vuestra amistad. Gracias también a Marina Romera, por hacer que los trámites

administrativos sean siempre fáciles y sobre todo por tu apoyo y los ánimos que siempre me das.

Quiero también agradecer especialmente el apoyo que recibí de la Profesora Angela Chambers y de Stéphanie O'Riordan durante mi estancia de investigación en Limerick (Irlanda). Os agradeceré siempre vuestra ayuda.

A las profesoras Isabel Andrés y Emilia Iglesias por prestarme generosamente vuestras clases para el pilotaje de los tests. Y en esto, como en todo lo demás, gracias a todos los miembros de ADELEX y a Mª Jesús Blasco.

A Ana Azorín por su amistad incondicional, por ser la mejor amiga que se puede tener.

Y, por supuesto, gracias a mi familia. A Ramón, por todo. Por ocuparse de mí en los momentos en que la tesis absorbía toda mi vida, por su generosidad infinita con su tiempo, su trabajo y su esfuerzo, siempre con una sonrisa y sin pedir nada a cambio. Gracias, por ser la roca en la que siempre me apoyo y en la que encuentro refugio. Gracias a mis padres. A mi padre, por su aliento constante y por inculcarme desde pequeña la satisfacción que da el deber cumplido. Y a mi madre, por su ternura, por sus sabios consejos y sus caricias en tantas noches sin sueño. A los dos, gracias por vuestro cariño y vuestro esfuerzo sin los que no habría tenido la oportunidad de afrontar ni esta tesis ni nada de lo que he hecho en mi vida. Gracias a mi hermana, por ser mi compañera y compartir todo. Por enseñarme a ser mejor persona cada día. Esta tesis, y todo lo que tengo, es también vuestro.

A Carmen y a mi familia.

Siempre os llevo en el corazón.

ÍNDICE

INTRODUCCIÓN

Una reconocida investigadora en el ámbito del léxico en la enseñanza/aprendizaje del inglés, y la persona que inspiró esta tesis, la Dra. Pérez Basanta, suele decir que el vocabulario[1] es un monstruo de muchas cabezas que articula nuestra competencia comunicativa y que representa la piedra angular del aprendizaje de una lengua extranjera. Utilizando sus palabras, "lexis is at the heart of language acquisition". Sin duda, las colocaciones representan una de las cabezas de ese monstruo, tan complejas, y a la vez tan fundamentales, que merecen una profunda investigación que nos lleve a comprenderlas y, por tanto, a enseñarlas mejor.

Hoy en día está ampliamente reconocido que las colocaciones (y, en términos más generales, las unidades fraseológicas) son de gran importancia en el estudio de una lengua (Cowie, 1998; Schmitt, 2004; Butler, 2005; Luque y Pamies, 2005). Y, sin duda, una razón de peso para valorar estas unidades es su frecuencia, un aspecto de capital importancia en nuestra opinión, y que formará uno de los ejes del trabajo que aquí se presenta. Baste decir por ahora que, como algunos autores apuntan (Mel'čuk, 1998; Hill, 2000), las colocaciones son tan comunes entre los hablantes que deberían constituir un área prioritaria en los estudios dedicados a la enseñanza de lenguas extranjeras.

Pero la importancia de las colocaciones no sólo radica en su frecuencia, sino que además son elementos léxicos que contribuyen en gran medida a fomentar tanto la precisión como la fluidez del hablante. En lo que se refiere a la primera, estas

[1] En este trabajo se utilizarán los términos "léxico" y "vocabulario" de forma intercambiable.

combinaciones forman unidades más o menos estables que otorgan naturalidad al lenguaje. Cuando un alumno produce oraciones como "*I will tell you about my opinion*" (en lugar de "*I'll give you my opinion*"), no suele quebrarse la comunicación ya que el mensaje se proyecta de manera comprensible. Sin embargo, uno suena "raro" o poco natural a los oídos del hablante nativo simplemente porque carece de precisión en la elección de los términos utilizados, un problema que sólo se solventa mediante la mejora de la competencia colocacional.

Junto a la precisión, la fluidez es también un aspecto que depende en buena medida del dominio que el hablante posea de las colocaciones. Como algunos investigadores han sugerido (Wray, 2002; Woolard, 2005a), las colocaciones son unidades prefabricadas de la lengua que suponen una única elección para el hablante y que, por tanto, funcionan como lubricantes y reducen el esfuerzo cognitivo en el uso de las destrezas receptivas y productivas. Hill (2000: 54) lo expresa de la siguiente forma:

> Collocation allows us to think more quickly and communicate more efficiently. Native speakers can only speak at the speed they do because they are calling on a vast repertoire of ready-made language, immediately available from their mental lexicons. Similarly, they can listen at the speed of speech and read quickly because they are constantly recognising multi-word units rather than processing everything word-by-word.

Así pues, parece evidente que poseer una buena competencia colocacional mejoraría definitivamente la precisión y la fluidez del alumno, ya que éste no necesitaría seleccionar ni procesar cada una de las palabras que utiliza. Si a ello unimos el hecho de que las colocaciones, como ya dijimos, son especialmente comunes dentro del ámbito de las combinaciones fraseológicas, resulta evidente que se trata de un componente fundamental del lenguaje.

Por otro lado, es también un lugar común hablar de la dificultad que las colocaciones entrañan en la adquisición de una lengua extranjera. Wray (1999: 468)

declara que "knowing which subset of grammatically possible utterances is actually commonly used by native speakers is an immense problem for even the most proficient of non-natives". Desde una perspectiva puramente lingüística, parece razonable pensar que la naturaleza arbitraria de las colocaciones conlleva una mayor complejidad en su aprendizaje. También desde el punto de vista pedagógico se han postulado distintas hipótesis sobre los factores que inciden en la dificultad de este elemento fraseológico. Así, algunos estudios consideran que la falta de una adecuada concienciación sobre este fenómeno puede explicar el hecho de que los alumnos tiendan a confiar en exceso en su lengua materna traduciendo las colocaciones literalmente (Falghal y Obiedat, 1995). Otros, en cambio, estiman que la principal causa de esta dificultad es el hecho de que los hablantes no nativos adquieren y organizan mentalmente el vocabulario de forma distinta a como lo hace el hablante nativo. Según argumentan autores como Wray (2002), y Schmitt y Underwood (2004), los aprendices de una lengua extranjera, a diferencia de lo que sucede con los nativos, tienden a procesar y memorizar el vocabulario de forma individual, pasando sólo más tarde a construir unidades multiléxicas. Ello explicaría la dificultad que supone para los alumnos establecer asociaciones de palabras formando colocaciones, frente a la destreza que suelen mostrar los nativos.

Teniendo en cuenta los argumentos anteriores, sería de esperar que las colocaciones representasen un aspecto ampliamente investigado tanto en lo que se refiere a su pedagogía como a su evaluación. Por desgracia, éste no parece ser el caso. Hasta la fecha, son todavía escasos los estudios producidos en el área de la enseñanza/aprendizaje de lenguas extranjeras dedicados a abordar el componente colocacional mediante un enfoque verdaderamente sistemático y bien fundamentado. Pero es quizá en el campo de la evaluación donde las carencias se hacen todavía más palpables. Las colocaciones han estado tradicionalmente olvidadas en la evaluación

de segundas lenguas, y tan sólo en la última década se han comenzado a llevar a cabo estudios más rigurosos en este sentido.

Es precisamente en este ámbito donde la presente tesis pretende hacer su aportación. Partiendo del convencimiento fundamental de que para poder realizar un tratamiento pedagógico eficaz se hace imprescindible primero conocer cuál es la situación actual, y cuál debe ser, por tanto, nuestro punto de partida, nos planteamos los siguientes objetivos:

- Evaluar los niveles de competencia colocacional de los alumnos de Filología Inglesa (Universidades de Granada y Almería) y de Traducción e Interpretación (Universidades de Granada y Jaume I).

  - Analizar variables como la licenciatura, el curso en el que se encuentran los estudiantes y las distintas estructuras gramaticales de las colocaciones (en nuestro caso N+N, V+N, A+N y N+V) con el fin de evaluar su impacto en los distintos niveles de competencia que muestran los alumnos.

- Diseñar un test que nos permita evaluar la competencia colocacional de forma válida y fiable.

  - Ofrecer una definición del concepto de colocación que sea operativa y evaluable con el fin de garantizar la validez de constructo de nuestro test.

  - Compilar un listado de colocaciones frecuentes mediante técnicas de lingüística de corpus con el fin de contar con un banco del que poder

extraer las colocaciones que se incluirán en los ítems del test, procurando así su validez de contenido.

- Diseñar una prueba tomando como referencia el modelo de construcción de tests de López-Mezquita (2005), en el que se presta especial atención a la fiabilidad (formato y número de ítems, instrucciones, tiempo de la prueba y criterios de corrección) y a la validez (operatividad del constructo, selección de contenidos y diseño de ítems).

- Comprobar empíricamente la validez y la fiabilidad del test, mediante análisis estadísticos donde se examina el comportamiento de los métodos de evaluación y de los ítems del test.

Para llevar a cabo tales objetivos, hemos realizado la investigación que recogen los cinco capítulos de este trabajo y que esbozamos brevemente a continuación.

El **capítulo 1** está dedicado por entero al constructo de la colocación. Con el objetivo de exponer de manera clara el modo en que este fenómeno se ha tratado en nuestro trabajo, llevamos a cabo una completa revisión de las diferentes aproximaciones al estudio de la colocación, entre las cuales destacamos el enfoque estadístico y el fraseológico, explicando las aportaciones de los principales expertos en la materia. La última parte del capítulo está dedicada a ofrecer nuestra propia visión al respecto, materializada en la formulación de unas características formales y funcionales que pretenden configurar parámetros operativos para el posterior diseño de nuestro test.

El **capítulo 2** pretende poner de manifiesto la importancia que cobra la noción de frecuencia en el ámbito de la enseñanza del léxico, basándonos en la convicción de que los elementos más frecuentes en la lengua son los que un aprendiz

necesita conocer. A tal fin, este capítulo realiza un recorrido histórico por las diferentes etapas de la enseñanza del vocabulario, destacando el hecho de que el estudio léxico y el concepto de frecuencia han ido siempre de la mano. Este lazo de unión se ha hecho, si cabe, más estrecho desde la aparición de los estudios de corpus, dado el enorme potencial que presentan desde el punto de vista del análisis cuantitativo de datos. En este capítulo, pues, se lleva a cabo una revisión de los listados de frecuencias de vocabulario a los que ha dado lugar esta nueva forma de estudiar la lengua. Finalmente, se realiza un recorrido similar por los listados y compilaciones de colocaciones, evidenciando la escasez de investigación que ha existido hasta la fecha en este ámbito. Con el objetivo de aportar algo de luz en el largo camino que aún queda por recorrer en el estudio de la frecuencia de las colocaciones, llevamos a cabo un listado propio cuya elaboración constituye el objeto de estudio del siguiente capítulo.

El **capítulo 3** describe de manera pormenorizada el proceso de elaboración de nuestro listado de colocaciones frecuentes, utilizando las fuentes y técnicas más actualizadas y fiables que ofrece hoy en día la lingüística de corpus para la extracción y análisis de colocaciones, y en el que también se realiza un exhaustivo análisis manual para acomodarse a las coordenadas establecidas en el capítulo primero con respecto a nuestra particular concepción de la colocación. Este novedoso listado, basado en datos del *Bank of English* y el *British National Corpus*, constituirá el banco de colocaciones que nos servirá para la selección de contenidos de nuestro test.

En el **capítulo 4** se presenta nuestro diseño de investigación. Siguiendo el modelo de López-Mezquita (2005) para la elaboración de medidas de evaluación, se ha diseñado un test, el **Test de Colocaciones ADELEX Versión 1**, que pretende ser una prueba que mida el conocimiento colocacional de los alumnos de forma válida y fiable. Así, este capítulo recoge el riguroso procedimiento seguido para su planificación, construcción, administración y corrección.

El **último capítulo** de este trabajo tiene como principal objetivo comprobar empíricamente, por un lado, cuáles son los niveles de competencia del alumnado y, por otro, si nuestro test representa una medida válida y fiable del conocimiento colocacional. Para ello, se ofrecen los resultados obtenidos tras la administración de la prueba en su primer proceso de pilotaje y se lleva a cabo un análisis de estadística descriptiva e inferencial de los datos recabados, junto con el coeficiente de fiabilidad y un análisis de ítems y de distractores. Tras este exhaustivo análisis cuantitativo, se llevará a cabo un proceso de mejora del test que se materializará en su segunda versión: **Test de Colocaciones ADELEX Versión 2**. En este caso, además, se trata de un test digitalizado que está ya listo para su pilotaje a través de la plataforma Moodle.

Finalmente, ofrecemos las conclusiones alcanzadas tras nuestro estudio, así como una propuesta de la investigación que nos proponemos llevar a cabo en el futuro.

En este trabajo hemos incluido asimismo 6 apéndices (todos ellos se adjuntan en un CD-Rom aunque los **apéndices 3 y 4 pueden encontrarse también en formato impreso al final** de este trabajo). En primer lugar se ofrece el listado de las 1.000 primeras palabras recogidas en la lista de frecuencias de López-Mezquita (2005), a partir de la cual se han extraído los sustantivos que conforman las bases de las colocaciones de nuestro listado de frecuencias. Los anexos 2 y 3 presentan dicho listado en dos formatos diferentes. En el primero de ellos se ofrece información detallada de la frecuencia y el índice *t-score* de cada colocado según los datos del *Bank of English* y el *British National Corpus* por separado, mientras que en el segundo aparece un listado definitivo de las colocaciones recogidas en nuestro estudio, donde la información aportada por ambos corpus se ha cruzado dando lugar a una sola lista de colocaciones. Los anexos 4 y 5 presentan un modelo del Test de Colocaciones ADELEX Versión 1. En el caso del apéndice 4 se ofrece el test tal y como se

administró a los candidatos, mientras que en el 5 se incluyen las respuestas correctas de cada ítem. Por último, el Anexo 6 presenta un modelo de la segunda versión de nuestro test: Test de Colocaciones ADELEX Versión 2.

No nos gustaría concluir este apartado sin hacer una mención especial a los trabajos que han supuesto el punto de partida de esta tesis, sin los cuales nuestra investigación no habría sido posible. Bien por ser obras ya clásicas de obligada referencia en el ámbito de los estudios léxicos y/o colocacionales, o bien por su carácter innovador y por los nuevos caminos que están abriendo en nuestro campo, se trata en todo caso de estudios que han sido fuente de inspiración en esta tesis. Son, en definitiva, los textos a los que hemos vuelto una y otra vez, releyendo y redescubriendo aspectos de interés para nuestra investigación, constituyendo pues verdaderas fuentes inagotables de ideas. Estos trabajos son los siguientes:

CARTER, R. 1987. *Vocabulary: Applied linguistic perspectives*. Londres/Nueva York: Routledge.

FONTENELLE, T. 1994. "What on earth are collocations?". *English Today*, 40, 10, 4: 42-48.

GRANGER, S. Y **PAQUOT, M**. 2008. "Disentangling the phraseological web". En S. Granger y F. Meunier (eds.), *Phraseology. An interdisciplinary perspective*. (pp. 27-49). Ámsterdam: John Benjamins.

GYLLSTAD, H. 2007. *Testing English collocations. Developing receptive tests for use with advanced Swedish learners*. Tesis doctoral. Lund: Lund University.

HOWARTH, P. A. 1996. *Phraseology in English academic writing. Some implications for language learning and dictionary making*. Tübingen: Niemeyer.

LEWIS, M. 1993. *The lexical approach*. Hove: Language Teaching Publications.

LEWIS, M. (ed.). 2000. *Teaching collocation: Further developments in the lexical approach*. Hove: Language Teaching Publications

LÓPEZ-MEZQUITA MOLINA, M. T. 2005/2007. *La evaluación de la competencia léxica: Tests de vocabulario. Su fiabilidad y validez*. Tesis doctoral. Granada: Universidad de Granada.

MCCARTHY, M. 1990. *Vocabulary*. Cambridge: Cambridge University Press.

NATION, P. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

NESSELHAUF, N. 2005. *Collocations in a learner corpus*. Ámsterdam: John Benjamins.

PAWLEY, A. Y SYDER, F. 1983. "Two puzzles for linguistic theory: nativelike selection and nativelike fluency". En J. Richards y R. Schmidt (eds.), *Language and communication*. (pp. 191-226). Londres: Longman.

PÉREZ BASANTA, C. 1999. "La enseñanza del vocabulario desde una perspectiva lingüística y pedagógica". En S. Salaberri (ed.), *Lingüística aplicada a la enseñanza de lenguas extranjeras*. (pp. 262-306). Almería: Universidad de Almería.

PÉREZ BASANTA, C. 2005. "Assessing the receptive vocabulary of Spanish students of English philology: An empirical investigation". En J.M. Martínez-Dueñas Espejo, N. McLaren, C. Pérez Basanta y L. Quereda Rodríguez-Navarro (eds.), *Towards an understanding of the English language: Studies in honour of Fernando Serrano*. (pp. 456-477). Granada: Universidad de Granada.

PÉREZ FERNÁNDEZ, A. 2002. *Las colocaciones léxicas de los adverbios intensificadores ingleses. Propuesta de un diccionario*. Tesis doctoral no publicada. Universidad de Jaén.

PHILIP, G. 2007. "Decomposition and delexicalisation in learners' collocational (mis)behaviour". En *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham, julio 2007. (pp. 1-9).

READ, J. 2000. *Assessing vocabulary*. Cambridge: Cambridge University Press.

RICHARDS, J. 1976. "The role of vocabulary teaching". *TESOL Quarterly*, 10, 1: 77-89.

SINCLAIR, J. 1966/1996. "Beginning the study of lexis". En J. Foley (ed.), J.M. Sinclair on lexis and lexicography. (1-20). Singapur: Unipress.

SINCLAIR, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

CAPÍTULO 1

EL CONCEPTO DE COLOCACIÓN. DEFINICIÓN Y CLASIFICACIÓN

Collocation: hard to pin down, but bloody useful.
Schmid (2003: 235)

1.1. Introducción

El estudio de las colocaciones ha suscitado un considerable interés en los círculos académicos en las últimas décadas. Llevados principalmente por el convencimiento de que estas unidades léxicas conforman un elemento central del lenguaje y de su enseñanza/aprendizaje, muchos lingüistas han profundizado en este componente de la lengua desde disciplinas tan alejadas, y a la vez tan cercanas hoy día, como la gramática, la semántica, la lexicología y la fraseología, la lingüística aplicada y los estudios computacionales.

Sin embargo, a pesar de la ingente investigación llevada a cabo en los últimos años en torno a las colocaciones, o quizá debido precisamente a la profusión de

estudios que han analizado este fenómeno desde ángulos tan distintos, no son pocas las incógnitas que restan aún por resolver en lo que se refiere a la definición y clasificación del propio concepto de "colocación". Tanto el hecho de que la gran mayoría de publicaciones dedicadas a este componente lingüístico tome como punto de partida la definición del constructo de colocación, como la falta de consenso existente entre los investigadores ponen de manifiesto la necesidad de establecer un criterio claro que resulte adecuado y operativo para el propósito de este trabajo. El presente capítulo trata de presentar de manera sucinta las principales concepciones y clasificaciones de la noción de colocación existentes en la actualidad a la vez que ofrece nuestra propia percepción del fenómeno con vistas al diseño de un test para evaluar la competencia colocacional.

1.2. Los primeros estudios sobre la colocación

No son muchos los datos que existen en torno a los orígenes del concepto de colocación. Sabemos, no obstante, que ya en la Francia del s. XVI se compiló uno de los primeros diccionarios de combinaciones léxicas de que se tiene constancia, titulado *Les Epithètes* y publicado en 1571 por M. de la Porte (citado por Hausmann, 1979). Este diccionario recoge ochenta mil combinaciones formadas por cuatro mil entradas de sustantivos y los adjetivos con los que pueden asociarse. Como Hausmann destaca, esta obra lexicográfica pretendía ser "non seulement utile à ceux qui font profession de la Poësie, mais fort propre aussi pour illustrer toute autre composition françoise (…) le type même du dictionnaire destiné à aider ceux qui sont à la recherche du mot approprié" (1979: 187-188).

La primera referencia que tenemos del uso del término "*collocation*" data de 1627, cuando Francis Bacon la emplea en su obra *Natural History*, aunque no en su

sentido lingüístico (Gyllstad, 2007). Según el *Compact Oxford English Dictionary* (1994), la palabra se utiliza por primera vez en relación con el área del lenguaje en 1750 cuando Harris escribe: "The accusative…in modern languages…being subsequent to its verb, in the collocation of the words". Años más tarde, en 1873, Earle subraya que "[a]ll languages use greater freedom of collocation in poetry than in prose" (Oxford English Dictionary, citado por Howarth, 1996: 25). Ya en el s. XX, Mitchell (1971) documenta que Jespersen también hace uso de esta palabra en el año 1917.

Aunque muchos de los trabajos en torno a las colocaciones estiman que el origen de este término para designarlas como objeto de estudio lingüístico se remonta a Firth (1957), nos consta que ya en los años 30 los miembros del *Institute for Research in English Teaching* (IRET), dirigido por el profesor Harold E. Palmer[1], utilizaban el término "*collocation*" con este fin. En el *Second Interim Report on Collocations*, publicado por esta institución en el año 1933, Palmer declara lo siguiente:

> The word 'collocation' (…) is the occupant of an honourable place in standard dictionaries. It is respectable-looking and sounding. (…). In Linguistics it is already in use as a technical but conveniently vague word. It is of Latin parentage. It is easily transliterated into other European languages, and in all of them has the aspect of semifamiliarity. It is neither freakish nor smile-provoking. It is reasonably short and hyphened. In its semantic history it so far means nothing in particular but means it very well —symbolizing, as it does, 'a placing together', or 'that which results from a placing together'. (Palmer, 1933, citado por Pérez Fernández, 2002)

En este trabajo, pionero en el estudio de la colocación, Palmer introdujo dicho término en la disciplina lingüística definiéndolo de la siguiente manera: "A collocation is a succession of two or more words that must be learned as an integral whole, and not pieced together from its component parts" (Palmer, 1933, citado por

[1] La importante labor de los miembros de esta institución en el desarrollo de la enseñanza de lenguas se describirá con más detalle en el capítulo 2.

Cowie, 1998: 211). Por tanto, resulta evidente que cuando Firth empleó esta palabra en sus trabajos de los años 50 no estaba utilizando un vocablo desconocido.

En lo que respecta al uso de este término en los estudios realizados en español, cabe destacar que fue precisamente a partir de la obra de Firth que Seco (1978: 218) introdujo la palabra "colocación" en la terminología lingüística española. A pesar de no contar con la aceptación unánime de los expertos en un principio (Alonso Ramos, 1993), hoy en día es el término más utilizado para designar este fenómeno fraseológico en el español actual (Koike, 2001; Luque y Pamies, 2005).

1.3. Definición de colocación

Podemos afirmar que en la gran mayoría de los trabajos realizados por los expertos en el ámbito de las colocaciones, sea cual sea la aproximación que éstos adopten, subyace la idea básica de que la colocación es un aspecto fraseológico del lenguaje que cuestiona de manera evidente la noción generativista del lenguaje postulada por Chomsky (1957, 1965), ya que se trata de una unidad prefabricada, disponible para el hablante como un todo y diferente de la combinación libre de palabras. Coincidimos por tanto plenamente con Mel'čuk (1998: 23) cuando destaca que "collocations —no matter how one understands them— are a subclass of what are known as *set phrases*"[2].

Sin embargo, y como ya señalamos anteriormente, existen muy diversas definiciones del concepto de colocación. Si bien es cierto que esta falta de consenso se debe en parte a la naturaleza escurridiza e imprecisa que caracteriza a las

[2] El concepto aquí expresado por Mel'čuk mediante el término genérico "*set phrases*" equivale a lo que nosotros denominaremos unidades fraseológicas. Como se verá en la sección 1.3.1.2.2., este autor utiliza también el término "*phraseme*" para referirse a este tipo de unidades.

combinaciones léxicas, otra razón fundamental de esta variedad es el hecho de que la aproximación a este fenómeno se realiza a menudo desde diferentes perspectivas que llevan a los expertos a formular definiciones y clasificaciones acordes con su particular punto de vista y su metodología de trabajo. Hoy en día son ya mayoría los autores (Nesselhauf, 2005; Gyllstad, 2007, Granger y Paquot, 2008) que establecen dos enfoques fundamentales en los estudios dedicados al área colocacional, y en general a la fraseología: aquel que parte del ámbito estadístico y de la frecuencia de co-aparición léxica, y el que adopta una perspectiva fraseológica, más cercana tradicionalmente al área de la enseñanza de lenguas. A continuación realizaremos un breve repaso de las aportaciones más destacables dentro de ambos enfoques con el fin de situar de un modo claro nuestra propia concepción al respecto.

1.3.1. Aproximaciones al concepto de colocación

1.3.1.1. La aproximación estadística

La aproximación estadística al fenómeno colocacional viene dada por el deseo compartido por un buen número de investigadores de dejar atrás las teorías intuitivas y poco fundamentadas que hasta mitad del s. XX caracterizaban a los estudios lingüísticos. A raíz de las nuevas prestaciones ofrecidas por los corpus y la informática, la investigación de las lenguas se basa en análisis cuantitativos de datos que le otorgan un nuevo estatus como ciencia empírica. Básicamente, desde este enfoque la colocación se entiende como la combinación de dos (y según algunos autores, como veremos más adelante, más de dos) palabras que aparece en la lengua conjuntamente con una frecuencia estadísticamente significativa, es decir, formando unidades léxicas de palabras que no se combinan al azar.

Son muchos los autores que han adoptado una definición puramente estadística en su estudio de la colocación, o basado en gran medida en datos cuantitativos extraídos de corpus informatizados (Kjellmer, 1994; Stubbs, 1995; Moon, 1998a; Bernardini, 2007). Sin embargo, analizaremos únicamente las aportaciones hechas en este campo por John Sinclair puesto que es el autor, en nuestra opinión, más representativo de esta aproximación, no sin antes detenernos en las teorías precursoras del estudio estadístico colocacional promulgadas por John Firth y Michael Halliday.

1.3.1.1.1. Precursores: John R. Firth y Michael A. Halliday

Dado el interés que John R. Firth (1890-1960) mostró en sus trabajos por establecer una teoría lexicológica donde las relaciones sintagmáticas y paradigmáticas fueran equiparables y paralelas a las existentes en la teoría gramatical, los estudios de este lingüista británico en el ámbito del vocabulario enfatizan la importancia tanto del cotexto como del contexto. Este académico, cuya investigación giraba en torno al significado como valor fundamental del lenguaje, fue el primero en considerar la colocación como un aspecto integrante de la dimensión semántica de cada palabra. En este sentido, es ya famosa su afirmación "you shall know a word by the company it keeps" (1957: 195), donde se advierte que las palabras no se pueden considerar como entes individuales en la lengua, sino como unidades cuyo valor depende en muchas ocasiones de las relaciones sintagmáticas en las que participan. Éste es el concepto que subyace tras el fenómeno colocacional.

Firth es considerado como el primer precursor de la concepción estadística de las colocaciones debido a afirmaciones como la siguiente:

> Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of 'night' is its collocability with 'dark', and of 'dark', of course, its collocation with 'night' (ibid.: 196).

Aunque, como apunta Gyllstad (2007), Firth no dejó en su obra una definición clara y consistente de lo que él entendía por colocación, sí se puede apreciar el enfoque cuantitativo que proponía, ya que en algunos de sus ejemplos ofrecía el número de veces que aparecían ciertas palabras en los textos que trabajaba, además de utilizar en muchos casos palabras de marcado carácter cuantificable como "*habitual*, *commonest*, *frequently*, *not very common*, *general*, *usual* and *more restricted*" (McEnery, Xiao y Tono, 2006: 146).

Aunque las novedosas ideas propuestas por Firth fueron poco después desarrolladas por autores como McIntosh (1961), Greenbaum (1970) y Mitchell (1971), las investigaciones que mayores repercusiones tuvieron en la aproximación estadística al estudio colocacional fueron las llevadas a cabo por Halliday, y poco después por Sinclair. En su intento por perfeccionar las nociones introducidas por Firth, Michael A. K. Halliday (n. en 1925) utiliza la colocación como criterio de agrupación, tratando de organizar el vocabulario en conjuntos léxicos. Así, el hecho de que "*strong*" y "*powerful*" coloquen con "*argument*" justifica, según Halliday (1966), que ambos se agrupen en un mismo conjunto léxico. Sin embargo, si tenemos en cuenta que "*strong*" coloca con "*tea*" y no con "*car*", mientras que "*powerful*" coloca con "*car*" y no con "*tea*", estos dos adjetivos pertenecen también a otros dos conjuntos léxicos diferentes. Vemos así que Halliday emplea, de forma totalmente pionera, la dimensión paradigmática de la colocación para establecer relaciones léxicas en la lengua, extrapolando de esta manera al vocabulario las teorías tradicionalmente empleadas en el ámbito de la gramática.

Como destaca Malmkjær, para Halliday un conjunto léxico "can be demonstrated as a statistical reality" (2004: 344). Si dos palabras tienen una alta

probabilidad de ir acompañadas por los mismos colocados existen razones suficientes para considerarlas miembros del mismo conjunto léxico. Por tanto, en la misma línea que Firth pero ahondando en las nociones de frecuencia y probabilidad de forma explícita, Halliday define las colocaciones como

> the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at *n* removes (a distance of *n* lexical items) from an item *x*, the items *a, b, c* …Any given item thus enters into a range of collocation, the items with which it is collocated being ranged from more to less probable (1961: 276).

Podemos observar que la definición propuesta por Halliday se encuentra ya considerablemente más encaminada hacia los análisis cuantitativos, que años después cobrarían un enorme auge gracias al desarrollo de las técnicas computacionales y a la introducción del análisis de corpus y la estadística en los estudios lingüísticos. Su clarividencia a la hora de establecer el valor y la necesidad de contar con datos y análisis probabilísticos en los estudios de colocaciones fue tal que incluso llegó a afirmar que "the occurrence of an item in a collocational environment can only be discussed in terms of probability" (Halliday, 1966: 159).

La influencia que Firth ejerció en los trabajos de Halliday se dejó notar no sólo en la consideración de la colocación desde el punto de vista puramente léxico, sino también en la atención que ambos prestaban al análisis textual y a las cuestiones relativas al cotexto. El interés suscitado en Halliday por los aspectos textuales del lenguaje dio lugar a una evolución en su concepto de colocación, tal y como se expresa en la obra *Cohesion in English* (1976: 287). En ella sus autores, Halliday y Hasan, consideran la colocación como "a cover term for the cohesion that results from the co-occurrence of lexical items that are in some way or other typically

associated with one another, because they tend to occur in similar environments". Por tanto, en un texto como el que sigue[3],

Los estilos de decoración interior toman como referencia tanto las modas formales, diseños y técnicas de colores, como los imperativos de cada época. La ornamentación, además de motivos vegetales, animales y mitológicos, utilizará símbolos de poder durante siglos.

se podrían identificar como colocaciones palabras como "estilos", "decoración", "modas", "diseños", "técnicas", "colores", "ornamentación" y "motivos" porque tienden a aparecer juntas en textos relativos a la misma temática aportando, de este modo, cohesión al texto.

Obviamente y como destaca Herbst (1996), esta concepción está más relacionada con las teorías de los campos semánticos y las relaciones semánticas que con el fenómeno colocacional tal y como se concibe en el ámbito fraseológico. En nuestra opinión, el hecho de que en un mismo texto aparezcan varias palabras relacionadas con el mismo tema —lo cual, naturalmente, aporta cohesión a dicho texto— no se puede identificar en modo alguno con el fenómeno de la colocación puesto que esta mera co-aparición de unidades léxicas se debe a factores semánticos y discursivos que nada tienen que ver con la naturaleza arbitraria y fraseológica del uso del lenguaje que justifica el empleo de colocaciones. Así, coincidimos plenamente con la opinión de Vanallemeersch (1994, citado por Pérez Fernández, 2002: 41), que denomina "conceptual collocations" a este tipo de combinaciones y las define como "pairs of words not being true collocations, but frequently appearing together because of their related meanings (e.g. bomb-soldier)". De hecho, Hasan (1984) se desmarcaría años más tarde de la teoría anteriormente formulada denominando

[3] Texto extraído de Galy, M. (ed.). 2004. *El Gran Libro del Bricolaje.* Madrid: Ediciones Larousse. Pág. 95.

"cadenas léxicas" (*lexical chains*), a lo que anteriormente se referían con la palabra "colocación", reconociendo con ello el mal empleo que se había hecho del término.

1.3.1.1.2. John M. Sinclair

Uno de los más notables seguidores de las ideas expuestas por Firth y Halliday en lo que a la probabilidad de co-aparición léxica se refiere es, sin duda, John M. Sinclair (1933-2007). Sin embargo, a diferencia de sus precursores, para quienes la colocación representaba únicamente un aspecto más a emplear en su teoría contextual del significado, Sinclair (1966/1996) desarrolla por primera vez una teoría de la colocación como tal, otorgándole así el estatus lingüístico del que hoy goza tanto a nivel descriptivo como aplicado.

Esta nueva dimensión de la colocación no habría sido posible en modo alguno si las teorías de Sinclair no hubieran estado acompañadas por los avances tecnológicos que durante los años 70 y 80 dieron lugar al auge de la lingüística computacional y a los estudios basados en corpus. La posibilidad de manejar enormes cantidades de vocabulario dentro de sus contextos auténticos con una gran rapidez para obtener y comparar datos facilitó enormemente los estudios lingüísticos. En palabras de Sinclair, "the ability to examine large text corpora in a systematic manner allows access to a quality of evidence that has not been available before" (1991: 4). No es casualidad, por tanto, que uno de los proyectos computacionales más relevantes en el ámbito lingüístico sea el proyecto *COBUILD* (*Collins-Birmingham University International Language Database*), dirigido por el propio Sinclair (ver sección 2.3.1.3.2., capítulo siguiente).

Gracias a las nuevas posibilidades que los corpus informatizados ofrecían, Sinclair dedicó sus trabajos al estudio de la enorme cantidad de datos cuantitativos de que se disponía por primera vez. Como decíamos, una de las más importantes

aportaciones de la lingüística de corpus a los estudios descriptivos llevados a cabo por Sinclair fue, sin duda, el análisis de la colocación. Nunca antes se había comprobado con tanta claridad que el vocabulario tiende a combinarse siempre de la misma manera y a relacionarse entre sí siguiendo unos parámetros más léxico-semánticos que gramaticales. Con ayuda de las recopilaciones informatizadas de textos auténticos de la lengua inglesa, este lingüista británico llevó a cabo un análisis de la frecuencia de las palabras y comprobó modelos de combinaciones de éstas mediante cálculos numéricos. Sin lugar a dudas, Sinclair es el máximo exponente de la aproximación estadística al estudio de la colocación.

Para este autor, la colocación es básicamente un fenómeno estadístico que se puede definir como "the ocurrence of two or more words within a short space of each other in a text" (Sinclair, 1991: 170). Tal es la exclusividad de su percepción estadística que llega a especificar claramente que en su trabajo "the attention is concentrated on lexical co-occurrence, more or less independently of grammatical pattern or positional relationship" (ibid.). Por tanto, el único criterio de identificación utilizado por Sinclair es la frecuencia de co-aparición de los ítems, algo que, como veremos, no ha contado con la aceptación de muchos de los autores que parten de una concepción sinclairiana de la colocación.

El factor terminológico de la teoría colocacional formulada por Sinclair debe ser considerado con cierta atención puesto que emplea vocablos muy específicos. El "breve espacio" al que Sinclair hace referencia en la definición arriba citada se denomina "*span*", y fue fijado por Jones y Sinclair (1974) en un límite de cuatro palabras a cada lado de la palabra que se está estudiando, la cual se designa con el término "*node*". Este espacio, no obstante, ha sido ampliado hasta cinco palabras a cada lado del nodo por otros autores (Martin, Al y Van Sterkenburg, 1983; Smadja, 1993; van der Wouden, 2002). Sinclair llama "*collocates*" a las palabras que flanquean al nodo dentro del espacio o "*span*" delimitado. Sin embargo, debemos destacar que

"*node*" y "*collocate*" no se corresponden con la noción de "base" y "colocado" que conforman los elementos constituyentes de una colocación según la tradición fraseológica (ver sección 1.3.1.2.1.). Como los propios autores explican, "there is no difference in status between node and collocate; if word A is a node and word B one of its collocates, when word B is studied as a node, word A will be one of its collocates" (Jones y Sinclair, 1974: 16). Por tanto, como más tarde afirmara Sinclair, "each successive word in a text is thus both node and collocate, though never at the same time" (1991: 115). Se trata, como vemos, de dos términos más útiles en el proceso de análisis estadístico de las concordancias extraídas de un corpus que en la definición y estudio del fenómeno de la colocación.

Dos aspectos dicotómicos resultan cruciales a la hora de comprender la colocación tal y como fue formulada por Sinclair. Éstos son la **colocación significativa** y el **principio idiomático**. En cuanto a la primera de estas dicotomías, Sinclair establece una distinción entre **colocación casual** ("*casual collocation*") y **colocación significativa** ("*significant collocation*"), llegando incluso a desestimar la primera y considerar colocación sólo la segunda (Sinclair, 1991). Mientras que la colocación casual es la que puede resultar interesante por inesperada y desde luego poco frecuente (por ej. "*pronounce a speech*"), la colocación significativa es la realmente importante para el autor puesto que denota una alta frecuencia de co-aparición léxica que se refleja en resultados estadísticos significativos (por ej. "*give a speech*"). Jones y Sinclair definen la colocación significativa como una "regular collocation between items, such that they co-occur more often than their respective frequencies and the length of text in which they appear would predict" (1974: 19). De acuerdo con esta definición, la combinación "*the dog*" no se consideraría una colocación significativa ya que aunque la co-aparición de ambos ítems es altamente frecuente, la palabra "*the*" tiene una frecuencia individual muy elevada en cualquier tipo de texto. En este mismo sentido, en cambio, la combinación "*barking dog*" sí conformaría una

colocación significativa puesto que "*barking*" no es una palabra de muy alta frecuencia pero si aparece es muy probable que lo haga junto a "*dog*". Como veremos más adelante, esta definición de colocación significativa plantea un problema básico de fondo.

En lo que respecta a la segunda dicotomía, la distinción entre el **principio de elección libre** ("*open-choice principle*") y el **principio idiomático** ("*idiom principle*") resulta también fundamental en la teoría de Sinclair. A grandes rasgos podemos definir el principio de elección libre como la forma tradicional de entender y describir el lenguaje, según la cual el hablante tiene total libertad de elección en todas y cada una de las palabras y oraciones que formula, construidas mediante la continua toma de múltiples decisiones que sólo deben atenerse a las reglas de corrección gramatical y a la lógica semántica. Sin embargo, destaca Sinclair que, lejos de llevar a la práctica este principio teóricamente aceptable de elección libre, los hablantes de una lengua están limitados en sus posibilidades por el principio idiomático, el cual debe entenderse como la facultad del lenguaje por la que "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (1991: 110). Es por tanto dentro de este último principio donde se enmarca la colocación, constituyendo un fenómeno que, según Sinclair, contribuye a la idiomaticidad que caracteriza a la lengua.

En nuestra opinión, aunque la conclusión alcanzada por Sinclair en cuanto al papel que desempeña la idiomaticidad lingüística en el fenómeno de la colocación, que le lleva a considerar ésta última como una tendencia abstracta del lenguaje, coincide básicamente con nuestra percepción al respecto, el procedimiento de identificación y análisis estadístico propugnado por este autor adolece de una deficiencia fundamental, al menos cuando se aborda desde el ámbito de la enseñanza de segundas lenguas. El problema más importante que plantea la aproximación

estadística viene dado por el hecho de someter el criterio de colocación exclusivamente al de frecuencia. Desde un punto de vista puramente lingüístico, el uso de la frecuencia de co-aparición estadísticamente significativa como único criterio de definición e identificación provoca que combinaciones de palabras que por razones semánticas o por reflejar situaciones naturales de la vida tienden a aparecer juntas en el discurso se consideren colocaciones, pasando por alto el hecho de que carecen de ciertas características, en nuestra opinión inherentes a la colocación, tales como la arbitrariedad de uso (ver sección 1.3.2.2.) (Howarth, 1996; Pérez Fernández, 2002). Uno de los ejemplos más ilustrativos al respecto es el planteado por Van-Roey (1978: 158). Como se mencionó anteriormente, siguiendo criterios puramente estadísticos "*barking dog*" se debería considerar una colocación, dado que la probabilidad de que cada vez que aparezca la palabra "*barking*" ésta venga acompañada de la palabra "*dog*" es extremadamente alta. Sin embargo, como afirma Van-Roey, resulta evidente que la frecuencia de asociación de estas dos palabras responde a motivos semánticos ya que sólo los perros ladran, por lo que el hablante no utiliza la palabra "*barking*" debido a ninguna restricción arbitraria impuesta por el uso; es, sencillamente, la única palabra que puede utilizar para expresar esa idea. Nos parece en este sentido muy acertada la apreciación de Gaatone (1997), para quien la fraseología merece sin duda un lugar central en el estudio del lenguaje, pero advierte de que no todo en la lengua se debe considerar fraseológico.

Ahondando en la misma cuestión, Herbst (1996: 384) es de la opinión de que las afirmaciones estadísticas en torno al estudio colocacional tienen una capacidad limitada y señala:

> Is —to quote one of Firth's famous examples— *dark night* a significant collocation because nights tend to be dark and not bright? In other words: wouldn't it be true to say that the fact that certain words tend to co-occur must be attributed to certain facts of the world —together with the way this world is conceptualised in language?.

Como también han puesto de manifiesto algunos autores (Pérez Fernández, 2002; Wray, 2002), el uso de la frecuencia de co-aparición como criterio exclusivo de identificación de colocaciones presenta también el problema opuesto, es decir, si una colocación aparece con muy poca frecuencia en un corpus (algo que puede deberse a hechos tan diversos como el tipo de textos que conforman el propio corpus, los criterios de identificación preestablecidos, la medida estadística que estemos empleando, etc.) ésta pasará desapercibida para el investigador. Partiendo de un total acuerdo con esta apreciación desde el punto de vista teórico, consideramos, no obstante, que en este caso las limitaciones que impone el método estadístico no suponen un problema fundamental desde la perspectiva del profesor de segundas lenguas, puesto que el hecho de que una colocación aparezca con muy poca frecuencia —o incluso no aparezca en el corpus— es una evidencia del limitado uso que tiene en la lengua real, y por tanto, se puede considerar prescindible en un programa de lenguas extranjeras. Cuestión diferente, por supuesto, sería que estuviéramos abordando las colocaciones con un objetivo descriptivo o lexicográfico, en cuyo caso sí sería necesario poder identificar todas las colocaciones de la lengua, y no sólo las más frecuentes.

El hecho de que la aproximación estadística a la colocación recoja combinaciones léxicas frecuentes que no constituyen verdaderas colocaciones sí presenta, en cambio, importantes limitaciones desde el punto de vista de la enseñanza, puesto que resulta poco operativo prestar atención a aquellas combinaciones léxicas que el alumno puede crear de forma autónoma mediante el uso de las reglas lingüísticas. En este sentido coincidimos plenamente con la opinión de Philip (2007b: 2-3), fundamentada de manera clara en las teorías de Wray (2002), en cuanto a que los hablantes nativos no aplican los mismos procesos que los no nativos al utilizar el lenguaje:

> For native speakers, the vast majority of everyday language is phraseological and involves a greater or lesser degree of delexicalisation. In very general terms, the idiom principle (Sinclair, 1991) governs the meaning of conventionalised language, shifting the semantic focus away from the individual words (...), and prioritising the phraseological and pragmatic meaning of the utterance as a whole. (...) this is the native-speaker view of language: the learner's view is a little different. While the everyday language of the native speaker is delexicalised to a considerable extent, the same language when learned as a L2 is perceived as being fully salient. (...) Collocations, therefore, are initially seen as compositional combinations of words rather than as a phenomenon of co-selection.

A la luz de la afirmación anterior, observamos que los hablantes nativos utilizan las rutinas prefabricadas hasta tal punto que usan el "principio idiomático" incluso cuando no sería necesario hacerlo porque no hay ninguna restricción arbitraria —Wray (2000: 466) cita en este sentido el siguiente ejemplo: "It was lovely to see you". Por el contrario, los hablantes no nativos parecen usar el "principio de elección libre" como norma general en su producción, y sólo recurren a unidades formulaicas cuando las identifican como combinaciones inesperadas o arbitrariamente restringidas que exigen, por tanto, de una memorización global para poder producirlas con acierto (Schmitt y Underwood, 2004). A nuestro modo de ver, pues, éstas son las unidades fraseológicas, y en nuestro caso las colocaciones, sobre las que debemos llamar la atención de nuestros alumnos y sobre las que es prioritaria la actuación pedagógica. Ésta es, asimismo, la percepción más generalizada desde el enfoque fraseológico.

1.3.1.2. La aproximación fraseológica[4]

Como se verá a continuación, incluimos dentro de este enfoque a varios autores cuyas concepciones de la colocación difieren en algunos aspectos (lo cual no es de extrañar dada la falta de consenso existente entre los expertos en el ámbito de las colocaciones), pero que comparten la idea esencial de que estas combinaciones no son meras asociaciones de palabras cuantificables estadísticamente, sino que se caracterizan por una serie de propiedades que las convierten de una u otra manera en unidades fraseológicas en el sentido más tradicional del término, es decir, en entidades *con cierto grado* de fijación y arbitrariedad combinatoria. Esta idea de que existe una gradación entre las distintas unidades multiléxicas de la lengua en términos de fijación e idiomaticidad, noción que constituye una de las principales aportaciones de esta aproximación al estudio fraseológico, ha sido heredada de las teorías de los fraseólogos rusos Victor V. Vinogradov y Natalya N. Amosova, precursores de este enfoque. También de la tradición rusa proviene la segunda característica principal de este enfoque: la búsqueda de una clasificación exhaustiva de todas las unidades fraseológicas que existen en las lenguas, estableciendo criterios válidos para diferenciar entre unos elementos y otros.

Como ya hiciéramos en el enfoque estadístico, prestaremos aquí especial atención a las aportaciones de los autores que se consideran (Nesselhauf, 2005; Gyllstad, 2007) más representativos de la visión fraseológica de la colocación: Anthony Cowie, Peter Howarth e Igor Mel'čuk, si bien es cierto que son muchos

[4] También se suele emplear el adjetivo "lexicográfico" para designar este enfoque puesto que los autores que han adoptado esta aproximación fraseológica son comúnmente lexicógrafos y/o especialistas en lingüística aplicada. En nuestra opinión no resulta muy adecuado utilizar esta denominación ya que el proyecto COBUILD, dirigido por Sinclair, también está destinado a fines lexicográficos, pudiendo por lo tanto llevar a confusión (Nesselhauf, 2005).

más los que se podrían sumar a este estudio (por ej. Hausmann, 1989; Benson, Benson e Ilson, 1986; Corpas Pastor, 1996; Granger, 1998; Koike, 2001).

1.3.1.2.1. Anthony P. Cowie y Peter A. Howarth

A pesar de que las investigaciones de estos dos autores son de tal relevancia en lo que a las colocaciones se refiere que merecerían un tratamiento individual en esta breve revisión de los estudios colocacionales, los aspectos en los que sus concepciones del fenómeno y sus trabajos coinciden son tantos que se pueden tratar de manera conjunta (Pérez Fernández, 2002). En primer lugar, tanto Anthony Cowie como su discípulo en la Universidad de Leeds, Peter Howarth, se caracterizan por el carácter aplicado de sus investigaciones fraseológicas. Por un lado, los estudios llevados a cabo por Cowie tienen como objetivo principal la compilación de diccionarios fraseológicos, dada la orientación lexicográfica de todos sus trabajos. La investigación que Howarth realiza, por otro lado, está encaminada hacia la lingüística aplicada y, en cierta medida, el análisis de corpus, ya que sus trabajos se dedican al estudio de la fraseología en la escritura académica de los alumnos de inglés como lengua extranjera. Además, ambos comparten un gran interés por el aspecto psicolingüístico de la fraseología y, por ende, de la colocación (si bien es cierto que este interés está más acentuado en los estudios de Howarth que en los de Cowie), estando sus apreciaciones muy en consonancia con las de Bolinger (1976) y Pawley y Syder (1983). Otro de los elementos comunes en sus estudios en materia colocacional es la reticencia que ambos muestran al enfoque estadístico. En palabras de Cowie (1998: 226),

> as regards recurrence as a measure of restrictedness, there is no clear evidence to date of a close correlation between measured frequency of occurrence and collocational restriction. Fruitful discussion of the issue is to some extent

> frustrated at present by the insistence of some scholars involved in the computerized analysis of large corpora that frequency of co-occurrence is the only significant measure of 'conventionality' in language (e.g. Sinclair, 1991).

Por su parte, Howarth (1996: 74) reconoce el valor de la investigación mediante corpus, pero advierte de que la frecuencia de co-aparición como único criterio de identificación es insuficiente:

> [T]he pure form of a computational analysis depending on statistical significance is not appropriate. (...) However, the argument is not against an empirical approach making use of corpus data, but against a reliance on measures of relative frequency derived from the automatic recognition of identical co-occurring forms.

Pero si existe un aspecto fundamental en el que estos dos investigadores se muestran especialmente de acuerdo, por el cual merecen ser tratados conjuntamente, éste es su concepto teórico de colocación. Como ya dijimos, uno de los objetivos más importantes para los estudiosos de esta corriente es la clasificación de las distintas unidades multiléxicas que existen. Así, para estos autores, las unidades fraseológicas se dividen en las llamadas "*formulae*" y "*composites*"[5]. Las primeras son expresiones, generalmente oracionales, cuyo significado en el lenguaje refleja su valor discursivo (saludos, invitaciones, preguntas, etc.), siendo por tanto unidades con una función principalmente pragmática. Ejemplos de este tipo de expresiones son "*good morning*", o "*are you following me?*". Las combinaciones denominadas *composites*, por su parte, son unidades que han desarrollado un valor semántico referencial más o menos unitario o idiomático, dentro de las cuales se encontrarían las colocaciones. El término genérico *composite* engloba, según ambos autores, cuatro tipos diferentes de combinaciones

[5] Como el propio Cowie (1988: 133-4) indica, las "*formulae*" ya fueron recogidas por otros autores aunque bajo diferentes denominaciones. Keller (1979) se refería a ellas como "*gambits*" mientras que Coulmas (1981) prefería la voz "*routine formulas*". En cuanto al término "*composites*", éste proviene de la obra de Mitchell (1971).

léxicas distribuidas a lo largo de una escala continua dependiendo de su grado de idiomaticidad. Howarth (1996: 48), ilustra la clasificación de estas combinaciones de la siguiente forma (Tabla 1.1.):

Tipo de combinación	**Definición**	**Ejemplo**
Free collocation	Combinaciones de dos o más palabras donde los elementos se usan en su sentido literal. Cada componente se puede sustituir sin afectar al significado del otro.	*"blow a trumpet"*
Restricted collocation	Combinaciones en las que un componente se usa en su sentido literal, y el otro en un sentido especializado. Este sentido especializado puede ser figurativo, delexicalizado o técnico de alguna forma, y determina de manera importante la limitación en su colocabilidad con el otro elemento. Son combinaciones, sin embargo, totalmente motivadas[6].	*"blow a fuse"*
Figurative idiom	Combinaciones que tienen significados figurados en términos del conjunto. Pueden permitir la sustitución arbitraria de uno o varios de sus elementos por sinónimos. Tienen una interpretación literal y están claramente motivadas.	*"blow your own trumpet"*
Pure idiom	Combinaciones que tienen un significado global que no se puede derivar de los significados de sus componentes. No permiten casi ninguna sustitución, y no están motivadas.	*"blow the gaff"*

Tabla 1.1.: Clasificación de los *composites*

[6] Los términos "*motivated*" y "*unmotivated*" que Howarth (ibid.) usa en sus definiciones provienen de las teorías de Vinogradov (1947, citado por Durrant, 2008) y se deben interpretar como "semánticamente transparente" y "semánticamente opaco" respectivamente.

Lo primero que llama nuestra atención en esta clasificación idiomática es la distinción que se plantea entre lo que a simple vista parecen dos tipos diferentes de colocaciones: *free collocations* y *restricted collocations*. Sin embargo, tal y como explican ambos autores, ellos utilizan indistintamente los términos "*free collocations*" y "*open collocations*" para designar lo que otros investigadores denominan "*free word-combinations*" (Aisenstadt, 1979) o "*free combinations*" (Benson, Benson e Ilson, 1986)[7]. Se trata, por tanto, de combinaciones libres de palabras y no de un tipo de colocaciones diferente del que comúnmente se estudia en fraseología y que ellos denominan "*restricted collocation*". De acuerdo con Howarth (1996), la razón por la cual prefieren el término "colocación" para referirse a la combinación libre es que éste les permite situarla dentro del continuo idiomático como combinación léxica, reforzando así el concepto de gradación. De este modo, co-apariciones de palabras que no cumplen las mínimas reglas sintácticas de las combinaciones léxicas (por ejemplo "*again but*" en la oración "*I have gone again but he isn't there*") quedan automáticamente excluidas de sus teorías. En nuestro caso, no obstante, preferimos reservar el término "colocación" para aquellas unidades en las que realmente existe algún grado de restricción arbitraria, es decir, para las denominadas "*restricted collocations*" por estos autores. Las estructuras, sea cual sea su naturaleza sintáctica, que no estén sujetas a ningún tipo de restricción idiomática se designarán "combinaciones libres".

La definición de colocación propuesta por Cowie y Howarth viene dada en contraposición a los otros tres elementos que componen el continuo fraseológico según su propia tipología (Tabla 1.1.). Aunque, como acabamos de ver, toman en consideración criterios como la sintaxis o la institucionalización para distinguir lo que

[7] A la vista de que incluso un mismo autor emplea dos expresiones diferentes para designar el mismo concepto, resulta comprensible la falta de consenso que existe en la terminología referente a las colocaciones. Cowie, por ejemplo, utiliza comúnmente el término "*free collocation*" (1998), pero en algunas ocasiones las denomina "*free combinations*" (1994).

es un *composite* de lo que no lo es, estos aspectos no son discriminantes a la hora de distinguir el grado de restricción idiomática de una combinación léxica. Esto se debe a que, en su opinión, prácticamente todas son igualmente correctas y aceptables desde el punto de vista sintáctico y gramatical y todas están en mayor o menor medida institucionalizadas en su uso, a excepción de algunas combinaciones libres que se utilizan de manera improvisada en situaciones muy concretas y que no es probable volver a oír.

Así, los dos criterios fundamentales de identificación y clasificación de *composites* son la **transparencia semántica** y la **conmutabilidad**, dos aspectos que, según estos autores, interactúan entre sí. Mediante la noción de transparencia semántica, Cowie y Howarth se refieren a si las palabras que integran una combinación y/o la combinación en su conjunto tienen un significado literal o no. En cuanto a la conmutabilidad, ésta designa la posibilidad de sustituir un elemento de la combinación por otro sin que se produzca una alteración en el significado del resto de elementos de la combinación o de la expresión en su conjunto.

Como se desprende de la tabla 1.1., de acuerdo con estos criterios una colocación es, a diferencia de una combinación libre donde todos sus elementos se utilizan en sentido literal, una expresión en la que al menos uno de los elementos que la componen no tiene un significado literal o primario mientras que el otro elemento sí se emplea de modo semánticamente transparente. Además, mientras que una combinación libre permite la sustitución de todos sus elementos con la única restricción de la lógica y la semántica, en una colocación existen ciertas restricciones de conmutabilidad que sólo responden a la arbitrariedad del uso lingüístico. Por ejemplo, la combinación "*to cut bread*" es claramente una combinación libre puesto que, por un lado, tanto sus componentes como la expresión en su conjunto son semánticamente transparentes y, por otro, permite una conmutabilidad total de elementos sólo limitada por las leyes semánticas: "*to cut cheese*", "*to eat bread*", etc. En

cambio, la expresión "*to foot the bill*" se debe identificar como colocación dado que "*foot*" adopta un significado secundario o figurado y su conmutabilidad está arbitrariamente restringida.

Por otro lado, una colocación se diferencia de una locución idiomática en que la primera es una combinación léxica que no forma una unidad semántica independiente, sino que su significado es derivable de los significados de sus partes integrantes (es decir, tiene un significado composicional), mientras que en una locución sí se identifica un significado global diferente del que daría la suma de los significados de los elementos que la integran. En cuanto a la conmutabilidad, mientras que en una locución la sustitución de los elementos integrantes es imposible o casi imposible (dependiendo de si se trata de una locución idiomática pura o figurada respectivamente), en el caso de la colocación esta sustitución léxica sí es posible en la mayoría de los casos —aunque dentro de las propias colocaciones se establecen varios grados de conmutabilidad desde las más flexibles donde ambos elementos son sustituibles, como por ejemplo *carry out/conduct an experiment/a test/a survey*, hasta las más restringidas donde ni el elemento literal ni el figurado se pueden sustituir, por ejemplo *curry favour* (Howarth, 1996: 43). En cualquiera de estos casos, se puede apreciar que la conmutabilidad que suele caracterizar a la colocación siempre está sujeta a unas restricciones arbitrarias impuestas por el uso. A modo de resumen, podríamos esquematizar la aplicación de estos dos criterios como sigue (Tabla. 1. 2.):

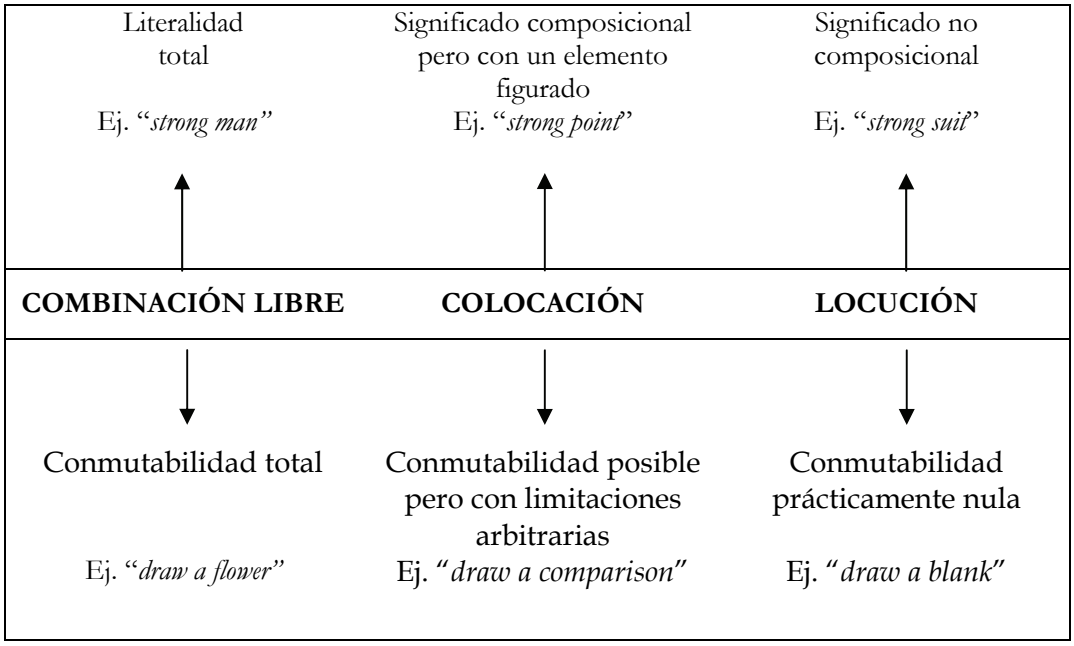


Tabla 1.2.: Combinación libre, colocación y locución

Finalmente, nos gustaría destacar que otro de los aspectos en el que tanto Cowie como Howarth se muestran de acuerdo en lo que respecta a la colocación es su estructura formal, compuesta de dos partes: la **base** y el **colocado**. Desde que Hausmann (1979) estableciera esta terminología[8], son mayoría los autores de la tradición fraseológica que la han adoptado (Cop, 1988; Corpas, 1992; Nesselhauf, 2005), si bien cabe destacar que otros investigadores utilizan una denominación distinta como es el caso de Mel'čuk, que, como veremos a continuación, prefiere designarlos "palabra clave" y "valor" respectivamente. A diferencia de lo que ocurría con el "nodo" y el "colocado" tal y como se entienden en la tradición estadística, la base y el colocado fraseológico son dos elementos que se encuentran a distinto nivel, siendo uno el elemento dominante, y restringiendo el uso del otro. Así, la base se

[8] En sus trabajos, Hausmann empleó los términos "base" y "collocatif" en francés y "base" y "kollokator" en alemán.

caracteriza por su autonomía semántica dado que es la parte de la colocación que tiene un valor propio totalmente independiente del colocado y es el elemento que selecciona el hablante en primer lugar y libremente en el uso de la lengua. Podemos decir, pues, que es el núcleo de la colocación y el elemento a partir del cual se puede predecir la presencia del otro. Por su parte, el colocado es el elemento cuya elección está determinada en gran medida por la base y, por tanto, no cobra su valor completo sin su dimensión sintagmática. Especialmente desde un punto de vista didáctico, esto hace que, como destaca Pérez Fernández (2002: 35), sea importante "insistir en que las colocaciones no se pueden aprender por separado sino como un 'todo'".

A modo de conclusión, creemos que a la vista de la definición y clasificación del fenómeno fraseológico descritas por Cowie y Howarth, resulta evidente desde nuestro punto de vista que las unidades fraseológicas se distribuyen a lo largo de un continuo en el que no existen fronteras claramente definidas sino una constante gradación que no debe medirse en términos de sí o no, sino de más o menos. Éste es, de hecho, uno de los aspectos en los que mayor énfasis ponen tanto Cowie como Howarth, destacando que lo difuso de las fronteras entre las distintas unidades es una de sus características inherentes y un aspecto omnipresente en la fraseología.

1.3.1.2.2. Igor Mel'čuk

La clasificación fraseológica establecida por Igor Mel'čuk coincide, aunque sólo en su nivel más general, con la propuesta por Cowie y Howarth. Este fraseólogo, cuya principal motivación es también lexicográfica, considera que las unidades fraseológicas se dividen en dos grandes esferas, que él denomina "pragmatemas" y "frasemas semánticos". De una manera simplificada podríamos decir que los pragmatemas son combinaciones léxicas cuya restricción viene dada por razones de

propiedad pragmática, siendo por tanto un elemento muy similar a las "*formulae*" de Cowie. Mel'čuk (1998: 28) cita como ejemplo la expresión "*Caesar Salad: All you can eat*", donde otras expresiones igualmente correctas desde el punto de vista gramatical y semántico como, por ejemplo, "#*Caesar Salad: As much as you like*"[9], no se emplean debido a razones puramente pragmáticas.

De otro lado, los frasemas semánticos, de acuerdo con Mel'čuk (ibid: 29-30), son unidades en las que existe una restricción sintagmática de naturaleza semántica, lo que sin duda refleja un concepto muy similar a las "*composites*" de Cowie. Dependiendo del tipo de restricción semántica, este tipo de frasemas se puede clasificar en:

- **Frasemas completos** o **locuciones idiomáticas**: unidades en las que el significado final no incluye el significado de ninguna de sus partes individuales. La fórmula de Mel'čuk (expresada aquí de manera simplificada) para representar esta unidad léxica es: A+B= C. Ejemplos: *[to] spill the beans*, *of course*, *[a] red herring*, etc.
- **Semi-frasemas** o **colocaciones**: unidades en las que un elemento A conserva su significado intacto y es elegido libremente por el hablante mientras que el elemento B está limitado semánticamente cuando aparece junto con A y su elección depende de la elección de A. La fórmula propuesta por Mel'čuk en este caso es A+B= A+C. Ejemplos: *[to] crack a JOKE*[10], *strong COFFEE*, *[to] launch an ATTACK*, etc.
- **Quasi-frasemas** o **quasi-locuciones idiomáticas**: unidades en las que los significados de ambos elementos se mantienen, pero existe además un elemento semántico añadido de naturaleza impredecible. La fórmula que lo expresa sería

---

[9] Empleamos el símbolo # tal y como lo usa Mel'čuk (ibid.) para indicar que una expresión es pragmáticamente inapropiada.

[10] Las palabras en letras versales son el elemento A. Expresado en versales en el texto original.

A+B= A+B+C. Ejemplos: *[to] give the breast [to N]*, *bacon and eggs*, *shopping centre*, etc.

En lo que respecta a las colocaciones de manera más concreta, Mel'čuk destaca que el hecho de que B signifique C sólo cuando aparece en combinación con A se puede dar en una serie de casos:

- B no tiene en el diccionario el significado C que adopta en la colocación y:
    - B es una palabra vacía seleccionada por A como auxiliar. Es lo que se denomina verbos "ligeros" (*light verbs*) o "construcciones de verbo soporte" que sólo sirven de apoyo al nombre sin poseer una carga semántica propia (Gross, 1981; Catell, 1984 citado por Mel'čuk, ibid: 34). Ejemplos: *[to] give a LOOK*, *[to] lay SIEGE [to N]*, etc.
    - B tiene su propio significado pero sólo adopta el significado C cuando se combina con A. Ejemplos: *black COFFEE*, *French WINDOW*, etc.
- B tiene en el diccionario el significado C y:
    - cuando B aparece con A no existe un sinónimo de B por el que éste se pueda intercambiar. En este caso, como vemos, Mel'čuk no considera el factor semántico sino que en realidad sólo toma la conmutabilidad de los elementos como criterio identificativo de la colocación. Ejemplos: *strong (*powerful) COFFEE, heavy (*weighty) SMOKER*, etc.
    - el significado de B incluye una parte importante de A por lo que B es muy específico y está totalmente ligado a A. En nuestra opinión, éste es un fenómeno más cercano a la solidaridad léxica que a la

colocación (ver sección 1.5.). Ejemplos: *rancid BUTTER*[11]*, the HORSE neighs*, etc.

Con el objetivo de alcanzar una mayor rigurosidad y sistematicidad en su definición del concepto de colocación, Mel'čuk estableció, dentro de la Teoría Sentido-Texto (*Meaning-Text Theory*), su ya famosa noción de las Funciones Léxicas. La Teoría Sentido-Texto, desarrollada por Mel'čuk, Žolkovskij y Apresjan en los años sesenta, tiene como objetivo principal la descripción de la lengua

> as a kind of logical device which associates with any given meaning M the set of all the texts in this language which are the expression of M (and which are consequently synonymous with one another), and with any text T, the set of all the meanings which are expressed by T (and which are so to speak, homonymous with one another) (Gérardy, 1995 citado por Pérez Fernández, 2002: 55).

Mediante esta teoría se pretenden crear reglas que permitan obtener de manera general el mayor número de significados expresados por una sola oración y, lo que es más relevante desde el punto de vista colocacional, el mayor número de expresiones para articular un significado considerablemente abstracto. Es dentro de esta teoría lingüística donde nace, ya a nivel de la palabra, la noción de **función léxica** introducida por primera vez por Žolkovskij y Mel'čuk (1967, citado por Mel'čuk, 1998) para el estudio, definición y clasificación de las relaciones paradigmáticas (de quasi-sinonimia) y sintagmáticas (colocacionales) existentes en el léxico.

La función léxica, basada en el sentido matemático de función donde **f**(x) = y, es definida por Mel'čuk (1998: 32) como

---

[11] Este ejemplo, citado por el autor, no se corresponde realmente con esta categoría y para nosotros no conformaría una solidaridad léxica, puesto que "*rancid*" no está restringido de forma exclusiva a "*butter*". También se combina en colocaciones como "*rancid fat*" o "*rancid oil*".

> a function that associates with a specific lexical unit (...), L, which is the 'argument', or 'keyword', of f, a set $\{L_i\}$ of (more or less) synonymous lexical expressions — the 'value' of f— that are selected contingent on L to manifest the meaning corresponding to f:
> $f(L) = \{L_i\}$.
> To put it differently, an LF [Lexical Function], (...), is a very general and abstract meaning, (...), which can be lexically expressed in a large variety of ways depending on the lexical unit to which this meaning applies.

La función léxica expresa la relación de dependencia semántica existente entre su "palabra clave" y su "valor", elementos que equivalen respectivamente a la base y el colocado que componen una colocación (ver sección 1.3.1.2.1.). La finalidad principal de las funciones léxicas es llevar a cabo una clasificación sistemática de todas las colocaciones existentes (además de incluir también otras relaciones léxicas como las paradigmáticas) mediante la distribución de combinaciones según el significado general que expresen y que represente su denominador común. Sin embargo, existen significados que Mel'čuk considera "no estándar" por ser demasiado específicos y limitados en su uso. En cuanto a las funciones léxicas estándar, este autor ha establecido un total de 64 (Mel'čuk, 1996), todas ellas con nombre de raíz griega o latina. A continuación ofrecemos algunos ejemplos de funciones estándar acompañadas de varias de sus manifestaciones:

Magn [Lat. magnus]: "*intense(ly), very*".

Magn *(thin)* = *as a rake* **Magn** *(naked)* = *stark*

Magn *(to rely)* = *heavily* **Magn** *(thirst)* = *unquenchable*

Func [Lat. functionare]: "*to function*".

Func *(snow)* = *falls* **Func** *(analysis)* = *concerns*

Func *(silence)* = *reigns* **Func** *(proposal)* = *comes, stems*

Oper [Lat. operari]: "*to do, to carry out*".

Oper *(support)* = *lend* **Oper** *(order)* = *give*

Oper *(resistence)* = *put up* **Oper** *(sacrifice)* = *make*

De acuerdo con Mel'čuk (1998), esta tipología abre el camino a una generalización semántica de las colocaciones que podría ser muy útil en los ámbitos de la enseñanza de lenguas, la lexicografía y la lingüística computacional (especialmente en la traducción automática). En la misma línea, Cowie (1998: 8) subraya que las funciones léxicas "have much to contribute to the design of collocational dictionaries, where a persistent weakness is the failure to specify adequately the semantic categories to which collocates belong".

Debemos destacar, sin embargo, que la aproximación de Mel'čuk al fenómeno colocacional ha suscitado también considerables críticas por parte de algunos autores. El principal inconveniente de esta teoría es, desde nuestro punto de vista, que algunas funciones léxicas producen combinaciones libres en lugar de colocaciones. Benson (1985) cita como ejemplo la función **Caus** [Lat. causare] ("*to cause, to bring about*"), que produce combinaciones como "*to grow corn*" o "*to build houses*", que son claramente predecibles y restringidas únicamente por la semántica. Asimismo, las funciones léxicas no sólo recogen relaciones colocacionales sino que dan lugar a relaciones morfológicas y también léxico-semánticas como son la sinonimia, la antonimia o la hiponimia. Como apunta Fontenelle (1994: 46), "[l]exical functions then cover the whole of the paradigmatic and syntagmatic relations on an entry word". Por citar sólo un ejemplo, resulta evidente que en la función que sigue,

S_0 ("*nominalization*")

S_0 (reject) = rejection

S_0 (apply) = application

las combinaciones obtenidas no se pueden considerar en modo alguno colocaciones sino que establece una relación morfológica derivativa entre lexemas.

Mencionaremos, por último, la apreciación hecha por Howarth a propósito de la aproximación de Mel'čuk, opinión que compartimos plenamente a la vista de los inconvenientes que acabamos de registrar. Para Howarth (1996: 45),

> these functions seem to be too general and applied too broadly. They are not sufficiently associated with specific lexical items to reflect the co-occurrence potential of such lexical items. It must be noted that Mel'čuk's interest is in a highly formalized lexicographic approach rather than the description of natural language data.

En términos generales, podemos comprobar que, como ya adelantamos, uno de los aspectos más destacables de esta corriente es su interés por establecer criterios válidos para definir las colocaciones, siendo los más utilizados los de restricción semántica y conmutabilidad. Sin embargo, al igual que sucedía con la aproximación estadística, esta visión no está, en nuestra opinión, exenta de ciertas áreas problemáticas. Para empezar, no todos los autores otorgan el mismo valor a ambos criterios. Así, podemos encontrar investigadores, entre los cuales debemos destacar de forma especial a Howarth (1996: 39), que sí consideran que ambos criterios van siempre de la mano e interactúan entre sí:

> [I]t is in practice very difficult to discuss the semantic nature of restricted collocations in isolation from the question of commutability. (…) If the continuum is accepted as a valid model, it is clear that in the central area there will be a tension between these two features: as the semantic transparency of an individual item decreases, its sense becomes more specialized, which coincides with a limit on the lexical contexts in which that sense is found and therefore the number of collocates.

A diferencia de Howarth, otros autores consideran que uno de los dos criterios predomina sobre el otro, aunque existen también discrepancias sobre cuál de los dos es el principal. Por un lado, los representantes de la tradición lexicológica soviética favorecían el valor semántico ya que uno de sus principios básicos es que la mayoría de las palabras son polisémicas, y sus diferentes significados se definen "according to the grammatical and phraseological context in which they occur" (Weinreich, 1963: 67). Así pues, estos autores entienden la colocación como un binomio que contiene "one component used in its direct meaning while the other is used figuratively: *meet the demand*, *meet the necessity*, *meet the requirements*" (Arnold, 1986, citado por Cowie, 1998: 215).

Por otro lado, autores como el propio Cowie (1994: 3169) son de la opinión de que la principal característica de la colocación es la restricción de su conmutabilidad: "[collocations are] characterized by arbitrary limitation of choice at one or more points". Partiendo de esta base, este investigador considera que el significado figurado de uno de sus componentes es un rasgo frecuente, pero no necesario, en la colocación. Así, reconoce como colocaciones aquellas combinaciones donde un elemento tiene un sentido figurado, pero también aquellas en las que todos sus elementos se usan de forma literal, como por ejemplo "*cut (*slash) one's throat*" o "*slash (*cut) one's wrists*". Mel'čuk, por su parte, muestra una aproximación claramente semanticista al fenómeno fraseológico. Sin embargo, como acabamos de ver, incluye en su taxonomía combinaciones de palabras que mantienen su significado literal (aunque quizá no sea el significado primario) pero que no se pueden sustituir por otros sinónimos, lo cual indica claramente que, en su opinión, no siempre se tienen que dar las dos restricciones a la vez para formar una colocación. Por último, cabe citar en este sentido la aportación de Nesselhauf (2005), quien, dando un paso más allá, aboga por eliminar definitivamente el criterio semántico. Esta autora argumenta que no siempre coinciden ambos criterios (por ejemplo en la combinación "*commit a*

crime" existe restricción en la conmutabilidad, pero ambos elementos se emplean en su sentido literal según reflejan los diccionarios). Además, y como el propio Howarth (1998a) reconoce, la aplicación del criterio semántico presenta otro problema importante ya que es verdaderamente difícil en ciertos casos determinar cuándo se está usando una palabra en su significado literal y cuándo en sentido figurado. Así pues, Nesselhauf (ibid.) utiliza la restricción arbitraria en la conmutabilidad de los elementos como único criterio para diferenciar las colocaciones de las combinaciones libres de palabras por un lado, y de las locuciones idiomáticas por otro.

A la luz de todo lo anterior, resulta evidente que nos encontramos, sin duda, ante un fenómeno muy problemático de la lengua, de naturaleza y fronteras difusas, que sigue planteando numerosas incógnitas a los lingüistas. Es también muy interesante notar cómo, a pesar de que las dos aproximaciones recogidas hasta el momento se han interpretado como dos posturas totalmente contrapuestas en muchos de los estudios dedicados a este fenómeno (Herbst, 1996; Nesselhauf, 2004, 2005), el papel predominante que parece estar tomando el criterio de conmutabilidad en los estudios fraseológicos está relacionado en cierta medida con los análisis de probabilidad combinatoria que priman en la aproximación estadística. De hecho, como destaca Durrant (2008: 32), "restrictions on substitution must ultimately be evidenced empirically, and it may be that frequency measures which are capable of measuring the degree of mutual predictability between words (...) will prove a reliable way of achieving this". Este posible punto de encuentro entre ambas corrientes habrá de ser investigado con mayor profundidad en el futuro, aunque son ya varios los autores que han observado la existencia de estas conexiones en algunos estudios recientes (Granger y Paquot, 2008). En el próximo apartado describiremos muy brevemente este fenómeno.

1.3.1.3. La tercera vía: Una aproximación mixta

En los estudios de fraseología más recientes (Gyllstad, 2007; Granger y Paquot, 2008) se están comenzando a poner de manifiesto las ventajas de una nueva aproximación al área de las colocaciones, originada por aquellos autores que, partiendo de uno de los dos enfoques mencionados anteriormente, hacen uso también de criterios o metodologías de trabajo propias del otro. La aparición de esta aproximación mixta parece, en realidad, una evolución natural en los estudios de colocaciones, si se tiene en cuenta las limitaciones que ambos enfoques presentan por separado. En este sentido, y como acabamos de ver, la aproximación puramente estadística presenta claras dificultades a la hora de identificar las colocaciones, puesto que su definición se fundamenta de forma exclusiva en el criterio de frecuencia, dejando pues al margen las cuestiones relacionadas con la naturaleza idiosincrásica de las colocaciones que, en nuestra opinión, forman una parte esencial de este fenómeno. Por otro lado, el enfoque fraseológico tampoco parece haber dado lugar por ahora a una definición consensuada de lo que se debe entender por colocación ni de las características que le son propias. A esto debemos sumar, además, que, a diferencia de lo que ocurre en los estudios basados en el enfoque sinclairiano, esta aproximación no ofrece un método operativo y sistemático de identificación y recopilación de colocaciones, siendo pues la labor manual el único modo de recopilarlas de forma exhaustiva. A tenor de esta situación, como decimos, parece muy comprensible que se estén desarrollando esfuerzos por aunar las ventajas de ambos enfoques o, en palabras de Gyllstad (2007: 16), por recoger "the best of two worlds". Ésta es, de hecho, la tendencia que sigue el trabajo que presentamos en esta tesis como se verá en las secciones siguientes.

En este tercer enfoque al estudio de las colocaciones, podemos distinguir entre aquellos autores que partiendo de una visión principalmente estadística, contemplan sin embargo algunos aspectos lingüísticos más en consonancia con la

aproximación tradicional, y los que realizan el recorrido inverso, es decir, que partiendo de una concepción fraseológica hacen uso de los procedimientos y herramientas que ofrece el enfoque distribucional. Entre los primeros, Gyllstad (ibid.) cita los trabajos de Mitchell (1966, 1971), Greenbaum (1970, 1974), Kjellmer (1984, 1994), Stubbs (1995) y Altenberg (1993). Aunque estos autores se pueden clasificar fundamentalmente en el enfoque estadístico, todos ellos plantean de manera explícita o implícita la necesidad de reparar en cuestiones gramaticales y sintácticas a la hora de identificar y/o clasificar las colocaciones. Se trata, como vemos, de trabajos donde los investigadores van más allá del análisis cuantitativo desprovisto de nociones teóricas preconcebidas por el que abogaba Sinclair. (Recordemos, en este sentido, la afirmación recogida más arriba, en la que este autor (1991: 170) declara: "The attention is concentrated on lexical co-occurrence, more or less independently of grammatical pattern or positional relationship"). Como Greenbaum (1970: 10-11) argumenta, Sinclair defiende una postura demasiado sujeta a la palabra (o "*item-oriented*" en sus propias palabras) que no contempla por tanto restricciones sintácticas de las colocaciones como por ejemplo el hecho de que "*much*" coloque con "*prefer*" cuando el primero precede al segundo (como en "*Some people much prefer wine*") pero no cuando la posición es inversa (como en "**Some people prefer wine much*").

En lo que se refiere a aquellos investigadores que parten de posiciones cercanas a la fraseología tradicional pero que adoptan algunas nociones o procedimientos propios del enfoque estadístico, podemos citar, por ejemplo, el trabajo de Nesselhauf (2005: 1). Esta autora parte de una marcada visión fraseológica, definiendo las colocaciones como "arbitrarily restricted lexeme combinations". A pesar de que, como ya vimos, el criterio primordial para esta investigadora es la arbitrariedad en la conmutación, en su estudio también se contempla la frecuencia de las combinaciones como un parámetro digno de estudio.

Pero si hay un estudio que merece especial atención por la novedosa fusión de criterios que plantea, ése es el llevado a cabo por Handl (2008). Esta autora indica al comienzo de su trabajo que su punto de partida es el enfoque estadístico, algo que se corrobora al comprobar que utiliza terminología propia de esta tendencia como "*node*" y "*co-text*". Sin embargo, en su estudio, Handl demuestra un verdadero eclecticismo, ya que propone estudiar las colocaciones de acuerdo con tres criterios diferentes distribuidos en tres escalas. Ya de partida, el hecho de plantear los parámetros de análisis en escalas graduadas supone un claro acercamiento a la tradición fraseológica. Más aún, los tres criterios que esta autora contempla son la transparencia semántica, la conmutabilidad, que Handl (ibid.: 54) denomina "collocational range" y la distribución y atracción estadística. Como podemos comprobar, para esta investigadora los principales criterios fraseológicos conviven con la importancia de la frecuencia de co-aparición de los elementos y las técnicas cuantitativas. Especial mención merece el último criterio de los mencionados, la atracción estadística, puesto que representa en sí mismo un novedoso intento de aproximar ambas tradiciones. A diferencia de los análisis de frecuencia significativa empleados comúnmente en el ámbito de los estudios estadísticos, Handl propone integrar la noción de asimetría que existe entre los dos elementos de la colocación (lo cual nos recuerda a los conceptos de base y colocado de la tradición fraseológica) para comprobar estadísticamente cuál de los componentes es el que ejerce atracción sobre el otro y si dicha atracción se repite además con la suficiente frecuencia en el corpus. En nuestra opinión, este tipo de análisis parece estar encaminado hacia una verdadera fusión de las dos corrientes, y sin duda abre un camino muy prometedor en el estudio de la colocación.

A nuestro juicio, de todo lo anterior se desprende la necesidad de establecer unos criterios claros que nos sirvan tanto para definir la propia naturaleza de las

colocaciones como para poder distinguirlas de otras combinaciones fronterizas. Sin duda, y como salta a la vista teniendo en cuenta la situación dibujada en las páginas anteriores, ésta no es tarea fácil en absoluto, pero sí es un paso imprescindible para poder llevar a cabo un estudio exhaustivo de este tipo de combinaciones léxicas. Por otro lado, en lo que respecta a los distintos enfoques desde los que se puede abordar este fenómeno, nos parece interesante destacar nuestro total acuerdo con Granger y Paquot (2008: 41) cuando afirman que "[t]he emergence of a new approach to phraseology is proving to be of immense value to the field. (…) both sides have a great deal to gain from a rapprochement". En la sección que sigue a continuación, trataremos de exponer de forma clara nuestra posición en este sentido.

1.3.2. Concepción de la colocación en este estudio

En este trabajo, la perspectiva que adoptamos a la hora de abordar las colocaciones es eminentemente fraseológica, si bien podríamos adscribirnos a la aproximación mixta dado que la frecuencia, como se verá, conforma también un aspecto fundamental de este trabajo. Partiendo de este enfoque, y analizando las diferentes tipologías postuladas por sus principales representantes, Granger y Paquot (2008) consideran que las más influyentes, aunque de ninguna manera las únicas, son las taxonomías propuestas por Cowie (1988), Mel'čuk (1998) y Burger (1998, citado por Granger y Paquot, 2008). Ya conocemos las categorizaciones establecidas por los dos primeros autores (ver sección 1.3.1.2.), por lo que sólo nos detendremos en la clasificación postulada por Burger. Para este autor, las unidades multiléxicas se pueden agrupar de la siguiente forma, atendiendo a la función que cumplen en el discurso (Tabla 1.3.):

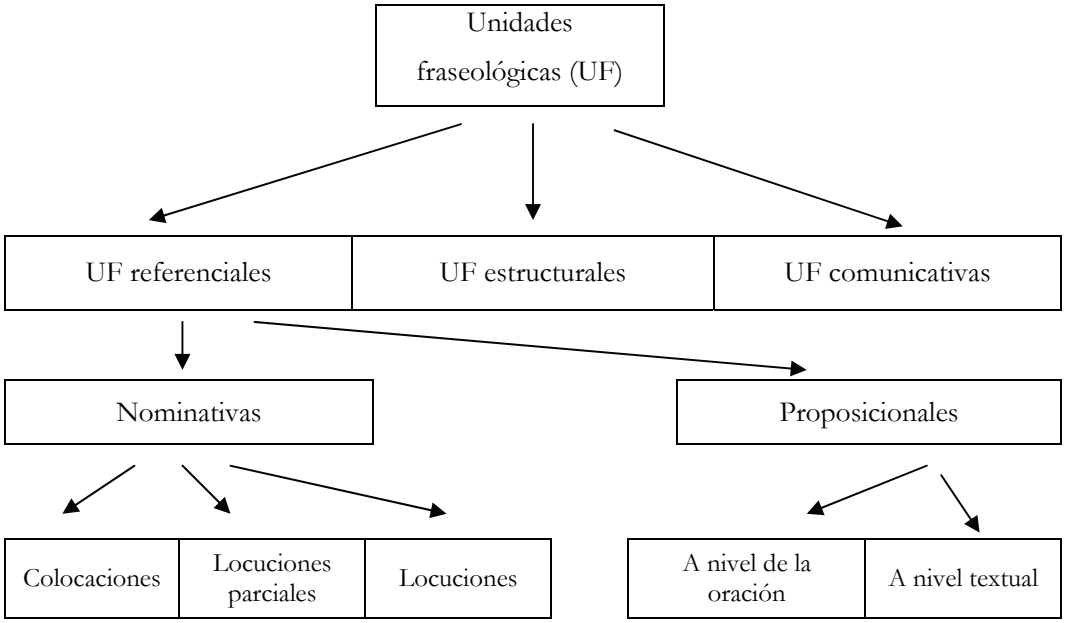


Tabla 1.3.: Tipología fraseológica de Burger (1998, citado por Granger y Paquot, 2008: 38)

Como vemos, y como apuntan Granger y Paquot, Burger considera que las unidades fraseológicas se pueden dividir en unidades referenciales, donde a su vez se distinguen unidades nominativas (elementos oracionales que se refieren a objetos, fenómenos o hechos de la vida, donde, siguiendo la tradición fraseológica, se incluyen las colocaciones, las locuciones y las locuciones parciales. Por ejemplo "*to rip somebody off*") y unidades proposicionales (afirmaciones u oraciones que se refieren a los objetos, fenómenos o hechos de la vida y donde se incluyen proverbios y oraciones idiomáticas del tipo "*the early bird catches the worm*"); unidades comunicativas, donde se incluyen las fórmulas utilizadas en la interacción con una función principalmente pragmática como por ejemplo "*Good morning*" o "*Well, I mean*…"; y, por último, unidades estructurales, una categoría creada de forma original por este autor para incluir unidades multiléxicas que establecen relaciones gramaticales tales como conectores del tipo "*as well as*" o "*concerning*". En términos generales, se trata de

una clasificación que se asemeja en buena medida al resto de tipologías postuladas por los autores del enfoque fraseológico (principalmente las de Cowie y Mel'čuk) si bien incluye algunos elementos novedosos que le otorgan un carácter más funcional, o menos formal, que los de otros autores, y también más completo.

Tomando como referencia la tipología establecida por este autor, Granger y Paquot (ibid.) han creado muy recientemente una clasificación propia que, desde nuestro punto de vista, es la más completa que existe actualmente, siendo por tanto la que adoptaremos en el presente trabajo. Para la elaboración de su planteamiento, estas autoras parten, como decimos, de la taxonomía de Burger, muy enraizada en la tradición fraseológica, pero llevan a cabo una notable mejora dado que incluyen "the new insights derived form the corpus-based approach" (ibid.: 42). Así, el marco fraseológico que postulan, y en el que se enmarcaría nuestro concepto de colocación es el siguiente (Tabla 1.4.):

Frasemas[12]

Función referencial	**Función textual**	**Función comunicativa**
Frasemas referenciales	Frasemas textuales	Frasemas comunicativos
Colocaciones (léxicas) Locuciones Bi- y trinomios irreversibles Símiles Compuestos Verbos frasales Colocaciones gramaticales	Preposiciones complejas Conjunciones complejas Adverbios conectores Fragmentos oracionales de organización textual (*Textual sentence stems*)	Fórmulas de actos de habla Fórmulas actitudinales Proverbios y fragmentos proverbiales Lugares comunes Lemas (*slogans*) Oraciones idiomáticas Citas

Tabla 1.4.: Tipología fraseológica de Granger y Paquot (2008: 42)

[12] A pesar de que estas autoras prefieren utilizar el término genérico "frasemas" ("*phrasemes*") en este trabajo se utilizarán los términos "unidades fraseológicas" y "unidades multiléxicas" por estar más generalizados en los estudios fraseológicos.

A la vista de la tabla anterior podemos comprobar que efectivamente se trata de una clasificación muy amplia y completa, donde se incluyen los distintos tipos de unidades que se contemplan en los estudios actuales. Como las propias autoras escriben (ibid.), las **unidades referenciales** son las que se emplean para expresar el contenido de un mensaje, refiriéndose a objetos así como a fenómenos y hechos de la vida real. Para mejor ilustrar el tipo de unidades incluidas en este apartado nos parece útil ofrecer una tabla con los ejemplos que Granger y Paquot citan para cada uno de los elementos (Tabla 1.5.):

Colocaciones (léxicas):	*heavy rain*; *closely linked*; *apologize profusely*
Locuciones:	*to spill the beans*; *to let the cat out of the bag*; *to bark up the wrong tree*
Bi- y trinomios irreversibles:	*bed and breakfast*; *kith and kin*; *left, right and centre*.
Símiles:	*as old as the hills*; *to swear like a trooper*.
Compuestos:	*black hole*; *goldfish*; *blow-dry*.
Verbos frasales:	*blow up*; *make out*; *crop up*.
Colocaciones gramaticales:	*depend on*; *cope with*; *a contribution to*; *afraid of*

Tabla 1.5.: Unidades referenciales ejemplificadas (Granger y Paquot, ibid.: 43)

En lo que se refiere a las **unidades textuales**, éstas contemplan secuencias gramaticalizadas que se usan para estructurar y organizar el contenido de un mensaje. En este apartado se comprenden las siguientes categorías (Tabla 1.6.):

Preposiciones complejas:	*with respect to*; *in addition to*; *apart from*; *irrespective of*
Conjunciones complejas:	*so that*; *as if*; *even though*; *as soon as*; *given that*
Adverbios conectores:	*in other words*; *last but not least*; *more accurately*; *what is more*; *to conclude*
Fragmentos oracionales de organización textual:	*the final point is*…; *another thing is*…; *it will be shown that*…; *I will discuss...* .

Tabla 1.6.: Unidades textuales ejemplificadas (ibid.: 44)

Por último, las **unidades comunicativas** se utilizan para expresar sentimientos o creencias acerca del contenido proposicional, o para dirigirnos a los interlocutores para llamar su atención, incluirlos en la conversación o influir en ellos de algún modo. Las categorías de este tipo de unidades se ejemplifican del siguiente modo (Tabla 1.7.):

Fórmulas de actos de habla:	*good morning!*; *take care!*; *happy birthday!*; *you're welcome*; *how do you do?*
Fórmulas actitudinales:	*in fact*; *to be honest*; *it is clear that*; *I think that*.
Proverbios:	*A bird in the hand is worth two in the bush*; *When in Rome*.
Lugares comunes:	*Enough is enough*; *We only live once*; *It's a small world*.
Lemas:	*Make love, not war*.

Tabla 1.7.: Unidades comunicativas ejemplificadas (ibid.)

Concretamente en lo que respecta a las colocaciones, el aspecto que aquí nos ocupa, Granger y Paquot (ibid.: 43) lo definen de la siguiente forma:

> (Lexical) collocations are usage-determined or preferred syntagmatic relations between two lexemes in a specific syntactic pattern. Both lexemes make an isolable semantic contribution to the word combination but they do not have the same status. Semantically autonomous, the 'base' of a collocation is selected first by a language user for its independent meaning. The second element, i.e. the 'collocate' or 'collocator', is selected by and semantically dependent on the base.

Como podemos observar, estas autoras tienen una concepción marcadamente fraseológica de la colocación. Aunque, en términos generales, nuestra definición se corresponde con la recogida en la cita anterior, el hecho de someter la diferencia entre base y colocado al criterio semántico presenta, como ya vimos, ciertas dificultades. Pero para tratar de todos los aspectos que conforman la naturaleza de la colocación tal y como se concibe en el presente trabajo de forma más detallada, vamos ahora a analizar las características formales y funcionales de la colocación.

1.3.2.1. Características formales

La principal característica de la colocación desde el punto de vista formal es, como ya hemos adelantado, su composición interna, formada por dos elementos, base y colocado, que no comparten el mismo estatus. Así, siempre existe un elemento, la base, que el hablante elige libremente y del cual parte a la hora de crear su mensaje, y otro, el colocado, que está arbitrariamente restringido por la base y es seleccionado por ésta.

Aunque no parece que existan dudas respecto al hecho de que la colocación es una combinación bipartita a nivel abstracto, es decir, desde el punto de vista de la asociación psicolingüística que se establece en nuestra memoria entre los dos elementos, los expertos no han alcanzado tal consenso a la hora de discernir si se trata de una combinación bimembre, es decir, si tanto la base como el colocado constan cada uno de una sola palabra. Por un lado, autores como Aisenstadt (1981),

Cop (1988), Cowie (1994) o Nesselhauf (2005) comparten la visión de Sinclair (1991) a este respecto cuando define la colocación como la co-aparición de "dos o más palabras" en un texto. Sin embargo, otros autores como Roos (1976) y Hausmann (1979) defienden que en la colocación sólo intervienen dos palabras. Quizá una de las razones de tal divergencia de opiniones es que, en muchos casos, cuando una colocación está formada por más de dos palabras se trata de combinaciones donde se solapan una colocación léxica (formada por dos palabras léxicas) y una gramatical (formada por una palabra léxica y otra funcional) (ver sección 1.4). Éste es el caso, por ejemplo, de la colocación "*take an interest in*" donde en realidad están presentes dos colocaciones: "*take an interest*" e "*interest in*". Si además de ello tenemos en cuenta que algunos autores no consideran los artículos determinados e indeterminados y las preposiciones como parte de la combinación, resulta evidente por qué restringen su concepción del fenómeno a combinaciones de dos palabras. En nuestro caso, en cambio, somos partidarios de una definición más flexible donde la colocación no tiene por qué estar limitada a dos ítems sino que puede incluir más elementos asociados, habida cuenta de que son elementos que siempre suelen estar presentes en el empleo de las palabras léxicas con las que colocan.

Otra de las cuestiones relativas a la colocación que suscita cierta controversia es la naturaleza de los ítems que componen la colocación. La opinión más generalizada entre los autores (Halliday, 1966; Mitchell, 1971; Greenbaum, 1974) es que los elementos que componen la colocación son lexemas o raíces, lo cual permite considerar que todas las inflexiones y derivaciones formadas a partir de un mismo lexema dan lugar a realizaciones de la misma colocación. Halliday afirma a este respecto: "(…) *strong*, *strongly*, *strength* and *strengthened* can all be regarded as the same item; and *a strong argument*, *he argued strongly*, *the strength of his argument* and *his argument was strengthened* all as instances of one and the same syntagmatic relation" (1966: 151).

Otros estudiosos (Gyllstad, 2007), sin embargo, estiman que las colocaciones están formadas por lemas, por lo que sólo consideran las inflexiones como realizaciones de una colocación. Según esta opinión, "*implement a method*" e "*implementing methods*" sí son ejemplos de la misma colocación, mientras que "*the implementation of a method*" no lo es. Una tercera visión, la más restrictiva, es la que considera que los elementos que componen la colocación son formas léxicas, lo que podríamos llamar sencillamente palabras. Esto implica que "*confirmed bachelor*" es una colocación diferente a "*confirmed bachelors*" (Kjellmer, 1994; Nesselhauf, 2005).

A nuestro juicio, en todos estos análisis parece perderse de vista la noción de que la base y el colocado son dos elementos distintos, que no se encuentran al mismo nivel léxico dado que uno funciona de forma libre mientras que el otro está restringido por el primero. En este sentido, las bases sí varían en ocasiones su forma de combinarse dependiendo de la forma léxica en la que aparezcan, confirmando con ello su primacía y su dominio sobre el elemento colocado. Así, y como están evidenciando los estudios de corpus más recientes (Sinclair, 2003; Evert, Heid y Spranger, 2004, Partington, 2007), podemos observar casos como por ejemplo "*eye*", que no suele exigir la presencia de los mismos colocados cuando se usa en singular y en plural, dando lugar a combinaciones que sólo se suelen dar cuando la base aparece en singular ("*to the naked eye*") y otras en las que la base suele utilizarse únicamente en plural ("*to roll your eyes*"). Teniendo este hecho en cuenta, nos parece importante destacar que, en nuestra opinión, las unidades que conforman las bases de las colocaciones son formas léxicas y no lemas o lexemas.

Cuestión distinta es, sin embargo, el caso del colocado. En lo que respecta a este componente opinamos que sí se puede considerar en su forma lematizada puesto que el hecho de que éste adopte distintas realizaciones morfosintácticas no implica en modo alguno que se trate de colocaciones diferentes. En este sentido, opinamos que las bases ejercen su influencia sobre lemas, y no sobre formas léxicas

concretas. Ejemplificando lo anterior, no nos parece acertado tratar las expresiones "*I took a decision*", "*she takes a decision*" y "*the decision was taken*" como tres colocaciones diferentes, puesto que parece evidente que en todos los casos el hablante está haciendo uso de la misma asociación léxica (Aitchison, 1994). Asimismo, desde el punto de vista de la enseñanza y evaluación de lenguas extranjeras, tampoco parece conveniente tratarlas como combinaciones distintas, puesto que cabe suponer que un alumno que conoce la combinación *"take + decision"* será capaz de producirla en distintas formas siempre que conozca las normas gramaticales pertinentes.

Finalmente, un último factor que nos parece interesante mencionar en esta descripción de la naturaleza formal de la colocación es el que se refiere a la distancia que existe entre los elementos que la conforman, lo que en la bibliografía anglosajona se denomina "*span*". Con relación a este aspecto, parece existir un consenso general en considerar que las palabras que integran la colocación no tienen que aparecer necesariamente de forma contigua en el lenguaje, sino que puede existir un determinado espacio (normalmente expresado en número de formas léxicas) entre la base y el colocado. Quizá la excepción más notable en este sentido sea Kjellmer (1994), para quien una colocación sólo es tal cuando los elementos que la integran son adyacentes y aparecen repetidamente de forma totalmente fija en el corpus.

Como vimos más arriba, la colocación admite cierto grado de variación morfosintáctica (incluyendo en algunos casos modificaciones como por ejemplo la pasivización y la nominalización), lo cual implica que su estructura puede ser variable permitiendo así la existencia de varias palabras entre sus miembros. En una colocación como "*make a complaint*" vemos que aparece un artículo indeterminado entre los elementos de la colocación, pero esta distancia puede aumentar como en el caso de "*make a formal complaint*" o "*complaints continue to be frequently made*". Por tanto, podemos decir que en términos teóricos, consideramos que la distancia entre los

colocados no es fija, pudiendo ser de una longitud considerable y sobrepasando incluso los límites de la oración.

A la hora de operativizar esta distancia para poder identificar las colocaciones de forma automatizada, algunos autores (Evert y Krenn, 2003, citado por Gyllstad, 2007) han puesto de manifiesto las dificultades que entraña establecer el número concreto de unidades que se deben contemplar entre la base y el colocado, puesto que un espacio demasiado pequeño no permitirá reconocer colocaciones interesantes dando lugar a estimaciones de frecuencia erróneas, mientras que un espacio demasiado amplio incluirá un número enorme de combinaciones que no son de interés en nuestros estudios. En cualquier caso, dado que se trata de una especificación necesaria en los estudios basados en procedimientos estadísticos, varios autores han tratado de establecer el tamaño más adecuado de este espacio. Como ya vimos en la sección 1.3.1.1.2., mientras que algunos autores (Jones y Sinclair, 1974) han fijado este espacio en cuatro unidades a cada lado de la base, estudios posteriores han ampliado este número hasta cinco (Martin, Al y Van Sterkenburg, 1983; Smadja, 1993), siendo ambas distancias utilizadas de forma muy común en los estudios colocacionales en la actualidad. En esta tesis, y debido fundamentalmente a las restricciones impuestas por defecto por los programas informáticos utilizados para la identificación de colocaciones, el margen que se contempla es de cinco palabras a derecha e izquierda de nuestras bases (ver sección 3.2.2.4.1.).

1.3.2.2. Características funcionales

Además de los rasgos formales que acabamos de mencionar, toda colocación consta asimismo de unas características funcionales que le otorgan el estatus de unidad fraseológica, a la vez que nos permiten diferenciarla de otras categorías como las

combinaciones libres de palabras o las locuciones. Así, los tres factores fundamentales que nos sirven para definir este fenómeno son: 1) arbitrariedad en la conmutabilidad, 2) composicionalidad semántica, y 3) frecuencia e institucionalización.

1.3.2.2.1. Arbitrariedad en la combinación

La arbitrariedad que caracteriza a la colocación viene dada por el hecho de que algunas palabras se combinan porque sí, mientras que otras combinaciones que en teoría son igualmente posibles para expresar la misma idea no se utilizan nunca. Así pues, consideraremos colocaciones aquellas combinaciones en las que existe una restricción totalmente arbitraria en cuanto a la aparición conjunta de palabras que nada tiene que ver con las restricciones semánticas impuestas por el sistema de una lengua. En lo que sin duda sirve como un claro ejemplo de lo que supone la arbitrariedad colocacional, Benson (1989: 4) afirma: "one says *make an estimate* but not **make an estimation*; *commit treason* but not **commit treachery*; *a running commentary* but not **a running discussion*; *warmest greetings* but not **hot greetings*; *a strong protest* but not **a strong complaint*".

Cabe destacar en este sentido que, de acuerdo con Cowie (1988) y Howarth (1996), la restricción en la posible sustitución de palabras se puede dar en cualquiera de los dos elementos de la colocación o en los dos a la vez, siendo la fijación en este último caso más acusada. El principal problema que se plantea al considerar la arbitrariedad de esta forma es que no parece fácil conjugar la idea de que la base es el elemento independiente que el hablante selecciona libremente a la hora de hablar, con la noción de que esta base está restringida de alguna manera en su conmutabilidad. Como afirma Nesselhauf (2005: 32), "[i]f the noun in a collocation is selected on the basis of its meaning (...), this means at the same time that nouns are

not arbitrarily restricted in collocations". A nuestro modo de ver, no cabe duda de que la base es elegida de forma totalmente libre por el hablante, por lo que no está restringida en su uso. Sin embargo, existen colocaciones como "*central issue*", en las que no parece existir una restricción en la conmutabilidad del colocado (puesto que "*key issue*", "*vital issue*" o "*crucial issue*" son igualmente válidas). En nuestra opinión, lo que sucede es que el colocado no está restringido en su conmutabilidad, pero sí en su combinabilidad con el sustantivo. Así, "*issue*" es la base que elegimos libremente, pero al utilizar el colocado "*central*" estamos seleccionando un elemento que no se combina libremente con cualquier otro sustantivo semánticamente válido ("**central view*" o "**central act*").

Así pues, lo anterior nos lleva a distinguir los siguientes tipos de unidades, que en ocasiones se solapan entre sí:

a) Colocaciones en las que el colocado está restringido en su conmutabilidad, por lo que la base tiende a ir acompañada de uno o varios colocados por los que muestra especial preferencia, mientras que otros sinónimos o expresiones posibles que podrían acompañar a la base para expresar la misma noción no se suelen usar nunca o se hace con muy poca frecuencia. Éste es el caso de combinaciones como "*make a mistake*", donde la base, "*mistake*", no coloca con otros posibles verbos como "**do*" o "**take*".
b) Colocaciones en las que el colocado está restringido en su combinabilidad, por lo que éste no acompaña libremente a cualquier base con la que en teoría podría combinarse siguiendo las leyes semánticas. Así, por ejemplo, usamos la colocación "*commit a crime*", donde "*commit*" puede combinarse con otros sustantivos de significado negativo ("*suicide/a murder/atrocity*") pero donde también existen excepciones, por lo que no acompaña a cualquier base (**commit a lie/*commit mischief*).

c) Colocaciones en las que el colocado está restringido tanto en su conmutabilidad como en su combinabilidad. En este caso se trata de unidades en las que tanto la base como el colocado muestran una preferencia casi exclusiva por el otro elemento para expresar un significado concreto. En combinaciones como "*shrug your shoulders*" o "*curry favour*", vemos que el sustantivo exige la compañía del colocado de forma totalmente arbitraria, mientras que el colocado, por su parte, no acompaña prácticamente a ningún otro nombre con el mismo significado.

Como podemos observar a la luz de lo anterior, la arbitrariedad que presentan las colocaciones en su combinación da lugar a unidades con distinto grado de fijación o fosilización. Ello indica claramente que, tal y como argumentan los autores de la tradición fraseológica, estas combinaciones parecen distribuirse a lo largo de una escala, desde las más flexibles en su cohesión hasta las más fijas e idiosincrásicas. El hecho de que esta fosilización esté presente en todos los casos, independiente del grado en el que se dé, siendo un rasgo definitorio, nos permite establecer esta característica como criterio para poder discernir entre la colocación y la combinación libre de palabras, que no forma parte del componente fraseológico de las lenguas. Así pues, en este trabajo consideramos que un binomio léxico formará una colocación cuando se aprecie la existencia de cierta restricción arbitraria en su combinatoria, mientras que se tomará como una combinación libre de palabras cuando éstas formen parte de una unidad que responde únicamente a las leyes gramaticales y semánticas.

Desde un punto de vista pedagógico, cabe añadir a este respecto que, como podemos suponer, la idiosincrasia propia de las colocaciones, que hace que los hablantes las almacenen de forma asociada en su lexicón mental (Bolinger, 1976; Aitchison, 1994; Lewis, 2000a; Nesselhauf, 2005), es en buena medida la causante de

las enormes dificultades que causan al aprendiz de una lengua extranjera. Este hecho, además, es el que más claramente justifica la necesidad de desarrollar herramientas adecuadas tanto para su enseñanza como para su evaluación.

1.3.2.2.2. "Composicionalidad" semántica[13]

El segundo aspecto que caracteriza a la colocación es su composicionalidad semántica, es decir, el hecho de que cada una de las palabras que la integran contribuye de forma individual a formar el significado global de la colocación (Howarth, 1996). Este aspecto, que claramente lleva implícita la noción de que cada una de las palabras que conforman la colocación conserva su valor léxico individual a pesar de formar parte a su vez de una unidad fraseológica, nos permite establecer la frontera entre las colocaciones y las locuciones por un lado, y entre las colocaciones y las palabras compuestas por otro. De esta forma, mientras que las colocaciones se caracterizan, como decimos, por tener un significado derivable a partir de los diferentes elementos que la integran, las locuciones forman una unidad semántica no composicional, es decir, en la que el significado de la expresión resultante no corresponde a la suma de los significados de las palabras individuales que la integran (Cowie, 1988). A título ilustrativo, podemos observar que mientras que en la combinación "*to cast an eye*" cada elemento conserva su significado individual formando así una colocación, la expresión "*to turn a blind eye*" pierde su significado literal y adopta un valor semántico global, distinto del que aporta cada palabra por separado. Como decíamos, esta diferencia nos servirá en el presente estudio para establecer la línea divisoria entre las colocaciones y las locuciones idiomáticas.

---

[13] Aunque el término "composicionalidad" no está recogido en el Diccionario de la Real Academia Española de la Lengua y se trata claramente de una palabra tomada de la voz inglesa "*compositionality*", sí parece estar en uso en la bibliografía española (Koike, 2001).

La composicionalidad semántica propia de las colocaciones también nos sirve para establecer la frontera entre éstas y las palabras compuestas. Así, cuando una combinación léxica esté formada por palabras que conservan su carga semántica individual la consideraremos una colocación, mientras que en aquellos casos en que las distintas unidades tengan un significado literal pero designen un único referente como si de una sola palabra se tratara hablaremos de palabra compuesta. Así pues, combinaciones como "*ironing board*", "*shopping centre*" o "*video game*" se clasificarán como nombre compuestos en este trabajo. Debemos destacar en este sentido que, a pesar de que en principio esta distinción parece clara, resulta sin embargo verdaderamente complicado ponerla en práctica en determinadas ocasiones, como se comprobará cuando abordemos el proceso de obtención de colocaciones llevado a cabo en nuestra investigación (ver sección 3.2.2.5.).

Antes de pasar al siguiente apartado nos gustaría destacar que en nuestro estudio no se considera una característica necesaria el hecho de que el colocado adopte un significado figurado o secundario al combinarse con la base de la colocación. Coincidimos por tanto plenamente con los autores citados anteriormente (Cowie, 1988; Nesselhauf, 2005), para quienes éste es un rasgo frecuente, pero no imprescindible, de las colocaciones. En este sentido, somos de la opinión de que, como suele suceder con todos los parámetros que existen en fraseología, el factor transparencia-opacidad semántica se manifiesta con desigual intensidad en la lengua, siendo pues un fenómeno graduable. Sin embargo, el hecho de que existan colocaciones como "*sleepless night*" o "*safe place*", que no presentan alteración ni opacidad semántica alguna de sus elementos, nos indica que en este caso se trata de un parámetro secundario que no siempre aparece en las unidades colocacionales. Estas combinaciones, no obstante, se pueden identificar como colocaciones porque muestran una conmutabilidad restringida ("**wakeful night*", "**security/*sure place*"). Éste es, pues, el rasgo verdaderamente definitorio de las colocaciones.

1.3.2.2.3. Frecuencia e institucionalización

Por último, otro de los factores que consolidan a una colocación como tal y la diferencian de una combinación cualquiera de palabras es la frecuencia con que la usan los hablantes de una lengua, es decir, el hecho de que la colocación pertenece a la lengua entendida como "norma" y no sólo como "sistema". Así, mientras que la combinación "*to finish a war*" es aceptable como parte del sistema puesto que cumple todas las reglas gramaticales y semánticas de la lengua inglesa, un hablante nativo emplearía la expresión "*to end a war*", debido a que ésta es la combinación institucionalizada en la norma por la frecuencia de uso. En este sentido, Pawley y Syder (1983: 209) afirman:

> What makes an expression a lexical item, what makes it part of the speech community's common dictionary, is (…) that it is a *social institution*. This (...) characteristic is sometimes overlooked, but is basic to the distinction between lexicalized and non-lexicalized sequence. (...) Rather that being a 'nonce form', a spontaneous creation of the individual speaker, the usage bears the authority of regular and accepted use by members of the speech community.

Y si la frecuencia de uso es una de las principales características que definen a la colocación, no resulta extraño que las técnicas de análisis aportadas por la lingüística de corpus hayan supuesto una revolución en el estudio de este fenómeno. Los cálculos estadísticos realizados sobre grandes compilaciones de textos nos han permitido medir de manera objetiva la frecuencia de co-aparición de las palabras en el discurso, y con ello establecer en gran medida qué combinaciones se pueden considerar co-apariciones estadísticamente significativas y cuáles no. Sin embargo y como ya apuntamos anteriormente, uno de los aspectos que el enfoque estadístico ha obviado tradicionalmente es el hecho de que, si bien la frecuencia es un factor fundamental, éste no es suficiente por sí solo para delimitar el concepto de colocación por completo. Si así fuera, la expresión "*nice house*" o cualquier otra

combinación de palabras que se usara de manera recurrente en una lengua tendría que clasificarse como colocación. En nuestra opinión, resulta evidente, pues, que los tres criterios aquí mencionados, la arbitrariedad combinatoria, la composicionalidad semántica y la frecuencia de uso, son igualmente imprescindibles a la hora de definir e identificar una colocación.

1.4. Clasificación de las colocaciones

Existen diversas clasificaciones de la colocación dependiendo de los parámetros que se tomen para organizarlas (Nesselhauf, 2005, apunta principalmente a las taxonomías basadas en rasgos sintácticos, semánticos o combinatorios). En nuestro caso, partiremos de la clasificación establecida por Benson, Benson e Ilson (1986), de carácter fundamentalmente sintáctico, dado que nos parece la más completa, clara y estable de cuantas se han postulado. Ésta es, además, la clasificación más generalizada en los estudios de colocaciones, quizá por ser la más amplia que existe sobre el fenómeno, dando cabida por tanto a todas las colocaciones y al resto de posibles categorizaciones.

Según estos autores, las colocaciones se pueden dividir en dos grandes grupos: gramaticales y léxicas. Nuestro trabajo se centra, como ya adelantamos, en las colocaciones de tipo léxico. Así pues, tras realizar una revisión general de las diferentes categorías que Benson, Benson e Ilson establecen, plantearemos la posibilidad de establecer una subdivisión dentro de la categoría de las colocaciones léxicas que resulte operativa para los propósitos de este trabajo y en la que se entable una relación congruente entre las diferentes estructuras sintácticas que pueden adoptar las colocaciones y su propia estructura interna formada por base y colocado.

1.4.1. Colocaciones gramaticales

Este tipo de combinaciones ha sido muy poco tratado en los estudios de este ámbito, centrándose la gran mayoría de ellos en las colocaciones léxicas. Las colocaciones gramaticales son combinaciones en las que la base es una palabra léxica y el colocado lo constituye por lo general, aunque no necesariamente, una preposición. Para los creadores de esta clasificación (Benson, Benson e Ilson, ibid.), esta clase de colocaciones incluye aspectos relativos a la complementación sintáctica del tipo "*avoid + -ing*", mientras que para otros (Granger y Paquot, 2008), entre los cuales nos incluimos, este tipo de complementación no da lugar a una colocación, sino que se trata de coligaciones (ver sección 1.5.). Según estos autores, las colocaciones gramaticales se pueden subdividir en 8 tipos diferentes (denominados G1, G2, G3, etc.) que pasamos a describir y ejemplificar brevemente a continuación[14]:

G1: Colocaciones de **nombre + preposición** (excepto *of* y *by*, que se consideran constituyentes de combinaciones libres): *blockade against*, *apathy towards*, etc.

G2: Colocaciones formadas por **nombre + infinitivo con *to*** (excepto infinitivos de finalidad que forman combinaciones libres): *a pleasure to*, *a compulsion to*, etc.

G3: Colocaciones que consisten en **nombre + oración con *that*** (excepto cuando *that* se utiliza como pronombre relativo): *an agreement that she would…*, *he took an oath that he would do his duty*, etc.

G4: Colocaciones de **preposición + nombre**: *in agony*, *by accident*, etc.

[14] Véase Benson, Benson e Ilson (1986: xv-xxix) para un análisis más detallado.

G5: Colocaciones formadas por **adjetivo + preposición** utilizados con verbos copulativos o que funcionan como oraciones sin verbo: *to be angry at everyone*, *fond of children*, etc.

G6: Colocaciones formadas por **adjetivo predicativo + infinitivo con *to***: *it was necessary to work*, *he is likely to be late*, etc.

G7: Colocaciones de **adjetivo + oración con *that***: *to be afraid that*…, *it was nice that*…, etc.

G8: Colocaciones que consisten en diferentes **estructuras verbales** dependiendo de su transitividad, su pasivización, si van seguidas de infinitivo o gerundio, etc.: *to ask someone to come*, *to be authorized to use the room*, *to decide to go, to enjoy watching TV*, etc.

1.4.2. Colocaciones léxicas

A diferencia de las anteriores, las colocaciones léxicas han sido ampliamente tratadas en los estudios de este ámbito. El rasgo distintivo de este tipo de colocaciones es que tanto la base como el colocado son palabras léxicas, es decir, verbos, sustantivos, adjetivos o adverbios.

1.4.2.1. Taxonomía de Benson, Benson e Ilson (1986)

Benson et al. (1986) distinguen 7 tipos de estructuras diferentes:

L1: Colocaciones cuya estructura es **verbo** (generalmente transitivo, aunque no necesariamente) **+ nombre/pronombre**. En su mayor parte el verbo denota Creación y/o Activación, de ahí que se las denomine "*CA collocations*": *issue a warning*; *launch a missile*, etc.

L2: Colocaciones con una estructura de **verbo + nombre** en las que el verbo denota fundamentalmente Erradicación y/o Nulificación por lo que se las conoce como "*EN collocations*": *ease tension*; *annul a marriage*, etc.

L3: Colocaciones constituidas bien por **adjetivo + nombre**, bien por **nombre (atributivo) + nombre** (excepto en aquellos casos en que el segundo elemento no mantiene su significado básico): *strong tea*; *chronic alcoholic*; *land reform*; *aptitude test*, etc.

L4: Colocaciones con una estructura de **nombre + verbo**, que indica una acción propia de la persona o cosa designada por el nombre: *blood circulates*; *bees buzz*, etc.

L5: Colocaciones con estructura **nombre1 *of* nombre2**, que indican la unidad de medida asociada con el nombre2. Ésta puede ser la unidad superior a la que pertenece cada miembro o la unidad inferior, concreta o específica de algo superior o más general: *a pride of lions*; *an act of violence*, etc.

L6: Colocaciones de **adverbio + adjetivo**: *strictly accurate*; *sound asleep*, etc.

L7: Colocaciones con una estructura de **verbo + adverbio**: *appreciate sincerely*; *affect deeply*, etc.

Podemos observar que, a pesar de ser una clasificación sintáctica, Benson et al. introdujeron factores semánticos para establecer distinciones entre algunos de los tipos. Concretamente el factor que diferencia las colocaciones L1 y L2 es el significado del verbo. Este aspecto, que evidentemente resta consistencia a la clasificación propuesta, ha sido evitado por otros autores que han establecido taxonomías atendiendo también a la estructura sintáctica de la colocación. De este modo, Hausmann (1989) y Corpas Pastor (1996) defienden tipologías muy similares a

la de Benson et al., pero sin contemplar la diferencia hecha por éstos entre los dos primeros tipos[15].

Otro problema que se plantea, en nuestra opinión, en esta clasificación aparece en la categoría L5. A nuestro juicio, el tipo de unidades que se recoge en esta sección no conforma, al menos en la mayor parte de los casos, colocaciones tal y como se conciben en el presente trabajo. Combinaciones como "*panes of glass*", "*flocks of sheep*" o "*bar of chocolate*" representan unidades donde, si bien es cierto que el primer nombre tiene un significado muy específico que está claramente ligado al segundo, no nos parece sin embargo que exista arbitrariedad a la hora de seleccionarlo. Según nuestro parecer, ambos nombres se usan en su significado literal, siendo los únicos elementos que se pueden utilizar para expresar ese referente concreto. Se trata, pues, de un fenómeno similar al que ya observamos con relación a la combinación "*barking dog*". Expresiones como "*the dog barks*" o "*the horse neighs*" no establecen ninguna restricción en la conmutabilidad, puesto que no hay otras posibles opciones que puedan "competir" con la seleccionada. Este fenómeno, que se conoce como "solidaridad léxica" y que trataremos en la última sección de este capítulo, se suele incluir como un tipo de colocación, aunque desde nuestro punto de vista se trata de un fenómeno de naturaleza diferente. Si en un test de colocaciones, por ejemplo, se incluyera este tipo de unidades, no se evaluaría el conocimiento colocacional del candidato, es decir, si conoce las restricciones combinatorias de la lengua. En nuestra opinión, lo único que se estaría valorando es su riqueza léxica, algo que no se corresponde con el constructo colocacional.

15 Ambos autores afirman que una de las principales ventajas de esta clasificación es que es válida tanto para las lenguas germánicas como para las romances y las eslavas.

1.4.2.2. Propuesta de taxonomía de colocaciones léxicas

Con el objetivo de crear una clasificación que establezca, en primer lugar, una relación coherente entre la naturaleza formal de la colocación (que consta de base y colocado) y las diferentes estructuras sintácticas en que se manifiesta, y para eliminar, en segundo lugar, los inconvenientes que, según hemos apuntado más arriba, presenta la clasificación de Benson et al., plantearemos a continuación una taxonomía de las colocaciones léxicas, que nos resulte operativa, asimismo, para la elaboración de nuestro test:

1) Colocaciones donde la **base es un sustantivo** y se combina con colocados nominales, adjetivales y verbales dando lugar a las estructuras nombre+nombre (N+N, donde el primer elemento es el colocado y el segundo elemento la base), adjetivo+nombre (A+N), verbo+nombre (V+N) y nombre+verbo (N+V). Este tipo de colocaciones comprende las categorías L1, L2, L3 y L4 de Benson et al.

2) Colocaciones formadas por una **base adjetival** acompañada de un adverbio formando combinaciones adverbio+adjetivo (Adv+A). Este tipo de colocaciones comprende la categoría L6 de Benson et al.

3) Colocaciones que contienen un **verbo como base** acompañado de un adverbio en estructuras verbo+adverbio (V+Adv). Este tipo de colocaciones comprende la categoría L7 de Benson et al.

1.5. Conceptos relacionados con la colocación

Muy brevemente, no nos gustaría concluir este primer capítulo sin hacer referencia a algunos de los aspectos que con mayor frecuencia se relacionan (y a menudo se confunden) con las colocaciones. Dada la falta de consenso que todavía hoy existe en torno a las cuestiones fraseológicas, y por supuesto a su terminología, nos parece pertinente llevar a cabo esta brevísima revisión que esperamos clarifique en mayor medida el concepto de colocación que planteamos en este trabajo. Así pues, trataremos de definir los términos "coligación", "solidaridad léxica" y "construcción de verbo soporte", todos ellos relacionados de una u otra forma con el fenómeno que conforma el objeto de estudio en la presente tesis.

a) **Coligación:** Los conceptos de colocación y coligación han estado siempre muy relacionados. En realidad, ambas nociones hacen referencia al mismo fenómeno pero desde perspectivas diferentes. Mientras que la colocación es la relación que se establece entre dos entidades léxicas concretas, la coligación es la generalización sintáctica de dicha combinación (Howarth, 1996). Así, desde un punto de vista léxico la colocación es la co-aparición frecuente de palabras, en tanto que desde el punto de vista gramatical la coligación es la estructura formada por la combinación de elementos sintácticos. Howarth (ibid.: 29) lo explica de forma muy clara en las siguientes palabras: "The term 'colligation' refers to a particular structural pattern that is realized by one or more collocations: a collocation is always an exponent of a colligation".

b) **Solidaridad léxica:** Generalmente se considera a Eugenio Coseriu como el fundador de la noción de solidaridad léxica en el año 1967 (Corpas Pastor, 1996; Nesselhauf, 2005). Este término designa un tipo de relación

"orientada" ya que se compone de una palabra determinante (cuyos rasgos distintivos forman parte de la otra palabra que forma la solidaridad) y otra palabra determinada (que posee rasgos distintivos de la palabra determinante). Esto supone que en una solidaridad léxica existe una relación en la que una palabra implica semánticamente a la otra pero no a la inversa (Corpas Pastor, 1996). Así, por ejemplo, en la combinación "*the lion roars*", la palabra "*lion*" determina semánticamente a "*roars*" por lo que "*roars*" posee unos rasgos distintivos que hacen que siempre implique "*lion*". Sin embargo, esta relación no se da a la inversa puesto que "*lion*" no implica necesariamente "*roars*".

Como vemos, este concepto se encuentra ciertamente relacionado con el de colocación pero en ningún caso se deben confundir. En nuestra opinión, si bien es cierto que constituye un elemento importante desde el punto de vista pedagógico (especialmente en la secuenciación de contenidos), no lo es tanto como la colocación puesto que la solidaridad no cuenta con la dificultad añadida de la arbitrariedad de uso que caracteriza a la colocación y que la hace merecedora de una atención especial por parte del estudiante.

c) **Construcción de verbo soporte *(Stretched verb constructions)***: Este tipo de combinaciones léxicas está más restringido que los anteriores dado que sólo se da en construcciones de verbo+nombre. Se trata de unidades en las que el verbo no aporta prácticamente información semántica a la combinación, por lo que se le suele llamar "verbo ligero" o "verbo vacío" ("*light verb*" o "*empty verb*") (Radford, 1997: 201), y en las que el nombre pertenece a la misma familia léxica que el verbo que podría sustituir a la combinación en su conjunto (Nesselhauf, 2005). Por ejemplo, la expresión "*to give an answer*" constituye una construcción de verbo soporte puesto que

"*give*" no posee un valor semántico pleno, además de que la combinación se podría sustituir simplemente por "*to answer*". A la vista de esta definición cabe preguntarse cómo es posible que la lengua, con su marcada tendencia a la economía y la eliminación de elementos innecesarios, posea este tipo de construcciones. Tal y como Allerton (2002) destaca, esto se debe a que ambas alternativas ofrecen matices diferentes en su significado puesto que el papel semántico de la acción lo realiza un nombre en el caso de la construcción de verbo soporte mientras que en la construcción estándar lo lleva a cabo un verbo.

Debemos destacar que este tipo de construcciones, que se restringen a combinaciones formadas por un número muy reducido de verbos —en concreto "*make*", "*take*", "*give*" y "*have*" (Labuhn, 2001, citado por Nesselhauf, 2005)— no sólo forman parte del espectro colocacional, sino que conforman uno de los tipos de colocaciones más frecuentes de la lengua, por lo que serán incluidas en nuestro estudio.

1.6. Conclusión

Como esperamos haber hecho patente a lo largo de este capítulo, no existe todavía un concepto de colocación claramente definido y aceptado unánimemente por la comunidad académica, y por tanto resulta primordial tratar de recoger las principales aportaciones que se han hecho en este ámbito para establecer, a partir de éstas, nuestra propia visión del fenómeno. Esta aproximación establece así los fundamentos teóricos necesarios como paso previo para poder diseñar un test de colocaciones con una adecuada validez de constructo y de contenido, fin último de esta investigación. En este sentido, todos los parámetros formales y funcionales que

integran la colocación tal y como se entiende en este estudio fueron tenidos en cuenta a la hora de seleccionar las colocaciones y diseñar los ítems que constituirían nuestra prueba de evaluación colocacional.

Pero antes de pasar a mostrar los procedimientos y resultados de nuestra investigación, nos detendremos en la exploración del concepto de frecuencia como factor fundamental en la enseñanza del vocabulario, y muy particularmente en el estudio de las colocaciones, aspecto éste muy relevante puesto que constituye uno de los pilares fundamentales de esta tesis. En definitiva, pretendemos recopilar y evaluar las colocaciones más frecuentes del inglés general, partiendo del convencimiento de que aquello que aparece con mayor frecuencia en la lengua será lo que con mayor probabilidad necesiten utilizar productiva y receptivamente nuestros alumnos.

CAPÍTULO 2

LA FRECUENCIA EN LA ENSEÑANZA DEL VOCABULARIO: LA LINGÜÍSTICA DE CORPUS Y LOS LISTADOS DE PALABRAS Y COLOCACIONES

> *Whatever the imperfections of the simple equation 'most frequent'= 'most important to learn', it is difficult to deny that frequency information becoming available from corpora has an important empirical input to language learning materials.*
>
> Leech (1997: 16)

2.1. Introducción

Nadie parece cuestionar en la actualidad que aquello que es frecuente en la lengua es interesante y merece nuestra atención. Hoy en día somos testigos de que la frecuencia se ha erigido en uno de los principios fundamentales de muchos de los estudios lingüísticos actuales gracias en gran medida, qué duda cabe, a las enormes posibilidades que la informática ha venido a ofrecer en este sentido. También desde el punto de vista del aprendizaje de lenguas extranjeras, y más concretamente en lo

que se refiere al vocabulario, la frecuencia es hoy en día un factor de capital importancia. Prácticamente la totalidad de los diccionarios y los libros de texto que se han publicado en los últimos años, no sólo en inglés sino también en otras lenguas, están diseñados con relación a la frecuencia de sus contenidos, un aspecto que constituye de hecho el principal reclamo comercial en la mayoría de los casos, dado el rigor y la fiabilidad que se le suponen a este tipo de trabajos.

Uno de los indicadores más evidentes del valor que tiene la frecuencia en los estudios de vocabulario es que los modelos de competencia léxica postulados en las últimas décadas por los investigadores más reconocidos incluyen este factor en su noción de lo que implica conocer una palabra. En este sentido, la primera definición de competencia léxica que conocemos, introducida de manera totalmente innovadora por Richards (1976) en su artículo "*The Role of Vocabulary Teaching*", este autor enumera siete aspectos importantes a la hora de aprender una palabra:

1) El hablante nativo continúa aumentando su vocabulario durante la madurez, mientras que la gramática experimenta muy poco desarrollo.
2) Conocer una palabra significa conocer la frecuencia con que se utiliza.
3) Conocer una palabra significa conocer sus condiciones de uso según las distintas funciones y situaciones.
4) Conocer una palabra significa conocer su comportamiento sintáctico.
5) Conocer una palabra significa conocer su morfología y sus posibles derivaciones.
6) Conocer una palabra significa conocer las redes asociativas que se establecen con otras palabras.
7) Conocer una palabra significa conocer su valor semántico.

Así pues, podemos observar que uno de los siete principios fundamentales de lo que implica conocer una palabra se refiere expresamente a su frecuencia. Más detenidamente, en el principio número 2, este autor (ibid.: 79) escribe:

> ASSUMPTION 2: *Knowing a word means knowing the degree of probability of encountering that word in speech or print. For many words we also "know" the sort of words most likely to be found associated with the word.*
> The speaker of a language recognizes that some words are common and familiar while other words are rare, unfamiliar or even totally unknown to him. Our knowledge of the general probability of occurrence of a word means that we recognize that a word like *book* is more frequent than *manual* or *directory* while both these words strike us as more frequent than *thesaurus*. Given a list of words, with the exception of concrete nouns a native speaker can classify them into "frequent," "moderately frequent," "not frequent," to a degree of accuracy reasonably close to their actual frequencies (Noble 1953; Richards 1974).

Como podemos comprobar, este principio expresa el convencimiento de que conocer y ser conscientes de la frecuencia de una palabra es un aspecto fundamental para poder afirmar que se conoce y para ser capaz de interpretarla correctamente. No sólo se trata, pues, de que el profesor conozca y establezca las palabras por orden de frecuencia para una selección racional de contenidos, sino que la propia consciencia que el alumno tenga de la frecuencia de una palabra es un aspecto integrante de su competencia y su habilidad lingüística.

Continuando pues con esta tradición en los estudios de competencia léxica, Carter (1987) y Nation (2001) han formulado asimismo sus propias definiciones de lo que significa conocer una palabra, siendo muy significativo el hecho de que ambos autores han conservado la noción de frecuencia como un componente fundamental de nuestro conocimiento del vocabulario. Carter (1987) estima que existen también siete principios básicos implicados en conocer una palabra, todos los cuales coinciden con los postulados por Richards excepto el primero de ellos, que Carter sustituye por la habilidad para usar la palabra productiva y receptivamente en el discurso, y el quinto, que Carter convierte en el conocimiento de las expresiones fijas

en las que interviene la palabra y su capacidad para usarlas como un todo. Por su parte, Nation (2001) establece que conocer una palabra implica dominar su forma (a nivel oral, escrito y morfológico), su significado (en relación a la forma, en cuanto al concepto y los referentes lingüísticos y en lo que se refiere a las asociaciones semánticas con otras palabras) y ser capaz de usarla adecuadamente (en estructuras gramaticales y colocacionales adecuadas y conociendo su valor pragmático y su frecuencia). Podemos comprobar, como decíamos, que los tres grandes modelos que definen lo que significa conocer una palabra están de acuerdo en considerar la frecuencia como un aspecto fundamental.

También los estudios realizados desde el campo de la psicolingüística han venido a confirmar que la frecuencia juega un papel fundamental en la adquisición y el uso del vocabulario. Aquello que con más frecuencia nos encontramos en nuestro uso diario de la lengua es lo que se encuentra más "disponible" en nuestro lexicón mental para poder volver a utilizarlo, lo cual implica que su frecuencia seguirá reforzándose (López Morales, 1993). Especialmente en lo que se refiere a las unidades fraseológicas, la investigación en este campo ha demostrado que buena parte de la comunicación se nutre de expresiones fijas memorizadas como un todo prefabricado que se repiten una y otra vez. El uso de estas secuencias repetidas tiene como una de sus principales finalidades y ventajas facilitar el procesamiento y el esfuerzo cognitivo necesario para comunicarnos, algo que a su vez redunda en una mayor fluidez y precisión en nuestro uso de la lengua (Wray, 2002).

A la luz de estos argumentos, por tanto, parece evidente que la frecuencia es un valor esencial en nuestro conocimiento y nuestra capacidad para utilizar el vocabulario de forma adecuada y eficaz. Si conocer la frecuencia de una palabra es un aspecto integrante de la propia competencia léxica, si aquello que es más frecuente es lo que con más probabilidad y más a menudo necesitaremos procesar y producir y es además un factor que facilita el procesamiento y la comunicación efectiva y fluida, no

cabe duda de que nos encontramos ante un aspecto que merece una atención especial en los estudios léxicos y en particular en su enseñanza.

En el capítulo que ahora comienza veremos que el interés por los elementos más frecuentes del lenguaje no es nuevo. Desde principios del siglo XX y debido sobre todo a la necesidad de seleccionar el vocabulario más relevante para los alumnos, la noción de que la frecuencia debe ser un factor determinante del diseño pedagógico comenzó a ser aceptada de forma general en el ámbito de la enseñanza de lenguas. Teniendo en cuenta, pues, que ambas nociones, la importancia del léxico y de la frecuencia, nacieron juntas, no debe extrañarnos que su progresión haya corrido paralela a lo largo de los años. En este capítulo, pretendemos llevar a cabo un recorrido por las etapas más importantes de la evolución de la enseñanza del vocabulario, y poder así comprobar hasta qué punto estos dos aspectos han ido siempre de la mano en el campo de la enseñanza de lenguas. Así pues, comenzaremos describiendo los primeros estudios que abordaron la frecuencia léxica en el apartado 2.2., para pasar a continuación a tratar del auge de la lingüística de corpus en la sección 2.3., y de las enormes repercusiones que ello tuvo en los estudios sobre la frecuencia del vocabulario y de la fraseología. De forma particular, centraremos nuestra atención en los listados de frecuencias de palabras basados en técnicas computacionales, para pasar finalmente a explorar las repercusiones que los estudios sobre la frecuencia han tenido en el área de las colocaciones.

Cabe añadir que este capítulo tiene como principal finalidad destacar la importancia y la necesidad de conocer cuáles son las colocaciones más frecuentes de la lengua, puesto que éstas deben ser las que conformen el núcleo de la actuación pedagógica. Antes de emprender tal actuación, y ante la ausencia casi total de estudios que traten este aspecto, se hace necesario conocer en qué situación de partida se encuentran nuestros alumnos, qué problemas se le plantean con respecto a

las colocaciones y si su competencia colocacional comprende al menos las unidades más frecuentes como sería deseable. La evaluación de diagnóstico llevada a cabo en el presente trabajo de investigación responde a tal fin.

2.2. Los primeros estudios de frecuencia léxica en la enseñanza de lenguas

Durante el siglo XIX, cuando el método que primaba era el de Gramática-Traducción, el léxico no pasaba de ser un mero instrumento que permitía al alumno leer las obras literarias de la segunda lengua y comprender los ejemplos de las reglas gramaticales propuestas. Se trataba, evidentemente, de un vocabulario artificial y obsoleto que no se correspondía con las necesidades comunicativas reales (Zimmerman, 1997). A la vista de esta situación, algunos autores comenzaron a alertar sobre los peligros que conllevaba el olvido del lenguaje cotidiano, a la vez que abogaban por un vocabulario más concreto y ajustado a las necesidades de los hablantes en situaciones reales (Gouin, 1892, citado por Rivers, 1983).

Fue en este contexto donde, de una forma totalmente pionera y avanzada para su época, Thomas Prendergast (1806-1886) produjo el primer listado de palabras frecuentes del que tenemos constancia. Esta lista fue publicada en 1864 en un libro titulado *The Mastery of Languages; or, the Art of Speaking Foreign Tongues Idiomatically*[1]. Se trata de un listado que contiene un total de 214 términos, en su mayoría palabras gramaticales entre las cuales abundan las preposiciones,

[1] La obra completa se encuentra disponible en la siguiente dirección de Internet: http://books.google.es/books?id=a5gFAAAAQAAJ&dq=%22thomas+prendergast%22+speaking+tongues&printsec=frontcover&source=bl&ots=EORtUsLs9V&sig=FSgNUIn9IjYp1Kf2-kweInBD1E4&hl=es&ei=G1ufSbWMI9TIjAfEn63LCw&sa=X&oi=book_result&resnum=10&ct=result#PPA48,M1 [Último acceso: 21.02.2009]

pronombres personales, adjetivos posesivos, verbos auxiliares y modales, etc. y la única fuente de información que Prendergast utilizó para elaborarlo fue su propia intuición. Su juicio, sin embargo, resultó bastante fiable como confirmaron estudios posteriores, ya que al compararlo con la lista producida por Thorndike y Lorge (1944), se pudo comprobar que el 82% de las palabras del listado de Prendergast se encuentran entre las primeras 500 palabras de la lista de Thorndike y Lorge (Howatt, 1984). A pesar de ello, y como cabe esperar teniendo en cuenta las pautas dominantes en la enseñanza de lenguas extranjeras de la época, esta primera llamada de atención sobre la importancia de la frecuencia como factor a considerar en la selección de vocabulario no tuvo ninguna repercusión en términos prácticos (Sweet, 1899/1964). En cualquier caso, no cabe duda de que se trata de un listado de gran valor, principalmente por su carácter pionero. Como apunta Zimmerman (1997: 7), "Prendergast's list was an important innovation because it came at a time when simplicity and everyday language were scorned and before it was normal to think in terms of common words".

No fue, sin embargo, hasta tres décadas después, en 1898, cuando Friedrich W. Kaeding publicó la siguiente lista de palabras frecuentes que conocemos. Este listado recogía las 2.402 palabras más frecuentes del alemán, compiladas manualmente a partir de un corpus de lenguaje escrito de casi 11 millones de palabras extraídas de diversas fuentes (periodísticas, históricas, militares, comerciales, políticas, etc.) y para cuyo procesamiento Kaeding precisó de la ayuda de más de 5.000 personas (Bongers, 1947). Como el propio autor advertía, la elección de este material respondía a las necesidades de los alumnos de taquigrafía a los que estaba destinada esta compilación (Fotos, 1931). Aunque, como podemos observar, esta lista no fue diseñada pensando en la enseñanza del alemán como lengua extranjera, sí que sirvió como referencia para futuros listados orientados a tal fin, tales como el llevado a cabo por Morgan (1930).

Pero fue sin duda el trabajo de los lingüistas británicos Harold E. Palmer (1877-1949), Albert S. Hornby (1898-1978) y Michael P. West (1888-1973) y de los psicólogos americanos Edward L. Thorndike (1874-1949) e Irving Lorge (1906-1961), junto con los esfuerzos de los *American and Canadian Committees on Modern Languages*, quienes en los años 20 y 30 dieron un impulso sustancial al papel del vocabulario y al concepto de frecuencia como aspectos fundamentales de la enseñanza de lenguas. Ya desde su primera publicación, Palmer (1917) asentó las bases del estudio de frecuencia léxica, destacando, de una forma totalmente innovadora, la importancia que se le debía conceder a la frecuencia como factor decisivo en la selección del vocabulario en los programas de enseñanza de lenguas. Este novedoso interés por las cuestiones léxicas, y especialmente por seleccionar el vocabulario de forma sistemática basándose en criterios de frecuencia, tuvo un gran calado tanto entre investigadores y docentes como a nivel institucional, lo cual indica que se trató de una tendencia verdaderamente trascendente.

Johnson (1927: 291) expresa claramente cómo la dificultad léxica se comenzó a explicar en términos de frecuencia: "It was necessary first to determine what could justly be termed 'difficult' vocabulary. The assumption was made that the students will know best those words which occur most frequently". Como vemos, por primera vez el vocabulario pasó a considerarse como un elemento importante que exigía por tanto una rigurosa selección y gradación de acuerdo con su dificultad, lo cual conllevaba la elaboración de listados de frecuencias de los cuales poder extraer el vocabulario necesario e idóneo para los alumnos según los distintos niveles de competencia. En resumen, el principio fundamental que todos compartían era expresado por Freeman (1931: 373) con estas palabras:

> Since it is almost axiomatic that students of a foreign language should be taught first those words which occur most often and which they must, therefore, be expected to encounter and know, frequency counts of vocabulary would seem to be a first condition for curriculum improvement.

Asimismo, es también relevante destacar el hecho de que esta corriente desarrollada en torno a la importancia del vocabulario y la necesidad de filtrarlo para facilitar el aprendizaje de lenguas, que dio lugar a lo que más tarde se conocería como el "vocabulary control movement" (Carter, 1987: 162), se enmarcó dentro del Método de Lectura (*Reading Method*) que predominaba en la época. Este método, nacido en parte como una reacción frente a las enormes deficiencias que se habían evidenciado entre los escolares norteamericanos en el informe Coleman (1929, citado por Zimmerman, 1997), tenía como principal finalidad el fomento, ante todo, de la destreza lectora como forma de mejorar el aprendizaje de una lengua extranjera. En este sentido, la necesidad de ordenar el vocabulario de las lenguas de acuerdo con su frecuencia se hacía más patente si cabe, dado que ya existía la noción de que conocer las palabras más frecuentes a nivel receptivo permitía alcanzar una buena compresión lectora: "A strong argument in favor of the word-frequency list is the fact that the first two thousand words of any frequency list covers 80 per cent of word occurrences in ordinary material taken at random" (West, 1937: 434). Así pues, una gran mayoría de los esfuerzos por mejorar la competencia lectora se centraron en controlar el léxico no sólo de los materiales pedagógicos (libros de texto, diccionarios monolingües, lecturas graduadas o simplificadas cuyo vocabulario se limitaba según los distintos niveles de dificultad) sino también de los tests de evaluación (Fotos, 1931). Sin duda, las listas de frecuencia léxica constituían una ayuda inestimable para llevar a cabo tal graduación. En este sentido West (1937: 434) afirmaría: "As a guide in selection of a *reading* vocabulary the word-frequency list is unassailable; for obviously that word which is most frequently met in reading material is the most valuable word for one to learn to read".

Con estas premisas en mente, comenzaron a producirse los primeros listados de palabras frecuentes del inglés. Podemos decir que el primer listado de frecuencias en lengua inglesa reconocido en el ámbito académico y profesional fue el elaborado

por Edward Thorndike, publicado con el título *The Teacher's Word Book* en el año 1921 (Fotos, 1931). En él se recogían las 10.000 palabras más frecuentes extraídas de un corpus de 4 millones y medio de palabras compuesto por textos de distinta naturaleza pero donde abundaban los de carácter literario y sobre todo bíblico. Se trataba, pues, de un listado de vocabulario dirigido principalmente al fomento de la destreza lectora, en el que no sólo se contempló la frecuencia de aparición total de las palabras, sino su rango, es decir, el número de textos de diferente naturaleza en el que aparecían (ibid.). Posteriormente, este mismo autor amplió este listado en *The Teacher's Word Book of 20,000 Words* (1931, citado por Howatt, 1984), para finalmente culminar su investigación en 1944 junto a Lorge con la publicación de *The Teacher's Word Book of 30,000 Words*. Esta última versión, que contenía 30.000 palabras (o 13.000 familias de palabras según Goulden, Nation y Read, 1990), se extrajo a partir de un corpus de 18 millones de palabras del inglés escrito y, durante años, supuso el punto de partida para la elaboración de otros listados posteriores así como de materiales pedagógicos (Howatt, 1984). Tras la muerte de Thorndike en 1949, Lorge realizó en ese mismo año una revisión de los trabajos sobre frecuencia semántica que de forma totalmente pionera habían llevado a cabo conjuntamente, publicando *The Semantic Count of the 570 Commonest English Words*, en la que se ofrecía la frecuencia de cada uno de los significados de las primeras 570 palabras del inglés, extraídas de un corpus de 5 millones de palabras compuesto por enciclopedias, novelas, revistas, ensayos, biografías, libros sobre ciencia, poesía, etc. (Lorge, 1953). Como veremos, este trabajo supondría pocos años después un gran punto de referencia para West en la elaboración de su *General Service List*.

De manera simultánea a los trabajos de Thorndike y Lorge en Estados Unidos, los lingüistas británicos Palmer y Hornby, desde su posición como investigadores en el *Institute for Research in English Teaching* (IRET) en Japón, y West, que trabajaba en Bengala (India), realizaron también una ingente labor con el fin de

establecer cuáles eran las palabras más frecuentes del inglés y, por tanto, las que sus alumnos necesitaban para mejorar su conocimiento de la lengua de forma progresiva.

Palmer y Hornby publicaron una serie de listados de frecuencias como por ejemplo *IRET's 600-Word Vocabulary for Story-Telling Purposes*, dedicado a lecturas de nivel elemental y *Thousand-Word English*, que fue utilizada en la elaboración de diccionarios y materiales pedagógicos (Smith, 2003). Además de estos listados de palabras, el trabajo de estos investigadores es especialmente reseñable en lo que se refiere a la frecuencia de las colocaciones. Como ya mencionamos en el capítulo anterior, estos autores, y muy especialmente Palmer, se pueden considerar pioneros en el estudio de estas unidades multiléxicas. Sin embargo, su innovación no se limitó únicamente a la consideración de las colocaciones como elementos importantes que forman parte del léxico de una lengua, sino que, dando un paso más allá y dentro del interés generalizado por la noción de frecuencia, también trató de establecer cuáles eran las colocaciones más frecuentes de la lengua inglesa y, por tanto, las que merecían especial atención desde el punto de vista pedagógico. Este listado de colocaciones frecuentes fue publicado en el año 1933 en el *Second Interim Report on English Collocations*, y constaba, según afirma Kennedy (2008), de varios miles de colocaciones, identificadas de forma subjetiva por el propio Palmer. No tenemos constancia de que este listado tuviera grandes repercusiones en el ámbito de la enseñanza del inglés como lengua extranjera, pero, sin duda, su valor es incalculable, no sólo por ser la primera compilación de colocaciones de la historia, sino porque, como veremos más adelante, fue el único que existió durante décadas.

Por su parte, West (1930), muy consciente de que los listados de frecuencias son extraordinariamente sensibles al tipo de material en el que están basados, consideraba que este tipo de listas eran especialmente útiles para el desarrollo de la lectura, frente al resto de las destrezas para las que la frecuencia era, en su opinión, de mucha menor utilidad (algo fácilmente explicable si consideramos que los únicos

corpus que se podían compilar en la época constaban de textos escritos dados los medios disponibles). Quizá por esta razón, unida al hecho de que, como ya comentamos, estos estudios se desarrollaron durante el periodo de mayor auge del Método de Lectura, West volcó buena parte de sus esfuerzos en la creación de una serie de lecturas graduadas, basada en unos principios objetivos y sistemáticos de selección léxica y donde el vocabulario nuevo se iba integrando en los distintos niveles de la serie de forma muy controlada. Esta novedosa serie como se indica incluso en su título, *New Method Reader Scheme*, que comenzó a publicarse en 1927, estaba basada fundamentalmente en el primer listado de frecuencias de Thorndike, una lista que era del agrado de West (1930) puesto que no sólo se fundamentaba en criterios de frecuencia sino que también se había considerado el rango de las palabras a la hora de incluirlas.

Así, a mediados de los años 30 y tras más de una década de investigaciones en torno a la frecuencia de las palabras, el trabajo de todos estos autores dio lugar a la publicación conjunta de *The Interim Report on Vocabulary Selection for the Teaching of English as a Foreign Language* (Faucett et al. 1936), también conocido como el *Carnegie Report*. En este informe se publicó un listado de palabras recomendadas para la elaboración de lecturas graduadas, que más tarde se utilizaría asimismo para producir gramáticas y diccionarios monolingües. Cabe destacar que la frecuencia no fue el único criterio fundamental de selección en este listado; también se consideraron aspectos como el valor estructural (por lo que incluyeron todas las palabras funcionales), el rango (es decir, su utilidad en distintos tipos de textos), la utilidad para definir otras palabras (con vistas a su función en trabajos lexicográficos), la capacidad productiva (para crear nuevas palabras) y el registro. Tras su publicación, y con el evidente retraso que ocasionó la Segunda Guerra Mundial, West realizó una revisión y edición de este listado, añadiendo información relativa a la frecuencia

tomada a partir del listado de Thorndike y Lorge (1944), y lo publicó en 1953 como *A General Service List of English Words* (GSL).

La GSL es una lista que contiene 2.000 lemas (o *headwords*), cada uno de los cuales representa una familia de palabras. De nuevo, y dado que, como ya hemos visto, West (1953: ix) creía firmemente que "[f]requency is not, of course, the only point to be considered in selecting items for teaching English", la frecuencia fue uno de los criterios que se tuvieron en cuenta en la elaboración de este listado, pero no el único. En su compilación West también consideró el esfuerzo de aprendizaje que requería la palabra, el grado de redundancia que suponía teniendo en cuenta otras palabras del listado, su eficacia a la hora de expresar distintas ideas (es decir, su cobertura), su registro (se excluyeron palabras de registro formal y coloquial, preservando las neutras) y su valor expresivo (se excluyeron palabras con fuerte carga emocional puesto que no eran prioritarias para la expresión de ideas generales y neutras). Es interesante destacar que en algunos casos West añade unas anotaciones finales con el fin de justificar la inclusión o exclusión de ciertas palabras a pesar de su alta o baja frecuencia (Tabla 2.1.).

La GSL está organizada a modo de diccionario. Las palabras están ordenadas alfabéticamente y para cada una de ellas se ofrece una breve definición y ejemplos de su uso dentro de un contexto. Asimismo, por cada lema se incluye su frecuencia estimada según un banco de 5 millones de palabras, y sus distintos significados se expresan de forma separada, ofreciendo un porcentaje de la frecuencia que tiene cada significado con respecto a la frecuencia total del lema. Asimismo, también aparecen las formas derivadas de cada lema, expresadas en negrita minúscula.

Mostramos a continuación la entrada del lema "*desire*" con el fin de ilustrar lo anterior.

DESIRE, n.	1032	(1) *(wish, request)*	
		A desire to visit Egypt	
		Expressed a desire to see the papers	**50%**
		(2) *(lust)*	
		Animal desires	**2%**
		(3) *(thing desired)*	
		She was called the World's Desire	**4%**
desire, v.		*(wish)*	
		I desire happiness; I desire to be happy; I desire that you may be happy; I desire my son to be happy	
		It shall be as you desire	**33%**
		[*request*, I desire you to, 5%]	
desired, adj.		Long-desired	**5%**
(This word may be taught rather late; but it is difficult to find a substitute for meaning (2)*.)*			

Tabla 2.1.: Lema "*desire*" en la GSL (West, 1953: 124)

La GSL es sin duda uno de los listados más conocidos en el ámbito académico, y su enorme influencia ha quedado reflejada en la gran cantidad de materiales pedagógicos de todo tipo (libros de texto, lecturas graduadas, tests de vocabulario e incluso programas informáticos de análisis de dificultad textual como *Web VocabProfile*[2]) que se han elaborado partiendo de la información que contiene. Evidentemente, los datos que se refieren a la frecuencia de las palabras son cada vez más cuestionables teniendo en cuenta la antigüedad del corpus del que fueron extraídos (según West (1953) los textos que lo componen fueron publicados en 1938 y 1949), pero a pesar de ello, actualmente sigue siendo considerado como uno de los listados de frecuencias más importantes y útiles de cuantos se han producido en inglés,

2 Analizador de textos, creado por Tom Cobb, disponible en http://www.lextutor.ca/vp/eng/ [Último acceso: 16.02.2009]

especialmente por la información que ofrece acerca de la frecuencia relativa de cada significado (Nation y Waring, 1997).

Finalmente, un último trabajo que merece nuestra atención en esta revisión de los primeros estudios sobre frecuencia léxica, y que pone también de manifiesto la importancia que tuvo esta corriente por lo mucho que contribuyó a la exploración lingüística es el llevado a cabo por George K. Zipf (1902-1950) en torno a la relación entre la frecuencia de las palabras y su posición en el listado. La hipótesis principal de sus estudios, que este lingüista pudo demostrar estadísticamente, dando lugar a una ley empírica conocida como "la ley de Zipf", establece que la frecuencia de cualquier palabra es inversamente proporcional al lugar que ocupa en la tabla de frecuencia (Zipf, 1932, 1935). Por tanto, la frecuencia de la primera palabra de un listado será doble que la de la segunda palabra, dos tercios superior a la de la tercera, y así sucesivamente.

Zipf no sólo formuló su ley en función a las palabras aisladas, sino que pudo comprobar que esta ley también se cumple con sintagmas e incluso con oraciones completas. Así pues, desde el punto de vista puramente lingüístico, el valor de este hallazgo es ciertamente importante puesto que abrió nuevas vías para la comprensión del lenguaje, y quizá incluso de la mente humana, basadas en términos de frecuencia y de probabilidad estadística. Como el propio Zipf (1935: 48) destacara:

> We select and arrange our words according to their meanings with little or no conscious reference to the relative frequency of occurrence of those words in the stream of speech, yet we find that words thus selected and arranged have a frequency distribution of great orderliness which (...) seems to be constant for language in general. The question arises as to the nature of meaning or meanings which leads automatically to this orderly frequency distribution. Whether this question can ever be completely solved quantitatively is probably doubtful, for meaning or meanings do not lend themselves to quantitative measurement. Yet, by the isolation of other factors which can be measured, we may gain a considerable insight into the nature of meaning, and perhaps finally apprehend something of its nature and behavior.

Desde la perspectiva pedagógica, esta ley no viene sino a confirmar la enorme importancia que tienen las primeras palabras de un listado para el alumno ya que su extraordinaria frecuencia, muy superior a la del resto de palabras de la lengua, supone una cobertura que sin duda le resultará muy rentable a la hora de comunicarse.

En resumen, uno de los valores más destacables de todos los estudios mencionados es el objetivo compartido de instaurar unos procedimientos objetivos y fiables para la exploración del vocabulario y/o la elaboración de materiales pedagógicos. En palabras del propio West (1937: 433),

> the results of a word-count are objective; a word-count proves that a certain word has value because it was actually encountered a large number of times in the representative material used in the study. Any divergence from this objective standard is liable to plunge the teacher into a sea of conflicting opinions.

Así pues, podemos decir que esta corriente de control del vocabulario supuso el primer intento por establecer principios racionales y científicos en el diseño de un sílabo para la enseñanza de lenguas, basándose fundamentalmente, y de forma pionera, en el valor de la frecuencia cuantificable de las palabras (Richards y Rodgers, 1986).

2.3. La frecuencia del vocabulario en la era computacional: El desarrollo de la lingüística de corpus

Si no fuera por la más que notable actividad desarrollada por académicos de la talla de John R. Firth, John M. Sinclair y Michael A. Halliday, podríamos afirmar que los años que siguieron a la etapa del "movimiento de control del vocabulario", e incluso hasta bien entrada la década de los 70, significaron un vacío total tanto en lo que se

refiere al interés por las cuestiones relacionadas con el léxico, como en cuanto a la noción de frecuencia, confirmando nuestra apreciación inicial de que ambos fenómenos han ido tradicionalmente de la mano a lo largo de los años.

Desde finales de los años 50 y hasta principios de los 70, el panorama de la lingüística internacional estuvo dominado por las teorías formuladas por Noam Chomsky (n. en 1928). Este lingüista estadounidense, padre de la Lingüística Generativo-Transformacional, aplicó las teorías cognitivistas de la época al estudio del lenguaje, con el fin de establecer la existencia de unos principios lingüísticos universales basados en estructuras sintácticas. Así, situando la sintaxis en el centro del estudio lingüístico y filológico, postuló que todas las lenguas del mundo comparten los mismos parámetros básicos, heredados de la estructura profunda e innata que existe en la mente humana (Chomsky, 1965). En sus teorías, por tanto, las cuestiones léxicas y los patrones lingüísticos, así como las relaciones entre ambos que se habían empezado a descubrir años antes gracias a los estudios estructuralistas (Fries, 1952), quedaban en un segundo plano, relegadas ahora a la periferia del lenguaje.

Otro de los aspectos importantes en los estudios llevados a cabo por Chomsky (1965: 3) fue la distinción que estableció entre *competence* y *performance*[3]:

> Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its (the speech community's) language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of this language in actual performance.

Así pues, según sus teorías, *competence* es el conocimiento innato que todos los hablantes tenemos de la lengua, la estructura profunda que se corresponde con una

[3] Esta dicotomía ha sido reformulada en publicaciones más recientes (Chomsky, 1992) como *I-Language* (*internalized language*) y *E-Language* (*externalized language*), aunque la distinción básica continúa siendo la misma que se estableció entre *competence* y *performance*.

serie de reglas sintácticas idealizadas y conceptuales que conforman el núcleo de la lengua, lo que el investigador debe estudiar. Por otro lado, *performance* es la materialización real de esas reglas, es decir, el lenguaje que producimos en situaciones comunicativas reales y bajo circunstancias concretas. Para Chomsky, este último aspecto del lenguaje no son sino ejecuciones imperfectas y superficiales de las reglas teóricas que existen en nuestra mente, por lo que no merecen la atención del lingüista. Todos los elementos que pertenecen al ámbito del "uso" lingüístico y que sólo se pueden analizar a través de la observación del lenguaje real como es el caso de los patrones léxicos, las estructuras recurrentes, las fórmulas que los hablantes tienden a repetir una y otra vez, en definitiva, todos los aspectos relativos a la naturaleza prefabricada del lenguaje y a la influencia que la frecuencia de uso ejerce sobre la realidad lingüística, para Chomsky no son sino elementos periféricos, puesto que el léxico no es más que un elemento de relleno que sirve para dar forma a la estructura gramatical. El propio autor (1962: 128) cuestionaba los estudios basados en la frecuencia léxica con estas palabras: "It seems that probabilistic considerations have nothing to do with grammar, e.g. surely it is not a matter of concern for the grammar of English that 'New York' is more probable than 'Nevada' in the context 'I come from–'".

Las teorías de Chomsky, como decíamos anteriormente, se desarrollaron de forma paralela a los trabajos realizados por Firth y sus discípulos. Firth, muy influenciado tanto por la obra de su profesor, el antropólogo B. Malinowski, como por los estudios de la Escuela de Praga en torno a la importancia del contexto social y cultural del lenguaje, consideraba que el aspecto central de la lengua es el significado, entendido como un fenómeno que depende del contexto situacional. Para él, todos los niveles lingüísticos establecidos por la tradición estructuralista (el fonético, el léxico, el gramatical y el semántico) confluyen a la hora de crear el

significado, a la vez que todos están delimitados por las elecciones que realiza el hablante atendiendo al contexto. Los elementos que componen el contexto situacional —los participantes, sus acciones verbales y no verbales y las consecuencias o efectos del acto verbal— constituyen así el eje central de su noción de significado. Como vemos, de acuerdo con su teoría y en total contraposición con la opinión de Chomsky, el significado, y por tanto el interés lingüístico, está totalmente ligado al uso de la lengua, a la *performance*, en términos chomskianos. Para Firth, la lengua sólo puede estudiarse a partir del análisis de ejemplos concretos, es decir, de muestras donde el lenguaje haya sido producido dentro de un contexto dado y con una función social determinada. Como el propio autor (1957: 53) escribiera:

> We must take our facts from speech sequences, verbally complete in themselves and operating in contexts of situation which are typical, recurrent, and repeatedly observable. Such contexts of situation should themselves be placed in categories of some sort, sociological and linguistic, within the wider context of culture.

La teoría contextual del significado de Firth y el interés que mostrara por el estudio del uso lingüístico caló muy hondamente entre sus discípulos. Son especialmente destacables los trabajos llevados a cabo por John Sinclair y Michael Halliday en torno al análisis de la lengua real tal y como la usan los hablantes en contextos concretos, estudios que volvieron a evidenciar la existencia de elementos y patrones recurrentes que demuestran el enorme valor de la frecuencia en (el uso de) la lengua. Para llevar a cabo este tipo de estudios, se hacía claramente imprescindible contar con una gran cantidad de textos auténticos que poder explorar, lo cual condujo irremediablemente a recuperar la tradición de compilar bancos de datos. Para entonces, a mitad de los años 60, se empezaban a producir los primeros avances en el campo de la informática, lo cual facilitaría enormemente la labor de compilación y análisis de textos. Estaba naciendo la lingüística de corpus computerizada y, con ella, la revalorización de la noción de frecuencia en los estudios léxicos.

2.3.1. La lingüística de corpus en la era informatizada

Aunque, como esperamos haya quedado comprobado en las secciones anteriores, el uso de los corpus lingüísticos no era en modo alguno novedoso en los años 60, sí parece acertado afirmar que fue hacia mediados de esta década cuando esta técnica de exploración de la lengua comenzó a tomar mayor fuerza y a consolidarse como metodología de análisis lingüístico[4]. Gracias, como decíamos, a la importancia que algunos lingüistas, con Sinclair a la cabeza, otorgaban al análisis de la lengua real producida por los hablantes en situaciones concretas, así como al desarrollo que paralelamente se estaba produciendo en el ámbito de la informática con la creación de los primeros ordenadores, esta época marcó el inicio de los estudios de corpus tal y como se entienden hoy en día. Son muchísimos los investigadores que desde aquellos años hasta la actualidad se han venido ocupando de esta nueva forma de estudiar el lenguaje, tratando de establecer las características que todo corpus debe tener, diseñando corpus de distinta naturaleza así como herramientas computacionales para su procesamiento y análisis, y evaluando las posibilidades que ofrecen estos bancos de datos en ámbitos tan diversos como la descripción lingüística, la enseñanza de lenguas, la traducción, el análisis literario, etc. Brevemente, pasamos a describir algunas de estas cuestiones generales sobre el diseño y las características de los corpus informatizados, antes de abundar en las repercusiones que esta metodología ha tenido en el área de los estudios de frecuencia léxica.

[4] Aunque algunos autores conciben la lingüística de corpus como una disciplina con un estatus teórico que la convierte en una rama independiente de la lingüística (Tognini-Bonelli, 2001), nosotros compartimos la opinión de McEnery, Xiao y Tono (2006) para quienes, a diferencia de disciplinas consolidadas como la semántica o la sintaxis que describen un aspecto determinado del lenguaje, la lingüística de corpus es una metodología que se puede utilizar en cualquiera de estas disciplinas para explorar casi cualquier aspecto de la lengua.

2.3.1.1. Definición y características de los corpus

De forma general y en términos sencillos, podemos definir un corpus como un gran banco de textos que se emplea para representar el lenguaje natural y que sirve a los investigadores para obtener muy diversos tipos de información lingüística. Según Crystal (1991: s.v.), un corpus es "a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language". Otra definición interesante es la que ofrece Sinclair (1996, citado por McEnery, Xiao y Tono, 2006: 4), para quien "[a] corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language". Los criterios lingüísticos a los que se refiere Sinclair son externos a la propia lengua y deben contribuir a organizar y diseñar el corpus de forma racional, respondiendo a la finalidad con que se ha compilado. En este sentido, para que un corpus se considere fiable y bien fundamentado debe haberse elaborado atendiendo a los siguientes parámetros (Kennedy, 1998; McEnery, Xiao y Tono, 2006):

a) **Representatividad:** Un corpus debe ser una muestra representativa de la lengua o lenguaje específico que se pretende estudiar, con el fin de que los resultados derivados de su análisis se puedan generalizar en la mayor medida posible para que sean válidos. Los factores que son en buena parte responsables de la representatividad de un corpus son el muestreo y el equilibrio:

   - **Muestreo:** A la hora de realizar la selección de textos para su inclusión en el corpus, se debe llevar a cabo un cuidadoso proceso de muestreo que recoja verdaderamente la amplia variedad de elementos

que contribuyen a la riqueza de la lengua o variedad lingüística que se quiere estudiar.

- **Equilibrio:** Muy relacionado con lo anterior, en la elaboración de un corpus se deben considerar las distintas variables de los textos (hablados y escritos, formales y no formales, de ficción y no ficción, producidos por hablantes de distintas edades, lugares de origen, etc.) ya que, en la medida de lo posible, deben distribuirse en igual proporción a la que existe en la lengua o en la variedad lingüística que se pretende representar.

b) **Tamaño:** Para que un corpus se pueda considerar representativo debe tener un tamaño considerable, aunque debemos ser conscientes de que un corpus, por muy grande que sea, nunca pasará de ser una minúscula muestra de toda la lengua que se produce diariamente en el mundo. Desde los años 80 y gracias a las posibilidades de almacenamiento y procesamiento que comenzaron a ofrecer los ordenadores, empezaron a proliferar corpus de millones de palabras. Es interesante comprobar en este sentido que ya a principios de los 90 Sinclair (1991: 20) consideraba que 20 millones de palabras no eran suficientes para lograr "a reliable description of the language as a whole".

Sin embargo, como ya apuntara Kennedy (1998: 68), "[a] huge corpus does not necessarily 'represent' a language or a variety of a language any better than a smaller corpus". Para este autor, el tamaño depende muy directamente del tipo de análisis que se pretenda llevar a cabo, por lo que un estudio de colocaciones exigirá un corpus de mayores dimensiones que un estudio de palabras frecuentes puesto que las colocaciones son un fenómeno mucho menos frecuente que las unidades léxicas individuales. Del mismo

modo, si lo que pretendemos es analizar un área específica del lenguaje, el corpus que necesitaremos será menor (siendo 25.000 palabras un tamaño aceptable según Tribble, 1997) que si deseamos estudiar la lengua en su totalidad.

2.3.1.2. Tipos de corpus

Según los distintos parámetros que se consideren en la elaboración de un corpus y, sobre todo, según la finalidad que se persiga en cada caso, éste puede ser de diversos tipos:

a) De referencia o especializados:

- Un corpus de referencia es aquel que ha sido creado para ser una muestra representativa de una lengua, de forma que ofrezca la más amplia información posible sobre dicha lengua y pueda servir como base, por ejemplo, para la elaboración de gramáticas y diccionarios generales.
- Por su parte, un corpus especializado es el que se diseña para representar una variedad lingüística concreta o un tipo de lenguaje especializado. Se considera de este tipo, por ejemplo, un corpus que sólo contenga textos académicos, jurídicos o periodísticos.

b) Fijos (cerrados) o "monitores" (abiertos):

- Son fijos o cerrados los corpus que se compilan durante un periodo de tiempo concreto y una vez concluidos no se le vuelven a añadir textos nuevos. Son corpus de dimensiones fijas, representativos del periodo de tiempo concreto al que pertenecen los textos que lo

integran y por tanto susceptibles de quedar obsoletos con el paso del tiempo.

- Siguiendo la terminología de Sinclair (1991: 26), un corpus "monitor" es aquel cuyo tamaño no deja de aumentar dado que se siguen añadiendo textos nuevos cada cierto tiempo. Evidentemente, la principal ventaja de este tipo de corpus es que está más actualizado que los fijos y permite observar la evolución de las lenguas, aunque también se debe considerar que, al sufrir constantes alteraciones, el muestreo y el equilibrio de este tipo de fuentes no es siempre tan riguroso como el de los corpus cerrados y, si no es modificado cuidadosamente, puede no ser tan fiable desde el punto de vista de los análisis cuantitativos (Kennedy, 1998).

c) Sincrónicos o diacrónicos:

- Son corpus sincrónicos aquellos integrados por textos que pertenecen a una época concreta y están concebidos para representar el lenguaje que se utiliza en ese periodo de tiempo en particular.
- Por el contrario, los corpus diacrónicos están compuestos por textos de distintas épocas históricas, y su principal finalidad es recoger y servir como testigos de la variación lingüística que se ha producido en una lengua o variedad de lengua con el paso del tiempo.

d) Monolingües o bilingües/multilingües (paralelos o comparables):

- Los corpus monolingües están compuestos por textos producidos en una sola lengua y sirven como representación de esa lengua concreta.
- Existen también corpus bilingües o multilingües, integrados por textos de diferentes lenguas. Dentro de esta categoría existen dos

posibles clases de corpus: los paralelos y los comparables. Un corpus paralelo es el que contiene textos producidos en una lengua y su traducción a otra/s lengua/s, siendo de fácil comparación en la actualidad gracias a programas informáticos que permiten explorar el texto original y su equivalente traducido de forma simultánea. Por otro lado, los corpus comparables están integrados por textos producidos en dos o más lenguas pero muestreados y distribuidos de tal manera que las secciones correspondientes a los distintos idiomas posean características y composiciones similares y hagan posible, así, la comparación interlingüística.

e) De hablantes nativos o de alumnos:

- La mayor parte de los corpus con los que se trabaja en la actualidad, y los únicos hasta hace relativamente poco tiempo, son aquellos compuestos por textos producidos por hablantes nativos.
- Los corpus de alumnos (*learner corpora*), que constan de textos generados por hablantes no nativos de la lengua en la que están escritos, son una importante innovación en la lingüística de corpus, ya que están permitiendo analizar de forma empírica los procesos interlingüísticos que predominan en la adquisición de lenguas.

f) Escritos y/u orales:

- Un corpus escrito es aquel que está conformado por textos producidos originalmente en esta modalidad de la lengua. Se trata claramente del modo más sencillo y rápido de compilar información lingüística, dado que ya se encuentran en un formato fácilmente compilable y analizable.

- Los corpus orales son aquellos en los que los textos que lo integran fueron originalmente producidos de forma oral y han sido posteriormente transcritos para poder incluirlos en el corpus. Evidentemente, su compilación exige mayor tiempo y dedicación que en el caso de los escritos, por lo que son menos abundantes, pero sin duda supusieron una gran innovación sobre todo a partir de los años 80, ya que hasta entonces los lingüistas siempre habían tenido que conformarse con la recopilación y estudio del lenguaje escrito dada la falta de medios técnicos para almacenar textos orales (West, 1937). En muchos casos, existen corpus mixtos donde se recogen textos en ambas modalidades.

g) Textuales ("monomodales") o multimodales:

- Los corpus textuales son aquellos compuestos por textos registrados y almacenados en formato escrito (bien porque se produjeron de forma escrita o porque han sido transcritos para poder ser incluidos en el corpus) y que sólo se pueden analizar desde el plano de la palabra impresa.
- Gracias al enorme avance de las nuevas tecnologías, hoy en día están comenzando a aparecer los primeros corpus multimodales, caracterizados por su contenido audiovisual. Se trata generalmente de corpus donde se recogen fragmentos de vídeo a los que se puede acceder desde tres vías o modalidades distintas: el texto escrito (transcripciones de la comunicación verbal), el audio y la imagen. Este tipo de corpus, que están suponiendo una verdadera revolución en el área del análisis lingüístico, permiten abordar cuestiones pragmáticas y contextuales que, hasta muy recientemente, eran difíciles de captar y

> estudiar, permitiéndonos así indagar ahora de forma empírica en los postulados formulados por Firth hace ya medio siglo.

En la actualidad, evidentemente, existen numerosos ejemplos de cada uno de los tipos de corpus que acabamos de enumerar. Gracias, como decíamos, a las ventajas que ofrecen hoy en día las nuevas tecnologías, resulta relativamente sencillo hallar y compilar una gran cantidad de textos de toda índole de forma informatizada, así como manipular y explorar posteriormente dichos corpus para muy diversos fines, por lo que en los últimos años han proliferado los proyectos de compilación de corpus de todo tipo. No nos detendremos, sin embargo, a describir todos estos corpus de forma detallada puesto que ello se alejaría del objetivo de nuestro trabajo, en el que pretendemos utilizar los corpus de referencia más representativos del inglés, con el fin de obtener un listado de colocaciones frecuentes y pedagógicamente interesantes que integrarán nuestro test. Así pues, dedicaremos el siguiente apartado a la descripción de dichos corpus de referencia, para pasar seguidamente a ocuparnos de las repercusiones que han tenido en los estudios de frecuencia léxica.

2.3.1.3. Breve revisión de los principales corpus electrónicos de referencia del inglés

2.3.1.3.1. Corpus de primera generación

a. *Brown Corpus*

El primer corpus informatizado que existió de la lengua inglesa fue el ***Brown University Standard Corpus of Present-Day American English***, conocido

popularmente como ***Brown Corpus*** ya que fue compilado entre 1961 y 1964 por Nelson Francis y Henry Kučera en la *Brown University* (Providence, Rhode Island). Se trata, sin duda, de un corpus de enorme valor no sólo por ser pionero sino porque se desarrolló en la época en la que los postulados de Chomsky, claramente contrarios a las recopilaciones de datos lingüísticos, gozaban de mayor aceptación en los círculos académicos estadounidenses (Leech, 1993). Este corpus, con un tamaño total de 1.014.300 palabras, está formado por 500 textos escritos de inglés americano de unas 2.000 palabras de longitud cada uno, todos ellos publicados en 1961 y seleccionados mediante un cuidadoso proceso de muestreo (Francis y Kučera, 1964). La composición y distribución de los textos que lo integran es la siguiente (Tabla 2.2.):

Tipo de texto	Total de textos	Proporción en el corpus (%)
Prosa de no ficción	**374**	**75**
A. Reportajes de prensa	44	8,8
B. Editoriales	27	5,4
C. Crítica de prensa (cine, literatura, etc.)	17	3,4
D. Textos religiosos	17	3,4
E. Aficiones y pasatiempos	36	7,2
F. Tradiciones populares	48	9,6
G. Biografías, memorias, etc.	75	15
H. Miscelánea (principalmente documentos gubernamentales)	30	6
I. Textos académicos	80	16
Prosa de ficción	**126**	**25**
J. Ficción en general	29	5,8
K. Misterio y policíaca	24	4,8
L. Ciencia ficción	6	1,2
M. Aventuras	29	5,8
N. Romántica	29	5,8
O. De humor	9	1,8
TOTAL	**500**	**100**

Tabla 2.2.: Estructura del *Brown Corpus* (Francis y Kučera, 1964)

El *Brown Corpus* sirvió como ejemplo para la recopilación de otros corpus posteriores, tanto en lo que se refiere a la planificación de su diseño y estructura como en cuanto a la voluntad de servicio público demostrada por el hecho de que durante 30 años

aproximadamente estuvo disponible gratuitamente para la comunidad académica[5]. A pesar de que hoy en día se considera un corpus pequeño y relativamente anticuado, su valor principal reside en que supuso el punto de partida para la creación de corpus electrónicos, de los que tanto nos estamos beneficiando todos en la actualidad.

b. *Lancaster-Oslo/Bergen Corpus*

Con el fin de crear una réplica británica del *Brown Corpus*, entre los años 1970 y 1978 se llevó a cabo un proyecto para la compilación del ***Lancaster-Oslo/Bergen Corpus (LOB Corpus)***, liderado por Geoffrey Leech en la Universidad de Lancaster y Stig Johansson en la Universidad de Oslo, y con la colaboración de Knut Hofland en el Centro Informático Noruego para las Humanidades de Bergen. Como su equivalente norteamericano, se trata de una compilación de un millón de palabras aproximadamente (en el caso de este corpus son exactamente 1.006.825 palabras, lo cual supone casi 7.500 palabras menos que el *Brown Corpus*, una diferencia que se debe a que la longitud de 2.000 palabras de los textos incluidos es aproximada, pero no exacta) tomadas de 500 textos escritos de inglés británico y publicados en 1961. Tanto el proceso de muestreo como las proporciones utilizadas en su elaboración fueron prácticamente las mismas que se habían empleado en el *Brown Corpus* (a excepción de variaciones mínimas en el número de textos pertenecientes a algunas categorías) por lo que pronto se pudo confirmar que se trataba verdaderamente de dos corpus gemelos que permitían llevar a cabo estudios comparativos fiables de ambas variedades de la lengua. El hecho de que, por ejemplo, los listados de frecuencias obtenidos de ambos corpus mostrasen resultados muy similares en las

[5] Actualmente varias versiones de este corpus, así como también de los dos corpus que describiremos en las secciones siguientes (el *LOB Corpus* y el *London-Lund Corpus*) se encuentran disponibles en CD-Rom previa subscripción, a través del *International Computer Archive of Modern and Medieval English* (ICAME): http://icame.uib.no/newcd.htm [Último acceso: 23.02.2009]

palabras gramaticales, demostraba claramente el grado de similitud que existía entre ambos (Johansson, 1980).

Estos dos corpus marcaron las pautas que se siguieron en un buen número de proyectos durante la década de los 80 para la elaboración de corpus similares en otros puntos del mundo[6]. Mientras tanto, y durante años, el *Brown Corpus* y el *LOB Corpus* fueron ampliamente analizados y comparados en distintos estudios (por ejemplo, Kennedy, 1987; Johansson y Norheim, 1988), mostrando con ello las posibilidades que ofrecían los medios informáticos para la exploración del lenguaje.

c. *London-Lund Corpus*

El último de los corpus electrónicos de primera generación que citaremos en este trabajo por su enorme valor como el primer corpus oral en formato informatizado, y el mayor y más fiable que existió hasta bien entrados los 90, es el ***London-Lund Corpus*** **(LLC)**. Durante los años 60, Randolph Quirk y sus colaboradores en el *University College London* habían creado el *Survey of English Usage Corpus* (SEU), un corpus en papel que contaba con un millón de palabras aproximadamente, de las cuales la mitad pertenecían al lenguaje hablado y la otra mitad a la modalidad escrita. En 1975, la parte oral de este corpus fue informatizada en la Universidad de Lund gracias al trabajo de Jan Svartvik, quien añadió 13 textos a los 87 que originalmente contenía el SEU logrando así completar las 500.000 palabras deseadas. El corpus resultante, llamado *London-Lund Corpus*, no sólo contenía la transcripción ortográfica de los textos, sino que también incluía información fonética y prosódica que permitió realizar grandes avances en el estudio del lenguaje oral.

[6] Algunos de los corpus más conocidos de este tipo son el *Kolhapur Corpus of Indian English*, el *Wellington Corpus of Written New Zealand English*, el *Australian Corpus of English* y el *Corpus of English-Canadian Writing*.

2.3.1.3.2. Corpus de segunda generación

El salto cualitativo hacia la producción de los llamados "second generation mega-corpora" (Kennedy, 1998: 45), que cuentan ya con cientos de millones de palabras, lo marcó el proyecto COBUILD (*Collins-Birmingham University International Language Database*), basado en el ***Bank of English*** y dirigido por el Profesor John Sinclair. En la actualidad, los corpus más representativos de la lengua inglesa y sin duda los más desarrollados son, además del ya mencionado *Bank of English*, el ***British National Corpus***, el ***Longman Corpus Network*** y el ***Cambridge International Corpus***. Estos serán, pues, los corpus en los que nos detendremos en el resto de esta sección.

a. *Bank of English*

Como acabamos de mencionar, el ***Bank of English***[7] (BoE) es el corpus recopilado por el proyecto COBUILD en la Universidad de Birmingham. Este trabajo comenzó en 1980 con el objetivo de crear un gran banco de datos para la elaboración de un diccionario y la observación y estudio de la lengua. Con este fin, se recogieron textos en prosa tanto escritos como orales que reflejaban el lenguaje estándar y donde se eliminaron los rasgos dialectales y el lenguaje técnico (Renouf, 1987). En 1982 el corpus contaba con más de 7 millones de palabras, cifra que aumentó hasta los 20 millones en tan solo tres años. En 1991, cuando el corpus alcanzaba los 200 millones de palabras, se decidió seguir incrementando el tamaño del corpus incorporando nuevas entradas de forma constante dado que los datos y resultados extraídos a partir de un corpus adquieren una fiabilidad mayor cuanto más extenso sea el corpus (López-Mezquita, 2005). Se trata por tanto de un corpus que, a diferencia de la gran mayoría, es abierto o "monitor".

[7] http://www.collins.co.uk/books.aspx?group=153 [Último acceso: 23.02.2009]

En la actualidad y desde su última actualización en 2004 el BoE consta de 524 millones de palabras. Está compuesto por textos escritos y orales obtenidos de miles de fuentes distintas que reflejan el lenguaje real que utilizan los hablantes del inglés hoy en día. Dos terceras partes de los textos escritos corresponden a publicaciones hechas por los medios de comunicación, aunque también contiene textos provenientes de libros de ficción y no ficción, folletos y páginas web. Por su parte, los textos orales suman más de 20 millones de palabras del total del corpus y han sido extraídos de transcripciones de conversaciones informales, entrevistas, emisiones de radio y televisión, reuniones y debates. Por otro lado, en este corpus también están representadas distintas variedades del inglés ya que, aunque la mayoría de los datos (el 70%) provienen de fuentes británicas, el 20% de los textos proceden del inglés americano y aproximadamente un 5% corresponde a otras variedades de la lengua como son el australiano o el canadiense.

Es importante destacar que, por ahora, este corpus sólo está disponible de forma íntegra para los investigadores de la Universidad de Birmingham. Fuera de esta Universidad, tan sólo se permite el acceso a un subcorpus de 56 millones de palabras, aunque recientemente HarperCollins ha hecho pública su intención de ampliar este límite y de mejorar las prestaciones del programa informático necesario para la exploración de este corpus (ver sección 3.2.2.3.).

b. *British National Corpus*

El ***British National Corpus***[8] (BNC) es una recopilación de 4.124 textos tanto escritos como orales que suma 100.046.235 palabras, diseñada para representar de manera fiable el inglés británico actual. Este proyecto, financiado en gran medida por el gobierno británico y por editoriales e instituciones académicas, fue llevado a cabo

[8] http://www.natcorp.ox.ac.uk/ [Último acceso: 24.02.2009]

por un consorcio liderado por la editorial *Oxford University Press* y en el que el resto de miembros son las editoriales *Addison-Wesley Longman* y *Larousse Kingfisher Chambers*, las Universidades de Oxford y Lancaster y la Biblioteca Británica.

La creación del BNC comenzó en el año 1991 y concluyó en 1994, publicándose la primera versión en 1995. Dado que se trata de un corpus cerrado, no se han vuelto a añadir textos desde entonces. Sin embargo, en los años 1995/96 el equipo de la Universidad de Lancaster realizó una revisión del corpus reclasificando algunos textos (casi un millón de palabras de la parte escrita pasaron de la sección de textos de ficción a la de textos informativos) y una mejora en el etiquetado de las categorías gramaticales que dio lugar a la publicación en 2001 de *BNC World*, la segunda versión del BNC. Por último, en 2007 apareció la *BNC XML Edition*, la tercera y por el momento última versión de este corpus.

Las fuentes de las que se extrajo el BNC fueron seleccionadas de tal manera que se garantizara la representatividad del inglés británico actual, con un número equilibrado de textos pertenecientes a los distintos estilos y registros de la lengua. Para lograr que el corpus sirviera como muestra del inglés de nuestra época, la gran mayoría de los textos que se recogieron (más del 93%) fueron producidos entre los años 1985 y 1994 e incluso en el caso de la parte oral del corpus no hay ninguno anterior al año 1991 (Leech, Rayson y Wilson, 2001).

Este corpus está compuesto de un 90% de textos escritos y un 10% de textos orales. Obviamente, la proporción no está todo lo equilibrada que cabría desear. Sin embargo, como Leech, Rayson y Wilson (ibid.: 1) destacan,

> it is a skilled and very time-consuming task to transcribe speech into the computer-readable orthographic text that can be processed to extract linguistic information. In view of this problem, these proportions were chosen as realistic targets which, given the size of the BNC, are also sufficiently large to be broadly representative.

En lo que respecta a la parte escrita del corpus, ésta se compone de dos tipos de textos: por un lado, los de carácter "informativo" (escritos sobre ciencias, actualidad mundial, economía, artes, pensamiento, deportes, etc.), suman un 75% del total y todos fueron producidos a partir de 1975. La distribución de este tipo de textos se hizo también atendiendo a las tendencias de publicación de los distintos géneros durante los años 80 y 90. Por otro lado, los textos de naturaleza "imaginativa" (obras literarias y de creación), completan el 25% restante de la parte escrita del BNC y recoge textos producidos entre 1960 y 1994.

De forma más gráfica, podemos observar la composición de la parte escrita del BNC en la tabla 2.3.

Inglés escrito (90%)	
Textos informativos (75%)	
Temas: ▪ Ciencias naturales (5%) ▪ Aspectos sociales (15%) ▪ Comercio y economía (10%) ▪ Pensamiento y religión (5%) ▪ Ciencias aplicadas (5%) ▪ Actualidad mundial (15%) ▪ Artes (15%) ▪ Entretenimiento y tiempo libre (15%)	*Nivel:* ▪ Especialista (30%) ▪ Lego (50%) ▪ Popular (20%) *Fecha:* ▪ 1975-1994
Géneros: ▪ Libros (55-65%) ▪ Publicaciones periódicas (20-30%) ▪ Miscelánea publicada (5-10%) ▪ Miscelánea no publicada (5-10%) ▪ Escritos para ser hablados (2-7%)	

Textos de ficción (25%)	
Nivel:	*Fecha:*
▪ Literario (33%)	▪ 1960-1974 (25%)
▪ Medio (33%)	▪ 1975-1994 (75%)
▪ Popular (33%)	

Tabla 2.3.: Sección escrita del BNC, versión 1.0 (adaptada de Leech, 1993: 13)

En cuanto al componente oral del corpus, el mayor existente en la época en que se elaboró, éste también consta de dos tipos de grabaciones. La primera mitad, tradicionalmente denominada "demographic" (Kilgarriff, 1997: 138) aunque Leech, Rayson y Wilson (2001: 2) prefieren referirse a ella como "conversational part", son conversaciones informales que comprenden más de 2.000 horas de grabación. Este material fue recogido con la colaboración de 124 voluntarios (y sus interlocutores) en cuya selección se cuidaron especialmente las cuestiones de edad, sexo, procedencia geográfica y clase social para lograr un equilibrio en los datos. El 50% restante de la variedad oral lo componen los textos denominados "context-governed" (Kilgarriff, 1997: 138) o "task-oriented" (Leech, Rayson y Wilson, 2001: 2), que fueron grabados en reuniones, discursos, debates parlamentarios, entrevistas, programas de radio donde existe interacción entre varias personas, etc. Este tipo de textos también se recogió atendiendo a la variedad geográfica, realizando grabaciones en doce regiones diferentes del Reino Unido (Kennedy, 1998). Posteriormente, todo este material fue transcrito para ser incluido en el corpus[9].

Como sostienen Leech, Rayson y Wilson (2001), una de las principales ventajas que ofrece el BNC respecto al resto de corpus de grandes dimensiones es que es un corpus cerrado, equilibrado y basado en muestras. En consecuencia se trata de un corpus estable en el que, debido al esfuerzo por recopilar datos de manera

[9] La procedencia detallada de cada archivo del corpus se puede consultar en: ftp://ftp.itri.bton.ac.uk/bnc/bib-dbase [Último acceso: 24.02.2009]

equitativa que sirvieran de muestra de los diversos tipos de lenguaje oral y escrito, la representatividad y fiabilidad de sus resultados parecen palpables.

c. *Longman Corpus Network*

El ***Longman Corpus Network***[10] consta de cinco bases de datos distintas y suma un total de 155 millones de palabras. En primer lugar, contiene el *Longman/Lancaster English Language Corpus*, de 30 millones de palabras, diseñado en la Universidad de Lancaster a partir de textos de muy diversas fuentes, de ficción y no ficción, producidas a partir del año 1900. Su principal finalidad era ser una muestra representativa del inglés general del siglo XX, sin estar restringida a variedades geográficas ni áreas específicas (Summers, 1991).

En segundo lugar, cuenta con un corpus de alumnos, el *Longman Learners' Corpus*, de 10 millones de palabras de inglés escrito, para cuya elaboración se recogieron textos procedentes de respuestas de exámenes, cartas, ensayos y diarios producidos por estudiantes de inglés de 160 nacionalidades y ocho niveles de dominio lingüístico diferentes.

El tercer corpus con el que cuenta es el *Longman Written American Corpus*, un corpus "monitor" que en la actualidad cuenta con 100 millones de palabras de inglés escrito en la variedad americana. Está compuesto por textos extraídos de periódicos, revistas, novelas, y escritos de carácter científico y técnico, entre otros, a los que constantemente se añade material nuevo.

Asimismo, el *Longman Corpus Network* también recoge un corpus de lenguaje oral en la variedad americana, el *Longman Spoken American Corpus*, cuyos 5 millones de palabras han sido compilados en la Universidad de California con la colaboración de

[10] http://www.pearsonlongman.com/dictionaries/corpus/ [Último acceso: 24.02.2009]

más de 1.000 participantes de diversas edades, razas, niveles sociales y procedencias geográficas.

Finalmente, en este corpus también se incluyeron los 10 millones de palabras que integran el componente oral del BNC, dado que fue inicialmente elaborado por la editorial Longman.

d. *Cambridge International Corpus*

El último corpus que revisaremos en esta sección, el ***Cambridge International Corpus*** (CIC)[11], es también un corpus "monitor" que en la actualidad consta de más de 1.100 millones de palabras de inglés oral y escrito. Esta colección está siendo elaborada por Cambridge University Press desde hace más de 10 años con el fin de contar con una fuente de información a la que poder recurrir para diseñar materiales pedagógicos de distinto tipo. De hecho, debemos decir que, a diferencia de otros corpus, el CIC sólo está disponible para los investigadores de la editorial Cambridge University Press y de la Universidad de Cambridge.

Este corpus está integrado por miles de textos de muy diversa procedencia: prensa, libros de ficción y no ficción de distintos temas, radio y televisión, grabaciones de conversaciones reales, páginas web, correo basura, etc. En la actualidad, la composición de este corpus es la siguiente (Tabla 2.4.):

[11] http://www.cambridge.org/elt/corpus/international_corpus.htm [Último acceso: 24.02.2009]

Inglés británico	
Corpus	**Número de palabras**
Inglés británico escrito	700 millones
Inglés británico oral incluyendo el corpus CANCODE	18 millones
Inglés británico escrito (lenguaje académico)	20 millones
Inglés británico escrito (lenguaje comercial)	60 millones
The Cambridge and Nottingham spoken Business English Corpus (CANBEC)	1 millón
Inglés americano	
Corpus	**Número de palabras**
Inglés americano escrito	275 millones
Inglés americano oral incluyendo el *Cambridge-Cornell Corpus of Spoken North American English*	30 millones
Inglés americano escrito (lenguaje académico)	9 millones
Inglés americano escrito (lenguaje comercial)	40 millones
Inglés de alumnos (no nativos)	
Corpus	**Número de palabras**
Cambridge Learner Corpus	30 millones
Corpus de alumnos con errores codificados	15 millones

Tabla 2.4.: Composición del *Cambridge International Corpus* en la actualidad

En la variedad británica, es especialmente destacable la sección de inglés oral compuesta en parte por el Corpus CANCODE (*Cambridge and Nottingham Corpus of Discourse in English*). CANCODE consta de 5 millones de palabras recogidas entre los años 1995 y 2000, mediante grabaciones realizadas en distintas zonas del Reino Unido. Todo este corpus contiene interacciones naturales producidas en muy diversas situaciones: conversaciones espontáneas informales, conversaciones entre personas en su lugar de trabajo, transacciones comerciales, solicitudes de información, debates, etc. Un rasgo distintivo y muy interesante del CANCODE que explica en gran medida su prestigio es que las grabaciones se codificaron atendiendo

al contexto en que se produjeron. Así, se especifica la relación existente entre los interlocutores, distinguiendo si se trata de personas que guardan una estrecha relación, que viven juntas, si son conocidos, compañeros de trabajo o desconocidos; este rasgo, evidentemente, permite analizar en qué sentido el lenguaje se ve afectado por los distintos grados de familiaridad o de formalidad que existen entre los interlocutores.

Finalmente, nos parece interesante comprobar que, al igual que sucede con el *Longman Corpus Network*, se trata de un corpus que incluye no sólo inglés británico y americano, sino que también tiene un componente de textos escritos por hablantes no nativos. Este corpus, denominado *Cambridge Learner Corpus*, cuenta actualmente con 30 millones de palabras y está siendo elaborado a partir de los textos producidos por alumnos que realizan los exámenes ESOL de Cambridge, de todas las edades y nacionalidades del mundo. Sin duda, la información que aporta este corpus es de enorme valía desde el punto de vista del diseño de materiales para la enseñanza del inglés, ya que permite a los investigadores y autores conocer cuáles son los errores y los aspectos más problemáticos para los alumnos de las distintas nacionalidades, pudiendo así elaborar materiales más adecuados y eficaces para cada tipo de alumnado.

Como ya mencionamos anteriormente, los aquí incluidos no son en modo alguno los únicos corpus del inglés que existen en la actualidad, pero sí nos parecen los más interesantes y representativos de cuantos se han elaborado hasta la fecha. Estos corpus son especialmente relevantes, asimismo, ya que han dado lugar a nuevos análisis y descubrimientos sobre el funcionamiento y la naturaleza del lenguaje. Esto es particularmente notable en el área de los estudios léxicos, que han tomado una nueva dimensión gracias al papel de los corpus en el estudio lingüístico, y sin duda en lo que respecta a la frecuencia como factor fundamental no sólo del

comportamiento lingüístico, sino también, y muy especialmente, en la enseñanza y aprendizaje de idiomas. En la próxima sección pasamos pues a ocuparnos de este aspecto.

2.3.2. La frecuencia léxica y la enseñanza de lenguas en la era computacional

2.3.2.1. Los años 60 y 70

Hemos mencionado anteriormente que el interés por las cuestiones léxicas, claramente marcado por motivaciones pedagógicas durante las primeras décadas del s. XX, decreció hasta casi desaparecer a mediados de siglo en los círculos académicos y no recobró verdadero valor hasta los años 80. Sin embargo, durante las décadas de los 60 y 70 se publicaron algunos trabajos que sin duda merecen nuestra atención ya que marcaron las pautas de la corriente que más tarde se desarrollaría en torno al vocabulario y su frecuencia en la enseñanza de lenguas extranjeras.

En este sentido, como ya adelantamos, cabe destacar en primer lugar los estudios realizados por John Sinclair a partir de mitad de los 60, reivindicando desde sus primeras publicaciones y en todo momento la importancia del léxico y de las colocaciones en la descripción lingüística. Adoptando el enfoque distribucional y estadístico que había sido insinuado primero por Firth (1957) y más tarde desarrollado por Halliday (1961) para el estudio del léxico y de las colocaciones, Sinclair comenzó a recopilar en los 60 pequeños bancos de datos informatizados con los que poder llevar a cabo análisis cuantitativos. Llevado por el convencimiento de que la frecuencia constituía un elemento fundamental en el estudio del vocabulario, Sinclair (1966/1996: 2-3) ya advertía en estos primeros estudios de la necesidad de

explorar los patrones recurrentes en el uso de la lengua, y de establecer nuevas teorías léxicas, inexistentes entonces, que pudieran explicar la naturaleza de esos patrones:

> It is at present impossible (…) to say anything at all objective about the lexical structure of a language. Whereas we can collect a few thousand examples of the verbal group and notice what patterns recur, there is no such general lexical category as 'verbal group' turning up several times a sentence, and there is no easy way of collecting a few thousand occurrences of any lexical item. Furthermore, it seems likely that the more common items will be much more difficult to describe than the rare ones. Consequently the theory of lexis is fairly rudimentary; (…) we have yet to see what a comprehensive description of the lexis of a language looks like.

En éste y otros trabajos posteriores (Jones y Sinclair, 1974; Sinclair, 1991) este autor profundizó en la noción de frecuencia, haciendo notar la importancia de las estructuras léxicas que tienden a co-aparecer en la lengua de forma significativamente frecuente y que conforman el núcleo del vocabulario de las lenguas, una apreciación que representa la base del enfoque estadístico a la fraseología y que Sinclair culminó con la proclamación de su "principio idiomático" (ver sección 1.3.1.1.2.). Es interesante destacar en este sentido que para su autor este principio es fundamental porque es el que se aplica en la mayor parte del lenguaje, mientras que la elección libre de elementos lingüísticos es secundaria y sólo se pone en funcionamiento cuando nos falla el principio idiomático. Esta observación fue más tarde corroborada por otros estudios: Erman y Warren (2000) estimaron que aproximadamente el 50% del lenguaje producido por un nativo es idiomático mientras que Altenberg (1998) aumenta este porcentaje hasta el 80%. Estos datos ponen claramente de manifiesto la enorme relevancia de la frecuencia en el uso de las lenguas, ya que demuestran hasta qué punto los hablantes nativos utilizan reiteradamente los patrones prefabricados. Con esto no queremos decir, por supuesto, que los hablantes no nativos utilicen o deban utilizar la misma proporción de unidades fraseológicas que los nativos, pero sí indica, en nuestra opinión, que se trata de unidades que merecen especial atención y

que deben ser seleccionadas atendiendo también a su frecuencia en la lengua puesto que pueden ser enormemente rentables en términos de cobertura lingüística.

En consonancia con los avances que ya se venían produciendo, como vemos, en el estudio de la frecuencia léxica en los 60 y principios de los 70, y continuando asimismo con la tradición instaurada por los lingüistas del movimiento de control de vocabulario, Richards (1976) confirmó la importancia de la frecuencia en el aprendizaje de vocabulario en su artículo "*The role of vocabulary teaching*", al que ya hicimos referencia al comienzo de este capítulo.

Además de asentar las bases de lo que significa conocer una palabra, un aspecto muy interesante de este pionero trabajo es el hecho de que en el segundo principio, donde se refería a la importancia de la frecuencia, Richards asocia esta noción al conocimiento de las colocaciones. La descripción del segundo principio citado anteriormente continúa (y concluye) de la siguiente manera (ibid.):

> The speaker of a language recognizes not only the general probability of occurrence of a word but also the probability of words being associated together with other words. Knowledge of collocation means that on encountering the word *fruit* we can expect the words, *ripe, green (*= not ripe), *sweet, bitter* etc; that for *meat* we might expect, *tender, tough.*

En nuestra opinión, el hecho de que Richards relacione ambos aspectos (la frecuencia léxica con el conocimiento colocacional) parece reflejar la enorme influencia que ya en aquella época ejercían los postulados de Sinclair, dado que la colocación se concibe aquí en su dimensión probabilística, siendo la co-aparición de dos palabras una consecuencia de la frecuencia de la palabra individual.

Estos primeros esfuerzos por situar el vocabulario en un lugar central tanto de la descripción lingüística (Sinclair) como de la enseñanza de lenguas (Richards) comenzaron a dar sus frutos algunos años más tarde, a partir de la década de los 80. Esta será la etapa que nos ocupará a continuación.

2.3.2.2. Desde los años 80 hasta la actualidad: La vuelta a la enseñanza del vocabulario y a los listados de frecuencias

Desde los años 70, la enseñanza de lenguas estuvo (y en ciertos casos sigue estando) dominada por el método comunicativo. Este enfoque pedagógico, fundamentado sobre las bases del funcionalismo y que aboga, por tanto, por un claro énfasis en el significado y el intercambio de mensajes de forma comunicativa, supuso, al menos en su concepción teórica, un cierto rechazo de los aspectos formales del lenguaje (principalmente de la gramática y el vocabulario). Con esta situación, y a pesar de los enormes avances que se habían estado produciendo en el campo de la descripción lingüística gracias a los estudios de corpus desde los 60, no fue hasta finales de los años 80 y en gran medida a partir de los 90 cuando el componente léxico de las lenguas recobró un verdadero auge en el área de la lingüística aplicada, impulsado por el renovado interés que se estaba despertando en torno a los aspectos formales y a la necesidad de fomentar un equilibrio entre fluidez y precisión en el uso de la lengua (Long, 1991).

Fue pues en estos años cuando comenzaron a proliferar los trabajos dedicados al estudio del vocabulario, estando muchos de ellos enfocados particularmente al terreno de la enseñanza (Carter, 1987; Gairns y Redman, 1986; Carter y McCarthy, 1988; Sinclair y Renouf, 1988; McCarthy, 1990; Nation, 1990; Willis, 1990; Lewis, 1993). Entre todos ellos, son quizá especialmente destacables los llevados a cabo por Willis (1990) y Lewis (1993), dado que dieron un paso más allá fundando el denominado Enfoque Léxico (*Lexical Approach*). Lo que estos autores pretendían, en términos generales, era establecer una nueva corriente pedagógica en la que el sílabo de idiomas estuviese fundamentado y organizado en base al vocabulario. La consigna que predominaba para estos autores era: "Words carry more meaning than grammar, so words determine grammar" (Lewis, 1993: 38).

No cabe duda de que los estudios que por esa época se estaban realizando en el ámbito de la lingüística de corpus (particularmente en el seno del proyecto COBUILD), así como todo el bagaje acumulado, como decíamos, en este campo desde veinte años antes, ejercieron una importante influencia en esta nueva etapa de auge de los estudios sobre la enseñanza del léxico. Sin lugar a dudas, y como apunta Schmitt (2000), una de las más destacables aportaciones de los estudios de corpus computacionales a la enseñanza del vocabulario fueron los listados de frecuencias, siendo por tanto la frecuencia un aspecto que volvía a ocupar un lugar central en la enseñanza del vocabulario. En los siguientes apartados, llevaremos a cabo una revisión de los principales listados a los que han dado lugar los corpus de segunda generación y las nuevas técnicas informáticas, gracias a los cuales se están obteniendo índices mucho más precisos y fiables de la frecuencia de las palabras en el lenguaje natural.

2.3.2.2.1. Listados de frecuencias de palabras basados en corpus computacionales

Prácticamente todos los corpus de referencia han dado lugar a listados de frecuencias, no sólo en inglés sino también, y cada vez en mayor medida, en otras lenguas[12]. Teniendo en cuenta el creciente número de listados que se ha venido

[12] En castellano, por ejemplo, se han compilado recientemente varios listados de gran rigurosidad y fiabilidad. Cabe destacar el recopilado por Mark Davies (cuya versión completa está disponible previo pago) a partir de un corpus de 20 millones de palabras del español del s. XX y que contiene 20.000 lemas (una muestra del listado con una de cada diez palabras se ofrece en http://www.corpusdelespanol.org/files/lemmas1-20000.txt). Los primeros 5.000 lemas de este listado se utilizaron en la elaboración del *Frequency Dictionary of Spanish: Core Vocabulary for Learners* (2005). También es interesante el listado elaborado en el año 2008 por la Real Academia Española a partir del Corpus de Referencia del Español Actual (CREA), donde se recogen las 737.799 formas distintas que aparecen en el corpus ordenadas por

produciendo en los últimos años, debemos ser conscientes de que, debido principalmente a las diferencias existentes entre las fuentes de datos de las que se obtienen y a los distintos criterios de elaboración y metodologías de trabajo contempladas en cada proceso de compilación, estos listados son noblemente dispares en cuanto a sus contenidos.

Al trabajar con un listado, se debe tener en cuenta, en primer lugar, las características y la composición del corpus del que procede, puesto que ello explicará en gran medida el tipo de palabras que contenga la lista. Como señala Coxhead (2000, citado por López-Mezquita, 2005), sólo teniendo en cuenta que el BoE está compuesto por un importante número de textos extraídos de periódicos de principios de los 90, cuando Yemen y Lituania estaban de actualidad, podremos comprender que el diccionario Collins COBUILD (1995) incluyera palabras como *Yemeni* o *Lithuanian* en la banda 3 de frecuencia, es decir, la que se encuentra entre las 1.720 y 3.300 palabras más frecuentes del inglés.

En segundo lugar, debemos observar qué tipo de unidad se ha considerado una palabra. En este sentido, suelen darse tres posibilidades (Nation, 2001):

a) listas cuya unidad está compuesta por toda una **familia de palabras**, incluyendo pues todas las inflexiones y todas las formas derivadas de la familia en un solo puesto de la lista (por ejemplo la forma base "*desire*" incluye sus inflexiones "*desires*", "*desired*" y "*desiring*" y sus derivados "*desirable*", "*desirability*", "*undesired*", "*undesirable*" y "*undesirability*").
b) listas integradas por **lemas**, es decir, por la forma básica de un lexema que engloba al conjunto de las distintas formas léxicas que tienen la misma base y categoría gramatical, aunque tengan inflexiones diferentes (por ejemplo el

---

frecuencia y con datos de su frecuencia normalizada por millón (listado disponible en: http://corpus.rae.es/lfrecuencias.html).

lema "*go*" incluye en la misma entrada del listado las formas "*go*", "*goes*", "*went*", "*gone*" y "*going*").

c) listas cuyas unidades son **formas léxicas** (denominadas *types* en inglés) y en las que, continuando con el ejemplo anterior, las formas "*go*", "*goes*", "*went*", "*gone*" y "*going*" se consideran palabras diferentes y cada una ocupa su propio lugar en el listado.

Por último, también se debe considerar qué criterios se han empleado en la elaboración del listado, sobre todo cuando su compilación responde a fines pedagógicos. Como ya vimos, en muchos de los primeros listados de palabras la frecuencia era el primer criterio contemplado a la hora de seleccionar el vocabulario, pero no el único. Más recientemente y en la misma línea, Nation (1990) recomienda tener muy presentes las necesidades de los alumnos a la hora de recopilar o utilizar un listado pedagógico puesto que en algunas ocasiones podemos encontrar palabras de muy alta frecuencia que realmente no son necesarias para un alumno de nivel inicial. En este sentido, este autor menciona vocablos como *bill*, *labor*, *stock*, *thee* y *thou*, que se encuentran entre las mil primeras palabras de Thorndike y Lorge (1944) pero que evidentemente no parecen ser unidades léxicas que un alumno principiante necesite.

A tenor de todo lo anterior, resulta ciertamente importante seleccionar aquel listado que más se adecue a las necesidades concretas de cada trabajo. En la actualidad, además, y quizá debido a que el proceso de obtención de listados se ha mecanizado de forma evidente, cada vez parecen usarse menos otros criterios que no sean el de frecuencia. La mayor parte de los listados que se han generado en los últimos años a partir de corpus de referencia no suelen hacer uso de criterios pedagógicos, sin duda necesarios como se hizo constar anteriormente. Como se verá

más adelante, éste aspecto ha sido muy tenido en cuenta en la presente tesis a la hora de seleccionar el listado de palabras a partir del cual diseñamos nuestro propio listado de colocaciones, puesto que el objetivo de este trabajo es eminentemente pedagógico.

Pasaremos pues a continuación a tratar de los distintos listados de frecuencias de palabras generados a partir de los corpus computacionales de referencia más representativos del inglés[13].

a. Primeros listados computacionales

El primer listado de frecuencias obtenido mediante el uso de herramientas y técnicas informáticas fue *The American Heritage Word Frequency Book* (Carroll, Davies y Richman, 1971). Este listado fue compilado a partir de un corpus de 5.088.721 palabras procedentes de textos usados en Estados Unidos en distintos niveles escolares y materias. En este listado no sólo se ofrecía la frecuencia de cada palabra, sino su frecuencia relativa en cada uno de los niveles, de las materias y también por millón de palabras (Nation y Waring, 1997).

Algunos años más tarde, en 1982, se publicó otro listado, *Frequency Analysis of English Usage: Lexicon and Grammar,* una lista de 50.000 palabras basada en el *Brown Corpus*. Por su parte y de forma paralela, como cabría esperar, el LOB Corpus también dio lugar a su propio listado, publicado por Johansson y Hofland (1989) en *Frequency Analysis of English Vocabulary and Grammar*.

También en los 80 se publicó *Cambridge English Lexicon* (Hindmarsh, 1980), un listado de 4.470 palabras resultado de combinar el listado de Thorndike y Lorge

[13] No se llevará a cabo en este apartado una descripción exhaustiva de todos los listados de frecuencias que se han producido hasta la fecha, algo que ciertamente excede el propósito principal de este trabajo. Nos limitaremos a tratar las listas más representativas del inglés general que han sido asimismo elaboradas a partir de los corpus de referencia más reconocidos de esta lengua.

(1944), la GSL (West, 1953), *The American Heritage Word Frequency Book* (Carroll, Davies y Richman, 1971), un listado basado en el *Brown Corpus* (Francis y Kučera, 1964) y varias listas más (Nation, 1990). *Cambridge English Lexicon* estaba dividido en cinco bandas distintas donde las dos primeras contenían 600 palabras cada una aproximadamente, mientras que las tres últimas constaban de unas 1.000 palabras. Es también interesante destacar que este listado, basándose en la información semántica que ofrecía la GSL, diferenciaba entre los distintos significados de las palabras y los clasificaba también atendiendo a su respectiva frecuencia.

Pero fue sin duda en los 90 cuando se produjo un verdadero avance en el campo de la frecuencia léxica, gracias, como decíamos anteriormente, a la aparición de "mega-corpus" de millones de palabras. Son especialmente significativos en este sentido los listados producidos a partir del BoE, el BNC y el CIC. No es posible por el momento, sin embargo, acceder a los listados obtenidos mediante este último corpus, puesto que sólo están disponibles para los investigadores de Cambridge University Press.

b. Listado de frecuencias del *Bank of English*

El corpus *Bank of English* ha dado lugar a una de las listas de frecuencias de más alta fiabilidad de entre las existentes hoy en día. Este listado, elaborado dentro del proyecto Collins-COBUILD, ha hecho posible la inclusión de información relativa a la frecuencia de las palabras en el diccionario del mismo nombre (Sinclair, 2001). El listado, que cubre las 14.600 palabras más frecuentes del inglés, se distribuye en 5 bandas de frecuencias identificadas con rombos en el diccionario. Así, la banda número 5 recoge las primeras 680 palabras de la lista, las cuales aparecen marcadas en el diccionario con 5 rombos negros (◆◆◆◆◆) simbolizando su alta frecuencia. La banda 4 consta de 1.040 palabras y viene marcada por 4 rombos negros

(◆◆◆◆◇), la banda 3 se compone de 1.580 palabras (◆◆◆◇◇), la banda 2 contiene 3.200 (◆◆◇◇◇) y la banda 1 consta de 8.100 palabras (◆◇◇◇◇) (ibid.).

El listado con las primeras 1.720 palabras, las cuales componen las bandas 5 y 4, aparece en la introducción del diccionario Collins COBUILD (ibid.). Sin embargo, hace relativamente poco tiempo, hemos podido contar con el listado completo en su versión lematizada. Éste consta de 10.000 lemas, cada uno de los cuales aparece acompañado de su categoría gramatical —producto del etiquetado automático del BoE realizado mediante los programas *English Two-Level Morphological Analyser*, ENGTWOL, y *English Constraint Grammar Analyser*, ENGCG, capaces de aplicar 140 etiquetas morfológicas diferentes— y de su número total de ocurrencias en el corpus (calculadas según la versión de 450 millones de palabras del BoE, anterior a la última actualización de datos que se realizó en 2004). Debemos también señalar que, a la vista de los elementos incluidos en este listado, el único criterio seguido para su elaboración fue el de frecuencia, puesto que en él podemos encontrar nombres propios, días de la semana, meses, numerales cardinales y ordinales, letras, siglas, acrónimos, etc., elementos éstos que, como veremos, se han suprimido en otros listados atendiendo a criterios pedagógicos. En el siguiente ejemplo, mostramos las primeras quince palabras de este listado a modo de ilustración (Tabla 2.5.).

1. **the** DT 24773218
2. **be** V 19238890
3. **of** IN 11555597
4. **and** CC 10605027
5. **a** DT 9914455
6. **in** IN 8093754
7. **to** TO 7181480
8. **have** V 5826161
9. **to** IN 4031776
10. **for** IN 3972094
11. **i** PPS 3312765
12. **on** IN 3149028
13. **with** IN 2912040
14. **he** PPS 2851584
15. **that** CS 2661678

Tabla 2.5.: Primeros 15 lemas del listado del BoE

c. Listado de frecuencias del *British National Corpus*: Kilgarriff (1995)

Los primeros listados de frecuencias extraídos a partir del BNC fueron elaborados por Adam Kilgarriff (1997: 135), uno de los académicos que más han defendido el enorme valor que tiene la frecuencia en el aprendizaje de vocabulario:

> A central fact about a word is how common it is. The more common it is, the more important it is to know it. All else being equal, more common words should be taught to foreign learners first, both so that they understand them and so that they know how to, and are inclined to, use them.

Las dos listas de frecuencias elaboradas por Kilgarriff (1995), que se obtuvieron, evidentemente, a partir de la primera versión del corpus, están disponibles de forma gratuita en Internet[14]. Así pues, entre otro tipo de información, en esta página nos podemos descargar sus listados de frecuencias en forma tanto lematizada como no lematizada en diferentes formatos.

Comenzaremos por describir la lista no lematizada. Ésta recoge las 939.028 palabras diferentes que aparecen en el BNC según este recuento (aunque, como veremos más adelante, Leech, Rayson y Wilson, 2001, hallaron un total de 757.087 palabras distintas), entendiendo, pues, por "palabra" cada una de las formas léxicas, inflexiones y derivaciones de la lengua. Este listado se presenta en varios formatos: por un lado Kilgarriff ofrece una lista con todas las palabras y por otro también las recoge separando las palabras del corpus escrito y las del oral, haciendo a su vez una subdivisión en estas últimas entre las pertenecientes a las dos secciones del componente oral del BNC (ver apartado 2.3.1.3.2.). Asimismo, también se pueden descargar los listados ordenados tanto alfabéticamente como por frecuencia.

La lista no lematizada, además, fue etiquetada usando los códigos del programa CLAWS (*Constituent Likelihood Automatic Word-tagging System*), que adjudica de forma automatizada la categoría gramatical a cada palabra utilizando un rango de 134 etiquetas diferentes[15]. Así pues, la información que ofrece este listado, como podemos observar en la muestra que aparece en la tabla 2.6., es el número total de veces que la palabra aparece en el corpus, la propia forma léxica, la etiqueta de su categoría gramatical y el número de archivos en los que aparece del total de 4.124 que contiene el BNC (información ésta última que da una clara indicación del rango de la palabra, es decir, de su distribución entre los distintos textos del corpus).

---

[14] http://www.kilgarriff.co.uk/bnc-readme.html [Último acceso: 28.02.2009]

[15] La lista completa de códigos está disponible en http://www.kilgarriff.co.uk/BNClists/poscodes.html [Último acceso: 28.02.2009]

1. 6187267 **the** at0 4120
2. 2941444 **of** prf 4108
3. 2682863 **and** cjc 4120
4. 2126369 **a** at0 4113
5. 1812609 **in** prp 4109
6. 1620850 **to** to0 4115
7. 1089186 **it** pnp 4097
8. 998389 **is** vbz 4097
9. 923948 **was** vbd 4005
10. 917579 **to** prp 4099
11. 884599 **i** pnp 3746
12. 833360 **for** prp 4104
13. 695498 **you** pnp 3696
14. 681255 **he** pnp 3817
15. 662516 **be** vbi 4080

Tabla 2.6.: Primeras 15 palabras del listado no lematizado (Kilgarriff, 1995)

Especialmente interesante, sin embargo, es la lista lematizada que Kilgarriff elaboró a partir de la no lematizada, ya que se trata de la única con que contamos actualmente que esté ordenada por frecuencia a partir de datos del BNC. Aquella se creó siguiendo el procedimiento que se había empleado en la extracción de la lista lematizada utilizada en un proyecto de Longman llevado a cabo con el fin de poder incluir información sobre la frecuencia de las palabras en el *Longman Dictionary of Contemporary English*, 3ª edición (LDOCE, 1995). Se trata de un listado integrado por los 6.318 lemas que aparecen con una frecuencia superior a las 800 veces en el BNC. En la elaboración de este listado se excluyeron números, nombres propios, interjecciones que no forman una palabra (por ejemplo *ah*, *er*, *um*) y conjuntos "cerrados" de palabras como son los días de la semana, meses, unidades monetarias,

países, nacionalidades y religiones (los tres últimos en sus formas nominales y adjetivales). Dado que estos conjuntos de palabras funcionan sintácticamente como los nombres propios, Kilgarriff (1997: 142) consideró que, desde una perspectiva lexicográfica y pedagógica, no era adecuado incluirlos en su listado de frecuencias puesto que "they were insufficiently word-like". En opinión de este autor (ibid.: 143), "pedagogically, they are not the sort of item where frequency information is of any interest".

En cuanto a sus etiquetas gramaticales, en el listado lematizado no se contempló una variedad tan extensa de categorías como en el no lematizado, ya que Kilgarriff las redujo hasta dejarlas en once elementos. En la tabla 2.7., mostramos cuáles son estas once etiquetas utilizadas así como el número total de elementos de cada categoría que se recogen en el listado (López-Mezquita, 2005):

Etiquetas gramaticales	Número de palabras y porcentaje
1. **a** - adjetivo	1.124 palabras (17,8%)
2. **adv** - adverbio	427 palabras (6,75%)
3. **conj** - conjunción	34 palabras (0,53%)
4. **det** - determinante	47 palabras (0,74%)
5. **infinitive marker** - marcador de infinitivo	1 palabra (0,01%)
6. **interjection** - interjección	13 palabras (0,2%)
7. **modal** - verbo modal	12 palabras (0,18%)
8. **n** - nombre	71 palabras (1,12%)
9. **prep** - preposición	46 palabras (0,72%)
10. **pron** - pronombre	1.281 palabras (20,27%)
11. **v** - verbo	3.262 palabras (51,63%)
	Total: 6.318 palabras (100%)

Tabla 2.7.: Proporciones de categorías gramaticales en el listado lematizado (Kilgarriff, 1995)

En la tabla que aparece a continuación (Tabla 2.8.) se puede apreciar que todos los lemas se incluyen en una de estas once categorías gramaticales, y también se ofrece, al igual que sucedía en la lista no lematizada, el número total de ocasiones en que el lema aparece en el BNC. No obtenemos, sin embargo, información sobre el rango de cada lema, como ocurría en el listado deslematizado.

1. 6187267 **the** det
2. 4239632 **be** v
3. 3093444 **of** prep
4. 2687863 **and** conj
5. 2186369 **a** det
6. 1924315 **in** prep
7. 1620850 **to** infinitive-marker
8. 1375636 **have** v
9. 1090186 **it** pron
10. 1039323 **to** prep
11. 887877 **for** prep
12. 884599 **i** pron
13. 760399 **that** conj
14. 695498 **you** pron
15. 681255 **he** pron

Tabla 2.8.: Primeras 15 entradas del listado lematizado (Kilgarriff, 1995)

Si comparamos este listado con el no lematizado ofrecido en la tabla 2.6., podemos observar, por ejemplo, que cuando calculamos la frecuencia de forma lematizada el lema "*be*" adquiere un valor máximo, ocupando como vemos el segundo puesto de la lista, mientras que si tratamos el vocabulario de forma deslematizada, la primera forma del lema "*be*" no aparece hasta el puesto número 8 ("*is*"). Esta disparidad pone

de manifiesto nuestra apreciación anterior de que debemos ser muy conscientes de las características y criterios de elaboración de cada listado, puesto que dependiendo de los parámetros empleados los resultados serán más o menos diferentes.

d. Listado de frecuencias del *British National Corpus*: Leech, Rayson y Wilson (2001)

Al igual que hiciera Kilgarriff varios años antes, Leech, Rayson y Wilson (2001) elaboraron sus propios listados de frecuencias lematizados y no lematizados a partir del BNC, aunque en esta ocasión no sólo se utilizó la versión 1.0 del corpus sino también las mejoras que se habían introducido en los años 1995/96 en cuanto a etiquetado y categorización de textos a las que tenían acceso a pesar de que no estar aún publicadas. Estas listas están disponibles tanto en formato impreso (ibid.) como de forma gratuita en Internet[16].

Es interesante destacar con respecto a estos listados que, con el objetivo de trabajar con un número manejable de datos, los autores limitaron el número de palabras tanto en la versión impresa como en la digital. Así, en las listas que se ofrecen en Internet, solamente se recogen las palabras que aparecen al menos 10 veces por cada millón de palabras del corpus, mientras que en la versión impresa, y por razones obvias, muchos de los listados que presentan están limitados a las palabras que aparecen al menos 100 veces por millón, y el resto de ellos muestran los datos completos que aparecen en la página web, siendo pues el punto de corte 10 veces por millón. Como los propios autores indican (Leech, Rayson y Wilson, 2001), del total de 757.087 palabras diferentes que aparecen en el corpus según su recuento (como vemos, una cifra de casi 200.000 unidades menos que las que incluye Kilgarriff, 1995), 397.041 aparecen solamente en una ocasión, 98.774 muestran dos

[16] http://ucrel.lancs.ac.uk/bncfreq/flists.html [Último acceso: 01.03.2009]

casos en el corpus, 46.459 aparecen tres veces, 28.770 lo hacen cuatro veces y 62.041 palabras aparecen entre cinco y nueve veces. Así pues, sólo 124.002 palabras según sus cálculos aparecen al menos 10 veces en el conjunto del BNC. Dado que, como decimos, estas cantidades de palabras se consideraron demasiado elevadas para la escasa representación que suponen en el corpus, los listados que ofrecen contienen 7.726 palabras en el caso del listado no lematizado que aparece en Internet, de las cuales las primeras 1.052 (que aparecen al menos 100 veces por millón) conforman el listado impreso. En palabras de los autores (ibid.: 9), "[t]his book shows only the tip of the iceberg: to keep it within a manageable size, only headwords with an overall frequency of **10 per million words** or more are inlcuded in the lists".

En cuanto al etiquetado de la categoría gramatical de las palabras, el BNC fue etiquetado mediante el programa CLAWS al que aludimos más arriba, el cual procuraba una precisión del 96,5%. Con el fin de mejorar este margen de error para la extracción de sus listados, este proceso fue más tarde revisado mediante el uso de la herramienta *Template Tagger*, con el cual se alcanzó un grado de acierto del 98%, que da una idea de la exactitud y fiabilidad de sus resultados. Como vimos anteriormente, la versión más completa de los programas de etiquetado automático ofrecen un total de 143 categorías distintas que permiten clasificar las diferentes palabras de un corpus a un nivel muy profundo. En este estudio de frecuencias, sin embargo, se consideró que tal grado de precisión no era verdaderamente necesario, por lo que en este caso los autores decidieron reducir el número de etiquetas gramaticales hasta dejarlas en 23, que corresponden a las siguientes categorías (Tabla 2.9.):

Adj	adjective (e.g. *good*, *old*, *fine*, *early*, *regional*)
Adv	adverb (e.g. *now*, *well*, *suddenly*, *early*, *further*)
ClO	clause opener (e.g. *in order* [*that/to*], *so as* [*to*])
Conj	conjunction (e.g. *and*, *but*, *if*, *because*, *so that*)
Det	determiner (e.g. *a*, *an*, *every*, *no*, *the*)
Det/P	determiner/pronoun (e.g. *this*, *these*, *those*, *some*, *all*)
Ex	existential particle (*there* in *there is*, *there are*, etc.)
Fore	foreign word (e.g. *de*, *du*, *la*)
Form	formula (e.g. *2x + z*)
Gen	genitive (*'s, '*)
Inf	infinitive marker (*to*)
Int	interjection or discourse marker (e.g. *oh*, *aha, oops*, *yep*, *no*)
Lett	letter of the alphabet, treated as a word (e.g. *p*, *P*, *Q*, *r*, *z*)
Neg	negative marker (*not*, *~n't*)
NoC	common noun (e.g. *wealth*, *walls*, *child*, *times*, *mission*)
NoP	proper noun (e.g. *Malaysia*, *Paris*, *Susan*, *Roberts*, *Tuesday*)
NoP-	word which is normally part of a proper noun (e.g. *San* in *San Diego*)
Num	(cardinal) number (e.g. one, *four*, *forty*, *viii*, *8*, *55*, *1969*)
Ord	ordinal (e.g. *first*, *1st*, *9th*, *twenty-first*, *next*, *last*)
Prep	preposition (e.g. *of*, *in*, *without*, *up to*, *in charge of*)
Pron	pronoun (e.g. *I*, *you*, *she*, *him*, *theirs*, *none*, *something*)
Verb	verb – excluding modal auxiliaries (e.g. *tell*, *find*, *increase*, *realize*)
VMod	modal auxiliary verb (e.g. *can*, *will*, *would*, *could*, *may*, *must*, *should*)

Tabla 2.9.: Etiquetas gramaticales en la lista no lematizada (Leech et al., 2001: 13)

Leech, Rayson y Wilson presentan los datos relativos a la frecuencia de las palabras en diversos formatos, ofreciendo una información muy completa y variada. Así, estos autores crearon listados lematizados y no lematizados tanto del conjunto del corpus como de las distintas secciones que lo componen (observamos estudios

comparativos de la parte oral y escrita, así como de las subdivisiones que a su vez se establecen en cada una de ellas).

Lo primero que llama nuestra atención en este sentido, sin embargo, es que mientras que los listados no lematizados aparecen ordenados por frecuencia, los lematizados sólo parecen estar ordenados alfabéticamente, lo cual dificulta en buena medida la comparación de ambos tipos de listas. Sería deseable, en nuestra opinión, poder contar con un listado de lemas ordenado por frecuencia, no sólo con el fin de poder observar claramente las diferencias con respecto al listado no lematizado sino para poder utilizarlo en actividades de docencia e investigación con mayor facilidad[17]. En las tablas siguientes ofrecemos una muestra del listado no lematizado (ordenado por frecuencia) y el lematizado (ordenado alfabéticamente) correspondientes al BNC completo, donde se puede apreciar claramente la disparidad existente entre ellos (Tablas 2.10. y 2.11.).

1.	**the**	Det	61847
2.	**of**	Prep	29391
3.	**and**	Conj	26817
4.	**a**	Det	21626
5.	**in**	Prep	18214
6.	**to**	Inf	16284
7.	**it**	Pron	10875
8.	**is**	Verb	9982
9.	**to**	Prep	9343
10.	**was**	Verb	9236

Tabla 2.10.: Listado no lematizado ordenado por frecuencia (Leech et al., 2001: 120)

[17] López-Mezquita (2005: 350) ya pone de manifiesto esta carencia y ofrece un listado de los primeros 50 lemas ordenados por frecuencia, elaborado manualmente a partir del recuento alfabético.

Lema	Categoría gramatical	Formas del lema	Frecuencia	Rango	Dispersión
abandon	Verb		44	99	0.96
		abandon	12	98	0.94
		abandoned	26	97	0.96
		abandoning	5	90	0.93
		abandons	1	47	0.87
abbey	NoC		20	95	0.90
		abbey	19	95	0.90
		abbeys	1	34	0.75
Aberdeen	NoP		14	88	0.80
		Aberdeen	14	88	0.80
ability	NoC		105	100	0.94
		abilities	13	96	0.91
		ability	91	100	0.94
able	Adj		304	100	0.97
abolish	Verb		19	92	0.89
		abolish	6	85	0.88
		abolished	11	87	0.89
		abolishes	0	17	0.68
		abolishing	2	67	0.85
abolition	NoC		12	86	0.88
abortion	NoC		15	97	0.86
		abortion	12	94	0.86
		abortions	3	58	0.83
about	Adv		447	100	0.97
about	Prep		1524	100	0.96

Tabla 2.11.: Listado alfabético lematizado (Leech et al., 2001: 26)

Además de ilustrar las dificultades que estos dos tipos de listados confieren a la hora de establecer alguna comparación, las dos tablas anteriores nos permiten obtener más información acerca de ambas listas. En primer lugar, en lo que se refiere a la lista no lematizada, que en su versión más extensa cuenta con 7.726 palabras como dijimos, podemos comprobar que ésta ofrece la posición que cada forma léxica ocupa en la lengua según su frecuencia, la categoría gramatical de cada palabra y el número total de veces por millón que aparece en el BNC.

En cuanto a la lista lematizada, ofrecida únicamente por orden alfabético, observamos que ésta incluye, como ya hiciera la GSL, información sobre la frecuencia individual de cada una de las formas del lema, así como sobre su rango y su dispersión. El rango expresa el número de sectores del corpus (de un total de 100 donde cada sector incluye un millón de palabras aproximadamente) en los que aparece la palabra en cuestión, mientras que la dispersión viene expresada por el coeficiente D de Juilland, que determina en valores de 0.00 a 1.00 el grado de distribución de cada palabra con respecto al total del corpus, lo cual nos ayuda a observar estadísticamente si la frecuencia de la palabra se debe a que está incluida en muchos sectores del corpus o a que aparece muy repetidamente en un número determinado de textos.

A diferencia de lo que ocurría en los listados elaborados por Kilgarriff, tanto la lista lematizada como la no lematizada incluyen en este caso números y nombres propios. Podemos observar, por ejemplo, en la tabla 2.11., la inclusión de la palabra "*Aberdeen*", que ilustra claramente este aspecto. En nuestra opinión, sin embargo, y coincidiendo con las apreciaciones de Kilgarriff (1997), este tipo de palabras no resultan especialmente útiles en un listado de estas características desde un punto de vista pedagógico, puesto que su frecuencia no responde a las mismas causas ni debe tener por tanto las mismas repercusiones didácticas que en el resto de palabras de la lengua.

Por otro lado, estos autores presentan diferentes listados donde las distintas variedades recogidas en el corpus se pueden comparar. Así, se presentan listas de frecuencias comparando la parte oral y escrita del corpus, los dos sectores (informativo y de ficción) de la parte escrita, también los dos sectores (conversacional y enfocado a tareas) de la parte oral, e incluso listas organizadas según las diferentes categorías gramaticales de las palabras. Como decíamos, esta gran variedad de formatos permite establecer comparaciones entre los diversos subcorpus

que contiene el BNC, comparaciones que arrojan datos interesantes acerca del comportamiento de la lengua. A la hora de realizar estos estudios comparativos, y dado que, por un lado, los diferentes sectores del corpus no están representados de manera totalmente equilibrada y, por otro, es necesario saber si las diferencias entre ambos son estadísticamente significativas, Leech, Rayson y Wilson (ibid.) emplean la medida estadística conocida como coeficiente de verosimilitud (ver sección 3.2.2.2.). Con ello, estos autores comprueban si las diferencias que puedan existir entre los distintos listados se deben al azar o si, por el contrario, reflejan una tendencia real de la palabra a aparecer en unos sectores con más frecuencia que en otros. En la tabla que mostramos a continuación (Tabla 2.12.) se observa a modo de ilustración la comparación que estos investigadores establecen entre la parte oral y escrita del BNC.

Palabra	**Categoría gramatical**	**Frecuencia oral**	**Coef. de verosimilitud**	**Frecuencia escrito**
er	Uncl	8542	+390869.9	11
you	Pron	25957	+385328.3	4755
's	Verb	17677	+384464.6	1848
I	Pron	29448	+369238.5	6494
yeah	Int	7890	+356172.5	17
erm	Uncl	6029	+281015.6	2
that	DetP	14252	+213613.5	2581
n't	Neg	12212	+177089.3	2300
oh	Int	5052	+166592.5	179
it	Pron	24508	+151913.5	9298

Tabla 2.12.: Frecuencias de los subcorpus oral y escrito del BNC

Como se aprecia en la tabla anterior, estos listados están ordenados de acuerdo con índice del coeficiente de verosimilitud, por lo que las palabras de los primeros

puestos no son las más frecuentes, sino las que muestran una tendencia más marcada a aparecer en una variedad de la lengua frente a la otra. En este sentido es interesante observar, por ejemplo, cómo la expresión "*er*", primera en este listado comparativo, es marcadamente más frecuente en los textos orales, con una frecuencia de 8.542 apariciones por millón de palabras, que en los escritos, donde sólo aparece 11 veces por millón. El coeficiente de verosimilitud en este caso, pues, es notablemente elevado (+390869.9). Este tipo de listados es, sin duda, una de las aportaciones más relevantes de estos autores al estudio de la frecuencia léxica.

A tenor de la revisión llevada a cabo hasta este punto resulta fácil comprender que los distintos listados de frecuencias elaborados hasta la fecha son claramente dispares y presentan discrepancias en muy distintos sentidos. Es muy interesante por este y otros motivos el trabajo llevado a cabo por María Teresa López-Mezquita (2005) en el seno del proyecto ADELEX, ya que realizó una comparación y análisis enormemente exhaustivos de los distintos listados mencionados más arriba. Esta investigación concluyó finalmente con la elaboración de un listado propio donde se aúnan las aportaciones de los listados anteriores y donde el criterio pedagógico fue el elemento central en su diseño. No nos gustaría dejar de mencionar en este sentido que la labor llevada a cabo por esta autora fue galardonada con el **Primer Premio Nacional de Investigación Educativa en el año 2005**, un reconocimiento que sin duda viene a avalar el rigor de este listado, que pasamos a describir a continuación.

e. Listado de frecuencias de López-Mezquita (2005). Una solución ecléctica

En su tesis doctoral, López-Mezquita (2005) llevó a cabo una comparación enormemente exhaustiva entre los listados obtenidos de las siguientes fuentes: 1)

General Service List, 2) *Brown Corpus*, 3) *Cambridge International Corpus*, 4) *British National Corpus*, versiones de Kilgarriff y de Leech, Rayson y Wilson, 5) *Bank of English*, y 6) *Cambridge and Nottingham Corpus of Discourse in English*. López-Mezquita (2005) demostró mediante este análisis que, además de algunos puntos en común, estos listados presentan, por lo general, notables discrepancias entre sí. Esto es debido principalmente a las diferencias existentes entre los distintos corpus empleados como fuente de datos (en cuanto a tipo, tamaño, variedades del inglés que recogen, proporción de materiales que los componen, procedimientos de etiquetado utilizado, etc.) y a los distintos criterios y procedimientos de elaboración de los propios listados (tratamiento de los números, los nombres propios, los homógrafos, las abreviaturas, los acrónimos, etc.). Asimismo, se hallaron en todos ellos aspectos mejorables si se consideran desde un punto de vista pedagógico.

Debido fundamentalmente a la enorme representatividad y valía del BNC y el BoE como los dos grandes corpus de referencia del inglés actual a cuyos listados se tiene acceso, López-Mezquita centró su investigación de forma especial en sus tres listas. En primer lugar se llevó a cabo una comparación de los recuentos producidos a partir del BNC, tras la cual esta autora (2005: 354) concluyó que mientras que los de Leech, Rayson y Wilson "son más modernos y exactos que los listados de Kilgarriff", a la vez que suministran información de gran valor sobre las distintas secciones del BNC, el hecho de que sólo se ofrezcan las palabras que aparecen al menos 10 veces por millón, hace del listado de Kilgarriff un recurso más extenso y de mayor alcance que el de Leech et al. Asimismo, y dada la finalidad eminentemente pedagógica que se perseguía en la investigación de López-Mezquita, se consideró que un listado lematizado respondía en mayor medida a las necesidades del alumnado, puesto que se entiende que un estudiante que conoce el lema "*do*" también debe conocer las formas "*does*", "*did*", "*done*" y "*doing*". En consecuencia, tanto desde el punto de vista de la enseñanza como de la evaluación, no resultaría muy adecuado

tratar por separado las formas que componen un mismo paradigma. Así pues, entre los dos listados del BNC se optó por utilizar fundamentalmente el recuento lematizado de Kilgarriff.

El siguiente paso en este estudio fue la comparación entre dicho listado y el producido a partir del BoE. Como cabe suponer teniendo en cuenta las claras diferencias existentes entre ambos corpus, así como entre los criterios de elaboración de los listados, este análisis arrojó datos muy dispares. Al comparar ambas listas se observó que el nivel de coincidencia entre ellas era del 77,65%, es decir, había un 22,35% de lemas que se encontraban en uno de estos listados pero no en el otro. Tras un detallado análisis, López-Mezquita concluye que existen una serie de aspectos mejorables desde el punto de vista pedagógico tanto en el listado basado en el BNC como en el del BoE. En el caso del primero, el principal inconveniente que se observa es su tamaño, que no alcanza el umbral de 10.000 palabras deseable cuando se trabaja con alumnos de nivel avanzado (Hazenberg y Hulstijn, 1996). En cuanto al segundo, las principales limitaciones se encuentran, según esta autora, en su heterogeneidad y en la inclusión de términos claramente prescindibles en un listado con finalidad educativa como son los nombres propios o los términos pertenecientes a grupos cerrados de palabras. Así pues, una vez concluido este laborioso proceso analítico, López-Mezquita (2005: 364-365) concluyó que se hacía necesario "elaborar un nuevo listado que aunase los aspectos positivos de cada listado y prescindiese de los negativos".

El listado de frecuencias propuesto por esta autora se compone de 7.125 palabras. Se trata de un listado lematizado por lo que su contenido se centra en palabras base. Para la elaboración de esta lista, la autora estableció en primer lugar el listado de 6.318 lemas de Kilgarriff (1995) como punto de partida y lo comparó con la lista de los 10.000 primeros lemas recogidos a partir del BoE. Con el objetivo de homogeneizar ambos listados y atendiendo a criterios pedagógicos, se eliminaron

manualmente todas las categorías del listado del BoE que no estaban contempladas en el de Kilgarriff (nombres propios, días de la semana, meses, numerales cardinales y ordinales, adjetivos de nacionalidad y de religiones, adverbios en grado comparativo, formas verbales negativas contraídas, letras, siglas, acrónimos, prefijos e interjecciones), dando como resultado la supresión de 1.348 palabras. Al volver a comparar ambos listados una vez realizada esta supresión, se comprobó que el grado de coincidencia entre ellos había aumentado hasta prácticamente un 86%, existiendo 733 palabras en el del BNC que no aparecían en el nuevo listado del BoE y 727 elementos que sí estaban en el listado del BoE pero no en el del BNC. Sin embargo, al comparar el listado de Kilgarriff con los primeros 6.318 lemas del nuevo listado obtenido del BoE se observó que sólo un elemento de cada listado no aparecía en el otro ("*cent*" no se encontraba en el del BNC mientras que "*including*" no aparecía en el del BoE), por lo que su coincidencia era máxima.

Con estos datos, López-Mezquita procedió a la elaboración de una nueva lista donde se incluyeron las 6.318 palabras del listado de Kilgarriff más las 727 palabras procedentes del listado del BoE que no estaban incluidas en aquél. Finalmente, también se contrastó este listado resultante con la lista de las 2.000 palabras más frecuentes recogidas en el *Longman Dictionary of Contemporary English* extraídas a partir del *Longman Corpus Network*. Esta comparación dio lugar a la inclusión de 80 palabras que no se encontraban presentes en el listado obtenido hasta entonces. El producto final fue una lista que contenía las 7.125 palabras más frecuentes del inglés (el Anexo I recoge las primeras 1.000 palabras de este listado ya que son las que han servido de base para la confección de nuestro test de colocaciones) y que, como la propia autora expone, se elaboró

> utilizando las cuatro fuentes más solventes y actualizadas que se encuentran disponibles hoy en día —*British National Corpus* (listados de Kilgarriff y de Leech *et al.*), *Bank of English* y *Longman Corpus Network*— siguiendo un riguroso procedimiento manual que ha estudiado cuidadosamente todos los detalles con

> objeto de conseguir un resultado final que constituyese un registro completo y fiable de las primeras palabras más frecuentes de la lengua. (López-Mezquita, 2005: 372)

A la luz de los resultados obtenidos en su investigación y de la evidente mejora que su listado supone en el área de los estudios de frecuencias con fines pedagógicos, este listado fue el que utilizamos como base para la compilación de nuestra lista de colocaciones. Más concretamente, la lista de López-Mezquita supuso la fuente de referencia de la que se extrajeron los sustantivos que conforman las bases de las colocaciones que integran nuestro test (ver apartado 3.2.1.).

2.3.2.2.2. La frecuencia y los umbrales de competencia léxica

A partir de los años 90 y gracias, en buena medida, a los nuevos listados de frecuencia léxica que se estaban produciendo desde el terreno de los estudios de corpus, surgió un considerable interés por conocer cuántas palabras debe poseer un alumno para poder comunicarse en la segunda lengua. Como vimos, ya desde los años 30 y gracias, entre otros estudios, a la ley de Zipf, los investigadores eran conscientes de que un número relativamente reducido de palabras cubren un porcentaje muy alto del lenguaje que utilizamos en nuestro día a día debido a su enorme frecuencia (West, 1937). En consecuencia, será muy rentable para el alumno aprender estas palabras antes que las de menor frecuencia puesto que ello repercutirá de forma notable en su competencia comunicativa. Como afirma Nation (2001: 16), "high-frequency words are so important that anything that teachers and learners can do to make sure they are learned is worth doing". Partiendo de este consenso en torno a la importancia de la frecuencia como factor decisivo a la hora de estimar qué vocabulario debe aprender un alumno, la siguiente pregunta que se planteaba era cuántas palabras debe conocer este alumno para poder manejarse con soltura en la

lengua extranjera o, dicho de otro modo, qué porcentaje de cobertura suponen las palabras más frecuentes de la lengua. Los estudios de corpus, una vez más, supusieron una fuente inestimable de información para poder dar respuesta a esta cuestión.

Según datos del *Brown Corpus*, Nation y Waring (1997) ofrecen las siguientes cifras, referidas a la cobertura de textos escritos y donde según los propios autores especifican, la última cifra fue tomada de Kučera (1982, citado por Nation y Waring, 1997) (Tabla 2.13.).

Número de lemas	**Cobertura textual**
1.000	72,0%
2.000	79,7%
3.000	84,0%
4.000	86,8%
5.000	88,7%
6.000	89,9%
15.852	97,8%

Tabla 2.13.: Cobertura textual por número de lemas (Nation y Waring, 1997)

Como se puede observar en este listado, las primeras 2.000 palabras de la lengua ofrecen una cobertura de prácticamente el 80% de los textos, una apreciación que coincide, como veremos, con otros estudios similares y que hoy en día constituye la estimación generalizada.

En la misma línea, Cobb (2003) ofrece los datos incluidos en la tabla 2.14., donde también se han tomado como principal referencia las frecuencias extraídas del *Brown Corpus* (Francis y Kučera, 1982) pero donde se ha añadido información complementaria a partir del listado de Carroll, Davies y Richman (1971), citado por Nation (2001: 15).

Número de lemas	Cobertura textual
10	23,7%
1.000	72,0%
2.000	79,7%
3.000	84,0%
4.000	86,7%
5.000	88,6%
6.000	89,9%
43.831	99,0%
86.741	100%

Tabla 2.14.: Cobertura textual por número de lemas (Cobb, 2003)

En el diccionario Collins-COBUILD, por otro lado, se estima que las primeras 15.000 palabras del BoE ofrecen una cobertura del 95% de un texto de nivel intermedio, mientras que Nation (1990) considera que las 2.000 palabras más frecuentes de la lengua cubren hasta un 87% de los textos, y algunos años después (Nation, 2001: 14) declara que "the 2,000-word level has been set as the most suitable limit for high-frequency words". Como ya mencionamos, podemos decir que en términos generales éste es el umbral que se suele establecer como requisito mínimo que todo alumno debe alcanzar si desea poder utilizar la lengua de forma eficaz.

Sin embargo, en lo que respecta a la competencia lectora, algunos estudios han sugerido que conocer el 80% de las palabras de un texto no es suficiente para poder comprenderlo. Estos trabajos consideran que el porcentaje debe aumentar hasta el 95% si se quiere lograr una comprensión general, e incluso hasta el 98% si lo que se pretende es llegar a una lectura placentera (Laufer, 1989, 1992; Hirsh y Nation, 1992). Para lograr el 95% mínimo imprescindible para comprender un texto no

simplificado, Laufer (1992) opina que el umbral mínimo que el alumno debe alcanzar es de 3.000 familias de palabras, lo que equivale a 5.000 palabras o formas léxicas diferentes. Por otro lado, en lo que se refiere a niveles avanzados, Groot (2000) estima que para comprender un texto académico o especializado son necesarias 7.000 palabras, mientras que Hazenberg y Hulstijn (1996) sitúan este nivel en las 10.000.

Partiendo de trabajos como los que acabamos de mencionar y mostrando el enorme potencial de estos estudios y de la propia noción de frecuencia en términos prácticos, Laufer y Nation (1995) desarrollaron lo que se conoce como Perfil de Frecuencia Léxica (PFL) (*Lexical Frequency Profile*). El PFL se puede definir como un procedimiento que distribuye las palabras de un texto según el nivel de frecuencia al que pertenezcan y donde, generalmente, cada nivel corresponde a una banda de 1.000 palabras. Como los propios autores explican, esta medida se lleva a cabo mediante un programa informático, el cual realiza un cálculo que compara distintas listas de frecuencias de 1.000 palabras cada una con el texto particular que se desee analizar. El programa comprobará qué porcentaje de las palabras del texto se encuentran en los distintos listados, y cuáles están fuera de los niveles de dificultad que contemplan las listas.

El PFL fue originado en un principio con el fin de analizar los textos producidos por los alumnos, es decir, como una herramienta de análisis de la escritura y de la riqueza léxica de los estudiantes. Sin embargo, muy pronto comenzó a aplicarse para analizar los textos que los alumnos han de leer, función ésta en la que mayor aplicación se le ha dado y que ha propiciado de hecho la aparición de varios programas informáticos capaces de aplicar este cálculo[18]. Quizá el más conocido de

18 Además de los programas *VocabProfile* y ADA que se citan en este apartado, otras herramientas, quizá de menor alcance en la actualidad, son *Range* y *Frequency* (http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx), *Frequency Level Checker* (http://language.tiu.ac.jp/flc/index.html#WHAT1) y *Word Frequency Text Profiler* (http://www.edict.com.hk/textanalyser/). En Moreno Jaén (2006) se puede ver una detallada descripción de cada uno de estos programas.

entre estos analizadores léxicos sea *VocabProfile*[19], diseñado por Tom Cobb e incluido como una de las herramientas que se incluyen en su programa *Compleat Lexical Tutor* [20]. Hasta muy recientemente, este programa ofrecía un perfil basado fundamentalmente en las frecuencias de la GSL de West (1953), lo cual suponía que sus bases de datos estaban algo anticuadas y la información que ofrecía era, por tanto, poco útil en cierta medida. Sin embargo, en octubre de 2008 esta herramienta se actualizó mediante la inclusión de 20 listados de frecuencias (de 1.000 palabras cada uno) extraídos a partir del BNC, con lo cual se ha convertido en uno de los analizadores léxicos más potentes y fiables que existen actualmente (Fig. 2.1.).

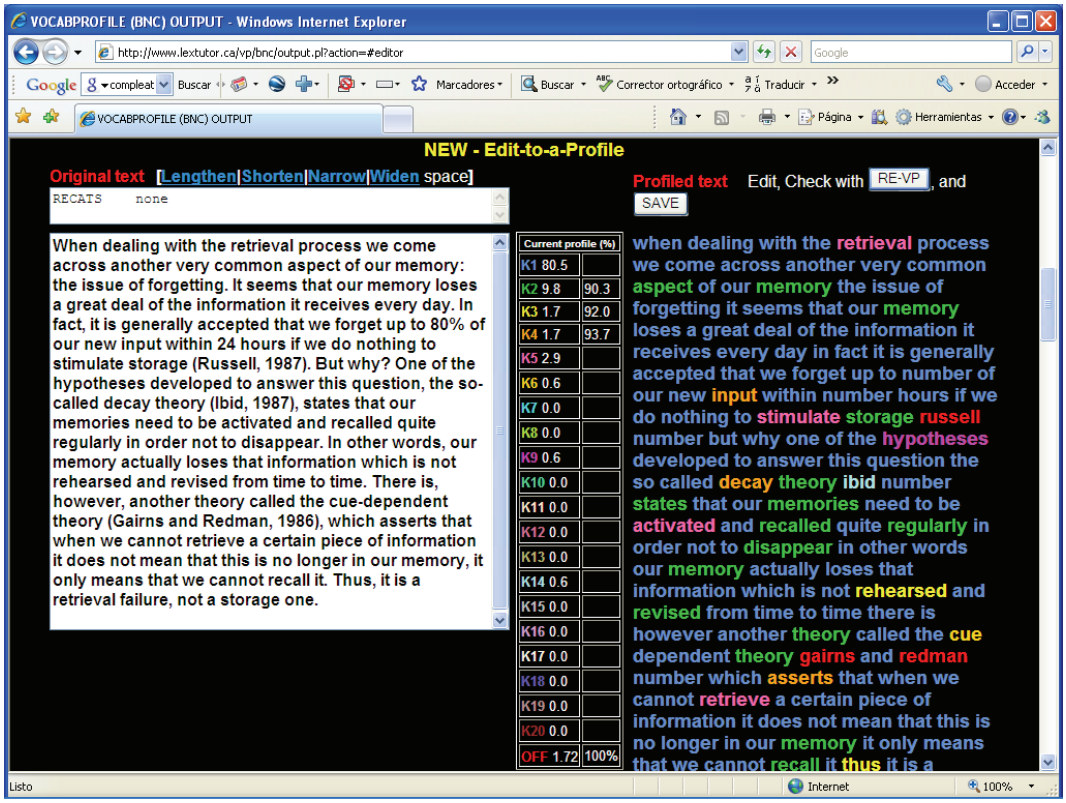


Fig. 2.1.: Texto analizado con *VocabProfile*

19 http://www.lextutor.ca/vp/ [Último acceso: 02.03.2009]
20 http://www.lextutor.ca/ [Último acceso: 02.03.2009]

No podemos dejar de citar en este sentido otra herramienta diseñada con los mismos objetivos dentro del proyecto ADELEX. Se trata del programa ADELEX ANALYSER (ADA)[21], y su principal contribución en el campo del análisis léxico de textos es que su base de datos está conformada por siete bandas de frecuencias extraídas a partir del listado de López-Mezquita (2005), un recuento que, como ya hemos mencionado, se caracteriza por su fiabilidad y por su orientación pedagógica.

Para trabajar con esta herramienta debemos introducir el texto que deseamos analizar (tal como aparece en la figura 2.2.), y tras compararlo con sus siete listados de frecuencias, el programa nos ofrece el número y porcentaje de palabras de dicho texto que corresponden a cada banda de frecuencia (Fig. 2.3.). Así pues, aquellas palabras que se encuentren en la primera banda del programa, es decir, entre las 1.000 palabras más frecuentes del inglés, aparecerán clasificadas en la sección "*Level 1*" y se colorearán en rojo en el texto de la parte superior, las que pertenezca al segundo nivel, es decir, entre las palabras 1.001 y 2.000, aparecerán en la sección "*Level 2*" y en azul, y así sucesivamente. De este modo, si, como ya dijimos, consideramos que un alumno debe conocer el 95% de las palabras de un texto para lograr una adecuada comprensión, habremos de observar qué niveles o bandas de frecuencia contienen las palabras necesarias para alcanzar dicho porcentaje. Si, como se observa en nuestro ejemplo (Fig. 2.3.), es necesario llegar a la quinta banda para lograr una cobertura aproximada del 95%, podemos considerar que se trata de un texto demasiado complejo para un nivel que no sea avanzado ya que exige del alumno conocer palabras de baja frecuencia.

[21] http://www.ugr.es/~inped/ada/ [Último acceso: 02.03.2009]

Fig. 2.2.: Herramienta *Lexical Profile* de ADA

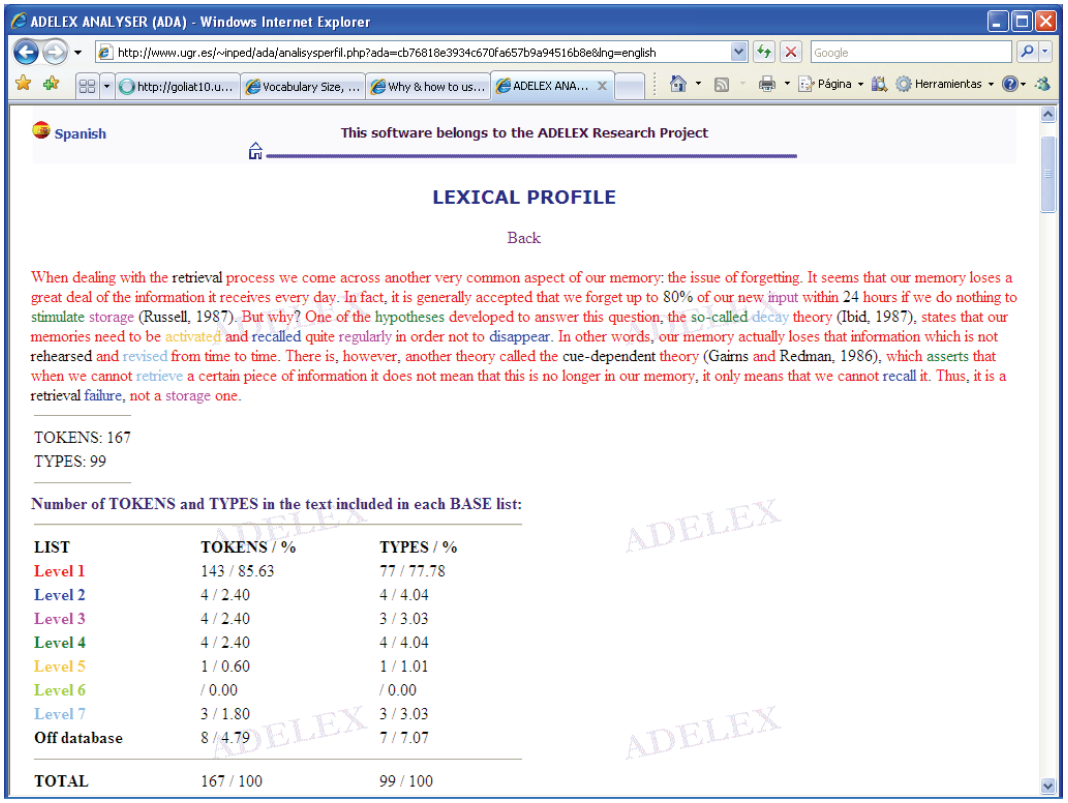


Fig. 2.3.: Análisis léxico con ADA

Como podemos comprobar, pues, este tipo de herramientas son de enorme utilidad tanto para el profesor como para el diseñador de materiales pedagógicos, puesto que ofrecen una información muy clara y rápida acerca de si un texto es adecuado para un cierto nivel de competencia lingüística, o si por el contrario, su nivel de dificultad es inadecuado, por ejemplo por ser demasiado elevado al contener palabras poco frecuentes.

Llegados a este punto y teniendo en cuenta que el principal tema de atención en el presente trabajo son las colocaciones, cabe preguntarse qué tratamiento han recibido éstas en el campo de los estudios de frecuencias. Así pues, en el siguiente

apartado llevaremos a cabo una revisión de este aspecto, de fundamental importancia para nuestro propósito.

2.3.2.2.3. Listados de frecuencias de colocaciones

Ya mencionamos que los estudios de corpus llevados a cabo desde que arrancara la era computacional tuvieron un enorme impacto en el análisis y desarrollo de las cuestiones léxicas. Uno de los aspectos en los que la lingüística de corpus ha tenido una mayor repercusión en las últimas décadas es sin duda en el estudio de la fraseología, que a partir de los 90 pasó a convertirse en un aspecto central de los estudios de vocabulario y de su enseñanza. Nunca antes se había podido observar con tanta claridad que la lengua no se articula en base a unidades léxicas individuales, sino que son las unidades multiléxicas prefabricadas las que la conforman en mayor medida. En palabras de Nesselhauf y Tschichold (2002: 252), "[w]e know today more than ever that multi-word expressions such as compounds, phrasal verbs, formulae, and collocations are an important part of a native speaker's knowledge of language and that they are possibly even more numerous than simple lexemes". Con la ayuda de programas de concordancias capaces de mostrar cualquier palabra dentro de sus diferentes cotextos (ver sección 3.2.2.3.), podemos ahora observar de forma muy gráfica hasta qué punto los patrones léxicos son recurrentes en la lengua y determinan el significado y en general el uso lingüístico.

Más aún, algunos autores han puesto de manifiesto que no todas las unidades fraseológicas comparten la misma frecuencia, sino que las colocaciones destacan sobre otros tipos de combinaciones léxicas (Farghal y Obiedat, 1995; Howarth, 1996; Moon, 1998b). Cabe mencionar a este respecto la opinion de Mel'čuk (1998: 24),

para quien "a phraseme[22] (...) is the numerically predominant lexical unit: in any language —i.e. in its lexicon— phrasemes outnumber words roughly ten to one. Collocations make up the lion's share of the phraseme inventory, and thus deserve our special attention". Al interpretar las palabras de Mel'čuk se debe tener en cuenta, por supuesto, que su definición de colocación es más amplia que la de otros estudios fraseológicos (incluyendo el nuestro). Sin embargo, su opinión ha sido también corroborada por autores cuya aproximación a este fenómeno se corresponde plenamente con la nuestra. Así, Lewis (2000a: 8) afirma:

> The single most important task facing language learners is acquiring a sufficiently large vocabulary. We now recognise that much of our 'vocabulary' consists of prefabricated chunks of different kinds. The single most important kind of chunk is collocation. Self-evidently, then, teaching collocation should be a top priority in every language course.

Desde la perspectiva del profesor, teniendo pues en cuenta que el léxico está integrado no sólo por palabras o lemas sino también por unidades fraseológicas, y de forma muy significativa por colocaciones, parece evidente que éstas también requieren de una especial atención en cuanto a su frecuencia en la lengua, con el fin de poder llevar a cabo una cuidada selección y un tratamiento pedagógico adecuado. Refiriéndose particularmente a las colocaciones, Nation (2001: 325) afirma a este respecto:

> [J]ust because a collocation exists does not mean that it deserves attention from a teacher. In order to decide if classroom time and effort should be spent on an item the criteria of frequency and range need to be considered. If the frequency of a collocation is high and it occurs in many different uses of the language, it deserves attention.

[22] Éste es uno de los términos empleados por Mel'čuk para referirse a las unidades fraseológicas.

La noción de que es importante conocer la frecuencia, y también el rango, de las colocaciones para poder trabajar de forma adecuada con ellas ha sido compartida por diferentes autores del área de los estudios léxicos (Carter, 1987; Hill, 2000; Nesselhauf, 2005). Quizá debido a la existencia de este claro consenso en torno a la importancia de la frecuencia de las colocaciones, se hace muy paradójico comprobar la escasa atención que se le ha otorgado a este aspecto en términos prácticos. Como veremos a continuación, es verdaderamente sorprendente el escaso número de estudios que han tratado de establecer cuáles son las colocaciones más frecuentes del inglés.

a. Palmer (1933)

Al hablar del movimiento de control del vocabulario, ya mencionamos que la primera recopilación de colocaciones de la que se tiene constancia es la que publicara Palmer en 1933 en el *Second Interim Report on English Collocations* producido en el IRET (Japón). Debido a las escasas copias que hoy en día se conservan de esta publicación ha sido imposible para la autora de este trabajo poder consultar la obra original. Sin embargo, según indica Kennedy (2008) y como ya adelantamos, se trata de un listado donde se recogen varios miles de unidades multiléxicas del inglés y donde la única fuente de información que Palmer y Hornby utilizaran para su elaboración fue su propia intuición, teniendo en cuenta las dificultades que los alumnos mostraban. En este sentido, pues, se trata de un listado de un gran valor desde el punto de vista pedagógico (Howarth, 1998b).

Una de las interesantes aportaciones que este listado supuso en el estudio de las colocaciones es la meticulosa clasificación gramatical de las unidades que incluye. Así, en su lista, Palmer (1933, citado por Cowie, 1998: 211) diferencia entre lo que él mismo denomina "Verb-collocations, Noun-collocations, Adverb-collocations,

Prepositions-collocations". Resulta evidente que en la época en que Palmer (con la ayuda de Hornby) llevara a cabo este trabajo, no existía la noción de que las unidades fraseológicas se mueven a lo largo de una escala donde existen distintos grados de fijación sintáctica y de opacidad semántica. Así pues, y como apunta Cowie (ibid.: 212), se incluyen en su listado un buen número de unidades que difícilmente se considerarían colocaciones en la actualidad sino que estarían más cercanas a la coligación como son las estructuras del tipo "prep + *one's* + nombre" como en "*at one's ease*", "*at one's feet*", "*at one's leisure*", "*at one's post*" y "*at one's time of life*". A pesar de que, como vemos, se incluyen unidades multiléxicas de diferente naturaleza, Palmer era plenamente consciente de que este tipo de unidades son distintas a las locuciones tradicionalmente recogidas en los diccionarios (como puedan ser "*skate on thin ice*" o "*buy a pig in a poke*") y por las que, según afirma Cowie (ibid.), Palmer sentía casi aversión.

Este listado de colocaciones, pionero en muchos sentidos en el campo de los estudios fraseológicos, y de especial relevancia dado que fue compilado con una clara orientación pedagógica, sigue siendo considerado por muchos en la actualidad como el único banco de colocaciones que existe para la selección de contenidos didácticos: "Lists of individual collocations which are worth learning do not seem to have been compiled since the *Second Interim Report on English Collocations*" (Nesselhauf, 2005: 255). Esta apreciación, aunque no del todo exacta, deja muy patente hasta qué punto nos encontramos en un terreno todavía inexplorado y donde es necesaria una mayor investigación.

Acabamos de mencionar que no es totalmente correcto decir que no han existido más listados de colocaciones, si bien es cierto que la gran mayoría de ellos son compilaciones lexicográficas y ninguno de ellos ofrece las colocaciones ordenadas por frecuencia. Dado que el objeto de análisis en este apartado son los estudios de frecuencia colocacional, no nos detendremos pues en todos los

diccionarios de colocaciones que se han producido hasta la fecha, sino sólo en aquellos en los que la frecuencia ha sido un factor de diseño y elaboración decisivo, a pesar de que evidentemente no se trate de listados de frecuencias como tal.

b. Perebeinos et al. (1986)

Según nuestros datos, el primer diccionario de colocaciones que hizo uso de la frecuencia como criterio de selección de contenidos fue *A Deskbook of Most Frequent English Collocations*, compilado en Rusia en el año 1986 por Perebeinos et al. Al citar este trabajo, Sinclair (2008: xvi) dice "if memory serves me it was assembled by hand by a group of language teachers who went through a great deal of text, noting frequencies, which are precisely recorded". Este interesante trabajo, sin embargo, no tuvo repercusión alguna en el área de los estudios lexicográficos ni colocacionales y, según afirma Nuccorini (2003: 366), "this dictionary is unknown to most and practically impossible to find".

Como esta autora explica (ibid.), se trata de un diccionario que recoge 426 entradas (187 sustantivos, 183 verbos y 56 adjetivos) a partir de las cuales se incluyen 8.000 colocaciones. Éstas fueron obtenidas mediante la exploración manual de un corpus de textos literarios, científicos y socio-políticos, en la que se fue anotando la frecuencia de cada una de las colocaciones, información que fue incluida en el diccionario con el objetivo de ayudar a los profesores y autores de materiales a seleccionar aquellas más útiles. Cada una de las entradas nos informa sobre la frecuencia de la palabra base así como de la frecuencia de las colocaciones de cada estructura gramatical. Así mismo, estos autores compararon la frecuencia de sus colocaciones con los datos de la GSL y concluyeron que su listado contenía 1.500 colocaciones de muy alta frecuencia que eran, por tanto, prioritarias en la enseñanza. En la figura 2.4. podemos observar una entrada de este diccionario.

ADD [æd] v 2028 40, 20, 10, 5

VN 793: a. smb 14, a. smth 779. Acid is added 16; a. nothing 10; a. a suffix 12; a. water 15; a. a word 19. I added water to it. Fruit acid may also be added if required. Now let me add a few things.

VS 1198 A. as... 24; a. that... 126; a. ø ... 1022. He added that it was a change for me anyway. Then, turning to the family he added, "Excuse me."

VprpN 140

VtoN 110: a. to smth 110. A. to it 12. Even this added to his anger. I had to add to the story.

VNV- 67 Add new words to complete the phrase.

VNprpN 463

VNatN 22: a. smth at smth 22. At my suggestion they also added some sugar. Nothing was added to this at the moment.

VNinN 26: a. smth in smth 26. He had added a postscript in his own writing.

VNtoN 355: a. smb to smth 5, a. smth to smth 350. He added some wood to the fire. I'd like to add a word of my own to this report. A number of coloured pictures added interest to the text.

Fig. 2.4.: Entrada de "*add*" en *A Deskbook of Most Frequent English Collocations* (adaptada de Nuccorini, 2003: 384)

Como vemos, la palabra "*add*" aparece 2.028 veces en su corpus, de las cuales 793 veces aparece en la combinación V+N, en 1.198 ocasiones el verbo aparece seguido de oración (V+S), 140 colocaciones son V+prep+N, "add" aparece seguido de sustantivo+verbo en 67 ocasiones y, por último, la estructura V+N+prep+N se da en 463 casos del corpus. También podemos apreciar en el ejemplo anterior que esta frecuencia aparece desglosada por cada una de las subestructuras de la entrada, lo cual indica el grado de precisión con el que se llevó a cabo el análisis y recuento de frecuencias.

Como podemos apreciar, la enorme valía de este listado no sólo radica en la información que ofrece sobre la frecuencia de las colocaciones y en que éste fue el principal criterio de selección de colocados, sino también, y especialmente, en que la frecuencia se concebía como un valor pedagógico elemental fundamentado en el convencimiento de que "more frequent combinations are to be taught first" (Nuccorini, 2003: 384).

c. Dzierżanowska y Kozłowska (1982 y 1991/1997)

A diferencia de los listados mencionados hasta ahora, otras compilaciones de colocaciones frecuentes que se han producido en las últimas décadas con fines lexicográficos están basadas en estudios de corpus. El primer estudio de estas características que conocemos fue el que llevaran a cabo Dzierżanowska y Kozłowska a finales de los 70 y que culminaría en 1982 con la publicación del diccionario *Selected English Collocations*, basado en colocaciones frecuentes de sustantivos. Durante los años 80 estos autores polacos continuaron su investigación y en 1991 apareció su segundo diccionario, *English Adverbial Collocations*, en el cual se recogían colocaciones frecuentes con la estructura verbo-adverbio y adjetivo-adverbio. Según indican Hill y Lewis (1997) en su reedición y ampliación de ambos trabajos publicada con el título *LTP Dictionary of Selected Collocations*, los dos diccionarios producidos por Dzierżanowska y Kozłowska fueron elaborados a partir de un corpus escrito de inglés británico compilado con material producido en los años 60. Se trataba, no obstante, de un corpus no computerizado, compuesto en su mayoría por obras literarias, prensa británica y publicaciones académicas y de referencia. Sin embargo, a pesar de que la frecuencia fue un aspecto importante en la elaboración de este diccionario, el hecho de que el corpus utilizado no estuviese almacenado en formato digital explica por qué no se incluyeron datos sobre la

frecuencia de las colocaciones, algo que Dzierżanowska y Kozłowska ya reconocían importante y que Nuccorini (2003) estima habría sido deseable.

La obra reeditada por Hill y Lewis varios años después, que contiene 50.000 colocaciones de 2.000 nombres frecuentes y 1.200 verbos y adjetivos de 5.000 adverbios, tiene un marcado carácter pedagógico, siendo la frecuencia un criterio importante en este trabajo pero no definitivo. Así, Hill y Lewis (1997: 7) explican: "The most frequent collocations are almost all made with some of the most common words in English (…). The lists in this dictionary help you to find combinations like these". Sin embargo, más abajo añaden: "The most common words are not included and neither are the most common collocations made with common words —*a fast car*, *have dinner*, *a bit tired*. If you are using this dictionary, we are confident you will already know these". Resulta evidente, sobre todo a la vista de que se eliminó una colocación tan marcada como "*fast car*" y pudiendo comprobar además que palabras tan básicas como "*man*", "*woman*", "*boy*" o "*girl*" no se encuentran en este listado, que esta compilación se llevó a cabo siguiendo criterios subjetivos en buena medida, lo cual puede quizá conducir a un resultado final de dudosa fiabilidad en algunos casos.

d. Kjellmer (1994)

Cabe destacar asimismo la extracción de combinaciones multiléxicas llevada a cabo por Kjellmer (1994) a partir del *Brown Corpus* para la producción de *A Dictionary of English Collocations: Based on the Brown Corpus*. Esta obra, dividida en tres volúmenes, recoge todas las combinaciones del mencionado corpus, sumando un total de 85.000. Uno de los rasgos más interesantes de este diccionario, cuya finalidad es más académica que pedagógica, es que cada una de sus entradas viene acompañada de datos estadísticos que muestran el número de combinaciones en los que interviene la palabra base, la frecuencia de cada una de las colocaciones y del total de colocaciones

de dicha base en el corpus y su distribución en las distintas secciones del corpus. En la figura 2.5. se muestra una de las entradas del diccionario a modo de ilustración.

Fig. 2.5.: Entrada de la palabra "*bend*" en el *Dictionary of English Collocations*

	EF	IF	RF	TC	DI
BEND CTy 8; CF 20; CTe 83					
BEND DOWN d	2	2		2	2
BEND IN b	2	2		N	0
A *BEND* a	3	3		2	2
CATFISH *BEND* a	2	2		C	1
THE *BEND* a	2	4		N	0
THE *BEND* OF ab	2	2		2	
TO *BEND* f	4	7		3	
HAD TO *BEND* edf	3	3		K	

CTy: Número de colocaciones distintas en las que aparece la palabra
CF: Frecuencia total de todas las colocaciones de la palabra en el corpus
CTe: Tendencia colocacional, es decir, el porcentaje de veces en que forma una colocación en el corpus
a-u: Tipo estructural al que pertenece la colocación[23]
EF: Frecuencia exclusiva, es decir, la frecuencia de la colocación repetida exactamente palabra por palabra, sin formar parte a su vez de colocaciones más extensas.
IF: Frecuencia inclusiva, es decir, la frecuencia de la colocación sumando los casos en que aparece sola y en los que forma parte de otra colocación más extensa.
F: Frecuencia relativa, es decir, la relación entre la frecuencia observada y la esperada (ver sección 3.2.2.2.). "The collocations with a relative frequency of 40.576 or more are marked *, while those with a relative frequency of less than 40.576 are unmarked" (ibid.: xxxviii).
TC: Categorías textuales del corpus en las que aparece. Cuando se incluye una letra, ésta indica la categoría concreta del texto (ver tabla 2.2.) y cuando aparece un número sabemos que la colocación se encuentra en varias categorías distintas.
DI: Índice distintivo en el corpus, relacionado con el grado de prominencia de algunas colocaciones sobre otras.

[23] Kjellmer (1994: xxii-xxx) distingue 19 categorías gramaticales en este diccionario:
a: Frase nominal: *the big question*, *evening service*, etc.
b: Núcleo nominal + estructura relacionada: *question whether*, *job as*, etc.
c: Verbo + objeto: *loved him*, *weighed anchor*, etc.

Uno de los aspectos que llama inmediatamente nuestra atención con respecto a este diccionario es que se recogen unidades fraseológicas que no son colocaciones tal y como se entienden en el presente trabajo. En la introducción de su diccionario, Kjellmer (1994: xiv) define las colocaciones como "such recurring sequences of items as are grammatically well formed". Resulta evidente que según esta definición, toda combinación que se pueda clasificar dentro de una de las no pocas estructuras gramaticales que este autor considera y cuya aparición en el corpus sea frecuente es susceptible de considerarse una colocación. En nuestra opinión, esta definición lleva a Kjellmer a incluir no sólo colocaciones sino también locuciones y un gran número de combinaciones libres de palabras. La finalidad que este autor perseguía es loable en cuanto a que reconocía el valor que debía darse a las colocaciones más frecuentes de la lengua. En la introducción de su diccionario, Kjellmer (ibid.: xiii) expresa que su principal intención es dar respuesta a cuestiones como: "Does the collocation XX exist in the corpus?; How frequent is the collocation XX in the corpus?; How

---

d: Verbo + estructura relacionada: *lied down on*, *paid for by*, etc.
e: Verbo + verbo(s): *will come*, *had been given*, etc.
f: Infinitivo con *to*: *to be*, *to reply*, etc.
g: Verbo + predicativo: *was cold*, *keep warm*, etc.
h: Adverbio + núcleo adjetival o adverbial: *very young*, *just above*, etc.
i: Adverbio + conjunción subordinante: *now that*, *as if*, etc.
k: Conjunción + adverbio: *but again*, *and yet*, etc.
l: Preposición + conjunction subordinante: *except that*, *in that*, etc.
m: Adverbio o preposición + preposición: *out from*, *away to*, etc.
n: Oración de verbo en forma personal, opcionalmente + estructura relacionada: *he said*, *when he was shot*, etc.
o: Construcción de *it-* o *there-* + estructura relacionada*: it is impossible to*, *there is no reason to*, etc.
p: Oración de verbo en forma no personal u oración sin verbo: *lips compressed*, *hands off*, etc.
r: *As* o *like* + frase nominal o adverbio: *as rector*, *like myself*, etc.
s: Interjecciones, exclamaciones, expresiones de vocativo: *hey there*, *oh dear*, etc.
t: Elementos coordinados: *openly and honestly*, *actual or potential*, etc.
u: Expresiones no inglesas: *deja vu*, *per se*, etc.
Como podemos observar en la entrada de la figura 2.5., cuando una combinación pertenece a varias categorías a la vez, se incluyen las distintas letras consecutivas.

frequent is the collocation XX in relation to YY?". En este sentido, su marcada orientación cuantitativa a la hora de recoger y plasmar los datos en esta obra llevó a Johansson (1998: 341) a afirmar que "this dictionary (…) is a frequency list rather than a dictionary in the ordinary sense".

Debemos también añadir, sin embargo, que el hecho de que su definición del propio fenómeno colocacional sea también tan marcadamente estadística dio lugar a la siguiente crítica por parte de Benson (1995: 65-67), que nosotros compartimos plenamente:

> Kjellmer has disregarded three major developments in collocational lexicography. First, Chomsky demonstrated that collocations or semi-fixed phrases should not be confused with free combinations. Second, Hausmann made the emphatic point that dictionaries should not be cluttered with unnecessary, free combinations. Third, Apresyan, Mel'čuk and Zolkovsky pointed out the significance of lexical collocations. (...). The result is a work that few lexicographers will agree to call a collocational dictionary. (...) even a quick look at the contents results in deep disappointment (...). The Dictionary is cluttered with a mass of unneeded word combinations (...). On the other hand, many real collocations —both lexical and grammatical— have not been entered.

En total consonancia con la última apreciación de Benson, Granger (1998) destaca asimismo que una importante consecuencia de basar su definición y selección exclusivamente en la recurrencia de las combinaciones fue que, debido al limitado tamaño del corpus en el que se basaba, este diccionario no recoge colocaciones tan frecuentes como "*highly significant*" o "*seriously ill*".

e. *Collins COBUILD Dictionary of English Collocations* (1995)

Este diccionario de colocaciones, distribuido en formato CD-Rom, contiene 10.000 palabras base, o nodos, según la terminología de este trabajo, y los 20 colocados más frecuentes de cada una de ellas. Uno de sus aspectos más meritorios es que fue el

primer diccionario de colocaciones elaborado a partir de los datos de uno de los grandes corpus informatizados, concretamente del *Bank of English* cuando éste contaba con 200 millones de palabras. Su recopilación trató de restringirse en la medida de lo posible a colocaciones léxicas, eliminando en muchos casos las palabras gramaticales que co-aparecían con una muy alta frecuencia junto al nodo en el corpus. Como ya sabemos, este diccionario se compiló desde un claro enfoque estadístico al fenómeno colocacional, por lo que las combinaciones que incluye fueron seleccionadas atendiendo exclusivamente a la frecuencia de co-aparición de sus elementos y extrayendo aquellas que formaban combinaciones estadísticamente significativas dentro de un margen de cuatro palabras a la derecha y la izquierda del nodo. Ello implica que en esta compilación se incluyen unidades que no conforman verdaderas colocaciones desde un punto de vista fraseológico, y que se podrían incluir en las combinaciones libres ("*more + emphasis*") o en las locuciones ("*spill + beans*"). No parece, por tanto, que este recurso aporte ninguna novedad con respecto a la información que se puede obtener a partir del propio corpus en lo que respecta a la frecuencia colocacional. Asimismo, su utilidad en el campo de la enseñanza ha sido también cuestionado desde distintos estudios, precisamente por su naturaleza puramente estadística (Pérez Fernández, 2002; Nuccorini, 2003).

f. *Oxford Collocations Dictionary for Students of English* (2002)

Este banco de colocaciones, creado a partir de los datos del *British National Corpus*, consta de 150.000 combinaciones basadas en 9.000 bases nominales, verbales y adjetivales. Se trata de un diccionario con un claro enfoque fraseológico, donde se especifica que sólo se han incluido colocaciones de distinto nivel de fijación (débiles como "*extremely complicated*", medias como "*direct equivalent*" y fuertes como "*burning ambition*"), excluyendo así combinaciones libres y locuciones (salvo en contadas

ocasiones donde la locución no era totalmente idiomática y se podría considerar fronteriza: los autores citan el ejemplo "*drive a hard bargain*"). Debemos decir, no obstante, que, aunque al revisar los contenidos del diccionario es fácil comprobar que contiene efectivamente una información muy útil basada en criterios fraseológicos y pedagógicos, resulta llamativo encontrar un número considerable de combinaciones libres de palabras como las formadas con "*very*", que puede acompañar prácticamente a cualquier adjetivo.

Este diccionario está claramente orientado a la enseñanza/aprendizaje de lenguas, como se hace evidente incluso desde el propio título. En la introducción también se señala que su finalidad es eminentemente práctica, por lo que la máxima prioridad a la hora de seleccionar los contenidos fueron las necesidades de los alumnos que puedan usar el diccionario. Así pues, y partiendo del convencimiento de que los usos más típicos de la lengua son los más necesarios para el alumno, se utilizó un corpus de enorme solvencia como es el BNC y se seleccionaron las colocaciones más frecuentes tanto del inglés general, como también de algunas áreas específicas como el inglés periodístico. Es sin embargo de lamentar que, a pesar de que la frecuencia fue el principal criterio de selección de sus contenidos, no se ofrezca información alguna a este respecto en las diferentes entradas del diccionario.

g. McCarthy y O'Dell (2005)

El más reciente de los listados de colocaciones del inglés general que conocemos es el elaborado por McCarthy y O'Dell (2005), publicado en el libro de texto *English Collocations in Use* y también básicamente enfocado al estudiante de inglés como lengua extranjera. Este listado, recogido al final del volumen, consta aproximadamente de 1.500 colocaciones léxicas, extraídas a partir del *Cambridge International Corpus* cuando éste contaba con 750 millones de palabras. Como indican

sus autores, en esta lista se incluyen aquellas combinaciones que son significativas según los datos arrojados por el corpus, siendo pues la frecuencia un aspecto fundamental. En este sentido, McCarthy y O'Dell señalan que su prioridad era recoger aquellas colocaciones que el alumno realmente necesite a la hora de usar la lengua comunicativamente. Por esta razón, aquellas colocaciones que, a pesar de su idiosincrasia, no son realmente comunes en el lenguaje se excluyeron de esta lista (quizá el caso más paradigmático sea "*rancid butter*", una colocación muy arbitraria pero de uso muy limitado según los autores).

Por otro lado, y también con el fin de seleccionar aquellas combinaciones realmente útiles para el alumno, expresiones fácilmente deducibles como "*friendly girl*" o "*to eat an apple*" no se incluyeron en el listado, a pesar de que se consideran colocaciones en esta obra. En nuestra opinión, sin embargo, la razón por la que estas combinaciones son deducibles y no requieren especial atención en un libro de colocaciones es porque se trata de combinaciones libres de palabras. En cualquier caso, y al margen de cuestiones teóricas, nos parece muy acertado que este tipo de unidades se excluyeran del listado final.

Sin embargo, y como ocurría con el trabajo anterior, no se ofrece ninguna indicación acerca de la frecuencia de las colocaciones aquí recogidas, limitándose pues a ofrecer un listado ordenado alfabéticamente.

h. Shin y Nation (2008)

El último de los listados de colocaciones producidos hasta la fecha según la información de que disponemos es el llevado a cabo por Shin y Nation. En este caso, y a diferencia de todos los anteriores, no se trata de una lista de colocaciones del inglés general, sino específicamente diseñada para explorar la vertiente oral de la lengua. Este listado, el único que conocemos donde las colocaciones se presentan

ordenadas por frecuencia, recoge las 4.698 colocaciones más frecuentes del inglés oral, extraídas a partir de los 10 millones de palabras que componen la parte oral del BNC. Con un enfoque puramente estadístico y ateniéndose por tanto únicamente al criterio de la frecuencia, los autores de este trabajo seleccionaron las 1.000 primeras palabras léxicas del corpus (entendiendo palabra como forma léxica, es decir, tratando las inflexiones y derivaciones como palabras distintas) y las utilizaron como nodos para explorar su comportamiento colocacional.

Mediante el uso de la herramienta *WordSmith Tools* 3.0 (ver sección 3.2.2.3.), extrajeron los colocados más frecuentes de dichos nodos, estableciendo 30 como el número mínimo de veces que la combinación debería co-aparecer en el corpus para asegurar su alta frecuencia. Es sin duda muy meritorio el hecho de que en la elaboración de este listado se contabilizaron por separado los distintos significados de las colocaciones polisémicas, dando lugar así a un mayor grado de precisión en los datos. Finalmente, se decidió incluir solamente aquellas combinaciones que conformaran una unidad sintagmática en la oración, lo cual permitía descartar combinaciones de palabras frecuentes como "*you at the place*", donde confluyen elementos pertenecientes a dos sintagmas diferentes.

Aunque sólo tenemos acceso a las primeras 100 colocaciones que integran este listado (Shin y Nation, 2008), podemos observar que el tipo de unidades de que se compone no responde a lo que en términos fraseológicos se considera una colocación. En la tabla 2.15. ofrecemos los primeros diez elementos de este listado, donde será fácil comprobar este punto.

RK	Collocations	FRE
1	you know	27348
2	I think (that)	25862
3	a bit	7766
4	(always [155], never [87]) used to {INF}	7663
5	as well	5754
6	a lot of {N}	5750
7	{No.} pounds	5598
8	thank you	4789
9	{No.} years	4237
10	in fact	3009

Tabla 2.15.: 10 primeras colocaciones del listado de Shin y Nation (2008: 346-347)

Como podemos observar, combinaciones del tipo "*I think that*" o "*never used to*", por ejemplo, no se corresponden con la noción fraseológica de colocación que contemplamos en esta tesis. No cabe duda de que el resultado obtenido en este caso es también una clara consecuencia del tipo de metodología empleado para su elaboración, donde únicamente se observaron criterios de frecuencia estadística y donde la arbitrariedad necesaria en una colocación verdaderamente interesante desde el punto de vista pedagógico no se consideró como criterio. A la vista de este listado, en el que se incluyen, como suele suceder con todos aquellos producidos a partir de análisis puramente estadísticos, un buen número de unidades libres, los propios autores concluyen que "frequency is not everything" (ibid.: 345). Al afirmar lo anterior, Shin y Nation no se refieren, sin embargo, a su metodología de extracción de combinaciones basada estrictamente en la frecuencia, sino a las limitaciones de los listados de frecuencias, manifestando que a la hora de utilizarlas en el aula se deben tener otros criterios en cuenta: "Although frequency in the language is an important criterion for selecting what to focus on, it is only one of several important criteria"

(ibid.: 345-346). En nuestra opinión, sin embargo, si un listado de frecuencias está verdaderamente bien construido e incluye por tanto aquellos elementos que son de pleno interés para el alumno, poder seleccionarlos de acuerdo a su frecuencia es no sólo útil sino posiblemente lo más esencial.

A la vista de todo lo anterior, podemos concluir que, si bien se han producido varias compilaciones pedagógicas y lexicográficas donde la frecuencia ha sido un factor fundamental, es evidente, en primer lugar, que estos trabajos son todavía muy escasos si consideramos el consenso que existe en la actualidad en cuanto a su enorme trascendencia en el aprendizaje de una lengua. Por otro lado, en algunos de ellos existen claros problemas de fiabilidad debido a que se compilaron a partir de fuentes poco representativas. También en algunas ocasiones estos listados presentan inconsistencias en cuanto a la propia naturaleza de las combinaciones que recogen. Finalmente, y quizá más importante, ninguno de los trabajos citados en este apartado constituye en realidad un listado de colocaciones frecuentes del inglés general donde éstas aparezcan por orden de frecuencia. Partiendo de esta realidad, parece que continúa aún muy vigente la llamada que ocho años atrás realizara Nation (2001: 328) en cuanto a la necesidad de contar con más investigaciones sobre la frecuencia de las colocaciones: "From a vocabulary learning point of view, we need research into collocation: 1) to tell us what the high-frequency collocations are, 2) to tell us what the unpredictable collocations of high-frequency words are". Así pues, con el objetivo de contribuir a aportar algo de luz en el estudio de las colocaciones más frecuentes de la lengua inglesa, y fundamentalmente para contar así con una fuente de datos lo suficientemente solvente que nos ayude a diseñar un test de colocaciones válido y fiable dirigido a nuestro alumnado, decidimos **elaborar un listado de colocaciones donde se tuviesen en cuenta criterios tanto de frecuencia como**

pedagógicos. El proceso de elaboración de dicho listado, que ya en sí mismo supone una investigación, será el objeto de estudio del capítulo siguiente.

2.4. Conclusión

La idea central que subyace en el capítulo que ahora concluye es que la frecuencia es un aspecto fundamental en el léxico de una lengua, y particularmente en su enseñanza y aprendizaje. No en vano, prácticamente todos los modelos de competencia léxica que existen en la actualidad consideran la frecuencia como uno de los factores que contribuyen a conformar el dominio de una palabra o expresión fraseológica. Desde un punto de vista más práctico, parece evidente que los elementos más frecuentes son aquellos que con mayor probabilidad necesitará el alumno en su uso tanto receptivo como productivo de la lengua, por lo que deben constituir una prioridad en cualquier sílabo de lenguas extranjeras. Nos gustaría destacar en este sentido que no pretendemos decir, en modo alguno, que la frecuencia sea el único criterio que se debe contemplar a la hora de seleccionar los contenidos de un programa pedagógico, pero sí creemos que se trata de un factor esencial que debería ocupar un lugar central en la actividad docente.

Quizá una de las más claras evidencias de la importancia que tiene la noción de frecuencia en el aprendizaje del vocabulario es el hecho de que estos dos aspectos han ido tradicionalmente de la mano y su evolución ha corrido paralela desde principios del siglo XX y hasta nuestros días. Como hemos podido comprobar en este capítulo, ya desde que los primeros lingüistas y profesores comenzaran a percibir que el léxico es de vital importancia en un programa de enseñanza de lenguas, la conciencia de que el vocabulario más frecuente es primordial ha estado presente en los trabajos lingüísticos y pedagógicos. Una vez superado el paréntesis que supusieron en este sentido las teorías chomskianas de los años 50 y 60, y gracias

también a la inestimable fuente de datos que suponían los nuevos corpus informatizados de los 70 y 80, las cuestiones relativas al vocabulario y a la frecuencia de los patrones léxicos volvieron a despertar interés entre los investigadores, de forma más silenciosa primero, y con un enorme estruendo a partir de los 90 y hasta la actualidad.

Los estudios de frecuencias han tenido, pues, una clara repercusión en la era computacional, dando lugar a la aparición de un buen número de listados de palabras frecuentes, que ofrecen sin duda un enorme potencial en el terreno de la enseñanza y que están aún por explotar en buena medida. Pero una vez analizado el dominio de la palabra, debemos dar un paso más y avanzar hacia el campo de la fraseología, por supuesto más complejo pero también quizá más prometedor. Como se ha visto en este capítulo, aquí es donde sin duda queda aún un largo camino por recorrer. En lo que respecta a las colocaciones, el elemento central de esta tesis, y sin pasar por tanto todavía al resto de unidades multiléxicas como puedan ser los nombres compuestos o las locuciones, son verdaderamente escasos los estudios dedicados a delimitar cuáles son las colocaciones más frecuentes del inglés, y que deben ser en consecuencia prioritarias en el aula, tanto en la enseñanza como en la evaluación. Es aquí, pues, donde el presente trabajo pretende realizar su aportación.

CAPÍTULO 3

LA COMPILACIÓN DE UN LISTADO DE FRECUENCIAS DE COLOCACIONES LÉXICAS

> *I have emphasized throughout that no procedures can ever be entirely automatic. We always start with intuitions about what is interesting to study, and intuition re-enters, in designing procedures and in interpreting findings.*
>
> Stubbs (1995: 48)

3.1. Introducción

Como ya mencionamos, uno de los objetivos fundamentales de este trabajo es elaborar un test válido y fiable que nos ayude a evaluar la competencia colocacional del alumnado universitario. Aunque todos los aspectos relativos a las características que debe tener un test se tratarán más detenidamente en el capítulo 4, nos parece necesario adelantar aquí que uno de los factores que mayor atención merece a la hora de diseñar y construir un instrumento de evaluación es su validez. Un test debe ser una muestra representativa del campo que pretende examinar, por lo que sus

contenidos han de ser cuidadosamente seleccionados a fin de que reflejen lo más fielmente posible el área que se va a evaluar y que sus resultados sean, por tanto, válidos. En un test de colocaciones como es nuestro caso, donde pretendemos evaluar el conocimiento que los alumnos tienen de las colocaciones más frecuentes de la lengua inglesa, es requisito imprescindible contar con un banco de colocaciones frecuentes suficientemente amplio y del que podamos extraer una muestra representativa y en proporciones semejantes a las que existen en la lengua en su conjunto.

Así pues, en este capítulo describiremos en primer lugar el procedimiento de recopilación de dicho banco de colocaciones. Como recordaremos, de acuerdo con nuestra definición de colocación, ésta está formada por dos elementos, la base y el colocado, que se encuentran a diferentes niveles semánticos, siendo la base el elemento independiente que el hablante elige libremente y del cual parte a la hora de usar la lengua, mientras que el colocado depende y está arbitrariamente restringido por la base. En consonancia con esta definición, el planteamiento para obtener nuestro listado de colocaciones debía partir de una unidad léxica preseleccionada a partir de la cual se buscarían las colocaciones a que da lugar al co-aparecer con otras palabras. Así, nuestro proceso de compilación comprendió dos etapas: en primer lugar se seleccionaron las bases de las colocaciones a partir de un listado de las 1.000 palabras más frecuentes del inglés, y en segundo lugar se extrajeron los colocados de dichas bases mediante el uso de técnicas computacionales y la posterior aplicación de unos criterios de selección necesarios para obtener un listado consecuente con nuestros objetivos y nuestro marco teórico. En este capítulo ofreceremos una descripción pormenorizada de los procedimientos que se siguieron para llevar a cabo este proceso.

Una vez obtenido nuestro listado, sin embargo, pudimos comprobar que la información que éste contenía iba más allá de lo que en principio cabía esperar.

Como resultado de esta compilación, saltaba a la vista que las colocaciones obtenidas parecen no ser meras combinaciones léxicas establecidas de manera totalmente arbitraria, sino que en muchos casos esconden procesos mucho más profundos y sutiles que a menudo responden a cuestiones semánticas, pragmáticas, socioculturales e incluso ideológicas. Este fenómeno, al que ya importantes lingüistas (Bernardini, 2002; Tognini-Bonelli, 2008) han denominado "el proceso de serendipia", se hace muy evidente en los resultados obtenidos en este listado y a su análisis dedicaremos la última sección del presente capítulo.

3.2. Procedimiento de elaboración del listado

3.2.1. Selección de la base

En el apartado 1.4.2.2. ya apuntamos que según la categoría gramatical del elemento que conforme la base de las colocaciones léxicas, éstas se pueden dividir en tres grupos diferentes:

1) Aquellas donde la base es un sustantivo y se combina con colocados nominales, adjetivales y verbales dando lugar a las estructuras nombre+nombre (N+N, donde el primer elemento es el colocado y el segundo elemento la base), adjetivo+nombre (A+N), verbo+nombre (V+N) y nombre+verbo (N+V).
2) Las formadas por una base adjetival acompañada de un adverbio formando combinaciones adverbio+adjetivo (Adv+A).
3) Las que contienen un verbo como base acompañado de un adverbio en estructuras verbo+adverbio (V+Adv).

De entre estos tres tipos de colocaciones, y partiendo de la base de que no era factible cubrirlos todos en este estudio, nuestro trabajo se dedicó exclusivamente a

las del primer apartado por dos razones. En primer lugar, resulta evidente que se trata del grupo que más combinaciones abarca dado que comprende cuatro estructuras gramaticales, a diferencia de los otros dos grupos que se limitan a un único tipo de combinación sintáctica. En segundo lugar, y sin duda la razón de más peso en nuestra elección, el hecho de que este grupo de colocaciones esté formado por una base nominal lo hace especialmente interesante desde el punto de vista pedagógico, ya que los nombres son los elementos que en mayor medida articulan los conceptos expresados cuando usamos el lenguaje y, por tanto, son el principal punto de referencia para el hablante a la hora de construir un mensaje. Como asegura Woolard (2000: 35), "the fact that nouns tend to be the focus of information in a text, that we tend to build the information up around the nouns, means that they are the most suitable headwords for collocation searches". Por ello, la primera tarea en la elaboración de nuestro listado fue seleccionar los sustantivos que conformarían las bases de nuestras colocaciones.

Como hicimos notar en el capítulo anterior, existe hoy en día un gran consenso en torno a la importancia de la frecuencia como factor decisivo en la enseñanza (y por ende en la evaluación) del vocabulario. Las palabras más frecuentes de la lengua son tan fundamentales que debemos hacer todo lo que esté en nuestra mano como profesores para garantizar que el alumno domina estas palabras en profundidad (Nation, 2001). En este sentido, ya mencionamos que conocer el comportamiento colocacional de aquellas palabras que ya son familiares para el alumno resulta tanto o más necesario que aprender vocabulario y colocaciones nuevas, menos frecuentes y, en consecuencia, menos útiles. Son varios los autores que se han pronunciado en esta misma línea. Así, Woolard (2005b: 46) opina que "learning new vocabulary is not just learning new words, it is often learning familiar words in new combinations" mientras que Shin y Nation (2008: 342) también consideran que "the learning of the collocations should strengthen and enrich words

students already know, not add an additional burden by adding unknown, lower frequency vocabulary".

Por otro lado, muchos de los sustantivos más frecuentes de la lengua son palabras con significados abstractos o en ocasiones incluso tan polisémicos que necesitan de un colocado para cobrar un valor semántico pleno. Nos referimos, por ejemplo, a términos como *idea*, *line*, *point*, *problem*, *question*, o *way* que, a pesar de ser nombres de altísima frecuencia y que se consideran básicos en el aprendizaje del inglés, suelen causar problemas incluso a los alumnos de niveles intermedio y avanzado a la hora de usarlos productivamente (Philip, 2007b). Como destaca Philip, dado que los alumnos aprenden las palabras de forma aislada, les resulta muy difícil utilizarlas en sus combinaciones fraseológicas naturales por lo que tienden a insertarlas en estructuras calcadas de su lengua materna. Esto repercute en una pobreza léxica y una falta de fluidez y precisión que limita claramente la competencia comunicativa del hablante, incluso en niveles universitarios. En palabras de la propia autora:

> Students generally encounter words in their literal sense first, match them to a translation equivalent in their L1, and from then on, unless instructed otherwise, use the word in calqued forms of the L1 phraseology. The relative success of this strategy effectively masks the underlying problem, which is more serious than simply getting collocations wrong. Persistent calquing actually prevents students from acquiring a sense of the word's conceptual range in the L2, negatively affecting textual fluency and cohesiveness (Philip, 2007a: 13).

Paradójicamente, a pesar de su frecuencia, de la dificultad que pueden entrañar para el alumno y de la trascendencia que tienen para un uso adecuado de la L2, estos sustantivos no suelen recibir, por lo general, una atención prioritaria por parte del profesor. Tanto nuestra propia experiencia en el aprendizaje de lenguas extranjeras como los estudios empíricos llevados a cabo para analizar el vocabulario incluido en libros de texto producidos por editoriales de prestigio nos demuestran que, en

muchas ocasiones, una vez que el alumno aprende el significado básico y más concreto de este tipo de sustantivos, raramente se vuelven a tratar en el aula, por lo que los aprendices no suelen profundizar en las restricciones de uso y los diferentes valores semánticos que adquieren en compañía de otras palabras a los que alude Philip (ibid.). A la vista de esta situación, en este estudio consideramos que los sustantivos más frecuentes de la lengua merecen una especial atención, particularmente en lo que toca a las combinaciones colocacionales en las que participan.

Con el fin de obtener un listado de los nombres más frecuentes de la lengua inglesa utilizamos la lista de frecuencias compilada por María Teresa López-Mezquita Molina (2005) en el seno del proyecto ADELEX (Universidad de Granada), por ser, a nuestro juicio, la más completa y pedagógica de cuantas conocemos. Como recordaremos, se trata de un listado que recoge las 7.125 palabras más frecuentes del inglés, cuya rigurosidad y fiabilidad viene dada principalmente por el hecho de que fue compilado a partir de un proceso de comparación y complementación de los datos aportados por el *British National Corpus*, el *Bank of English* y el *Longman Corpus Network* (ver sección 2.3.2.2.1.).

Para nuestro estudio seleccionamos las 1.000 primeras palabras de este listado, a partir de las cuales extrajimos los sustantivos, que sumaron un total de 412[1]. Dado que, como ya sabemos, el listado de López-Mezquita consta de lemas, los nombres que conformaban nuestra lista se encontraban en su forma base. Sin embargo, en el capítulo 1 de esta tesis dejamos constancia de que, de acuerdo con nuestro concepto de colocación, la información lematizada no es suficiente en lo que respecta a las bases de las colocaciones puesto que los sustantivos pueden tener una

[1] El listado completo sumaba 413 sustantivos pero se redujo en una unidad ya que eliminamos la palabra "*few*" que, aunque aparece como sustantivo en el listado de López-Mezquita, no está etiquetada como tal en ninguno de los dos corpus que utilizamos para su análisis (el *British National Corpus* y el *Bank of English*).

combinatoria marcadamente distinta en sus formas singular y plural (Sinclair, 2003; Evert, Heid y Spranger, 2004). Así, por ejemplo, parece claro que colocaciones como *to come to an end* o *mutual respect* sólo se dan con el nombre en singular, mientras que *to take sides* o *to roll your eyes* suelen aparecer únicamente en plural. Teniendo en cuenta lo anterior, en nuestro estudio se contemplaron las formas singular y plural como dos palabras distintas, por lo que nuestra lista inicial aumentó hasta contar con 803 sustantivos, resultado de añadir todas las formas plurales de los nombres, excepto en los 21 casos en que se trataba de nombres incontables que, por tanto, sólo se usan en su forma base[2].

3.2.2. Extracción de los colocados

Una vez obtenida nuestra lista inicial con los nombres más frecuentes del inglés, la siguiente fase de nuestra investigación se dedicó a la búsqueda y extracción de los colocados nominales, adjetivales y verbales de dichos sustantivos. Para llevar a cabo esta tarea existían dos posibles métodos a seguir: la "elicitación", basada en el juicio subjetivo del hablante nativo y por tanto propia de enfoques próximos a los estudios psicolingüísticos y/o experimentales, o la lingüística de corpus, que hace uso de un número amplísimo de ejemplos donde un fenómeno concreto ha sido producido por muchos y muy diferentes hablantes nativos en situaciones comunicativas reales. En nuestro caso, consideramos que la metodología que más ventajas aportaba era la lingüística de corpus.

Como acabamos de mencionar, la principal diferencia entre ambos enfoques radica en el tipo de información a que da lugar cada uno de ellos. A este respecto,

[2] Estos nombres son, por orden de aparición en la lista de frecuencias: *information*, *money*, *water*, *police*, *education*, *health*, *management*, *evidence*, *data*, *security*, *music*, *hair*, *knowledge*, *series*, *training*, *news*, *means*, *help*, *whole*, *advice* y *blood*.

aquellos autores que han puesto de manifiesto las limitaciones de los corpus como fuente de datos (Kaltenböck y Mehlmauer-Larcher, 2005) suelen argumentar que mientras que la introspección humana es capaz de aportar información no sólo sobre lo que se ha dicho, sino sobre lo que se puede decir, sobre todo lo que cabe y es posible en la lengua y sobre cómo se estructura el lenguaje en nuestra mente, los estudios de corpus se quedan un paso atrás, puesto que sólo recogen lo que se ha dicho, limitándose por tanto a la observación externa de la realidad lingüística. Widdowson (2000: 6) es quizá la figura más reconocida dentro de esta corriente de pensamiento:

> The perspective of corpus linguistics is that of the third person observer, not that of the first person experiencer of language... what these [corpus] descriptions represent is not insider first person reality as apprehended by language users but outsider third person reality, a construct of observation.

Según esta perspectiva, todo lo que resta al lingüista que usa corpus es mera especulación sobre lo que sería o no posible según los datos observados (Kaltenböck y Mehlmauer-Larcher, 2005).

Muy en desacuerdo con lo anterior, creemos firmemente que, aunque es cierto que un corpus no indica lo que en teoría se puede o no hacer en la lengua, la información que ofrece sobre lo que se ha usado en la práctica y, en especial, sobre aquello que es frecuente, lo que la mayoría de los hablantes nativos emplean en su uso diario del lenguaje, es sin duda muy relevante ya que, en definitiva, se trata de lo que un aprendiz de esa lengua necesita saber. Más aún, el uso de corpus como fuente de información presenta, a nuestro juicio, otra clara ventaja frente a aquellos fundamentados en la intuición humana: la objetividad. A este respecto nos parece muy interesante la observación de Kennedy (1998: 8):

> Any scientific enterprise must be empirical in the sense that it has to be supported or falsified on evidence and, in the final analysis, statements made about language have to stand up to the evidence of language use. The evidence can be based on the introspective judgement of the speakers of the language or on a corpus of text. The difference lies in the richness of the evidence and the confidence we can have in the generalizability of that evidence, in its validity and reliability.

A la luz de esta afirmación parece evidente que el empleo de fuentes basadas en miles de ejemplos de la lengua usada en contextos auténticos contribuye a garantizar la validez de cualquier estudio, a diferencia de la introspección, fundamentada en juicios subjetivos emitidos por un número siempre limitado de sujetos y en condiciones experimentales que se alejan normalmente de los contextos reales de uso de la lengua. A esto cabe añadir, además, que en no pocas ocasiones la intuición humana produce planteamientos equivocados que no se corresponden con la realidad que reflejan los corpus (Aarts, 1991; Aston y Burnard, 1998), algo que, evidentemente, también sucede en el ámbito de las colocaciones ya que éstas, en palabras de Hunston (2002: 109), "are often unavailable to intuition or conscious awareness". Concretamente en lo que se refiere a la frecuencia de las colocaciones, Stubbs (1995: 24-25) afirma que los hablantes nativos "certainly cannot document collocations with any thoroughness, and they cannot give accurate estimates of the frequency and distribution of different collocations". En la misma línea, Deignan (2005: 194) también opina que "native-speaker intuition is unreliable in identifying frequent collocates for many words (…), and (…) computerized corpora can provide more accurate and detailed information about collocational patterns".

Por otro lado, consideramos asimismo que era especialmente recomendable utilizar la lingüística de corpus para nuestro trabajo porque las técnicas que ofrece, fundamentadas principalmente en el análisis empírico de datos cuantificables y contrastables, se adecuan perfectamente a nuestro objetivo de recoger aquellos colocados que acompañan a los nombres de forma más frecuente y estadísticamente significativa. En nuestro caso, por tanto, la metodología basada en el estudio y

análisis de corpus nos pareció la mejor forma de elaborar una lista de las colocaciones más frecuentes que fuera la base para procurar la validez de contenido de nuestro test.

3.2.2.1. Los corpus utilizados

Para realizar el análisis y la extracción de los colocados era necesario utilizar corpus que fuesen representativos de la lengua inglesa en su conjunto, por lo que se emplearon corpus de referencia (ver sección 2.3.1.3.2.). Como ya vimos, existen diferentes corpus de este tipo en inglés, entre los cuales los más destacados por su tamaño y sus características son el *British National Corpus* (BNC), el *Bank of English* (BoE), el *Longman Corpus Network* y el *Cambridge International Corpus* (CIC). De entre ellos, solamente teníamos acceso a los dos primeros, ya que, por el momento, el *Longman Corpus Network* sólo es accesible para los lexicógrafos de la editorial Pearson/Longman, al igual que el CIC, que está disponible únicamente para los investigadores que trabajan para la editorial Cambridge University Press. El BNC y el BoE fueron pues los dos corpus que empleamos en nuestro estudio para la extracción de colocaciones.

Llegados a este punto cabe quizá preguntarse por qué razón se emplearon dos corpus en esta investigación en lugar de uno solo. A la hora de seleccionar el corpus que nos iba a servir como fuente de información, consideramos cuidadosamente las ventajas e inconvenientes que presentan tanto el BNC como el BoE en lo que se refiere a su naturaleza y contenido. En cuanto al primero, y como ya vimos, se trata de un corpus sobradamente reconocido entre la comunidad académica, siendo uno de los recursos más estandarizados y ampliamente utilizados en el área de los estudios de corpus en todo el mundo. No en vano existen innumerables trabajos de investigación lingüística, muchos de ellos producidos muy

recientemente (Kennedy, 2007; Butler, 2008; Pearce, 2008; Sievers, 2008; por nombrar sólo unos pocos), que siguen considerando el BNC como una muestra representativa del inglés, bien equilibrada y de tamaño suficiente como para otorgar validez y fiabilidad a las conclusiones que de su análisis se derivan. En nuestra opinión, la mayor ventaja que presenta el BNC para nuestro estudio es que se trata del único corpus de referencia del inglés compuesto por textos de diversa procedencia y no únicamente obtenidos de Internet al que se tiene acceso de forma íntegra, siendo pues todavía hoy en día el corpus de mayor extensión de entre los más representativos de esta lengua disponible para cualquier investigador.

Es sin embargo necesario considerar también las limitaciones que presenta este corpus. En este sentido cabe destacar que los textos orales y escritos que lo componen fueron producidos entre 1964 y 1993. Dado que se trata de un corpus cerrado que no ha sido modificado en su composición desde su creación inicial, no recoge ninguna de las novedades que se han producido en la lengua en los últimos quince años, un aspecto preocupante si tenemos además en cuenta que el inglés es una lengua que evoluciona muy rápidamente (Graddol, 1997). Algunos autores han destacado este hecho (Pearce, 2008), y pensamos que un buen ejemplo de ello es la escasa representación de términos relativos a la informática que existe en el BNC. Llama la atención, por poner algunos ejemplos, que la palabra *Internet* ofrezca tan sólo 74 resultados, *web site* solamente aparece en una ocasión y, evidentemente, la palabra *blog* no está recogida en este corpus.

Fijando ahora nuestra atención en el segundo banco de datos que se encontraba disponible, el *Bank of English*, comprobamos que este corpus también presentaba ventajas e inconvenientes para nuestro propósito concreto. La principal ventaja que pudimos observar es que se trata de un corpus "monitor", es decir, de un banco de datos que no deja de ser actualizado con nuevos textos cada cierto tiempo. Este hecho era de especial relevancia en nuestro caso puesto que, como dijimos más

arriba, el BNC no recoge los nuevos usos que se han creado en la lengua desde el año 1993. Por otro lado, también contemplamos los rigurosos criterios con que el BoE ha sido diseñado y periódicamente actualizado, logrando una fuente de datos bien equilibrada en su contenido y creada a partir de miles de fuentes diferentes, todo lo cual la convierte en un corpus altamente representativo del inglés.

Pero, al igual que sucedía con el BNC, el BoE también venía acompañado de ciertas limitaciones que debíamos considerar. Quizá la más evidente sea el hecho de que, a pesar de que este corpus cuenta en la actualidad con 524 millones de palabras en su versión completa, aquellos investigadores no pertenecientes a la Universidad de Birmingham que desean utilizarlo sólo obtienen acceso a un subcorpus de 56 millones de palabras, es decir, aproximadamente una décima parte de la versión íntegra. Sin embargo, aunque evidentemente se trata de una reducción notable en proporción al tamaño total del corpus, el número de palabras a las que se permite el acceso nos pareció lo suficientemente amplio como para utilizarlo como corpus de referencia a partir del cual poder extraer datos válidos y fiables para nuestro test.

Tras una profunda reflexión acerca de todo lo anterior, llegamos a la conclusión de que ambos corpus se complementaban claramente y, por tanto, debíamos utilizar los dos en nuestro análisis de colocaciones, ya que el BNC ofrecía las mejores condiciones en lo que a tamaño se refiere, mientras que el BoE aportaba información actualizada sobre las colocaciones más frecuentes del inglés que se emplea en la sociedad de los últimos años. Con el uso combinado de ambas herramientas obteníamos pues máxima representatividad y fiabilidad de cobertura de la lengua, lo cual contribuiría a aumentar la validez de contenido de nuestro test.

Una vez seleccionados los dos corpus que nos servirían como fuente de información, procedimos a considerar otros aspectos fundamentales como los análisis estadísticos y los programas informáticos que debíamos utilizar en el proceso de extracción de colocaciones.

3.2.2.2. Análisis estadísticos

Ya mencionamos anteriormente que una de las principales ventajas que la lingüística de corpus y los medios computacionales han venido a aportar a los estudios lingüísticos es la posibilidad de manejar ingentes cantidades de datos desde una aproximación cuantitativa aplicando cálculos matemáticos y estadísticos que permiten obtener nueva información sobre la naturaleza del lenguaje en general y de cada lengua en particular. Una de las áreas que mayor impacto ha recibido gracias a estas nuevas técnicas de análisis es la fraseología y, de un modo especialmente notable, el estudio de las colocaciones. Dado el carácter más o menos fijo y recurrente de estas combinaciones prefabricadas de la lengua, resulta ahora más sencillo que nunca para el investigador obtener combinaciones de palabras que aparecen en la lengua con una frecuencia mayor de la que cabría esperar si su aparición conjunta se debiera al azar y que son, por consiguiente, estadísticamente significativas.

A la hora de evaluar cuantitativamente el grado de restricción con que los dos miembros de una combinación co-aparecen en la lengua, existen dos posibilidades: considerar la frecuencia de co-aparición absoluta o la frecuencia relativa (Chung y Lee, 2001). En el primero de los casos, obviamente, dos palabras cuya frecuencia individual sea muy elevada tenderán a co-aparecer de forma más frecuente que las palabras menos comunes de la lengua. Pero resulta evidente que esto no nos indica nada sobre la fuerza de atracción que existe entre las unidades léxicas, puesto que la propia frecuencia de cada una de ellas justificaría que aparezcan de forma conjunta repetidamente. Lo que necesitaríamos conocer, por tanto, es el grado de asociación que existe entre estas palabras, es decir, hasta qué punto una palabra condiciona su entorno y exige la presencia de otra con la que formar combinación. Para evaluar esta atracción idiosincrásica, existen fórmulas estadísticas que consideran la frecuencia relativa de la combinación, teniendo en cuenta no sólo la frecuencia de la palabra

individual sino otros factores como el tamaño del corpus utilizado y la probabilidad combinatoria.

La gran mayoría de estos análisis parten de dos conceptos fundamentales: la frecuencia observada y la frecuencia esperada. Así, estos trabajos tienen como objetivo subyacente la comparación de aquello que observamos en un corpus con lo que cabría esperar si partimos de la siguiente hipótesis nula: *una palabra A no tiene ningún efecto sobre las palabras que la flanquean y la frecuencia de estas palabras sería exactamente la misma independientemente de la presencia de A*. Partiendo de esta premisa, se han planteado a lo largo de las últimas décadas distintos cálculos matemáticos capaces de identificar automáticamente combinaciones donde la diferencia entre la frecuencia observada y la esperada es estadísticamente significativa y que son, pues, firmes candidatas a formar colocaciones[3].

El hecho de que, como decimos, existan diferentes medidas estadísticas para la obtención de combinaciones léxicas, ha llevado recientemente a distintos autores a realizar estudios comparativos para delimitar los aspectos positivos y negativos de cada una de ellas (Evert 2004[4]; Paquot, 2007). Estos trabajos han puesto de manifiesto que los diferentes cálculos son claramente dispares y no dan lugar a un conjunto más o menos homogéneo de combinaciones. Muy al contrario, "a wide range of equally plausible association measures will extract entirely different sets of 'recurrent word combinations' from a given corpus" (Evert y Krenn, 2005: 3). Es necesario señalar que, si bien es cierto que algunos estudios (Anagnostou y Weir, 2006) han destacado el Coeficiente de Verosimilitud como la medida más adecuada

[3] Nótese que las combinaciones extraídas no siempre dan lugar a colocaciones tal y como se conciben en esta tesis por lo que, siguiendo la distinción terminológica que algunos autores están comenzando a establecer (Granger y Paquot, 2008; Seretan, 2008), nos referiremos a combinaciones léxicas, co-apariciones o binomios cuando se trate de conjuntos estadísticamente significativos y a colocaciones cuando se trate de unidades fraseológicas.

[4] En su tesis doctoral, Evert (2004) llega incluso a enumerar 47 medidas diferentes para extraer binomios de un corpus.

para extraer binomios, casi todos los autores parecen coincidir en que no existe una medida de asociación marcadamente mejor que el resto: la medida estadística más adecuada será aquella que mejor satisfaga las necesidades concretas en cada caso (Church et al., 1991; Biber, Conrad y Reppen, 1998).

Para poder decidir qué método estadístico era el más apropiado para nuestras necesidades concretas, llevamos a cabo una revisión de las diferentes posibilidades que nos ofrecen hoy día los estudios sobre extracción automatizada de combinaciones léxicas. Las medidas que se han empleado tradicionalmente con este fin son las de **Información Mutua (IM)**, **IM^3** y ***Log-log*** por un lado, y las de ***z-score***, ***t-score***, χ^2 y **Coeficiente de Verosimilitud** por otro[5]. En los párrafos siguientes llevaremos a cabo una breve revisión de estas medidas, aunque no profundizaremos en las imbricaciones matemáticas de carácter altamente especializado ya que no nos parece necesario para los objetivos de este trabajo[6].

- **Información Mutua**

Esta medida estadística, derivada por Church y Hanks (1989) de la Teoría de la Información (Shannon y Weaver, 1949), expresa la diferencia entre la co-aparición observada de dos palabras A y B y la co-aparición esperada según la hipótesis nula. Para calcular la frecuencia esperada de co-aparición, este método calcula la probabilidad de que A y B aparezcan juntas en la lengua considerando la frecuencia con que aparecen cada una de ellas por separado y el número de palabras totales del corpus, es decir, considerando la frecuencia relativa de cada una de las palabras. Si se cumpliera la hipótesis y no existiera entre ambas palabras más conexión que la que

[5] Hemos conservado los términos ingleses para *log-log*, *t-score* y *z-score* puesto que se suelen emplear sin traducir en la bibliografía española (Pazos Bretaña, 2005), pero hemos traducido aquellos conceptos que se suelen emplear en castellano en las publicaciones nacionales: Información Mutua (*Mutual Information*) y Coeficiente de Verosimilitud (*Log Likelihood*).

[6] Para obtener información estadística especializada sobre las medidas de asociación léxica ver Dunning (1993), Oakes (1998), Manning y Schütze (1999) y Evert (2004).

cabe esperar en base a la probabilidad matemática, la frecuencia esperada coincidiría exactamente con la frecuencia observada, siendo pues la diferencia entre ambas 0. Lógicamente, cuanto mayor sea la diferencia entre lo observado y lo esperado, mayor será el índice de Información Mutua y mayor la fuerza de atracción entre A y B. No obstante, es importante notar que esta medida determina la fuerza de asociación entre dos palabras, pero no indica si esa asociación, es decir, si la diferencia entre lo esperado y lo observado, es estadísticamente significativa (McEnery, 2006).

Existe, además, un problema fundamental en cuanto a los resultados que ofrece Información Mutua que no podemos obviar. Como numerosos autores han señalado (Manning y Schütze, 1999; Chung y Lee, 2001; McEnery, Xiao y Tono, 2006), este método tiende a dar prioridad a las palabras menos frecuentes de la lengua, destacando las combinaciones donde intervienen este tipo de unidades frente a otras en las que sus componentes co-aparecen más frecuentemente. Sirva un ejemplo para comprender la razón de lo anterior. La palabra "*manicured*" tiene una frecuencia muy baja en la lengua, razón por la cual se le estimará una probabilidad ínfima de que co-aparezca junto a otra palabra, como por ejemplo "*garden*". Por tanto, si la combinación "*manicured garden*" aparece en el corpus simplemente en una ocasión, este hecho supondrá una realidad muy superior a la estimación probabilística establecida y su índice de Información Mutua será muy alto, superponiéndose a índices de combinaciones que aparecen con mucha más frecuencia en la lengua pero que están compuestas de palabras también más frecuentes. Manning y Schütze (1999: 182) lo explican con gran claridad:

> [Mutual Information] is a bad measure of dependence because for dependence the score depends on the frequency of the individual words. Other things being equal, bigrams composed of low-frequency words will receive a higher score than bigrams composed of high-frequency words. That is the opposite of what we would want a good measure to do since higher frequency means more evidence and we would prefer a higher rank for bigrams for whose interestingness we have more evidence.

Resulta evidente que la clase de combinaciones que se pueden obtener aplicando esta medida estadística no son interesantes en modo alguno desde el punto de vista pedagógico, puesto que se da prioridad a las rarezas de la lengua frente a lo frecuente y en consecuencia verdaderamente necesario para el hablante no nativo.

- **IM^3**

Dada la limitación que presenta Información Mutua en cuanto al tipo de combinaciones que destaca, se llevó a cabo una modificación de la fórmula estadística con el fin de dar mayor peso a las combinaciones frecuentes. Para ello, la frecuencia de co-aparición observada de A y B se elevó a todas las potencias desde la segunda hasta la décima. Mediante estos ensayos se demostró empíricamente que la tercera potencia era la que ofrecía los coeficientes más eficaces, por lo que se instauró la Información Mutua cúbica como un nuevo índice de asociación léxica (Oakes, 1998). Sin embargo, como podemos suponer, el hecho de que la frecuencia estimada se siguiera calculando en base únicamente a la frecuencia individual de cada palabra suponía una clara limitación de esta medida.

- ***Log-log***

Al igual que en el caso de IM^3, *Log-log* (Kilgarriff y Tugwell, 2001), es una medida ideada para superar la limitación de IM en cuanto a la relevancia que cobran las combinaciones formadas por palabras poco frecuentes. Así, como una extensión de la fórmula de IM, este índice es multiplicado por el logaritmo de la frecuencia de co-aparición observada de las dos palabras. Con ello, Kilgarriff y Tugwell (ibid.) lograron obtener combinaciones más relevantes para los propósitos lexicográficos que las conseguidas mediante el uso de IM.

El hecho de que la frecuencia individual de cada una de las palabras de una combinación es insuficiente para estimar su frecuencia de aparición conjunta es tan evidente que, como acabamos de ver, ha sido necesario llevar a cabo modificaciones sobre la fórmula original de IM para poder mejorar sus resultados. Los métodos que siguen a continuación, todos derivados de la tradición iniciada con *z-score*, sin embargo, siempre han contado de una u otra forma con información relativa a la aparición conjunta de ambos elementos para calcular su co-aparición estimada.

- ***Z-score***

La medida *z-score*, diseñada por Berry-Roghe en 1972, se puede considerar como la primera incursión en el área de los estudios estadísticos aplicados al análisis de las combinaciones léxicas, dando lugar a una aproximación diferente a la que aportan los análisis derivados de la Teoría de la Información. Al igual que éstos, su objetivo fundamental es medir la diferencia entre la frecuencia observada y la frecuencia estimada en una combinación. Sin embargo, existen claras diferencias entre estos dos enfoques. Básicamente, *z-score* nos indica si la diferencia entre lo observado y lo esperado es estadísticamente significativa o no, es decir, si existe una atracción idiosincrásica entre A y B o si la co-aparición se debe al azar. Para ello, *z-score* realiza en primer lugar una estimación de la frecuencia de co-aparición que cabría esperar si se diera la hipótesis nula, para lo cual no sólo considera la frecuencia individual de cada una de las dos palabras y las relaciona con el tamaño total del corpus como hace IM, sino que también tiene en cuenta la probabilidad de la propia combinación. Más concretamente, *z-score* computa en primer lugar la probabilidad de que B (el colocado) co-aparezca junto a A (la base) si se diera la hipótesis nula, para lo cual se calcula la probabilidad de que B aparezca cuando A no está presente y se multiplica por la frecuencia observada de A (Oakes, 1998). Una vez obtenida de este modo la

frecuencia estimada de co-aparición, se compara con la frecuencia observada para comprobar si existe una diferencia significativa.

Esta medida ha sido utilizada con ciertas modificaciones desde los años 70 hasta la actualidad (Lafon, 1984, citado por Seretan, 2008; Smadja, 1993), si bien a principios de los 90 se desarrolló el índice de *t-score* como una nueva medida de análisis basada en la fórmula de *z-score* pero aportando ciertas mejoras.

- ***T-score***

El objetivo fundamental de *t-score* (medida derivada por Church et al. en 1991 a partir del índice *z-score*), es también el de determinar si la diferencia entre la frecuencia observada y la esperada es significativa. Para realizar este cálculo, se computa la media de la probabilidad de que A y B aparezcan conjuntamente y la media de la probabilidad de aparición independiente de A y B. Teniendo además en cuenta la desviación típica de este cálculo de probabilidad, se obtiene la media de la frecuencia esperada. Al compararla con la observada en el corpus, obtenemos el índice de *t-score*, que será mayor cuanto mayor sea la frecuencia de co-aparición de ambas palabras (Manning y Schütze, 1999). Resulta evidente, así pues, que, a diferencia de lo que ocurría con IM, obtendremos un alto grado de significación estadística en aquellas combinaciones que, además de mostrar una fuerza de asociación notable dada la diferencia entre la frecuencia estimada y la observada, aparezcan también conjuntamente con suficiente frecuencia como para ser fiables.

Como podemos comprobar, esta medida presenta también una clara diferencia con respecto a *z-score* puesto que mientras que *t-score* calcula la frecuencia de co-aparición estimada a partir de la probabilidad de que A y B aparezcan conjuntamente y también la probabilidad de que aparezca cada una de ellas por separado, *z-score* parte de la probabilidad de aparición independiente de una sola

palabra, el colocado, en lugar de las dos que intervienen en el binomio. Esto proporciona, desde nuestro punto de vista, mayor fiabilidad a la medida *t-score*.

Un último aspecto que cabe señalar en cuanto a *t-score* se refiere a las limitaciones que presenta en el sentido de que, al calcular la frecuencia media esperada, asume que las probabilidades cuentan con una distribución casi normal y por tanto puede resultar problemático cuando trabajamos con combinaciones poco frecuentes (Church y Mercer, 1993; Seretan, 2008). A pesar de ello, varios autores han destacado que se trata de una medida adecuada para el estudio y la extracción de combinaciones léxicas en comparación con otras medidas estadísticas (Evert y Krenn, 2001).

- **χ^2 test**

La medida χ^2 surgió como alternativa a la prueba *t-score* ya que aquella, al ser una prueba no paramétrica, no asume probabilidades con una distribución normal y, por tanto, es más fiable cuando se trabaja con probabilidades muy elevadas, es decir, con corpus de gran tamaño (Manning y Schütze, 1999). Este método estadístico, al igual que los ya descritos, pretende fundamentalmente comparar las frecuencias observadas en la lengua con las frecuencias esperadas según la hipótesis nula, con el objetivo de comprobar si ésta se puede rechazar. Para comprender el funcionamiento básico de esta medida debemos considerar las cuatro posibilidades de combinación que existen entre dos palabras en términos probabilísticos:

a) Que A y B aparezcan conjuntamente. Ejemplo: *strong tea*.
b) Que aparezca B pero no junto a A. Ejemplo: *good tea*.
c) Que aparezca A pero no junto a B. Ejemplo: *strong man*.
d) Que no aparezcan ni A ni B. Ejemplo: *good man*.

Estas probabilidades se suelen expresar en una tabla de contingencia de 2x2 del siguiente modo (Tabla 1):

	A = strong	**A ≠ strong**
B = tea	a) Número de casos E.g. *strong tea*	b) Número de casos E.g. *good tea*
B ≠ tea	c) Número de casos E.g. *strong man*	d) Número de casos E.g. *good man*

Tabla 1. Tabla de contingencia de 2x2 para el cálculo de probabilidades (adaptada de Manning y Schütze, ibid.: 169)

A partir de tablas como la mostrada más arriba donde se ofrece la frecuencia observada, es posible calcular la frecuencia estimada de cada una de las celdas, multiplicando los datos correspondientes a los de su fila por los de su columna, aunque convertidos en proporciones dividiendo cada uno por el número total de binomios del corpus, y finalmente multiplicando el producto de nuevo por el total de binomios. De este modo, por ejemplo, se puede calcular la frecuencia estimada de la combinación *strong tea*, que correspondería a la probabilidad marginal de que *strong* aparezca como la primera parte de un binomio [(a+c)/(a+b+c+d)], multiplicada por la probabilidad marginal de que *tea* aparezca como la segunda parte de un binomio [(a+b)/(a+b+c+d)], multiplicado por el número total de binomios del corpus (a+b+c+d) (Manning y Schütze, ibid.).

A pesar de lo somero de la descripción de esta medida estadística, nos parece suficiente para poder comprobar que se trata de una prueba más refinada que las anteriores desde el punto de vista matemático y probabilístico. Sin embargo, consideramos necesario añadir la apreciación de Manning y Schütze (ibid.: 170) acerca de esta medida, donde se asegura que no es más adecuada necesariamente para la extracción de combinaciones léxicas que la prueba *t-score*: "In general, for the problem of finding collocations, the differences between the *t* statistic and the χ^2

statistic do not seem to be large. For example, the 20 bigrams with the highest *t* scores in our corpus are also the 20 bigrams with the highest χ^2 scores".

- **Coeficiente de Verosimilitud**

La última de las medidas estadísticas que trataremos, y una de las más utilizadas en los trabajos más recientes ya que, como apuntamos anteriormente, algunos estudios comparativos han concluido que se trata de la fórmula más robusta y precisa para la obtención de combinaciones léxicas (Anagnostou y Weir, 2006), es el Coeficiente de Verosimilitud, introducido por Dunning (1993). Como descripción fundamental podemos afirmar que este método se utiliza para evaluar una hipótesis dada y, en el caso del estudio de binomios, se emplea para comprobar en qué medida una hipótesis es más probable que la otra. Como ya sabemos, las dos hipótesis que se confrontan en el análisis combinatorio son:

- **Hipótesis 1** (o **Hipótesis nula**): Ambas palabras son independientes y su aparición conjunta se debe al azar. No existe asociación léxica.
- **Hipótesis 2**: Existe una relación idiosincrásica entre ambas palabras y el uso de una predetermina en cierta medida la aparición de la otra, sobre la que ejerce una fuerza de atracción. Existe asociación léxica.

Mediante el uso del Coeficiente de Verosimilitud se establecen unas complejas estimaciones de probabilidad diferentes para cada una de las dos hipótesis y se comparan entre sí.

Esta medida presenta, según Manning y Schütze (ibid.), dos ventajas fundamentales con respecto a las ya mencionadas. En primer lugar, el índice que ofrece como resultado el cálculo del Coeficiente de Verosimilitud es más fácil de interpretar que los valores que aportan Información Mutua, *t-score* y χ^2, dado que éstos no significan nada por sí mismos y no cobran sentido si no consultamos una tabla, mientras que con el Coeficiente podemos entender perfectamente la

información que nos ofrece de modo intuitivo ya que se trata de la probabilidad de que algo suceda, expresada en números enteros.

Por otro lado, ya vimos que χ^2 era especialmente adecuado para trabajar con grandes cantidades de texto, disminuyendo su fiabilidad cuando la muestra es más pequeña. Esta limitación también ha sido superada por el Coeficiente de Verosimilitud, que puede ofrecer estimaciones más precisas con un corpus de menor tamaño. En cualquier caso, dado que en nuestro estudio particular estábamos trabajando con dos corpus de referencia, y por tanto de gran tamaño, este aspecto no suponía un riesgo para nosotros.

Tras la revisión de los métodos estadísticos anteriores y, antes de poder decidir definitivamente cuál de ellos emplearíamos en nuestro trabajo, era necesario considerar qué herramientas informáticas de exploración de corpus podíamos utilizar y comprobar en términos prácticos qué prestaciones ofrecía cada una de ellas en cuanto a medidas estadísticas para la extracción de combinaciones léxicas. Esta fue, pues, nuestra siguiente tarea.

3.2.2.3. Herramientas informáticas

Como es lógico, para poder manejar un corpus de las dimensiones del BNC y del BoE, y más aún poder realizar la extracción de las combinaciones más frecuentes de 803 sustantivos mediante análisis estadísticos fiables, resulta imprescindible contar con la ayuda de programas informáticos de exploración de corpus, también llamados programas de concordancias. Como apunta Partington (2006: html), "a corpus by itself is simply an inert archive. However, it can be 'interrogated' using dedicated software".

En la actualidad existe un buen número de herramientas de este tipo, algunas de ellas de acceso gratuito a través de Internet y otras a las que sólo se puede acceder

previa subscripción. En la Web se pueden encontrar diferentes listados de los programas de concordancias existentes y sus características[7]. Para nuestro propósito concreto, era esencial contar con herramientas que nos permitieran explorar el BNC y el BoE y, por razones obvias, nos interesaba especialmente hallar programas capaces de proporcionar binomios extraídos a partir de análisis estadísticos robustos como los descritos en el apartado anterior.

En primer lugar, realizamos una búsqueda de los programas que existen actualmente para acceder al BNC. Entre ellos cabe distinguir dos variedades: por un lado aquellos que han sido diseñados para explorar exclusivamente el BNC y solamente se pueden utilizar con este corpus, y, por otro lado, los que permiten explorar cualquier corpus que tengamos almacenado en la memoria de nuestro ordenador (bien porque sea de construcción propia o porque se trate de corpus distribuidos comercialmente para ser descargados en nuestro disco duro), entre los cuales se puede encontrar el BNC.

Entre los primeros encontramos programas que se utilizan bien a través de Internet como es el caso de *BNCweb*[8] (Lehmann, Hoffmann y Schneider, 2002) y las herramientas gratuitas *Phrases in English*[9] (Fletcher, 2008) y *BYU-BNC*[10] (Davies, 2008), o mediante un programa descargable en nuestro ordenador como *Xaira*[11] (y su versión más antigua, *Sara*), proporcionado junto con el propio corpus por la Universidad de Oxford. En cuanto a los del segundo tipo debemos destacar por sus muchas prestaciones y lo generalizado de su uso los programas *WordSmith Tools*[12]

[7] http://www.athel.com/corpus_software.html [Último acceso: 30.12.2008]
http://www.fi.muni.cz/~thomas/EAP/concordancers.htm [Último acceso: 30.12.2008]
[8] http://www.bncweb.info/ [Último acceso: 05.01.2009]
[9] http://phrasesinenglish.org/ [Último acceso: 30.12.2008]
[10] http://corpus.byu.edu/bnc/[Último acceso: 30.12.2008]
[11] http://www.oucs.ox.ac.uk/rts/xaira/ [Último acceso: 30.12.2008]
[12] http://www.lexically.net/wordsmith/ [Último acceso: 30.12.2008]

(Scott, 1999) y *Sketch Engine*[13] (Kilgarriff et al., 2004), ambos disponibles previo pago y capaces de proporcionar una información muy completa del corpus que se esté analizando.

Quizá debido en parte al creciente número de herramientas capaces de manejar el BNC que de un modo u otro compiten por ofrecer los mejores servicios, y en parte por la importancia del propio corpus como el banco de datos de mayor tamaño disponible para cualquier investigador, los programas de concordancias mencionados en el párrafo anterior han desarrollado una gran cantidad de prestaciones de enorme calidad, que están permitiendo a los investigadores obtener información cada vez más exhaustiva, fiable e interesante sobre los diversos aspectos del lenguaje. En este sentido, la mayoría de las herramientas citadas arriba (con la única excepción de *Phrases in English* como veremos más adelante) ofrecen hoy en día aplicaciones capaces de extraer concordancias, definidas como "a list of unconnected lines of text that have been summoned by the concordance program from a computer corpus, with the searchword located at the centre of each line. The rest of each line contains the immediate co-text to the left and right of the searchword" (Partington, 2006: html). Sobre estas concordancias, además, los programas pueden aplicar distintos filtros para ordenarlas según nos interese en cada momento o seleccionar las que contengan una determinada información (Fig. 3.1.). Asimismo, estas herramientas también ofrecen listas de palabras frecuentes (Fig. 3.2.), palabras clave donde se destacan aquellas palabras que aparecen en un corpus específico con una frecuencia significativamente mayor que en un corpus de referencia (Fig. 3.3.), combinaciones y paquetes léxicos[14] (Fig. 3.4.), así como información gráfica sobre la dispersión de una palabra concreta dentro de un gran número de textos (Fig. 3.5.)

[13] http://www.sketchengine.co.uk/ [Último acceso: 30.12.2008]

[14] Utilizamos aquí la traducción "paquetes léxicos" para referirnos a lo que Biber et al. (1999) denominan "*lexical bundles*", que más tarde Cortés (2006: 392) define como "combinations of three or more words which are identified empirically in a corpus of natural language".

(Partington, ibid.) Es importante también destacar la calidad de las interfaces de estos programas, cada vez más fáciles de utilizar y con un diseño cada vez más atractivo y cuidado. Como dijimos, la única excepción en este sentido es el programa *Phrases in English*, que se limita a ofrecer combinaciones frecuentes de entre 2 y 8 palabras que recupera de una base de datos donde han sido previamente almacenadas (Fig. 3.6.), y hasta 50 ejemplos de oraciones completas donde aparecen esas combinaciones en el BNC (Fig. 3.7.).

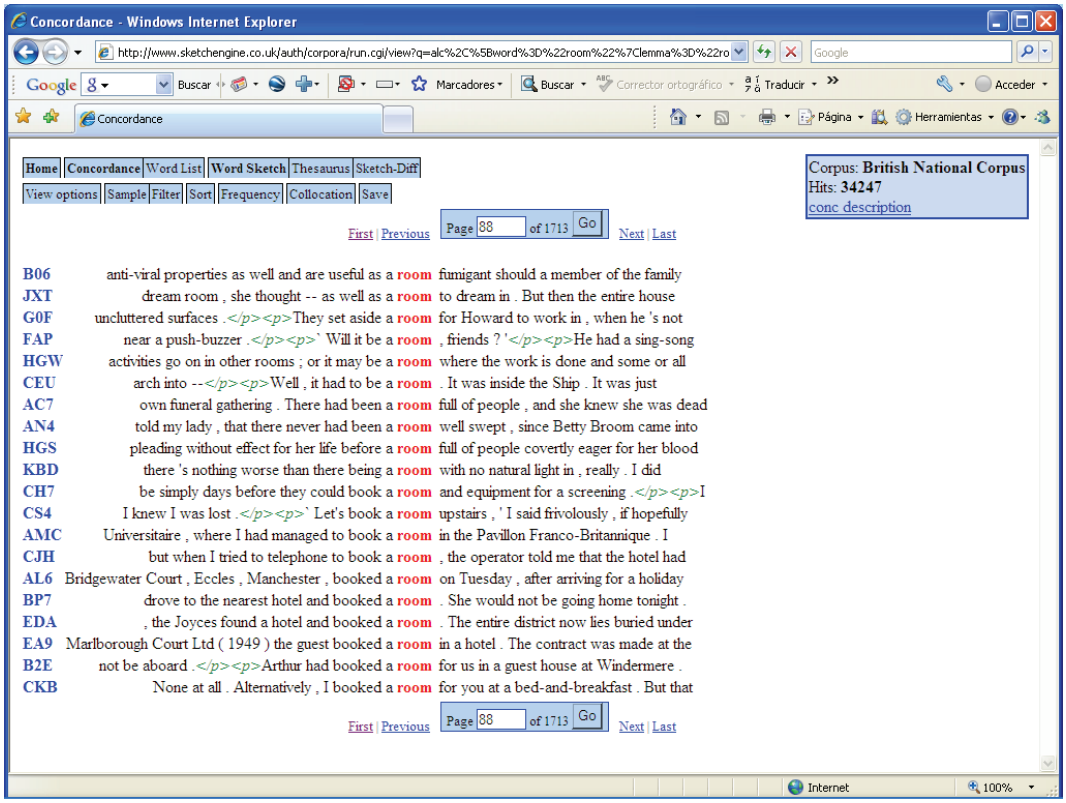
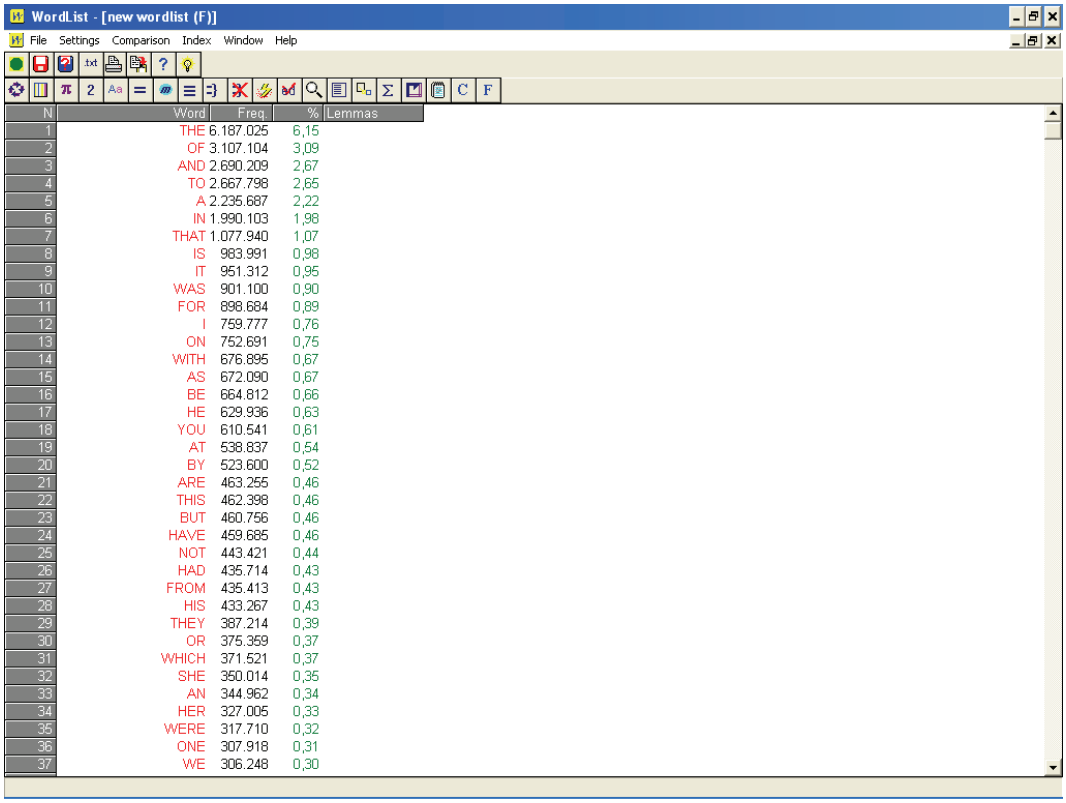
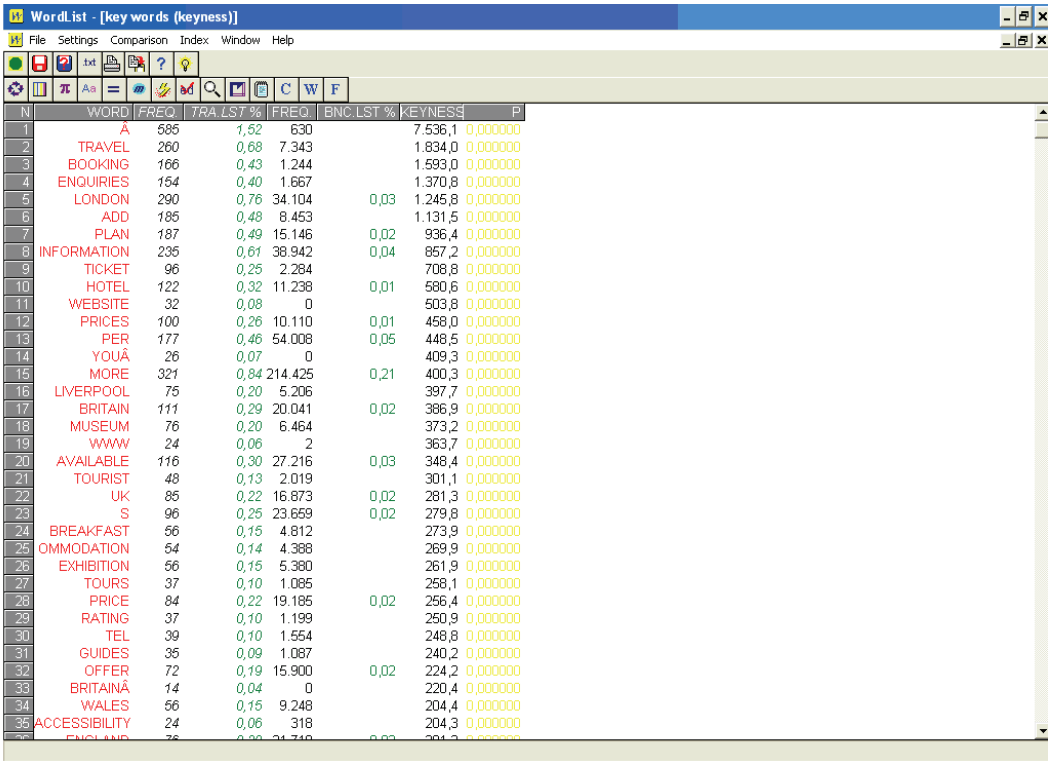


Fig. 3.1.: Concordancias de la palabra *room* obtenidas con *Sketch Engine* donde la palabra inmediatamente anterior está ordenada alfabéticamente



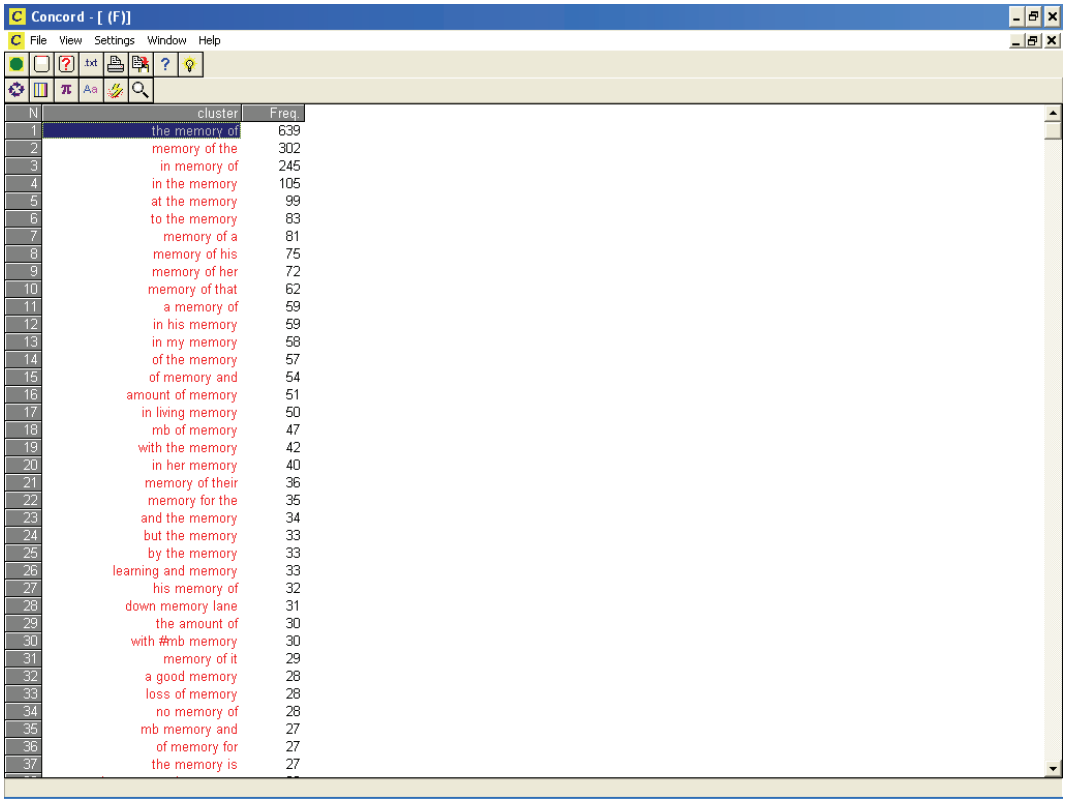
N	Word	Freq.	%	Lemmas
1	THE	6.187.025	6,15	
2	OF	3.107.104	3,09	
3	AND	2.690.209	2,67	
4	TO	2.667.798	2,65	
5	A	2.235.687	2,22	
6	IN	1.990.103	1,98	
7	THAT	1.077.940	1,07	
8	IS	983.991	0,98	
9	IT	951.312	0,95	
10	WAS	901.100	0,90	
11	FOR	898.684	0,89	
12	I	759.777	0,76	
13	ON	752.691	0,75	
14	WITH	676.895	0,67	
15	AS	672.090	0,67	
16	BE	664.812	0,66	
17	HE	629.936	0,63	
18	YOU	610.541	0,61	
19	AT	538.837	0,54	
20	BY	523.600	0,52	
21	ARE	463.255	0,46	
22	THIS	462.398	0,46	
23	BUT	460.756	0,46	
24	HAVE	459.685	0,46	
25	NOT	443.421	0,44	
26	HAD	435.714	0,43	
27	FROM	435.413	0,43	
28	HIS	433.267	0,43	
29	THEY	387.214	0,39	
30	OR	375.359	0,37	
31	WHICH	371.521	0,37	
32	SHE	350.014	0,35	
33	AN	344.962	0,34	
34	HER	327.005	0,33	
35	WERE	317.710	0,32	
36	ONE	307.918	0,31	
37	WE	306.248	0,30	

Fig. 3.2.: Lista de las palabras más frecuentes del BNC obtenida con *WordSmith Tools 3*



N	WORD	FREQ	TRA LST %	FREQ	BNC LST %	KEYNESS	P
1	Â	585	1,52	630		7.536,1	0,000000
2	TRAVEL	260	0,68	7.343		1.834,0	0,000000
3	BOOKING	166	0,43	1.244		1.593,0	0,000000
4	ENQUIRIES	154	0,40	1.667		1.370,8	0,000000
5	LONDON	290	0,76	34.104	0,03	1.245,8	0,000000
6	ADD	185	0,48	8.453		1.131,5	0,000000
7	PLAN	187	0,49	15.146	0,02	936,4	0,000000
8	INFORMATION	235	0,61	38.942	0,04	857,2	0,000000
9	TICKET	96	0,25	2.284		708,8	0,000000
10	HOTEL	122	0,32	11.238	0,01	580,6	0,000000
11	WEBSITE	32	0,08	0		503,8	0,000000
12	PRICES	100	0,26	10.110	0,01	458,0	0,000000
13	PER	177	0,46	54.008	0,05	448,5	0,000000
14	YOUÂ	26	0,07	0		409,3	0,000000
15	MORE	321	0,84	214.425	0,21	400,3	0,000000
16	LIVERPOOL	75	0,20	5.206		397,7	0,000000
17	BRITAIN	111	0,29	20.041	0,02	386,9	0,000000
18	MUSEUM	76	0,20	6.464		373,2	0,000000
19	WWW	24	0,06	2		363,7	0,000000
20	AVAILABLE	116	0,30	27.216	0,03	348,4	0,000000
21	TOURIST	48	0,13	2.019		301,1	0,000000
22	UK	85	0,22	16.873	0,02	281,3	0,000000
23	S	96	0,25	23.659	0,02	279,8	0,000000
24	BREAKFAST	56	0,15	4.812		273,9	0,000000
25	OMMODATION	54	0,14	4.388		269,9	0,000000
26	EXHIBITION	56	0,15	5.380		261,9	0,000000
27	TOURS	37	0,10	1.085		258,1	0,000000
28	PRICE	84	0,22	19.185	0,02	256,4	0,000000
29	RATING	37	0,10	1.199		250,9	0,000000
30	TEL	39	0,10	1.554		248,8	0,000000
31	GUIDES	35	0,09	1.087		240,2	0,000000
32	OFFER	72	0,19	15.900	0,02	224,2	0,000000
33	BRITAINÂ	14	0,04	0		220,4	0,000000
34	WALES	56	0,15	9.248		204,4	0,000000
35	ACCESSIBILITY	24	0,06	318		204,3	0,000000

Fig. 3.3.: Palabras clave de un corpus del área específica del turismo comparado con el BNC, obtenidas con *WordSmith Tools 3*



N	cluster	Freq
1	the memory of	639
2	memory of the	302
3	in memory of	245
4	in the memory	105
5	at the memory	99
6	to the memory	83
7	memory of a	81
8	memory of his	75
9	memory of her	72
10	memory of that	62
11	a memory of	59
12	in his memory	59
13	in my memory	58
14	of the memory	57
15	of memory and	54
16	amount of memory	51
17	in living memory	50
18	mb of memory	47
19	with the memory	42
20	in her memory	40
21	memory of their	36
22	memory for the	35
23	and the memory	34
24	but the memory	33
25	by the memory	33
26	learning and memory	33
27	his memory of	32
28	down memory lane	31
29	the amount of	30
30	with #mb memory	30
31	memory of it	29
32	a good memory	28
33	loss of memory	28
34	no memory of	28
35	mb memory and	27
36	of memory for	27
37	the memory is	27

Fig. 3.4.: Paquetes léxicos de tres palabras con el sustantivo *memory* ordenados por frecuencia, obtenidos con *WordSmith Tools 3*

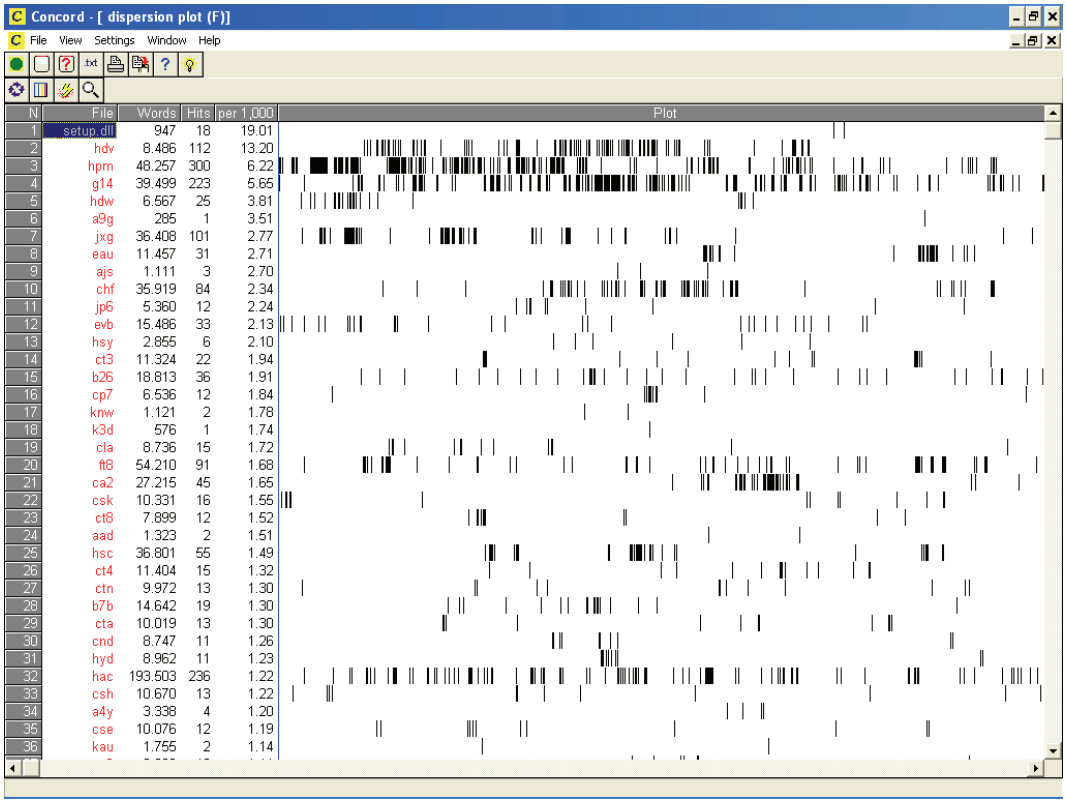


Fig. 3.5.: Dispersión de la palabra *memory* en los diferentes archivos del BNC, obtenida con *WordSmith Tools 3*

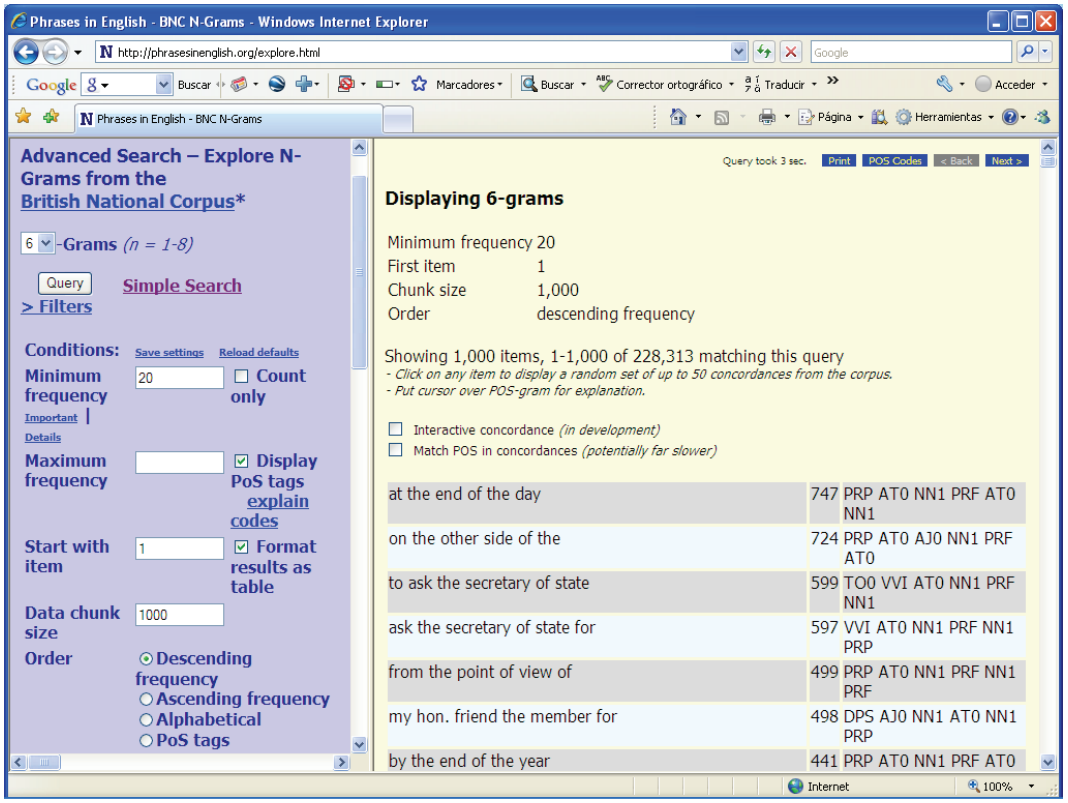


Fig. 3.6.: Paquetes léxicos de 6 palabras ordenados por frecuencia según el BNC, obtenidos con *Phrases in English*

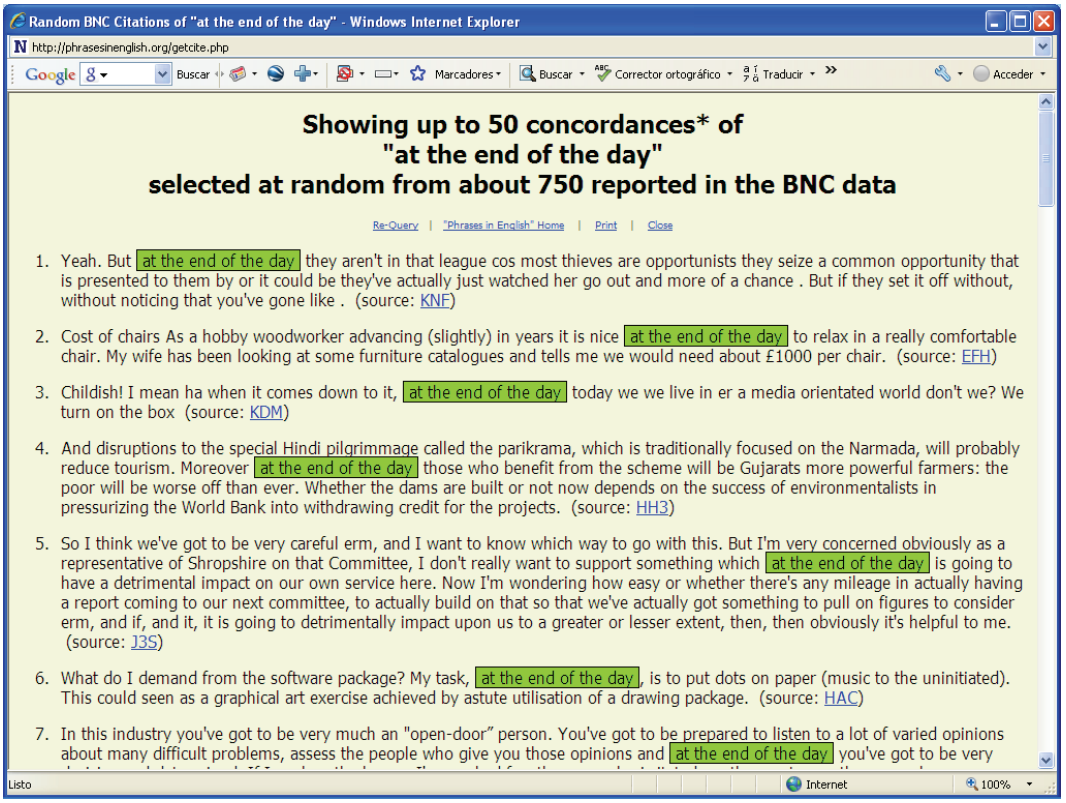


Fig. 3.7.: Oraciones que contienen *at the end of the day*, la combinación de 6 palabras más frecuente del BNC, obtenida con *Phrases in English*

En lo que respecta al BoE, el panorama es marcadamente distinto porque Collins Cobuild sólo permite el acceso y la exploración de este corpus mediante el programa *jLookup*[15], distribuido por la propia editorial previa subscripción y con el cual se trabaja en modo *online*, puesto que no se permite la descarga del corpus en el ordenador del usuario. Además, como ya mencionamos, este programa sólo proporciona acceso a un subcorpus de 56 millones de palabras extraído del corpus original. Si comparamos este programa con las herramientas disponibles para trabajar con el BNC, resulta evidente que *jLookup* es una aplicación más limitada, que no ofrece tantas prestaciones como los programas anteriormente descritos. En concreto,

[15] http://www.collins.co.uk/books.aspx?group=154 [Último acceso: 30.12.2008]

esta herramienta tan solo ofrece las concordancias de la palabra de búsqueda, con la posibilidad de aplicarle diferentes filtros, y un listado de sus combinaciones léxicas[16]. A pesar de estas limitaciones, ya que nos parecía necesario utilizar el BoE en nuestro estudio y *jLookup* nos ofrecía información suficiente para completar nuestro listado de colocaciones, consideramos oportuno utilizar esta herramienta.

En lo que respecta a la aplicación dedicada a la extracción de binomios, el programa *jLookup* ofrece al usuario la posibilidad de utilizar únicamente las medidas estadísticas de Información Mutua o *t-score* por lo que, a pesar de que el Coeficiente de Verosimilitud es la medida más empleada en los estudios actuales, nuestro trabajo debía utilizar una de estas dos fórmulas. Teniendo en cuenta las características de cada uno de ellas mencionadas más arriba, evidentemente decidimos emplear el índice de *t-score* para elaborar nuestro listado. Para poder complementar esta lista con las combinaciones obtenidas a partir del BNC, debíamos pues encontrar un programa que nos permitiera extraer binomios de este corpus basados también en la medida *t-score*.

Tras una exhaustiva exploración de las herramientas disponibles para el estudio del BNC (a excepción del programa *Phrases in English* dado que no emplea medidas estadísticas para la obtención de combinaciones), observamos que, como cabía esperar, cada una utilizaba fórmulas diferentes. En la siguiente tabla (Tabla 2), podemos observar de forma gráfica las distintas medidas de asociación que cada programa pone a disposición del usuario:

[16] En un mensaje enviado el día 16 de diciembre de 2008 por *Collins Cobuild* a través de la lista de distribución *Corpora List* se puso en conocimiento de la comunidad investigadora que la herramienta de exploración del BoE va a ser mejorada en sus distintas aplicaciones y se proporcionará acceso a un corpus etiquetado casi 10 veces mayor que el subcorpus de 56 millones que se ofrece actualmente. Se prevé que esta nueva interfaz esté disponible en la primavera de 2009.

	IM	**IM³**	***Log-log***	***z-score***	***t-score***	**χ^2**	**Coef de Veros.**
Xaira				√			
BNCweb	√	√		√	√	√	√
BYU-BNC	√						
WordSmith Tools 4.0	√	√		√			√
Sketch Engine	√	√	√		√		√

Tabla 3.2.: Comparativa de medidas de asociación de los programas de exploración del BNC

A la luz de la información que observamos en la tabla anterior, comprobamos que los dos únicos programas que podíamos utilizar para extraer los colocados del BNC eran *BNCweb* y *Sketch Engine*, puesto que son los únicos que ofrecen entre las distintas fórmulas estadísticas el índice de *t-score*. Ambos programas presentan unas características muy similares, con prestaciones muy completas e interfaces de muy alta calidad, y ambas se utilizan a través de Internet con una clave de acceso. Sin embargo, como ya mencionamos, *Sketch Engine* no está restringido exclusivamente a la exploración del BNC, sino que también permite acceder a una buena cantidad de corpus diferentes, entre ellos algunos de dimensiones tan gigantescas como el corpus ukWaC, extraído de páginas web y que cuenta con más de 1.500 millones de palabras. Puesto que esto suponía una ventaja añadida de este programa que nos serviría más adelante para la elaboración de algunos de los ítems de nuestro test que necesitaban de oraciones contextualizadas, decidimos emplear *Sketch Engine* en nuestro trabajo.

Así pues, una vez seleccionados *jLookup* y *Sketch Engine* como las herramientas a utilizar, comenzamos el proceso de extracción de combinaciones que pasamos a describir de forma pormenorizada.

3.2.2.4. Primera fase: Extracción automatizada

Como hemos venido comentando a lo largo de este capítulo, nuestro objetivo era recopilar un listado de las colocaciones de los nombres más frecuentes del inglés, a partir de los datos aportados por el BNC y BoE. Para hallar los colocados de los sustantivos previamente seleccionados, comenzamos con la exploración del BoE mediante *jLookup* puesto que, como veremos, su herramienta de extracción de colocados imponía más limitaciones que la de *Sketch Engine*, de uso más flexible y, por tanto, necesitábamos partir de las prestaciones que ofrecía *jLookup* para luego poder establecer las mismas condiciones al usar *Sketch Engine*.

3.2.2.4.1. Extracción a partir del BoE

Para llevar a cabo la extracción de los colocados con *jLookup*, el proceso seguido fue el siguiente: en primer lugar, introducíamos el sustantivo que servía de base de búsqueda para extraer sus concordancias. Puesto que en no pocas ocasiones el sustantivo adopta en inglés la misma forma que el verbo, delimitamos la categoría gramatical de la palabra para que seleccionara únicamente los casos en que la palabra actuaba como sustantivo y especificamos si se trataba de una forma singular o plural (recordemos que buscamos los colocados de los sustantivos en singular y en plural por separado puesto que las dos formas dan lugar a colocaciones distintas en muchos casos). En el programa *jLookup* estas especificaciones se añaden mediante el uso de dos etiquetas distintas: /NN si se trata de un nombre singular y /NNS si buscamos

una forma plural. En las imágenes siguientes ofrecemos un ejemplo de la búsqueda del sustantivo *work* en su forma singular (Fig. 3.8.) y de las concordancias que ofrece el programa al introducir este término de búsqueda (Fig. 3.9.).

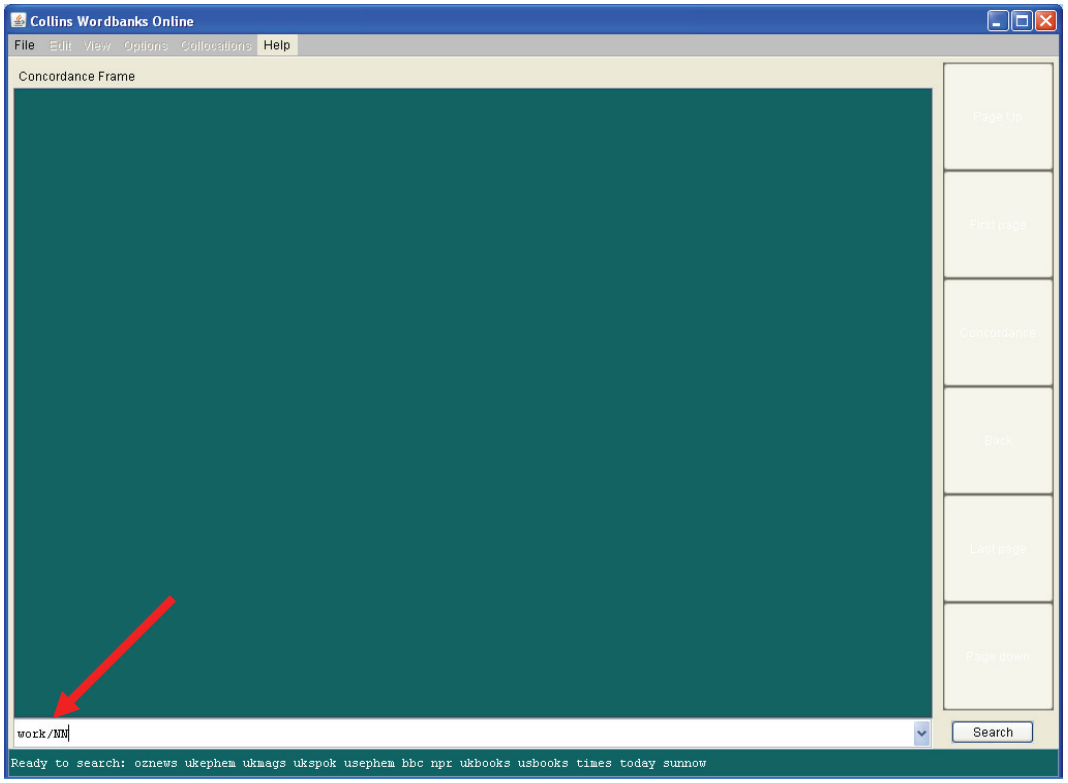


Fig. 3.8.: Búsqueda de las concordancias de *work* como sustantivo singular con *jLookup*

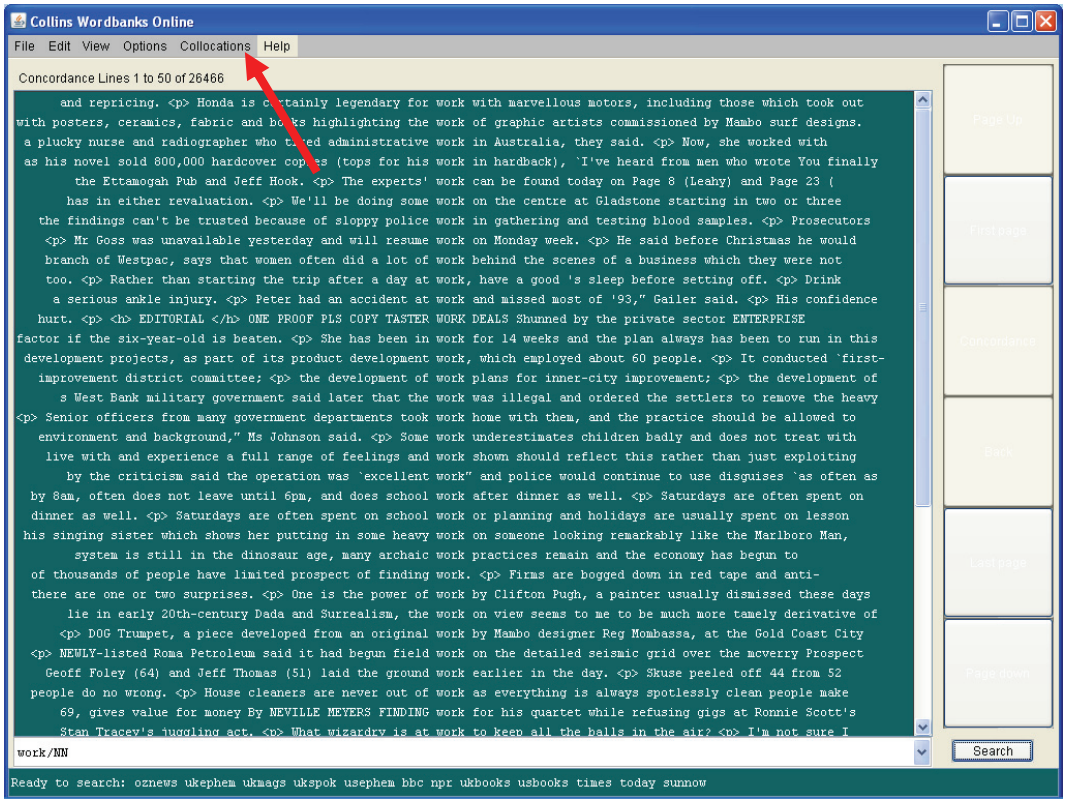


Fig. 3.9.: Concordancias de *work* obtenidas con *jLookup*

Como podemos observar en la Figura 9, al ofrecer las concordancias del sustantivo el programa activa diferentes aplicaciones en la barra de herramientas superior, entre las cuales se encuentra la dedicada a la extracción de binomios, llamada "Collocations". Esta aplicación ofrece dos alternativas a la hora de presentar la información relativa a las combinaciones: o bien mediante la herramienta "Collocation grid" o mediante "Collocation list" (Fig. 3.10.).

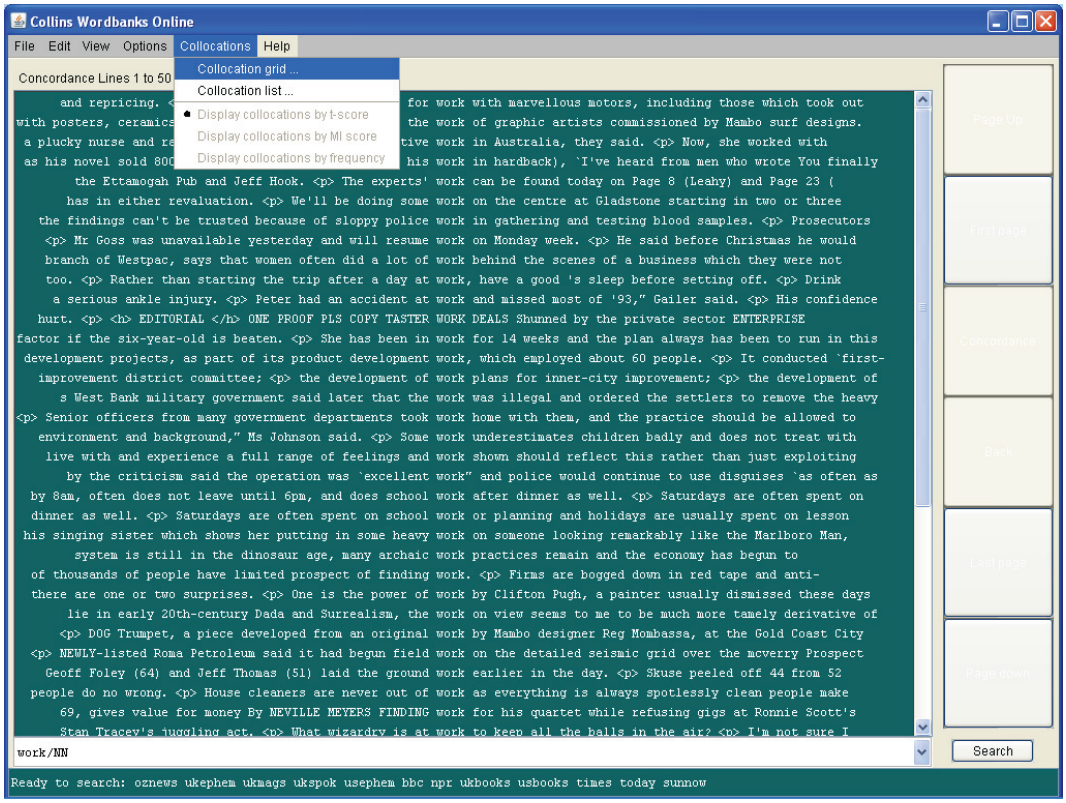
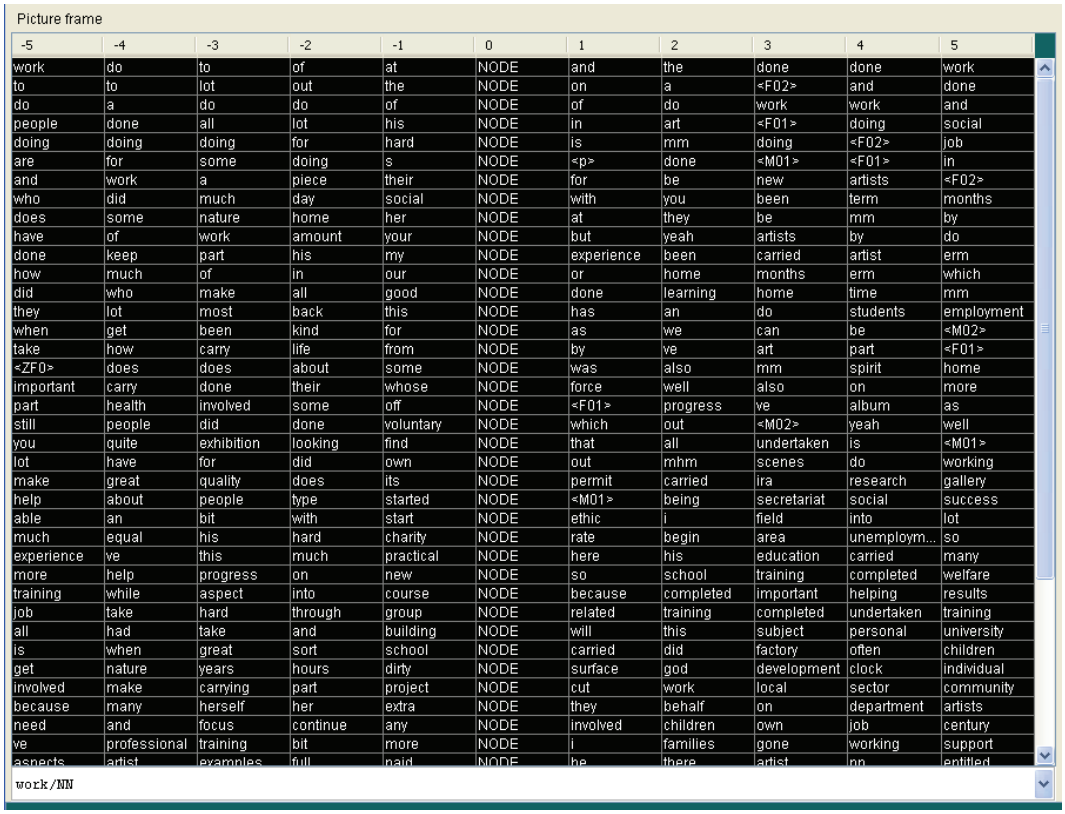


Fig. 3.10.: "Collocation grid" en *jLookup*

"Collocation grid" es una aplicación capaz de crear una tabla donde se distribuyen los colocados que más frecuentemente acompañan al sustantivo atendiendo a la posición que suelen ocupar en la oración con respecto a éste. Esta tabla, con los colocados más significativos del nombre según la medida *t-score* que se encuentran entre 2 y 5 posiciones a ambos lados del sustantivo (según las preferencias del usuario), especifica así la distancia que suele haber entre la base y cada colocado de la tabla y si suele aparecer a la derecha o a la izquierda de la base. En la figura 3.11. podemos observar la tabla con los colocados de *work* donde se han seleccionado 5 posiciones a cada uno de sus lados.


Picture frame

-5	-4	-3	-2	-1	0	1	2	3	4	5
work	do	to	of	at	NODE	and	the	done	done	work
to	to	lot	out	the	NODE	on	a	<F02>	and	done
do	a	do	do	of	NODE	of	do	work	work	and
people	done	all	lot	his	NODE	in	art	<F01>	doing	social
doing	doing	doing	for	hard	NODE	is	mm	doing	<F02>	job
are	for	some	doing	s	NODE	<p>	done	<M01>	<F01>	in
and	work	a	piece	their	NODE	for	be	new	artists	<F02>
who	did	much	day	social	NODE	with	you	been	term	months
does	some	nature	home	her	NODE	at	they	be	mm	by
have	of	work	amount	your	NODE	but	yeah	artists	by	do
done	keep	part	his	my	NODE	experience	been	carried	artist	erm
how	much	of	in	our	NODE	or	home	months	erm	which
did	who	make	all	good	NODE	done	learning	home	time	mm
they	lot	most	back	this	NODE	has	an	do	students	employment
when	get	been	kind	for	NODE	as	we	can	be	<M02>
take	how	carry	life	from	NODE	by	ve	art	part	<F01>
<ZF0>	does	does	about	some	NODE	was	also	mm	spirit	home
important	carry	done	their	whose	NODE	force	well	also	on	more
part	health	involved	some	off	NODE	<F01>	progress	ve	album	as
still	people	did	done	voluntary	NODE	which	out	<M02>	yeah	well
you	quite	exhibition	looking	find	NODE	that	all	undertaken	is	<M01>
lot	have	for	did	own	NODE	out	mhm	scenes	do	working
make	great	quality	does	its	NODE	permit	carried	ira	research	gallery
help	about	people	type	started	NODE	<M01>	being	secretariat	social	success
able	an	bit	with	start	NODE	ethic	i	field	into	lot
much	equal	his	hard	charity	NODE	rate	begin	area	unemploym...	so
experience	ve	this	much	practical	NODE	here	his	education	carried	many
more	help	progress	on	new	NODE	so	school	training	completed	welfare
training	while	aspect	into	course	NODE	because	completed	important	helping	results
job	take	hard	through	group	NODE	related	training	completed	undertaken	training
all	had	take	and	building	NODE	will	this	subject	personal	university
is	when	great	sort	school	NODE	carried	did	factory	often	children
get	nature	years	hours	dirty	NODE	surface	god	development	clock	individual
involved	make	carrying	part	project	NODE	cut	work	local	sector	community
because	many	herself	her	extra	NODE	they	behalf	on	department	artists
need	and	focus	continue	any	NODE	involved	children	own	job	century
ve	professional	training	bit	more	NODE	i	families	gone	working	support
aspects	artist	examples	full	paid	NODE	he	there	artist	no	entitled

work/NN

Fig. 3.11.: Tabla de los colocados de *work* en 5 posiciones a derecha e izquierda

Por otro lado, la segunda opción para la extracción de combinaciones es "Collocation list", que ofrece un único listado donde se ordenan por frecuencia todos los colocados estadísticamente significativos del sustantivo que aparecen dentro del margen de 10 posiciones, 5 a la derecha y 5 a la izquierda, de la base, en lo que se denomina "*span*" en la bibliografía anglosajona (ver sección 1.3.1.1.2.) (Fig. 3.12.). Ya que esta herramienta, como vemos, ofrecía en una sola lista y de forma más clara todos los colocados frecuentes de nuestros sustantivos considerando un espacio de 5 posiciones a cada lado (lo cual se ajustaba perfectamente a nuestro constructo de colocación según el cual los dos elementos que la integran no tienen que ser siempre necesariamente contiguos), decidimos utilizar esta aplicación para nuestra compilación. Con respecto al margen de 5 posiciones que contempla "Collocation

list" cabe añadir que, si bien es cierto que está predeterminado por el programa y no se puede alterar, se enmarca dentro de las recomendaciones de varios autores a este propósito, mencionadas en el primer capítulo de este trabajo (Sinclair, Jones y Daley, 1970/2004; Martin, Al y Van Sterkenburg, 1983; Smadja, 1993; van der Wouden, 2002). Por este motivo, consideramos que la inflexibilidad de *jLookup* en este sentido no supone un problema.

A la hora de extraer los colocados, este programa calcula tanto el índice de *t-score* como el de IM de cada uno de ellos, pero permite elegir la medida estadística mediante la cual preferimos que se ordenen los colocados en el listado. Así, como observamos en la figura 12, "Collocation list" crea una lista con 5 columnas: en la primera presenta los colocados del sustantivo, en la segunda muestra la frecuencia absoluta del colocado en el corpus, en la tercera muestra la frecuencia de co-aparición del colocado y la base, en la cuarta se ofrece el índice de *t-score* y en la quinta encontramos el índice de IM. En nuestro caso, como ya adelantamos, decidimos ordenar el listado mediante la medida *t-score* y, como se puede comprobar en la imagen, éste se ordena de forma descendente situando los colocados con mayor índice de *t-score* en las primeras posiciones de la lista.

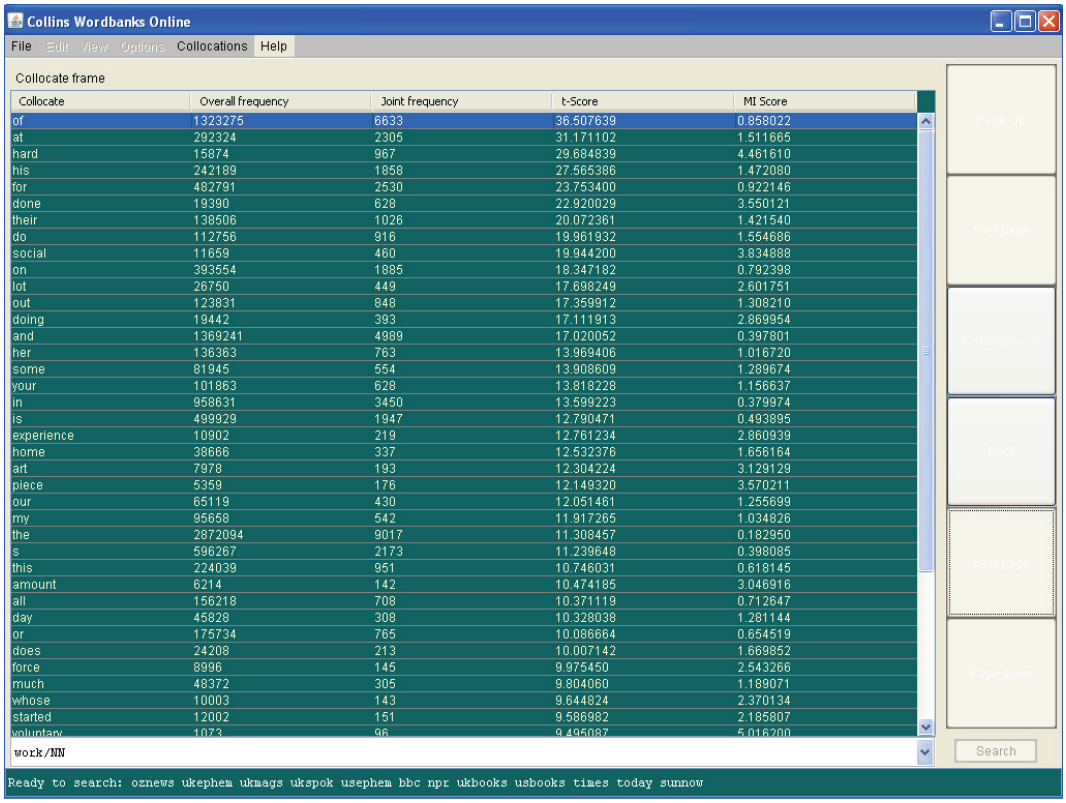


Fig. 3.12.: Listado de los colocados del sustantivo *work* ordenados según *t-score*

A la hora de realizar la compilación de los colocados surgieron algunas otras cuestiones que nos llevaron a establecer una serie de criterios prácticos. En primer lugar, consideramos necesario establecer un punto de corte en cuanto a la cantidad de colocados que se podían extraer para cada sustantivo ya que, por un lado, no era factible compilar docenas de colocados por cada una de las bases y, por otro, necesitábamos colocaciones verdaderamente frecuentes y significativas que fueran operativas para el diseño de un test, fin último de esta investigación.

Según indican McEnery, Xiao y Tono (2006), un *t-score* de 2 o superior se suele considerar estadísticamente significativo en los estudios de asociación léxica. Sin embargo, tras realizar nuestras primeras búsquedas pudimos observar que al establecer el índice mínimo de *t-score* en 2, el número de co-ocurrencias que

obteníamos era tan elevado que se hacía prácticamente inmanejable y, lo que resultaba más preocupante, un número muy elevado de las combinaciones extraídas resultaban poco relevantes desde el punto de vista pedagógico. Podríamos pensar que esta circunstancia se debía al hecho de que se trataba de nuestras primeras búsquedas y, por tanto, de los sustantivos más frecuentes del inglés. Para comprobar si esta era la razón de que obtuviéramos tal cantidad de combinaciones y algunas tan poco interesantes para nuestro propósito, decidimos realizar un pequeño estudio explorando los colocados obtenidos a partir de 30 sustantivos de nuestra lista elegidos al azar, y comprobar entre qué márgenes aproximados se suelen encontrar los índices de *t-score* de las combinaciones más interesantes desde el punto de vista pedagógico. Prestando especial atención a la parte baja de los listados, es decir, a los índices mínimos por debajo de los cuales siempre solían aparecer combinaciones propias de lenguajes especializados o técnicos y expresiones más raras y menos útiles para el alumno, observamos que en la mayor parte de los casos éstas aparecían a partir de un *t-score* 6, un índice bastante alejado de 2, lo que confirmaba nuestras observaciones iniciales. Pudimos comprobar, asimismo, que deteniendo nuestra búsqueda al alcanzar un *t-score* 6 había un margen suficiente como para obtener una amplia cantidad de combinaciones interesantes, pero no tantas como para que se escapasen a nuestra capacidad de procesamiento. Así, esta medida que es, a todas luces, notablemente alta y por tanto exigente, se estableció como punto de corte para nuestra extracción.

Cabe mencionar en este punto que un aspecto importante a tener en cuenta cuando se trabaja con la aplicación "Collocation list" de *jLookup* es que el número máximo de colocados que contiene la lista que genera es de 50, sin que exista la posibilidad de acceder a un listado mayor. En algunas ocasiones, se dio la circunstancia de que los 50 colocados que nos ofrecía el programa se encontraban por encima del 6 en su índice de *t-score* por lo que nos era imposible explorar todos

los colocados hasta alcanzar nuestro punto de corte. En cualquier caso, 50 nos parece un número más que suficiente de colocados a partir de los cuales poder extraer los que eran de interés para este trabajo.

Nos referimos a los colocados de interés ya que, como recordaremos, este estudio se dedicó exclusivamente a los colocados nominales, adjetivales y verbales de los sustantivos. Sin embargo, si observamos la lista extraída por el programa para el sustantivo *work* (Fig. 12), podemos comprobar que el programa realiza una extracción automática de todas las palabras que acompañan al nombre de forma significativa, sin ofrecerle al usuario la posibilidad de delimitar la búsqueda a las categorías gramaticales que le interesan en particular. Como no todos los colocados que aparecían en la información ofrecida por el programa eran pues relevantes para nuestro estudio, en esta primera fase del proceso de compilación de nuestro listado hubimos de llevar a cabo una selección basada en criterios estrictamente sintácticos, extrayendo de forma manual exclusivamente aquellos colocados que eran nombres, adjetivos o verbos para su inclusión en nuestra base de datos y descartando el resto. Debemos recordar asimismo que las combinaciones de tipo V+N no implican necesariamente que se trate de un verbo seguido de un complemento directo. Así, aquellos casos en los que la co-aparición de un verbo y un nombre daba lugar a estructuras de verbo y complemento circunstancial también se consideraron interesantes para nuestros propósitos. Por ello, elementos tales como *take (someone) to court*, *get into/out of bed*, *put to the test* o *jumped at the chance* fueron incluidos en nuestro listado.

Como resulta también evidente a la vista de la figura 12, el programa crea su listado en base a las diferentes inflexiones de las palabras, es decir, de forma deslematizada. Así, como vemos en la imagen, en la sexta posición de la lista encontramos el colocado *done*, dos puestos más abajo vemos *do*, en la posición decimotercera aparece *doing* y en el lugar número 33 se encuentra *does*. A la hora de

incluirlos en nuestra propia lista, sin embargo, decidimos recopilar únicamente la primera de las formas verbales o adjetivales que aparecían. Consideramos que, aunque desde un punto de vista puramente lingüístico puede resultar ciertamente interesante comprobar las diferencias existentes entre el comportamiento colocacional de las distintas formas de un lema, un listado concebido desde una aproximación pedagógica no necesita contemplar estas sutilezas ya que, utilizando el ejemplo anterior, una vez que el alumno conozca la colocación "*do* + *work*", éste será capaz de extrapolar este conocimiento y emplearlo con las distintas inflexiones del verbo, dando lugar en todos los casos a diferentes realizaciones de la misma colocación. Cuestión distinta es, sin embargo, el caso de los sustantivos que, como ya explicamos, forman la base de la colocación y por tanto pueden dar lugar a colocaciones distintas en su forma singular y plural, razón por la cual se hacía necesario considerarlas como dos términos de búsqueda diferentes.

Por último, nos gustaría mencionar que en el proceso de recopilación de nuestra lista, y más concretamente a la hora de separar las combinaciones N+N, A+N, V+N y N+V del resto, se hizo necesario en numerosas ocasiones consultar las concordancias originales de las que el programa había extraído los colocados. La herramienta *jLookup* ofrece esta posibilidad, lo cual suponía una gran ventaja en aquellos casos en que la relación sintáctica y a veces incluso semántica entre el sustantivo y un determinado colocado de la lista no resultaban claras si no se observaban dentro de un contexto. A modo de ejemplo, este fue el caso de la combinación *hour long* que, evidentemente no se encuadraba dentro de los tipos de colocaciones recogidos en este trabajo. Sin embargo, si observamos la palabra *long* dentro de una lista descontextualizada de colocados de la palabra *hour* (Fig. 3.13.), podemos suponer erróneamente que se trata de la combinación *long hour*, que sí corresponde a la estructura adjetivo-nombre que buscábamos. Fue necesario en todos estos casos comprobar el tipo de oraciones en que se utiliza cada combinación

accediendo a las concordancias del corpus para asegurarnos de que se trataba de una construcción válida.

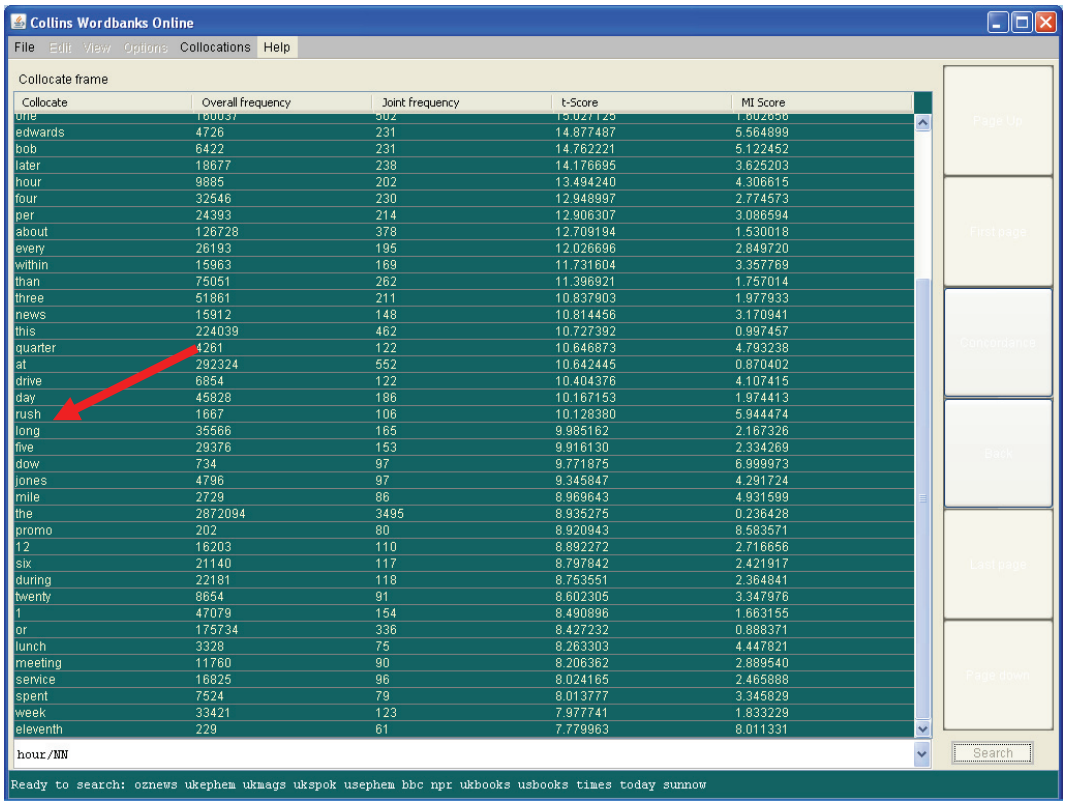


Fig. 3.13.: *Long* entre los colocados de *hour* en la colocación *hour long*

3.2.2.4.2. Extracción a partir del BNC

El proceso de obtención de las combinaciones sintácticas N+N, A+N, V+N y N+V a partir del BoE se simultaneó con el del BNC, dando lugar así al listado completo de los colocados de cada sustantivo según ambos corpus. Como cabe suponer, a la hora de extraer los colocados con *Sketch Engine* se puso especial cuidado en usar los

mismos procedimientos y criterios que se habían establecido para el BoE, con el fin de que ambos fueran equiparables y se pudieran combinar sin ningún inconveniente.

Así pues, en la extracción de combinaciones con *Sketch Engine*, introducíamos la base de la colocación en el programa para obtener en primer lugar sus concordancias, restringiendo el término de búsqueda a la categoría gramatical de sustantivo y especificando si se trataba de una forma singular o plural, del mismo modo que habíamos hecho con *jLookup* (Fig. 3.14.). Una vez generadas las concordancias del nombre en cuestión (en nuestro ejemplo, la palabra *work*), el programa ofrece la posibilidad de buscar sus colocados mediante la aplicación "Collocation" (Fig. 3.15.).

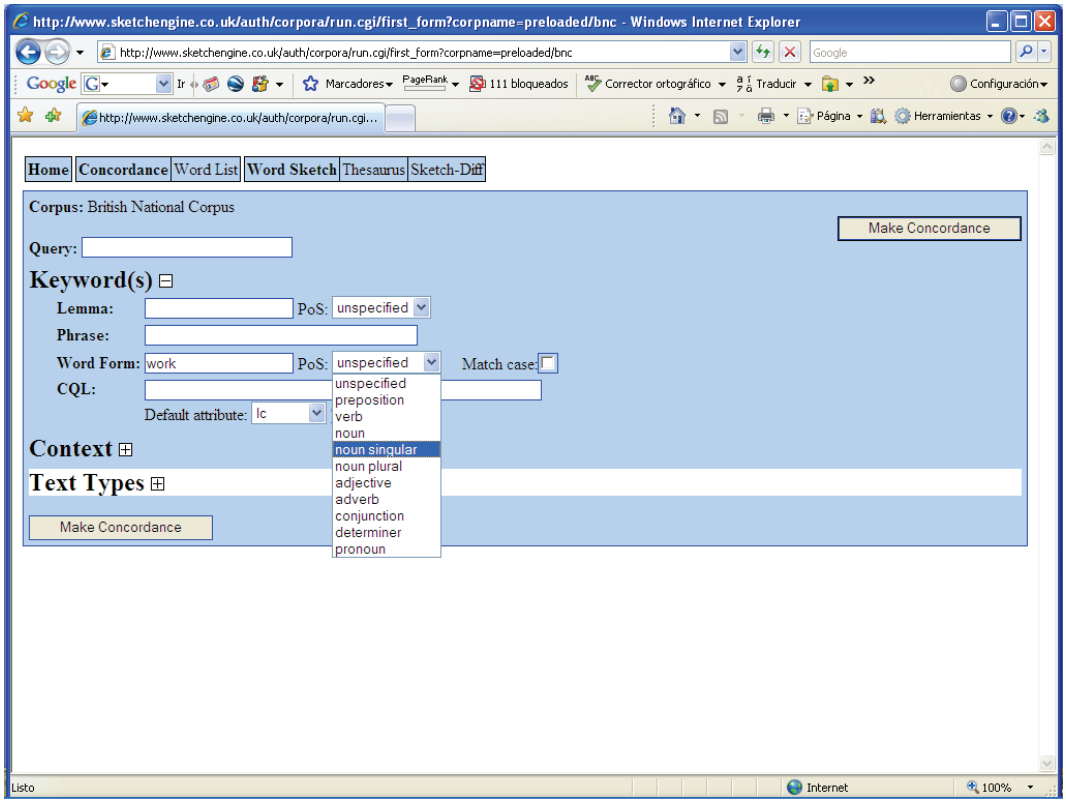


Fig. 3.14.: Búsqueda de las concordancias de *work* como nombre singular con *Sketch Engine*

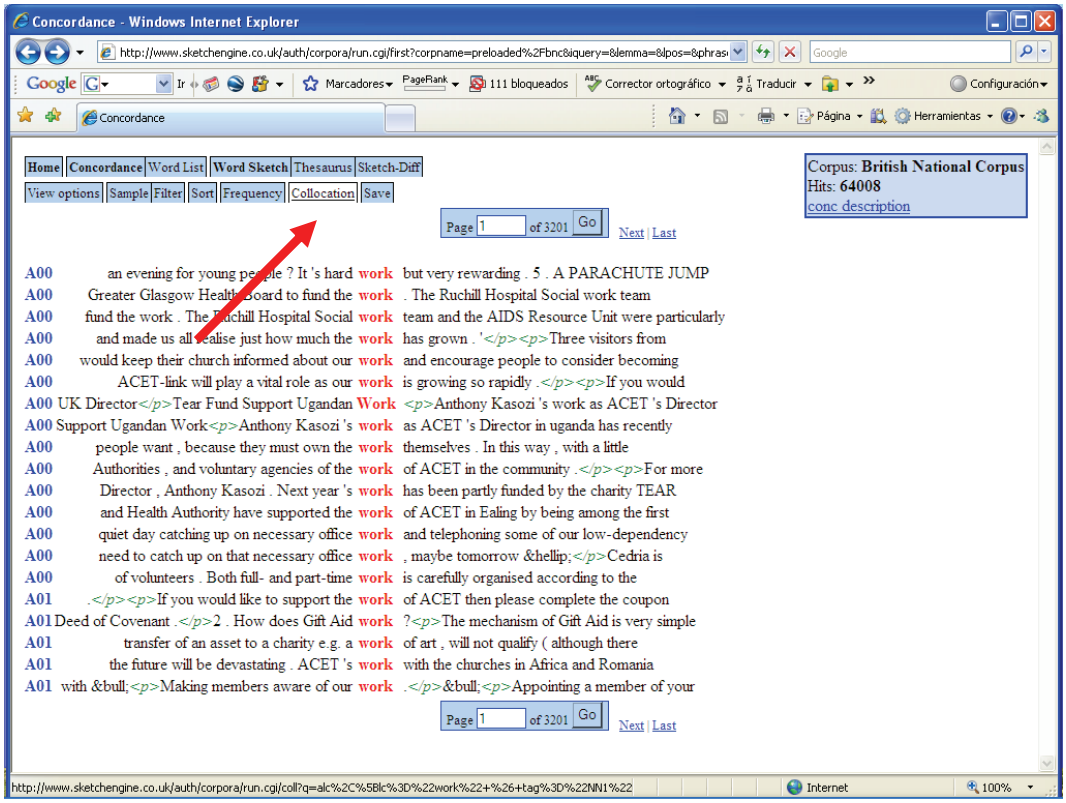


Fig. 3.15.: Concordancias de *work* en *Sketch Engine* y herramienta "Collocation"

La herramienta "Collocation" de *Sketch Engine* ofrece un mayor margen de maniobra al usuario que el que existe en *jLookup*. Como podemos comprobar en la figura 3.16., esta aplicación permite, en primer lugar, seleccionar el tipo de colocado que queremos extraer (bien una palabra deslematizada, un lema con o sin especificación de su categoría gramatical, o bien una categoría gramatical como conjunto sin especificar las palabras concretas que la componen). En nuestro caso, teniendo en cuenta que *jLookup* ofrecía por defecto colocados deslematizados, decidimos elegir esta opción, expresada con el atributo "*word*" en el programa (Fig. 3.16.).

En lo que respecta al número de posiciones que el programa considera a cada lado de la base para extraer los colocados frecuentes, *Sketch Engine* también es más flexible que *jLookup* puesto que permite al usuario elegir cuántas posiciones se desean

incluir. En nuestro caso, seleccionamos 5 palabras a cada lado del sustantivo para equiparar igualmente esta condición con la que ofrece *jLookup* (Fig. 3.16.).

Otro aspecto en el que era esencial hacer coincidir ambos programas por razones evidentes era la medida estadística empleada. Como vemos en la imagen (Fig. 3.16.), se seleccionó la fórmula de *t-score* de entre las que ofrece *Sketch Engine* para que los colocados extraídos se ordenaran de acuerdo con esta medida de asociación.

Por último, debemos también mencionar otras 3 opciones que el programa ofrece al usuario para realizar su búsqueda. La primera de ellas trata de la frecuencia mínima individual con que el colocado debe aparecer en el corpus. En segundo lugar, se establece la frecuencia mínima de co-aparición exigible entre el colocado y la base y la tercera se refiere al número de colocados que deseamos que ofrezca el programa. En lo que respecta a los dos primeros parámetros, decidimos conservar los valores que *Sketch Engine* estipula por defecto, 5 como frecuencia mínima del colocado en el corpus y 3 como frecuencia mínima de co-aparición, dado que no pudimos obtener información alguna sobre estos aspectos en el programa *jLookup* y no podíamos arriesgarnos, por tanto, a establecer unos valores más altos que implicaran un desequilibrio entre el nivel de exigencia impuesto a cada programa. En cuanto al tercer factor, considerando que el número predeterminado de colocados que ofrece *jLookup*, 50 concretamente, resultaba en ocasiones insuficiente para alcanzar el punto de corte de 6 en el índice de *t-score*, decidimos aprovechar la oportunidad que nos ofrecía *Sketch Engine* para aumentar el listado y establecimos que el programa extrajera 250 colocados de cada sustantivo (Fig. 3.16.).

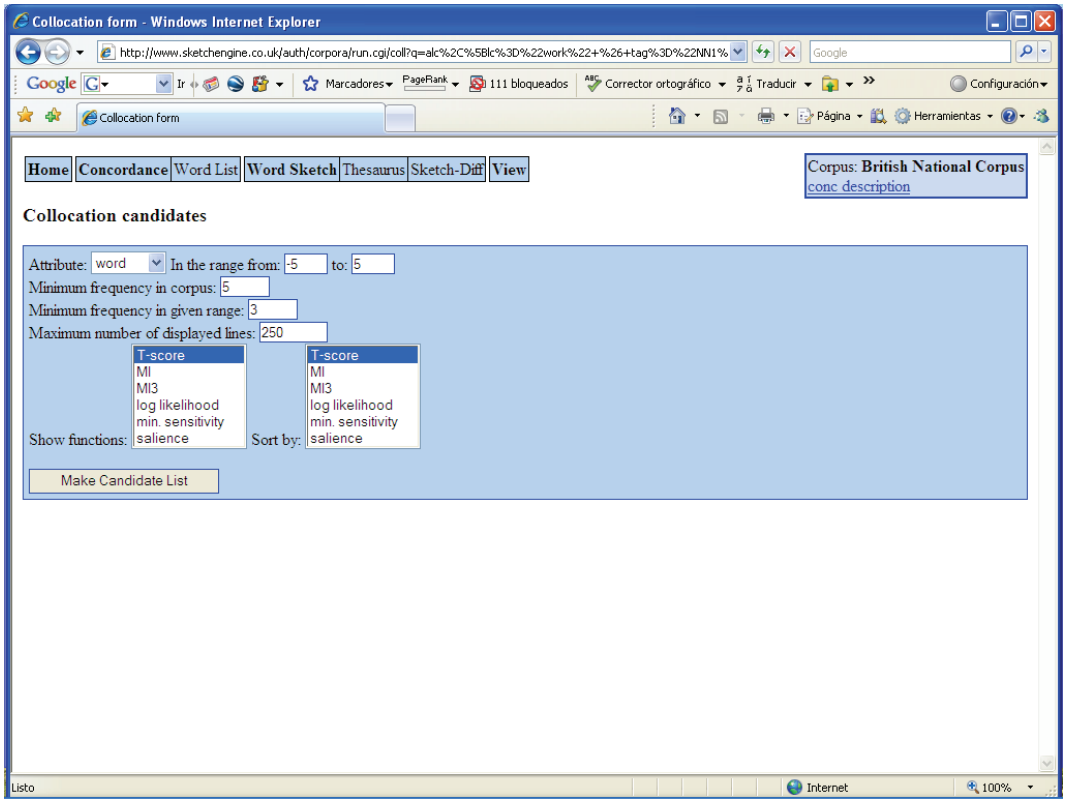


Fig. 3.16.: Opciones para la extracción de colocados de *Sketch Engine*

Tras introducir en el programa todas las especificaciones anteriores, obteníamos un listado de los colocados más frecuentes de nuestros sustantivos según los datos del BNC (Fig. 3.17.). Al igual que sucedía con *jLookup*, *Sketch Engine* generaba un listado de todos los colocados de una palabra entre los que aparecían elementos de todas las categorías gramaticales. Fue necesario pues filtrar también en este caso los colocados adjetivales, verbales y nominales para incluirlos en nuestra base de datos inicial, desestimando el resto de unidades léxicas que aparecían en el listado informático.

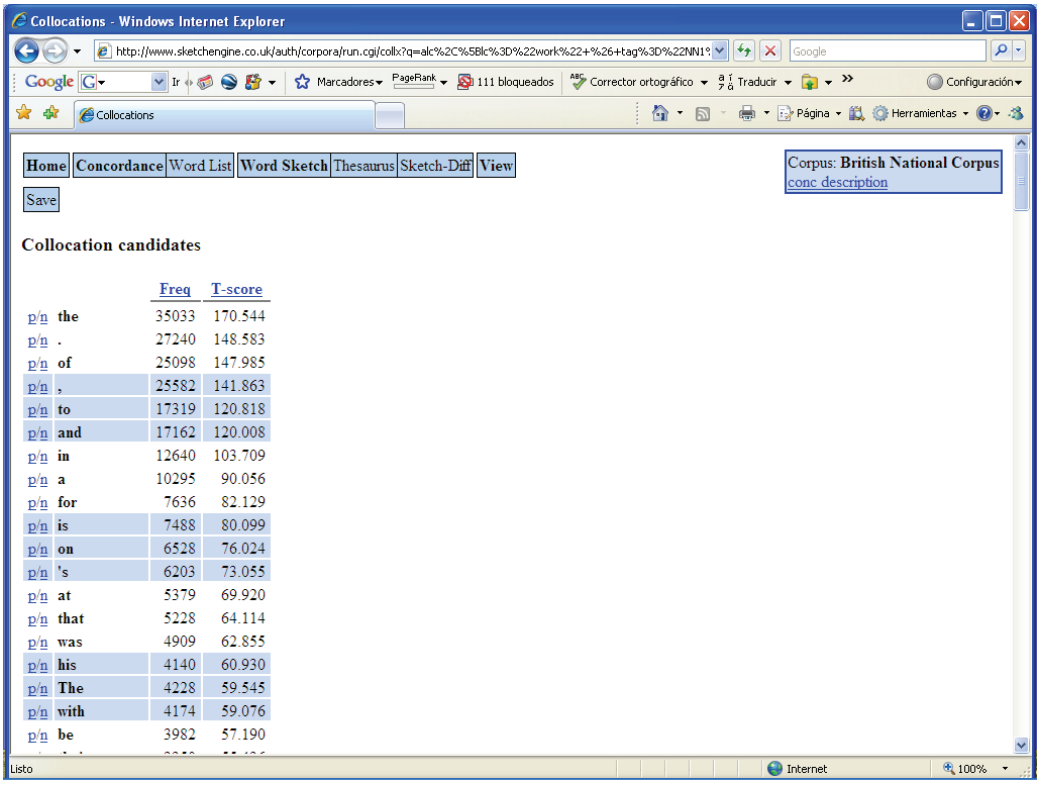


Fig. 3.17.: Lista de colocados de *work* obtenida con *Sketch Engine*

3.2.2.4.3. *jLookup* y *Sketch Engine*: comparación

Si observamos con detenimiento una tabla comparativa con los primeros 25 colocados extraídos por ambos programas a partir del sustantivo *work* (Tabla 3.3.), podemos comprobar que existen ciertas diferencias entre ellos. Como vemos, la lista de colocados de *jLookup* contiene un mayor número de palabras léxicas en sus puestos más altos, mientras que en *Sketch Engine* la primera palabra léxica, *is*, no aparece hasta el lugar número 10 del listado. De hecho, mientras que en los 25 colocados más frecuentes de *work* según *jLookup* aparecen 10 palabras léxicas, en *Sketch Engine* solamente aparecen 4, de las cuales 3 son distintas formas del verbo *to be* y 1 es la forma *has*, todas ellas palabras que también pueden ejercer de verbos

auxiliares, siendo en este supuesto palabras funcionales en lugar de léxicas (por ejemplo en las oraciones *work* ***was*** *imposed on the prisoners* o *work* ***has*** *been done*[17]). Esta diferencia, que en nuestra opinión puede ser atribuible al hecho de que se trata de dos corpus diferentes tanto por su tamaño como por su contenido y composición, no implica en modo alguno que la lista obtenida a partir del BoE fuera más informativa que la generada desde el BNC. Muy al contrario, ya que *Sketch Engine* nos ofrecía hasta 250 colocados de cada sustantivo, solía darse la circunstancia de que este listado presentara más combinaciones válidas (es decir nominales, adjetivales y verbales) para nuestro trabajo dentro del límite de *t-score* 6 que las que podíamos obtener con *jLookup*. En justicia a éste, sin embargo, debemos destacar que a pesar de ofrecer únicamente 50 colocados, éstos eran en la mayoría de los casos palabras léxicas que conformaban combinaciones relevantes para nuestro listado, ya que no se "desaprovechaban" puestos de la lista con palabras funcionales ni signos de puntuación, que sí eran incluidos por *Sketch Engine*.

La segunda diferencia que nos llama la atención se refiere a la disparidad que existe entre los índices de *t-score* que arrojan ambos programas. Si comparamos, por ejemplo, el primer colocado de cada una de las listas, veremos que mientras que el índice más alto en *jLookup* es de 36,507 para la preposición *of*, en el caso de *Sketch Engine* encontramos que *the* aparece con un *t-score* de 170,544. De nuevo, opinamos que esta clara diferencia se debe al hecho de que se están manejando dos corpus de tamaños tan marcadamente distintos que uno, el BNC, posee prácticamente el doble de palabras que el otro, lo cual afecta irremediablemente a los resultados de los cálculos estadísticos. Con respecto a esta cuestión, debemos señalar que la diferencia entre los índices numéricos producidos no suponía un inconveniente para la compilación de nuestro propio listado, dado que, como sucede en la mayoría de los estudios que utilizan un índice estadístico para establecer si la asociación léxica es

[17] Ambos ejemplos extraídos del BNC.

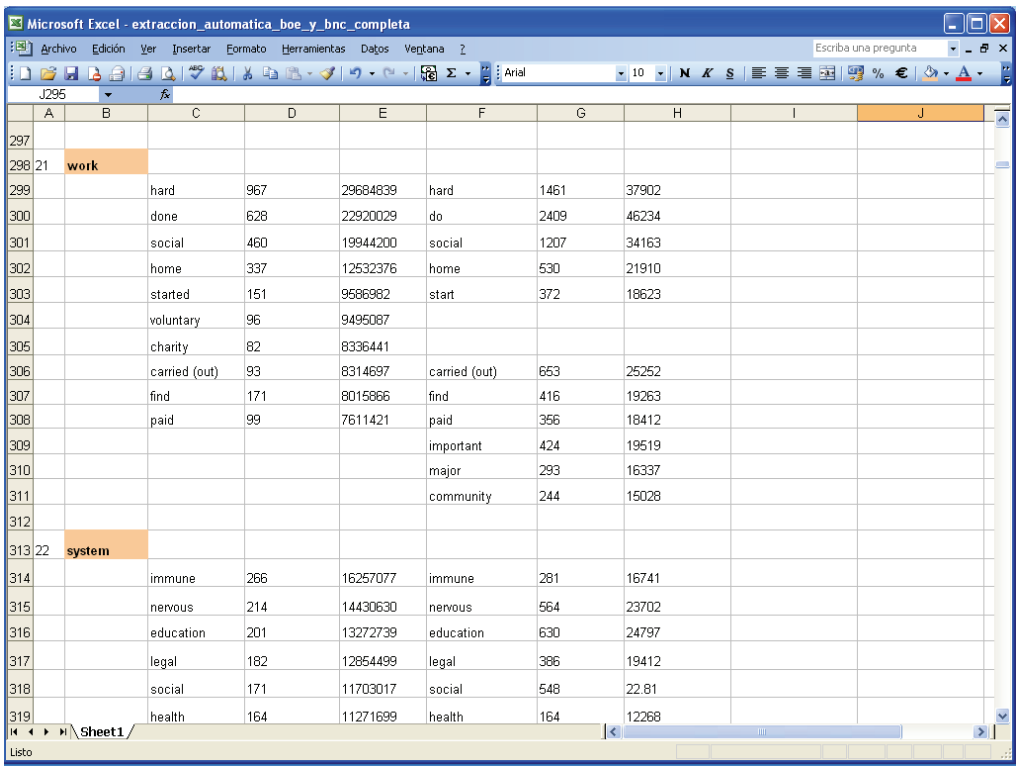
significativa (Manning y Schütze, 1999), los datos obtenidos se emplearon únicamente para establecer un orden de frecuencia o ranking en nuestro listado, no por su propio valor numérico. Por esta misma razón tampoco fue necesario en nuestro caso normalizar y hacer así proporcionales los resultados obtenidos a partir de ambos corpus a pesar de su diferente tamaño.

***jLookup* (BoE)**		***Sketch Engine* (BNC)**	
Colocados	***t-score***	**Colocados**	***t-score***
of	36,507	the	170,544
at	31,171	.	148,583
hard	29,684	of	147,985
his	27,565	,	141,863
for	23,753	to	120,818
done	22,920	and	120,008
their	20,072	in	103,709
do	19,961	a	90,056
social	19,944	for	82,129
on	18,347	is	80,099
lot	17,698	on	76,024
out	17,359	‘s	73,055
doing	17,111	at	69,920
and	17,020	that	64,114
her	13,969	was	62,855
some	13,908	his	60,930
your	13,818	The	59,545
in	13,599	with	59,076
is	12,790	be	57,190

experience	12,761	their	55,436
home	12,532	by	52,086
art	12,304	I	50,794
piece	12,149	this	50,743
our	12,051	‘	50,606
my	11,917	has	46,797

Tabla 3.3.: Primeros 25 colocados de *work* extraídos con *jLookup* y *Sketch Engine*

Una vez finalizado el proceso de búsqueda automatizada mediante el procedimiento descrito, obtuvimos como primer resultado una base de datos de 4.284 palabras diferentes que co-aparecen de forma frecuente con nuestros sustantivos (Fig. 3.18.).



	A	B	C	D	E	F	G	H	I	J
297										
298	21	**work**								
299			hard	967	29684839	hard	1461	37902		
300			done	628	22920029	do	2409	46234		
301			social	460	19944200	social	1207	34163		
302			home	337	12532376	home	530	21910		
303			started	151	9586982	start	372	18623		
304			voluntary	96	9495087					
305			charity	82	8336441					
306			carried (out)	93	8314697	carried (out)	653	25252		
307			find	171	8015866	find	416	19263		
308			paid	99	7611421	paid	356	18412		
309						important	424	19519		
310						major	293	16337		
311						community	244	15028		
312										
313	22	**system**								
314			immune	266	16257077	immune	281	16741		
315			nervous	214	14430630	nervous	564	23702		
316			education	201	13272739	education	630	24797		
317			legal	182	12854499	legal	386	19412		
318			social	171	11703017	social	548	22.81		
319			health	164	11271699	health	164	12268		

Fig. 18: Primer listado de sustantivos, adjetivos y verbos que acompañan frecuentemente a *work*, obtenido mediante extracción automatizada y selección sintáctica

Como vemos, para cada uno de nuestros sustantivos incluimos la información dividida en 6 columnas, donde las 3 primeras contienen los datos extraídos del BoE y las 3 siguientes corresponden a los del BNC. Así, en las columnas 1 y 4 introducimos sólo los nombres, adjetivos y verbos que formaban combinaciones N+N, A+N, V+N y N+V según los listados generados por *jLookup* y *Sketch Engine* (descartando construcciones como por ejemplo *work experience* o *work of art*), en las columnas 2 y 5 aparece la frecuencia de co-aparición de la combinación en el BoE y el BNC, y en las columnas 3 y 6 se ofrece el índice de *t-score*, también según cada corpus[18]. Dado que el primer corpus que consultamos fue el BoE, decidimos ordenar los colocados según el *t-score* ofrecido por este corpus (columna 3), mientras que los del BNC se incluyeron de forma paralela a los anteriores para que la comparación y las diferencias entre ambos corpus resultasen más evidentes. En el ejemplo de la figura 18 vemos que con esta disposición es sencillo comprobar que ambos corpus coinciden en la mayoría de los colocados que presentan para *work*, excepto en el caso de *voluntary* y *charity* que sólo aparecen en el BoE, e *important*, *major* y *community* que sólo se encuentran en el BNC.

Otro aspecto importante que debemos aclarar con respecto a la compilación de nuestra base de datos se refiere a las formas plurales de los sustantivos de búsqueda. Como se puede comprobar en la figura 3.18., no aparece el nombre *work* en plural y desde su forma singular pasamos al sustantivo siguiente (*system*). Esto se debe a que, a pesar de que se realizó la búsqueda y extracción automatizada de colocados para todos y cada uno de los sustantivos en sus formas singular y plural

[18] Como se observa en la imagen y en el Apéndice 1, los índices de *t-score* incluidos en nuestra base de datos no aparecen como una cifra con decimales sino como un número entero. Esto se debe a que en el proceso de copiar los datos desde *jLookup* y *Sketch Engine* a nuestro archivo de Excel, el sistema eliminaba el punto que indica el comienzo de las cifras decimales en inglés. Por tanto, para interpretar los datos correctamente debemos tener en cuenta que *jLookup* ofrece 6 cifras decimales (por lo que en el caso de la primera combinación, *hard work*, el *t-score* en el BoE es 29,684839) y *Sketch Engine* ofrece 3 decimales (*hard work* en el BNC tiene un *t-score* de 37,902).

(sumando un total de 803 términos de búsqueda), en aquellos casos en que la forma plural no presentaba ningún colocado distinto a los que ya se habían extraído para el singular, se decidió no incluirlos en nuestra base de datos. Tras completar el proceso de obtención automatizada de la lista, ésta contaba con un total de 461 sustantivos, resultado de sumar las 412 formas singulares y los 49 plurales que aportaban colocados novedosos y que por tanto se añadieron a la lista. En la tabla siguiente (Tabla 3.4.) podemos comprobar cuáles fueron concretamente los 49 plurales añadidos, especificando el lugar que ocupan en el listado de sustantivos.

5. ways	137. figures
8. days	146. ages
13. parts	156. forces
15. lives	158. conditions
17. cases	164. patients
30. hands	189. feet
40. ends	208. parents
44. families	221. materials
46. facts	226. sales
52. sides	246. details
54. nights	257. charges
56. eyes	284. benefits
64. changes	288. elections
66. interests	291. standards
81. needs	320. thoughts
83. effects	331. principles
85. uses	364. resources
87. ideas	377. funds
91. jobs	384. glasses
93. names	426. circumstances
97. friends	442. affairs
99. rights	445. respects
107. hours	450. goods
119. terms	459. memories
121. sorts	

Tabla 3.4.: Sustantivos plurales incluidos en la base de datos

Esta base de datos, sin embargo, presentaba un problema fundamental desde nuestro punto de vista porque contenía un número considerable de elementos que, a pesar de aparecer junto a nuestros nombres con una frecuencia estadísticamente significativa, no daban lugar a colocaciones tal y como se conciben en esta tesis. Así pues, tras concluir esta primera extracción automatizada de colocados potenciales, se consideró necesario llevar a cabo una selección manual de aquellas combinaciones que se correspondían con nuestro constructo de colocación y que nos parecían interesantes desde el punto de vista pedagógico.

3.2.2.5. Segunda fase: Selección manual

Teniendo en cuenta que el presente trabajo toma como punto de partida fundamental una definición fraseológica del concepto de colocación, en la que, como ya vimos en el primer capítulo, se contemplan aspectos relacionados tanto con la idiosincrasia y la arbitrariedad de la lengua como con su enseñanza/aprendizaje y evaluación, no resulta extraño comprobar que los procesos puramente computacionales de selección de combinaciones léxicas nos parezcan insuficientes para producir un listado fiable y eficaz. Por supuesto, esto no significa que las herramientas y técnicas de extracción informatizada de binomios no tengan un papel fundamental en la creación de un listado de frecuencias. Muy al contrario, nos parece que son recursos esenciales a la hora de determinar qué combinaciones son verdadera y significativamente frecuentes en la lengua real que los hablantes nativos usan en su día a día, prueba de lo cual es el hecho de que esta metodología de obtención de combinaciones supuso el punto de partida de nuestra investigación. Sin embargo, una vez obtenidas las combinaciones más frecuentes de nuestros sustantivos, cumpliendo con ello el criterio de frecuencia e institucionalización que define a las colocaciones (ver sección 1.3.2.2.3), consideramos que existen todavía

limitaciones en los mecanismos estadísticos tanto en lo que se refiere a la diferenciación de los distintos tipos de unidades multi-léxicas establecidos en base a su naturaleza semántica y a las restricciones arbitrarias de conmutación (hoy en día los ordenadores no son todavía capaces de distinguir entre una colocación, una combinación libre de palabras, una locución o un nombre compuesto), como en cuanto a las necesidades del hablante no nativo (para el cual no todo lo frecuente es necesariamente interesante en materia colocacional). Como afirma Nation (2001: 56),

> [t]he availability of large corpora (...) has helped research on collocation considerably. However, the research can only be done to a certain point by computer and then researcher judgement and analysis must be used. Studies of collocation which have relied solely on computing procedures have yielded results which are not very useful.

Así pues, por el momento, el único modo de superar estas limitaciones parece ser el análisis manual.

En nuestro caso en particular, pudimos comprobar que en el listado de nombres, adjetivos y verbos que habíamos compilado con la ayuda de herramientas informáticas existían combinaciones de naturaleza y características muy diversas que no se correspondían con nuestra definición de colocación. Por otro lado, también observamos que la lista contaba con colocaciones propias de lenguajes especializados que no respondían a las necesidades de un listado pedagógico para la evaluación del inglés general y que, por tanto, también debíamos descartar. Esto nos llevó a iniciar un proceso de selección manual para identificar las combinaciones que nos interesaban y rechazar las demás.

Para llevar a cabo esta tarea de discriminación, no exenta de una enorme complejidad dado lo difuso de las fronteras entre las colocaciones y el resto de combinaciones léxicas, establecimos una serie de criterios fundamentados en nuestra definición de colocación y en nuestras necesidades concretas. Así, todas las

combinaciones léxicas susceptibles de formar colocaciones interesantes para nuestro estudio debían cumplir los siguientes criterios:

a. Arbitrariedad de la combinación

Con el objetivo de descartar las estructuras que formaban combinaciones libres de palabras, establecimos como criterio que todos los binomios de nuestro listado debían caracterizarse por un cierto grado de arbitrariedad en su combinación. En consonancia con nuestra aproximación teórica (capítulo 1), la arbitrariedad podía darse en dos sentidos:

1) Combinabilidad restringida del colocado: Cuando el colocado está restringido a la base y no acompaña libremente a cualquier sustantivo de la lengua sin ceñirse a más limitación que la que imponen las leyes de la semántica y la lógica del lenguaje. De acuerdo con este criterio, combinaciones tales como *disabled people*, *elderly woman* o *a car crashed* debían aparecer en nuestro listado definitivo puesto que los colocados no se usan en combinación con otros sustantivos posibles: **disabled dog*, **elderly tree* o **a person crashed*. En cambio, binomios como *read a book*, *next month* o *important question* se descartaron por ser combinaciones libres.

2) Conmutabilidad restringida del colocado: Cuando la base tiende a ir acompañada de colocados por los que muestra especial preferencia formando combinaciones institucionalizadas en la lengua, mientras que otros sinónimos o expresiones posibles que podrían acompañar a la base para expresar una noción similar no se suelen usar nunca o se hace con muy poca frecuencia. Así, *full advantage*, *uncertain future* o *immediate answer* forman

colocaciones que debíamos aceptar en nuestro listado puesto que son expresiones institucionalizadas en la lengua frente a alternativas como **total advantage*, **doubtful future* o **hasty answer* que no se suelen utilizar.

b. "Composicionalidad" semántica

Otro aspecto importante a la hora de seleccionar nuestras colocaciones fue su "composicionalidad" semántica, dado que nos permitía discernir entre éstas y las locuciones. Como ya vimos, la principal diferencia entre estos dos tipos de unidades fraseológicas radica en el hecho de que las colocaciones tienen un significado transparente y su valor semántico se construye a partir de los significados individuales de cada una de las unidades léxicas que las componen, mientras que las locuciones forman una unidad semántica indivisible no composicional, es decir, que no se corresponde con la suma de los significados de las palabras que la integran. Teniendo estos conceptos en mente, establecimos como criterio que todas aquellas combinaciones seleccionadas debían ser semánticamente composicionales, descartando así las que no lo fueran. Siguiendo este criterio descubrimos una serie de locuciones que habían sido incluidas en nuestra base de datos y que debían ser eliminadas como, por ejemplo, *bottom line*, *brave face* (que corresponde a la expresión *to put on a brave face*), *to pull someone's leg* o *blind eye* (de la locución *to turn a blind eye*), por nombrar algunas de ellas.

Debemos mencionar también a este respecto que en nuestra lista existían casos en los que una combinación podía tener una interpretación literal, composicional, a la vez que otra no composicional. Este era el caso por ejemplo de binomios como *wash your hands* o *rub your hands*. Se decidió incluirlos en el listado puesto que su uso literal, es decir, como colocaciones, era frecuente y por tanto debían formar parte de nuestro banco de datos.

c. Estructura interna y valor semántico

Observando nuestra base de datos pudimos comprobar que otro tipo de combinaciones léxicas que aparecía con mucha frecuencia eran las palabras compuestas, algo que no debe extrañar si consideramos que la composición es un recurso morfológico muy frecuente en inglés (Katamba, 1994) y en el cual intervienen las construcciones sintácticas que nosotros habíamos seleccionado para nuestro listado, destacando especialmente las de N+N y A+N. No cabe duda de que los compuestos conforman un tipo de unidad léxica muy interesante desde el punto de vista de la enseñanza de lenguas, y quizá si cabe de manera más especial en el caso del inglés debido a la alta frecuencia de palabras de este tipo que contiene. Sin embargo, con el fin de ajustarnos estrictamente a nuestro objetivo inicial de estudiar el área de las colocaciones en su aspecto más prístino, decidimos seleccionar únicamente este tipo de combinaciones para su inclusión en nuestra lista final, descartando los compuestos.

Para poder establecer la diferencia entre las colocaciones y las palabras compuestas utilizamos como criterio fundamental su estructura interna y valor semántico. Como ya vimos en el capítulo 1, las colocaciones son combinaciones donde concurren dos miembros diferenciados que conservan sus características léxicas y su carga semántica individual. Por el contrario, los compuestos presentan una estructura interna mucho más fija y compacta que se refleja tanto en su fonología, dado que en una gran mayoría de casos los compuestos totalmente establecidos en la lengua sólo se acentúan en una de las palabras integrantes, como en su valor semántico ya que designan un solo concepto. Son combinaciones, pues, que a todos los efectos funcionan como una única palabra. Utilizando así este criterio seleccionamos aquellas combinaciones que formaban colocaciones en nuestro listado, frente a los compuestos que el procedimiento automatizado había incluido en

un primer momento. Algunos de los compuestos que pudimos identificar y descartar fueron *fast food*, *welfare state*, *ironing board* o *personal computer*.

Finalmente, nos gustaría hacer constar que en la práctica este proceso resultó complicado en algunas ocasiones debido a que la línea divisoria entre estos dos tipos de construcciones no es siempre fácil de establecer. Sirvan como ejemplo casos como *single parent*, *gift shop*, *general election* o *interior design*, donde resulta evidentemente difícil discernir dónde acaba la construcción de un significado composicional y empieza la creación de un único concepto, y donde, desde nuestra perspectiva, sería igualmente aceptable considerarlos como colocaciones que como nombres compuestos. En el caso concreto de estas cuatro combinaciones se optó por recogerlas en nuestro listado de colocaciones ya que consideramos que aún se puede percibir la carga semántica literal que cada elemento conserva y que se trata de la asociación de dos conceptos diferentes, a pesar de que se encuentran en un proceso avanzado de creación de una globalidad semántica y, por tanto, de fijación léxica.

d. Colocaciones de uso general

Finalmente, el último criterio que establecimos en esta selección manual se refería al tipo de colocaciones que queríamos incluir en nuestro listado para otorgarle la orientación pedagógica que deseábamos. En este sentido, teniendo en cuenta que nuestra intención era diseñar un test de colocaciones que presentan un uso frecuente en el inglés general, nuestro banco debía estar formado por este tipo de combinaciones, descartando así aquellas que están restringidas a lenguajes especializados y jergas de distintos campos del saber como por ejemplo el jurídico o el económico. Este aspecto está muy en consonancia con la opinión de muchos lingüistas (West, 1931; Carter, 1987; Nation, 2001) en torno a la necesidad de considerar no sólo la frecuencia sino también el rango en la elaboración de listados

pedagógicos: "As in studies of vocabulary, in the study of collocation range needs to be considered along with frequency" (Nation, 2001: 329). Colocaciones del tipo *housing provision*, *inward investment*, *marginal cost* o *registered land* fueron por tanto eliminadas de nuestra lista.

Así pues, una vez llevado a cabo este segundo procedimiento de selección de colocaciones mediante la aplicación de los criterios descritos, obtuvimos un listado definitivo de 2.688 colocaciones diferentes (ver Anexos 2 y 3[19]). Este gran banco de datos, inédito hasta el momento, supone, en nuestra opinión y como cabía esperar, una referencia inestimable no sólo para la elaboración de nuestro test de colocaciones sino, de forma más general, para la enseñanza del inglés en términos de selección de contenidos y diseño de materiales. Sin embargo, a medida que compilábamos este listado nos sorprendió comprobar que ofrecía asimismo información muy valiosa sobre el comportamiento de la lengua que nos permitía realizar análisis lingüísticos e incluso ideológicos más profundos. El siguiente apartado se dedicará a tratar estas y otras cuestiones.

3.3. Análisis y aplicaciones de la lista de colocaciones

Como mencionamos más arriba, el listado definitivo de colocaciones que obtuvimos como resultado del minucioso proceso de extracción automatizada y selección manual realizado contaba con un total de 2.688 colocaciones diferentes. Como recordaremos, nuestro estudio se realizó sobre una lista de 461 sustantivos que funcionaban como bases de las colocaciones, por lo que podemos concluir que el

[19] Como ya se adelantó en la introducción de este trabajo, el Anexo 2 presenta el listado con información completa de su frecuencia, índice *t-score* y corpus de procedencia, mientras que el Anexo 3 presenta un listado definitivo de colocaciones sin la información anterior.

listado final contaba con una media de 5,83 colocados por nombre, cifra que asciende hasta los 6,49 colocados si descontamos en nuestro cálculo los 45 sustantivos que no dieron lugar a ninguna colocación (como se puede observar en el Anexo 3, estos 45 sustantivos que no ofrecían ninguna colocación fueron eliminados de nuestro listado final). En nuestra modesta opinión, este índice refleja que estamos ante una base de datos muy completa que aporta una considerable cantidad de información sobre el comportamiento colocacional de cada sustantivo.

Otra observación que cabe hacer a la vista del resultado obtenido se refiere a las claras diferencias que existen entre los datos aportados por ambos corpus en términos cuantitativos. A simple vista se puede apreciar que el BNC aportó un número muy superior de colocados al del BoE, sin duda debido a la restricción de 50 colocados que *jLookup* tiene impuesta por defecto, como ya advertimos. Más concretamente, los datos que arroja el listado indican que de entre las 2.688 colocaciones totales, ambos corpus coinciden en 1.281 de ellas, mientras que del resto, el BoE aportó 96 que no se encontraban en los listados de *Sketch Engine*, y el BNC contribuyó con 1.311 colocaciones que no aparecían en las extraídas desde el BoE. Sin embargo, a pesar de que a juzgar por estos datos el uso del BoE puede parecer innecesario para compilar este listado, un análisis más detenido de las colocaciones que proceden exclusivamente de este corpus justifica claramente su papel en este estudio. Así, vemos que entre las colocaciones que sólo el BoE recoge como combinaciones significativas encontramos combinaciones tan idiosincrásicas y frecuentes como *expect a child*, *quit a job*, *superb quality*, *poor standard* o *missing person* por nombrar sólo algunas. Por otro lado, también observamos que algunas de las colocaciones que sólo aparecen en el BoE ponen claramente de manifiesto que se trata de un corpus más actualizado donde se recogen expresiones que se utilizan de forma más frecuente en la actualidad que dos o tres décadas atrás. Este es el caso, por ejemplo, de *quiz show*, *payment method* o *documentary series*. La naturaleza de este tipo

de colocaciones, claramente interesantes para nuestro estudio, nos lleva pues a concluir que la decisión de emplear ambos corpus de forma complementaria resultó no sólo justificada sino muy positiva para lograr un listado completo, fiable y eficaz.

En lo que se refiere a los tipos de colocaciones que incluye nuestra lista desde el punto de vista de su estructura sintáctica, las proporciones resultantes son las siguientes (Tabla 3.5.):

	Número de colocaciones	**Porcentaje**
A+N	1.588	59,09%
V+N	851	31,68%
N+N	186	6,94%
N+V	63	2,3%
Total	2.688	100%

Tabla 3.5.: Proporciones de estructuras sintácticas del listado

La figura 3.19. ofrece una representación gráfica de estas proporciones:

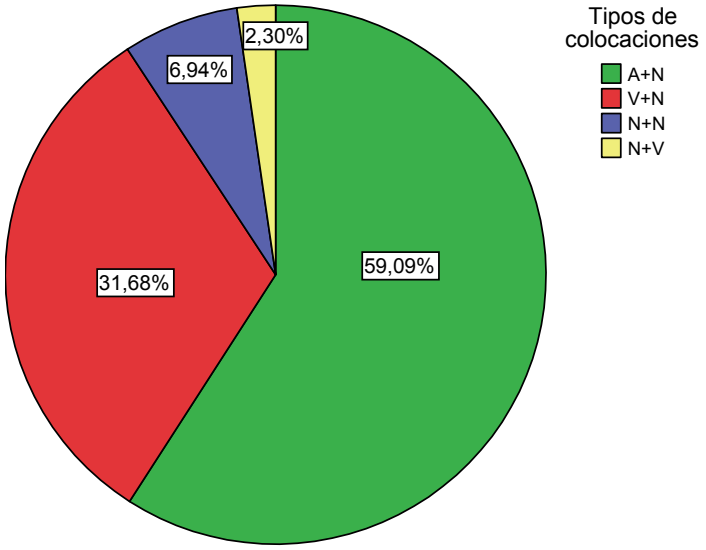


Fig. 3.19.: Gráfico de la proporción de estructuras sintácticas

A la vista de este gráfico resulta evidente que, como cabía esperar, las colocaciones de base nominal del tipo adjetivo-nombre y verbo-nombre representan la inmensa mayoría en nuestro listado (sumando un total de 90,77%) y, por extensión, en la lengua. Sí nos llama poderosamente la atención, sin embargo, el hecho de que la categoría más frecuente sea la de adjetivo-nombre, alcanzando de hecho un porcentaje que prácticamente dobla al de verbo-nombre. La gran mayoría de los trabajos dedicados al estudio de las colocaciones se dedican de forma exclusiva a la combinación verbo-nombre argumentando que se trata de la más frecuente y/o productiva (por ejemplo Aisenstadt, 1981; Howarth, 1996; Nesselhauf, 2005), lo cual contradice de forma clara nuestros resultados. Consideramos que esta disparidad de opiniones puede quizá deberse a que estos trabajos asumen criterios más estrictos a la hora de seleccionar las colocaciones, y en el caso de las combinaciones adjetivo-nombre, no consideran colocaciones aquellas donde varios adjetivos sinónimos son igualmente válidos y de uso frecuente junto a un determinado nombre. Siguiendo este criterio, construcciones como *key/central/vital/crucial role* se considerarían combinaciones libres, ya que el sustantivo no está restringido a uno de estos adjetivos en particular ni muestra preferencia por uno o varios de ellos, sino que los acepta todos. Sin embargo, desde nuestra perspectiva, y como ya argumentamos en el primer capítulo de este trabajo, en casos como éste sí suele existir cierta arbitrariedad en la co-aparición de los elementos, siendo normalmente el adjetivo el que está restringido a algunos nombres mientras que no suele acompañar a otros que en teoría también serían posibles. Por continuar con el ejemplo anterior, combinaciones como **central day* o **vital argument* son aceptables gramaticalmente pero en cambio no se suelen usar, lo cual demuestra que sí existe cierto grado de arbitrariedad.

Pero si hay un aspecto que verdaderamente llama nuestra atención a la vista del listado obtenido, éste se refiere sin lugar a dudas a la propia naturaleza de las combinaciones léxicas que recoge y las relaciones semánticas que se establecen entre

ellas, como si se tratara de un mapa que refleja la realidad e incluso la ideología y la cultura de una comunidad de habla. Esta información, que surge como consecuencia inesperada de la exploración de la lengua a través de los corpus en lo que se conoce como "proceso de serendipia", nos ocupará en el siguiente apartado.

3.3.1. Proceso de "serendipia"

Desde que el escritor Horace Walpole acuñara el término "*serendipity*" en 1754[20], traducido frecuentemente al español como "serendipia" y también en algunas ocasiones como "serendipidad"[21], este concepto ha ido afianzándose paulatinamente en el campo de la ciencia para referirse al proceso por el cual un hallazgo se produce de forma casual, es decir, como un efecto colateral de una investigación que tenía un objetivo distinto. Son famosos, por poner algunos ejemplos, los procesos de serendipia que se han producido en las distintas disciplinas científicas como el descubrimiento fortuito del planeta Urano en 1781 cuando el astrónomo William Herschel buscaba cometas, o en hechos históricos como el descubrimiento del continente americano por parte de Cristóbal Colón en 1492 cuando su intención era descubrir una nueva ruta hacia la India.

---

[20] Walpole creó esta palabra tras leer un cuento persa titulado "The Three Princes of Serendip", en el que sus tres protagonistas lograban resolver sus problemas gracias a la obtención de información de manera fortuita.

[21] Aunque ninguno de los dos términos está recogido en el *Diccionario de la Real Academia Española*, el *Diccionario del Español Actual* (Seco, Andrés y Ramos, 2005) contempla la palabra "serendipidad" mientras que el *Gran Diccionario de Uso del Español Actual* (Sánchez, 2001) incluye el término "serendipia". En cuanto a los diccionarios bilingües consultados, tanto el *Gran Diccionario Oxford Español-Inglés/Inglés-Español* (2003) como el *Collins Universal Español-Inglés/Inglés-Español* (2005) traducen el término como "serendipia". Por su parte, el Corpus de Referencia del Español Actual (CREA) (http://corpus.rae.es/creanet.html) muestra 3 resultados con "serendipia" frente a 1 solo con "serendipidad". Este dato se confirma con nuestras búsquedas en Internet, ya que Google muestra 56.300 resultados con el primero, y tan solo 7.220 con el segundo. A la vista de que "serendipia" parece ser claramente más frecuente, hemos preferido usar esta traducción.

En lo que a la lingüística se refiere, desde que esta disciplina se ha incorporado al área de las ciencias empíricas gracias, especialmente, a que hoy en día los estudios en este campo se fundamentan sobre datos fehacientes aportados por los corpus, el proceso de serendipia ha comenzado también a manifestarse en este campo del saber. Desde el punto de vista descriptivo, Tognini-Bonelli (2008) ha destacado recientemente las virtudes de la metodología de corpus a la hora de estudiar el lenguaje, entre otras razones porque al observar y analizar innumerables manifestaciones de la lengua, se produce muy a menudo un proceso de serendipia que nos lleva a descubrir conductas lingüísticas recurrentes de naturaleza insospechada. Este mecanismo de análisis ha sido también utilizado en el área de la lingüística aplicada en trabajos como los de Johns (1988), Higgins (1991), Zanettin (1998) o Bernardini (2000, 2002), para quienes este proceso de descubrimiento a que da lugar la exploración de corpus se puede explotar en el aula, mediante actividades que fomenten la reflexión y el aprendizaje heurístico. Como señala Bernardini (2002: 165), "one of the main functions of corpora in language pedagogy is that of being able to provide rich sources of autonomous learning activities of a serendipitous kind".

Este fenómeno tan recurrente en los estudios de corpus, como decíamos, se ha manifestado también de forma notable en nuestro trabajo. De forma inesperada, hemos podido comprobar que si observamos con detenimiento el listado de colocaciones producto de nuestro proceso de compilación, podemos identificar patrones más o menos fijos que reflejan claramente las interconexiones léxicas que existen en la lengua. No nos referimos aquí únicamente a las relaciones semánticas de sinonimia, antonimia y pertenencia al mismo campo semántico que se establecen entre los colocados de cada nombre, algo que no nos resulta sorprendente ya que sabemos que las palabras no se utilizan de forma totalmente aleatoria sino que responden a unos patrones semánticos, lógicos y contextuales que nos permiten la

comunicación. Lo que más llama nuestra atención a la vista de este listado se refiere a aspectos más sutiles y de estudio más reciente como son la preferencia y la prosodia semántica, que parecen regir la combinatoria léxica.

3.3.1.1. Preferencia semántica

La preferencia semántica, una noción introducida en la lingüística gracias a los estudios de corpus, ha sido definida por Stubbs (2001: 65) como "the relation, not between individual words, but between a lemma or word-form and a set of semantically related words". Esta definición, como vemos, establece claramente la diferencia entre colocación y preferencia semántica, ya que la primera se da en combinaciones de unidades léxicas concretas, mientras que la segunda supone un nivel superior de abstracción porque implica la combinación de una palabra con un conjunto de palabras que comparten ciertos rasgos semánticos. Uno de los ejemplos más paradigmáticos de los estudios relacionados con este aspecto se refiere a "*large*", palabra que suele ir acompañada de términos relativos a la cantidad y el tamaño como "*number*", "*scale*", "*part*", "*amounts*" o "*quantities*" (Stubbs, 2001). Otro claro exponente son los adverbios intensificadores "*utterly*", "*totally*", "*completely*" y "*entirely*" que muestran una clara preferencia por términos que denotan carencia o cambio de estado, como por ejemplo "*irrelevant*", "*unexpected*", "*altered*" o "*different*" (Partington, 2004).

Evidentemente, y como hacen notar los estudiosos en este campo (Sinclair, 2003; Hoey, 2004; Partington, 2004), el hecho de que una palabra presente una preferencia semántica por uno o varios grupos de unidades léxicas no implica en modo alguno que sólo se pueda combinar con elementos que encajen en dichos grupos semánticos. Como ya vimos en el capítulo 1, el lenguaje tiene una clara tendencia a estructurarse en patrones prefabricados, en consonancia con el "principio

idiomático" que postulara Sinclair (1991), pero esto no implica que todo sea fijo y esté predeterminado en el uso lingüístico. Existe también un considerable grado de flexibilidad en la lengua, contemplado asimismo por Sinclair (ibid.) en su "principio de elección libre", que nos otorga cierta libertad de movimiento dando lugar a la creatividad expresiva. En este sentido, pues, no debe resultar contradictorio que una palabra muestre una clara tendencia a combinarse con elementos de un determinado campo semántico, a la vez que también se combina de forma más aislada con palabras que no comparten los rasgos semánticos de ese campo. En palabras del propio Sinclair (2003: 178),

> [w]hile the majority of the choices will show the preference clearly, there may be a small number of marginal cases where the preferred meaning has to be interpreted in a rather elastic fashion, and some which appear to be exceptions. For this reason we do not use a word like "restriction" instead of "preference".

Partington (2004: 146) ofrece también una explicación muy gráfica: "language users are able to swim against the current —can 'switch off' primings— when they seek particular creative effects. Semantic prosody and preference do not ordain that counter examples *cannot* happen, just that they *seldom* happen".

Otro aspecto que merece nuestra atención con respecto a este tipo de relación léxica se refiere a su distribución dentro de la colocación. Como sabemos, la corriente sinclairiana no establece una distinción entre base y colocado tal y como se entiende en los estudios fraseológicos donde ambos elementos se consideran a distinto nivel, siendo el colocado el elemento que depende de la base. Para los autores de la aproximación estadística el nodo es la palabra que se esté estudiando en un momento preciso, por lo que cualquier término es pues susceptible de convertirse en el nodo de una combinación. Por esta razón, la preferencia semántica no se ha definido hasta ahora en términos de base y colocado, sino que se ha estudiado en distintas palabras, o nodos, independientemente de su posición fraseológica en la

colocación. En cualquier caso, como veremos a continuación, al observar este fenómeno considerando la diferencia entre base y colocado, no resulta sorprendente comprobar que la preferencia semántica se da en ambas direcciones, si bien es cierto que su frecuencia es marcadamente superior en las bases que en los colocados. Decimos que no nos sorprende observar que tanto las bases como los colocados muestran preferencias semánticas, puesto que es lo que cabría esperar si consideramos que ambos elementos muestran también arbitrariedad a la hora de combinarse con otras palabras. Es posible, en este sentido, que la preferencia (y también la prosodia semántica que trataremos en el siguiente apartado) sea un fenómeno paralelo en cierto grado a la arbitrariedad combinatoria, o quizá incluso la provoque en algunos casos. De cualquier modo, se trata de aspectos que requieren de mayor investigación empírica puesto que, como decimos, se trata de una cuestión que hasta ahora no ha sido estudiada a fondo.

En el listado recogido en los Anexos 2 y 3, podemos observar un número considerable de casos de preferencia semántica. En cuanto a aquellos donde es la base la que muestra preferencia por ciertos conjuntos de colocados, mostramos los ejemplos más notables en la tabla 3.6.:

Elemento concreto (base)	Conjunto(s) semántico(s) por el(los) que muestra preferencia
Control	Concepto de logro/tenencia: *exercise, gain, have, keep, maintain, retain, take* Concepto de totalidad: *complete, full, overall, total*
Factor	Concepto de importancia: *critical, crucial, deciding, decisive, determining, key, major, significant, vital*
Feature	Concepto de distinción/predominio: *attractive, characteristic, distinctive, dominant, major, notable, prominent, significant, striking, unique, unusual*
Hand	Acciones físicas: *extend, hold, lift, raise, shake, stretch out, take, wash, wave*
Head	Acciones físicas: *bend, bow, hold, jerk, lift, lower, nod, pull back, raise, shake, throw back, tilt, turn*
Information	Concepto de logro: *collect, exchange, gather, obtain, provide, supply*
Man	Aspecto físico: *big*, *dark*, *small*, *tall*
Part	Concepto de importancia: *essential, full, integral, major, significant, vital*
Population	Concepto de tamaño: *entire, large, overall, total, whole*
Position	Concepto de poder/predominio: *competitive, dominant, leading, privileged, strong*
Pressure	Concepto de cantidad: *considerable, enormous, growing, heavy, high, increasing, intense, low, severe, strong*
Problem	Concepto de resolución: *address, overcome, resolve, solve, sort out, tackle* Concepto de importancia: *big, major, real, serious*
Woman	Belleza física: *attractive, beautiful, good-looking, pretty, tall*

Tabla 3.6.: Preferencia semántica de algunas bases del listado

Por otro lado, también podemos observar este tipo de asociación léxica en la dirección opuesta, es decir, cuando el colocado tiende a ir acompañado de bases semánticamente relacionadas. La tabla 3.7. muestra algunos ejemplos:

Elemento concreto (colocado)	Conjunto(s) semántico(s) por el(los) que muestra preferencia
Low	Concepto de cantidad: *cost, figure, income, interest, price, rate* Concepto de clasificación: *level, quality, standard*
Serious	Situaciones de contingencia: *disease, doubt, loss, problem, risk* Concepto de potencialidad y acto: *attempt, business, cases, condition, issue, matter* Actividades que exigen esfuerzo cognitivo: *attention, thought*
Total	Concepto de cantidad: *cost, number, population, sales, value*
Whole	Nombres colectivos: *church, country, family, lot, population, school, series, set, society, world* Concepto de temporalidad: *day, life, week* Concepto de potencialidad y acto: *affair, business, idea, point, process, project, question, thing*

Tabla 3.7.: Preferencia semántica de algunas bases del listado

Observando las tablas anteriores podemos confirmar que muchas palabras tienden a asociarse de forma idiosincrásica respondiendo a patrones semánticos muy sutiles, lo cual casi siempre revierte en la acotación de la propia palabra, que pasa de tener un significado potencialmente amplio a una cierta especialización o restricción marcada por los elementos que la suelen acompañar. En este sentido, nos parece de interés

comprobar que términos como "*low*" o "*total*" tienen un marcado carácter cuantitativo debido a que muestran una preferencia semántica por términos relacionados con la noción de cantidad, del mismo modo que "*feature*", por ejemplo, también ve claramente acotado su propio significado debido a su tendencia a combinarse con adjetivos que denotan prominencia o distinción.

Pero la preferencia semántica puede ir incluso más allá, dado que no sólo nos permite indagar en la naturaleza y las sutilezas semánticas del léxico, sino que también nos ayuda a desvelar rasgos culturales e ideológicos de una determinada comunidad de habla, mostrando hasta qué punto la lengua es un claro reflejo de la sociedad en la que vivimos. Volviendo de nuevo a los resultados que ofrece nuestro listado, resulta ciertamente interesante descubrir que el concepto más relacionado con la palabra "*information*" es "conseguir", lo cual parece demostrar la importancia que tiene estar en posesión de la información en la sociedad actual, reflejando así que vivimos en la era de la información y la comunicación, en la que estar bien informado supone una clave importante para lograr el éxito.

Asimismo, consideramos que las palabras "*man*" y "*woman*" merecen una atención especial. Nos parece muy significativo, en primer lugar, el hecho de que ambas muestren una tendencia muy marcada a combinarse con elementos que pertenecen al campo semántico del aspecto físico. Esto, que en realidad no resulta sorprendente ya que se corresponde con lo que cabría esperar desde la intuición, parece confirmar que el aspecto físico es una característica verdaderamente importante que nos merece gran atención al referirnos a las personas, lo cual nos parece muy indicativo de los valores que priman en la sociedad actual.

Por otro lado, nos parece también muy interesante la información que se puede obtener al observar las diferencias entre la preferencia semántica que muestra cada una de estas dos palabras. Son varios los estudios que han abordado este tema a lo largo de los últimos años (Caldas-Coulthard y Moon, 1999, citado por Pearce,

2008; Romaine, 2000; Pearce, 2008), demostrando todos ellos que existen notables diferencias entre las asociaciones de "*man*" y "*woman*" y, en general, entre los conceptos masculino y femenino. En este sentido, todos estos trabajos han puesto de manifiesto que los corpus reflejan claramente los estereotipos sociales de hombre y mujer, ya que "*man*" se suele asociar con términos que denotan importancia, poder, fuerza física e incluso agresión, mientras que "*woman*" se combina con nociones como la debilidad, las emociones y la belleza física.

Al analizar los resultados obtenidos en nuestro listado, podemos comprobar que éste refleja las mismas tendencias que los estudios que acabamos de mencionar. Como se puede apreciar en la tabla 6, "*man*" parece colocar con adjetivos en su mayoría relacionados con la idea del tamaño y, cabe pensar, la fuerza física, como "*big*", "*small*" o "*tall*", mientras que "*woman*" muestra una clara tendencia a combinarse con adjetivos relativos a la estética y la belleza física como "*attractive*", "*beautiful*", "*good-looking*" o "*pretty*". Esta diferencia entre los conceptos que se asocian al hombre y la mujer, se confirman también en términos generales si observamos los colocados de "*boy*" y "*girl*" en nuestro listado. A pesar de que "*boy*" no presenta una preferencia semántica definida y "*girl*" sólo muestra dos colocados (razón por la que no se han incluido en la tabla 6), podemos observar que, de nuevo, "*boy*" está asociado a la idea de tamaño ("*big*" y "*small*") y de comportamiento ("*good*" y "*bad*"), mientras que "*girl*" parece asociarse exclusivamente con términos relativos a la belleza física ("*beautiful*" y "*pretty*"). En nuestra opinión, estas tendencias son muy sintomáticas de que los conceptos de masculinidad y feminidad siguen muy asociados a los valores tradicionales en la sociedad actual donde, podemos deducir, siguen persistiendo las actitudes sexistas.

Pensamos que los ejemplos referidos ilustran de forma clara las posibilidades de análisis y profundización en la naturaleza de las combinaciones léxicas que ofrece el estudio de la preferencia semántica en un listado como el nuestro. Éste no es el

único rasgo observable en materia de asociación léxica, sin embargo. Los estudios de corpus, como decíamos, han evidenciado la existencia de otro fenómeno igualmente interesante y del que pasamos a ocuparnos a continuación: la prosodia semántica.

3.3.1.2. Prosodia semántica

La prosodia semántica, al igual que sucedía con la preferencia semántica, es un concepto de muy reciente aparición desarrollado principalmente por Sinclair (1987), Louw (1993) y Stubbs (1995). Partington (2004: 149) define la prosodia semántica como "instances where an item shows a preference to co-occur with items that can be described as bad, unfavourable or unpleasant, or as good, favourable or pleasant" y continúa añadiendo que la verdadera importancia de este fenómeno radica en su valor discursivo y pragmático: "semantic prosodies are evaluative or *attitudinal* and are used to express the speaker's approval (good prosody) or disapproval (bad prosody) of whatever topic is momentarily the object of discourse (Sinclair 1996: 87)" (ibid.: 150).

Entre los ejemplos más conocidos de prosodia semántica podemos citar brevemente los dos primeros casos identificados por Sinclair (1987), "*happen*" y "*set in*", con relación a los cuales este autor descubrió que poseen prosodias claramente negativas ya que suelen combinarse con palabras de este signo. Otro estudio que supone un punto de referencia en este campo es el llevado a cabo por Stubbs (1995) sobre el término "*cause*" (tanto en su función de nombre como de verbo), que también muestra una clara tendencia a co-aparecer junto a elementos de valor negativo como "*anxiety*", "*concern*", "*crisis*", "*damage*", "*distress*" o "*trouble*", entre otras similares.

Pero, evidentemente y como apunta Louw (1993: 171), "[t]here seem, *prima facie*, to be more 'bad' prosodies than 'good' ones, but the latter certainly exist and the

principles on which they work are the same". Este autor cita, por ejemplo, "*build up*", que cuando se usa de forma transitiva tiene una clara prosodia positiva. Lo mismo sucede con palabras como "*provide*", cuya prosodia viene marcada, como demuestra Stubbs (1995), por términos como "*aid*", "*facilities*", "*help*", "*information*", "*services*" o "*support*", todos ellos de carácter claramente favorable.

A pesar de que Partington considera la prosodia semántica como un fenómeno dicotómico donde sólo se observa la evaluación negativa y positiva del hablante, esta visión ha dado lugar a cierta controversia. En primer lugar, varios autores (Louw, 1993; Stubbs, 1995) han documentado el hecho de que, por un lado, existen casos en los que no se da una inclinación positiva ni negativa sino que existe una prosodia neutral, y por otro, a veces podemos observar una prosodia mixta dado que existe un número similar de palabras negativas, positivas y neutras en el entorno de una palabra. Stubbs (ibid.) ha ejemplificado la prosodia neutral con la palabra "*reason*", dado que se combina con elementos como "*apparent*", "*different*", "*simple*" o "*discernible*", que no manifiestan ninguna orientación marcadamente favorable o desfavorable por parte del hablante. La prosodia mixta también ha sido recogida por este autor en referencia a palabras como "*result*" que coloca con "*disappointing*", "*inconclusive*" o "*losses*" por un lado, con elementos de carácter positivo como "*expected*" o "*positive*" por otro, y con palabras neutras como "*test*" o "*final*".

Dando un paso más allá, Hunston (2007) opina que una palabra no suele tener una prosodia que sirva por igual en todos los contextos y, sobre todo, para todos los hablantes, sino que se trata de un fenómeno más sutil que depende del punto de vista de cada interlocutor. En opinión de esta autora,

> the concept that a word 'has a (negative or positive) semantic prosody' (...) can involve taking a somewhat simplistic view of attitudinal meaning. Such meaning is often not reducible to a simple 'positive or negative'. It is essentially linked to point of view, so that there is often not one indisputable interpretation of attitude. ((...)

> *destruction* is a process which is often good for the destroyer but bad for the destroyed.) (Hunston, 2007: 256).

En términos generales podemos comprobar que la prosodia semántica es una relación menos restringida que la que se establece en la preferencia semántica, constituyendo de hecho, como afirman Sinclair (1996, 1998) y Stubbs (2001), el tipo de relación léxica más abstracta y abarcadora de las cuatro descritas hasta el día de hoy: 1) colocación (relación entre dos (o más) palabras concretas), 2) coligación (relación entre una palabra y una forma o estructura gramatical), 3) preferencia semántica (relación entre una palabra y un grupo de palabras pertenecientes al mismo campo semántico) y 4) prosodia semántica (relación entre una palabra y un grupo de palabras de significado positivo o negativo).

Aparte de por el hecho de que la relación semántica que se establece en la prosodia es más general, más indefinida, que la que observamos en la preferencia semántica, la razón por la que aquella tiene una naturaleza más abstracta es que no sólo afecta a las 2 palabras que intervienen en la combinación sino que su efecto se extiende por lo general a lo largo de una considerable parte de la oración o del texto (Partington, 2004). Esto se debe, sin duda, a que la característica fundamental de la prosodia, a diferencia de la preferencia cuya naturaleza es puramente semántica, es su valor discursivo. Como afirma Sinclair (1996: 88), la prosodia le indica al interlocutor cómo debe interpretar una oración desde el punto de vista funcional, ya que "[w]ithout it the string of words just 'means' —it is not put to use in a viable communication". La prosodia semántica (o prosodia discursiva como prefiere denominarla Stubbs dada la información funcional que ofrece) aporta un valor pragmático añadido similar al que observamos en la connotación de las palabras individuales. Resulta evidente, pues, que estamos ante un fenómeno de enorme trascendencia para el estudio, y por supuesto también la enseñanza, de la lengua.

Una vez definido el concepto de prosodia semántica, nos interesa comprobar los ejemplos que ofrece nuestro listado con relación a este fenómeno, de nuevo estableciendo la distinción entre las bases y los colocados de nuestra lista. Resultará también interesante comprobar si se confirman las apreciaciones que se han hecho en estudios anteriores sobre algunas de las palabras que contiene nuestra lista, como hemos visto en los ejemplos citados más arriba, puesto que esto contribuiría a otorgar validez a nuestra lista. Comenzaremos por la prosodia que reflejan algunas de las bases de nuestras colocaciones (Tabla 3.8.).

Elemento concreto (base)	**Prosodia semántica**
Attempt	Prosodia negativa: *desperate, fail, unsuccessful, vain*
Cases	Prosodia negativa: *criminal, extreme, injury, serious, severe*
Condition	Prosodia mixta: ▪ Positiva: *excellent, good, immaculate, perfect, stable* ▪ Negativa: *critical, deteriorated, poor, serious*
Effect	Prosodia negativa: *adverse*, *detrimental, devastating, dramatic, opposite*
Environment	Prosodia positiva: *healthy, improve, natural, protect, safe, stable*
Ideas	Prosodia positiva: *bright*, *fresh*, *original*
Opportunity	Prosodia positiva: *ample, equal, excellent*, *golden, ideal, perfect, real, unique*
Reason	Prosodia neutral: *apparent, major, real, simple, sufficient, unknown*
Step	Prosodia positiva: *big, essential, major, necessary, positive, significant*

Tabla 3.8.: Prosodia semántica de algunas bases del listado

En lo que respecta a la prosodia de los colocados de nuestra lista, mostramos algunos ejemplos en la tabla 3.9..

Elemento concreto (colocado)	Prosodia semántica
Cause	Prosodia negativa: *concern*, *death, disease, fire, loss, problem*
Provide	Prosodia positiva: *answer*, *care, evidence, help, means, opportunity, support, training*
Serious	Prosodia mixta: ▪ Negativa: *cases*, *condition*, *disease, doubt, loss, problem, risk* ▪ Neutral: *attempt, attention, business, issue, matter*, *thought*
Severe	Prosodia negativa: *cases, disease, loss, pressure*

Tabla 3.9.: Prosodia semántica de algunos colocados del listado

A la vista de la información recogida en las tablas anteriores se hace patente que la prosodia semántica es un fenómeno de considerable importancia en la lengua ya que da lugar a patrones sorprendentemente estables que otorgan a las palabras un valor pragmático añadido, imposible de imaginar desde la intuición y que en la mayoría de los casos no está todavía recogido en ningún diccionario ni manual de gramática (Partington, 2004). Así, podemos observar que palabras que en principio parecen neutras y sin ninguna carga connotativa propia presentan en realidad una fuerte tendencia a aparecer en contextos donde se aprecia una valoración positiva por parte del hablante, como es el caso de "*environment*", "*ideas*" o "*step*". En este mismo sentido, también podemos comprobar que, al igual que sucedía en el caso de la preferencia semántica, los casos de prosodia que observamos en nuestro listado confirman, en términos generales, las tendencias descritas en investigaciones anteriores. Palabras como "*cause*", "*provide*" o "*reason*", por ejemplo, muestran en

nuestro listado el mismo comportamiento combinatorio que en los estudios llevados a cabo por Stubbs (1995), lo cual no viene sino a avalar los resultados obtenidos en la lista obtenida en nuestra investigación.

Por último, cabe también mencionar que, al igual que en la preferencia semántica, a veces existen excepciones en la tendencia prosódica que presenta una palabra, lo cual confirma, a nuestro juicio, la opinión de Partington (2004: 153) en cuanto a que "prosodic meaning, though basic to an item, is more obviously *probabilistic* than some other kinds of meaning". Este es el caso, por ejemplo, del sustantivo "*effect*" que parece mostrar una tendencia prosódica negativa pero también se puede combinar con adjetivos positivos como "*beneficial*" o "*desired*". En la misma línea, vemos que las tendencias prosódicas están más marcadas en unas palabras que en otras, siendo por tanto la prosodia un fenómeno graduable. Entre los ejemplos incluidos en las tablas 8 y 9 resulta sencillo observar que unidades léxicas como "*cause*" muestran una mayor propensión a aparecer en un entorno prosódico bien delimitado que otras como "*condition*", de naturaleza más flexible. A nuestro modo de ver, este hecho confirma la opinión de Hunston (2007) a propósito de que una palabra no siempre muestra una prosodia única, bien delimitada, y aplicable indefectiblemente en todos los contextos, sino que se trata de un fenómeno muy abstracto que depende de distintos factores discursivos.

Una vez llevado a cabo el análisis de las colocaciones incluidas en nuestro listado en lo que respecta a su preferencia y prosódica semántica, nos parece oportuno mencionar brevemente algunas cuestiones relativas a este estudio. En primer lugar, somos conscientes de que en este apartado se ha realizado un análisis muy somero del contenido del listado y hemos presentado tan sólo una parte de la información que éste ofrece, pero lo consideramos suficiente para nuestro objetivo, que es el de ilustrar las posibilidades que presenta la lista resultado de nuestra

investigación en materia de análisis léxico-semántico, información que no pretendíamos obtener en un principio y que hemos hallado por el proceso de serendipia al que hacíamos referencia anteriormente. Por supuesto, pensamos que sería conveniente llevar a cabo un análisis más exhaustivo de todos estos aspectos ya que nos parece que puede conducir a resultados muy interesantes acerca de la propia naturaleza de las colocaciones, y ello constituirá sin duda el objeto de estudio de futuros trabajos.

Asimismo, también nos gustaría mencionar que, por un lado, dadas las características de las colocaciones que se han recogido en este trabajo, enmarcadas en nuestra definición fraseológica de partida, el número de colocados obtenido por cada nombre es inferior al que se suele encontrar en los estudios de análisis semántico y de análisis crítico del discurso basados en corpus. Esto redunda, evidentemente, en una menor cantidad de información que la que normalmente recogen estos trabajos, cuya orientación en cuanto al fenómeno colocacional suele ser más estadística. Por otro lado, sin embargo, teniendo en cuenta la finalidad eminentemente pedagógica que se persigue en esta tesis, consideramos que las colocaciones incluidas en nuestro listado son aquellas especialmente útiles para nuestro alumnado, y por tanto, las más interesantes en este trabajo. Asimismo, pensamos que el hecho de que, como hemos podido comprobar, las conclusiones a las que estos estudios han llegado en relación a las palabras que aparecen en nuestro listado se ven sistemáticamente confirmadas por éste, contribuye a reforzar la validez de nuestra lista.

En conclusión, nos parece interesante destacar la importancia de los patrones semánticos que los estudios de corpus han venido a revelar, ya que han pasado tradicionalmente inadvertidos para los lingüistas. Como han argumentado distintos autores (Louw, 1993; Partington, 1998; Hunston, 2002), la prosodia y la preferencia semántica son dos fenómenos tan inaccesibles a la intuición del hablante como lo es la propia colocación. Esta es, claramente, una de las más importantes contribuciones

que la lingüística de corpus, y por supuesto el admirable trabajo de John Sinclair, ha venido a aportar al estudio lingüístico. Gracias a esta nueva forma de mirar y de analizar, somos ahora capaces de advertir las sutilezas que existen en la lengua.

Además, otra interesante aportación de los estudios de corpus, es sin duda el proceso de serendipia al que dan lugar, ya que, a la vista de lo anterior, podemos afirmar que se trata de un proceso muy fructífero que nos ayuda a profundizar en la compleja naturaleza del lenguaje y a formularnos nuevos interrogantes para futuras investigaciones. A nuestro modo de ver, la serendipia es un aspecto que hace especialmente reveladores los estudios de corpus y contribuye en gran medida a que comprendamos mejor no sólo cómo se estructura y cómo funciona el lenguaje humano, sino también a que seamos conscientes de lo mucho que nos resta aún por descubrir mediante el estudio empírico de la lengua viva. En definitiva, es a través de este tipo de estudios cuando podemos comprender el verdadero alcance de las palabras de John Sinclair (2004): "trust the text". Sólo confiando en el texto, en los datos, podremos descubrir la riqueza que esconde el lenguaje.

3.4. Conclusión

En este capítulo se ha descrito de forma detallada el minucioso procedimiento llevado a cabo para la compilación de un listado de colocaciones frecuentes que nos sirviera como banco de datos para la posterior elaboración de un test. Un aspecto que nos parece importante destacar es que, a la luz del proceso descrito a lo largo de este capítulo, resulta evidente que, a diferencia de muchos estudios de corpus donde la frecuencia estadística es el criterio más importante, y a veces incluso el único, a la hora de seleccionar y analizar elementos léxicos o fraseológicos, en este estudio se propone un enfoque ecléctico donde las colocaciones se extraen a partir de datos

extraídos de corpus en primer lugar, pero donde las restricciones arbitrarias que impone su naturaleza fraseológica también se consideran un factor discriminante. Nos parece, pues, que se trata de un listado inédito que viene a ofrecer una valiosa fuente de datos tanto para la investigación lingüística como para la enseñanza y la evaluación pedagógica.

Por supuesto, somos conscientes de que en la compilación que hemos llevado a cabo en este trabajo la intervención humana ha jugado un papel importante, con el fin de utilizar criterios que no siempre han sido fáciles de aplicar y que se prestan a diferentes interpretaciones. Esto supone, sin duda, que el listado final obtenido debe considerarse sólo como una propuesta, que evidentemente será discutible en algunos de sus elementos. Debido a la compleja naturaleza del propio fenómeno colocacional, donde intervienen muy diversos factores, y donde confluyen aspectos que son compartidos a la vez por otros tipos de combinaciones léxicas, resulta siempre difícil establecer límites claros y es inevitable que aparezcan elementos fronterizos que requieren de una interpretación subjetiva, que por tanto puede ser opinable. En cualquier caso, en este trabajo hemos tratado de conjugar las técnicas estadísticas más avanzadas que existen hoy en día con los criterios fraseológicos que esperamos se hayan establecido de forma sistemática y rigurosa, lo cual ha dado lugar a un listado fiable, muy completo y desde luego muy útil en términos prácticos.

Por otro lado, debido al hecho de que la información incluida en nuestro listado parte en primer lugar de los datos que ofrecen los corpus lingüísticos, hemos podido comprobar que el resultado final obtenido constituye una fuente de información verdaderamente interesante desde el punto de vista lingüístico, cultural e ideológico. Gracias al proceso de serendipia que la exploración de corpus suele conllevar, podemos observar que las colocaciones no son simples combinaciones recurrentes de palabras, sino que responden a tendencias más complejas y de mayor

alcance en el discurso. En nuestra opinión, los estudios de corpus presentan como vemos un enorme potencial ya que nos conducen hacia un mayor y mejor entendimiento del lenguaje, a la vez que descubre nuevos caminos por recorrer.

Finalmente, otra de las vertientes que presenta el listado aquí compilado es su aplicabilidad en el plano del diseño de materiales pedagógicos. En este sentido, no cabe duda de que contar con un banco de datos que recoja las colocaciones más frecuentes y más relevantes de la lengua inglesa es enormemente útil no sólo a la hora de seleccionar los contenidos que deben incluir nuestras actividades, sino también para poder organizarlas y presentarlas de forma convenientemente relacionada y ayudar así al alumno en su proceso de aprendizaje.

En definitiva, consideramos que el listado aquí presentado constituye un recurso novedoso y necesario desde el punto de vista pedagógico, y que confiamos contribuya a la elaboración de nuestro test de colocaciones, para que su contenido sea válido y fiable.

CAPÍTULO 4

LA EVALUACIÓN DE LA COMPETENCIA COLOCACIONAL: DISEÑO DEL TEST DE COLOCACIONES ADELEX VERSIÓN 1.

> *Testing collocation is superficially easy, using questions such as:*
> *The government is trying to close the ..... between rich and poor.*
> *A. space B. gap C. distance D. door*
> *But if such testing is to avoid arbitrariness, simply demonstrating that the learner either does or does not know a particular collocation, steps have to be taken to ensure that the test items are devised in principled ways.*
>
> Lewis (2000b: 150-151)

4.1. Introducción

Como hemos destacado a lo largo de este trabajo, las colocaciones constituyen un elemento fundamental de la competencia léxica de un hablante dado que contribuyen enormemente a la fluidez y precisión de su discurso. Cada vez se oyen más voces argumentado que el conocimiento individual de las palabras no es suficiente ya que conduce a "lexical incompetence on the part of L2 learners" (Farghal y Obiedat,

1995: 326). Por ello, resulta imprescindible incluir las colocaciones en los sílabos de segundas lenguas de una manera sistemática y eficaz (Pérez Basanta, 1999; Lewis, 2000; Nesselhauf, 2005). A pesar de este hecho, sin embargo, los autores que las estudian desde el área de la lingüística aplicada coinciden en destacar el escaso tratamiento que de ellas se ha hecho por parte tanto de investigadores como de profesores (Bahns y Eldaw, 1993; Howarth, 1996; Hill, 2000; Nation, 2001; Nesselhauf, 2005). Quizá el único trabajo que merece ser reseñado por su aportación a este ámbito, ya que está dedicado por entero al desarrollo pedagógico de las colocaciones, es el volumen titulado "*Teaching Collocation: Further Developments in the Lexical Approach*" editado por Michael Lewis en el año 2000. Por supuesto, no se trata de la única obra dedicada a la enseñanza de las colocaciones (existen otras como "*English Collocations in Use*", de McCarthy y O'Dell, 2005, o "*Key Words for Fluency*", publicado por Woolard también en 2005), pero sí es la única que conocemos en la que se realiza una aproximación más profunda al fenómeno desde el punto de vista de la lingüística aplicada y donde se ofrece una fundamentación teórica sólida. A pesar de lo anterior, sin embargo, cabe destacar que ni siquiera en este trabajo encontramos una propuesta didáctica rigurosa y bien planificada en términos prácticos, sino que más bien recoge una serie de consejos y actividades dispersas que se ofrecen al profesor a modo de complemento para sus clases. Por tanto, resulta evidente que todavía en la actualidad **existe una carencia notable de investigaciones que ofrezcan propuestas prácticas para el desarrollo pedagógico de las colocaciones de una manera sistemática, rigurosa y eficaz**.

Sin embargo, y pese al alcance de la afirmación formulada más arriba en cuanto a la necesidad de un tratamiento pedagógico de las colocaciones, nosotros no vamos a dedicarnos en este trabajo a ello, sino que creemos que para abordar en profundidad las posibles carencias de la competencia colocacional, en nuestro caso con relación al alumnado español, debemos primero llevar a cabo un diagnóstico

certero de la situación en nuestro contexto educativo, ya que hasta la fecha no ha existido ninguna investigación rigurosa que nos informe sobre puntos de partida y niveles en el ámbito nacional.

Cabe destacar con respecto a lo anterior, no obstante, que esta carencia de trabajos que se dediquen a la evaluación de la competencia colocacional no es exclusiva de nuestro contexto, sino que también constituye un fenómeno muy escaso en el ámbito internacional, siendo una queja reiterada en el campo de la lingüística aplicada (Gyllstad, 2007: 2):

> Even though we know that collocations are challenging to L2 learners and that collocational knowledge is seen as something that normally distinguishes between L1 and L2 speakers of a language (Schmitt, 2000), there is a lack of reliable and properly validated test instruments with which learners' knowledge of collocations may be measured.

Por todo lo anterior, hemos decidido construir un test de colocaciones al que denominaremos **Test de Colocaciones ADELEX Versión 1 (TCA1)**, cuyo planteamiento es de doble vía. En primer lugar, pretendemos definir **qué es lo que hay que evaluar y cómo debe hacerse** y, en segundo lugar, tratamos de diseñar una prueba de diagnóstico que permita **medir niveles de conocimiento.** En cuanto al primer objetivo, pretendemos desarrollar un instrumento evaluador que resulte válido y fiable. En este sentido, y teniendo en cuenta que uno de los factores más descuidados en el campo de la evaluación colocacional es la sistematicidad en la elección de los contenidos, nos proponemos sentar las bases para una adecuada selección y gradación de colocaciones elaborada con criterios computacionales, prestando así especial atención a lo que en teoría clásica de la evaluación se conoce como la validez de contenido. En lo que respecta al segundo objetivo, nuestro test busca indagar de forma rigurosa y fiable en el conocimiento que nuestros alumnos poseen de un aspecto tan esencial como es la combinación arbitraria de las palabras,

algo que constituye un paso previo indispensable para poder proyectar un adecuado tratamiento pedagógico en el futuro. Ni qué decir tiene que en este trabajo se asume plenamente el argumento de Gyllstad (ibid.: 3), para quien "collocational knowledge is a separate skill which can be measured as a stand-alone trait, albeit potentially interdependent on other closely related lexical constructs".

Los dos objetivos fundamentales mencionados en el párrafo anterior constituyen la esencia de nuestra investigación y se abordarán en dos capítulos distintos. Así, mientras que el capítulo 5 se dedicará a la evaluación de la competencia colocacional de los alumnos universitarios españoles, el que ahora nos ocupa va a estar enfocado a ofrecer una descripción pormenorizada de los aspectos que se han observado a la hora de elaborar nuestro test de colocaciones. Para ello, realizaremos en primer lugar una breve revisión del estado de la cuestión en lo que se refiere a la evaluación de este fenómeno fraseológico. Seguidamente, abordaremos las cuestiones relativas al diseño de nuestro test, observando los preceptos de fiabilidad y validez que todo instrumento de medida debe contemplar. Por último, trataremos de explicar las condiciones en que se administró este test durante el proceso de pilotaje. De los resultados de dicho proceso, como dijimos, se ocupará el último capítulo de esta tesis.

4.2. La evaluación de la competencia colocacional: Revisión histórica

La evaluación del conocimiento colocacional, como ya hemos adelantado, ha sido un aspecto prácticamente olvidado en la investigación lingüística hasta fechas muy recientes. Incluso hoy en día, y a pesar de que ya existe un amplio consenso, al menos en el plano teórico, en torno a la importancia de las colocaciones como aspecto

prioritario en la adquisición de una adecuada competencia léxica y, por ende, comunicativa, este fenómeno sigue siendo relativamente minoritario en los estudios producidos en el campo de la evaluación. En lo que se refiere al uso que los estudiantes de inglés como lengua extranjera (L2) hacen de las colocaciones inglesas, área en la que nos concentraremos en esta breve revisión por ser la tocante al tema de esta tesis, debemos destacar, no obstante, que en las últimas dos décadas se han realizado varios trabajos dedicados a abordar este aspecto. La investigación en este ámbito se ha llevado a cabo principalmente desde dos aproximaciones distintas. Por un lado, diferentes autores (por ejemplo Gitsaki, 1999 y Nesselhauf, 2005) han estudiado el nivel colocacional que los alumnos demuestran en su propia escritura (no conocemos ningún estudio dedicado exclusivamente al ámbito de las colocaciones en el que se realice este tipo de análisis en el lenguaje oral), mediante el uso de corpus de textos producidos por los alumnos. Este tipo de trabajos está especialmente indicado si lo que se pretende es evaluar en qué medida y de qué forma se hace uso de las colocaciones en el lenguaje productivo, y si los alumnos son capaces de utilizarlas adecuadamente desde el punto de vista pragmático y comunicativo; de esta manera, en nuestra opinión, están más enfocados hacia la investigación lingüística aplicada en tanto en cuanto se pretende generalizar aspectos concernientes a la adquisición de la competencia colocacional.

Por otro lado, se ha realizado también una serie de estudios encaminados a evaluar la competencia colocacional del estudiante de forma más directa, mediante el diseño e implementación de pruebas de evaluación de diferentes características. En este caso, y a diferencia de la modalidad anterior, suele tratarse de trabajos en los que se diseña un instrumento que permita al profesor-investigador conocer el nivel de competencia de unos sujetos concretos, obteniendo resultados válidos y fiables. Este tipo de evaluación, además, permite seleccionar de antemano el área de conocimiento colocacional (o, dicho de otro modo, el constructo y el contenido) sobre el que se

desea actuar, obteniendo así una información de mayor validez y posiblemente más relevante para las necesidades concretas tanto del evaluador como de los alumnos. Dado que en esta tesis se ha optado por el segundo de estos procedimientos, nos detendremos particularmente en esta última cuestión. No obstante, no podemos dejar de hacer, en primer lugar, una breve mención a los estudios basados en corpus de alumnos.

4.2.1. La evaluación de la competencia colocacional mediante corpus de alumnos

Según apunta Gyllstad (2007) en su exhaustiva revisión de los estudios de evaluación colocacional llevados a cabo hasta el año 2005, los principales trabajos dedicados a medir el grado de competencia de los estudiantes de inglés como L2 a partir de sus propios ensayos son los de Howarth (1996), Granger (1998), Gitsaki (1999) y Nesselhauf (2005). En nuestra opinión, a este listado cabría añadir el estudio llevado a cabo por Lorenz (1999) [1].

En lo que se refiere al análisis realizado por Howarth (1996), éste estudia las colocaciones formadas por la estructura V+N tal y como las emplean 10 alumnos de distintas nacionalidades (sumando un corpus total de casi 23.000 palabras), comparándolas con un corpus de textos producidos por nativos de casi 250.000 palabras. Las principales conclusiones de este estudio, que sin duda hay que tomar con precaución dada la modesta dimensión de la muestra, son: 1) Los hablantes no

[1] Existen otros trabajos dedicados a la evaluación de la competencia fraseológica mediante el análisis del lenguaje que los hablantes no nativos utilizan tanto de forma oral (De Cock et al., 1998; Oppenheim, 2000; Foster, 2001; Adolphs y Durow, 2004) como en su producción escrita (Yorio, 1989; Hyland, 2008). Sin embargo, no nos detendremos en ellos en esta breve revisión ya que su campo de análisis no se reduce a las colocaciones, sino que contemplan distintos tipos de unidades fraseológicas que quedan fuera de nuestro objeto de estudio.

nativos utilizan las colocaciones en menor medida que los nativos (los alumnos utilizan un 25% frente al 34% que se contabilizó en el corpus de nativos); 2) Se encontró muy poca correlación entre la competencia colocacional de los estudiantes y su nivel de conocimiento del inglés general, resultado que merecería una investigación propia por su trascendencia para la adquisición de las colocaciones, y a la vez por la ausencia de investigación al respecto.

En cuanto al estudio realizado por Granger (1998), esta autora utilizó la sección del *International Corpus of Learner English* (ICLE)[2] producida por hablantes cuya lengua nativa es el francés, con 250.000 palabras aproximadamente, y lo comparó, como ya hiciera Howarth, con un corpus de referencia, en este caso un banco de casi 235.000 palabras formado por el *Lovain Essay Corpus* y distintas secciones del *International Corpus of English* y el LOB Corpus. En su trabajo, dedicado a las colocaciones formadas por Adverbio Intensificador + Adjetivo (como, por ejemplo, "*closely related*" y "*deeply convinced*"), Granger concluyó que: 1) Los hablantes no nativos utilizan una menor variedad y cantidad de adverbios intensificadores (por ejemplo "*highly*", "*strongly*" o "*deeply*") que los sujetos nativos; 2) Los alumnos tienden a usar un número limitado de intensificadores (particularmente "*completely*" y "*totally*") de forma marcadamente más frecuente que los nativos, algo atribuible a que se trata de unidades más flexibles en su combinación y muy frecuentes en francés (lengua nativa de los alumnos).

Lorenz (1999) lleva a cabo un estudio muy similar al de Granger ya que compara el uso que los hablantes nativos y no nativos hacen de las colocaciones Adverbio+Adjetivo mediante un corpus de referencia de 218.000 palabras y otro de 300.000 palabras producido por los estudiantes, aunque en este caso los sujetos eran de nacionalidad alemana. Las conclusiones alcanzadas en este trabajo confirman también los resultados de Granger: 1) El repertorio colocacional de los estudiantes es

[2] http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm [Último acceso: 23.02.2009]

muy inferior al de los hablantes nativos; 2) Los alumnos tienden a utilizar con una frecuencia desproporcionada un pequeño número de colocaciones.

Gitsaki (1999), por su parte, investigó el uso que los adolescentes griegos, con diferentes niveles de dominio del inglés, hacen de las colocaciones desde un punto de vista más genérico. Para ello, esta autora compiló un corpus de dimensiones relativamente reducidas (55.000 palabras aproximadamente) y comprobó en qué proporciones se utilizan las diferentes combinaciones sintácticas recogidas por Benson, Benson e Ilson (1986) en su diccionario, a las que Gitsaki añadió otras cuatro, consiguiendo un total de 37 estructuras. Sus principales conclusiones fueron las siguientes: 1) El tipo de combinaciones que los alumnos emplean depende en gran medida de su nivel de conocimiento del inglés. Los alumnos de nivel inicial utilizan preferentemente combinaciones Sujeto+Verbo y Adjetivo+Nombre mientras que los alumnos de nivel medio prefieren estructuras como Preposición+(Determinante)+Nombre o Verbo+Preposición y los de nivel medio-alto utilizan combinaciones como Nombre+Preposición y Sujeto+Verbo+(Objeto)+*that*; 2) La estructura Verbo+Nombre se usa raramente, independientemente del nivel del alumnado, lo que parece indicar que se trata de una combinación de mayor complejidad para el hablante no nativo.

Por último, el estudio llevado a cabo por Nesselhauf (2005), quizá el más amplio y de mayor alcance hasta la fecha de entre los dedicados al análisis de la competencia colocacional a través de los corpus de alumnos, utiliza una sección del ICLE formada por 155.000 palabras, con la cual esta autora analiza el uso que los alumnos universitarios alemanes y austriacos hacen de las colocaciones Verbo+Nombre en inglés. Su investigación arroja importante luz sobre los siguientes aspectos: 1) Dos tercios de las combinaciones V+N recogidas en el corpus constituyen colocaciones aceptables desde el punto de vista fraseológico, lo cual indica, como la propia autora destaca, que se trata de un fenómeno que causa ciertas

dificultades incluso a los alumnos de nivel avanzado; 2) El factor que más comúnmente incide en los errores de los alumnos es la interferencia de la lengua materna; 3) Otro factor que explica un amplio número de los errores es el grado de arbitrariedad y restricción de la combinación. Según su estudio, las colocaciones más idiomáticas tienden a producir menos errores en el lenguaje de los alumnos que las más flexibles.

A pesar de lo interesante de este trabajo y de su innegable valor empírico por el alto número de combinaciones que examina (hasta un total de 2.082), existen dos limitaciones que debemos también destacar. En primer lugar, y como apunta Philip (2007b), para poder valorar el conocimiento colocacional de los estudiantes de la manera que pretenden los trabajos basados en corpus de alumnos, no es suficiente con evaluar el número de errores que éstos producen en el uso de las colocaciones, sino que también es importante considerar la cantidad de colocaciones que se han utilizado, es decir, hasta qué punto se hace uso de este tipo de combinaciones o si se evitan de entrada: "It can (…) be misleading to imply that, e.g. adverb+verb collocations are relatively unproblematic because they are rarely misused, because it may be that adverbs are simply not being used; if so, absence is as significant as error" (ibid.: 1). En nuestra opinión, este aspecto debería ser fundamental en este tipo de trabajos, y sería por tanto deseable que Nesselhauf lo hubiese contemplado. Por otro lado, coincidimos plenamente con Gyllstad (2007) en su observación de que este estudio sería más completo si los datos se hubieran sometido a análisis estadísticos, en lugar de limitarse únicamente a valoraciones de carácter más cualitativo como es el caso.

4.2.2. Evaluación de la competencia colocacional mediante el diseño de tests

En lo que respecta a los estudios dedicados a la evaluación de la competencia colocacional mediante instrumentos de medida diseñados para la elicitación de unas colocaciones determinadas, cabe destacar que, aunque todavía escasos, son varios los trabajos que se han venido produciendo en los últimos años. Según los datos de que disponemos, los autores más reseñables en este sentido son Biskup (1992), Bahns y Eldaw (1993), Farghal y Obiedat (1995), Granger (1998), Gitsaki (1999), Bonk (2001), Mochizuki (2002), Barfield (2003), Gyllstad (2007) y Keshavarz y Salimi (2007)[3]. En términos generales, y como se comprobará seguidamente, los 10 trabajos aquí citados se pueden dividir en dos etapas que corresponderían a las dos décadas que comprenden. Así, mientras que los trabajos producidos durante los años 90 se pueden considerar como los primeros intentos de evaluar el conocimiento colocacional, todavía de forma muy tentativa y poco sistemática, los estudios llevados a cabo en nuestra década suponen un claro avance en este campo y representan en su mayoría investigaciones de mayor envergadura y solidez.

4.2.2.1. Primeros tests de colocaciones

4.2.2.1.1. Biskup (1992)

Comenzando por el primer trabajo que conocemos dedicado a la evaluación de las colocaciones, el realizado por Biskup (1992), debemos señalar que la autora no aporta

[3] Como ya hiciéramos notar en la sección anterior, en esta revisión sólo incluiremos aquellos trabajos dedicados exclusivamente a la evaluación colocacional. Estudios en los que las colocaciones se incluyen como uno de los elementos que conforman un constructo de naturaleza más amplia (por ejemplo Schmitt, 1999; Qian, 2002; Boers et al., 2006) no se tratarán.

una descripción detallada de los procedimientos seguidos en su elaboración —no se especifica cómo se llevó a cabo la selección de contenidos, el número de ítems incluidos, o si éstos constaban de colocaciones descontextualizadas o no—, lo cual, por un lado, muestra claramente que nos encontramos todavía en los inicios de los estudios de evaluación colocacional ya que no se presta atención a aspectos tan relevantes como los anteriores y, por otro, hace que resulte difícil valorar su alcance. Todo lo que sabemos sobre el test en sí es que se trataba de una prueba productiva donde los alumnos encontraban colocaciones de estructura V+N y A+N que debían traducir desde su lengua materna al inglés. En lo que respecta a los resultados obtenidos por esta autora tras administrar el test a 62 universitarios alemanes y polacos, no se apreciaron diferencias en cuanto al número de respuestas correctas obtenidas por los dos grupos de alumnos. Sin embargo, sí se pudo comprobar que en aquellos ítems donde los examinandos no conocían la respuesta, los estudiantes polacos preferían dejarla en blanco mientras que los alemanes eran más proclives a parafrasear la expresión en inglés, lo cual indica, según Biskup, una mayor audacia a la hora de correr riesgos en los sujetos de nacionalidad alemana.

4.2.2.1.2. Bahns y Eldaw (1993)

Un año más tarde, en 1993, Bahns y Eldaw publicaron otro estudio en el que se había llevado a cabo una evaluación de este aspecto fraseológico. Como en el resto de los tests de colocaciones producidos en los 90 (a excepción del propuesto por Granger, como veremos más adelante), se trataba de una prueba diseñada para medir el conocimiento productivo de los estudiantes. Bahns y Eldaw seleccionaron un banco de 40 colocaciones a partir de diferentes obras de referencia como por ejemplo el diccionario BBI y, tras su pilotaje con dos informantes nativos, escogieron las 15 colocaciones en las que éstos habían ofrecido el mismo colocado para un sustantivo

concreto. Así, las 15 combinaciones V+N que incluyeron en su test son las siguientes (por orden de aparición en el test) (Tabla 4.1.):

Keep + diary	Attend + lectures	Take + call
Admit + defeat	Withdraw + money	Do + damage
Cancel + order	Reject + proposal	Whip + cream
Serve + sentence	Refuse + admission	Pay + compliments
Set + watch	Arouse + compassion	Achieve + perfection

Tabla 4.1. Colocaciones incluidas en los tests de Bahns y Eldaw (1993)

Más concretamente, estos autores diseñaron dos pruebas, cada una de las cuales medía las 15 colocaciones anteriores en un formato diferente: en el primer test se empleaba la técnica de la traducción mientras que en el segundo se utilizó el formato "*cloze*"[4], donde los huecos siempre correspondían al colocado. Ambos tests, además, consistían en ítems contextualizados.

Tras administrar el test de traducción a 34 estudiantes y el "*cloze*" a otros 24 sujetos (todos ellos universitarios alemanes), estos investigadores comprobaron que no existían diferencias significativas entre los resultados de ambas medidas (el primero obtuvo una media de aciertos del 53,9% frente al 48,1% del segundo), indicando con ello que los dos formatos parecen funcionar de manera similar en la evaluación de colocaciones. Por otro lado, también realizaron una comparación entre el conocimiento general del vocabulario y el nivel de competencia colocacional de los alumnos, comparación que les llevó a concluir que ambos fenómenos no se

[4] Aunque la técnica "*cloze*" en su formato tradicional consiste en un texto donde se dejan huecos en intervalos regulares (es decir, cada 5 ó 6 palabras se elimina una para dejar un hueco que el alumno debe rellenar), existen distintas variaciones como en el caso del test de Bahns y Eldaw, donde el hueco en la oración no es fijo sino que el investigador decide qué palabras eliminar según distintos criterios (Read, 2000).

desarrollan de forma paralela. No obstante, y como apunta Gyllstad (2007), esta conclusión resulta poco fiable si tenemos en cuenta que la estimación del nivel de competencia léxica general se realizó a partir de los mismos instrumentos de medida (concretamente mediante el test de traducción y calculando el número de palabras individuales que se habían traducido correctamente) y no procedían, pues, de datos independientes.

4.2.2.1.3. Farghal y Obiedat (1995)

Farghal y Obiedat (1995) evaluaron un total de 22 colocaciones entre un grupo de 57 universitarios saudíes. De estos 57 alumnos, 34 (grupo A) recibieron un test de 11 ítems diseñado con la técnica de "rellenar huecos"[5] donde los sujetos debían aportar el colocado a partir de una oración en la que se proporcionaba la base de la colocación y donde la principal pista contextual venía dada por una colocación antónima (por ejemplo: "I prefer _____ tea to strong tea"). Mientras tanto, los 23 candidatos restantes (grupo B) realizaron el test en un formato de traducción, en el que las colocaciones se ofrecían como parte de una oración escrita en árabe y donde los alumnos debían traducir la colocación completa al inglés. Los autores describen este segundo formato como "basically an Arabic version of the English one" (ibid.: 319).

Tras administrar ambas pruebas, la actividad de rellenar huecos obtuvo un 18,3% de aciertos, mientras que la de traducción tan sólo logró un 5,5% de respuestas correctas. A través del análisis de las respuestas erróneas, estos autores también estudiaron el tipo de estrategias que utilizaban los alumnos cuando no conocían la colocación correcta. Así, concluyeron que el recurso más frecuente es el

[5] Este formato, descrito como "fill-in-the-blank form" por Farghal y Obiedat (1995: 319), es igual que el que usaran Bahns y Eldaw (1993), y al que estos autores se referían como "*cloze*".

empleo de un sinónimo (41% en el grupo A y 35% en el grupo B) (aunque en nuestra opinión es muy probable que en muchos casos se trate de un error colocacional en lugar de un recurso), seguido de la omisión ("*avoidance*") (27% y 21%), la transferencia de la lengua materna (10% y 13%) y, por último, la paráfrasis (4% y 25%).

Los resultados obtenidos en este estudio, verdaderamente bajos, no parecen tan sorprendentes, sin embargo, cuando se observa el tipo de oraciones incluidas en los ítems del test. Asumiendo, como decíamos, que existe una relación de antonimia entre las colocaciones seleccionadas, estos autores presentan ítems como "Some people like salty soup, but others like ____ soup" donde la palabra buscada es "bland" o "John is the one with the plain shirt, whereas George is the one with the ____ shirt", en donde la respuesta correcta sería "striped". A la vista de este tipo de preguntas, en las que el contexto no parece ser suficiente para su correcta cumplimentación (a pesar de que fue validado por dos hablantes nativos), y donde observamos que la noción de antonimia empleada de esta forma es muy cuestionable, resulta evidente que el test presenta ciertas limitaciones. Si además consideramos el reducido número de ítems propuesto y el hecho de que no se ofrece información alguna sobre el modo en que se seleccionaron las 22 colocaciones, nos parece que nos encontramos ante un test muy mejorable en términos generales.

4.2.2.1.4. Granger (1998)

En el trabajo llevado a cabo por Granger (1998) al que nos referimos en el apartado anterior, esta investigadora también diseñó un estudio en el que se comparaba la actuación de 56 hablantes nativos con la de otros tantos estudiantes de inglés (cuya lengua nativa era el francés). Para ello, elaboró un test en el que se incluyeron 11 adverbios (*highly*, *seriously*, *readily*, *blissfully*, *vitally*, *fully*, *perfectly*, *heavily*, *bitterly*, *absolutely*, *utterly*), cada uno de ellos seguido de una lista, siempre la misma, de 15 adjetivos

(*significant*, *reliable*, *ill*, *different*, *essential*, *aware*, *miserable*, *available*, *clear*, *happy*, *difficult*, *ignorant*, *impossible*, *cold*, *importat*) de entre los cuales los sujetos debían seleccionar los que colocaban con cada adverbio. Asimismo, se pidió a los participantes que marcaran con un asterisco aquellas asociaciones que le parecieran especialmente frecuentes, ya que, según Granger, de esta forma se podría evaluar el grado de prominencia de estas combinaciones en el lexicón mental de los sujetos.

Los resultados de su investigación fueron los que cabría esperar. Los hablantes no nativos sólo reconocieron como asociaciones especialmente frecuentes un total de 280 combinaciones, frente a las 384 que observaron los nativos (por ejemplo, ningún alumno destacó la combinación "*blissfully ignorant*", y sólo cuatro marcaron "*blissfully happy*"). Además, los estudiantes no sólo consideraron un menor número de colocaciones como frecuentes, sino que algunas de las que seleccionaron no fueron reconocidas por los nativos (por ejemplo "*fully different*" o "*fully impossible*"). Para Granger (ibid.: 152) estos resultados indican que "the learners' sense of salience is not only weak, but also partly misguided". A pesar de que estos resultados son, sin duda, interesantes, opinamos que para poder establecer de forma adecuada el nivel de conocimiento de los hablantes no nativos, es quizá más apropiado utilizar como valores de referencia unos contenidos (en este casos unas colocaciones) seleccionadas de acuerdo con unos objetivos y unos criterios concretos, en lugar de comparar nuestros resultados con los obtenidos por hablantes nativos, que, en términos generales, siempre serán mejores.

4.2.2.1.5. Gitsaki (1999)

Al igual que en el caso anterior, el trabajo de Gitsaki (1999) también se llevó a cabo haciendo uso de las dos aproximaciones: en primer lugar se utilizó un corpus de alumnos como vimos anteriormente y más tarde se empleó un test para evaluar el

conocimiento de los estudiantes sobre un número de colocaciones concretas. Esta prueba consistió en dos técnicas diferentes. Por un lado, Gitsaki propuso un test de traducción donde los alumnos tenían que producir en inglés las colocaciones incluidas en 10 oraciones escritas en griego. La segunda prueba consistía en una tarea de rellenar huecos donde se presentaban oraciones en inglés y se ofrecía una parte de la colocación para que el alumno la completara. Esta segunda tarea contaba con distinto número de ítems según el nivel de sus alumnos (recordemos que se separaron en tres grupos según su nivel de inglés general: inicial, medio y medio-alto), siendo el primero de 50 ítems, el segundo de 65 y el tercero de 90. En lo que respecta a la selección de las colocaciones incluidas, se trataba de combinaciones tomadas a partir de libros de texto usados en los institutos de secundaria griegos y pertenecían a distintas categorías de las enumeradas en el diccionario BBI (en concreto, la tarea de traducción recogía 6 estructuras distintas, mientras que la de rellenar huecos evaluaba 11 tipos de combinaciones diferentes). Asimismo, se estableció como criterio de selección que ninguna de las colocaciones tuviera una correspondencia literal, palabra por palabra, con su traducción griega.

En términos generales, los resultados obtenidos tras la administración de estas pruebas confirmaban los ya observados mediante el análisis del corpus de alumnos. Básicamente, las estructuras Sujeto+Verbo y Adjetivo+Nombre resultaron ser las más frecuentes y las utilizadas por alumnos desde niveles iniciales, mientras que se observó una tendencia a evitar otras combinaciones como Sujeto+Verbo+Objeto+Objeto, Nombre+*that* o Adverbio+Adjetivo. Nos parece especialmente interesante destacar que, como la propia autora explica, las colocaciones V+N volvieron a ser las que mayores dificultades plantearon para sus alumnos, tanto en la traducción como en la tarea de rellenar huecos. Finalmente, y dado que en su trabajo se comparaban grupos con distintos niveles de inglés, Gitsaki también concluyó que existe una relación entre la competencia colocacional y el

conocimiento lingüístico general. Sin embargo, y como apunta Gyllstad (2007), el hecho de que esta comparación se estableciera a partir de tests con distinto número de ítems y también con diferentes proporciones entre los distintos tipos de colocaciones plantea una cierta limitación a sus resultados.

Como apreciación general, podemos observar que los tests producidos durante los años 90 eran pruebas en su mayoría (quizá con la única excepción del trabajo de Gitsaki, que ya presenta una mayor elaboración) de naturaleza muy experimental, con un número de ítems muy reducido y donde los procesos de selección y/o elaboración de contenidos no se realizaron de forma sistemática. En justicia al trabajo de estos autores destacaríamos que se trata de los primeros intentos en este campo de estudio, por lo que su valor como punto de referencia es innegable.

4.2.2.2. Tests de colocaciones de la década actual

Como ya mencionamos, los trabajos producidos durante la década que estamos concluyendo presentan en general una mayor rigurosidad en la evaluación de la competencia colocacional. Nos parece también interesante notar que, mientras que los tests desarrollados en los años 90 eran en su mayoría pruebas de tipo productivo, todos los elaborados a partir del año 2000 están enfocados a medir el conocimiento receptivo, a excepción del trabajo de Bonk (2001) que estudia ambos aspectos.

4.2.2.2.1. Bonk (2001)

El primer estudio que se llevó a cabo en este ámbito después del año 2000 fue el de Bonk (2001), cuyo aspecto mas meritorio es sin duda la incorporación de un instrumento empírico de medida. El propio Bonk destaca que su objetivo es indagar

en la validez y la fiabilidad del test que propone, así como comprobar si existe una correlación entre la competencia lingüística y la colocacional. Todo ello se trató de avalar mediante un riguroso análisis estadístico que condujo a lo que, en nuestra opinión, se puede considerar como una investigación primigenia en el campo de la evaluación colocacional y, muy particularmente, en el estudio de instrumentos de medida eficaces.

Así, en un intento por evaluar empíricamente la utilidad de diferentes métodos de evaluación y las diferencias entre distintos tipos de colocaciones, el test que propone Bonk consta de tres secciones, llegando a un total de 50 ítems: 1) 17 ítems en los que se evalúan combinaciones V+N, con frases en las que aparece un hueco para que el alumno aporte el elemento verbal de la colocación; 2) 17 ítems también con formato de rellenar huecos, en los que se evalúan colocaciones gramaticales de estructura V+Prep y en las que el candidato debe añadir la partícula preposicional; 3) 16 ítems de formato receptivo, en cada uno de los cuales aparecen cuatro oraciones en las que se utiliza un mismo verbo pero en compañía de una base nominal distinta en cada caso. En esta actividad el alumno debe identificar cuál de las cuatro oraciones muestra un uso incorrecto del verbo desde el punto de vista colocacional. Este test fue validado por 10 hablantes nativos y se administró a 98 estudiantes universitarios de procedencia fundamentalmente asiática. Asimismo, y con el fin de poder correlacionar sus resultados con el nivel de competencia general de los alumnos, Bonk también administró una prueba de 49 ítems de inglés general.

Brevemente, en lo que se refiere a los resultados obtenidos por los alumnos, éstos alcanzaron una media de aproximadamente el 50% de aciertos en cada una de las tres secciones (siendo ésta también, por tanto, la media total de respuestas correctas en el test) mientras que esta cifra aumentó hasta el 76% en la prueba de inglés general. En cuanto a los análisis comparativos entre ambos tests, Bonk encontró una correlación de 0,73, lo cual indica que existía un 53% de coincidencia

entre los resultados de ambas pruebas[6]. Cabe destacar, no obstante, que teniendo en cuenta que se trata del único test que examina tanto el conocimiento receptivo como el productivo, era de esperar que este autor llevara a cabo correlaciones entre los dos aspectos. Sin embargo, no se realizó ninguna comparación en este sentido, sin duda cuestión todavía pendiente en el ámbito de la evaluación colocacional.

En lo que respecta a los análisis estadísticos sobre la eficacia del propio test como instrumento de medición, los resultados mostraron un índice de fiabilidad general de 0,83. Este resultado, que se podría calificar de muy satisfactorio[7], contrasta sin embargo con el bajo índice alcanzado por la sección 2 de su test, cuya fiabilidad individual sólo logró un 0,47. Bonk también llevó a cabo un análisis de ítems (tanto de facilidad como de discriminación), comprobando que la mayoría de las 50 preguntas de su test habían funcionado de forma adecuada aunque, de nuevo, la sección 2 aportaba datos algo inferiores a los de las otras dos partes del test. Aunque este autor concluyó que era recomendable descartar esa sección y ampliar las otras dos, en nuestra opinión no parece que la diferencia se deba exclusivamente al formato (puesto que es el mismo que se había empleado en la sección 1), sino que más bien parecería evidenciar una mayor dificultad de las colocaciones de tipo V+Prep.

En general, vemos que se trata de un trabajo riguroso y muy avalado estadísticamente, en el que se evalúa un número aceptable de ítems (50) y en el que se llevó a cabo un análisis de secciones con el fin de comparar la eficacia de distintos formatos en la evaluación colocacional. Cabría desear, no obstante, que se hubiera llevado a cabo una selección más rigurosa de los contenidos del test, ya que las colocaciones parecen haber sido escogidas de forma totalmente intuitiva.

---

[6] En el capítulo 5 (sección 5.3.1.6.), se mostrará una descripción más detallada de los índices de correlación y se ofrecerán los parámetros para su correcta interpretación.

[7] Los parámetros para poder interpretar los índices de fiabilidad adecuadamente se indicarán en la sección 5.2. (capítulo siguiente).

4.2.2.2.2. Mochizuki (2002)

El trabajo de Mochizuki (2002) tenía como principal finalidad comparar los resultados de un test de colocaciones con los obtenidos a partir de un test de relaciones paradigmáticas y otro de competencia léxica general. En lo que respecta al test de colocaciones, el que nos interesa en este caso, se trataba de una prueba que contenía 72 ítems en formato de elección múltiple y donde las palabras se ofrecían de forma totalmente descontextualizada. En cada ítem se presentaba una palabra (que en todos los casos era un nombre, un verbo o un adjetivo) y cuatro opciones entre las que el candidato debía seleccionar la correcta para formar una colocación aceptable. Es muy interesante notar que en este trabajo, y de forma totalmente novedosa, los 72 nombres, adjetivos y verbos que formaban la base de cada ítem fueron seleccionados al azar a partir de listados de frecuencias, algo que sin duda contribuye a añadir validez a este test.

Los resultados obtenidos tras administrar la prueba colocacional (además de las otras dos mencionadas más arriba) a un grupo de 54 universitarios japoneses en dos ocasiones distintas fueron, sin embargo, algo desalentadores. El índice de fiabilidad que obtuvo fue de 0,54 en la primera administración y 0,70 en la segunda, dos valores moderados, según el propio autor reconoce. A pesar de que se trata, como vemos, de una prueba cuya fiabilidad es poco satisfactoria, nos parece muy destacable el hecho de que se incluyera un número elevado de ítems, hasta 72, formados además por colocaciones seleccionadas de acuerdo con criterios de frecuencia.

4.2.2.2.3. Barfield (2003)

El test de colocaciones que propone Barfield consta de 100 ítems formados por colocaciones V+N, además de otros 20 (también de estructura V+N) que no forman

combinaciones aceptables (aunque debemos mencionar que el propio autor reconoce que algunas de estas 20 combinaciones sí se podrían utilizar en algunos contextos particulares como "*create+temperature*" en la oración "*to create a temperature at which certain solid elements become liquid*"). Así, se trata de un test en el que el alumno debe valorar cada uno de los 120 ítems que se le presentan de forma descontextualizada de acuerdo con una escala que va desde el nivel 1 ("*I don't know this combination at all*") hasta el nivel 4 ("*This is definitely a frequent combination*"). Al igual que ya hiciera Mochizuki (2002), Barfield presta especial atención a la selección de los contenidos. En primer lugar este autor seleccionó 40 verbos a partir del conocido listado de inglés académico *Academic Word List* (AWL) (Coxhead, 2000) y de la GSL de West (1953). Seguidamente, extrajo 3 nombres que colocaban con cada uno de los verbos según datos del BoE. Las 20 combinaciones formadas por estructuras V+N que el alumno debe identificar como inaceptables se crearon de forma intuitiva.

Barfield administró su test a un grupo de 93 alumnos japoneses, tras lo cual obtuvo una serie de resultados interesantes. En primer lugar, comprobó que la media de reconocimiento colocacional fue de 2,56 sobre 4 (recordemos que se puntuaba sobre una escala de 4 opciones). Asimismo, evaluó por separado la fiabilidad de los 100 ítems que contaban con colocaciones correctas y la de los 20 "distractores", encontrando valores muy similares y muy satisfactorios en ambos (0,97 en los primeros y 0,93 en los segundos). Merece también especial atención el hecho de que 19 de las 20 colocaciones que obtuvieron mejores resultados estaban compuestas por nombres y verbos que se encontraban entre las 3.000 palabras más frecuentes del BNC, lo cual indica, según apunta el propio autor, que la frecuencia de las palabras individuales que conforman la colocación es un factor que influye en buena medida en el conocimiento colocacional.

Por último, cabe mencionar que Barfield estableció una comparación entre los resultados de este test y los recogidos en otra prueba muy similar, en la que se

evaluaba el reconocimiento de los mismos verbos y nombres, pero de forma individual. Como cabría esperar, los alumnos mostraron mayores conocimientos en este último estudio que en el test de colocaciones, lo cual nos lleva a concluir que conocer las palabras de forma individual no implica necesariamente saber cómo combinarlas. El conocimiento colocacional representa, por tanto, una competencia distinta que merece atención propia tanto desde el punto de vista de la enseñanza como de la evaluación.

4.2.2.2.4. Keshavarz y Salimi (2007)

El test llevado a cabo por Keshavarz y Salimi consta de 50 ítems y se trata, como la mayoría de los elaborados en los últimos años, de una prueba diseñada para medir la competencia colocacional del alumnado a nivel receptivo. Durante el proceso de elaboración de este test, los investigadores llevaron a cabo en primer lugar un pilotaje de 60 ítems. De éstos, 30 eran colocaciones con estructura V+N (por ejemplo, "*peel a banana*" o "*express concern*") mientras que los 30 restantes eran colocaciones gramaticales formadas por Adjetivo+Preposición ("*replete with*"), Nombre+Preposición ("*attitude toward*") y Verbo+ Preposición ("*insist on*"). Estas colocaciones conformaban un test productivo de rellenar huecos que se administró a 30 estudiantes de forma experimental. Así, partiendo de esta primera prueba, Keshavarz y Salimi diseñaron su test definitivo con formato de elección múltiple, donde incluyeron 50 de los ítems pilotados y en el que los distractores estaban formados por las tres respuestas más frecuentes obtenidas en el proceso de pilotaje de la primera versión de su test.

Una vez finalizado el proceso de construcción, los autores administraron el test a otro grupo de 30 alumnos universitarios iraníes. Los datos obtenidos reflejaban una media de aciertos del 74% y un índice de discriminación medio de los ítems de

0,89, resultados en nuestra opinión muy positivos. Además, los resultados de este test se compararon con los de dos pruebas de formato "*cloze*" enfocadas a medir el nivel general de competencia de los alumnos en inglés, obteniendo una correlación de 0,683. Según los propios autores concluyen, estos resultados "strongly suggest that learners' collocational competence and proficiency level are closely and positively associated" (ibid.: 88).

Desde nuestra perspectiva, es interesante destacar el cuidado proceso de pilotaje llevado a cabo durante el periodo de construcción del test. Sin embargo, debemos también mencionar un aspecto que en nuestra opinión limita el alcance de sus resultados: la selección de los contenidos. Como ya sucediera con estudios anteriores, los autores no aportan información alguna sobre los criterios empleados a la hora de seleccionar las colocaciones que pretenden evaluar, algo que sin duda es de especial relevancia desde el punto de vista de la validez de la prueba.

4.2.2.2.5. Gyllstad (2007)

El último de los estudios sobre competencia colocacional que conocemos, y sin lugar a dudas el de mayor envergadura, es el que Gyllstad llevó a cabo en su tesis doctoral. Este autor diseñó dos tests de conocimiento receptivo (COLLEX 5 y COLLMATCH 3) que pretenden ser medidas complementarias que logren una evaluación precisa y fiable de la competencia de los candidatos (en su caso, en el contexto educativo sueco). El primero de ellos, COLLEX 5, (que pasó por cuatro procesos de revisión distintos como se puede inferir por su nombre), consiste en 50 ítems de formato elección múltiple en el que aparece un sustantivo acompañado de tres verbos diferentes de forma descontextualizada. La tarea del alumno, pues, es decidir cuál de las tres combinaciones V+N forma una colocación aceptable. Los ítems se presentan de la siguiente forma (ibid.: 305):

	a	b	c
1. a. do damage b. make damage c. run damage			

En lo que se refiere a la selección de los contenidos, Gyllstad eligió palabras con una alta frecuencia individual. Así, el 88% de los términos que componen tanto las colocaciones correctas como los distractores se encuentran entre las tres primeras bandas del listado JACET 8000, basado en el BNC (Ishikawa et al., 2003, citado por Gyllstad, 2007). En cuanto al 12% restante, éstas pertenecen al cuarto listado y, como el autor explica, hubo de seleccionarlas a pesar de su menor frecuencia para que formaran distractores plausibles en los ítems. Asimismo, Gyllstad realizó un análisis de la frecuencia que las combinaciones resultantes exhibían en el BNC con el fin de incluir colocaciones institucionalizadas como respuestas correctas y expresiones que verdaderamente no se utilizan como distractores.

Por su parte, el test denominado COLLMATCH 3, que también ha sido mejorado en diferentes ocasiones a la luz de los distintos resultados obtenidos durante su laborioso proceso de pilotaje, consiste en 100 ítems también constituidos por combinaciones V+N que se presentan, como sucedía con COLLEX 5, en ítems descontextualizados. En este caso, sin embargo, la tarea es diferente puesto que se ofrece cada una de las combinaciones y el alumno debe indicar si cree que la expresión se utiliza en inglés o no. Se trata, por tanto de un test que utiliza el método si/no (ja/nej en sueco) y que cuenta con un total de 70 colocaciones válidas y 30 distractores. Veamos un ejemplo de este formato (ibid.: 309):

1 have a say

	ja
	nej

En cuanto a la selección de las colocaciones incluidas en este test, el proceso fue similar al anterior. Así, se seleccionaron verbos de entre las primeras 4.000 palabras del listado JACET 8000 y se analizaron los nombres que colocaban con dichos verbos con la ayuda de hablantes nativos y poniendo especial cuidado en seleccionar combinaciones en las que existiera algún grado de arbitrariedad en el uso del verbo (aunque este proceso no se explica en detalle cabe pensar que la selección, en último término, se llevó a cabo de forma intuitiva). Por último, en este caso también se comprobó la frecuencia de las combinaciones en su conjunto, con el fin de garantizar que las respuestas correctas formaran colocaciones verdaderamente institucionalizadas, mientras que los distractores no mostraban una frecuencia significativa en el BNC. Con respecto al proceso de selección de colocaciones llevado a cabo en este caso, parece responder a la concepción de que los dos miembros de una colocación se encuentran al mismo nivel. Como ya destacáramos en el primer capítulo de este trabajo, son muchos los autores (Hausmann, 1989; Mel'čuk, 1998; Granger y Paquot, 2008) que han destacado que, por el contrario, en una colocación suele haber un elemento totalmente independiente que el hablante elige libremente en su uso de la lengua (la base), mientras que el otro es el que está restringido a ésta (el colocado). Si partimos de esta base, no parece adecuado extraer las colocaciones atendiendo a la frecuencia individual de cada una de las palabras que las integran, sino que se debe partir de la frecuencia de la base, para proceder a extraer sus colocados más frecuentes (independientemente de la frecuencia individual de éstos). Por tanto, este proceso de selección de contenidos podría mejorar este trabajo, por otra parte, muy meritorio.

Tras el extenso pilotaje llevado a cabo por este autor, los resultados logrados en su investigación no han sido, sin embargo, los que cabría desear. Con respecto a COLLEX 5, su índice de fiabilidad tras administrarlo a varios grupos de alumnos y hablantes nativos (sumando un total de 269 examinandos) alcanzaba 0,89, aunque

resulta especialmente llamativo que al analizar los resultados de cada grupo de forma individual los índices más bajos se registraran en los tests en los que participaron los alumnos universitarios del curso más avanzado que utilizó (31 sujetos, $\alpha = 0{,}58$) e incluso entre el grupo de nativos (34 sujetos, $\alpha = -0{,}09$). Gyllstad apunta que estos pobres resultados pueden deberse a la homogeneidad de sus grupos (ya que los análisis de fiabilidad dependen en gran medida de la varianza que exista en la población), y se muestra optimista en conseguir futuras mejoras: "Although the low reliability is worrying, I feel confident in reaching higher reliability values with more heterogeneous groups in future test administrations" (ibid.: 174). Sin embargo, sí que parece preocupante en este sentido notar que la media de aciertos lograda por el grupo de hablantes ingleses fue de un 48,9%. En nuestra opinión, este resultado no apunta hacia la homogeneidad del grupo como posible causa de la baja fiabilidad, sino a un posible problema en el diseño del test.

En lo que respecta a COLLMATCH 3, la situación es muy similar. Mientras que el índice de fiabilidad general es de 0,89, cuando este análisis se realiza por cada grupo individual los valores disminuyen desde 0,52 en el caso de grupo de nativos hasta 0,88 entre alumnos universitarios que cursaban el primer semestre de estudios. Cabe destacar, no obstante, que en este caso la media de aciertos por grupo sí está dentro de lo que cabría esperar con respecto a los hablantes nativos (92,9%).

Para concluir, nos gustaría destacar que, a pesar de los inevitables problemas que todo test presenta en sus etapas de pilotaje (y que también se produjeron en nuestra propia investigación como se verá en el capítulo siguiente), el trabajo llevado a cabo por Gyllstad supone, en nuestra opinión, el esfuerzo más reseñable que se ha hecho hasta la fecha sobre el complejo proceso de la evaluación colocacional.

Como hemos podido observar en las páginas anteriores, el área de la evaluación de la competencia colocacional se encuentra todavía en una etapa

incipiente, en la que está todavía casi todo por hacer. Además, la mayoría de los trabajos llevados a cabo hasta ahora se han dedicado de forma casi exclusiva a las combinaciones V+N, obviando el resto de estructuras.

Por otro lado, mientras que los primeros tests estaban principalmente dedicados al análisis de la producción colocacional, los estudios más recientes se han dedicado casi exclusivamente al modo receptivo. Esto ha provocado que no se haya llevado a cabo hasta la fecha ningún análisis comparativo entre ambos aspectos de la competencia colocacional, ni se hayan realizado correlaciones para investigar el grado de coincidencia que existe entre ambos.

En lo que respecta a la selección de ítems, un aspecto verdaderamente importante si queremos dar cumplida cuenta de la validez de un test, son todavía muy escasos los estudios que han hecho uso de las enormes ventajas que ofrecen los corpus actuales para investigar la frecuencia de las colocaciones. Además, aquellos trabajos en los que sí se utilizan técnicas computacionales y/o listados de frecuencias (Mochizuki, 2002; Barfield, 2003; Gyllstad, 2007) sólo se hace uso de la información aportada por un corpus, lo que puede suponer ciertas limitaciones en términos de representatividad. Hasta donde conocemos, no se ha llevado a cabo hasta la fecha ningún intento por comparar la presencia de las colocaciones en más de un corpus. En este sentido, y con la honrosa excepción de Gyllstad, ningún autor aporta datos concretos sobre la frecuencia de las colocaciones empleadas. Esto supone una carencia importante en estos estudios puesto que "[i]n the absence of such information, it is impossible to tell whether learners do not know the items because (...) they are particularly bad at learning collocations or simple because they are so infrequently encountered" (Durrant, 2008: 140).

Por último, nos parece también interesante comentar que, según los datos de que disponemos, no existe tampoco ningún estudio de estas características en el contexto educativo español.

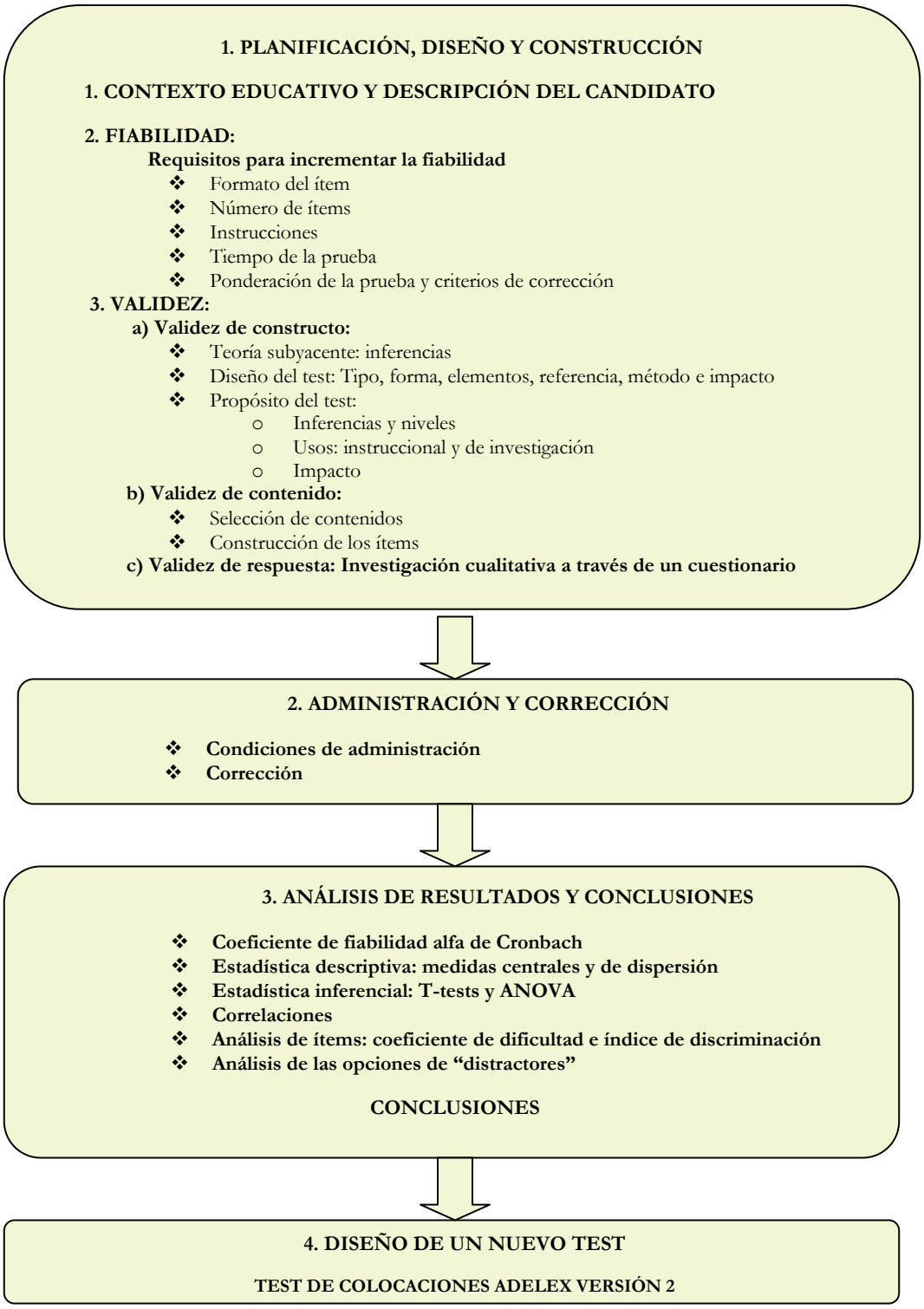
A la vista de esta situación y tomando como punto de partida las investigaciones existentes, emprendimos el proceso de diseño y construcción de nuestro test: Test de Colocaciones ADELEX Versión 1 (TCA1).

4.3. Diseño del Test de Colocaciones ADELEX Versión 1 (TCA1)

Son muchos los autores que proporcionan definiciones de qué se entiende por "test". En términos generales se puede definir como un instrumento de medida que recoge información sobre la actuación de un individuo. Carroll (1968: 46, citado por Bachman, 1990) considera que un test puede ser psicológico o educativo y lo define como "a procedure designed to elicit certain behaviour from which one can make inferences about certain characteristics of an individual". Por su parte, Abad et al. (2004: html) entienden un test como un "conjunto de preguntas (o ítems) administrados a un sujeto para estimar su nivel de conocimientos, inteligencia, actitudes o personalidad". A su vez, un "ítem" es una pregunta individual que forma parte de un test y requiere que el candidato dé una respuesta (Harris y McCann, 1994). Destacaremos por último la definición que ofrece Hughes (1989: 4) refiriéndose al ámbito concreto de la lingüística. Para este autor un test es "any structured attempt to measure language ability".

Los dos aspectos que caracterizan a un test de manera más fundamental son la validez y la fiabilidad (Henning, 1987). Ambos factores, estrechamente relacionados entre sí, resultan imprescindibles para garantizar la utilidad de una prueba. Por ello, dedicaremos las próximas páginas a explicar las actuaciones llevadas a cabo en el diseño y construcción de nuestro test para lograr una medida válida y fiable de la competencia colocacional de nuestros estudiantes.

Es común que el investigador que pretende diseñar un test, lo haga teniendo en cuenta una serie de parámetros que le obliguen a ocuparse de todas aquellas cuestiones que son fundamentales para la construcción de una prueba que dé cuenta de su validez y fiabilidad. Por tanto, para conseguir una cierta sistematización en el diseño de una prueba se puede recurrir a la formulación de un modelo propio que contemple unos criterios que se deben observar para una correcta construcción, o bien podemos atenernos a modelos ya experimentados. En nuestro caso, hemos creído que el modelo diseñado por López-Mezquita (2005), que se ha llevado a cabo dentro del proyecto ADELEX, es muy adecuado para nuestro test por dos razones. En primer lugar, está basado en fuentes muy reputadas y reconocidas en el campo de la evaluación (Thorndike y Hagen, 1980; Madsen, 1983; Carroll y Hall, 1985; Heaton, 1989; Bachman, 1990; Pérez Basanta et al., 1992; Alderson et al., 1995; Bachman y Palmer, 1996; Frary, 2000; Murray, 2002; Abad et al., 2004) y, en segundo lugar, está planificado teniendo en cuenta el contexto del alumnado español. Por todo ello, decidimos que las tres etapas ya clásicas en el diseño de un test, reflejadas en el diagrama de flujo de López-Mezquita (Fig. 4.1.), eran las idóneas puesto que recogen los principios subyacentes de la evaluación: la fiabilidad y validez (un test mide lo que pretende medir y lo mide de forma consistente y precisa). Así, y según López-Mezquita, las tres etapas para la construcción de un test son: 1) planificación, diseño y construcción; 2) administración y corrección; 3) análisis de resultados y conclusiones. En nuestro caso, como se verá, este proceso nos llevó a una cuarta etapa dedicada a la mejora de nuestro test y a la producción de una segunda versión (TCA2).


1. PLANIFICACIÓN, DISEÑO Y CONSTRUCCIÓN
1. CONTEXTO EDUCATIVO Y DESCRIPCIÓN DEL CANDIDATO
2. FIABILIDAD:
Requisitos para incrementar la fiabilidad
Formato del ítem
Número de ítems
Instrucciones
Tiempo de la prueba
Ponderación de la prueba y criterios de corrección
3. VALIDEZ:
a) Validez de constructo:
Teoría subyacente: inferencias
Diseño del test: Tipo, forma, elementos, referencia, método e impacto
Propósito del test:
Inferencias y niveles
Usos: instruccional y de investigación
Impacto
b) Validez de contenido:
Selección de contenidos
Construcción de los ítems
c) Validez de respuesta: Investigación cualitativa a través de un cuestionario
2. ADMINISTRACIÓN Y CORRECCIÓN
Condiciones de administración
Corrección
3. ANÁLISIS DE RESULTADOS Y CONCLUSIONES
Coeficiente de fiabilidad alfa de Cronbach
Estadística descriptiva: medidas centrales y de dispersión
Estadística inferencial: T-tests y ANOVA
Correlaciones
Análisis de ítems: coeficiente de dificultad e índice de discriminación
Análisis de las opciones de "distractores"
CONCLUSIONES
4. DISEÑO DE UN NUEVO TEST
TEST DE COLOCACIONES ADELEX VERSIÓN 2

Estas cuatro etapas, en definitiva, pretenden trazar una hoja de ruta que observe de forma consistente todas aquellas cuestiones que redunden en la consecución de "un buen test". Por tanto, en este capítulo analizaremos las cuestiones referentes a la planificación y administración de un test de colocaciones, para en el capítulo siguiente abordar sus resultados, conclusiones y la elaboración de una versión mejorada.

4.3.1. Planificación, diseño y construcción del Test de Colocaciones ADELEX Versión 1

El nacimiento de un test surge de la necesidad de cubrir un hueco en algún área del campo de la evaluación. En nuestro caso en particular, esta necesidad se planteó en el año 2006, cuando la investigadora llevó a cabo su trabajo para la obtención del DEA (Diploma de Estudios Avanzados). Mediante un test de 80 ítems que pretendía medir el conocimiento colocacional, observamos que éste era muy deficiente entre nuestros alumnos, lo cual les plantearía muchos problemas a la hora de afrontar sus necesidades académicas, sociales y profesionales. Muy someramente estos fueron los resultados obtenidos (Tabla 4.2.):

		Competencia receptiva	Competencia productiva
N	Válidos	63	63
	Perdidos	0	0
Media		46,0317%	30,5159%
Mediana		47,50%	30,00%
Moda		47,50%	30,00%
Desv. Típ.		12,3019%	10,86166%
Varianza		151,338%	117,976%
Rango		65,00%	52,50%
Mínimo		15,00%	7,50%
Máximo		80,00%	60,00%

Tabla 4.2.: Resultados del test de colocaciones administrado en 2006

A pesar de que la prueba tenía importantes limitaciones, y por tanto las conclusiones podrían no ser del todo fiables, los resultados eran tan bajos que nos hicieron cuestionarnos seriamente la competencia colocacional del alumnado español. Este estudio fue el que nos animó a seguir investigando a fin de confirmar o rechazar las tendencias apuntadas en aquel trabajo.

Por tanto, vamos a pasar ahora al detalle de todos aquellos factores que hemos tenido en cuenta para la elaboración de este test y que nos llevan en primer lugar a la descripción del contexto educativo en donde surge el conflicto, y el perfil del candidato sobre el que esta investigación se va a centrar.

4.3.1.1. Contexto educativo y descripción del candidato

Al tratar del contexto educativo, no podemos por menos que referirnos al proyecto ADELEX —"Evaluación y Desarrollo de la Competencia Léxica"— dentro del cual se enmarca nuestra investigación en la Universidad de Granada. Este programa,

desarrollado en dos proyectos nacionales de I+D (2003-2006 y 2007-2010) y coordinado por Carmen Pérez Basanta, tiene como uno de sus principales objetivos medir los niveles de competencia léxica del alumnado universitario español y mejorarlos a través de la enseñanza virtual y las Nuevas Tecnologías gestionadas por plataformas de teleformación (WebCT, Ilias y Moodle). Pero vamos a retrotraernos un poco más atrás para alcanzar a ver las raíces de este trabajo y el contexto en el que se desarrolla. En el año 1999, el Profesor Norbert Schmitt, de la Universidad de Nottingham (Reino Unido) (Schmitt et al., 2001), pretendía validar un test de vocabulario elaborado siguiendo el modelo del conocido *Vocabulary Levels Test* (Nation, 1990). Para ello necesitaba la colaboración de otras universidades ─tanto de países de lenguas romances como germánicas e incluso orientales─ por lo que solicitó la colaboración de la Dra. Pérez Basanta para contar con un amplio alumnado de la Universidad de Granada que coadyuvase a validar dicho test. Los resultados obtenidos de esta investigación, publicada en la revista *Language Testing* en el año 2001, mostraban claramente las deficiencias léxicas del alumnado español.

Posteriormente, la Dra. Pérez Basanta decidió replicar este trabajo en el 2001 (Pérez Basanta, 2005), esta vez con alumnos de último curso de la licenciatura de Filología Inglesa. Los resultados volvieron a confirmar los datos de Schmitt et al. (2001). Era muy evidente que con los niveles de competencia léxica que los alumnos mostraban no podrían hacer frente ni a las demandas académicas ni a las profesionales con que se encontrarían en el futuro. La situación parecía tan preocupante que un grupo de profesores, coordinados por la Dra. Pérez Basanta, solicitó y obtuvo financiación del Vicerrectorado de Planificación, Calidad y Evaluación Docente de la Universidad de Granada para iniciar el Proyecto de Innovación Pedagógica "ADELEX: A Course for Assessing and Developing Lexical Competence on the Internet". En octubre de 2002 el proyecto fue finalista de los Premios de Innovación Pedagógica, convocados por dicho Vicerrectorado, y obtuvo

una Mención Honorífica. Posteriormente el proyecto ha sido galardonado con el premio Sello Europeo a la Innovación en la Enseñanza y Aprendizaje de Lenguas Extranjeras en la convocatoria de 2003. En ese mismo año, la investigación se convirtió en un Proyecto de Investigación y Desarrollo (Ref: BFF2003-02561) financiado por el Ministerio de Ciencia y Tecnología, con la finalidad de trabajar el campo de la evaluación y desarrollo del "conocimiento de la palabra" en sus diferentes facetas. En la actualidad, este proyecto ha dado un paso más y se dirige tanto a la evaluación de la competencia colocacional como a su posterior mejora mediante un tratamiento pedagógico a través de las Nuevas Tecnologías.

En cuanto al perfil del candidato en este trabajo, no quisimos constreñir nuestro campo de actuación al alumnado de Filología Inglesa de nuestra universidad, sino que se pensó que los alumnos de Traducción e Interpretación con Inglés como primera lengua meta podían aportar datos muy iluminadores sobre la situación actual en torno al conocimiento colocacional del alumnado español. Así pues, los sujetos elegidos para realizar este test de diagnóstico y emplazamiento fueron alumnos de la Licenciatura de Filología Inglesa de la Universidad de Granada y de Almería, y también de la Licenciatura de Traducción e Interpretación de la Universidad de Granada y de la Universitat Jaume I (Castellón).

En este estudio colaboraron 311 alumnos, 243 mujeres y 68 hombres, cuyas edades estaban comprendidas entre los 18 y los 25 años en la mayoría de los casos (295 sujetos entre 18 y 25; 11 sujetos entre 26 y 35; 4 sujetos entre 36 y 45 y tan solo un alumno mayor de 46). Aunque el test fue completado por alumnos de una gran cantidad de nacionalidades distintas, y dado que el propósito de este estudio era investigar la situación en lo que respecta a alumnos que tienen el español como lengua materna, sólo los resultados de éstos últimos fueron tenidos en cuenta en nuestros análisis. Dado que este trabajo pretendía asimismo comprobar la

competencia colocacional de los alumnos en diferentes niveles de aprendizaje, el test se administró a estudiantes de los cuatro cursos de ambas licenciaturas. Cabe añadir con respecto a este particular que en el caso de Filología Inglesa, y con el fin de poder comparar los resultados con los de Traducción, no se utilizó el 5º curso. Por otro lado, dada la flexibilidad del 3er curso de Filología (compuesto únicamente por asignaturas optativas), los alumnos que se encuentran en su tercer año de estudios normalmente cursan las asignaturas de 4º o se encuentran en su estancia Erasmus. Por este motivo, al administrar nuestro test en las clases de 4º, se pidió a los alumnos que indicaran si se encontraban en su tercer o cuarto año de estudios para así diferenciar ambos niveles. La distribución de alumnos por cursos y licenciaturas es la siguiente (Tabla 4.3.):

	Filología Inglesa	Traducción e Interpretación	Total
Primero	46	34	**80**
Segundo	29	42	**71**
Tercero	38	57	**95**
Cuarto	23	42	**65**
Total	**136**	**175**	**311**

Tabla 4.3.: Distribución de candidatos por cursos y licenciaturas

4.3.1.2. La fiabilidad de los tests

La fiabilidad es un concepto clásico de la teoría de la evaluación que se fundamenta en el hecho de que las puntuaciones de un test deben ser consistentes, es decir deben producir resultados similares si el test se aplica a los mismos alumnos en distintas

ocasiones bajo condiciones de administración parecidas. Por tanto, un test es fiable cuando mide las habilidades de los alumnos de forma precisa y exacta (Hughes, 1989; Pérez Basanta, 1995).

Fundamentalmente, la fiabilidad tiene que ver con cuestiones intrínsecas del test como sus **contenidos** en cuanto a la elaboración de los ítems, el factor azar, la claridad de las instrucciones, etc. —fiabilidad intrínseca—; y por otro lado, existen factores externos como el propio **candidato**, la **administración**, la **corrección** y la **calificación del examen** —fiabilidad extrínseca— que también afectan a su medida.

Existe ya un listado clásico elaborado por Hughes (1989) que especifica una serie de cuestiones que se deberían tener en cuenta para incrementar la fiabilidad:

- ✓ El formato del ítem debe fomentar una **puntuación objetiva**.
- ✓ El test debe tener un **número suficiente** de ítems.
- ✓ La confección de los ítems debe ser **revisada por más de un evaluador**, es decir, se debe someter al juicio de otros colegas para evitar el riesgo de ítems mal construidos o que se pueden prestar a una cierta ambigüedad.
- ✓ Se debe prestar especial atención a que las **instrucciones** sean claras.
- ✓ El **diseño tampoco debe inducir a ningún tipo de confusión o mala interpretación**, siendo siempre claramente legible.
- ✓ Es importante que los candidatos tengan una **cierta familiaridad** con el tipo de ítem y no pierdan tiempo en interpretar qué requiere una determinada actividad o tarea.
- ✓ Las **condiciones de administración deben ser muy similares** en caso de que se apliquen a **grupos diferentes**.
- ✓ Siempre es recomendable **identificar a los candidatos mediante un número** y no por el nombre, con objeto de garantizar el anonimato.

Estos requisitos fueron rigurosamente observados en el diseño de nuestro test, a excepción del último en que, a diferencia de lo propuesto por Hughes, se les pidió a los candidatos que incluyesen sus nombres. La razón de esta decisión tuvo que ver con el riesgo de que ante un test de una importante longitud, como era el que se les presentaba, y sin aparentes repercusiones en las calificaciones de los candidatos, no hicieran el esfuerzo de cumplimentarlo correctamente. Así, consideramos que la constatación de la autoría podía minimizar dicho factor perturbador.

Además de estos factores, y puesto que existe un procedimiento estadístico que establece la fiabilidad interna de un test, se decidió que, una vez corregido, aplicaríamos la prueba conocida como **alfa de Cronbach**, que indica el grado de precisión y consistencia interna del examen; en este caso y puesto que toda la estadística se calcularía con el programa informático SPSS v.15 (*Statistical Package for Social Sciences*), ese índice alfa de consistencia interna que oscila entre 0 y 1, se tendría muy en cuenta para la comprobación de la fiabilidad. Según todos los estudios solventes consultados (Abad et al., 2004; Hughes, 1989), los valores ideales deberían estar por encima de 0,9 (sobre esta cuestión volveremos en el capítulo siguiente).

A continuación, nos vamos a ocupar más detenidamente de aspectos estrechamente relacionados con la fiabilidad: el formato y el número de ítems, la ponderación de la prueba, el tiempo y las instrucciones.

4.3.1.2.1. El formato del ítem

Los métodos utilizados en la evaluación léxica han sido objeto de estudio durante muchos años, bien como tests receptivos (Nation, 1983; Meara y Buxton, 1987), productivos (Arnaud, 1992; Laufer y Nation, 1995; Laufer y Nation, 1999), o de asociación de palabras (Read, 1993). En términos generales, podemos decir que existe un consenso sobre aquellas técnicas que se consideran más eficaces para

evaluar el vocabulario, ya que, si existe una opinión unánime en este campo de estudio, ésta se refiere a la necesidad de presentar un número importante de ítems para que la prueba se acepte como muestra representativa. Quizá ésta es la razón de que los ítems objetivos y, particularmente, los de elección múltiple, hayan sido los más utilizados tradicionalmente en la evaluación del léxico ya que se prestan a una mayor practicabilidad, es decir, a crear un test económico en su proceso de administración y corrección.

El experto por excelencia en evaluación del vocabulario, John Read (2000), contempla tres parámetros principales: 1) evaluación a través de ítems independientes ("*discrete*") o integrados ("*embedded*"), 2) si el vocabulario se evalúa aisladamente ("*selective*") o dentro de otras destrezas ("*comprehensive*"), y, por último, 3) teniendo en cuenta la presencia ("*context-dependent*") o ausencia de un contexto ("*context-independent*"). Así, este autor ofrece la siguiente tabla (Tabla 4.4.):

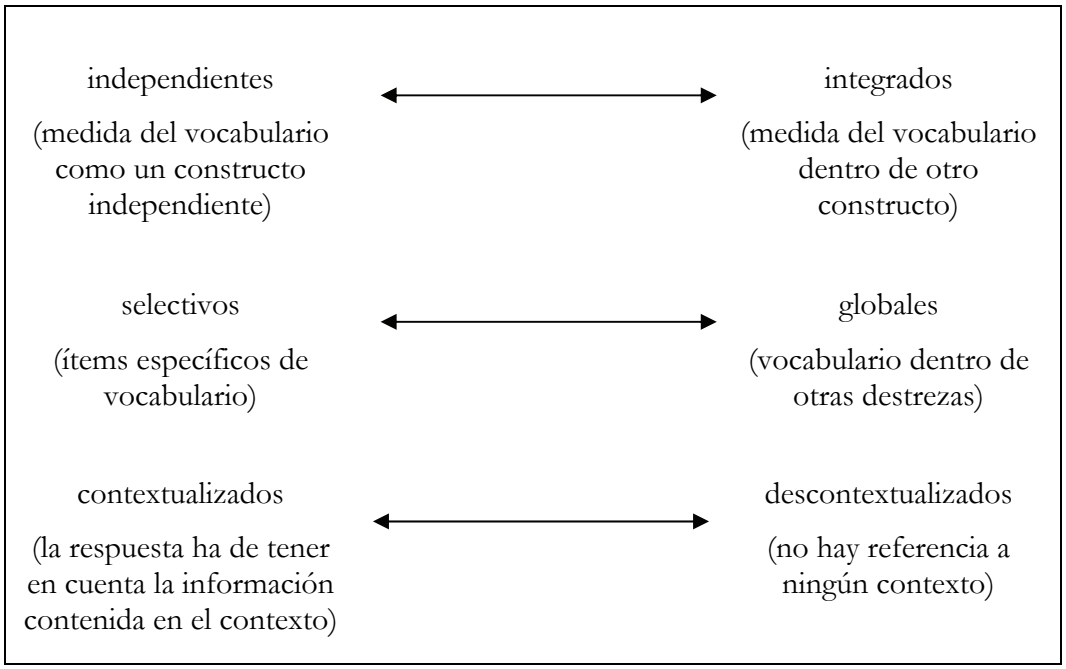


Tabla 4.4.: Dimensiones en la evaluación léxica (Read, 2000, citado por López-Mezquita, 2005: 584)

En nuestro caso, y atendiendo especialmente a las recomendaciones de Hughes (1989) enumeradas más arriba, partíamos de una serie de consideraciones que no queríamos dejar de observar. En primer lugar, y a la vista de la poca investigación existente en torno a la evaluación del conocimiento de las colocaciones, no queríamos utilizar un solo formato, sino experimentar con varios de ellos. Además, a ser posible, buscábamos métodos objetivos con ítems dicotómicos que contribuyeran a otorgar fiabilidad a nuestra prueba. En tercer lugar, nos decantábamos por los métodos que en la tabla 4.4. aparecen a la izquierda, excepto en el tercer apartado donde preferimos también investigar ambas facetas: es decir, consideramos el constructo de la colocación como un aspecto independiente, por lo cual íbamos a formular ítems específicos y no integrados en otras destrezas, y ofreceríamos ítems en contexto y descontextualizados. Por último, buscamos formatos que reprodujeran las actividades con las que se practican las colocaciones en clase para que resultaran familiares para los alumnos. Con estas premisas en mente, elegimos cuatro métodos que ahora pasamos a describir y que aparecieron como cuatro secciones diferentes del test (Secciones A, B, C y D):

1. *C-test.*
2. Traducción del colocado.
3. Detección del intruso.
4. Elección múltiple.

A continuación, explicaremos en qué consiste cada una de estas técnicas y las razones que nos llevaron a escogerlas.

a. Formato "*c-test*" (Sección A)

Para evaluar el conocimiento productivo colocacional elegimos en primer lugar un formato adaptado del conocido método *c-test*, al que nos referiremos con el mismo nombre. La técnica *c-test*, creada por los investigadores alemanes Klein-Braley y Raatz (1984), consiste en presentar párrafos en los que una de cada dos palabras es mutilada por la mitad (dejando la cifra de letras más reducida cuando el número total de letras de la palabra es impar; por ejemplo dejando las dos primeras letras en una palabra de cinco). Este formato, que es a su vez una adaptación de la conocida técnica "*cloze*" en la que se eliminan palabras completas a intervalos fijos (es decir, cada 6 ó 7 palabras del texto, por ejemplo), fue utilizado por Laufer y Nation (1999) con ligeras variaciones para el propósito de la evaluación léxica. En la versión utilizada por estos autores, se prefirió utilizar oraciones en lugar de párrafos y sólo una palabra, la palabra meta, era mutilada. Además, y puesto que el suyo era un test productivo, Laufer y Nation redujeron la cantidad de letras que se conservaban de las palabras mutiladas hasta dejarla en el número necesario para que la respuesta no presentara ambigüedad. Mostramos a continuación dos ítems producidos por estos autores para ilustrar lo anterior (ibid.: 49):

1. You must be awa____ that very few jobs are available.
2. The organisers li____ the number of participants to fifty.

Tras el pilotaje de este tipo de formato, Laufer y Nation (ibid.: 33) concluyeron lo siguiente:

> The controlled-production vocabulary-levels test was found to be reliable, valid (in that the levels distinguished between different proficiency groups) and practical. There was a satisfactory degree of equivalence between two equivalent forms of the test.

Basándonos, pues, en esta investigación, decidimos utilizar en nuestro test un formato similar para evaluar la capacidad productiva del alumno. En nuestro caso, sin embargo, se trataba de medir el conocimiento colocacional, por lo que no parecía conveniente incluir todas las letras necesarias para que la respuesta (el colocado) fuese evidente por sí misma, sino que pretendíamos comprobar si la base de la colocación llevaba al candidato al colocado correcto. Por otro lado, tampoco consideramos apropiado utilizar un formato de rellenar huecos, sin ninguna letra como pista y en el que el alumno pudiese incluir cualquier palabra semánticamente correcta. Dada la flexibilidad combinatoria de muchas colocaciones, esto llevaría a producir ítems en los que cabrían muchas respuestas correctas y, por tanto, poco fiables. Así, concluimos que ofrecer solamente la **primera letra del colocado** sería suficiente (para que nuestro test no se convirtiera en una prueba de competencia léxica) y a la vez necesario (para limitar las respuestas y contar así con un test más objetivo).

Se consideró asimismo que esta técnica era muy apropiada ya que presentaba los ítems de forma contextualizada, algo necesario en un formato productivo de este tipo dada la polisemia de los sustantivos que estábamos empleando como bases de nuestras colocaciones. Además, Nation (2001: 353) apunta que "[t]he value of context may be to orient the learner to the correct part of speech and, by more closely resembling conditions of normal use, encourage normal access to the meaning". En nuestro caso, opinamos que las dos ventajas señaladas por este autor serían especialmente útiles.

Así pues, el primer formato utilizado en nuestro test constaría de oraciones que incluirían una base de nuestro listado de colocaciones y un hueco proporcionando la primera letra del colocado que buscábamos. El procedimiento detallado de la construcción de los ítems tanto de esta como de las otras tres

secciones del test se abordará más adelante, cuando tratemos de la validez de contenido (sección 4.3.1.3.2.).

b. Traducción del colocado (Sección B)

El segundo formato que utilizamos en nuestro test fue el de la traducción. Somos conscientes de que, desde el campo de la evaluación de vocabulario en general, la traducción constituye un método tradicionalmente poco valorado. Muy posiblemente como consecuencia del gran impacto que ha tenido el enfoque comunicativo en las últimas décadas, los profesores se han mostrado generalmente reacios a utilizar la lengua materna tanto en la enseñanza como en la evaluación de segundas lenguas. En este sentido, sin embargo, nos parece muy destacable que una autoridad en el área de los estudios léxicos como es Paul Nation (2001: 351) asegure: "There is a general feeling that first language translations should not be used in the teaching and testing of vocabulary. This attitude is quite wrong".

Por otro lado, una de las razones fundamentales que nos hizo decantarnos por este método fue el hecho de que, como vimos en la revisión de los estudios realizados en este campo, un considerable número de los trabajos en los que se ha tratado de evaluar la dimensión productiva de la competencia colocacional ha hecho uso de esta técnica. Así pues, dado que parece ser un método que, de manera más o menos intuitiva, se ha considerado siempre como una forma adecuada de medir esta faceta léxica, nos pareció oportuno incluirlo en nuestra investigación y tratar de comprobar de forma empírica si verdaderamente se trata de una técnica eficaz.

Además, una última consideración que nos animaba a hacer uso de esta técnica es que pensábamos producir un test que fuese adecuado no sólo para los alumnos de Filología sino también para los de Traducción y opinamos que este método podría ser no sólo motivador para los estudiantes sino también oportuno

desde el punto de vista de la fiabilidad, puesto que se trata de una actividad claramente familiar para ellos.

Por tanto, en esta sección incluimos ítems en los que el alumno contaba con una colocación en español y con la base del equivalente de dicha colocación en inglés, siendo la tarea del alumno completar la combinación inglesa con el colocado. La razón por la que decidimos aportar la base inglesa, es decir, el sustantivo, y no pedir al examinando que tradujera la combinación completa tiene que ver con la validez de constructo de nuestro test. Dado que, a nuestro entender, la base de la colocación es un elemento libremente seleccionado por el hablante, mientras que el colocado es el elemento restringido, lo que pretendíamos medir en este test es si el alumno conoce las restricciones que existen en la combinación, o, dicho de otro modo, el colocado que cada base exige. En nuestra opinión, pedirle al candidato que aporte no sólo el colocado sino también la base de la colocación implica un constructo diferente, puesto que se estaría evaluando su conocimiento léxico, es decir, si conoce el sustantivo en cuestión de forma productiva, y no sólo su dominio de la combinatoria colocacional o, en otras palabras, su capacidad para combinar el nombre con unos colocados frecuentes. Como ya dijimos, volveremos sobre los pormenores de la construcción de los ítems cuando abordemos las cuestiones relativas a la validez**.**

c. Detección del intruso (Sección C)

El método "detección del intruso", conocido popularmente en inglés como "*odd-one-out*", es una técnica en la que el alumno cuenta con un número de opciones, entre las cuales tiene que identificar la incorrecta o la que no encaja con el resto. En el caso de las colocaciones, se presenta una base junto a una serie de posibles colocados entre los cuales hay uno erróneo, que el alumno seleccionará.

En esta técnica, empleada para medir la competencia colocacional receptiva, nuestro principal objetivo era evaluar el espectro colocacional de nuestros sustantivos. Dado que, gracias al banco de colocaciones compilado para esta investigación, contábamos con listados muy completos de los diferentes colocados que acompañan a las bases de forma frecuente, consideramos que sería interesante evaluar varios de los colocados de cada nombre, o lo que podríamos denominar el grado de "riqueza colocacional" de los candidatos. La importancia de este aspecto, además, ha sido también manifestada por los autores de las pruebas diseñadas en el prestigioso centro UCLES (*University of Cambridge Local Examinations Syndicate*). Como apunta Hargreaves (2000: 220-221), uno de los miembros de este centro, al evaluar el nivel de competencia léxica de los examinandos es necesario analizar "the learner's ability to recognise not just one collocation (…), but a **range** of collocations in which a word occurs" (énfasis en el original).

Por otro lado, debemos también destacar que otra de las principales razones que nos impulsó a utilizar este método fue que se trata de una técnica muy común en los libros de texto y los manuales para la enseñanza de las colocaciones. Por ejemplo, en la obra *Teaching Collocation: Further Developments in the Lexical Approach*, al que ya nos referimos al inicio de este capítulo, ésta es una de las técnicas que se recomiendan para trabajar las colocaciones en el aula. Los autores lo ilustran con el siguiente ejemplo (Hill, Lewis y Lewis, 2000: 113) (Fig. 4.2.):

Odd verb out

One verb in each line does not collocate with the noun. Cross out the one which does not fit.

1. accept, act on, disregard, follow, ignore, make, solicit, take	ADVICE
2. come up with, do, expect, get, require, supply	AN ANSWER
3. build up, close down, set up, put off, take over, wind up	A BUSINESS
4. deal with, do, examine, ignore, reject, respond to	A COMPLAINT
5. accept, answer, come in for, give rise to, make, reject	CRITICISM
6. describe, do, enjoy, have, recall, share	AN EXPERIENCE
7. crash, finish, hire, park, repair, run, service, start, write off	A CAR
8. arrange, do, gatecrash, go to, have, throw	A PARTY

Now try these more difficult words:

1. acclaim, disparage, exaggerate, praise, reduce	AN ACHIEVEMENT
2. come to, decide, endorse, implement, reach, sign	AN AGREEMENT
3. analyse, determine, establish, make, study, trace	THE CAUSE

…

Fig. 4.2.: Actividad de detección del intruso (Hill, Lewis y Lewis, 2000: 113)

Así pues, consideramos que el hecho de que este formato resultase familiar para los estudiantes supondría un aspecto positivo de cara a lograr un instrumento de medida fiable. Otras razones que hacían de esta técnica un método atractivo para nuestra investigación es que se trata de un formato objetivo que presenta varias ventajas desde el punto de vista de la practicabilidad: 1) es un método de construcción relativamente sencilla desde el punto de vista de la búsqueda de distractores ya que, evidentemente, sólo es necesario uno, y 2) permite una rápida corrección y una calificación totalmente objetiva puesto que se trata de ítems dicotómicos.

Por todas las consideraciones anteriores, seleccionamos este método en nuestro estudio. Además, en este caso decidimos que consistiera en ítems

descontextualizados puesto que el contexto no se hacía realmente necesario para poder seleccionar la combinación "intrusa" ya que ésta sería una expresión que no se utiliza en inglés (y no una colocación aceptable en ciertos contextos pero no en el del ítem en concreto). Esto nos llevó a considerar que era una buena oportunidad para reducir el esfuerzo del alumno y abreviar su tarea. El segundo motivo por el que en esta ocasión decidimos no utilizar un contexto fue porque, como ya dijimos, nuestra intención era experimentar con ambas variables (con y sin contexto) y tratar de determinar si una ofrece mejores resultados que la otra.

Por último, debemos destacar que, dado que este formato es, al fin y al cabo, una prueba de elección múltiple que presentaba por tanto ciertos riesgos, decidimos añadir una opción extra en nuestros ítems: "*no wrong collocation*" ("ninguna es incorrecta"). Con ello, pretendíamos evitar que el alumno pudiera utilizar la contestación por descarte de forma indiscriminada, lo cual contribuiría a aumentar la fiabilidad de la prueba. Además, en este caso se optó por utilizar la expresión en forma negativa "*no wrong collocation*" en lugar de la positiva "*all of them*" ("todas son correctas") para que el alumno no tuviera que cambiar de un enfoque negativo (buscar la incorrecta) a uno positivo en la última opción (todas son correctas), tratando de evitar así posibles confusiones. Pero pasemos a tratar de la técnica de opción múltiple tradicional, donde explicaremos estos aspectos de forma más detallada.

d. Elección múltiple (Sección D)

Como ya dijimos, el método de elección múltiple es quizá el formato por excelencia en la evaluación léxica. Se trata, como sabemos, de una técnica que permite una evaluación objetiva, donde no interviene el juicio del corrector y, por tanto, altamente fiable, fácil y rápido de corregir. Asimismo, y por esta misma razón, es también un

tipo de test fácilmente virtualizable y que permite evaluar un alto número de ítems mediante una administración relativamente rápida. En opinion de Read (1997: 307), "since well-designed multiple-choice vocabulary items have excellent technical characteristics, they are desirable items to include in a language test if one gives priority to reliability and to purely correlational measures of validity".

No obstante, debemos mencionar que este formato presenta también una serie de problemas a los que se debe prestar especial atención. Wesche y Paribakht (1996) destacan, entre otros aspectos, que no resulta sencillo construir ítems en este formato, y a menudo se convierte en un proceso laborioso que exige de un adecuado pilotaje y refinamiento. Una de las principales razones de esta dificultad es que debemos asegurarnos de que los distractores estén bien seleccionados para que, por un lado, sean plausibles y supongan una verdadera medición del nivel de conocimiento del alumno y, por otro, sean indudablemente incorrectos y no den lugar a ambigüedad. Por otro lado, también destacan que el principal problema que se plantea en esta actividad es que existe una alta probabilidad de que influya el factor azar. En este sentido, en un test con tres opciones, siempre existirá un 33% de posibilidades de responder correctamente aunque no se conozca la respuesta. Muy en relación con lo anterior, este formato también se presta al acierto por descarte que mencionábamos anteriormente, donde en realidad estaríamos evaluando el conocimiento de lo que no es correcto, en lugar de lo que el alumno verdaderamente conoce de la lengua.

Como destaca López-Mezquita (2005), distintos autores han tomado diferentes medidas para tratar de prevenir estos riesgos. En primer lugar, algunos investigadores consideran que se debe aumentar el número de opciones del ítem, puesto que resulta evidente que el porcentaje de aciertos debidos al azar dependerá del número de alternativas que presente. Así, Alderson et al. (1995) consideran recomendable que los ítems de opción múltiple incluyan hasta cuatro posibles

respuestas, mientras que Heaton (1989) estima oportuno incluso aumentarlo hasta cinco. No obstante, expertos en psicometría como Muñiz (1998) y Abad, Olea y Ponsoda (2001), declaran que el número óptimo de alternativas que debe contener un ítem de opción múltiple es tres.

Además del número de opciones, también se ha postulado (López-Mezquita, 2005) que la inclusión de la respuesta "ninguna de las anteriores" ("*none of these*") o "todas las anteriores" ("*all of them*", o, en nuestro caso "*no wrong collocation*") como última opción de cada ítem puede contribuir a minimizar la incidencia del factor azar. Dado que en un principio no todos los autores parecían estar de acuerdo y se consideraba dudosa la capacidad de esta opción para reducir la probabilidad de acierto por azar (Abad et al., 2004), López-Mezquita llevó a cabo una interesante investigación en la que se compararon los resultados de un grupo experimental que realizó un test que incluía la opción "ninguna de las anteriores" en la mitad de sus preguntas y un grupo de control que lo llevó a cabo sin esa opción. Mediante un análisis estadístico de regresión lineal, que permite predecir cuál sería el hipotético comportamiento de una variable (variable dependiente) una vez que se conoce el valor de otra (variable independiente), esta autora pudo determinar que los resultados del grupo experimental habrían sido superiores (es decir habrían acertado más preguntas por azar) en la mitad del test que contenía la opción "*none of these*" si esta opción no hubiese estado presente (concretamente habría existido una diferencia de 6,7 puntos en la media).

Así pues, teniendo en cuenta las ventajas de este método en cuanto a objetividad, validez y practicabilidad, y considerando además que existen ya métodos empíricamente validados para minimizar el efecto del factor azar, parecía claro que se trata de un método que puede resultar eficaz para la evaluación colocacional. En nuestro caso, y atendiendo a las recomendaciones de Abad, Olea y Ponsoda (2001), decidimos incluir tres opciones tanto en la sección de detección del intruso como en

la de opción múltiple, aunque, dados los concluyentes resultados aportados por López-Mezquita (2005), decidimos añadir también una cuarta opción. En la detección del intruso sería "*no wrong collocation*" y en la opción múltiple "*none of these*". Como se ha comprobado, esta última opción contribuiría a reducir la probabilidad de aciertos por azar.

Aunque este aspecto se tratará con más detenimiento cuando veamos cómo se construyeron los ítems de nuestro test, no nos gustaría concluir esta sección sin mencionar que en este caso no se utilizaron ítems descontextualizados, como sucedía en el formato anterior, sino que se presentaban dentro de una definición que pretendía funcionar a modo de contexto informando al alumno sobre la situación en la que se requeriría la colocación meta (por ejemplo: "*When you gain money by selling things you* _____ *money*"). De esta manera, aunque los distractores eran combinaciones que no se utilizan en la lengua, la inclusión del contexto parecía facilitar en cierta medida la comprensión del ítem haciéndolo más natural (Nation, 2001) y, además, nos permitía comparar, como ya dijimos, formatos contextualizados y no contextualizados.

Un último apunte en cuanto a los distintos formatos de nuestro test que nos es obligado destacar es la secuenciación. Considerando que, en principio, partíamos de la hipótesis de que las tareas receptivas resultarían más sencillas para el candidato y necesitarían un menor esfuerzo cognitivo para su cumplimentación, decidimos que constituyeran la segunda mitad del test para que el alumno las realizara cuando ya estuviera algo más cansado, mientras que las tareas productivas se evaluarían mejor si el alumno las realizaba al comienzo. Por esta razón, los formatos *c-test* y traducción corresponden a las Secciones A y B, mientras que la detección del intruso y la opción múltiple constituyen las Secciones C y D. Teniendo también en cuenta el factor cansancio y atendiendo a las recomendaciones de Dörnyei (2003), la información

personal que los alumnos tenían que aportar en la forma de un breve cuestionario se incluyeron al final del test, puesto que requiere menos concentración.

4.3.1.2.2. El número de ítems

Como indicaba Hughes en sus recomendaciones para aumentar el índice de consistencia interna, el número de ítems de que consta un test no es una cuestión baladí. Nation (2001: 345) advierte que "a good vocabulary test has plenty of items (around 30 is probable a minimum for a reliable test)" mientras que otros autores (Meara, 2006, citado por Gyllstad, 2007) consideran que 50 ítems sería un número cercano a lo ideal. Así pues se consideró que, con el fin de lograr pruebas fiables en cada una de las cuatro secciones de nuestro test, 50 ítems por cada una de ellas sería un número adecuado. En total, nuestro test contaba por tanto con 200 ítems.

Esta decisión, como veremos más adelante, no fue tan acertada como en un principio pudo parecer porque, en el caso de un test de colocaciones que supone un gran esfuerzo cognitivo, produjo un efecto cansancio muy acusado que quizá subestimamos en un principio. En el capítulo siguiente comprobaremos que, a la vista de nuestros resultados, parece más adecuado reducir el número total de ítems para lograr un test más práctico y fiable.

4.3.1.2.3. Las instrucciones

Otro de los aspectos que puede influir de forma decisiva en la fiabilidad de una prueba son las instrucciones, puesto que si éstas no son lo suficientemente claras y precisas, pueden condicionar la actuación del alumno haciendo así que intervengan factores externos en nuestra medición. En este caso, optamos por incluir las instrucciones de cada sección en inglés para que existiese continuidad en la prueba,

pero durante la administración del test nos aseguramos de que los examinandos comprendieran claramente (con explicaciones en español) lo que debían hacer en cada sección, a fin de evitar confusiones.

En lo que respecta a la Sección A, se hizo especial hincapié —en instrucciones previas— en que la palabra que buscábamos estaba estrechamente relacionada con el sustantivo que aparecía junto al hueco marcado en negrita y también se les advirtió de que la palabra meta sólo podía ser un sustantivo, un adjetivo o un verbo. Las instrucciones de la prueba especificaban asimismo este hecho como podemos comprobar a continuación (Fig. 4.3.):

TASK 1

Fill in the blanks in each sentence by adding only one word. The word you need to add can only be an adjective, a noun or a verb. The first letter of each word is provided to help you.

Example:

0. **N_____ light** is preferable to artificial light.

Answer:

0. Natural **light** is preferable to artificial light.

Fig. 4.3.: Instrucciones de la Sección A

En el caso de la Sección B se especificaba que sólo se podía incluir una palabra en el hueco que se les presentaba, aunque se puntualizaba que esta palabra podía ser única o compuesta (en cuyo caso iría unida con un guión) (Fig. 4.4.). Esto se hacía necesario puesto que la respuesta del ítem b10 había de ser un adjetivo compuesto (*single-parent*, *lone-parent* o *one-parent*) (ver Anexo 5):

TASK 2

Translate the following collocations into English. Add <u>either one single word or one hyphenated word</u> in each case.

Example:

0. Prestar atención: To <u>__pay___</u> **attention**

Fig. 4.4.: Instrucciones de la Sección B

En lo que respecta a las instrucciones de las dos últimas secciones, y aunque evidentemente se ofrecía una explicación en las instrucciones escritas al comienzo de cada una (Fig. 4.5. y 4.6.), los procesos de validación iniciales que llevamos a cabo con cinco sujetos antes de proceder a la administración general del test (este primer proceso de pilotaje se describirá en detalle más adelante) demostraron que era especialmente relevante en este caso explicar la diferencia entre ambas tareas de forma oral, ya que a simple vista parecían iguales y los alumnos asumían que en ambos casos se trataba de un formato de elección múltiple tradicional. Durante la administración, se consideró por tanto especialmente importante insistir de forma oral en la diferencia entre ambas (en la opción C debían marcar la combinación "incorrecta" mientras que en la D se marcaría la "correcta").

TASK 3

Choose the word which <u>does not</u> collocate with the noun given (there is <u>only one wrong collocation</u> in each case). If all of them are correct, tick the option "no wrong collocation".

Example:

0. To ___ **a list**

□ compile ☒ compose □ make □ no wrong collocation

Fig. 4.5.: Instrucciones de la Sección C

TASK 4

Choose the correct collocation and tick the appropriate box. There is <u>only one correct</u> collocation in each case. If none of the 3 first options is correct, tick the option "none of these".

Example:

0. A place devoted to entertainment is called **a/an ____ area**.

□ break □ game ☒ play □ none of these

Fig. 4.6.: Instrucciones de la Sección D

Finalmente, nos parece también muy importante destacar que, como se puede apreciar en las figuras anteriores, las instrucciones de cada una de las secciones del

test venían acompañadas de un ejemplo, algo que, sin duda, contribuye a la mejor comprensión de la tarea por parte del candidato.

4.3.1.2.4. Tiempo de la prueba

A la hora de diseñar y administrar un test es también necesario tener muy en cuenta el tiempo que necesitarán los candidatos para completarlo. Al administrarlo, los alumnos deben tener tiempo suficiente para poder contestar todos los ítems pero también debemos vigilar que no sobre demasiado tiempo puesto que cualquiera de las dos condiciones influiría en la fiabilidad de los resultados. A título orientativo, Alderson (comunicación personal, 2008) considera que se debe calcular que un alumno tardará aproximadamente el doble de tiempo que el diseñador de la prueba en poder completarlo.

4.3.1.2.5. Ponderación de la prueba y criterios de corrección

Otro aspecto que influye de manera decisiva en la fiabilidad de un test se refiere a la ponderación de la prueba, es decir, a la puntuación que se le dará a cada uno de los ítems. Si un test consta de ítems objetivos que se pueden puntuar sin la intervención del juicio del corrector, ello contribuirá sin duda a garantizar la fiabilidad. Si, por el contrario, se trata de tests en los que es necesario realizar una valoración de carácter más subjetivo, es recomendable elaborar previamente una escala en la que se contemple una serie de descriptores con el fin de que la ponderación se realice de forma sistemática y en base a unos criterios de referencia.

En nuestro caso, la puntuación se realizaría de forma objetiva dado que nuestros formatos así lo permitían, otorgando un punto a cada respuesta correcta y cero a las respuestas incorrectas o sin contestar. Especial mención merecen los

criterios establecidos para la corrección de las secciones A y B, de competencia productiva. En estas dos tareas, se estipuló que no se penalizarían aquellas respuestas en las que existiese un error gramatical u ortográfico, puesto que nuestro test no pretendía medir estos aspectos sino únicamente la capacidad para combinar las palabras de forma adecuada.

En estas dos secciones, además, se contemplaron todas las posibles respuestas correctas con el fin de que los resultados fueran válidos. Así, existían ítems como por ejemplo el nº 7 de la Sección B,

7. Un problema grave: A/An _____ problem.

que no tenían sólo una respuesta correcta sino que cabían varias posibilidades igualmente aceptables en la lengua. En estos casos decidimos que se darían por válidas las posibles combinaciones que diesen lugar a una expresión con un significado similar al de la colocación española. Por continuar con el ejemplo anterior, en este ítem se puntuaron como correctas "*big*", "*huge*", "*important*", "*serious*" y "*severe*", al considerar que todas se emplean en inglés para expresar la misma idea, mientras que "*difficult*", "*hard*" o "*high*" se consideraron erróneas porque no eran válidas o su significado era diferente. Nos gustaría destacar que, aunque no cabe duda de que algunas respuestas eran más acertadas que otras desde un punto de vista estrictamente colocacional, no consideramos de justicia penalizar a un alumno por utilizar una expresión más genérica o menos restringida en su conmutabilidad, si realmente era aceptable, e incluso frecuente en inglés y transmitía el mensaje adecuado. En definitiva, es razonable pensar que un test debe reflejar y dar cabida a la flexibilidad que la propia lengua permite, algo que constituye además una característica inherente a la naturaleza de las colocaciones y que sin duda incrementa la dificultad de su evaluación.

Todas las posibilidades consideradas como correctas en nuestro test se pueden consultar en el Anexo 5.

4.3.1.3. La validez de los tests

La validez de un test es su capacidad para medir aquello que se quiere medir, es decir, el grado de precisión con el que una prueba mide lo que se propone medir (Henning, 1987; Hughes, 1989; Bachman, 1990; Alderson et al., 1995; López-Mezquita, 2007). Al hablar de validez se han distinguido aspectos muy distintos, pero sin duda los dos conceptos fundamentales que todo test debe observar son la validez de constructo y la de contenido. En nuestro trabajo se incluirá además otro tipo de validez de aparición muy posterior y que no contemplan todos los autores: la validez de respuesta. Veamos que suponen estas nociones, ya tan establecidas en el campo de la evaluación.

4.3.1.3.1. La validez de constructo

El constructo de un test se puede definir como "the underlying ability or trait being measured by the test" (McNamara, 2000: 52). Así, la validez de constructo se debe entender como la validez conceptual de un test (Thorndike y Hagen, 1980), y su importancia radica en que una prueba debe estar fundamentada en los principios teóricos que constituyen aquello que se desea medir. En esencia, la validez de constructo se refiere "a la teoría lingüística subyacente, cualidad no observable sobre la que se sustentan a su vez todos los demás tipos de validez, y que está esencialmente condicionada por el propósito del test, el nivel y el *washback* o impacto del test (las implicaciones y aplicaciones pedagógicas del test, tanto para los docentes como para los investigadores)" (López-Mezquita, 2005: 655).

Establecer la validez de constructo en el terreno léxico es una cuestión peliaguda debido a la "ill-defined nature of vocabulary as a construct" (Read y Chapelle, 2001: 1), algo que resulta quizá todavía más notable en el campo de la competencia colocacional debido a la falta de consenso que existe en torno a su definición (ver capítulo 1). No obstante, cabe señalar que no es infrecuente que haya discrepancias acerca de cuál es el constructo de una determinada habilidad (Alderson et al., 1995).

Como destaca López-Mezquita (2005: 674), una vez establecida la capacidad o rasgo que nuestro test pretende medir "debemos formular una red de consideraciones teóricas que conduzcan a predicciones claras y definidas susceptibles de comprobación". Con el fin de abordar este aspecto en materia colocacional, particularmente en cuanto a la teoría subyacente, se debe considerar, en primer lugar, la naturaleza de la colocación, así como también cuestiones más concretas como número y combinabilidad de elementos que la componen (A+N, N+N, V+N o N+V), grados en el conocimiento colocacional —escalabilidad de los resultados—, dificultad de las colocaciones según su carácter —frecuencia, arbitrariedad, fosilización, etc.[8]— e incluso explorar la naturaleza de los ítems en cuanto a su presentación como elementos aislados o integrados, ausencia o presencia del contexto, uso de diferentes formatos, etc. Todas estas consideraciones deben establecerse para atestiguar la validez de constructo y según Gyllstad (2007: 33) se pueden concretar en tres apartados:

a. The construct needs to be defined theoretically;
b. The construct needs to be defined operationally;
c. Procedures must be established for the quantification of observations.

[8] Debemos hacer notar que este aspecto no formó parte de la presente investigación y sin duda se abordará en futuros trabajos, una vez que nuestro test esté suficientemente validado y perfilado.

En definitiva, creemos que estas indicaciones se formulan de forma más práctica en: 1) la teoría subyacente, 2) la forma en que hacemos operativo el constructo por medio de unas especificaciones previas y 3) la cuantificación de los resultados por medio de una investigación cuantitativa y estadística. A continuación desgranaremos los dos primeros aspectos mencionados. El tercer elemento constituye el objeto de estudio del capítulo siguiente.

a. La teoría subyacente

Una de las premisas fundamentales sobre las que se debe asentar un test es la formulación del constructo, es decir, la habilidad que queremos medir. Como recordaremos, el capítulo primero de este trabajo estuvo dedicado a delimitar nuestra concepción teórica (es decir, nuestro constructo) en torno al fenómeno colocacional, y, según indicábamos, las características fundamentales que definen a la colocación son:

- ✓ Es una combinación bipartita (aunque no necesariamente de dos palabras sino que puede incluir un número mayor) formada por dos elementos que se encuentran a distinto nivel. Por un lado, la base es libremente elegida por el hablante y no está sujeta a ninguna restricción o atadura fraseológica, mientras que el colocado está restringido por la base y no se utiliza de forma libre respondiendo únicamente a su valor semántico y gramatical.

- ✓ Debido a la diferencia de estatus existente entre base y colocado, consideramos que la base de una colocación es una forma léxica, mientras que los colocados son lemas.

- ✓ Es una combinación en la que existe una restricción totalmente arbitraria en cuanto a la aparición conjunta de palabras que nada tiene que ver con las restricciones semánticas impuestas por el sistema de una lengua. Esta restricción, que implica únicamente al colocado, se puede manifestar en su conmutabilidad (cuando no se puede sustituir libremente por un sinónimo) o en su combinabilidad (cuando no acompaña libremente a cualquier base con la que teóricamente podría co-aparecer). Esta restricción no es fija sino que se da en una escala donde se distribuyen las colocaciones, desde las más flexibles a las más fijas y estables.

- ✓ Es una combinación semánticamente composicional, es decir, cuyo significado se puede derivar a partir de los elementos léxicos que las integran.

- ✓ Se trata de una combinación que se usa repetidamente en la lengua, lo cual implica que se ha erigido como un uso institucionalizado y ampliamente aceptado por una comunidad de habla.

Brevemente, y teniendo en cuenta lo anterior, el constructo que establecemos como objetivo de nuestro test de colocaciones es el que responde a las siguientes preguntas:

1. ¿El sujeto es capaz de reconocer el colocado que está arbitrariamente restringido a la base, es decir, el que ésta exige o el que suele acompañarla?
2. ¿El sujeto es capaz de producir el colocado que la base exige?

De estos interrogantes se deriva que la competencia colocacional tiene dos facetas: la receptiva y la productiva. Debemos destacar a este respecto que la dimensión receptiva del conocimiento colocacional tal y como se recoge en nuestro constructo

no se corresponde con la noción tradicional del conocimiento receptivo del vocabulario, que implica ser capaz de reconocer y comprender una palabra (es decir, acceder al significado a partir de la forma) (Nation, 2001). En nuestra opinión, y como varios autores han puesto de manifiesto (Marton, 1977; Gitsaki, 1999), no cabe duda de que la comprensión semántica de las colocaciones no presenta ningún problema para el hablante de una segunda lengua, siempre que éste conozca las palabras individuales que las componen. Lo que se evalúa, pues, desde el punto de vista receptivo en un test de colocaciones no es la capacidad de decodificar las colocaciones semánticamente, como es el caso en los tests de vocabulario, sino su capacidad para reconocer los "lazos colocacionales" que se establecen entre las palabras; no es pues una cuestión semántica, sino sintagmática o combinatoria.

En lo que toca a la segunda dimensión, la productiva, se trata de una faceta que entraña una verdadera dificultad para los alumnos, que suele motivar una diferencia abismal entre los nativos y los no nativos. Así, para Meara (1996: 48) éstos últimos no han adquirido "connections between words that are obvious to native speakers", ya que para los nativos la colocación es una parte de su competencia comunicativa, a la que acceden por una intuición que los no nativos no han podido desarrollar.

Teniendo en cuenta que las colocaciones presentan serios problemas a los alumnos a nivel productivo, y considerando, por otro lado, que los estudios más recientes y rigurosos sobre evaluación colocacional parecen estar indagando más exhaustivamente la dimensión receptiva por el momento (Barfield, 2003; Gyllstad, 2007), nuestro trabajo está más encaminado hacia el estudio y la posterior mejora de la competencia colocacional a nivel productivo. Sobre este aspecto, no obstante, debemos hacer la siguiente puntualización: el trabajo que se presenta en esta tesis pretende ser sólo un primer paso para lograr un instrumento evaluador válido y fiable de la competencia colocacional. En este caso, nuestros esfuerzos se han concentrado

en procurar una metodología sistemática para evaluar las colocaciones de los sustantivos contenidos en las 1.000 primeras palabras de la lengua, en principio tanto a nivel receptivo como productivo. El análisis exhaustivo de la competencia productiva de los hablantes, y de los factores que inciden en su dificultad, serán el objeto de futuras investigaciones.

Dos cuestiones más deben ser apuntadas. La primera tiene que ver con la concepción personal, apoyada por muchos investigadores, de que la competencia colocacional es un constructo independiente susceptible de medirse como un aspecto léxico en sí (Bonk, 2001; Barfield, 2003, 2006; Gyllstad, 2007). En segundo lugar, esta investigación no hubiera sido factible sin la noción de frecuencia y la forma en que la lingüística de corpus y computacional la puede hacer operativa.

b. Las especificaciones operativas del constructo

Una vez que se establece el constructo de forma teórica, la siguiente etapa consiste en hacerlo operativo. Es muy conveniente que nos planteemos qué especificaciones debemos establecer para convertir la teoría subyacente en rasgos que se puedan medir de forma observable. En este sentido, López-Mezquita (2005), siguiendo a Henning (1987), afirma que para comprobar la validez de constructo debemos recurrir a **la suma de otras evidencias**. En primer lugar, esta investigadora destaca la validez de contenido, que trataremos seguidamente pero que en nuestro caso supone llevar a cabo una adecuada selección de las colocaciones que se incluirán en la prueba. Además, la adecuación del ítem al constructo que se investiga es otro importante aspecto a considerar a la hora de operativizar nuestra teoría. Refiriéndonos de nuevo a nuestro test, esto implica que nuestros ítems deben medir de forma rigurosa el conocimiento receptivo y productivo que los hablantes tienen de las colocaciones que contemplamos. Otra cuestión que plantea López-Mezquita es la

necesidad de realizar un análisis de ANOVA, que demuestre la existencia de escalabilidad, es decir, la evidencia de que grupos con distintos niveles de competencia lingüística obtengan también distintos resultados en nuestro test. En este sentido, y para corroborar que existe una verdadera escalabilidad, se hace necesario que un test muestre una diferencia significativa entre los grupos de diferentes niveles (Schmitt et al., 2001). Por último, otra evidencia más que debemos tener muy en cuenta para poder hacer de nuestro test una prueba con validez de constructo es el coeficiente de fiabilidad, pues es bien sabido que la fiabilidad es la primera condición para corroborar la validez de un test (Hughes, 1989).

No queremos, sin embargo, atenernos sólo a estas consideraciones de carácter general sobre la validez de constructo. Afortunadamente, en el año 2001 dos reconocidos investigadores en el área de los estudios léxicos, Read y Chapelle (2001), acotaron aspectos concretos para la consecución de una validez todavía poco conocida. A tal fin, contemplan una serie de coordenadas que habría que tener en cuenta para mejorar el constructo del vocabulario de cara al diseño de las pruebas y que se concretan en la tabla 4.5.

Validez de constructo	
1. Diseño del test:	**2. Propósito del test:**
a. Tipo de información	a. Inferencias y niveles
b. Forma	b. Usos
c. Nº de elementos a evaluar	c. Impacto
d. Referencia calificadora	
e. Método de corrección	
f. Impacto	

Tabla 4.5.: Condiciones para la consecución de la validez de constructo (Read y Chapelle, 2001)

Aunque Read y Chapelle se refieren en todo momento a la competencia léxica con relación al conocimiento de las palabras por separado, nosotros vamos a tratar de adaptar este modelo al diseño de un test de colocaciones para la consecución de la validez de constructo. Vamos ahora a ver las especificaciones que estos autores han hecho y sobre las que nosotros planteamos las preguntas que figuran a continuación.

DISEÑO DEL TEST:

a. El *tipo de información* que deseamos obtener. En nuestro caso, la prueba es claramente de **diagnóstico**, aunque se podría utilizar también como de progreso o aprovechamiento. Como axioma general, este test procura recabar información sobre el conocimiento de las colocaciones formadas a partir de los sustantivos incluidos en las 1.000 palabras más frecuentes del inglés.

   Pregunta 1. ¿Cuál es el nivel de conocimiento de las colocaciones de un grupo de alumnos (311) procedentes de 4 centros universitarios españoles: Filología Inglesa (Univ. de Granda), Traducción e Interpretación (Univ. de Granada), Filología Inglesa (Univ. de Almería) y Traducción e Interpretación (Univ. Jaume I)?

   No obstante, además de la medición a nivel general, se buscaba hacer un diagnóstico más profundo, intentando contestar de forma más detallada a otras preguntas que se formularon de la siguiente manera.

   Pregunta 2. ¿Existe alguna diferencia de nivel entre los estudios de Filología Inglesa y Traducción e Interpretación?

Pregunta 3. ¿Existen diferencias de nivel entre los diferentes cursos de las licenciaturas?

b. La *forma en la que el test está construido*. En este sentido hay que abordar los tests de colocaciones desde dos vertientes distintas. Por una parte se deben considerar las estructuras gramaticales de la propia colocación, y por otra, los métodos o técnicas de medición, que pretenden minimizar los efectos de un método único. Así nos interrogamos sobre:

Pregunta 4. ¿Existen diferencias en los resultados de las colocaciones según las estructuras gramaticales de las mismas (N+N, A+N, V+N, N+V)? ¿Resultan unas estructuras más complejas que otras?

Pregunta 5. ¿Existen diferencias en los resultados achacables al método o tarea de la medición (Secciones del test: A, B, C y D)? ¿Son unos métodos más adecuados que otros para evaluar la competencia colocacional?

c. El *número de elementos evaluados*. En nuestra opinión, en este apartado se podría añadir también la bondad de los ítems y su repercusión en la dificultad y discriminación del test.

Pregunta 6. ¿Cómo contribuye el diseño de los ítems a la fiabilidad y validez del test? ¿Cuáles son los ítems que debemos descartar?

Pregunta 7. ¿Cómo han funcionado los distractores según el análisis de los ítems?

Éstas son las formulaciones que este diseño investigador intentó contestar, siempre teniendo en cuenta las consideraciones que Read y Chapelle se hacen con relación a otros aspectos que reseñamos a continuación.

d. La *referencia calificadora.* En nuestro caso, se trata de una **prueba referida al criterio** (*criterion-referenced test*) y no a la norma, es decir, los resultados del candidato aluden a sus habilidades y no al establecimiento de un orden en el grupo comparando la actuación de unos sujetos con otros.

e. El *método de corrección* de nuestro test está basado en una puntuación **objetiva.** Cabe mencionar la sección de la traducción donde se aceptaron diferentes opciones pero siempre siguiendo unos criterios claros y establecidos de antemano que no restaron, por tanto, objetividad y con los que evaluadores diferentes no hubieran dado resultados distintos.

f. El *impacto.* En nuestro caso se trata claramente de una **prueba de bajo impacto** (*low stakes)* para los candidatos. Precisamente por esa razón se les indicó que incluyesen su nombre, ante el riesgo de que este bajo impacto pudiera producir falta de interés en el candidato al ser anónimo.

PROPÓSITO DEL TEST:

a. Las **inferencias** acerca del conocimiento de los ítems con carácter individual y colectivo son "estimates of a learner's vocabulary size at each of the different levels" (Read y Chapelle, 2001: 11) y se materializan en extrapolaciones en cuanto a los resultados obtenidos como, por otra parte, es usual en los tests de niveles. Es decir, inferimos que el candidato con un resultado del 50% tendrá un nivel

colocacional medio en cuanto a las 2.688 colocaciones extraídas de los sustantivos de la primera banda de frecuencias del inglés (1.000 palabras).

b. Las **utilidades del test,** es decir, el objetivo para el cual está concebido, es el de diagnosticar niveles de competencia colocacional; no obstante, también se podría usar como test de emplazamiento e incluso como prueba de progreso en tratamientos pedagógicos.

c. Por último, el **impacto** o ***washback*** de este test es triple: a) tiene un importante impacto investigador en tanto en cuanto ha dado lugar a la elaboración de un listado de frecuencias de colocaciones inédito en el campo de la lingüística aplicada; b) ofrece a la comunidad académica información de cuáles son los puntos de partida y de llegada en cuanto a los umbrales que los alumnos deben conocer; c) estamos convencidos de que el impacto de esta prueba puede llevar a un *washback* muy positivo en el sentido de que si los alumnos son evaluados del aspecto colocacional, tanto los profesores como los aprendices se acostumbrarán a tomar las colocaciones mucho más en consideración, con lo cual a la postre este proceso de concienciación producirá una clara mejora tanto en la enseñanza como en el aprendizaje. En este sentido, consideramos que tiene un carácter muy instrumental.

Todas las observaciones que hemos contemplado sobre el constructo del test y los requisitos operantes que se deben tomar en consideración, y que a simple vista pueden parecer obvios, nos han servido, sin duda, para concretar los aspectos a tener en cuenta en el diseño de un test que mida correctamente el conocimiento colocacional. Como ya avanzamos, la última etapa de la validez del constructo se refiere a la cuantificación de los resultados por medio de una investigación

cuantitativa y estadística, algo que se verá en el análisis de resultados recogido en el siguiente capítulo. Ahí es donde esta investigación empírica nos hará observables y palpables muchos de los aspectos mencionados en los dos apartados anteriores.

4.3.1.3.2. La validez de contenido

Una prueba tiene validez de contenido cuando es una **muestra representativa** del campo de conocimiento que se pretende evaluar y lo hace en las proporciones correctas (Hughes, 1989). Sin duda, este tipo de validez tiene que ver directamente con una adecuada selección de contenidos y un cuidado proceso de elaboración de los ítems.

a. Selección de contenidos

Todos los evaluadores (Meara, 1996; Nation y Waring, 1997; López-Mezquita, 2007) destacan la necesidad de incluir muestras representativas de lo que pretende evaluar, atendiendo al criterio de frecuencia. A tal fin, cabe destacar que tests de tanto prestigio en el área de la evaluación léxica como el *Vocabulary Levels Test* de Nation (1983) o el de Schmitt et al. (2001) se elaboraron utilizando los listados de Thorndike y Lorge (1944), Kučera y Francis (1967), y la *General Service List* (West, 1953).

En nuestro caso, el listado de frecuencias de colocaciones que hemos empleado para la elaboración de nuestra prueba ha sido ampliamente detallado en el capítulo 3. Baste recordar a este respecto que esta lista consta de un total de 2.688 colocaciones, todas con un sustantivo como base y acompañadas por colocados nominales, verbales o adjetivales. Para su compilación se seleccionaron en primer lugar los sustantivos incluidos en las 1.000 primeras palabras de un exhaustivo listado de frecuencias, compilado mediante una rigurosa investigación (López-Mezquita,

2005). Seguidamente, se extrajeron de forma automática los colocados más frecuentes de dichos sustantivos (en su forma singular y plural) según la medida *t-score* y a partir de los datos aportados por dos de los corpus más representativos del inglés, el BoE y el BNC. Tras un posterior proceso de selección y filtrado manual en el que se observó minuciosamente nuestro constructo de colocación para descartar las combinaciones léxicas que en nuestra opinión no conformaban verdaderas colocaciones, obtuvimos un listado definitivo con cuatro tipos diferentes de colocaciones en términos estructurales, también en proporciones diferentes (A+N = 59,09%, V+N = 31,68%, N+N = 6,94%, N+V = 2,3%,). Se trata, pues, de un recurso inédito hasta ahora, en el que se recogen las colocaciones más frecuentes de los sustantivos también más frecuentes: en definitiva, las que nuestros alumnos deben conocer para poder comunicarse de forma fluida y precisa.

Así pues, tomando este banco de colocaciones como punto de partida para la selección de los contenidos de nuestro test, y poniendo especial cuidado en cubrir las distintas estructuras gramaticales en sus correctas proporciones para que nuestra prueba fuese verdaderamente representativa de la realidad lingüística, procedimos al diseño de los ítems. Durante este proceso, se fueron seleccionando las colocaciones que conformarían nuestro test atendiendo a dos criterios fundamentales. En primer lugar, fue prioritario en todo momento atenernos a las proporciones que nuestro listado contenía en cuanto a las estructuras gramaticales. Así pues, nuestros ítems debían reflejar unos porcentajes semejantes a los incluidos en la lista de referencia. Esta similitud de proporciones no sólo se daría en el test en su conjunto, sino también en cada una de las secciones de forma individual. Así, la distribución de cada una de las cuatro estructuras gramaticales contempladas en nuestro test es la siguiente (Tabla 4.6.):

	Sección A	Sección B	Sección C	Sección D	**Total test**
N+N	3 (6%)	3 (6%)	4 (8%)	4 (8%)	14 (7%)
A+N	30 (60%)	29 (58%)	29 (58%)	29 (58%)	117 (58,5%)
V+N	16 (32%)	16 (32%)	16 (32%)	16 (32%)	64 (32%)
N+V	1 (2%)	2 (4%)	1 (2%)	1 (2%)	5 (2,5%)
Total sección	50 (100%)	50 (100%)	50 (100%)	50 (100%)	200 (100%)

Tabla 4.6.: Distribución de estructuras gramaticales por secciones

En segundo lugar, las colocaciones se seleccionaron de forma aleatoria dado que partíamos de la base de que todas provenían de la misma banda de frecuencias. No obstante, durante el proceso de diseño del test se fue haciendo una selección de acuerdo con las colocaciones que más se prestaban a cada formato. Así, los métodos productivos hacían necesario incluir colocaciones de mayor restricción combinatoria, es decir, más idiosincrásicas, para reducir al máximo el número de posibles respuestas correctas y fomentar un resultado objetivo. Por otro lado, la Sección C requería de colocaciones donde contáramos con un espectro colocacional suficiente para incluir dos (y en algunos casos tres) combinaciones correctas formadas con el mismo sustantivo. Por último, en la Sección D también parecía más conveniente incluir colocaciones que no fueran demasiado flexibles, es decir, que no permitieran un gran número de sinónimos como colocados, porque esto agravaría la ya de por sí ardua tarea de encontrar distractores plausibles pero definitivamente incorrectos o inaceptables como colocaciones.

b. Construcción de los ítems

La construcción de los ítems es otra cuestión esencial para la consecución tanto de la validez como de la fiabilidad de un test. Aquí vamos a detenernos en el estricto proceso de construcción de ítems que hemos llevado a cabo, puesto que ya se ha realizado en este capítulo una descripción de los distintos formatos contemplados (ver sección 4.3.1.2.1.).

Debemos destacar que, dado que pretendíamos alcanzar un total de 50 ítems para cada sección, decidimos construir el doble, es decir, 100 por sección, para poder más tarde utilizar los 50 que mejores resultados ofrecieran, tras un primer proceso de validación del que daremos cuenta al tratar de la validez de respuesta.

b.1. Ítems de la Sección A (*c-test*)

Como recordaremos, la primera sección de nuestro test contenía 50 ítems diseñados en formato de *c-test*. Mostramos un ejemplo del primer ítem de esta sección para ilustrarla:

1. Applications are particularly welcome from women, people from ethnic minority backgrounds and **d_____ people.**

Para la construcción de los ítems de la Sección A, y tras seleccionar las colocaciones que los integrarían siguiendo los dos criterios mencionados más arriba, debíamos proporcionar un contexto adecuado, considerando distintos aspectos: 1) que resultase claro y sencillo para el candidato y no contuviese referencias a elementos

externos o un vocabulario complejo, 2) que le aportase suficiente información para poder completar el ítem, 3) que no tuviese una extensión demasiado amplia para evitar el esfuerzo de procesamiento y la influencia de la destreza lectora en la respuesta, 4) que fuese auténtico en lugar de inventado por el diseñador del test.

Partiendo de estas premisas, consideramos que lo más adecuado era extraer este contexto de un corpus, y así proporcionar usos reales de la lengua, es decir, como aparecen en el lenguaje natural. Afortunadamente contábamos con el programa *Sketch Engine* en su versión Beta, que recientemente ha incorporado una opción en donde las concordancias se ordenan ofreciendo en primer lugar los mejores ejemplos de un uso, según criterios pedagógicos y lexicográficos. Tal y como explican sus autores (Kilgarriff et al., 2008), esta nueva opción ha sido posible gracias a la puesta en marcha del sistema GDEX (*Good Dictionary Example*), que aplica una serie de criterios al extraer las concordancias de un corpus, con el fin de que constituyan los ejemplos más adecuados para ilustrar una palabra o una colocación concreta. Los criterios que aplica esta herramienta son los siguientes:

- Se priorizan las oraciones cuyas palabras se encuentran entre las 17.000 más frecuentes de la lengua, aplicando penalizaciones especiales a palabras de muy baja frecuencia.
- Se priorizan oraciones que no contienen pronombres y elementos anafóricos como "*this*", "*that*", "*it*" o "*one*" dado que suelen hacer referencia a contextos externos a la oración.
- Se priman oraciones donde la colocación meta aparece en la proposición principal.
- Se priorizan oraciones completas (identificadas por comenzar con una letra mayúscula y concluir con un punto, o signos de interrogación o exclamación).

- Se priorizan oraciones donde, además de la colocación buscada, aparece un "segundo colocado" con especial prominencia.
- Se priman oraciones donde la colocación aparece hacia el final de la oración, dado que se considera que un buen ejemplo introduce el contexto en primer lugar, creando así un "hueco" donde encajará la palabra o colocación, para que el significado de ésta se haga más explícito y fácil de comprender.

Así pues, a la vista de las ventajas que ofrece esta nueva herramienta, no cabe duda de que suponía una ayuda muy interesante y novedosa para esta investigación; revisar manualmente las concordancias para hallar una oración que ofreciera un contexto que evidenciase el uso de las colocaciones hubiera resultado realmente laborioso. Por otra parte, *Sketch Engine Beta* nos permitía también acceder al corpus ukWaC[9], una base de datos de enormes dimensiones (contiene más de 2.000 millones de palabras), diseñada por Adriano Ferraresi y Marco Baroni en el año 2007, que recoge textos extraídos de páginas web británicas. Un corpus de tal envergadura, pues, presentaba más usos de cada una de nuestras colocaciones, lo cual, unido a la organización según el "mejor ejemplo" aludida anteriormente, suponía una fuente inestimable de opciones para nuestros ítems.

Concretamente, el proceso de búsqueda de nuestras oraciones comenzaba accediendo al corpus ukWaC (desde el programa *Sketch Engine*) e introduciendo nuestra colocación (permitiendo que existiera un espacio de hasta cinco palabras entre la base y el colocado). Una vez que el programa extraía las concordancias, se delimitaba la opción de búsqueda para que éstas se reordenaran ofreciendo, en primer lugar, las que constituían los mejores ejemplos (Fig. 4.7.) y se establecía que

[9] Información disponible en http://clic.cimec.unitn.it/marco/publications/lrec2008/lrec08-ukwac.pdf [Último acceso: 25.01.2009]

apareciesen en formato de oración, y no en la disposición tradicional de KWIC (*Key Words in Context*) donde las oraciones aparecen incompletas o mutiladas (Fig. 4.8.).

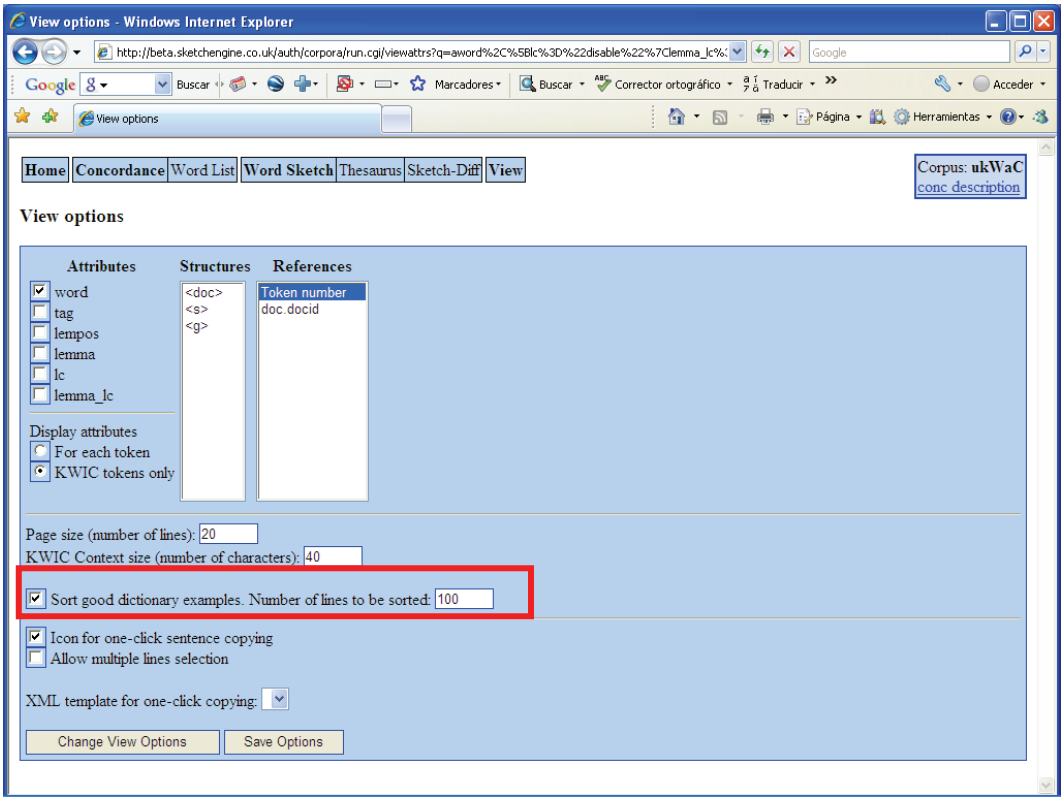


Fig. 4.7.: Ordenar concordancias según el criterio de "mejor ejemplo" (*Sketch Engine Beta*)

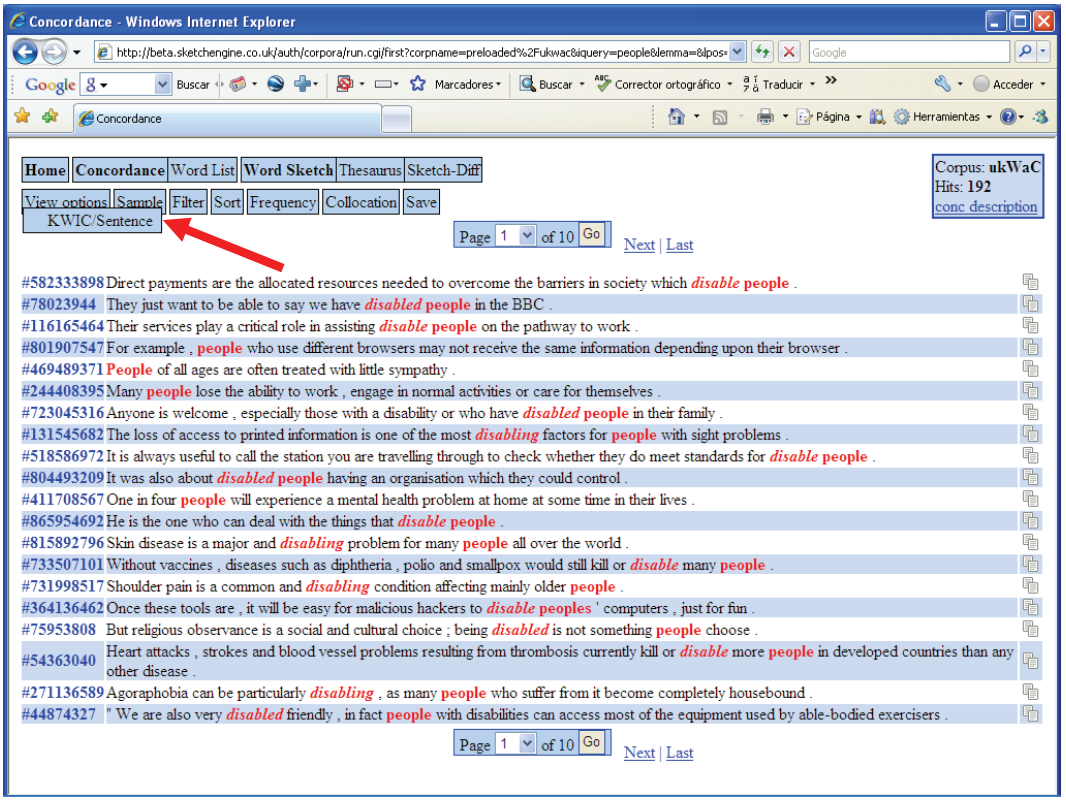


Fig. 4.8.: Concordancias en formato de oración, en lugar del tradicional KWIC

Una vez establecidos estos parámetros, fuimos extrayendo las oraciones de cada una de las colocaciones que habíamos seleccionado, como podemos apreciar en la figura 4.8. Cabe señalar que, además de contar ya con los ejemplos más idóneos para el alumnado gracias al filtro que se había aplicado, también se llevó a cabo una selección manual en la que se revisaron las primeras 40 ó 50 oraciones (cada página ofrece alrededor de 20 de ellas y se puede comprobar que a partir de la tercera página las oraciones se van haciendo más largas y, por tanto, menos útiles para nuestro propósito). Finalmente, se seleccionaba la oración que considerábamos más idónea y más explícita para cada caso.

Dos breves aclaraciones más merecen nuestra atención. Como se puede apreciar en el Anexo 4, decidimos ofrecer la palabra base en negrita a fin de destacar

la palabra clave que debían considerar a la hora de dar un colocado. Además, el espacio que se le dejaba al alumno para contestar la pregunta (marcado con una línea) era de la misma longitud en todos los casos para que esto no indujera a posibles deducciones o inferencias por parte del alumno.

b.2. Ítems de la Sección B (traducción)

Los ítems de la sección B fueron diseñados, como sabemos, con un formato en el que al alumno se les pedía traducir de la lengua materna (el español en este caso) al inglés. El primer ítem de esta sección, a modo de ilustración era el siguiente:

1. Empezar el día: To _____ **the day**

En este caso, como ya dijimos, se puso especial cuidado en seleccionar colocaciones que no presentaran una opción muy amplia de respuestas para que los ítems fueran lo más objetivos posible. Cabe asimismo destacar que, aunque la mayoría presentaban el sintagma de forma descontextualizada puesto que se consideró que la construcción en español era suficiente para indicar al examinando la colocación que se buscaba en cada caso, hubo ítems en los que se estimó oportuno ampliar la longitud de la oración hasta convertirla en una proposición para que no hubiese lugar a dudas en cuanto a la colocación que se estaba trabajando (por ejemplo: "Lo que nos depara el futuro: What the future _____"). Además, en el caso de los ítems b9 y b33 (ver más abajo), se incluyó una aclaración entre paréntesis puesto que los alumnos a los que se le administró el test durante el periodo de validación inicial indicaron que se prestaba a confusión o que hubieran agradecido esta aclaración para estar seguros de la respuesta que se les pedía.

b9. Dar ejemplo/servir como ejemplo (a los demás): To _____ an example
b33. Papel estelar (en cine, teatro,…): _____ role

Finalmente, y como ya señaláramos en el caso de la sección anterior, todos las líneas que marcaban el espacio en el que el alumno debía escribir la respuesta eran de la misma longitud para garantizar la validez del test.

b.3. Ítems de la Sección C (detección del intruso)

Los ítems de la Sección C consistían, como recordaremos, en preguntas diseñadas con la técnica de detección del intruso que se presentaban de manera descontextualizada. El primer ítem de esta sección nos servirá para ilustrar el proceso de construcción llevado a cabo:

1. To___ your hand

☐ hold ☐ rise ☐ shake ☐ no wrong collocation

Como vemos, en primer lugar aparecía la entrada o estímulo del ítem, en la que se ofrece la base de la colocación y una indicación del lugar en el que aparecerían los colocados que constituían las opciones (de nuevo con una línea siempre de la misma longitud). Esta indicación estaba especialmente encaminada a facilitar la comprensión de los ítems N+V en los que el alumno podría llevarse a confusión si no se le indica que el verbo aparecía en estos casos detrás del nombre.

En lo que respecta a la construcción de las opciones, éstas son, como ya explicamos, tres, además de la alternativa fija "*no wrong collocation*". Las tres primeras opciones se dispusieron por orden alfabético, mientras que "*no wrong collocation*" siempre era la cuarta opción. De las tres primeras casillas, dos estaban conformadas por colocados extraídos de nuestra lista de colocaciones, que en este caso funcionaban como distractores, mientras que la tercera era una combinación inválida, que constituía la respuesta correcta. Se puso especial cuidado en que todas las opciones perteneciesen a la misma categoría gramatical. Dado que los ítems en este caso eran descontextualizados, no se hacía necesario que las colocaciones compartieran un mismo significado (aunque sí ocurre en un buen número de los ítems). Además, nos cuidamos de no incluir en un mismo ítem colocaciones que alteraran el significado del sustantivo-base entre una y otra (por ejemplo, se evitaron casos como "*break a record*" y "*keep a record*" donde el nombre adopta un significado distinto dependiendo de la colocación en la que aparezca).

En lo que respecta a las respuestas correctas, éstas debían cumplir otros dos requisitos, además del de tener la misma categoría gramatical que las colocaciones de las otras dos alternativas. En primer lugar, debía tratarse de combinaciones plausibles, que verdaderamente discriminaran entre los alumnos con una buena competencia colocacional y los que no la tienen, habiendo en ocasiones casos en los que se incluían elementos que podían mostrar interferencias de la lengua materna (por ejemplo "**imperious need*" o "**the fire extends*"). Así, se incluyeron combinaciones que se ajustaran a las normas de la lógica y que no fueran incongruentes o que contuviesen palabras erróneas o inexistentes, puesto que en ese caso no se estaría evaluando el conocimiento colocacional de los alumnos, sino el léxico.

Un segundo requisito, en este caso evidente, que debían cumplir las combinaciones que conformaban la respuesta correcta era que, efectivamente, fuesen binomios de palabras que no se utilizan en inglés. Para comprobar si verdaderamente

se trataba de combinaciones inválidas, todas las combinaciones susceptibles de formar parte de nuestros ítems se comprobaron en el corpus BNC con la ayuda de *Sketch Engine* y, de nuevo, permitiendo un espacio de cinco palabras a la izquierda y la derecha de la base. Determinamos que si existía algún ejemplo en el que la combinación se utilizase, se rechazaría como posible opción y se buscaría otra alternativa.

Finalmente, en lo que toca a la opción "*no wrong collocation*", se dispuso que debía existir un 10% de las preguntas en las que ésta fuese la opción correcta. Por tanto, cinco de los ítems del test contendrían tres colocaciones aceptables: c15, c20, c33, c37 y c43.

b.4. Ítems de la Sección D (elección múltiple)

La última sección de nuestro test estaba compuesta por ítems de opción múltiple, el primero de cuales se ofrece a continuación:

1. If you say **in ____ fact**, you indicate that you are giving more detailed information about what you have just said.

   ☐ actual ☐ real ☐ true ☐ none of these

En este caso se ofrecía, como vemos, un estímulo en el que los alumnos encontraban la base de la colocación marcada en negrita (así como cualquier otra palabra que formara parte de la expresión: en el ejemplo anterior "*in*") incluida en una definición de la colocación a modo de contexto. Esta definición, que contribuye a naturalizar la

expresión y ayuda a su comprensión y contextualización, se extrajo del diccionario *Collins COBUILD English Dictionary for Advanced Learners* en CD-Rom (2001), que resultó ser de enorme valor para nuestros propósitos. Sólo en ocasiones tuvimos que completar la información con el *Longman Dictionary of Contemporary English*[10] y el *Cambridge Advanced Learner's Dictionary*[11], ambos en su versión en línea. Continuando con el ejemplo ofrecido más arriba, que corresponde al primer ítem de la Sección D, podemos comprobar el origen de su definición en la figura 4.9.:

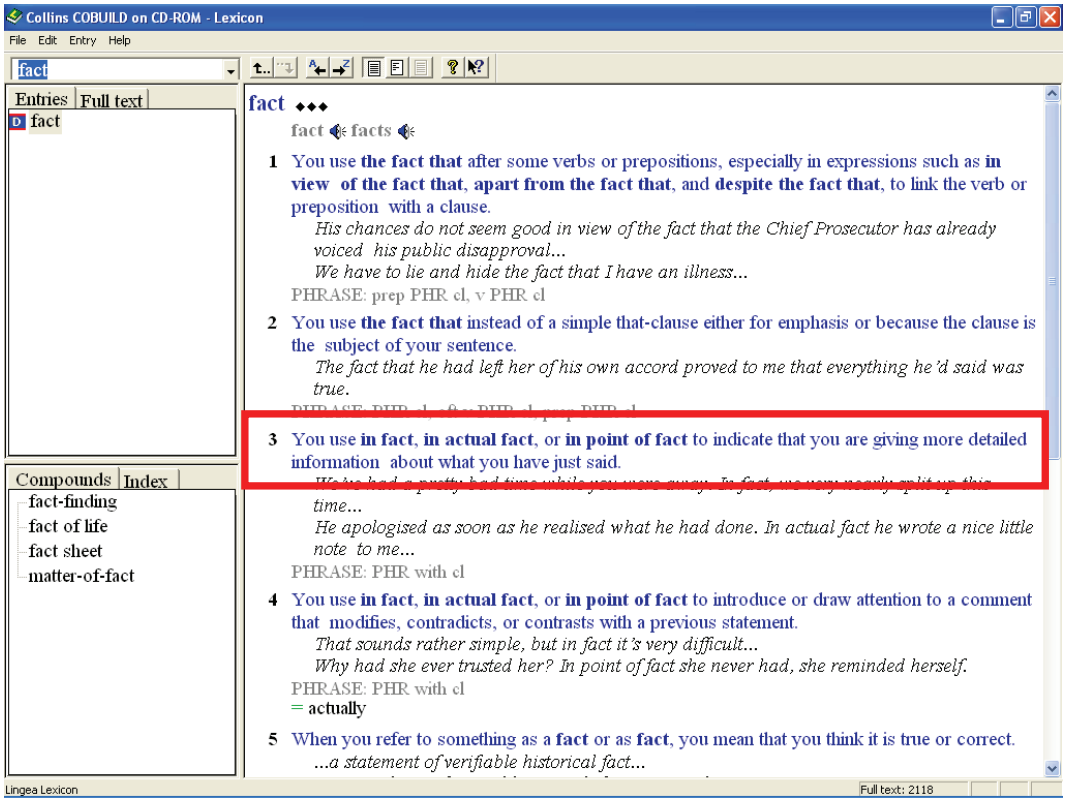


Fig. 4.9: Definición de "*in actual fact*" en Collins COBUILD

[10] http://www.ldoceonline.com/
[11] http://dictionary.cambridge.org/

En lo que respecta a las opciones de los ítems, en este caso se trataba de una respuesta correcta que correspondía a la colocación extraída de nuestro listado, dos distractores con combinaciones inaceptables y la opción "*none of these*" comentada anteriormente. De nuevo se situaron los tres primeros elementos en orden alfabético, siendo "*none of these*" la última opción, y de nuevo compartían categoría gramatical. Los dos distractores que se incluyeron en los ítems de esta sección eran, al igual que sucedía en la anterior, combinaciones plausibles y formadas por palabras que existen en inglés, para que discriminaran adecuadamente entre los sujetos en cuanto a su conocimiento colocacional. Sin embargo, se trataba de combinaciones inaceptables en inglés y, por tanto, que se pueden considerar incorrectas sin lugar a dudas. Para cerciorarnos de esto último, también se realizó la comprobación de todos los distractores en el BNC. En general, debemos hacer notar que el proceso de elaboración de estos ítems fue especialmente arduo, puesto que entraña verdadera dificultad encontrar distractores plausibles pero en los que no quepa duda sobre el hecho de que son inaceptables en la lengua. Como sabemos, y como ya hemos mencionado más arriba, una de las características propias de la colocación es su flexibilidad, lo cual las diferencia de las locuciones y otras construcciones fraseológicas de mayor fijación. Esta flexibilidad, sin duda, dificultaba nuestra tarea en buena medida, aunque el resultado final fue satisfactorio.

Un último apunte en cuanto a los distractores se refiere a la frecuencia de las palabras. En este sentido, somos conscientes de que algunos de los términos que se incluyeron como opciones erróneas son palabras de baja frecuencia en inglés (por ejemplo, "*denigration*" o "*luxurious*"). Sin embargo, en estos casos constituyen cognados que no presentan problemas para un hablante nativo de español, por lo que no se consideró que supusieran una dificultad añadida.

Por lo que respecta a la opción "*none of these*", en este caso también conformaba la respuesta correcta de un 10% de nuestros ítems: d8, d25, d32, d36 y d43.

4.3.1.3.3. La validez de respuesta

La validez de respuesta, establecida por Henning en1987, supone la introducción de una serie de técnicas cualitativas, que se analizan *a posteriori*, para comprender la actuación de los candidatos en la prueba. En otras palabras, se pretende conocer los juicios personales e impresionistas del individuo, mediante técnicas como los protocolos introspectivos orales y escritos o las entrevistas personales, para confirmar las bondades o deficiencias de una prueba. Estos testimonios se refieren tanto a la evaluación del test en general como a cuestiones más concretas (adecuación de los métodos, la construcción de los ítems, la duración de la prueba, etc.). Estos datos suelen ser de la mayor relevancia porque nos permiten llevar a cabo la triangulación de los datos, o el contrate entre ambas perspectivas investigadoras ─cualitativa y cuantitava─ (McDonough y McDonough, 1997). Para McDonough y McDonough (ibid.), los datos suelen ser coincidentes y no hacen más que reforzar las conclusiones que se deducen de la propia investigación a nivel general.

En nuestra investigación, y como ya se ha adelantado en distintas secciones de este capítulo, llevamos a cabo en primer lugar una administración inicial del test una vez construido, con el fin de validar sus contenidos y asegurarnos de que estaba correctamente diseñado para su administración a mayor escala. Esta primera validación se realizó con cuatro alumnos de doctorado de la Universidad de Granada y con un hablante nativo. A estos cinco sujetos se les administró el test en su versión íntegra, es decir, con los 400 ítems diseñados en un principio, pero en dos sesiones diferentes para evitar un cansancio extremo que afectara a los resultados. Tras estas

sesiones, se llevó a cabo un protocolo oral en el que también se recabó más información acerca de la prueba. Los comentarios de estos sujetos contribuyeron principalmente, aunque no únicamente, a:

- Mejorar las instrucciones de la Sección A, indicando las posibles categorías gramaticales que se buscan.
- Mejorar varios ítems de la Sección B, incluyendo aclaraciones entre paréntesis para facilitar la comprensión.
- Comprobar que era más adecuado utilizar la expresión "*no wrong collocation*" que "*all of them*" en la Sección C.
- Comprobar que los distintos sujetos encontraban especialmente difícil secciones diferentes, por lo que no destacaba una como particularmente compleja o mal diseñada frente al resto.

Tras este proceso de validación, se comprobaron los resultados obtenidos por estos sujetos y se seleccionaron los 200 ítems definitivos de nuestro test, descartando aquellos que habían presentado problemas para varios sujetos, y teniendo también especial cuidado en incluir solamente aquellos que habían sido correctamente contestados por el sujeto nativo.

La segunda forma en que se aplicó la validez de respuesta fue mediante la realización de protocolos orales con los sujetos que habían realizado el test definitivo. Aunque en este caso no hemos acometido una investigación de la envergadura de la tabulación y análisis de resultados cuantitativos, hemos querido, al menos, recabar la información aportada por los alumnos en cuanto a la impresión general del test. En términos prácticos, y puesto que los grupos eran muy numerosos, se pidió a 5 ó 6 sujetos de cada grupo que ofrecieran sus apreciaciones sobre el test a nivel general y también sobre cuestiones más específicas como la dificultad de las secciones y los

ítems, la duración de la prueba, etc. Aunque, como decimos, no se ha llevado a cabo una investigación estadística, ofrecemos datos reiteradamente aportados por los diferentes grupos (Tabla 4.7.). Ofrecemos estos datos en esta primera parte del diseño de investigación porque creemos que arrojarán una importante luz a la hora de entender e interpretar la compleja investigación cuantitativa, que mostramos en el siguiente capítulo.

- El test es demasiado largo, no debería contener más de 100 ítems.
- La duración fue adecuada: hemos tenido tiempo de completar el test sin prisas.
- La construcción del test parece muy cuidada.
- Las instrucciones son claras.
- La Sección C es especialmente difícil porque resulta confusa. Incluso cuando creías conocer una colocación, el encontrarla junto a opciones tan similares y totalmente fuera de contexto hacía muy difícil estar seguro. En general, todas las alternativas parecían igualmente posibles.
- El contexto resulta de gran ayuda y se echa en falta en la Sección C.
- La primera letra en la Sección A es de mucha ayuda para poder responder la pregunta.
- En general, el test es demasiado difícil para nuestro nivel (esta apreciación provenía especialmente de alumnos de los dos primeros cursos de las licenciaturas).
- En general, el test es muy interesante. Contiene expresiones que deberíamos dominar para poder utilizar el inglés fluidamente.
- Sería de mucha utilidad contar con el listado de colocaciones y poder estudiarlas. Son muy útiles.

- Los contenidos del test no se enseñan en clase y debería hacerse porque es una de las cuestiones en las que más deficiencias tenemos.
- Hasta ahora no habíamos reparado en la importancia de aprender las palabras "en grupos". Es algo a lo que prestaremos más atención en el futuro.

Tabla 4.7.: Respuestas más frecuentes en los protocolos orales

Aunque, como decimos, no se llevó a cabo una investigación estadística de este protocolo, las opiniones de la lista anterior se repitieron una y otra vez en todos los grupos. Son especialmente destacables las apreciaciones relativas a cuestiones como la importancia del contexto o la mayor dificultad de la Sección C. Pero debemos destacar que, como se puede apreciar, el test les parecía que medía un aspecto del lenguaje en que ni habían sido instruidos ni tan siquiera eran conscientes de su existencia e/o importancia. Esto nos lleva a pensar que el impacto del test (o *washback*) en sí puede constituir una aportación valiosa de esta investigación. En el capítulo siguiente comprobaremos que los resultados cuantitativos parecen confirmar estas impresiones, lo cual contribuye sin duda a reforzar la validez de respuesta de nuestro test.

4.3.2. Administración y corrección

4.3.2.1. Condiciones de administración

Una vez concluida la planificación, diseño y construcción del test, la siguiente etapa de nuestra investigación se dedicó a la administración de la prueba. Como ya dijimos, el TCA1 se pilotó con 311 sujetos de tres universidades distintas. Así, fue enviado al

Departamento de Traducción y Comunicación de la Universitat Jaume I (Castellón) y al Departamento de Filología Inglesa de la Universidad de Almería. En estas dos Universidades, los tests fueron administrados a los alumnos durante sesiones de clase ordinarias y los sujetos fueron supervisados por los profesores, asegurando así la fiabilidad del proceso. Este mismo proceso se realizó en la Universidad de Granada (tanto en Filología Inglesa como en Traducción e Interpretación), y en este caso además la autora de este trabajo llevó a cabo la administración de todos los tests, contando siempre con la ayuda de los profesores de cada uno de los grupos.

Por razones prácticas fue imposible que todos los tests se administraran en franjas horarias similares (unos se llevaron a cabo por la mañana y otros por la tarde) pero creemos que tanto el hecho de que las sesiones fueran supervisadas en todo momento por los profesores, como el pedirles a los alumnos que se identificaran en el test, contribuyó a evitar el desinterés y a incrementar el rigor.

4.3.2.2. Corrección de la prueba

La corrección de la prueba también se llevó a cabo de forma objetiva, dado que los ítems así lo permitían. Así, se otorgó un punto a cada respuesta correcta y cero puntos si era incorrecta. En cuanto a la parte productiva del test, y puesto que la prueba fue corregida por la autora de este trabajo, hablante no nativa de inglés, las respuestas que no correspondían a lo que esperábamos se compararon con los datos del BNC para comprobar que constituían combinaciones posibles en inglés y, en su caso, para asegurarnos de que su significado coincidía con el planteado en el ítem mediante la consulta de concordancias. Si no se cumplían ambos requisitos, la respuesta se consideraba incorrecta. Este riguroso proceso de comprobación contribuye, en nuestra opinión, a otorgar validez y fiabilidad a los resultados.

En lo que respecta al apartado receptivo (Secciones C y D), se consideraron incorrectos aquellos ítems en los que se había marcado más de una opción (incluso aunque una de ellas fuese la correcta), aunque debemos destacar que esto sucedió en muy contadas ocasiones.

Asimismo, el proceso se revisó en dos ocasiones para comprobar que se habían seguido unos criterios uniformes en todos los casos. Una vez concluido este proceso, se procedió a preparar los datos para su análisis estadístico mediante el programa SPSS v.15. Pero las cuestiones relativas a este análisis de datos, sus resultados y conclusiones serán abordados ya en el capítulo 5 de este trabajo.

4.4. Conclusión

En el capítulo que ahora concluye se ha abordado el diseño investigador de esta tesis, conducente a la planificación, construcción y pilotaje de un test de colocaciones: Test de Colocaciones ADELEX Versión 1. Siguiendo el modelo planteado por López-Mezquita (2005) dentro del proyecto ADELEX para la construcción de un test, hemos contemplado las distintas etapas que deben configurar el planteamiento y desarrollo de un instrumento de medida que pretenda ser válido, fiable y eficaz. Así, consideramos en primer lugar las cuestiones relativas a la planificación de la prueba, atendiendo al contexto educativo en el que se deseaba actuar y prestando también especial atención a las dos facetas que todo test debe contemplar si queremos que sea verdaderamente riguroso y, en nuestro caso, útil tanto para la investigación como para la pedagogía de las colocaciones: la fiabilidad y la validez. En cuestiones de fiabilidad, hemos sido especialmente cuidadosos a la hora de plantear los distintos formatos de nuestros ítems, hemos tratado de diseñar un test con un número de preguntas que lo hagan una muestra representativa del conocimiento colocacional, y

también hemos procurado establecer unos criterios de ponderación y corrección que conduzcan a una evaluación rigurosa.

En lo tocante al segundo pilar de la evaluación, la validez, hemos asimismo contemplado los aspectos relativos a la validez de constructo de nuestro test, con el fin de garantizar que se trate de un instrumento que nos aporte datos sobre el conocimiento colocacional de los sujetos, y no sobre cuestiones léxicas más generales que no se corresponden con el fenómeno colocacional. En este mismo sentido, también se ha tratado de reflejar escrupulosamente las distintas consideraciones teóricas que se abordaron en el primer capítulo de esta tesis, integrando así plenamente nuestro marco teórico en el diseño investigador. La validez de contenido, por otro lado, venía también avalada por la investigación que llevamos a cabo en torno a la frecuencia de las colocaciones, y que finalmente se materializó en un listado de colocaciones frecuentes, exhaustivo, completo y enfocado a las necesidades concretas de nuestros alumnos. Haciendo uso de este valioso banco de datos para la selección de nuestros ítems, estábamos garantizando sin duda la validez de contenido de nuestra prueba. Finalmente, la construcción de los ítems del test, rigurosamente elaborado con la ayuda de recursos y técnicas lexicográficas y computacionales para la selección de contextos auténticos y pedagógicamente fundamentados, y la elaboración de distractores basada también en herramientas lexicográficas y de corpus, nos hace concebir la esperanza de haber diseñado un test eficaz.

En conclusión, confiamos en que el proceso descrito en este capítulo para la consecución de una medida válida y fiable contribuya a aportar luz en el todavía poco iluminado camino de la evaluación colocacional. Sin embargo, esta tan ansiada validez y fiabilidad no puede estar únicamente basada en el proceso de construcción del test, necesitamos datos que nos ayuden a confirmar o refutar empíricamente la idoneidad del TCA1. Ésta será pues la tarea que llevaremos a cabo en el quinto y último capítulo de nuestro trabajo.

CAPÍTULO 5

EL DISEÑO DE INVESTIGACIÓN. TCA1: ANÁLISIS DE RESULTADOS, CONCLUSIONES Y PROPUESTA DE UNA NUEVA VERSIÓN (TCA2)

> *The data are only as good as the instrument that you used to collect them and the research framework that guided their collection.*
>
> Pallant (2007: 3)

5.1. Introducción

Este capítulo va a estar dedicado al análisis estadístico de los resultados del Test de Colocaciones ADELEX Versión 1 (TCA1), según la Teoría Clásica de los tests. Así pues, llevamos a cabo análisis estadísticos descriptivos para calcular las medidas de tendencia central —la media, la mediana, la moda, los valores mínimos y máximos—, y de dispersión —desviación típica y varianza—, del mismo modo que calculamos el índice de fiabilidad mediante el coeficiente alfa de Cronbach. También se realizaron

exploraciones de estadística inferencial o muestral a través de t-tests de muestras independientes y análisis de la varianza (ANOVA) para refutar o confirmar la existencia de similitudes y diferencias entre diferentes grupos. Asimismo, fue necesario efectuar correlaciones para encontrar aspectos parecidos o discrepancias entre muestras relacionadas. Por último, se acometió un análisis de los ítems para averiguar la dificultad y la discriminación de cada uno de ellos, al mismo tiempo que se localizaron los distractores que no cumplían con su función de forma fiable. Para la obtención de todos estos resultados estadísticos se utilizó el programa informático SPSS v.15 (*Statistical Package for Social Sciences, version 15*).

Nos parece pertinente recordar que el objetivo de este diseño de investigación fue llevar a cabo un diagnóstico certero del conocimiento de las colocaciones de los sustantivos incluidos entre las 1.000 palabras más frecuentes del inglés, con relación a una población constituida por alumnos españoles de las licenciaturas de Filología Inglesa y Traducción e Interpretación. La principal pregunta que nos planteamos fue la siguiente:

A. ¿Cuál es el nivel de conocimiento colocacional de un grupo de alumnos (311) procedentes de 4 centros universitarios españoles —Filología Inglesa (Univ. de Granda), Traducción e Interpretación (Univ. de Granada), Filología Inglesa (Univ. de Almería) y Traducción e Interpretación (Univ. Jaume I)—, con relación a 2.688 colocaciones provenientes de una base de datos donde se recogen las combinaciones más frecuentes de los primeros 412 sustantivos de la lengua, en sus formas singular y plural?

Sin embargo, a este propósito inicial se incorporaron otras hipótesis para mejor acotar y delimitar la competencia colocacional, y abajo las enunciamos en detalle:

B. ¿Existe alguna diferencia de nivel en el campo de las colocaciones entre los alumnos de Filología Inglesa y Traducción e Interpretación?
C. ¿Existen diferencias de nivel entre los diferentes cursos de las licenciaturas?
D. ¿Existen diferencias en el nivel de competencia de los alumnos según la estructura sintáctica de las colocaciones (N+N, A+N, V+N, N+V)? Es decir, ¿resultan unas estructuras más complejas que otras?
E. ¿Existen diferencias en los resultados achacables al método o tarea de la medición (Secciones del test: A, B, C y D)? Es decir, ¿son unos formatos más adecuados que otros para evaluar la competencia colocacional?
F. ¿Cómo contribuye el diseño de los ítems a la fiabilidad y validez del test? ¿Cuáles son los ítems que debemos descartar tras el proceso de pilotaje?
G. ¿Cómo han funcionado los distractores según el análisis de los ítems y cuáles hay que descartar?

En las siguientes secciones mostraremos los resultados estadísticos obtenidos con el fin de dar cumplida cuenta de la competencia colocacional del grupo en cuestión. No obstante, como premisa previa prestaremos atención a la fiabilidad del test, cualidad esencial de partida. A continuación, pasaremos a abordar las respuestas de los interrogantes formulados en el diseño de investigación, para poder así corroborar tanto la validez de constructo como la de contenido. Seguidamente, llevaremos a cabo un análisis de ítems para comprobar el coeficiente de dificultad (CD) y el índice de discriminación (ID), seguido de una profunda revisión psicométrica de las alternativas de los ítems de elección múltiple; esta revisión de ítems busca ante todo detectar y paliar los posibles fallos de su construcción. Finalmente, mostraremos una nueva versión de la prueba, en la que se ha tratado de conseguir un test más válido y fiable, que ha sido ya construido en formato electrónico y está preparado para su futura utilización.

5.2. La fiabilidad

Como ya señalamos en el capítulo anterior, la fiabilidad muestra la ausencia de error. Recordando una definición ya clásica, la fiabilidad "has to do with the stability of scores for the same individuals. If the scores of students are stable the test is reliable; if the scores tend to fluctuate for no apparent reason, the test is unreliable" (Lado, 1961: 330). La razón de abordar la fiabilidad antes que la validez es que, como ya hemos mencionado anteriormente, un test no puede ser válido si no es fiable. Así, según Bachman (1990: 160), "When we increase the reliability of our measures, we are also satisfying a necessary condition for validity: in order for a test score to be valid, it must be reliable".

Para comprobar la fiabilidad del test de una forma empírica, hemos recurrido al coeficiente **alfa de Cronbach**, comprendido entre 0,00 y 1,00 (Lado, 1961; Hughes, 1989). Para ilustrar al lector sobre la interpretación de este índice estadístico, ofrecemos la siguiente tabla de referencia (Tabla 5.1.):

Coeficiente alfa de Cronbach	Descriptor (interpretación del coeficiente alfa)
< ,60	Inaceptable
,60-,65	Indeseable
,65-,70	Mínimamente aceptable
,70-,80	Respetable
,80-,90	Muy buena
> ,90	Excelente

Tabla 5.1.: Índices de referencia para interpretar el coeficiente alfa de Cronbach (DeVellis, 1991, citado por Gyllstad, 2007: 227)

A la vista de estos valores, el resultado de 0,968 que ha obtenido nuestro test se puede calificar de "excelente" (Tabla 5.2.).

Alfa de Cronbach	N de elementos
,968	200

Tabla 5.2: Coeficiente de fiabilidad de TCA1

Es curioso y muy alentador recordar que un test de competencia colocacional receptiva como el de Gyllstad (2007), test que ha pasado por cuatro versiones diferentes con el fin de incrementar su fiabilidad, consiguió finalmente un alfa de Cronbach de 0,89, constatando con ello la dificultad que implica lograr una fiabilidad verdaderamente satisfactoria en la evaluación colocacional. Entendemos que nuestro magnífico resultado ha tenido mucho que ver con la atenta observación de los factores de Hughes (1989), que ya mencionamos en el capítulo anterior, y que sin duda repercutieron en la fiabilidad intrínseca y extrínseca del test por, entre otras circunstancias, el amplísimo número de ítems utilizado (200), el aceptable número de alumnos para su pilotaje (311), la atenta construcción de la prueba —revisada por sujetos ingleses y españoles—, la variedad de métodos evaluativos, el cuidado al expresar las instrucciones de forma comprensible, la inclusión de tareas que fuesen familiares para los alumnos, la búsqueda de una puntuación objetiva (1/0), condiciones de administración similares para todos los grupos, identificación de la autoría para minimizar el efecto "desinterés", etc.

5.3. La validez

Aunque para la validez no contamos con un índice estadístico semejante al coeficiente alfa de Cronbach, se pueden obtener datos cuantitativos relevantes que evidentemente sustentan la validez de un test en sus dos grandes aspectos: constructo y contenido.

5.3.1. Los resultados del test y la validez de constructo

Empezando por la validez de constructo, debemos recordar que las decisiones tomadas en la etapa de la planificación de nuestra prueba (ver capítulo 4) tuvieron como finalidad, en primer lugar, delimitar el test a la luz de la teoría subyacente del dominio lingüístico que se pretende medir, la colocación. Por otra parte, se trató de hacer operativos aspectos del test que nos permitieran cuantificarlos mediante una investigación empírica. Por tanto, y siguiendo los pasos sugeridos por Read y Chapelle (2001) para conseguir la operatividad de la prueba, delimitamos en primer lugar el tipo de test y su finalidad. Como ya mencionamos, nuestro test se encuadra dentro de la evaluación de diagnóstico y busca medir el conocimiento colocacional de un grupo de alumnos de Filología Inglesa y Traducción e Interpretación a nivel general. También examinamos otras hipótesis más específicas relacionadas con las posibles diferencias entre las licenciaturas, las diferencias entre los niveles o cursos, la homogeneidad o heterogeneidad de los grupos, etc. Asimismo, se intentó detectar los efectos que las distintas estructuras gramaticales de la colocación pudieran tener en sus resultados, sin olvidarnos de la repercusión del uso de distintos métodos o formatos de evaluación (lo que Alderson et al., 1995, denominan *method effect*) y sus consecuencias para la validez de constructo.

A la hora de evaluar todos estos aspectos empíricamente, la estadística descriptiva se convierte en una inestimable ayuda tanto en la obtención de las medidas centrales como de dispersión, ya que tratándose de un test de diagnóstico, su principal objetivo era dar a conocer dichos resultados para, a continuación, establecer unas inferencias pertinentes sobre el grado de conocimiento de los candidatos. Además, con ello lograríamos, a la postre, crear un efecto *washback* que contribuyese a una mejora tanto de la enseñanza como del aprendizaje de las colocaciones. Éstos, sin embargo, no fueron los únicos análisis estadísticos realizados ya que para la obtención de las comparaciones arriba apuntadas tuvimos que recurrir a la estadística muestral (t-test y ANOVA) al igual que a las correlaciones de Pearson. A continuación mostraremos los resultados en detalle.

5.3.1.1. Medidas centrales y de dispersión del grupo en su totalidad

Primeramente, examinaremos los datos de la estadística descriptiva, comenzando por las medidas centrales y de dispersión del grupo total de 311 alumnos. En cuanto a los valores de tendencia central, "índices estadísticos que permiten situar la posición de una distribución y ofrecen el valor de la variable x hacia el cual tienden a agruparse los datos" (Tejada Fernández, 1997: 128), comprobaremos la **media**, el valor que resulta de sumar todos los resultados de una prueba y dividirlos por el número de resultados que hay, la **mediana,** es decir, la puntuación central de la distribución, y por último **la moda**, o el valor más repetido. Por otra parte la obtención de las medidas de dispersión se justifica ya que medidas centrales muy similares pueden poseer características de homogeneidad muy diferentes. Así pues, son las medidas de dispersión (desviación típica y varianza) las que "proporcionan el grado de variabilidad de los datos de una distribución" (ibid.: 130), y por tanto se convierten en los claros indicadores de la homogeneidad o heterogeneidad de un grupo.

Como recordaremos, el principal objetivo de nuestro test de diagnóstico era medir la competencia colocacional de los alumnos a nivel general (Tabla 5.3. y Fig. 5.1.):

ESTADÍSTICA DESCRIPTIVA		
N	Válidos	311
	Perdidos	0
Media		45,6756
Mediana		48,6250
Moda		45,38
Desv. típ.		14,45641
Rango		61,88
Mínimo		10,00
Máximo		71,88

Tabla 5.3.: Estadística descriptiva del total de la población, expresada en porcentajes

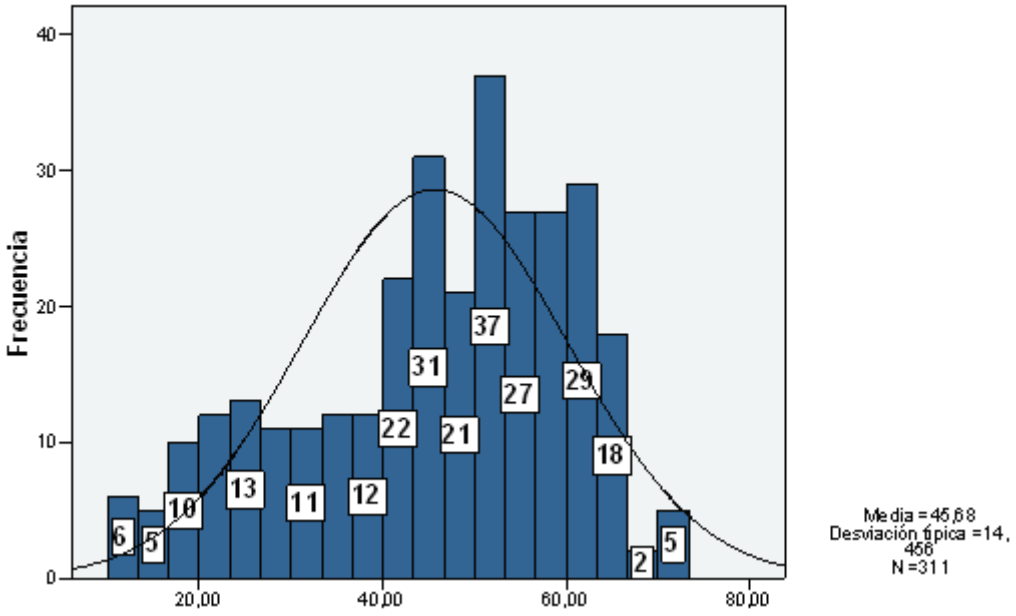


Fig. 5.1.: Histograma de la distribución de resultados del total de la población

De los resultados anteriores podemos colegir lo siguiente:

a) La media de aciertos no alcanza ni siquiera un 50%, lo cual indica que la competencia colocacional del grupo en general se puede considerar muy deficiente y preocupante.

b) Tanto la mediana (48,62%) como la moda (45,38%) no se alejan mucho de la media, medida central por excelencia, por lo que están dentro de valores normales con relación a ésta.

c) La desviación típica (14,45%) es muy alta, dato que indica la gran heterogeneidad del grupo, como suele ocurrir por lo general con la medición de la competencia léxica. Esta cuestión, la diferencia de nivel en el conocimiento del vocabulario, es un hecho ya muy conocido que incluso se manifiesta de forma muy palpable en los hablantes nativos, siendo ésta una de las razones por las que se sigue considerando hoy en día que la etiología de la adquisición del léxico es un verdadero misterio (Pérez Basanta, 2005).

d) La diferencia de conocimiento entre el alumnado (mínimo de 10% y máximo de 71,88%) muestra un rango llamativo de 61,88%. El hecho de que un alumno sólo haya conseguido un 10% nos lleva a pensar que esto puede deberse a que posee una competencia colocacional prácticamente nula, o bien a que ha tenido nulo interés en la cumplimentación del test, aún a riesgo de que el profesor llegase a conocer este hecho. Este fenómeno, que no deja de ser curioso, fue absolutamente minoritario, muy posiblemente debido a que, con el objetivo de evitarlo y contraviniendo los consejos de Hughes (1989) en favor del anonimato, en nuestro test los candidatos estaban obligados a identificarse.

e) El histograma muestra una curva con una campana de Gaus muy aceptable y dentro de parámetros normales. Este dato es absolutamente crucial para validar un test porque al poseer una varianza normal podremos recurrir a todo tipo de

pruebas paramétricas al aplicar la estadística muestral o comparación entre resultados.

5.3.1.2. Medidas centrales y de dispersión de los estudios de Filología Inglesa y Traducción e Interpretación

Analizando, a continuación, los resultados por licenciaturas, nos encontramos con lo siguiente (Tabla 5.4 y Figura 5.2.):

Licenciatura	N	Media	Desv. típ.
Filología Inglesa	136	41,4256	14,75518
Traducción e Interpretación	175	48,9786	13,35390
Total	311	45,6756	14,45641

Tabla 5.4.: Resultados globales por licenciatura expresado en porcentajes

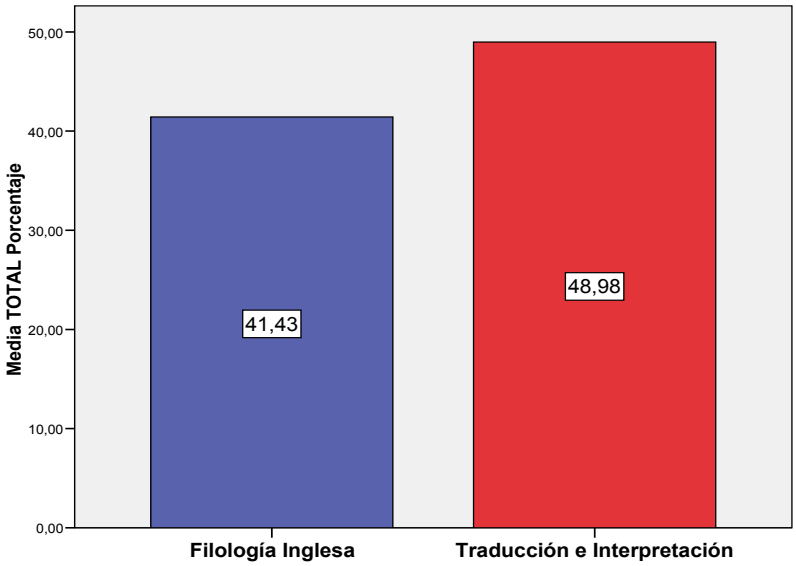


Fig. 5.2.: Representación gráfica de la media por licenciatura

Las siguientes conclusiones se derivan de la información anterior:

a) Las media de los alumnos de Traducción e Interpretación (48,98%) es claramente superior a la de los alumnos de Filología Inglesa (41,43%), siendo la diferencia de siete puntos y medio aproximadamente. Este dato no es extraño en absoluto porque indudablemente el *numerus clausus* al que los alumnos de Traducción están sometidos hace que en términos generales tengan niveles de dominio superiores a los estudiantes de Filología, que son admitidos con calificaciones muy inferiores.
b) La desviación típica es ligeramente más baja en Traducción (13,35%) que en Filología (14,75%). De nuevo, el *numerus clausus* también puede ser la causa de la homogeneidad del grupo.

Hasta ahora, hemos ofrecido una panorámica descriptiva del comportamiento del test con relación a la totalidad de la población y a sus dos grandes grupos (licenciaturas de Filología Inglesa y Traducción e Interpretación). Pasamos ahora a confirmar si las diferencias observadas son estadísticamente significativas, es decir, si, lejos de deberse al azar, son relevantes y podrían generalizarse a grupos semejantes de población. Para llevar a cabo esta comprobación, emplearemos la llamada prueba "t" o t-test, que se usa generalmente para contrastar dos medias procedentes de distintas muestras (t-test no pareado) o de idénticos sujetos (t-test pareados) (García Roldán, 1995). Al realizar esta prueba hay que asumir siempre un valor de error, que en nuestro caso es de <0,05. No creemos necesario extendernos más en estos conceptos que son de estadística básica, pero sí nos gustaría mencionar que a la prueba "t" se le pueden aplicar pruebas paramétricas, en donde se asume que la distribución de los resultados es normal (campana gausiana en un histograma), o bien pruebas no paramétricas, en el caso de que la distribución esté sesgada a la derecha o a la izquierda (es decir, cuando no hay una distribución de campana gausiana porque la mayoría de los alumnos obtuvieron una puntuación marcadamente alta o baja en el

test). En nuestro caso y como ya hemos comprobado en el histograma de la figura 5.1., la distribución es normal, por lo que la prueba paramétrica sería la más indicada y la más robusta en cuanto a sus resultados.

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
		Inferior	Superior	Inferior	Superior	Inferior	Superior	Inferior	Superior	Inferior
TOTAL Porcentaje	Se han asumido varianzas iguales	4,103	,044	-4,725	309	,000	-7,55302	1,59847	-10,69828	-4,40776
	No se han asumido varianzas iguales			-4,666	275,088	,000	-7,55302	1,61860	-10,73943	-4,36661

Tabla 5.5.: Comparación de medias de Filología y Traducción mediante t-test de muestras independientes

Como muestra la tabla anterior, la diferencia entre los grupos es absolutamente significativa con un valor de p=0,00, y por tanto podemos afirmar que existe una diferencia que no se debe al azar. Así, esta afirmación se podría generalizar a una población semejante, en este caso los alumnos de Filología Inglesa y Traducción, con suficiente grado de confianza.

5.3.1.3. Comparaciones de las medias totales por cursos

Una de las cuestiones que más claramente afectan a la validez de constructo es el concepto de escalabilidad: el hecho de que un test obtenga resultantes diferentes según los diferentes grados de dominio, en este caso con referencia a los cursos de las licenciaturas. Así pues, la hipótesis de la escalabilidad se formula como el lógico

incremento en los resultados según los diferentes niveles de competencia. En nuestro test, los resultados en este sentido fueron los siguientes (Tabla 5.6. y Figura 5.3.):

Curso	N	Media	Desv. típ.
Primero	80	28,6828	13,33364
Segundo	71	47,9824	9,53488
Tercero	95	50,5342	8,58531
Cuarto	65	56,9692	7,32858
Total	311	45,6756	14,45641

Tabla 5.6.: Medias por cursos expresadas en porcentajes

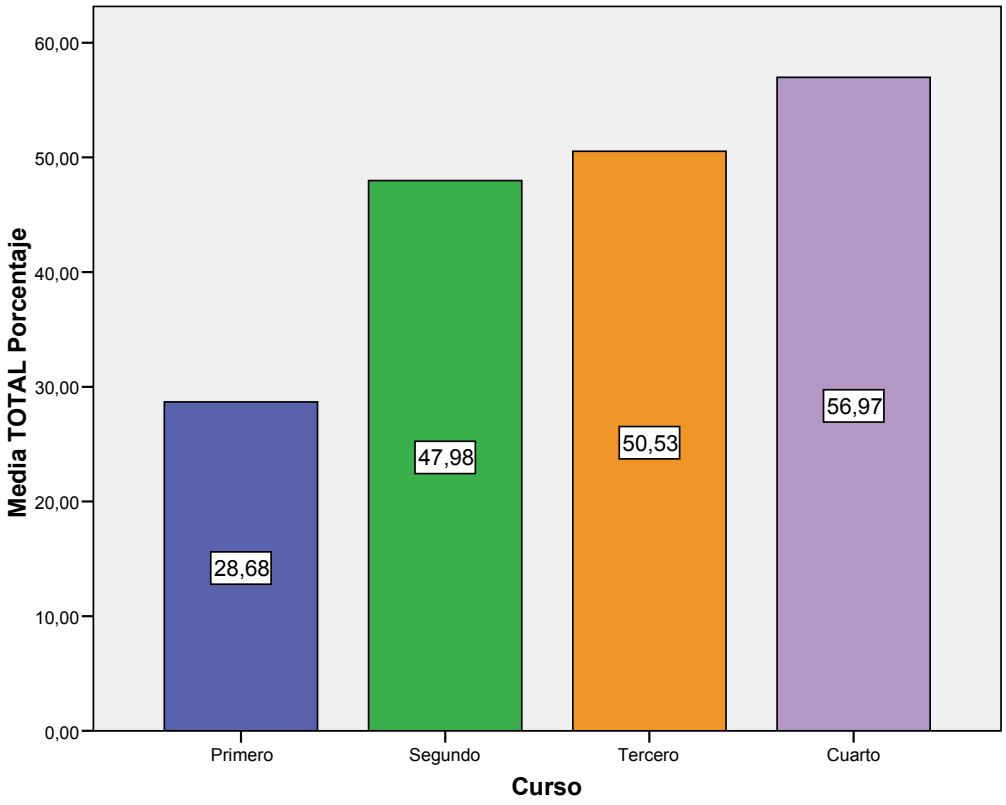


Fig. 5.3.: Representación gráfica de las medias de cada curso

La Tabla 5.6. y la Figura 5.3. arrojan los siguientes datos:

a) A la vista de esta información, se observa que sí existe diferencia de niveles en cuanto a los 4 cursos, que van desde una media de 28,68% en 1er curso hasta un 56,97% para el 4º curso, pasando por medias de 47,98% y 50,53% para el 2º y el 3er curso respectivamente. Podemos comprobar, por tanto, que se produce una escalabilidad.

b) La diferencia más abultada se da entre los cursos 1º y 2º, quizá debido a que existe un filtro natural por el tradicional abandono de los alumnos principiantes ante el fracaso inicial.

c) Se registra también un incremento superior entre 3er y 4º curso, en este caso muy posiblemente motivado por la estancia Erasmus de los alumnos de 4º que, como recordaremos, en el caso de Filología les hacía compartir asignatura con alumnos de 3º.

d) Cabe subrayar que la diferencia más baja se da entre el 2º y el 3er curso, donde casi no se puede apreciar un incremento de la competencia colocacional.

Para corroborar estos datos, recurrimos de nuevo a las comparaciones estadísticas que nos facilita la prueba ANOVA (Tabla 5.7.).

Scheffé

(I) Curso	(J) Curso	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Primero	Segundo	-19,29958*	1,63246	,000	-23,8885	-14,7106
	Tercero	-21,85140*	1,51929	,000	-26,1222	-17,5806
	Cuarto	-28,28642*	1,67190	,000	-32,9862	-23,5866
Segundo	Primero	19,29958*	1,63246	,000	14,7106	23,8885
	Tercero	-2,55182	1,57070	**,452**	-6,9671	1,8635
	Cuarto	-8,98684*	1,71875	,000	-13,8183	-4,1553
Tercero	Primero	21,85140*	1,51929	,000	17,5806	26,1222
	Segundo	2,55182	1,57070	**,452**	-1,8635	6,9671
	Cuarto	-6,43502*	1,61165	,001	-10,9655	-1,9046
Cuarto	Primero	28,28642*	1,67190	,000	23,5866	32,9862
	Segundo	8,98684*	1,71875	,000	4,1553	13,8183
	Tercero	6,43502*	1,61165	,001	1,9046	10,9655

*. La diferencia de medias es significativa al nivel .05.

Tabla 5.7.: Comparación de las medias de los cursos mediante ANOVA

Esta tabla nos muestra cómo las diferencias son significativas entre todos los cursos a excepción de 2º y 3º, donde el valor p=0,452 no es significativo. Este resultado que en principio nos pareció un tanto incomprensible y preocupante, coincide plenamente con los resultados del test de colocaciones de Gyllstad (2007), en donde, como en nuestro caso, no obtiene una diferencia significativa entre 2º y 3º (incluso en su caso es más grave porque no hay diferencia alguna). Este autor, que ha llevado a cabo numerosos pilotajes de su test, lo achaca al fenómeno de que para conseguir ganancias en la competencia colocacional se necesita más de un curso académico ya que el conocimiento que se adquiere en el periodo de un año "does not develop to the extent that a difference is measurable" (ibid.: 126).

5.3.1.4. Comparaciones de las medias por cursos y licenciaturas

A fin de apoyar la hipótesis anterior en el sentido de que el conocimiento colocacional requiere de plazos más extensos para apreciar claras diferencias, hemos

realizado análisis estadísticos semejantes al anterior, pero ahora concretándolos en los cursos y las licenciaturas. En lo que se refiere a Filología Inglesa los resultados son los que se muestran a continuación (Tablas 5.8. y 5.9. y Figura 5.4.):

Curso	N	Media	Desv. típ.
Primero	46	25,4837	8,66793
Segundo	29	44,3319	8,96392
Tercero	38	49,5691	8,73545
Cuarto	23	56,1902	8,16832
Total	136	41,4256	14,75518

Tabla 5.8.: Medias por cursos en Filología Inglesa expresadas en porcentajes

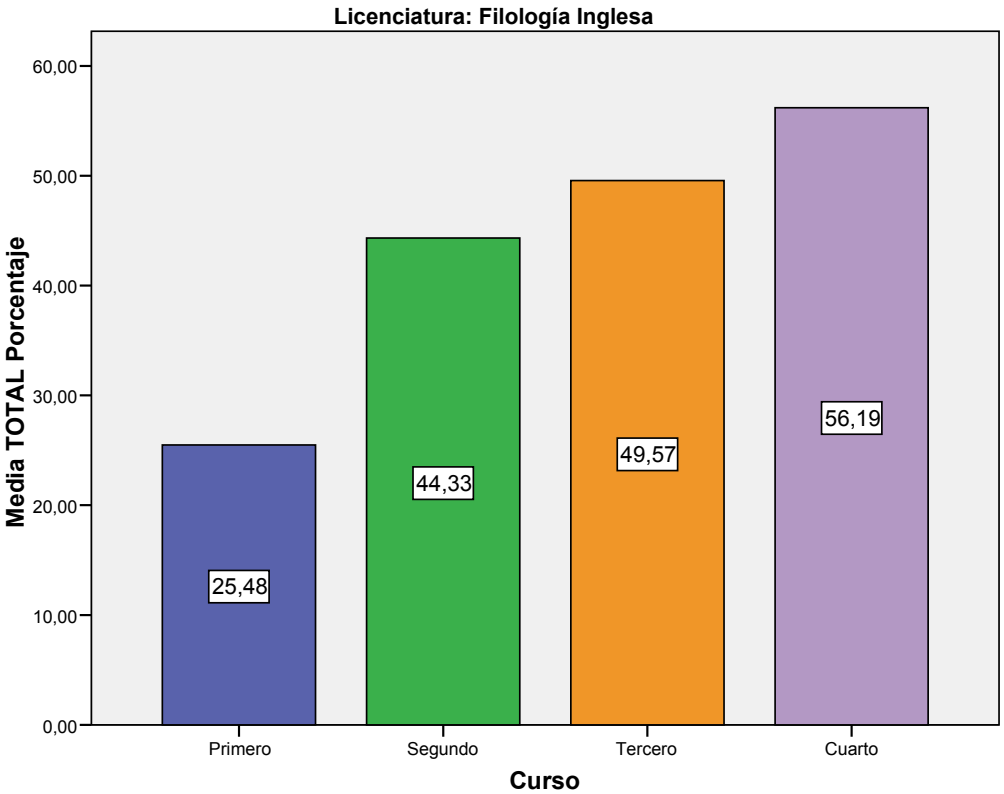


Fig. 5.4.: Representación gráfica de las medias por cursos en Filología Inglesa

Scheffé

(I) Curso	(J) Curso	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Primero	Segundo	-18,84820*	2,05574	,000	-24,6698	-13,0266
	Tercero	-24,08538*	1,90057	,000	-29,4676	-18,7032
	Cuarto	-30,70652*	2,21410	,000	-36,9766	-24,4364
Segundo	Primero	18,84820*	2,05574	,000	13,0266	24,6698
	Tercero	-5,23718	2,13777	,117	-11,2911	,8168
	Cuarto	-11,85832*	2,42077	,000	-18,7137	-5,0029
Tercero	Primero	24,08538*	1,90057	,000	18,7032	29,4676
	Segundo	5,23718	2,13777	,117	-,8168	11,2911
	Cuarto	-6,62114*	2,29047	,043	-13,1075	-,1348
Cuarto	Primero	30,70652*	2,21410	,000	24,4364	36,9766
	Segundo	11,85832*	2,42077	,000	5,0029	18,7137
	Tercero	6,62114*	2,29047	,043	,1348	13,1075

*. La diferencia de medias es significativa al nivel .05.

Tabla 5.9.: Comparación de medias por cursos en Filología Inglesa mediante ANOVA

A continuación veremos los resultados correspondientes a la licenciatura de Traducción e Interpretación:

Curso	N	Media	Desv. típ.
Primero	34	33,0110	17,01727
Segundo	42	50,5030	9,18752
Tercero	57	51,1776	8,50013
Cuarto	42	57,3958	6,89263
Total	175	48,9786	13,35390

Tabla 5.10.: Medias por cursos en Traducción e Interpretación expresadas en porcentajes

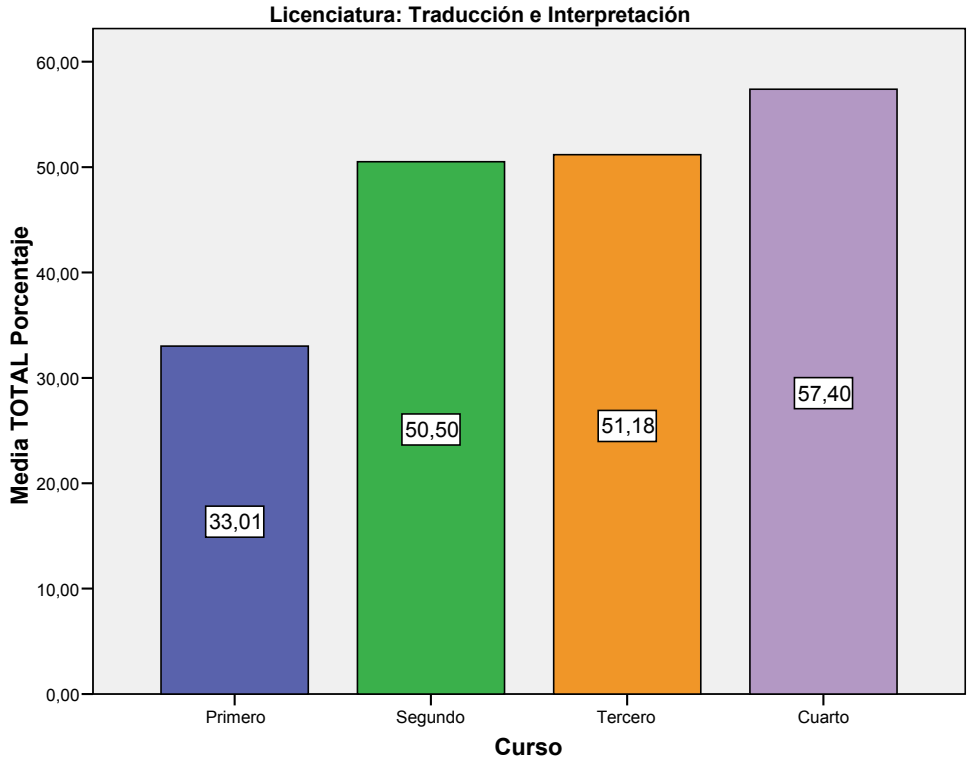


Fig. 5.5.: Representación gráfica de las medias por cursos en Traducción e Interpretación

Scheffé

(I) Curso	(J) Curso	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Primero	Segundo	-17,49195*	2,43248	,000	-24,3601	-10,6238
	Tercero	-18,16660*	2,28481	,000	-24,6178	-11,7154
	Cuarto	-24,38480*	2,43248	,000	-31,2530	-17,5166
Segundo	Primero	17,49195*	2,43248	,000	10,6238	24,3601
	Tercero	-,67466	2,14418	**,992**	-6,7288	5,3795
	Cuarto	-6,89286*	2,30090	,032	-13,3895	-,3962
Tercero	Primero	18,16660*	2,28481	,000	11,7154	24,6178
	Segundo	,67466	2,14418	**,992**	-5,3795	6,7288
	Cuarto	-6,21820*	2,14418	,041	-12,2724	-,1640
Cuarto	Primero	24,38480*	2,43248	,000	17,5166	31,2530
	Segundo	6,89286*	2,30090	,032	,3962	13,3895
	Tercero	6,21820*	2,14418	,041	,1640	12,2724

*. La diferencia de medias es significativa al nivel .05.

Tabla 5.11.: Comparación de medias por cursos en Traducción mediante ANOVA

Por último, la figura 5.6. muestra las diferencias entre los cursos de las respectivas licenciaturas de forma comparativa.

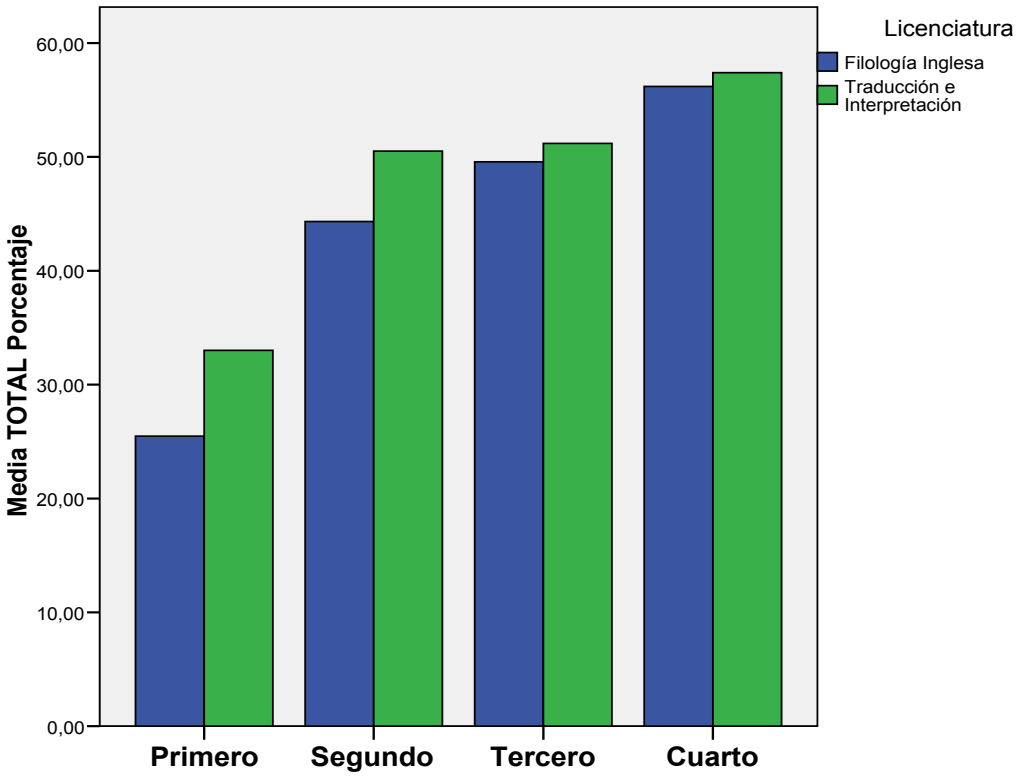


Fig. 5.6.: Comparativa de las medias por curso en Filología y Traducción

Los datos mostrados en las páginas anteriores nos indican lo siguiente:

a) En el caso de Filología Inglesa, la diferencia entre 2º y 3º es considerable (5 puntos separan los resultados de ambos cursos) pero no consigue ser estadísticamente significativa, como indica el valor de la ANOVA (p=0,117).

b) En Traducción, por su parte, existe sólo medio punto de diferencia entre 2º y 3º, con lo cual la comparación tampoco es significativa, consiguiendo un valor de p=0,992.

c) Se repiten los patrones que obtuvimos en la comparación a nivel general, mostrando la superioridad de los resultados de la licenciatura de Traducción frente a los de Filología.

d) La razón que apunta Gyllstad (2007) en cuanto al periodo requerido para mejorar las colocaciones se corrobora con estos datos y los anteriores.

5.3.1.5. Medidas centrales y de dispersión con relación a las estructuras gramaticales de las colocaciones

Como ya hemos apuntado, pretendemos también recabar información con relación a cómo las diferentes modalidades combinatorias afectan al conocimiento de las colocaciones. Estas estructuras son: nombre+nombre (N+N), adjetivo+nombre (A+N), verbo+nombre (V+N) y nombre+verbo (N+V). Recordemos que las proporciones en las que nos basamos a la hora de confeccionar los ítems fueron las obtenidas en nuestro listado de frecuencias extraído del BNC y el BoE. Debemos puntualizar, sin embargo, que el número real varió ligeramente debido a la necesidad de cuadrarlas en números enteros. En la tabla 5.12. mostramos las proporciones que aparecen en el test:

		Frecuencia	Porcentaje	Porcentaje acumulado
Válidos	N+N	14	7,0	7,0
	A+N	117	58,5	65,5
	V+N	64	32,0	97,5
	N+V	5	2,5	100,0
	Total	200	100,0	

Tabla 5.12.: Proporciones de tipos de colocaciones según su estructura gramatical

Como dijimos al hablar de la validez de constructo, explorar las posibilidades combinatorias de la colocación en cuanto a su aparición real en los textos del BNC y

el BoE es también una forma de profundizar en la trascendencia de la colocación en el lenguaje real y, por ende, en su validez de constructo. Como apuntamos anteriormente, la frecuencia con que se dan las colocaciones en el inglés actual según nuestro listado coloca a las de A+N entre las más numerosas seguidas de las de V+N. Ya a una distancia considerable aparecen las de N+N y, finalmente, las de N+V son prácticamente testimoniales.

No obstante, si la frecuencia aporta datos hasta ahora desconocidos, la dificultad que muestran las estructuras combinatorias es de un interés excepcional, por lo novedoso, para la pedagogía de las colocaciones. Es muy revelador que no existan estudios serios y fundamentados en un número suficiente de datos, en nuestro conocimiento, que hayan investigado este importante hecho y sin embargo, es un lugar común subrayar la mayor dificultad e interés de las combinaciones N+V (como recordaremos, ésta es una de las conclusiones alcanzada por Gitsaki, 1999). En la tabla y la figura que siguen, mostramos los resultados obtenidos en nuestro test en relación a las diferentes estructuras gramaticales contempladas en TCA1 (Tabla 5.12. y Figura 5.7.).

Estructura gramatical	N	Media	Desv. típ.
N+N	14	47,2143	15,14128
A+N	117	54,1111	21,21799
V+N	64	59,3906	21,18147
N+V	5	52,4000	23,56480
Total	200	55,2750	21,01494

Tabla 5.12.: Medias por categorías gramaticales

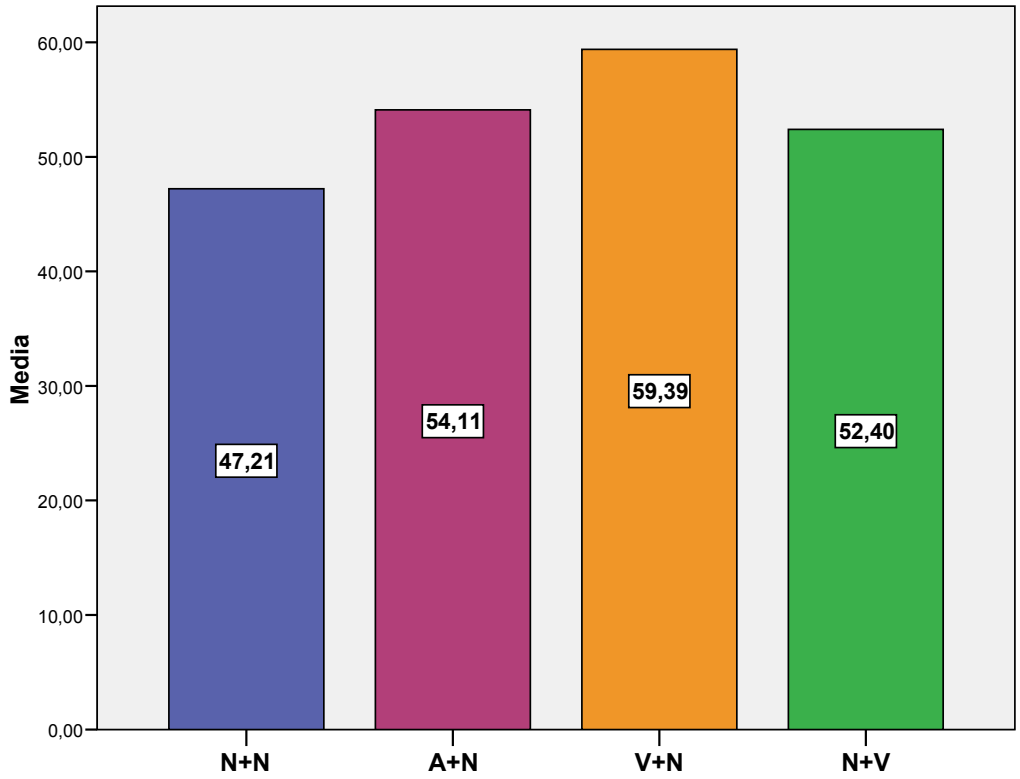


Fig. 5.7.: Representación gráfica de las medias por categorías gramaticales

Si examinamos ahora las distintas categorías en términos de dificultad de adquisición podemos concluir que:

a) Efectivamente existe, según este estudio, diferencia en la adquisición de las colocaciones según los modelos combinatorios.

b) Se aprecia que las de N+N (47,21%) son las más difíciles seguidas de las de N+V (52,40%), aunque debemos advertir que esta casuística es muy poco fiable por el ínfimo número de casos con los que hemos operado (14 ítems de N+N y sólo 5 de N+V). Por tanto, la diferencia entre A+N y V+N constituye el dato más sólido dado que el número de casos es aceptable. A diferencia de las expectativas que teníamos, basadas en la creencia popular de que las colocaciones V+N son

las más difíciles, este estudio apunta todo lo contrario: la combinación A+N (54,11%) muestra una mayor dificultad que la V+N (59,39%).

Al realizar una prueba de t-test para comprobar si existe una diferencia significativa entre los resultados obtenidos en las colocaciones A+N y V+N, nos encontramos con lo siguiente (Tabla 5.13.):

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de diferencia	95% Intervalo de confianza para la diferencia	
		Inferior	Superior	Inferior	Superior	Inferior	Superior	Inferior	Superior	Inferior
media	Se han asumido varianzas iguales	,084	,772	-1,601	179	,111	-5,27951	3,29684	-11,78518	1,22615
	No se han asumido varianzas iguales			-1,602	129,888	**,112**	-5,27951	3,29517	-11,79866	1,23963

Tabla 5.13.: Comparación de medias de A+N y V+N mediante t-test

Al comprobar la relación entre las dos muestras asumiendo varianzas diferentes, vemos que la diferencia no es significativa (p=0,112). A la vista de estos resultados, sólo podemos hablar de "tendencia" en la dificultad de las A+V. Quizá en el futuro con un número de casos superior se podrá corroborar este dato.

5.3.1.6. Comparación de los diferentes métodos de evaluación que componen el test (análisis por secciones)

A la hora de diseñar los distintos métodos para la construcción de los ítems, no nos preocupó tanto que hubiera una diferencia de dificultad entre ellos (pensábamos que el que algunos resultaran de mayor o menor dificultad era de hecho positivo para un correcto pilotaje) sino mostrar una correlación aceptable entre ellos. Recordemos que el test estaba constituido por 4 métodos distintos, marcando las 4 secciones descritas en el capítulo anterior: Sección A con un formato de tipo *c-test* donde el alumno produce el colocado partiendo de la primera letra ya dada, y dentro de un contexto; Sección B en la que el alumno ha de traducir el colocado del español al inglés; Sección C con formato *odd-one-out*, donde el alumno debe seleccionar la opción incorrecta o, en su caso, declarar que todas son correctas; Sección D de elección múltiple donde el alumno elige la opción correcta o, en su caso, afirmar que todas son incorrectas.

En primer lugar, contemplaremos los resultados obtenidos y, por tanto, el nivel de dificultad de estos cuatro métodos de evaluación, es decir, de cada sección del TCA1 (Tabla 5.14. y Figura 5.8.):

Secciones	N	Media	Desv. típ.
A	50	67,320	16,51770
B	50	65,900	16,76762
C	50	37,720	16,08021
D	50	50,160	19,40583
Total	200	55,275	21,01494

Tabla 5.14.: Medias según secciones expresadas en porcentajes

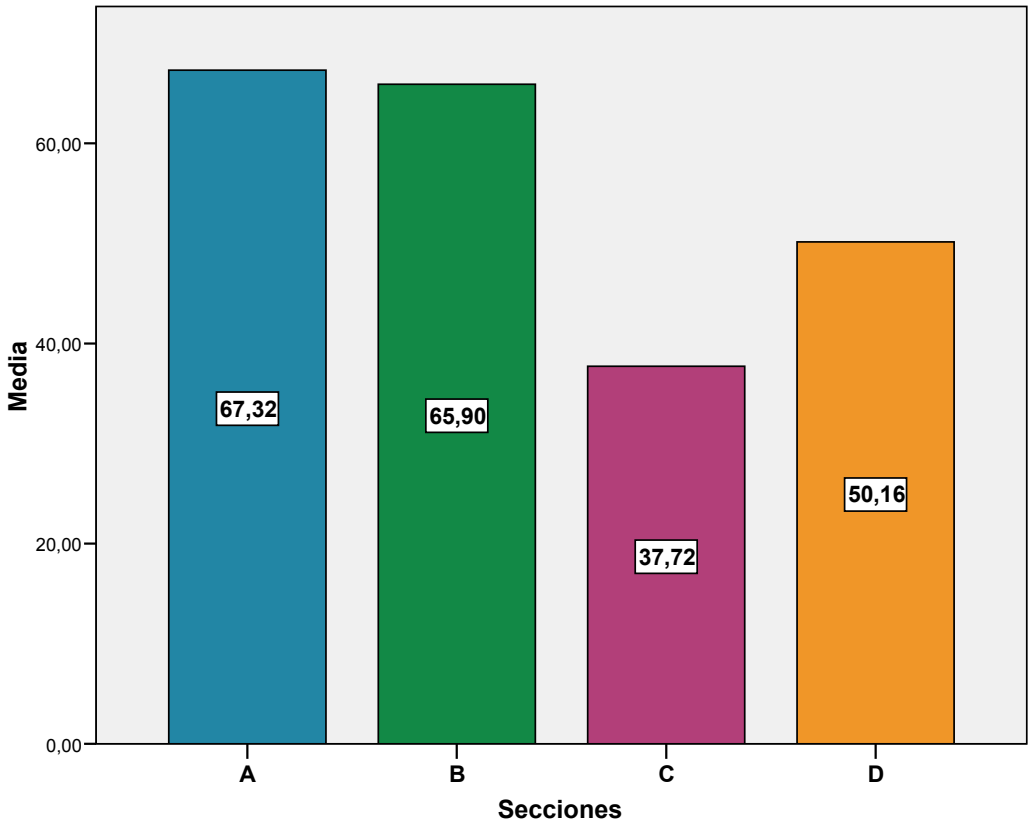


Fig. 5.8.: Representación gráfica de las medias por secciones

A la luz de los resultados anteriores concluimos que:

a) Existen claras diferencias en cuanto a la dificultad, sobre todo, entre el grupo formado por las secciones A y B con respecto a C y D.

b) La Sección A consigue los mejores resultados (67,32%), probablemente porque además de aparecer la colocación contextualizada, se le ofrece la primera letra del colocado a modo de *c-test*. Esto corrobora el hecho de que la inclusión del contexto siempre ayuda tanto al reconocimiento como a la producción en la adquisición de una lengua (Read, 2000); al mismo tiempo la técnica del *c-test* ha sido ampliamente defendida y sus ventajas con relación al método *cloze*

tradicional y otras técnicas evaluadoras han sido puestas de manifiesto por varios autores (Connelly, 1997; Weir, 1990).

c) La opción B, en donde el alumno tiene que traducir el colocado a la lengua inglesa, obtiene unos resultados muy similares a la anterior (65,90%), corroborando que la traducción parece ser una técnica adecuada a la hora de medir la competencia colocacional.

d) La sección C es la que resultó marcadamente más difícil para el alumnado (37,72%), y la explicación parece lógica: el alumno tiene que descartar el ítem que no es correcto de entre 3 o incluso 4 opciones —en el caso de que la opción correcta sea "*no wrong collocation*"— muy similares entre sí, por tanto se le evalúa de varios elementos a la vez. Además, es muy probable que la total ausencia de contexto contribuyera a incrementar la dificultad de los ítems.

e) Por último, la opción D ha obtenido unos resultados más bajos (50,16%) de los esperados, ya que es una tarea que *a priori* puede parecer la más sencilla puesto que el alumno tiene que elegir una colocación de entre 4 (con la opción de "*none of these*") para que encaje en un contexto. En nuestra opinión, el efecto cansancio es muy probablemente el responsable de este ligero descenso en la media. Esto además lo corroboraríamos con los datos de la desviación típica, que es la más alta de las cuatro secciones (19,4%), lo cual indica que en el grupo hubo puntuaciones muy altas en esta sección, mientras que otras fueron muy bajas. Esta fluctuación, que no es tan marcada en el resto de secciones, parece responder a la influencia de un factor externo, que, como decimos, en nuestra opinión puede ser el cansancio. Por otro lado, nos parece también interesante destacar que, a pesar de ser la última sección, obtuvo mejores resultados que la Sección C, lo cual puede también indicar que el contexto en este caso resultó ser de ayuda para el alumno.

Pero, como ya adelantamos, quizá lo más interesante no es tanto la dificultad como la evidencia de que las secciones se correlacionan bien entre sí, es decir, los alumnos con mejores resultados contestan los ítems más difíciles mientras que los de nivel más bajo encuentran mayor dificultad en todas las secciones. Veamos la tabla 5.15., obtenida mediante una de las pruebas de correlación más usuales cuando se trabaja con un mismo grupo de sujetos, el test de Pearson:

		SECCIÓN A	SECCIÓN B	SECCIÓN C	SECCIÓN D
SECCIÓN A	Correlación de Pearson	1	,904**	,502**	,677**
	Sig. (bilateral)		,000	,000	,000
	N	311	311	311	311
SECCIÓN B	Correlación de Pearson	,904**	1	,474**	,637**
	Sig. (bilateral)	,000		,000	,000
	N	311	311	311	311
SECCIÓN C	Correlación de Pearson	,502**	,474**	1	,602**
	Sig. (bilateral)	,000	,000		,000
	N	311	311	311	311
SECCIÓN D	Correlación de Pearson	,677**	,637**	,602**	1
	Sig. (bilateral)	,000	,000	,000	
	N	311	311	311	311

**. La correlación es significativa al nivel 0,01 (bilateral).

Tabla 5.15.: Correlación entre las distintas secciones de TCA1

Antes de comentar la información de la tabla anterior nos parece también interesante mostrar los diagramas de dispersión, puesto que creemos que representan muy gráficamente las diferencias entre secciones. Además, estos diagramas nos muestran el coeficiente de determinación, es decir, el valor r (índice de correlación de Pearson expresado en la tabla anterior) elevado al cuadrado (por lo que se denomina "*sq r lineal*" como vemos en los diagramas siguientes). Al multiplicar este coeficiente por 100 obtendremos el porcentaje de varianza, es decir, el porcentaje de resultados que muestran un comportamiento similar en las dos secciones que se están comparando en cada caso y que, por tanto, indican el grado de correlación entre ambas:

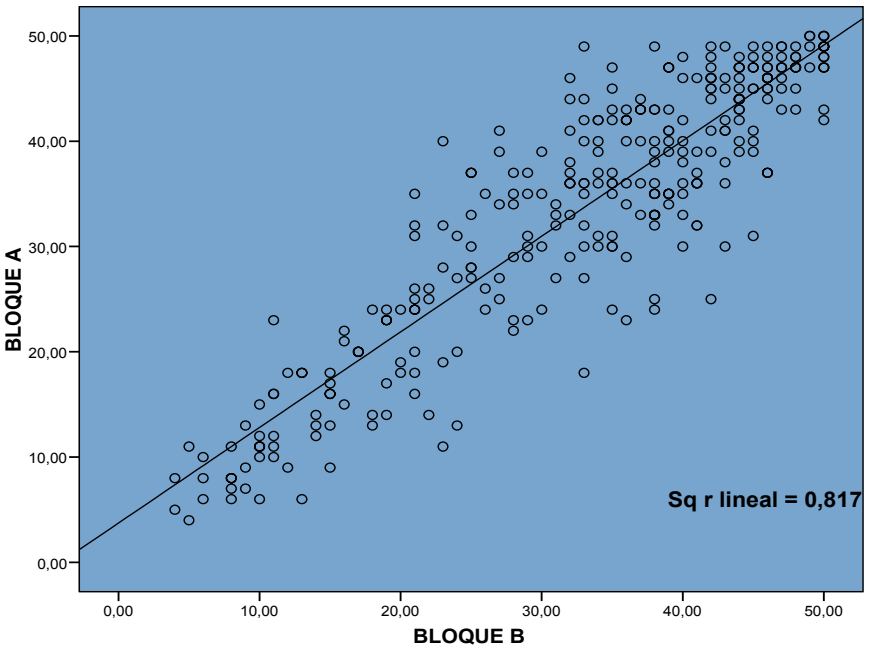


Fig. 5.9.: Diagrama de dispersión: Correlación secciones A y B

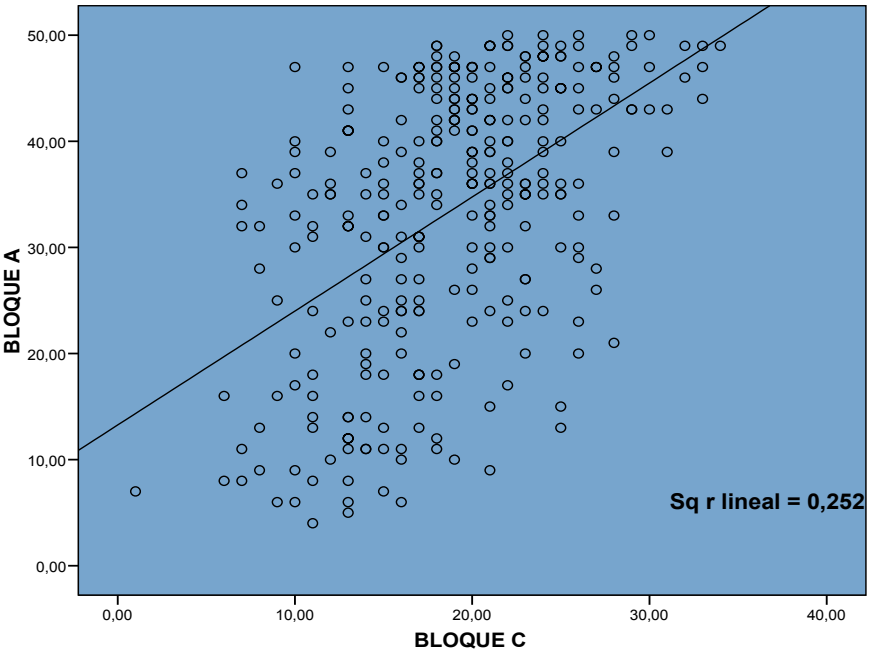


Fig. 5.10.: Diagrama de dispersión: Correlación secciones A y C

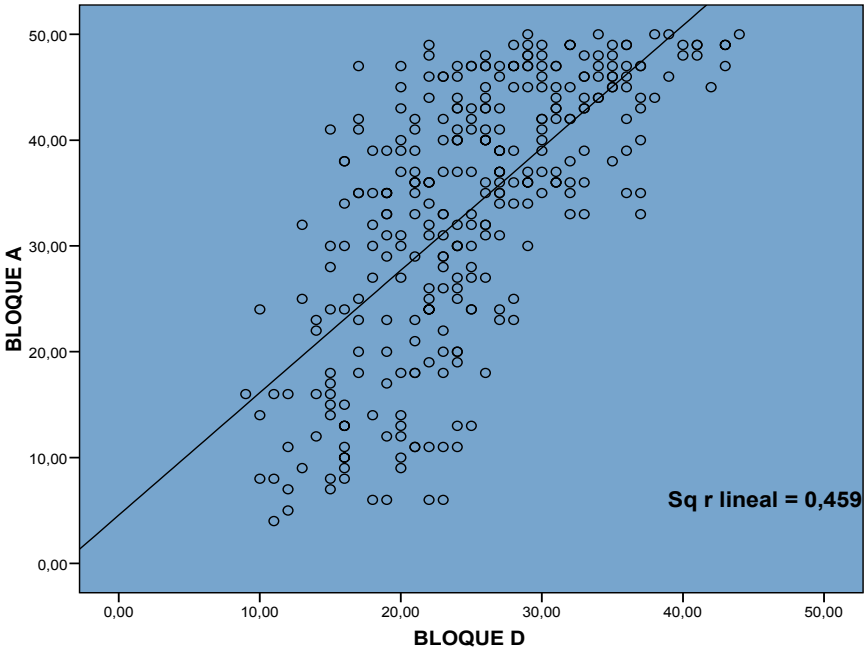


Fig. 5.11.: Diagrama de dispersión: Correlación secciones A y D

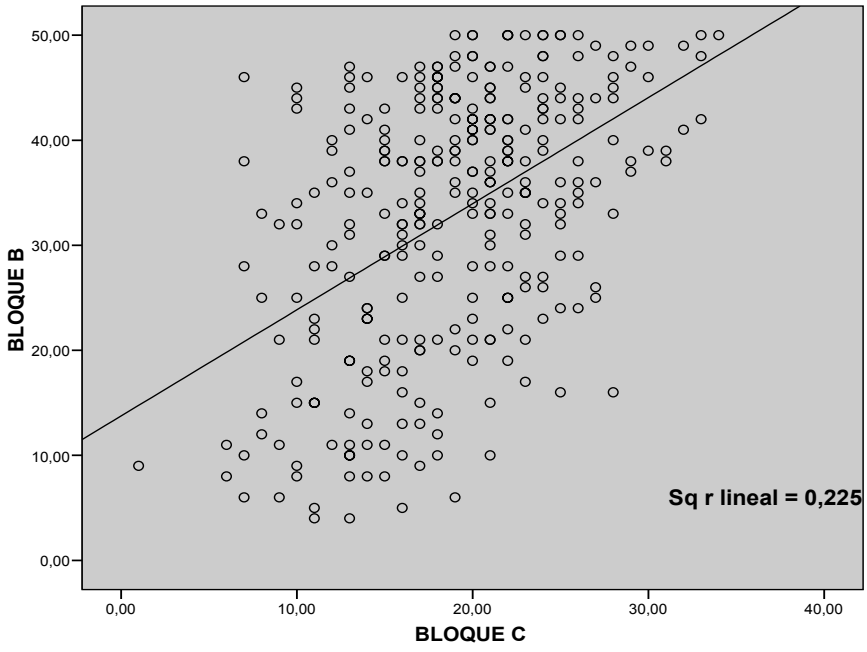


Fig. 5.12.: Diagrama de dispersión: Correlación secciones B y C

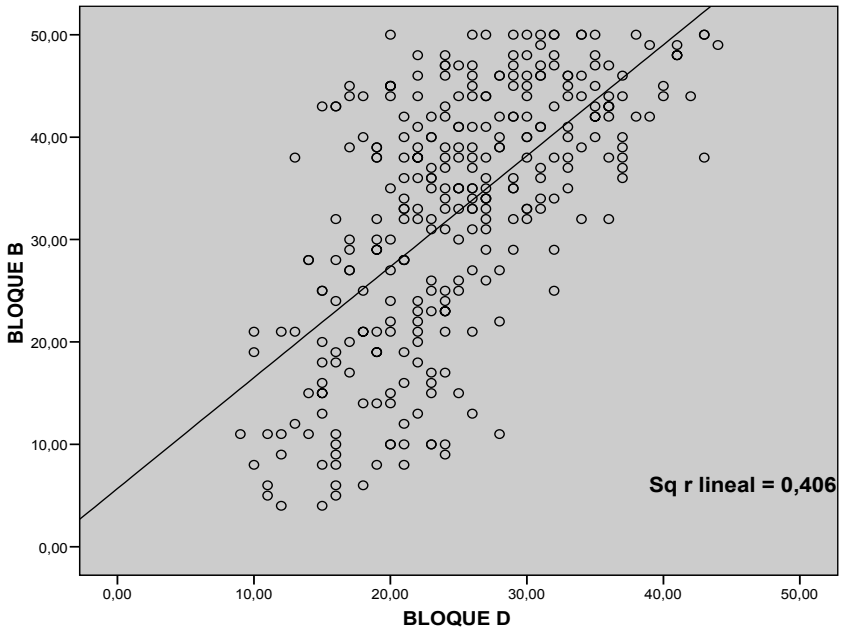


Fig. 5.13.: Diagrama de dispersión: Correlación secciones B y D

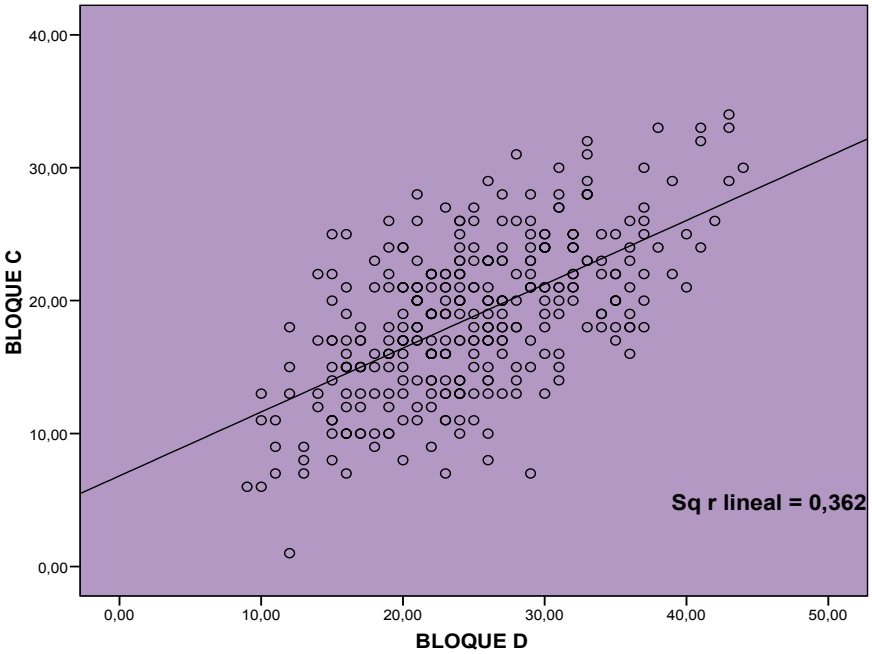


Fig. 5.14.: Diagrama de dispersión: Correlación secciones C y D

Los resultados anteriores nos informan de lo siguiente:

a) Como hemos visto en la tabla 5.15., existe una correlación no sólo positiva sino también significativa (marcada con dos asteriscos) —es decir no es debida a error o al azar— entre las cuatro secciones. Para mejor interpretar los datos de esta tabla es conveniente tener en cuenta los valores de referencia que apunta Cohen (1988, citado por Pallant, 2007: 132) (Tabla 5.16.):

Valores según análisis de Pearson (valores r)	Grado de correlación
r= ,10 a ,29	Baja
r= ,30 a ,49	Media
r= ,50 a 1,0	Alta

Tabla 5.16.: Valores de referencia para interpretar el índice r de Pearson

Así pues, podemos observar que entre la Sección A y la B existe una correlación verdaderamente alta (r = 0,904, cifra que al elevarla al cuadrado y multiplicarla por 100 indica una coincidencia de puntuaciones del 81,7%). Este dato confirma asimismo los resultados obtenidos por Bahns y Eldaw (1993), quienes, como recordaremos, utilizaron dos formatos muy similares a los que conforman nuestras dos primeras secciones y no hallaron diferencias significativas (ver sección 4.2.2.1). Aunque algo más baja, esta correlación también es perceptible entre A y D (r = 0,677, es decir, 45,9% de coincidencia) y entre B y D (r = 0,637, lo que indica un 40,6% de solapamiento).

b) Las correlaciones de A, B y D con la sección C son, sin embargo, claramente inferiores (r = 0,502, r = 0,474 y r = 0,602 respectivamente, lo cual arroja unos porcentajes de coincidencia de 25,2%, 22,5% y 36,2%). Si bien es cierto que

todos los valores se encuentran entre los parámetros de correlación media y alta según Cohen (ibid.), consideramos, no obstante, que estos índices evidencian una extraña desviación de la Sección C con respecto a las demás.

c) Los diagramas de dispersión confirman lo observado anteriormente. Vemos que en los diagramas que comparan A y C, por un lado, y B y C, por otro, existe un menor agrupamiento de los resultados y un ligero incremento en el número de elementos que tienden hacia la esquina superior izquierda y la inferior derecha, todo lo cual nos indica un menor grado de coincidencia o similitud en la actuación de los sujetos en las dos secciones comparadas. En lo que se refiere a la figura 5.14., en la que se comparan C y D, observamos que el agrupamiento es algo mayor, pero la línea que se traza tiene una tendencia más horizontal que el resto, lo cual también evidencia menor correlación que en las comparaciones donde no intervine la Sección C.

Todo lo anterior nos lleva a pensar que el formato de la sección C, "averiguar el intruso" (*odd-one-out*), no parece ser un método adecuado para la evaluación de colocaciones y quizá hubo un error en nuestro planteamiento inicial al escogerlo. Por un lado, la familiaridad que los alumnos suelen tener con este método —es una de las técnicas más utilizadas tradicionalmente en la enseñanza de las colocaciones (Lewis, 2000c)— nos hacía pensar que podría ser un importante factor para incrementar la fiabilidad. Por otro lado, el método *odd-one-out* se eligió con el fin de evaluar un amplio espectro colocacional de la palabra (algo que, según Hargreaves, 2000, debe formar parte del constructo de un test de colocaciones), pero al hacerlo estábamos también desoyendo la razón de peso muy repetida desde el campo de la evaluación de que no se debe medir más de un elemento a la vez (Heaton, 1989). Así pues, ésta puede ser otra de las razones más probables que explican su dificultad y su extraño comportamiento con respecto al resto de métodos (sobre esto volveremos más

adelante cuando analicemos el análisis individual de los ítems con relación a su dificultad y discriminación).

En términos generales, los aspectos arriba señalados nos parecen un ejemplo revelador de cómo la cuantificación de los resultados nos conduce a la detección de errores en el diseño de un test que deben ser subsanados en la etapa del pilotaje, si verdaderamente queremos conseguir esa tan ansiada validez de constructo.

5.3.2. Análisis de ítems: la validez de contenido

La validez de contenido se ha definido como la cualidad de un test para medir lo que debe medir (los contenidos) y en las proporciones adecuadas (Pérez Basanta et al., 1992). Como destaca Hughes (1989), podemos considerar que un test posee validez de contenido cuando constituye una muestra representativa de las habilidades y conocimientos lingüísticos que trata de comprobar. Por ello, resulta fundamental llevar a cabo una selección de contenidos ─lo que en evaluación se conoce como las especificaciones de un test─ y una elaboración de ítems correcta y adecuada para lograr que el resultado final no se vea afectado por la inadecuación de las técnicas o métodos utilizados ─el conocido y ya mencionado *method effect* (Alderson et al., 1995)─. Ya en el capítulo anterior describimos cómo habíamos llevado a cabo ambos procesos en la construcción de nuestro test de colocaciones para lograr la validez de contenido que una prueba debe atestiguar. Con el fin de comprobar si el test es satisfactorio en este sentido, mostraremos a continuación los resultados del estudio sobre la dificultad y discriminación de los ítems, para finalmente explorar cómo han funcionado los distractores en la sección con formato de opción múltiple.

5.3.2.1. Análisis de ítems

Estudiar la calidad métrica de un test y de cada uno de sus ítems es fundamental en la construcción de instrumentos de medida ya que nos informa de las características estadísticas y el comportamiento individual de cada ítem, así como del grado en el que contribuye a fomentar la validez de constructo y de contenido. Así, el análisis de ítems es útil, por un lado, para la construcción y mejora de un test, y, por otro lado, para maximizar la fiabilidad y validez del mismo. Por tanto, la finalidad de este análisis es seleccionar aquellos ítems que presentan una mayor calidad psicométrica y que, al mismo tiempo, se ajustan a los criterios de fiabilidad y validez establecidos previamente. En nuestro caso esto se llevó a cabo en dos fases: en el estudio piloto y en la confección de la prueba final. De esta forma, realizamos primero un examen pormenorizado del funcionamiento de los ítems de TCA1, con la idea de eliminar aquellos que resultan demasiado fáciles, demasiado difíciles o que no contaban con un índice de discriminación adecuado. Finalmente, rediseñamos nuestro test dando lugar al Test de Colocaciones ADELEX Versión 2 (TCA2), con las modificaciones pertinentes.

Así pues, las medidas de las que ahora nos ocuparemos son el coeficiente de dificultad (CD) y el índice de discriminación (ID). Empezando por el CD, debemos observar que existen diferentes criterios para evaluar la bondad de los ítems, aunque nosotros utilizaremos la tabla ya clásica de Ebel (1965) como referencia (Tabla 5.17.):

Coeficiente de dificultad	Valoración del ítem
,86 - 1,00	Muy Fácil
,71 - ,85	Fácil
,40 - ,70	Satisfactorio
,15 - ,39	Difícil
,01 - ,14	Muy Difícil

Tabla 5.17.: Niveles de dificultad (Ebel, 1965)

Según estos criterios, aquellos ítems que tengan resultados por encima de 0,86 deben ser eliminados por ser demasiado fáciles, así como los ítems cuyos valores sean inferiores a 0,15, en este caso por ser demasiado difíciles. Por tanto, los valores deseables deben estar comprendidos entre 0,15 y 0,85. Teniendo esta referencia presente, revisaremos a continuación los resultados del CD analizando los ítems en sus respectivas secciones.

Sección A:

La primera sección del test estaba constituida por 50 ítems que pretendían medir las colocaciones en contextos reales, para lo cual, como ya mencionamos, se extrajeron ejemplos del corpus ukWaC mediante el programa *Sketch Engine*. Recordemos que el hecho de que el método utilizado fuera semejante al *c-test* se debió a que se pretendía obtener ítems totalmente objetivos y la inclusión de la primera letra de la palabra que buscábamos facilitaba que los ítems fueran dicotómicos (creemos que este particular fue suficientemente debatido en el capítulo anterior). El ejemplo siguiente nos recuerda la extensión del contexto y cómo estaba formulada la pregunta:

TASK 1

Fill in the blanks in each sentence by adding only one word. The word you need to add can only be an adjective, a noun or a verb. The first letter of each word is provided to help you.

Example:

0. **N**_____ **light** is preferable to artificial light.

Answer:

0. **N**atural **light** is preferable to artificial light.

Vamos ahora a mostrar la tabla del CD de la Sección A subrayando aquellos ítems que estarían en las categorías extremas y que, por tanto, siempre según Ebel (1965), deberían descartarse. Hay que reseñar que la tabla aportada por SPSS, muestra resultados globales, es decir, con relación a los 200 ítems del test y no a los 50 de cada sección.

Ítems Sección A	Media	Desviación típica
a1	,41	,493
a2	**,94**	,235
a3	,61	,488
a4	,85	,360
a5	,68	,469
a6	**,92**	,274
a7	,44	,497
a8	,60	,490
a9	,57	,496
a10	,58	,494
a11	,52	,500
a12	,79	,409
a13	,67	,471
a14	**,93**	,252

a15	,48	,501
a16	,66	,476
a17	,60	,490
a18	,66	,476
a19	**,86**	,347
a20	,64	,480
a21	,71	,454
a22	,64	,481
a23	,63	,484
a24	,81	,392
a25	,85	,354
a26	,66	,474
a27	,84	,369
a28	**,87**	,337
a29	,49	,501
a30	,56	,498
a31	**,92**	,268
a32	,70	,460
a33	**,90**	,297
a34	,77	,422
a35	,46	,500
a36	,57	,496
a37	,66	,475
a38	,48	,500
a39	,74	,439
a40	,60	,491
a41	,65	,477
a42	,61	,489
a43	,34	,476
a44	**,98**	,138
a45	,35	,479
a46	,77	,420
a47	,60	,491
a48	**,94**	,235
a49	,69	,461
a50	,46	,500

Tabla 5.18.: Coeficiente de dificultad por ítem (Sección A)

Observando los resultados de nuestros ítems nos encontramos con la siguiente información (Tabla 5.19.):

Muy Fácil ,86 -1,00	a2, a6, a14, a19, a28, a31, a33, a44, a48 = 9 (18%)
Fácil ,71 - ,85	a4, a12, a21, a24, a25, a27, a34, a39, a46 = 9 (18%)
Satisfactorio ,40 - ,70	a1, a3, a5, a7, a8, a9, a10, a11, a13, a15, a16, a17, a18, a20, a22, a23, a26, a29, a30, a32, a35, a36, a37, a38, a40, a41, a42, a47, a49, a50 = 30 (60%)
Difícil ,15 - ,39	a43, a45 =2 (4%)
Muy Difícil ,01 - ,14	

Tabla 5.19.: Distribución de ítems Sección A según su dificultad

Como se puede observar en la tabla 5.19., el 60% de los ítems entra en la categoría de Satisfactorios, un 18% está entre los Fáciles y un 4% entre los Difíciles. Por tanto, se deben eliminar los 9 de la categoría de Muy Fáciles que suman un 18%. No existe, como muestra la tabla, ningún ítem que se deba considerar Muy Difícil.

A título ilustrativo, mostraremos a continuación ejemplos de las 4 categorías:

Muy Fácil

a2. Use a map to help you **f_____ your way** along the trail.

Fácil

a4. Everything I read and studied for many years of very **h_____ work** explicitly rejected God as an explanation of anything.

Satisfactorio

a1. Applications are particularly welcome from women, people from ethnic minority backgrounds and **d_____ people**.

Difícil

a43. Why are violence and **d_____ behaviour** in schools a growing problem in Britain?

Por consiguiente, en esta Sección A, según la tabla 5.19., rechazaríamos **9 ítems, de la categoría Muy Fácil.**

SECCIÓN B:

Empecemos de nuevo por la tabla aportada por SPSS.

Ítems Sección B	Media	Desviación típica
b1	,96	,186
b2	,83	,372
b3	,78	,413
b4	,76	,426
b5	,77	,424
b6	,67	,470
b7	,77	,424
b8	,51	,501
b9	,78	,418
b10	,45	,498
b11	,69	,465
b12	,57	,495
b13	,69	,464
b14	,76	,430
b15	,64	,480

b16	,78	,415
b17	,31	,463
b18	,46	,499
b19	,78	,415
b20	,63	,483
b21	,73	,444
b22	,83	,378
b23	,56	,497
b24	,62	,485
b25	**,92**	,274
b26	,47	,500
b27	,38	,487
b28	,34	,474
b29	,59	,492
b30	**,93**	,252
b31	,73	,444
b32	,83	,378
b33	,67	,471
b34	,73	,443
b35	,66	,474
b36	,43	,496
b37	,39	,488
b38	,48	,500
b39	,48	,500
b40	**,91**	,288
b41	,75	,434
b42	**,86**	,347
b43	,79	,409
b44	,65	,479
b45	,38	,487
b46	,51	,501
b47	,53	,500
b48	,64	,481
b49	,82	,381
b50	,75	,436

Tabla 5.20.: Coeficiente de dificultad por ítem (Sección B)

Muy Fácil ,86 -1,00	b1, b25, b30, b40, b42 = 5 (10%)
Fácil ,71 - ,85	b2, b3, b4, b5, b7, b9, b14, b16, b19, b21, b22, b31, b32, b34, b41, b43, b49, b50 = 18 (36%)
Satisfactorio ,40 - ,70	b6, b8, b10, b11 b12, b13, b15, b18, b20, b23, b24, b26, b29, b33, b35, b36, b38, b39, b44, b46, b47, b48 = 22 (44%)
Difícil ,15 - ,39	b17, b27, b28, b37, b45 = 5 (10%)
Muy Difícil ,01 - ,14	

Tabla 5.21.: Distribución de ítems Sección B según su dificultad

Esta sección esta constituida por ítems de traducción, cuyo método ha sido ampliamente debatido en secciones anteriores. No obstante, creemos que traer un ejemplo, a título recordatorio, puede facilitar la interpretación de los resultados.

TASK 2

Translate the following collocations into English. Add <u>either one single word or one hyphenated word</u> in each case.

Example:

0. Prestar atención: To <u>pay</u> attention

Como en la tabla anterior, los resultados nos parecen muy alentadores, tratándose del diseño de unos ítems de traducción ─actividad muy denostada en estos últimos años por la influencia de la metodología comunicativa, lo cual ha generado muy poca o nula investigación sobre su efectividad─. Opinamos que los resultados de esta Sección B muestran que ha sido elaborada de una manera bastante correcta con: sólo

5 ítems en la categoría de Muy Fáciles, consiguiendo un 44% de Satisfactorios, un 36% de Fáciles y 10% en la categoría de Difícil. En los siguientes ejemplos se puede ver una muestra de 4 ítems según los niveles de dificultad que obtuvieron:

Muy Fácil

b1. Empezar el día: To _____ the day

Fácil

b2. La vida cotidiana: _____ life

Satisfactorio

b6. El mundo en vías de desarrollo: The _____ world

Difícil

b17. Desarrollo sostenible: _____ development

Por tanto, en esta sección se rechazaron **5 ítems de la categoría Muy Fácil.**

SECCIÓN C:

Esta es la siguiente tabla en donde, a primera vista, ya podemos observar que al contrario de las dos anteriores, la dificultad de los ítems es la raíz del problema que debemos afrontar.

Ítems Sección C	Media	Desviación típica
c1	,29	,454
c2	,76	,426
c3	,45	,499
c4	,55	,499
c5	,47	,500
c6	,31	,463
c7	,55	,499
c8	,75	,434
c9	,21	,409
c10	,37	,483
c11	,57	,496
c12	,15	,354
c13	,46	,500
c14	,24	,428
c15	,39	,489
c16	,52	,500
c17	,41	,492
c18	,27	,444
c19	,52	,500
c20	,34	,475
c21	,52	,500
c22	,23	,422
c23	,58	,495
c24	,44	,497
c25	,40	,491
c26	,50	,501
c27	,18	,386
c28	,44	,497
c29	,58	,494
c30	,18	,381
c31	,12	,326
c32	,25	,432
c33	,28	,449
c34	,46	,499

c35	,34	,476
c36	,27	,443
c37	,52	,501
c38	,55	,499
c39	,18	,386
c40	,33	,470
c41	,56	,497
c42	,22	,413
c43	**,11**	,318
c44	**,13**	,333
c45	,22	,415
c46	,31	,464
c47	,32	,467
c48	,50	,501
c49	,36	,481
c50	,20	,399

Tabla 5.22.: Coeficiente de dificultad por ítem (Sección C)

En esta Sección C, se puede apreciar la dificultad de una gran cantidad que ítems, hecho ya recogido cuando analizamos el test por secciones e incluso al estudiar las correlaciones entre dichas secciones. Esperamos que este nuevo análisis nos ofrezca datos complementarios que confirmen o refuten la posible inadecuación del diseño de esta sección. Como ya dijimos, el método empleado en la tarea C de nuestro test es el *odd-one-out,* ejercicio muy frecuente e incluso aconsejado en la metodología de las colocaciones, y que resulta un referente clásico en los libros de texto dedicados a desarrollar el aspecto colocacional del vocabulario. El ejemplo de abajo ilustra esta tarea y a continuación se ofrecen sus resultados.

TASK 3

Choose the word which <u>does not</u> collocate with the noun given (there is <u>only one wrong collocation</u> in each case). If all of them are correct, tick the option "no wrong collocation".

Example:

0. To ___ a list
□ compile ☒ compose □ make □ no wrong collocation

Muy Fácil ,86 -1,00	
Fácil ,71 - ,85	c2, c8 = 2 (4%)
Satisfactorio ,40 - ,70	c3, c4, c5, c7, c11, c13, c16, c17, c19, c21, c23, c24, c25, c26, c28, c29, c34, c37, c38, c41, c48 = 21 (42%)
Difícil ,15 - ,39	c1, c6, c9, c10, c12, c14, c15, c18, c20, c22, c27, c30, c32, c33, c35, c36, c39, c40, c42, c45, c46, c47, c49, c50 = 24 (48%)
Muy Difícil ,01 - ,14	c31, c43, c44 = 3 (6%)

Tabla 5.23.: Distribución de ítems Sección C según su dificultad

Al observar la tabla anterior, la dificultad de esta sección salta a la vista. La causa más probable, como ya se comentó en los análisis previos, es el uso del método *odd-one-out*, que si bien tiene la ventaja a nivel teórico de comprobar un espectro de varias colocaciones (al menos 2), la confusión que produce al ofrecérsele al alumno varias alternativas muy semejantes incrementa sobremanera la dificultad del ítem. Así, se contabiliza un 42% de ítems Satisfactorios por un 48% de ítems Difíciles y un 6% de la categoría Muy Difícil. No obstante, con estos datos de dificultad no podemos

todavía considerar este método como "perverso" y tendremos que esperar al índice de discriminación para obtener datos más concluyentes. Ejemplos de esta actividad con los índices anteriores son:

Fácil

c2. ___ place

□ safe □ secure □ sure □ no wrong collocation

Satisfactorio

c3. ___ quality

□ big □ high □ superb □ no wrong collocation

Difícil

c1. To___ your hand

□ hold □ rise □ shake □ no wrong collocation

Muy Difícil

c31. ___ future

□ foreseeable □ imprecise □ long-term □ no wrong collocation

SECCIÓN D:

Como recordaremos, el método empleado en esta sección es el de elección múltiple, sin duda el más ampliamente utilizado tanto para enseñar como para evaluar las colocaciones. En nuestro caso cada ítem consta de 3 opciones y un distractor adicional, "*none of these*", que pretende reducir el factor azar.

TASK 4

Choose the correct collocation and tick the appropriate box. There is only one correct collocation in each case. If none of the 3 first options is correct, tick the option "none of these".

Example:

0. A place devoted to entertainment is called **a/an ____ area**.
□ break □ game ☒ play □ none of these

Los resultados de esta última sección en cuanto al CD fueron los siguientes:

Items Sección D	Media	Desviación típica
d1	,30	,457
d2	,53	,500
d3	,54	,499
d4	,47	,500
d5	,16	,363
d6	,32	,468
d7	,76	,428
d8	,50	,501
d9	,84	,366
d10	,22	,418
d11	,42	,494
d12	,50	,501

d13	,73	,446
d14	,71	,456
d15	,84	,369
d16	,32	,468
d17	,21	,411
d18	,61	,489
d19	,77	,422
d20	,45	,499
d21	,75	,436
d22	,61	,488
d23	,61	,488
d24	,61	,488
d25	,36	,480
d26	,51	,501
d27	,44	,497
d28	,78	,415
d29	,48	,500
d30	,19	,389
d31	,42	,494
d32	,20	,402
d33	,38	,485
d34	,39	,489
d35	,65	,478
d36	,37	,484
d37	,46	,499
d38	,81	,394
d39	,62	,486
d40	,45	,498
d41	,39	,488
d42	,48	,500
d43	,44	,498
d44	,24	,430
d45	,78	,418
d46	,43	,496
d47	**,14**	,344
d48	,64	,481
d49	,78	,415
d50	,47	,500

Tabla 5.24.: Coeficiente de dificultad por ítem (Sección D)

Muy Fácil ,86 -1,00	
Fácil ,71 - ,85	d7, d9, d13, d14, d15, d19, d21, d28, d38, d45, d49 = 11 (22 %)
Satisfactorio ,40 - ,70	d2, d3, d4, d8, d11, d12, d18, d20, d22, d23, d24, d26, d27, d29, d31, d35, d37, d39, d40, d42, d43, d46, d48, d50 = 24 (48 %)
Difícil ,15 - ,39	d1, d5, d6, d10, d16, d17, d25, d30, d32, d33, d34, d36, d41, d44 = 14 (28 %)
Muy Difícil ,01 - ,14	d47 = 1 (2%)

Tabla 5.25.: Distribución de ítems Sección D según su dificultad

Antes de comentar los resultados de este bloque, debemos recordar que una de las quejas que los alumnos apuntaron en los protocolos orales fue la longitud del test, algo que sin duda y como los propios estudiantes señalaban, ha afectado a la última sección por el cansancio acumulado. Sin embargo, creemos que los resultados no han sido realmente decepcionantes, ocupando los ítems de nivel Satisfactorio la mayor parte del espectro (48%), con además un número razonable de ítems Fáciles (22%) y otro número aceptable de ítems Difíciles (28%). Nos parece muy positivo asimismo que sólo un ítem aparezca como Muy Difícil, y ninguno como Muy Fácil. Ejemplos de lo anterior son los siguientes:

Fácil

d7. When you refer to the world and life in general, in contrast to a particular person's own life, experience and ideas, which may seem untypical, you talk about **the____ world**.

□ authentic □ real □ true □ none of these

Satisfactorio

d2. If you ____ **your way** somewhere, you walk or travel there.

□ begin □ make □ reach □ none of these

Difícil

d1. If you say **in ____ fact**, you indicate that you are giving more detailed information about what you have just said.

□ actual □ real □ true □ none of these

Muy Difícil

d47. When you quickly and eagerly do something when you have the chance to do it, you ____ **the opportunity**.

□ catch □ hold □ seize □ none of these

A la luz de los datos anteriores habría que descartar **1 ítem de la Categoría Muy Difícil.**

Sin embargo, para mejor estudiar los ítems de nuestro test, debemos analizar asimismo su ID, tarea ésta a la que nos dedicaremos a continuación.

5.3.2.2. El índice de discriminación (ID)

El índice de discriminación es un indicador que establece la correlación que existe entre el resultado obtenido en cada ítem particular y el del test en general, siendo su

finalidad principal la de discriminar entre los candidatos de alto y bajo nivel. Este coeficiente es, por tanto, de suma importancia para determinar la fiabilidad y la validez de una prueba. Podemos decir que un test discrimina bien, y por tanto será válido y fiable, cuando diferencia correctamente los distintos niveles de conocimiento de los alumnos. La tabla 5.26. nos muestra los criterios que hemos empleado para valorar el ID de los ítems:

Índices de discriminación	Valoración del ítem
,40 o superior	Muy bueno
,30 - ,39	Bueno
,19 - ,29	Pasable, aunque puede necesitar modificaciones
,18 o inferior	Malo, debe en general rechazarse o replantearse

Tabla 5.26.: Niveles de discriminación (Ebel, 1965)

A continuación procederemos al análisis del ID, analizando cada una de las 4 secciones, de la misma manera que llevamos a cabo anteriormente el estudio del CD. El parámetro que nos interesa en las tablas que siguen, ofrecidas por SPSS, es la **cuarta columna, en donde el programa presenta la "Correlación elemento-total",** es decir, la discriminación de cada ítem comparando sus resultados con los del total del test. Como ya recordamos al hablar del CD, la discriminación está calculada con relación a la globalidad del test (200 ítems) y no de cada sección (50 ítems).

SECCIÓN A:

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	**Correlación elemento-total corregida**	Alfa de Cronbach si se elimina el elemento
a1	110,13	1074,917	,539	,967
a2	109,60	1088,676	,248	,967
a3	109,94	1073,970	,574	,967
a4	109,70	1086,303	,259	,967
a5	109,87	1072,648	,641	,967
a6	109,63	1084,196	,461	,967
a7	110,11	1073,349	,583	,967
a8	109,94	1077,436	,463	,967
a9	109,97	1072,260	,618	,967
a10	109,96	1072,845	,603	,967
a11	110,02	1071,742	,628	,967
a12	109,76	1078,595	,515	,967
a13	109,88	1076,063	,527	,967
a14	109,61	1086,433	,365	,967
a15	110,06	1073,250	,582	,967
a16	109,89	1076,248	,516	,967
a17	109,94	1072,700	,612	,967
a18	109,89	1074,033	,587	,967
a19	109,69	1083,083	,410	,967
a20	109,90	1071,430	,666	,967
a21	109,83	1075,630	,562	,967
a22	109,91	1073,310	,604	,967
a23	109,92	1073,006	,609	,967
a24	109,73	1075,062	,676	,967
a25	109,69	1081,986	,450	,967
a26	109,88	1073,999	,591	,967
a27	109,71	1080,449	,494	,967
a28	109,68	1082,878	,432	,967
a29	110,05	1075,053	,526	,967
a30	109,99	1073,834	,567	,967
a31	109,62	1085,076	,420	,967
a32	109,85	1079,530	,425	,967
a33	109,64	1084,986	,383	,967
a34	109,78	1080,500	,429	,967
a35	110,08	1074,114	,556	,967
a36	109,97	1074,104	,561	,967

a37	109,89	1078,577	,442	,967
a38	110,07	1073,477	,575	,967
a39	109,81	1078,073	,496	,967
a40	109,95	1071,965	,633	,967
a41	109,89	1075,614	,535	,967
a42	109,94	1071,368	,655	,967
a43	110,20	1078,109	,456	,967
a44	109,56	1089,862	,295	,967
a45	110,19	1077,615	,468	,967
a46	109,77	1083,531	,321	,967
a47	109,94	1071,896	,636	,967
a48	109,60	1086,689	,376	,967
a49	109,85	1072,838	,646	,967
a50	110,08	1073,234	,583	,967

Tabla 5.27.: Índice de discriminación Sección A

Según las categorías de Ebel (1965), los resultados son los siguientes:

Muy buenos ,40 o superior	a1, a3, a5, a6, a7, a8, a9, 10, a11, a12, a13, a15, a16, a17, a18, a19, a20, a21, a22, a23, a24, a25, a26, a27, a28, a29, a30, a31, a32, a34, a35, a36, a37, a38, a39, a40, a41, a42, a43, a45, a47, a49, a50 = 43 (86%)
Buenos ,30 - ,39	a14, a33, a46, a48 = 4 (8%)
Pasables ,19 - ,29	a2, a4, a44 = 3 (6%)
Malos ,18 o inferior	

Tabla 5.28.: Distribución de ítems Sección A según su discriminación

Es evidente que la discriminación de esta sección ha sido excelente, con un 86% de ítems enmarcados en la categoría de Muy Buenos, un 8% en Buenos y sólo 3 serían Pasables o regulares. No habría que descartar ningún ítem según el ID. A continuación vamos a ofrecer algún ejemplo:

Muy Bueno

a6. The programme gives students the possibility to s_____ a problem step by step.

Bueno

a14. They lost their lives when their car c_____ into a tree.

Pasable

a44. In this book the m_____ character is Sir Charles Baskervilles.

SECCIÓN B:

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	**Correlación elemento-total corregida**	Alfa de Cronbach si se elimina el elemento
b1	109,58	1090,257	**,186**	,967
b2	109,71	1078,564	,567	,967
b3	109,76	1077,406	,553	,967
b4	109,78	1082,868	,340	,967
b5	109,78	1076,322	,578	,967
b6	109,87	1076,150	,525	,967
b7	109,78	1077,202	,546	,967
b8	110,04	1075,562	,510	,967
b9	109,77	1077,064	,559	,967
b10	110,10	1076,929	,471	,967

b11	109,86	1079,417	,423	,967
b12	109,97	1074,810	,540	,967
b13	109,86	1071,953	,672	,967
b14	109,79	1080,577	,418	,967
b15	109,90	1074,714	,560	,967
b16	109,77	1091,176	**,045**	,968
b17	110,24	1079,295	,430	,967
b18	110,08	1073,954	,561	,967
b19	109,77	1072,571	,729	,967
b20	109,91	1078,165	,447	,967
b21	109,81	1073,311	,655	,967
b22	109,72	1079,747	,511	,967
b23	109,98	1072,794	,600	,967
b24	109,92	1073,382	,596	,967
b25	109,63	1084,880	,423	,967
b26	110,08	1074,385	,548	,967
b27	110,16	1077,661	,459	,967
b28	110,21	1076,120	,522	,967
b29	109,95	1077,095	,472	,967
b30	109,61	1085,723	,408	,967
b31	109,81	1072,887	,669	,967
b32	109,72	1088,796	**,146**	,968
b33	109,88	1075,946	,531	,967
b34	109,81	1074,948	,600	,967
b35	109,88	1072,273	,647	,967
b36	110,12	1074,918	,536	,967
b37	110,16	1075,989	,510	,967
b38	110,06	1074,745	,536	,967
b39	110,07	1072,578	,603	,967
b40	109,64	1085,678	,359	,967
b41	109,80	1077,759	,514	,967
b42	109,69	1085,070	,323	,967
b43	109,76	1079,077	,496	,967
b44	109,90	1073,120	,613	,967
b45	110,16	1074,957	,544	,967
b46	110,04	1075,129	,524	,967
b47	110,02	1073,276	,581	,967
b48	109,91	1072,776	,621	,967
b49	109,72	1080,847	,463	,967
b50	109,80	1078,663	,480	,967

Tabla 5.29.: Índice de discriminación Sección B

Los resultados de los ítems según las categorías del ID han sido:

Muy bueno ,40 o superior	b2, b3, b5, b6, b7, b8, b9, b10, b11, b12, b13, b14, b15, b17, b18, b19, b20, b21, b22, b23, b24, b25, b26, b27, b28, b29, b30, b31, b33, b34, b35, b36, b37, b38, b39, b41, b43, b44, b45, b46, b47, b48, b49, b50 = 44 (88%)
Bueno ,30 - ,39	b4, b40, b42, = 3 (6%)
Pasable ,19 - ,29	
Malo ,18 o inferior	b1, b16, b32 = 3 (6%)

Tabla 5.30.: Distribución de ítems Sección B según su discriminación

Como ya comentamos al analizar el índice de dificultad, nos resultaba difícil prever el comportamiento de los ítems en donde se pedía a los alumnos que tradujeran el colocado de una base que se les presentaba en español. Reiteramos que los resultados demuestran que la traducción puede ser un método muy indicado para la evaluación de la competencia colocacional porque la discriminación ha sido francamente positiva con un 88% de ítems Muy Buenos (un buen número de los cuales muestran una discriminación superior a 0,60, valor éste verdaderamente alto), un 6% Buenos y tan solo 3 ítems aparecen por debajo de la barrera del 0,19 y, por tanto, no discriminan entre los alumnos con mejores y peores actuaciones. Ejemplos:

Muy Buenos

b19. Hacer un examen: _____ a test

Buenos

b4. Trabajo de voluntariado: _____ work

Malos

b16. El lugar equivocado: The _____ place

SECCIÓN C:

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	**Correlación elemento-total corregida**	Alfa de Cronbach si se elimina el elemento
c1	110,26	1097,559	**-,173**	,968
c2	109,78	1084,711	,274	,967
c3	110,09	1080,969	,346	,967
c4	110,00	1086,430	**,179**	,968
c5	110,07	1080,119	,371	,967
c6	110,24	1086,677	**,186**	,968
c7	110,00	1087,824	**,137**	,968
c8	109,80	1089,883	**,087**	,968
c9	110,33	1088,738	**,136**	,968
c10	110,18	1097,665	**-,166**	,968
c11	109,97	1089,700	**,080**	,968
c12	110,40	1090,605	**,079**	,968
c13	110,08	1081,612	,326	,967
c14	110,31	1091,561	**,029**	,968
c15	110,15	1079,628	,395	,967
c16	110,02	1084,706	,231	,968
c17	110,14	1083,805	,263	,967
c18	110,28	1085,529	,234	,967
c19	110,02	1093,104	**-,024**	,968
c20	110,20	1096,945	**-,146**	,968

c21	110,02	1078,895	,409	,967
c22	110,31	1086,783	,202	,968
c23	109,97	1083,484	,272	,967
c24	110,11	1085,043	,223	,968
c25	110,14	1088,677	**,113**	,968
c26	110,05	1085,737	,200	,968
c27	110,36	1097,789	**-,210**	,968
c28	110,11	1083,343	,275	,967
c29	109,96	1081,842	,323	,967
c30	110,37	1087,367	,202	,967
c31	110,43	1088,571	**,181**	,968
c32	110,30	1088,686	**,130**	,968
c33	110,27	1085,186	,243	,967
c34	110,08	1085,107	,219	,968
c35	110,20	1080,552	,377	,967
c36	110,28	1083,661	,299	,967
c37	110,03	1095,794	**-,105**	,968
c38	110,00	1086,274	**,184**	,968
c39	110,36	1084,415	,315	,967
c40	110,22	1095,128	**-,089**	,968
c41	109,98	1084,325	,245	,967
c42	110,33	1094,019	**-,059**	,968
c43	110,43	1095,360	**-,137**	,968
c44	110,42	1092,075	**,018**	,968
c45	110,32	1091,979	**,015**	,968
c46	110,23	1086,668	**,186**	,968
c47	110,23	1082,684	,315	,967
c48	110,05	1083,939	,254	,967
c49	110,19	1089,213	**,099**	,968
c50	110,35	1088,996	**,130**	,968

Tabla 5.31.: Índice de discriminación Sección C

Los resultados por categorías han sido los siguientes:

Muy bueno ,40 o superior	c21 = 1 (2%)
Bueno ,30 - ,39	c3, c5, c13, c15, c29, c35, c39, c47 = 8 (16%)
Pasable ,19 - ,29	c2, c16, c17, c18, c22, c23, c24, c26, c28, c30, c33, c34, c36, c41, c48 = 15 (30%)
Malo ,18 o inferior	c1, c4, c6, c7, c8, c9, c10, c11, c12, c14, c19, c20, c25, c27, c31, c32, c37, c38, c40, c42, c43, c44, c45, c46, c49, c50 = 26 (52%)

Tabla 5.32.: Distribución de ítems Sección C según su discriminación

Como hemos venido apuntado, esta Sección C se presenta como ciertamente problemática. En primer lugar, vimos que ofrece un resultado bajo en cuanto a la media de respuestas correctas obtenidas (37,72%). También comprobamos que muestra una correlación inferior con el resto de secciones que entre la que éstas establecen entre sí. En tercer lugar, acabamos también de observar mediante el CD que la mayoría de los ítems de esta sección pertenece a la categoría de Difíciles. Sin embargo, no podíamos extraer conclusiones definitivas hasta conocer el ID de los ítems. Por desgracia, éste ha confirmado los peores augurios. Más de la mitad de los ítems (52%) no discriminan, mientras que un 30% están entre los Pasables y sólo un 16% se pueden considerar Buenos. Esto nos hace pensar que el método "*odd-one-out*" no ha funcionado como debería. Las razones más probables son: a) El hecho de que se evaluaba más de un elemento (en realidad hasta tres en algunos casos). b) Las colocaciones son tan similares y tan plausibles que se necesita un conocimiento casi de nativo para poder discernir la correcta de entre las distintas opciones. En este sentido, cabe señalar que los propios alumnos apuntaban que incluso ante algunas

que no parecían difíciles o que creían conocer, el encontrarlas entre otras opciones muy parecidas en esta sección les confundía. c) La total ausencia de contexto también puede haber incrementado la confusión haciendo más difícil la identificación de la colocación incorrecta.

Con todos los elementos de juicio que hemos acumulado, no nos cabe ninguna duda de que este tipo de formato no es apropiado para evaluar el conocimiento de los candidatos en materia colocacional y, por consiguiente, esta sección del test debe eliminarse en su totalidad. Los siguientes ejemplos pueden ilustrar estas consideraciones.

Muy Buenos

c21. ___ **moment**

□ adequate □ precise □ right □ no wrong collocation

Buenos

c13. ___ **matter**

□ serious □ severe □ simple □ no wrong collocation

Pasables

c18. ___ **letter**

□ business □ reception □ resignation □ no wrong collocation

Malos

c20. **To ___ evidence**

□ give □ grant □ provide □ no wrong collocation

SECCIÓN D:

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	**Correlación elemento-total corregida**	Alfa de Cronbach si se elimina el elemento
d1	110,25	1087,374	**,166**	,968
d2	110,02	1080,472	,361	,967
d3	110,00	1085,098	,220	,968
d4	110,07	1082,779	,290	,967
d5	110,39	1087,385	,211	,967
d6	110,22	1086,129	,202	,968
d7	109,79	1084,566	,278	,967
d8	110,05	1081,747	,321	,967
d9	109,70	1082,919	,395	,967
d10	110,32	1081,646	,392	,967
d11	110,13	1089,188	**,096**	,968
d12	110,05	1084,248	,245	,967
d13	109,82	1082,931	,322	,967
d14	109,84	1094,110	**-,058**	,968
d15	109,71	1089,667	**,114**	,968
d16	110,22	1082,051	,335	,967
d17	110,33	1091,056	**,050**	,968
d18	109,94	1085,133	,223	,968
d19	109,78	1088,950	**,124**	,968
d20	110,09	1078,317	,428	,967
d21	109,80	1086,396	,209	,968
d22	109,93	1080,422	,372	,967
d23	109,93	1082,533	,305	,967
d24	109,93	1076,051	,509	,967
d25	110,19	1083,990	,264	,967
d26	110,04	1078,129	,432	,967
d27	110,11	1093,011	**-,021**	,968
d28	109,77	1083,150	,338	,967

d29	110,07	1086,715	**,170**	,968
d30	110,36	1088,759	**,143**	,968
d31	110,13	1083,827	,262	,967
d32	110,34	1085,901	,246	,967
d33	110,17	1079,268	,410	,967
d34	110,15	1082,508	,305	,967
d35	109,90	1084,817	,239	,967
d36	110,18	1083,676	,272	,967
d37	110,09	1084,966	,224	,968
d38	109,74	1091,243	**,045**	,968
d39	109,93	1088,447	**,121**	,968
d40	110,09	1079,734	,384	,967
d41	110,16	1084,036	,258	,967
d42	110,06	1086,843	**,166**	,968
d43	110,10	1085,531	,207	,968
d44	110,30	1083,436	,316	,967
d45	109,77	1085,285	,259	,967
d46	110,12	1082,403	,304	,967
d47	110,41	1085,793	,294	,967
d48	109,91	1086,691	**,178**	,968
d49	109,77	1089,795	**,095**	,968
d50	110,07	1085,995	,192	,968

Tabla 5.33.: Índice de discriminación Sección D

Los resultados han sido:

Muy bueno ,40 o superior	d20, d24, d26, d33 = 4 (8%)
Bueno ,30 - ,39	d2, d8, d9, d10, d13, d16, d22, d23, d28, d34, d40, d44, d46 = 13 (26%)
Pasable ,19 - ,29	d3, d4, d5, d6, d7, d12, d18, d21, d25, d31, d32, d35, d36, d37, d41, d43, d45, d47, d50 = 19 (38%)
Malo ,18 o inferior	d1, d11, d14, d15, d17, d19, d27, d29, d30, d38, d39, d42, d48, d49 = 14 (28%)

Tabla 5.34.: Distribución de ítems Sección D según su discriminación

Por último, la opción D presenta también una problemática muy particular en cuanto a la discriminación, ya que si bien las tres cuartas partes del test son aprovechables (puesto que el 72% de los ítems discriminan en mayor o menor medida), existe un 38% de ítems que se sitúan en la categoría de Pasable, indicando pues que son mejorables. Es también de destacar que el 28% de los ítems debía ser eliminado o replanteado. Estos datos, en nuestra opinión, son peores de lo que en sí se podría esperar teniendo en cuenta la naturaleza de la actividad propuesta (elección múltiple), pueden deberse a la razón apuntada anteriormente: al final la prueba resultó muy larga y el efecto cansancio se hizo notar. Algunos ejemplos de las diferentes categorías son los siguientes:

Muy Buenos

d20. If you are not wearing anything on your feet you have ____ **feet**.

□ bare □ nude □ undressed □ none of these

Buenos

d8. The most important part of a meal is called **the ____ course**.

□ full □ principal □ strong □ none of these

Pasables

d4. A child who has no brothers or sisters is called **a/an ____ child**.

□ alone □ only □ unique □ none of these

Malos

d11. A family group which includes relatives such as uncles, aunts, and grandparents as well as parents, children, brothers and sisters is referred to as ____ **family**.

□ broad □ extended □ inclusive □ none of these

5.3.2.3. Análisis de distractores

Además del análisis de los ítems en términos del CD y el ID, la información sobre el comportamiento de las opciones en los ítems de elección múltiple aporta siempre nociones interesantes ya que lo que buscamos es conocer estadísticamente el funcionamiento diferencial del ítem. Aunque es todavía muy poco frecuente en los estudios de evaluación encontrar esta información, nos ha parecido que podría arrojar luz no sólo al diseño de un test de colocaciones, un campo en el que este análisis se puede considerar, hasta donde sabemos, pionero, sino también al tratamiento pedagógico de las colocaciones, un tema también hasta el momento muy poco estudiado de forma empírica.

A continuación, y basándonos en la información aportada por la Dra. Rita Green (2008, curso de estadística aplicada a la evaluación con SPSS, Universidad de Lancaster), reconocida especialista en el campo de la evaluación lingüística, vamos a proceder a ofrecer algunos parámetros estadísticos que las alternativas o distractores de los ítems deben cumplir. La información proporcionada por estos parámetros, basada en el número de alumnos que han seleccionado la opción correcta y las alternativas incorrectas o distractores, es sin duda de importancia capital para la calidad del ítem. La tabla 5.35. recoge los criterios en que nos hemos basado para analizar el funcionamiento diferencial de los ítems.

1. La opción correcta de un ítem debe ser la más elegida, y debe estar entre el 30% y 70%, siendo el 50% absolutamente ideal.
2. Los distractores deben atraer al menos un 7% de las respuestas, siendo un 10% el deseado.
3. Cuando el ítem tiene un buen índice de discriminación (> ,30), no se deben modificar los distractores aunque no se cumplan los parámetros anteriores.
4. No debe haber más de un 10% de respuestas en blanco en un ítem, es decir, los candidatos que no respondieron al ítem no deben sobrepasar el 10% del grupo.

Tabla 5.35.: Criterios para la evaluación de los ítems y sus distractores (Green, 2008)

Dado que la Sección C va a ser eliminada por los datos negativos que se han ido acumulando, el único método de elección múltiple que nos queda en el test es la Sección D, sobre la cual llevaremos a cabo un detallado análisis psicométrico de cada una de las alternativas de elección de los ítems. En este sentido, nos parece además muy pertinente llevar a cabo este análisis tan profundo precisamente en esta sección, a la vista de que, junto a la C, fue la que peores resultados obtuvo. Pensamos, pues, que la revisión y replanteamiento de los distractores en esta última parte del test puede suponer una importante mejora en su fiabilidad y validez.

A título informativo y para mejor comprender las tablas en que se ofrece la información sobre cada ítem, nos gustaría mencionar tres cuestiones: 1) Los ítems que vayan acompañados de una E* se refieren a los previamente eliminados por su inadecuado índice de dificultad (CD por debajo de ,15 o superior a ,86) o por el bajo índice de discriminación (inferior a ,19), por lo que no es necesario realizar

modificación alguna. 2) Los que presenten una A* se conservarán por las razones contrarias: poseen un índice de discriminación alto (superior a ,30) y por tanto no se deben alterar. 3) La "x" en la información del ítem indica el número de alumnos que no lo contestaron.

Item D1 ID: ,166 E*

	Recuento	% tabla
a = key	92	29,6%
b	85	27,3%
c	36	11,6%
d	73	23,5%
x	25	8,0%

Item D2 ID: ,361 A*

	Recuento	% tabla
a	48	15,4%
b = key	165	53,1%
c	51	16,4%
d	27	8,7%
x	20	6,4%

Item D3 ID: ,220

	Recuento	% tabla
a	5	1,6%
b	115	37,0%
c = key	169	54,3%
d	14	4,5%
x	8	2,6%

Item D4 ID: ,290

	Recuento	% tabla
a	6	1,9%
b = key	145	46,6%
c	94	30,2%
d	60	19,3%
x	6	1,9%

Item D5 ID: ,211

	Recuento	% tabla
a	147	47,3%
b = key	48	15,4%
c	26	8,4%
d	71	22,8%
x	19	6,1%

Item D6 ID: ,202

	Recuento	% tabla
a	107	34,4%
b = key	100	32,2%
c	12	3,9%
d	67	21,5%
x	25	8,0%

Item D7 ID: ,278

	Recuento	% tabla
a	21	6,8%
b = key	237	76,2%
c	18	5,8%
d	19	6,1%
x	16	5,1%

Item D8 ID: ,321 A*

	Recuento	% tabla
a	20	6,4%
b	105	33,8%
c	25	8,0%
d = key	153	49,2%
x	8	2,6%

Item D9 ID: ,395 A*

	Recuento	% tabla
a	13	4,2%
b	15	4,8%
c = key	260	83,6%
d	17	5,5%
x	6	1,9%

Item D10 ID:, ,392 A*

	Recuento	% tabla
a	41	13,2%
b	179	57,6%
c = key	69	22,2%
d	17	5,5%
x	5	1,6%

Item D11 ID: ,096 E*

	Recuento	% tabla
a	72	23,2%
b = key	130	41,8%
c	10	3,2%
d	78	25,1%
x	21	6,8%

Item D12 ID: ,245

	Recuento	% tabla
a	79	25,4%
b = key	153	49,2%
c	29	9,3%
d	36	11,6%
x	14	4,5%

Item D13 ID: ,322 A*

	Recuento	% tabla
a	26	8,4%
b = key	225	72,3%
c	22	7,1%
d	18	5,8%
x	20	6,4%

Item D14 ID: -,058 E*

	Recuento	% tabla
a	14	4,5%
b	63	20,3%
c = key	221	71,1%
d	6	1,9%
x	7	2,3%

Item D15 ID: ,114 E*

	Recuento	% tabla
a	5	1,6%
b	16	5,1%
c = key	260	83,6%
d	25	8,0%
x	5	1,6%

Item D16 ID: ,335 A*

	Recuento	% tabla
a	38	12,2%
b	67	21,5%
c = key	100	32,2%
d	90	28,9%
x	16	5,1%

Item D17 ID: ,050 E*

	Recuento	% tabla
a = key	67	21,5%
b	58	18,6%
c	133	42,8%
d	47	15,1%
x	6	1,9%

Item D18 ID: ,223

	Recuento	% tabla
a = key	189	60,8%
b	27	8,7%
c	49	15,8%
d	30	9,6%
x	16	5,1%

Item D19 ID: ,124 E*

	Recuento	% tabla
a	16	5,1%
b = key	191	61,4%
c	103	33,1%
d	1	,3%

Item D20 ID: ,428 A*

	Recuento	% tabla
a = key	141	45,3%
b	82	26,4%
c	55	17,7%
d	24	7,7%
x	9	2,9%

Item D21 ID: ,209

	Recuento	% tabla
a = key	231	74,3%
b	23	7,4%
c	15	4,8%
d	34	10,9%
x	8	2,6%

Item D22 ID: ,372 A*

	Recuento	% tabla
a = key	190	61,1%
b	20	6,4%
c	32	10,3%
d	50	16,1%
x	19	6,1%

Item D23 ID: ,305 A*

	Recuento	% tabla
a = key	191	61,4%
b	10	3,2%
c	79	25,4%
d	27	8,7%
x	4	1,3%

Item D24 ID: ,509 A*

	Recuento	% tabla
a	16	5,1%
b = key	191	61,4%
c	103	33,1%
d	1	,3%

Item D25 ID: , 264

	Recuento	% tabla
a	136	43,7%
b	34	10,9%
c	19	6,1%
d = key	110	35,4%
x	12	3,9%

Item D26 ID: ,432 A*

	Recuento	% tabla
a = key	156	50,2%
b	61	19,6%
c	64	20,6%
d	19	6,1%
x	11	3,5%

Item D27 ID: -,021 E*

	Recuento	% tabla
a	12	3,9%
b = key	137	44,1%
c	127	40,8%
d	28	9,0%
x	7	2,3%

Item D28 ID: ,338 A*

	Recuento	% tabla
a	17	5,5%
b = key	242	77,8%
c	12	3,9%
d	27	8,7%
x	13	4,2%

Item D29 ID: ,170 E*

	Recuento	% tabla
a	38	12,2%
b	58	18,6%
c = key	147	47,3%
d	48	15,4%
x	20	6,4%

Item D30 ID: ,143 E*

	Recuento	% tabla
a = key	57	18,3%
b	83	26,7%
c	69	22,2%
d	79	25,4%
x	23	7,4%

Item D31 ID: ,262

	Recuento	% tabla
a = key	128	41,2%
b	73	23,5%
c	19	6,1%
d	71	22,8%
x	20	6,4%

Item D32 ID: ,246

	Recuento	% tabla
a	30	9,6%
b	164	52,7%
c	28	9,0%
d = key	63	20,3%
x	26	8,4%

Item D33 ID: ,410 A*

	Recuento	% tabla
a = key	116	37,3%
b	9	2,9%
c	130	41,8%
d	43	13,8%
x	13	4,2%

Item D34 ID: ,305 A*

	Recuento	% tabla
a	36	11,6%
b = key	121	38,9%
c	54	17,4%
d	64	20,6%
x	36	11,6%

Item D35 ID: ,239

	Recuento	% tabla
a = key	203	65,3%
b	53	17,0%
c	11	3,5%
d	22	7,1%
x	22	7,1%

Item D36 ID: ,272

	Recuento	% tabla
a	51	16,4%
b	72	23,2%
c	33	10,6%
d = key	116	37,3%
x	39	12,5%

Item D37 ID: ,224

	Recuento	% tabla
a = key	144	46,3%
b	53	17,0%
c	56	18,0%
d	48	15,4%
x	10	3,2%

Item D38 ID: ,045 E*

	Recuento	% tabla
a	29	9,3%
b = key	252	81,0%
c	2	,6%
d	24	7,7%
x	4	1,3%

Item D39 ID: ,121 E*

	Recuento	% tabla
a = key	194	62,4%
b	40	12,9%
c	38	12,2%
d	35	11,3%
x	4	1,3%

Item D40 ID: ,384 A*

	Recuento	% tabla
a = key	139	44,7%
b	53	17,0%
c	36	11,6%
d	60	19,3%
x	23	7,4%

Item D41 ID: ,258

	Recuento	% tabla
a = key	119	38,3%
b	52	16,7%
c	38	12,2%
d	74	23,8%
x	28	9,0%

Item D42 ID: ,166 E*

	Recuento	% tabla
a = key	149	47,9%
b	35	11,3%
c	99	31,8%
d	19	6,1%
x	9	2,9%

Item D43 ID: ,207

	Recuento	% tabla
a	26	8,4%
b	47	15,1%
c	76	24,4%
d = key	140	45,0%
x	22	7,1%

Item D44 ID: ,316 A*

	Recuento	% tabla
a	2	,6%
b = key	75	24,1%
c	171	55,0%
d	51	16,4%
x	12	3,9%

Item D45 ID: ,259

	Recuento	% tabla
a	27	8,7%
b = key	242	77,8%
c	10	3,2%
d	19	6,1%
x	13	4,2%

Item D46 ID: ,304 A*

	Recuento	% tabla
a = key	27	8,7%
b	41	13,2%
c	173	55,6%
d	45	14,5%
x	25	8,0%

Item D47 CD: ,14 E*

	Recuento	% tabla
a	186	59,8%
b	36	11,6%
c = key	42	13,5%
d	34	10,9%
x	13	4,2%

Item D48 ID: ,178 E*

	Recuento	% tabla
a	61	19,6%
b	14	4,5%
c = key	197	63,3%
d	22	7,1%
x	17	5,5%

Item D49 ID: ,095 E*

	Recuento	% tabla
a	32	10,3%
b = key	243	78,1%
c	13	4,2%
d	15	4,8%
x	8	2,6%

Item D50 ID: ,192

	Recuento	% tabla
a	24	7,7%
b = key	146	46,9%
c	51	16,4%
d	48	15,4%
x	42	13,5%

El análisis de las tablas anteriores nos indica los ítems que deben ser aceptados o rechazados según sus CD e ID, pero además nos ofrece un análisis pormenorizado de las opciones de la Sección D, con la finalidad de incrementar la fiabilidad y validez en versiones posteriores del test.

A continuación, recordaremos de manera sintetizada algunas de las cuestiones que hemos ido comentando a lo largo del análisis de los ítems de la Sección D:

1. Teniendo en cuenta el CD, sólo **1 ítem** (d47) debe ser rechazado por su alta dificultad (media: 0,14).
2. A la vista de la discriminación, escasa o incluso negativa, **14 ítems** deben también descartarse (d1, d11, d14, d15, d17, d19, d27, d29, d30, d38, d39, d42, d48, d49)
3. Los **4 ítems** cuya discriminación es muy buena (d20, d24, d26, d33) y los **13** con discriminación buena (d2, d8, d9, d10, d13, d16, d22, d23, d28, d34, d40, d44, d46) no necesitan un análisis de distractores puesto que funcionan adecuadamente y no se deben alterar.
4. Los ítems que están en valores regulares o pasables de discriminación, serán revisados en cuanto a la psicometría de sus opciones para una mejora posterior. Estos ítems son **18** (excluyendo el d47 que se descartó por el CD): d3, d4, d5, d6, d7, d12, d18, d21, d25, d31, d32, d35, d36, d37, d41, d43, d45, d50.

Por tanto, a continuación vamos a revisar uno por uno los ítems susceptibles de modificaciones futuras, a la vista de los resultados de sus opciones. Queremos destacar que las nuevas opciones que se ofrecen en los ítems que requieren de alguna modificación en sus distractores han sido seleccionadas con el mismo cuidado que se tuvo durante el proceso de elaboración de la primera versión. Todos los distractores nuevos han sido pues comprobados con la ayuda del BNC para asegurarnos de que no son opciones válidas en la lengua. (En la tabla de abajo, la "x" significa "sin respuesta" y no debe tener valores superiores al 10% mientras que la "d" es la opción "*none of these*" y su medida no va a tenerse en cuenta como las anteriores ya que su

finalidad principal fue minimizar el efecto azar y el número de respuestas que atrajo no responde, evidentemente, a la buena o mala elección del distractor).

Item d3 (ID: ,220)

A day on which people go to work is called **a/an ____ day**.			
□ job	□ labour	☒ working	□ none of these

	Recuento	% tabla
a	5	1,6%
b	115	37,0%
c = key	169	54,3%
d	14	4,5%
x	8	2,6%

Problema: La opción "a" no ha sido elegida, al menos, por un 7% de los candidatos.
Modificaciones: En lugar de "*job*", creemos que "*routine*" puede ser una mejor alternativa.

Item d4 (ID: ,290)

A child who has no brothers or sisters is called **a/an ____ child**.			
□ alone	☒ only	□ unique	□ none of these

	Recuento	% tabla
a	6	1,9%
b = key	145	46,6%
c	94	30,2%
d	60	19,3%
x	6	1,9%

Problema: La opción "a" no ha sido suficientemente atractiva y por tanto no fue seleccionada por al menos el 7% de los candidatos.

Modificaciones: En lugar de "*alone*", sugerimos que "*sole*" puede ser una mejor alternativa.

Item d5 (ID: ,211)

When you ____ **a report**, you make it publicly available.

□ announce ☒ issue □ proclaim □ none of these

	Recuento	% tabla
a	147	47,3%
b = key	48	15,4%
c	26	8,4%
d	71	22,8%
x	19	6,1%

Problema: El distractor "a" ha resultado ser demasiado atractivo, superando claramente al número de respuestas correctas.

Modificaciones: En lugar de "*announce*", sugerimos "*declare*".

Item d6 (ID: ,202)

A/an ____ process lasts for a long time. □ durable ☒ lengthy □ stretched □ none of these

	Recuento	% tabla
a	107	34,4%
b = key	100	32,2%
c	12	3,9%
d	67	21,5%
x	25	8,0%

Problema: El distractor "c" no ha sido elegido, al menos, por un 7% de los candidatos.

Modificaciones: En lugar de "*stretched*", sugerimos "*persistent*" como mejor alternativa.

Item d7 (ID: ,278)

When you refer to the world and life in general, in contrast to a particular person's own life, experience and ideas, which may seem untypical, you talk about **the____ world**. □ authentic ☒ real □ true □ none of these

	Recuento	% tabla
a	21	6,8%
b = key	237	76,2%
c	18	5,8%
d	19	6,1%
x	16	5,1%

Problema: La opción "c" no ha llegado al 7% de elección. Aunque el distractor "a" tampoco ha llegado a este punto de corte, consideramos que se encuentra lo suficientemente cerca como para considerarlo válido.

Modificaciones: En lugar de "*true*", sugerimos "*existing*" puede ser una mejor alternativa.

Item d12 (ID: ,245)

Nights during which you don't sleep are ____ **nights**.
□ awake ☒ sleepless □ wakeful □ none of these

	Recuento	% tabla
a	79	25,4%
b = key	153	49,2%
c	29	9,3%
d	36	11,6%
x	14	4,5%

Problema: En este ítem todos los elementos se encuentran en parámetros no sólo normales sino muy deseables.

Item d18 (ID: ,223)

Rich, powerful and fashionable people are called ____ **society**.
☒ high □ peak □ top □ none of these

	Recuento	% tabla
a = key	189	60,8%
b	27	8,7%
c	49	15,8%
d	30	9,6%
x	16	5,1%

Problema: En este ítem todos los elementos se encuentran en parámetros normales.

Modificaciones: Sin embargo, con el fin de aumentar el porcentaje de la opción "b" hasta el deseado 10%, podríamos sustituir el distractor "*peak*" por "*grand*" y examinar si verdaderamente mejora el resultado obtenido.

Item d21 (ID: ,209)

When you have good reasons to show that something is true or untrue, right or wrong, you have **a/an____ argument**.

☒ convincing □ sure □ true □ none of these

	Recuento	% tabla
a = key	231	74,3%
b	23	7,4%
c	15	4,8%
d	34	10,9%
x	8	2,6%

Problema: La opción "c" no ha sido elegida, al menos, por un 7% de los candidatos. También sería aconsejable alterar la opción "b" con el fin de lograr un 10% de atracción.

Modificaciones: En lugar de "*sure*" ofertaremos "*faithful*", y en lugar de "*true*", "*likely*".

Item d31 (ID: ,262)

If your **face** ____, it goes red because you are feeling a strong emotion such as embarrassment or anger.

☒ flushes □ heats □ illuminates □ none of these

	Recuento	% tabla
a = key	128	41,2%
b	73	23,5%
c	19	6,1%
d	71	22,8%
x	20	6,4%

Problema: La opción "c" no ha sido elegida, al menos, por un 7% de los candidatos.

Modificaciones: En lugar de "*illuminates*", ofertaremos "*burns*".

Item d32 (ID: ,246)

The ____ cost is an approximate judgement or calculation of a value or price.

□ guessed □ speculated □ vague ☒ none of these

	Recuento	% tabla
a	30	9,6%
b	164	52,7%
c	28	9,0%
d = key	63	20,3%
x	26	8,4%

Problema: El distractor "b" es demasiado atractivo, ya que obtiene un resultado muy superior al de la opción correcta.

Modificaciones: En lugar de "*speculated*" ofreceremos "*inexact*" como alternativa.

Item d35 (ID: ,239)

____ **age** is the period when you stop working, usually because of your age. ☒ retirement □ retreat □ withdrawal □ none of these

	Recuento	% tabla
a = key	203	65,3%
b	53	17,0%
c	11	3,5%
d	22	7,1%
x	22	7,1%

Problema: La opción "c" no alcanza el 7% de respuestas.

Modificaciones: En lugar de "*withdrawal*", brindaríamos "*rest*".

Item d36 (ID: ,272)

If you successfully agree on something with other people, you ____ **a decision**. □ accomplish □ achieve □ grasp ☒ none of these

	Recuento	% tabla
a	51	16,4%
b	72	23,2%
c	33	10,6%
d = key	116	37,3%
x	39	12,5%

Problema: En este ítem el número de personas que no lo contestan es ligeramente superior al aconsejado (x = 12,5%). Sin embargo, ante el buen funcionamiento de las otras opciones y aunque es un ítem con una dificultad superior a la media (porcentaje de aciertos: 37,3%), decidimos no realizar ninguna modificación.

Item d37 (ID: ,224)

A/an _____ **road** is a place full of vehicles and movement.

☒ busy □ heavy □ packed □ none of these

	Recuento	% tabla
a = key	144	46,3%
b	53	17,0%
c	56	18,0%
d	48	15,4%
x	10	3,2%

Problema: Todos los distractores son también correctos; no es necesario llevar a cabo ningún cambio.

Item d41 (ID: ,258)

If you get someone's respect because you deserve it, you ____ **his/her respect**.

☒ earn □ gather □ take □ none of these

	Recuento	% tabla
a = key	119	38,3%
b	52	16,7%
c	38	12,2%
d	74	23,8%
x	28	9,0%

Problema: Todas las opciones muestran resultados correctos; no es necesario llevar a cabo ningún cambio.

Item d43 (ID: ,207)

When something happens normally or is usually true, it happens as **a/an ____ rule**.

□ average □ extended □ standard ☒ none of these

	Recuento	% tabla
a	26	8,4%
b	47	15,1%
c	76	24,4%
d = key	140	45,0%
x	22	7,1%

Problema: En este caso todos los distractores funcionan correctamente, aunque sería positivo mejorar la opción "a" para que su elección llegara al 10%.

Modificaciones: En lugar de "*average*", se ofrecería "*regular*".

Item d45 (ID: ,259)

To ____ **treatment** is to say what medicine or treatment a sick person should have.

□ order ☒ prescribe □ stipulate □ none of these

	Recuento	% tabla
a	27	8,7%
b = key	242	77,8%
c	10	3,2%
d	19	6,1%
x	13	4,2%

Problema: Las opción "c" no presenta un funcionamiento correcto.

Modificaciones: En lugar de "*stipulate*", un distractor más creíble puede ser "*command*".

Item d50 (ID: ,197)

An effect badly damaging something is called **a/an ____ effect**.

□ appalling ☒ devastating □ overwhelming □ none of these

	Recuento	% tabla
a	24	7,7%
b = key	146	46,9%
c	51	16,4%
d	48	15,4%
x	42	13,5%

Problema: De nuevo, nos encontramos con un ítem donde el porcentaje de sujetos que lo dejaron sin contestar es demasiado elevado (13,5%). En este caso consideramos que presenta un comportamiento correcto de todos sus distractores, quizá con una opción "a" que tiende a ser poco atractiva. Consideramos, no obstante, que dado el elevado porcentaje de alumnos que dejaron esta pregunta en blanco no sería conveniente sustituir este distractor por otro más atractivo, dado que ello aumentaría aún más la dificultad del ítem. Por otro lado, también pensamos que el elevado número de candidatos que no contestaron esta pregunta puede deberse en parte a que se trata del último ítem del test, el número 200, y muy posiblemente el cansancio influyera en este caso.

5.4. Revisión del test y propuesta de una nueva versión: Test de Colocaciones ADELEX Versión 2 (TCA2)

Una vez realizados todos los análisis anteriores y a la vista de los fallos que hemos venido observando tras el pilotaje de TCA1, decidimos, por una parte, reducir sustancialmente el número de ítems y, por otra, corregir las opciones o alternativas de los ítems que mostraban alguna irregularidad psicométrica, es decir, no se atenían a criterios de la Teoría Clásica de la evaluación.

TCA2, el nuevo test resultante de estas consideraciones, es pues una versión reducida del anterior que consiste en 110 ítems:

a) 40 ítems se tomaron de la Sección A. Se eliminaron los 9 ítems que obtuvieron un CD inadecuado (a2, a6, a14, a19, a28, a31, a33, a44, a48) así como el a4, que mostraba un CD rayano en el valor de Muy Fácil (CD = ,85) y era, además, el ítem con peor discriminación de toda la sección (ID = ,259).
b) Otros 40 ítems provienen de la Sección B. Se eliminaron los 5 ítems que resultaron Muy Fáciles en el análisis de dificultad (b1, b25, b30, b40, b42) y aquellos que mostraban poca discriminación (b16, b32, además del b1 ya eliminado). Para redondear el número de ítems igualándolo a la sección A y mejorar aún más la calidad de la nueva medida, decidimos eliminar otros 3 ítems que habían obtenido un CD demasiado elevado (b2 y b22, con un CD de ,83 y el b49 con ,82 de CD). Así pues, un total de 10 ítems fueron descartados de esta sección.
c) Los últimos 30 ítems proceden de la Sección D. Además de eliminar el ítem d47 por su alto índice de dificultad y los 14 ítems que habían mostrado poca discriminación (d1, d11, d14, d15, d17, d19, d27, d29, d30, d38, d39, d42, d48, d49), se decidió eliminar otros 5 más también para obtener un nuevo test más fiable y un número más redondo de preguntas. Los 5 ítems eliminados fueron el d5, d10 y d32 por su alto grado de dificultad (,16, ,22 y ,20 respectivamente), y el d6 y d50, por su bajo nivel de discriminación (,202 y ,192 respectivamente).
d) La Sección C se eliminó por completo por las razones ya expuestas.

En definitiva, TCA2 mantiene tres bloques con un total de 110 ítems (ver Anexo 6), un número que nos parece más adecuado y que pretendemos reduzca el efecto del cansancio en los resultados. Aunque no se ha hecho ningún pilotaje con nuevos

sujetos, algo que sin duda formará parte del trabajo de futuras investigaciones, creemos necesario acreditar la fiabilidad y la validez ─tanto en cuanto al CD como al ID─ de esta nueva versión. Para ello, utilizaremos los resultados obtenidos a partir del pilotaje de TCA1, aunque, como decimos, somos conscientes de que es necesario llevar a cabo una nueva administración para este segundo test.

Empezaremos por mostrar en la tabla 5.36. el nuevo índice de fiabilidad de esta prueba, que contendría ahora sólo 110 ítems.

Alfa de Cronbach	N de elementos
,973	110

Tabla 5.36.: Coeficiente de fiabilidad de TCA2

Aunque normalmente la reducción del número de ítems afecta gravemente a la fiabilidad de una prueba, en este caso y al eliminar todos los ítems perturbadores, TCA2 queda reforzado en 5 centésimas con relación a la versión 1 que mostraba un coeficiente alfa de Cronbach de 0,968.

En cuanto a las medidas centrales y de dispersión, la estadística del nuevo test ha quedado alterada de la siguiente manera:

	N	Mínimo	Máximo	Media	Desv. típ.
TCA2	311	7,27	98,18	58,5004	23,74983
N válido (según lista)	311				

Tabla 5.37.: Estadística descriptiva de TCA2

Si recordamos que la media de TCA1 era 45,67%, ahora se encuentra en 58,50%, lo cual indica que esta versión del test es más fácil y quizá más adecuada para el nivel del alumnado en general. En cuanto al análisis por secciones que ofrecemos a

continuación, también habría algún cambio en las medias: la Sección A del TCA1 (67,32%) bajaría su media hasta 61,01% en la nueva versión, la B cambiaría de 65,90% a 60,28% y la D, por el contrario, aumentaría su facilidad de 50,16% a 52,76%. Ante estos datos podemos concluir que la dificultad de las distintas secciones del test se homogeneizaría, haciendo así todos los bloques más equiparables y equilibrados.

	N	Media	Desv. típ.
Sección A	311	67,32	16,518
Sección B	311	65,90	16,768
Sección C	311	37,72	16,080
Sección D	311	50,16	19,406
N válido (según lista)	311		

Tabla 5.38.: Media por secciones en TCA1

	N	Media	Desv. típ.
Sección A	311	61,0129	28,33731
Sección B	311	60,2894	27,96305
Sección D	311	52,7653	18,28681
N válido (según lista)	311		

Tabla 5.39: Media por secciones en TCA2

Pasando ahora al análisis de CD de los ítems, los resultados son los siguientes (Tabla 5.40.):

ÍTEMS	Media	Desviación típica
a1	,41	,493
a3	,61	,488
a5	,68	,469
a7	,44	,497

a8	,60	,490
a9	,57	,496
a10	,58	,494
a11	,52	,500
a12	,79	,409
a13	,67	,471
a15	,48	,501
a16	,66	,476
a17	,60	,490
a18	,66	,476
a20	,64	,480
a21	,71	,454
a22	,64	,481
a23	,63	,484
a24	,81	,392
a25	**,85**	,354
a26	,66	,474
a27	,84	,369
a29	,49	,501
a30	,56	,498
a32	,70	,460
a34	,77	,422
a35	,46	,500
a36	,57	,496
a37	,66	,475
a38	,48	,500
a39	,74	,439
a40	,60	,491
a41	,65	,477
a42	,61	,489
a43	,34	,476
a45	,35	,479
a46	,77	,420
a47	,60	,491
a49	,69	,461
a50	,46	,500
b3	,78	,413
b4	,76	,426
b5	,77	,424
b6	,67	,470

b7	,77	,424
b8	,51	,501
b9	,78	,418
b10	,45	,498
b11	,69	,465
b12	,57	,495
b13	,69	,464
b14	,76	,430
b15	,64	,480
b17	,31	,463
b18	,46	,499
b19	,78	,415
b20	,63	,483
b21	,73	,444
b23	,56	,497
b24	,62	,485
b26	,47	,500
b27	,38	,487
b28	,34	,474
b29	,59	,492
b31	,73	,444
b33	,67	,471
b34	,73	,443
b35	,66	,474
b36	,43	,496
b37	,39	,488
b38	,48	,500
b39	,48	,500
b41	,75	,434
b43	,79	,409
b44	,65	,479
b45	,38	,487
b46	,51	,501
b47	,53	,500
b48	,64	,481
b50	,75	,436
d2	,53	,500
d3	,54	,499
d4	,47	,500
d7	,76	,428

d8	,50	,501
d9	,84	,366
d12	,50	,501
d13	,73	,446
d16	,32	,468
d18	,61	,489
d20	,45	,499
d21	,75	,436
d22	,61	,488
d23	,61	,488
d24	,61	,488
d25	,36	,480
d26	,51	,501
d28	,78	,415
d31	,42	,494
d33	,38	,485
d34	,39	,489
d35	,65	,478
d36	,37	,484
d37	,46	,499
d40	,45	,498
d41	,39	,488
d43	,44	,498
d44	**,24**	,430
d45	,78	,418
d46	,43	,496

Tabla 5.40.: Coeficiente de dificultad de TCA2

Podemos observar a partir de la tabla anterior que todos los ítems quedan ahora en parámetros de normalidad, yendo desde un mínimo de ,24 a un máximo de ,85 y evitando así las dos categorías marginales (,86-1,00 para Muy Fáciles y ,01-,14 para Muy Difíciles). En suma, los porcentajes totales han sido: 27 Fáciles (24,54%), 69 Satisfactorios (62,72%) y 14 Difíciles (12,72%).

En cuanto al análisis del ID, estos han sido los resultados:

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	**Correlación elemento-total corregida**	Alfa de Cronbach si se elimina el elemento
a1	64,31	661,053	,540	,972
a3	64,11	660,265	,577	,972
a5	64,04	658,523	,675	,972
a7	64,28	659,728	,589	,972
a8	64,11	663,261	,456	,972
a9	64,15	658,829	,626	,972
a10	64,13	659,021	,621	,972
a11	64,19	657,988	,653	,972
a12	63,93	664,490	,491	,972
a13	64,05	662,333	,513	,972
a15	64,23	659,756	,583	,972
a16	64,06	662,351	,507	,972
a17	64,11	659,033	,625	,972
a18	64,06	659,504	,625	,972
a20	64,07	657,763	,691	,972
a21	64,01	661,427	,572	,972
a22	64,08	659,147	,633	,972
a23	64,09	658,962	,635	,972
a24	63,91	661,024	,686	,972
a25	63,86	666,587	,454	,972
a26	64,06	660,000	,607	,972
a27	63,88	665,494	,492	,972
a29	64,22	660,832	,540	,972
a30	64,16	659,505	,596	,972
a32	64,02	664,514	,433	,972
a34	63,95	665,574	,425	,972
a35	64,25	660,014	,574	,972
a36	64,15	660,138	,574	,972
a37	64,06	663,580	,458	,972
a38	64,24	659,219	,604	,972
a39	63,98	663,684	,491	,972
a40	64,12	658,406	,649	,972
a41	64,06	661,442	,543	,972

a42	64,11	658,164	,661	,972
a43	64,37	663,023	,479	,972
a45	64,36	662,851	,483	,972
a46	63,94	667,811	,323	,973
a47	64,12	658,768	,635	,972
a49	64,02	659,540	,644	,972
a50	64,25	658,874	,619	,972
b3	63,94	662,628	,573	,972
b4	63,95	667,197	,346	,973
b5	63,95	661,766	,598	,972
b6	64,05	661,601	,545	,972
b7	63,95	662,398	,569	,972
b8	64,21	660,936	,536	,972
b9	63,94	662,642	,567	,972
b10	64,27	662,152	,491	,972
b11	64,03	664,149	,443	,972
b12	64,14	660,735	,551	,972
b13	64,03	658,471	,685	,972
b14	63,96	664,930	,446	,972
b15	64,07	660,590	,575	,972
b17	64,41	664,021	,451	,972
b18	64,26	659,983	,576	,972
b19	63,94	658,742	**,754**	,972
b20	64,08	663,433	,456	,972
b21	63,99	659,127	,687	,972
b23	64,15	659,140	,612	,972
b24	64,09	659,799	,600	,972
b26	64,25	660,364	,560	,972
b27	64,33	662,803	,477	,972
b28	64,38	661,617	,540	,972
b29	64,12	663,183	,457	,972
b31	63,99	659,010	,692	,972
b33	64,05	661,421	,551	,972
b34	63,98	660,856	,613	,972
b35	64,06	658,515	,669	,972
b36	64,29	660,649	,553	,972
b37	64,33	661,420	,531	,972
b38	64,24	660,624	,549	,972
b39	64,24	658,815	,620	,972
b41	63,97	663,126	,523	,972

b43	63,93	664,158	,507	,972
b44	64,07	659,265	,630	,972
b45	64,33	660,829	,556	,972
b46	64,21	661,064	,531	,972
b47	64,19	659,667	,587	,972
b48	64,08	659,323	,625	,972
b50	63,97	663,983	,482	,972
d2	64,19	665,854	,345	,973
d3	64,18	669,643	,198	,973
d4	64,25	667,809	,269	,973
d7	63,96	668,861	,269	,973
d8	64,22	667,156	,293	,973
d9	63,88	667,451	,392	,972
d12	64,22	669,280	,211	,973
d13	63,99	667,886	,300	,973
d16	64,40	667,067	,319	,973
d18	64,11	669,695	,200	,973
d20	64,26	664,344	,405	,972
d21	63,97	670,126	,208	,973
d22	64,10	665,898	,352	,973
d23	64,10	667,520	,287	,973
d24	64,10	661,110	,544	,972
d25	64,36	668,537	,251	,973
d26	64,21	663,372	,441	,972
d28	63,94	667,472	,343	,973
d31	64,30	668,381	,250	,973
d33	64,34	665,111	,386	,972
d34	64,32	667,582	,284	,973
d35	64,07	669,393	,217	,973
d36	64,35	668,358	,256	,973
d37	64,26	669,541	,202	,973
d40	64,27	664,926	,382	,973
d41	64,33	668,848	,234	,973
d43	64,27	669,847	**,190**	,973
d44	64,47	668,635	,278	,973
d45	63,94	669,417	,250	,973
d46	64,29	667,535	,282	,973

Tabla 5.41.: Índice de discriminación de TCA2

La discriminación queda también en coordenadas admisibles y va desde 0,190, Pasable, a 0,754, una magnífica discriminación. Los 81 ítems (73,63%) con discriminación Muy Buena son claramente mayoría, hay además 10 (9,09%) en la categoría de Buena y los de Pasable suman 19 (17,27%). No debemos olvidar, sin embargo, que en la Sección D, aquella que sigue mostrando los índices de discriminación más bajos, se han realizado modificaciones en los distractores estadísticamente problemáticos, lo cual confiamos contribuya a incrementar la discriminación de dichos ítems en futuras administraciones del test.

Por último, nos parecía absolutamente necesario que este test respondiese a las proporciones que la base de datos aportó sobre las combinaciones gramaticales de las colocaciones ((N+N, A+N, V+N y N+N) y que tiene que ver directamente con la validez de constructo y de contenido de nuestra prueba. Esta ratio ha sido muy difícil de conseguir ya que se ha "tenido que jugar" con los 10 ítems que hemos quitado de las Secciones A y B para obtener una distribución similar a la anterior, sin perder de vista la importancia de que el CD y el ID fuera el adecuado en cada ítem. Sin embargo, observando la tabla 5.43., vemos que aunque no hay una exactitud completa, TCA2 se acerca mucho a la anterior, mostrando sólo una pequeña diferencia en la combinación N+N y V+N. Compárense las dos versiones:

		Frecuencia	Porcentaje	Porcentaje acumulado
Válidos	N+N	14	7,0	7,0
	A+N	117	58,5	65,5
	V+N	64	32,0	97,5
	N+V	5	2,5	100,0
	Total	200	100,0	

Tabla 5.42.: Proporción de estructuras gramaticales en TCA1

		Frecuencia	Porcentaje	Porcentaje acumulado
Válidos	N+N	9	8,2	8,2
	A+N	63	57,3	65,5
	V+N	35	31,8	97,3
	N+V	3	2,7	100,0
	Total	110	100,0	

Tabla 5.43.: Proporción de estructuras gramaticales en TCA2

5.4.1. Virtualización de TCA2

Una vez llevada a cabo la profunda revisión de la nueva versión de nuestro test, decidimos dar un paso más allá y proceder a su virtualización. El motivo por el que optamos por este método de administración es que presenta una serie de ventajas con respecto al formato tradicional de lápiz y papel que no queríamos dejar de aprovechar. López-Mezquita (2005) enumera una serie de aspectos técnicos y humanos que inciden muy positivamente sobre la fiabilidad y la practicabilidad de los instrumentos de evaluación:

- En lo que respecta a la administración, los tests informatizados contribuyen a aumentar la seguridad de las pruebas (por ejemplo, mediante la presentación de las preguntas en distinto orden para cada nueva administración), permiten homogeneizar las condiciones de aplicación, hacen posible presentar cada ítem de forma individual, lo cual evita, en cierta medida, la frustración que le produce al examinando encontrar un test demasiado largo, y se pueden administrar de forma más rápida, siendo por tanto más eficaces. Este tipo de tests, además, ofrecen mayor flexibilidad en su administración, puesto que pueden ser cumplimentados

en línea, con lo que desaparecen las limitaciones espacio-temporales que presentan los tests tradicionales.

- En los tests de formato objetivo, los ordenadores son mucho más fiables que los humanos en el proceso de corrección de las pruebas, siendo capaces de ofrecer una puntuación y también una retroalimentación inmediata al examinando. Asimismo, pueden aportar una información muy completa sobre la actuación de cada uno de los candidatos (tiempo empleado en dar la respuesta, acceso a recursos y materiales de referencia, número de intentos antes de dar con la respuesta correcta en el caso de que se habilite la opción de varios intentos, etc.).
- El formato electrónico tiene también, por lo general, un impacto muy positivo en el candidato dado que a los alumnos le suele parecer motivador trabajar con el ordenador, en lugar de responder a un test en formato papel.

En nuestro caso, la virtualización de TCA2 se ha llevado a cabo en la plataforma Moodle[1]. En la actualidad, el Centro de Enseñanzas Virtuales de la Universidad de Granada (CEVUG[2]) utiliza esta herramienta para el diseño de sus cursos virtuales, siendo uno de dichos cursos el programa ADELEX, dentro del cual se enmarca nuestra investigación. Para no extendernos demasiado en este apartado, baste decir que Moodle es un complejo sistema de teleformación capaz de recrear un entorno didáctico a través de Internet. Se trata, pues, de una plataforma que ofrece una serie de herramientas capaces de recrear un aula virtual (con prestaciones que permiten presentar materiales didácticos, actividades, tests autocorregibles, foros, e-mail, chats, calendarios, etc.).

[1] http://moodle.org/ [Último acceso: 21.02.2009]
[2] http://cevug.ugr.es/ [Último acceso: 21.02.2009]

En lo que se refiere a los tests digitalizados, Moodle cuenta con plantillas que permiten diseñar pruebas de distinto formato (elección múltiple, preguntas cortas de respuesta abierta, unir parejas, rellenar huecos, etc.). Dadas sus distintas posibilidades, esta plataforma nos brindaba la oportunidad de convertir las tres secciones de nuestro test a formato digital, lo cual nos permitiría poder pilotarlo también dentro de la asignatura virtual ADELEX. Así pues, una vez llevado a cabo el proceso de virtualización, obtuvimos tres pruebas distintas, una para cada sección del test, cada una de las cuales viene acompañada de sus instrucciones. Ofrecemos a continuación una muestra de cada una de las secciones tal y como aparecen en Moodle (Figuras 5.15, 5.16. y 5.17.):

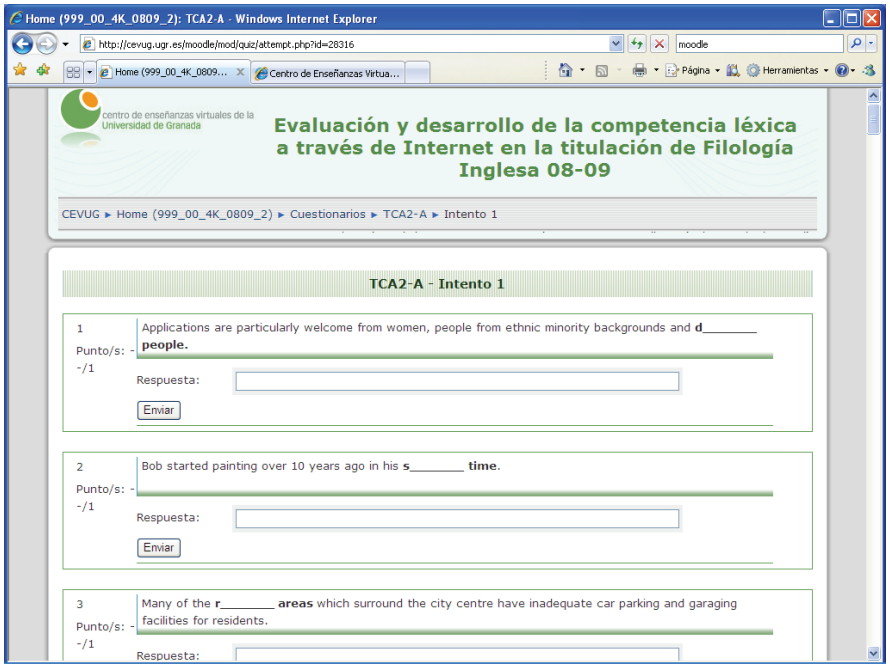


Fig. 5.15.: TCA2, Sección A

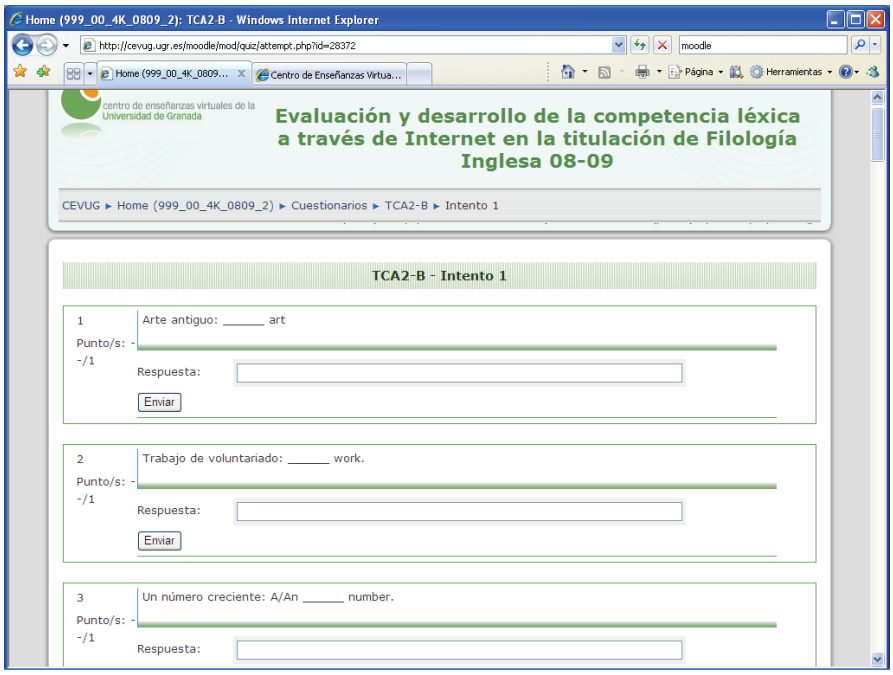


Fig. 5.16.: TCA2, Sección B

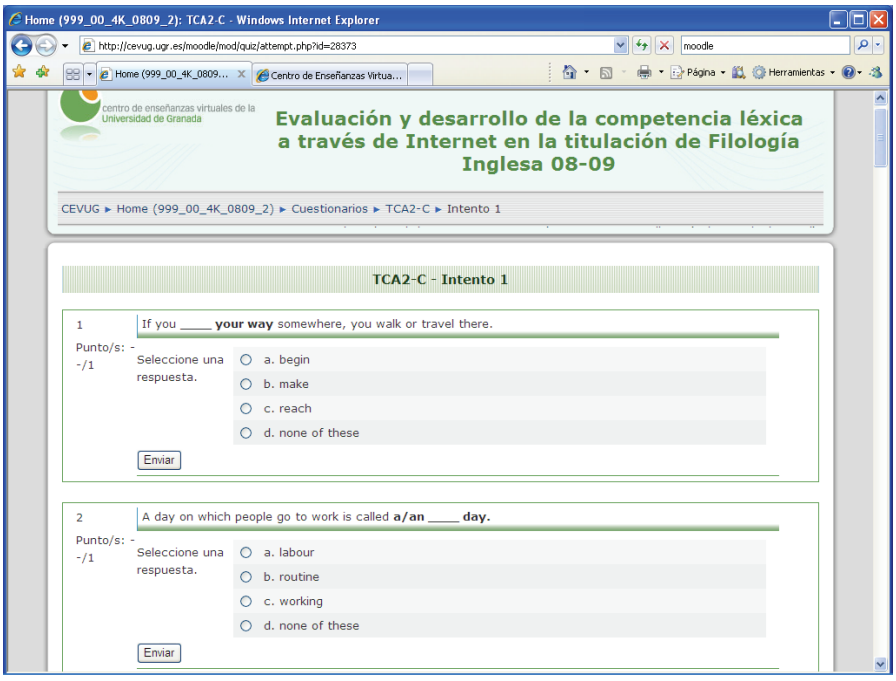


Fig. 5.17.: TCA2, Sección C

Finalmente, sólo nos gustaría subrayar que tras el cuidadoso análisis llevado a cabo y las modificaciones realizadas en el TCA2, el nuevo test informatizado está ya preparado para su pilotaje.

5.5. Conclusión

En este capítulo hemos mostrado los resultados de nuestra investigación. El objetivo de este diseño de investigación era llevar a cabo un diagnóstico riguroso del conocimiento de las colocaciones de los sustantivos contenidos en las 1.000 palabras más frecuentes con relación a una población constituida por alumnos españoles de las licenciaturas de Filología Inglesa de las Universidades de Granada y de Almería y de las licenciaturas de Traducción e Interpretación de la Universidad de Granada y la Universitat Jaume I (Castellón). A tal fin, se diseñó el Test de Colocaciones ADELEX Versión 1 sobre el que ya hablamos con profusión en el capítulo anterior. Los resultados, en general, muestran de forma muy contundente el bajo nivel del alumnado, sobre todo si tenemos en cuenta que las colocaciones evaluadas se encuentran entre las más frecuentes de la lengua, por lo que nos parece urgente que se establezca un tratamiento pedagógico de este aspecto tan esencial para lograr una buena competencia léxica.

En cuanto a la investigación cuantitativa, se llevaron a cabo medidas estadísticas de tendencia central —la media, la mediana, la moda, los valores mínimos y máximos—, y de dispersión —desviación típica y varianza—, al mismo tiempo que se calculó el índice de fiabilidad mediante el coeficiente alfa de Cronbach. También se realizaron exploraciones de estadística inferencial o muestral a través de t-tests de muestras independientes y análisis de la varianza (ANOVA) para refutar o confirmar

hipótesis planteadas en esta investigación en torno a las diferencias de nivel en: las licenciaturas, los cursos de las licenciaturas entre sí, las secciones del test o los métodos evaluadores y, por último, las estructuras combinatorias de la colocación (N+N, A+N, V+N, N+V). Acudimos, asimismo, a correlaciones para encontrar similitudes o discrepancias entre muestras relacionadas. Por último, se acometió un análisis de ítems para averiguar el coeficiente de dificultad (CD) y de discriminación (ID) del ítem, a la vez que se realizó un profundo examen de los distractores que no cumplían con su función de forma fiable. Para la obtención de todos estos resultados estadísticos se utilizó el programa estadístico SPSS v.15 (*Statistical Package for Social Sciences*).

A la vista de los errores y distorsiones encontrados en el TCA1, se procedió a hacer las modificaciones oportunas, y se construyó un segundo test, TCA2, que esperamos haya mejorado considerablemente este instrumento de medida para comprobar el conocimiento colocacional. Este segundo test, además, ha sido construido en formato digitalizado, ya que, entre otras ventajas, ello contribuirá a reducir el tiempo de administración, lo cual confiamos mejore los resultados obtenidos. Así pues, en este momento el test se encuentra preparado para futuras investigaciones.

CONCLUSIONS

[The conclusions of this thesis will be written in English complying with current regulations concerning the European Doctorate Distinction].

The piece of research we have reported in this work has addressed the complex issue of second language collocation assessment. This study stems from the results of a previous investigation the author conducted during her Masters Degree where both the scarcity of efficient measures for testing collocational knowledge existing to date, and the problems our students show when dealing with these lexical elements called for further research. Thus, this thesis is an attempt to make a contribution in this field.

The overall aim of our study was two-fold. First, we intended to assess whether our students are proficient in collocations or, on the contrary, we should take some course of action to ensure a better pedagogical treatment of this important aspect of linguistic competence in our classrooms. Second, we tried to design a new measure which might lead to a valid and reliable assessment of students' productive and receptive knowledge of collocations, ADELEX Test of Collocations v.1. More specifically, the **aims** we set at the beginning of our study were:

- To assess students' knowledge of collocations.

  - To analyse variables such as university degree, year of studies and grammatical structure of collocations in order to fully evaluate their impact on learners' collocational competence.

- To design a valid and reliable test for collocational assessment.

  - To offer a definition of the concept of collocation, and one which is operational in order to secure construct validity.
  - To compile a list of frequent collocations by means of corpus-based techniques in order to create a bank of data for item selection which may ensure content validity.
  - To design a test bearing upon a validated model for test construction (López-Mezquita, 2005), where issues concerning reliability (item format, number of items, instructions, timing and scoring criteria) and validity (construct, selection of contents and items) are carefully addressed.
  - To empirically verify test validity and reliability, by running statistical analyses on test performance regarding different test formats and items.

In order to tackle the issues detailed above, we first endeavoured to offer an account of main tendencies and approaches to the concept of collocation from a theoretical viewpoint. Thus, **Chapter 1** aimed at describing the state of the art in the field of collocation studies and to put forward our **definition of the construct**, basically rooted in the phraseological tradition (i.e. the **arbitrarily restricted commutability of words**), but with an emphasis on the crucial role of frequency and the invaluable

contribution corpus-based statistical techniques can make to the study of collocations.

Being particularly concerned with the role that frequency has played in the field of second language vocabulary teaching/learning, in **Chapter 2** we carried out a chronological analysis of the notion of frequency in Applied Linguistics. The basic assumption made in this chapter was that frequency, though not the only criterion to bear in mind when teaching or testing lexical contents, is a crucial factor in second language pedagogy since **those elements which are most frequent in natural language are the ones our students definitely need to know**.

In **Chapter 3** we reported on the careful procedure to compile a corpus-driven list of frequent collocations. By making use of two of the most representative English corpora (the Bank of English and the British National Corpus) and also with the help of recent technological advances —in the way of computerised resources currently available for the study of collocations— we first drew a list of statistically significant word combinations. However, we did not aim at a radical statistically-based extraction of collocations. Given their phraseological nature and, again, in order to secure construct validity, we carried out a manual analysis of the resulting list, and performed a further selection of those N+N, V+N, A+N and N+V combinations which show the inherent arbitrariness which characterises and ultimately determines the nature of collocations, as opposed to free combinations and idioms. All in all, we have hopefully produced **a comprehensive, reliable and innovative list of frequent collocations** which may have an enormous potential in the field of ELT.

Chapter 4 was devoted to the **construction of ADELEX Test of Collocations v.1,** a 200 item test devoted to tap into students' receptive and productive knowledge of collocations. Following López-Mezquita's (2005) model for test construction, we designed a measure where the **fundamental issues of validity**

and reliability were taken into consideration. Bearing on the theoretical framework established in Chapters 1 and 2, we were able to produce a test which tries to lie on a well-defined and operational construct. Furthermore, the list of frequent collocations produced in our study served as the basis for the principled selection of contents which constitute a representative sample of the field, thus accounting for content validity. Relevant issues concerning test reliability were also taken into account, especially those related to practical aspects of test design (formats and construction of items, timing, conditions of administration and scoring). In general, all the requisites for test planning, construction and administration were carefully met, in an attempt to produce a test which could be confidently used as an adequate indicator of candidates' knowledge in this field.

In order to tap, on the one hand, into students' knowledge of collocations, and also with the aim of assessing the adequacy of our measure as a valid and reliable instrument for collocational evaluation, we performed a thorough statistical analysis of our results. Consequently, **Chapter 5** was devoted to give a full account of this process. After being administered to 311 subjects, we first run a reliability analysis on our test, yielding a total score of .968 as estimated through Cronbach's alpha. Once reliability was thus attested, we evaluated students' performance. Statistical analyses were run in order to answer our first four research questions:

- **What is Spanish university students' knowledge of collocations, as measured in four different university institutions: University of Granada (English Philology and Translation Studies), University of Almería (English Philology) and University Jaume I (Translation Studies)?**

On the whole, the results yielded by our pilot test lead us to conclude that the overall collocational competence of Spanish university students is insufficient (mean =

45.67%) and this indicates that students may fall short in the social and academic demands made on their command of L2. The major implication of these results, moreover, would seem to be the urgent need to carry out an efficient pedagogical intervention to overcome students' collocational deficiencies.

- **Is there any difference between English Philology and Translation Studies students in terms of knowledge of collocations?**

When comparing results obtained from both university degrees, students of Translation Studies performed better than English Philology (48.97% vs. 41.42%), being this difference statistically significant (p =.000). We can conclude, thus, that these results can be confidently generalised to a similar population.

- **Are there differences between students from different university levels?**

As can be observed below, there existed differences between students from the four years of the university degrees, showing thus the **scalability** of our measure. This difference was statistically significant between groups in all cases except between 2nd and 3rd year students, where no significance was found:

Year of studies	Overall mean (%)	English Philology (%)	Translation Studies (%)
First	28.6828	25.4837	33.0110
Second	47.9824	44.3319	50.5030
Third	50.5342	49.5691	51.1776
Fourth	56.9692	56.1902	57.3958
Total	45.6756	41.4256	48.9786

- **Are there differences in students' performances on different grammatical structures of collocations (N+N, A+N, V+N, N+V)? Are some grammatical combinations more difficult than others for L2 learners?**

Given the small sample we used for N+N and N+V collocations (in accordance to the proportions retrieved from our corpus-based list of collocations), we could only compare A+N and V+N results. Although the former seemed to be more difficult for learners (54.11% vs. 59.39%), this difference was not statistically significant, so we can only affirm that there seems to be a "tendency" in this respect. Further studies will try to confirm or refute this contention.

Regarding our second aim, assessing whether our test was a valid and reliable instrument for collocation assessment, we performed comparative analyses between the different sections, as well as an item analysis, and finally we examined the way distractors behaved, in an attempt to answer the research questions posed at the outset of this study in relation to test performance (research questions 5, 6 and 7):

- **Are there significant differences between scores attributable to tests formats? Are some formats better than others to measure collocational knowledge?**

By analysing and comparing results on the four sections of our test, made up of four different formats (c-test, translation, odd-one-out technique and multiple choice), it was empirically proven that the odd-one-out technique did not seem to be an effective task for collocation assessment. On the other hand, c-test, translation and multiple choice methods seem to be valid measures for testing collocations.

- **How do items contribute to test validity and reliability? Which are the ones we need to revise or reject?**

After performing facility and discrimination item analyses, 31 items out of 150 (the total remaining after rejecting the odd-one-out section) were rejected —9 from the c-test section, 7 from the translation one and 15 from the multiple choice task. In other words, 59.5% of the items remained.

- **How did distractors perform in multiple choice items?**

Following Green (personal communication, 2008), we carried out a careful examination of distractors. This analysis revealed that 12 of the items called for improvement and thus, we designed new distractors.

In the light of the results reported above, we decided to produce an improved version of our test, **ADELEX Test of Collocations v.2**, which intends to overcome the main problems identified in the previous one. The observation that our results would have probably been better had this new version been administered to our group (overall reliability would have increased to .973 and all items would have been within acceptable facility and discrimination parameters) seems very promising to us. Furthermore, the new test has been **computerised**, and it is now ready for piloting via a digital platform (Moodle).

To conclude, a further note should be added concerning the **limitations** of our study. Firstly, we believe that an in-depth analysis concerning the intrinsic difficulty

of collocations might provide new insights into SLA, insofar as there is hardly any research on the way students process and retrieve collocations. It was, however, beyond the scope of this thesis to investigate these aspects.

Another path which could have been opened is how stays in an English-speaking country affected collocational thresholds. And last but not least, we would love to look for the Holy Grail and carry out a longitudinal study examining the role of instruction in collocational mastery.

As we will make clear in the next, and last, section of this thesis, these and other topics for future research projects will be suggested.

FUTURE RESEARCH

In this thesis an attempt has been made to delve into the area of collocation assessment by designing a valid and reliable measurement. However, there is no doubt that this is only the first of the many actions that need to be taken in order to come up with a definite tool which provides a reliable measure of students' performance. In order to make further contributions in this field, we intend to conduct the following studies in the future:

- **Pilot administrations of ADELEX Test of Collocations v.2** will be carried out in order to improve our measure. This will be done, moreover, through a computer-based methodology.

- **Concurrent validity** will be examined by comparing results obtained from our test with those yielded by measures of lexical competence (López-Mezquita, 2005) and general English proficiency.

- **Factors affecting collocation difficulty** for L2 learners will be carefully analysed by examining the nature of different collocations and results obtained from the administration of our test.

- **External variables** which might influence test performance and, ultimately, knowledge of collocations, especially time spent in English-speaking countries, will be taken into consideration.

- Our **list of collocations will be enlarged** so that it will cover those frequent combinations integrated by nouns included in the first 2,000 words of the language, as it is generally considered the basic threshold in the field of L2 vocabulary learning (Nation, 2001).

It goes without saying that there is still a good way ahead to offer conclusive evidence of the design of a definite test, but we hope this work has been a step forward in collocation assessment.

REFERENCIAS BIBLIOGRÁFICAS

AARTS, J. 1991. "Intuition-based and observation-based grammar". En K. Aijmer y B. Altenberg (eds.), *English corpus linguistics: Studies in honour of Jan Svartvik*. (pp. 44-62). Londres: Longman.

ABAD, F. J., ATENCIA, J., GARCÍA C., HONTANGAS, P., OLEA J., PONSODA, V., REVUELTA J., SUERO, M. Y XIMÉNEZ, C. 2004. *Ayuda a la creación de exámenes*. Disponible en http://www.uam.es/docencia/ace [Última consulta: 24.03.2009]

ABAD, F. J., OLEA J. Y PONSODA, V. 2001. "Analysis of the optimum number of alternatives from the Item Response Theory". *Psicothema*, 13, 1: 152-158.

ADOLPHS, S. Y DUROW, V. 2004. "Social-cultural integration and the development of formulaic sequences". En N. Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use*. (pp. 107-126). Ámsterdam/Filadelfia: John Benjamins.

AISENSTADT, E. 1979. "Collocability restrictions in dictionaries". En R. Hartmann (ed.), *Dictionaries and their users*. (pp. 71-74). Exeter Linguistics Studies, 4. University of Exeter.

AISENSTADT, E. 1981. "Restricted collocations in English lexicology and lexicography". *ITL Review of Applied Linguistics*, 53: 53-61.

AITCHISON, J. 1994. *Words in the mind: An introduction to the mental lexicon*. Oxford: Basil Blackwell.

ALDERSON, J.C., CLAPHAM, C. Y WALL, D. 1995. *Language test construction and evaluation*. Cambridge: Cambridge University Press.

ALLERTON, D. J. 2002. *Stretched verb constructions in English*. Londres: Routledge.

ALONSO RAMOS, M. 1993. *Las funciones léxicas en el modelo lexicográfico de I. Mel'cuk*. Tesis doctoral. Madrid: U.N.E.D.

ALTENBERG, B. 1993. "Recurrent verb-complement constructions in the London-Lund Corpus". En J. Aarts, P. de Haan y N. Oostdijk (eds.), *English language corpora: Design, analysis and exploitation*. (pp. 227-245). Ámsterdam: Rodopi.

ALTENBERG, B. 1998. "On the phraseology of spoken English: The evidence of recurrent word-combinations". En A. Cowie (ed.), *Phraseology. Theory, analysis, and applications*. (pp. 101-122). Oxford: Oxford University Press.

ANAGNOSTOU, N. Y WEIR, G. 2006. "Review of software applications for deriving collocations". En *ICT in the Analysis, Teaching and Learning of Languages, Preprints of the ICTATLL Workshop 2006*. (pp. 91-100). Glasgow, agosto 2006. Disponible en: http://www.cis.strath.ac.uk/cis/research/publications/papers/strath_cis_publication_1541.pdf [Último acceso: 05.01.2009]

ARNAUD, P. 1992. "Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate component tests". En P. Arnaud y H. Béjoint (eds.), *Vocabulary and applied linguistics*. (pp. 133-145). Londres: Macmillan.

ASTON, G. Y BURNARD, L. 1998. *The BNC handbook. Exploring the British National Corpus with Sara*. Edimburgo: Edinburgh University Press.

BACHMAN, L. F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

BACHMAN, L. F. Y PALMER, A. S. 1996. *Language testing in practice*. Oxford: Oxford University Press.

BAHNS, J. Y ELDAW, M. 1993. "Should we teach EFL students collocations?". *System*, 21, 1: 101-114.

BARFIELD, A. 2003. *Collocation recognition and production: Research insights*. Tokio: Chuo University.

BARFIELD, A. 2006. An exploration of second language collocation knowledge and development. Tesis doctoral no publicada. University of Wales: Swansea.

BENSON, M. 1985. "Collocations and idioms". En R. Ilson (ed.), *Dictionaries, lexicography and language learning. ELT Documents, 120*. (pp. 61-68). Oxford: Pergamon Press/British Council.

BENSON, M. 1989. "The structure of the collocational dictionary". *International Journal of Lexicography*, 2, 1: 1-14.

BENSON, M. 1995. "Review article of G. Kjellmer (1994) *A Dictionary of English Collocations: Based on the Brown Corpus*". *International Journal of Lexicography*, 8, 1: 65-67.

BENSON, M., BENSON, E. E ILSON, R. 1986. *The BBI combinatory dictionary of English*. Ámsterdam/Filadelfia: John Benjamins.

BERNARDINI, S. 2000. "Systematising serendipity: Proposals for concordancing large corpora with language learners." En L. Burnard y T. McEnery (eds.), *Rethinking language pedagogy from a corpus perspective. Papers from the Third International Conference on Teaching and Language Corpora*. (pp. 183-190). Hamburgo: Peter Lang.

BERNARDINI, S. 2002. "Serendipity expanded: Exploring new directions for discovering learning". En B. Kettermann y G. Marko (eds.), *Teaching and learning by doing corpus analysis. Papers from the Fourth International Conference on Teaching and Language Corpora*. (pp. 165-182). Ámsterdam: Rodopi.

BERNARDINI, S. 2007. "Collocations in translated language: Combining parallel, comparable and reference corpora". En *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham, julio 2007. Disponible en http://www.corpus.bham.ac.uk/corplingproceedings07/paper/15_Paper.pdf [Último acceso: 11.01.2009].

BERRY-ROGHE, G. L. M. 1972. "The computation of collocations and their relevance to lexical studies". Publicado en la página web del Laboratorio Informático Atlas (Chilton, Reino Unido). Disponible en: http://www.chilton-computing.org.uk/acl/applications/cocoa/p010.htm [Último acceso: 11.01.2009]. Reimpreso en 1973 en A. J. Aitken, R. W. Bailey y N. Hamilton-Smith (eds.), *The computer and literary studies*. (pp. 103-112). Edimburgo: Edinburgh University Press.

BIBER, D., CONRAD, S. Y REPPEN, R. 1998. *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. Y FINEGAN, E. 1999. *The Longman grammar of spoken and written English*. Londres: Longman.

BISKUP, D. 1992. "L1 influence on learners' renderings of English collocations. A Polish/German empirical study". En P. Arnaud y H. Béjoint (eds.), *Vocabulary and applied linguistics*. (pp. 85-93). Londres: Macmillan.

BOERS, F., EYCKMANS, J., KAPPEL, J., STENGERS, H. Y DEMECHELEER, M. 2006. "Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test". *Language Teaching Research*, 10, 3: 245-261.

BOLINGER, D. 1976. "Meaning and memory". *Forum Linguisticum*, 1, 1: 1-14.

BONGERS, H. 1947. *The history and principles of vocabulary control*. Woerden: Wocopi.

BONK, W.J. 2001. "Testing ESL learners' knowledge of collocations". En T. Hudson y J. D. Brown (eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests*. (Technical Report #21). (pp. 113-142). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

BROWN, J.D. 2000. "What is construct validity?" *Shiken: JALT, Testing & Evaluation SIG Newsletter*, 4, 2: 7-10. Disponible en http://www.jalt.org/test/bro_8.htm [Última consulta 23.03.2009]

BUTLER, C. 2005. "Functional approaches to language". En C. Butler, M. A. Gómez-González y S. Doval-Suárez (eds.), *The dynamics of language use*. (pp. 3-17). Ámsterdam/Filadelfia: John Benjamins.

BUTLER, C. 2008. "The very idea! A corpus-based comparison of IDEA, CONCEPT and NOTION and their formal equivalents in Spanish". *Atlantis*, 30, 2: 59-77.

CARROLL, B., DAVIES, P. Y RICHMAN, B. 1971. *The American heritage word frequency book*. Nueva York: American Heritage Publishing Co.

CARROLL, B.J, Y HALL, P.J. 1985. *Make your own language tests. A practical guide to writing language performance tests*. Oxford: Pergamon Press.

CARTER, R. 1987. *Vocabulary: Applied linguistic perspectives*. Londres/Nueva York: Routledge.

CARTER, R. Y MCCARTHY, M. (eds.). 1988. *Vocabulary and language teaching*. Londres: Longman.

CHOMSKY, N. 1957. *Syntactic structures*. La Haya: Mouton.

CHOMSKY, N. 1962. "Discussion". En A. Hill (ed.), *Third Texas conference on problems of linguistic analysis in English*. Austin: University of Texas.

CHOMSKY, N. 1965. *Aspects of the theory of syntax*. Cambridge (Mass.): MIT Press.

CHOMSKY, N. 1992. "On the nature, use and acquisition of language". En M. Pütz (ed.), *Thirty years of linguistic evolution. Studies in honour of René Dirven on the occasion of his sixtieth birthday*. (pp. 3-29). Ámsterdam/Filadelfia: John Benjamins.

CHUNG, Y. M. Y LEE, J. Y. 2001. "A corpus-based approach to comparative evaluation of statistical term association measures". *Journal of the American Society for Information Science and Technology*, 52, 4: 283-296.

CHURCH, K. Y HANKS, P. 1989. "Word association norms, mutual information and lexicography". En *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. (pp. 76-83). Vancouver, 1989.

CHURCH, K., GALE, W., HANKS, P. Y HINDLE, D. 1991. "Using statistics in lexical analysis". En U. Zernik (ed.), *Lexical acquisition: Exploiting on-line resources to build a lexicon*. (pp. 115-164). Hillsdale, NJ: Lawrence Erlbaum Associates.

CHURCH, K. Y MERCER, R. L. 1993. "Introduction to the special issue on computational linguistics using large corpora". *Computational Linguistics* 19, 1: 1–24.

COBB, T. 2003. "Why & how to use frequency lists to learn words". Disponible en http://www.wordhacker.com/en/article/usefrequency2learnwords.htm [Último acceso: 02.03.2009]

COBB, T. 2004. *Web Vocabprofile*. Disponible en http://www.lextutor.ca/vp/eng/ [Último acceso: 16.02.2009]

COLLINS COBUILD ENGLISH COLLOCATIONS ON CD-ROM 1995. Londres: HarperCollins Publishers.

COMPACT OXFORD ENGLISH DICTIONARY. 1994. Oxford: Oxford University Press.

CONNELLY, M. 1997. "Using C-tests in English with post-graduate students". *English for Specific Purposes*, 16, 2: 139-150.

COP, M. 1988. "The function of collocations in dictionaries". En T. Magay y J. Zigani (eds.), *Budalex proceedings: Papers from the Euralex third international congress*. (pp. 35-46). Budapest: Akademiai Kiado.

CORPAS PASTOR, G. 1992. "Las colocaciones como problema en la traducción actual Inglés/Español". *Revista del Departamento de Filología Moderna* 2, 3: 179-186.

CORPAS PASTOR, G. 1996. *Manual de fraseología española*. Madrid: Gredos.

CORTÉS, V. 2006. "Teaching lexical bundles in the disciplines: An example from a writing intensive history class". *Linguistics and Education*, 17: 391-406.

COULMAS, F. (ed.). 1981. *Conversational routine*. La Haya: Mouton de Gruyter.

COWIE, A. P. 1988. "Stable and creative aspects of vocabulary use". En R. Carter y M. McCarthy (eds.), *Vocabulary and language teaching*. (pp. 126-139). Londres: Longman.

COWIE, A. P. 1994. "Phraseology". En R. Asher (ed.), *The encyclopaedia of language and linguistics*. Vol. 6. (pp. 3168-3171). Oxford: Pergamon Press.

COWIE, A. P. 1998. "Phraseological dictionaries: Some east-west comparisons". En A. P. Cowie (ed.), *Phraseology. Theory, analysis, and applications*. (pp. 209-228). Oxford: Oxford University Press.

COXHEAD, A. 2000. "The academic word list". *TESOL Quarterly*, 34, 2: 213-238.

CRYSTAL, D. 1991. *A dictionary of linguistics and phonetics*. Oxford: Blackwell.

DAVIES, M. 2005. *Frequency dictionary of Spanish: Core vocabulary for learners*. Londres/Nueva York: Routledge.

DAVIES, M. 2008. *BYU-BNC*. Disponible en: http://corpus.byu.edu/bnc/ [Último acceso: 30.12.2008]

DE COCK, S., GRANGER, S., LEECH, G. Y MCENERY, T. 1998. "An automated approach to the phrasicon on EFL learners". En S. Granger (ed.), *Learner English on computer*. (pp. 67-79). Londres: Addison Wesley Longman.

DEIGNAN, A. 2005. *Metaphor and corpus linguistics*. Ámsterdam/Filadelfia: John Benjamins.

DICCIONARIO COLLINS UNIVERSAL ESPAÑOL-INGLÉS/INGLÉS-ESPAÑOL, 6ª edición. 2005. Barcelona: Grijalbo.

DÖRNYEI, Z. 2003. *Questionnaires in second language research: Construction, administration, and processing*. Mahwah: Lawrence Erlbaum Associates.

DUNNING, T. 1993. "Accurate methods for the statistics of surprise and coincidence". *Computational Linguistics*, 19, 1: 61-74.

DURRANT, P. 2008. *High frequency collocations and second language learning*. Tesis doctoral. University of Nottingham. Disponible en:

http://etheses.nottingham.ac.uk/622/1/final_thesis.pdf [Último acceso: 20.01.2009]

EBEL, R. L. 1965. *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice Hall.

ERMAN, B. Y WARREN, B. 2000. "The idiom principle and the open choice principle". *Text*, 20: 29-62.

EVERT, S. 2004. *The statistics of word cooccurrences: Word pairs and collocations*. Tesis doctoral. University of Stuttgart. Disponible en: http://www.collocations.de/EK/index.html#Evert_04 [Último acceso: 03.01.2009]

EVERT, S., HEID, U. Y SPRANGER, K. 2004. "Identifying morphosyntactic preferences in collocations". En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisboa, mayo 2004. Disponible en: http://www.ims.uni-stuttgart.de/projekte/corplex/paper/evert/EvertHeidSpranger2004.pdf [Último acceso: 18.12.2008]

EVERT, S. Y KRENN, B. 2001. "Methods for the qualitative evaluation of lexical association measures". En *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. (pp. 188-195). Toulouse, julio 2001. Disponible en http://www.collocations.de/EK/Articles/EvertKrenn2001.pdf [Último acceso: 04.01.2009]

EVERT, S. Y KRENN, B. 2005. "Exploratory collocation extraction". Comunicación presentada en el congreso Phraseology 2005: The Many Faces of Phraseology, Lovaina, octubre 2005. Disponible en: http://www.ofai.at/~brigitte.krenn/papers/evert-krenn.pdf [Último acceso: 02.01.2009]

FARGHAL, M. Y OBIEDAT, H. 1995. "Collocations: a neglected variable in EFL". *International Review of Applied Linguistics in Language Teaching*, 33, 4: 315-331.

FAUCETT, L., WEST, M., PALMER, H. Y THORNDIKE, E. 1936. *The interim report on vocabulary selection for the teaching of English as a foreign language*. Londres: P. S. King.

FIRTH, J. 1957. *Papers in linguistics 1934-1951*. Londres: Oxford University Press.

FLETCHER, W. 2008. *Phrases in English*. Disponible en: http://phrasesinenglish.org/ [Último acceso: 30.12.2008]

FONTENELLE, T. 1994. "What on earth are collocations?". *English Today*, 40, 10, 4: 42-48.

FOSTER, P. 2001. "Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers". En M. Bygate, P. Skehan y M. Swain (eds.), *Researching pedagogic tasks: Second language learning, teaching and testing*. (pp. 75-94). Londres: Longman.

FOTOS, J. 1931. "Word and idiom frequency counts in French and their value". *The Modern Language Journal*, 15, 5: 344-353.

FRANCIS, N. Y KUČERA, H. 1964/1979. *Manual of information to accompany 'A Standard Sample of Present-Day Edited American English, for Use with Digital Computers'*. Providence, R.I.: Department of Linguistics, Brown University. Disponible en: http://khnt.aksis.uib.no/icame/manuals/brown/ [Último acceso: 28.02.2009]

FRANCIS, N. Y KUČERA, H. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

FRARY R. B. 2000. "How difficult should a test be?" Disponible en http://www.testscoring.vt.edu/memo02.html [Última consulta: 27.03.2009]

FREEMAN, F. 1931. "Foreign Languages (Chapter VIII)". *Review of Educational Research*, 1, 5: 371-378.

FRIES, C. 1952. *The structure of English*. Nueva York: Harcourt, Brace and Co.

GAATONE, D. 1997. "La locution: Analyse interne et analyse global". En M. Martins-Baltar (ed.), *La locution entre langue et usages*. (pp. 165-177). Fontenay-Saint Cloud : ENS éditions.

GAIRNS, R. Y REDMAN, S. 1986. *Working with words: A guide to teaching and learning vocabulary*. Cambridge: Cambridge University Press.

GARCÍA ROLDÁN, J. L. 1995. *Cómo elaborar un proyecto de investigación*. Secretariado de Publicaciones. Universidad de Alicante.

GITSAKI, C. 1999. *Second language lexical acquisition: A study of the development of collocational knowledge*. San Francisco: International Scholars Publications.

GOULDEN, R., NATION, P. Y READ, J. 1990. "How large can a receptive vocabulary be?" *Applied Linguistics*, 11, 4: 341-363.

GRADDOL, D. 1997. *The future of English? A guide to forecasting the popularity of the English language in the 21st century*. Londres: The British Council.

GRAN DICCIONARIO OXFORD ESPAÑOL-INGLÉS/INGLÉS-ESPAÑOL, 3ª edición. (2003). Oxford: Oxford University Press.

GRANGER, S. 1998. "Prefabricated patterns in advanced EFL writing: Collocations and formulae". En A. Cowie (ed.), *Phraseology. Theory, analysis, and applications*. (pp. 145-160). Oxford: Oxford University Press.

GRANGER, S. Y PAQUOT, M. 2008. "Disentangling the phraseological web". En S. Granger y F. Meunier (eds.), *Phraseology. An interdisciplinary perspective*. (pp. 27-49). Ámsterdam/Filadelfia: John Benjamins.

GREENBAUM, S. 1970. *Verb-intensifier collocations in English. An experimental approach*. The Hague: Mouton.

GREENBAUM, S. 1974. "Some verb-intensifier collocations in American and British English". *American Speech*, 49, 1, 2: 79-89.

GROOT, P. 2000. "Computer assisted second language vocabulary acquisition". *Language Learning and Technology*, 4, 2: 60-81.

GYLLSTAD, H. 2007. *Testing English collocations. Developing receptive tests for use with advanced Swedish learners*. Tesis doctoral. Lund University.

HALLIDAY, M. A. 1961. "Categories of the theory of grammar". *Word*, 17: 241-292.

HALLIDAY, M. A. 1966. "Lexis as a linguistic level". En Ch. Bazell, J. Catford, M. Halliday y R. Robins (eds.), *In memory of J. R. Firth*. (pp. 148-162). Londres: Longman.

HALLIDAY, M. A. Y HASAN, R. 1976. *Cohesion in English*. Londres: Longman.

HANDL, S. 2008. "Essential collocations for learners of English. The role of collocational direction and weight". En F. Meunier y S. Granger (eds.), *Phraseology in foreign language learning and teaching*. (pp. 43-66). Ámsterdam/Filadelfia: John Benjamins.

HARGREAVES, P. 2000. "Collocation and testing". En M. Lewis (ed.), *Teaching collocation. Further developments in the lexical approach*. (pp. 205-223). Hove: Language Teaching Publications.

HARRIS, M. Y MCCANN, P. 1994 *Assessment*. Oxford: Heinemann

HASAN, R. 1984. "Coherence and cohesive harmony". En J. Flood (ed.), *Understanding reading comprehension. Cognition, language, and the structure of prose*. (pp. 181-219). Newark (International Reading Association).

HAUSMANN, F. 1979. "Un dictionnaire des collocations est-il possible?". *Travaux de Linguistique et de Littérature*, 17: 187-195.

HAUSMANN, F. 1989. "Le dictionnaire de collocations". En F. Hausmann, H. Wiegand y L. Zgusta (eds.), *Wörterbücher, Dictionaries, Dictionnaries. Ein internationales Handbuch zur Lexikographie*. (pp. 1010-1019). Berlín: Walter de Gruyter.

HAZENBERG, S. Y HULSTIJN, J. 1996. "Defining a minimal second-language vocabulary for non-native university students: An empirical investigation". *Applied Linguistics*, 17, 2: 145-163.

HEATON, J. B. 1989. *Writing English language tests*. Nueva York: Longman Group UK Limited.

HENNING, G. 1987. *A guide to language testing: Development, evaluation, research*. Cambridge, Mass.: Newbury House.

HERBST, T. 1996. "What are collocations: sandy beaches or false teeth?". *English Studies*, 4: 379-393.

HIGGINS, J. 1991. "Looking for patterns". En T. Johns y P. King (eds.), *Classroom concordancing*. (pp. 63-70). ELR Journal, 4. Birmingham: Birmingham University Press.

HILL, J. 2000. "Revising priorities: From grammatical failure to collocational success". En M. Lewis (ed.), *Teaching collocation. Further developments in the lexical approach*. (pp. 47-69). Hove: Language Teaching Publications.

HILL, J. Y LEWIS, M. (eds.). 1997. *LTP dictionary of selected collocations*. Hove: Language Teaching Publications.

HILL, J., LEWIS, M. Y LEWIS, M. 2000. "Classroom strategies, activities and exercises". En M. Lewis (ed.), *Teaching collocation. Further developments in the lexical approach*. (pp. 88-117). Hove: Language Teaching Publications.

HINDMARSH, R. 1980. *Cambridge English lexicon*. Cambridge: Cambridge University Press.

HIRSH, D. Y NATION, P. 1992. "What vocabulary size is needed to read unsimplified texts for pleasure?" *Reading in a Foreign Language*, 8, 2: 689-696.

HOEY, M. 2004. "Lexical priming and the properties of text". En A. Partington, J. Morley y L. Haarman (eds.), *Corpora and discourse*. (pp. 385-412). Berna: Peter Lang. Disponible en: http://www.monabaker.com/tsresources/LexicalPrimingandthePropertiesofText.htm [Último acceso: 27.01.2009]

HOEY, M. 2007. "What's in a word". Disponible en: http://www.onestopenglish.com/section.asp?docId=155130 [Último acceso: 26.01.2009]

HOWARTH, P. A. 1996. *Phraseology in English academic writing. Some implications for language learning and dictionary making*. Tübingen: Niemeyer.

HOWARTH, P. A. 1998a. "Phraseology and second language proficiency". *Applied Linguistics*, 19, 1: 24-44.

HOWARTH, P. A. 1998b. "The phraseology of learners' academic writing". En A. Cowie (ed.), *Phraseology. Theory, analysis, and applications*. (pp. 161-186). Oxford: Oxford University Press.

HOWATT, A. 1984. *A history of English language teaching*. Oxford: Oxford University Press.

HUGHES, A. 1989. *Testing for language teachers*. Cambridge: Cambridge University Press.

HUNSTON, S. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

HUNSTON, S. 2007. "Semantic prosody revisited". *International Journal of Corpus Linguistics*, 12, 2: 249-268.

HYLAND, K. 2008. "Academic clusters: Text patterning in published and postgraduate writing". *International Journal of Applied Linguistics*, 18, 1: 41-62.

JOHANSSON, S. 1980. "The LOB Corpus of British English texts: Presentation and comments". *ALLC Journal*, 1: 25-36.

JOHANSSON, S. 1998. "Review article of G. Kjellmer (1994) *A Dictionary of English Collocations: Based on the Brown Corpus*". *International Journal of Corpus Linguistics*, 3, 2: 338-348.

JOHANSSON, S. Y NORHEIM, E. 1988. "The subjunctive in British and American English". *ICAME Journal*, 12: 27-36.

JOHANSSON, S. Y HOFLAND, N. 1989. *Frequency Analysis of English Vocabulary and Grammar*. Oxford: Clarendon Press.

JOHNS, T. 1988. "Whence and whither classroom concordancing?". En T. Bongaerts, P. de Haan, S. Lobbe y H. Wekker (eds.), *Computer applications in language learning* (Foris). (pp. 9-33).

JOHNSON, C. 1927. "Vocabulary difficulty and textbook selection". *The Modern Language Journal*, 11, 5: 290-297.

JONES, S. Y SINCLAIR, J. 1974. "English lexical collocations". *Cahiers de Lexicologie*, 24: 15-61.

KALTENBÖCK, G. Y MEHLMAUER-LARCHER, B. 2005. "Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching". *ReCALL*, 17, 1: 65-84.

KATAMBA, F. 1994. *English words*. Londres: Routledge.

KELLER, E. 1979. "Gambits: Conversational strategy signals". *Journal of Pragmatics*, 3: 219-238.

KENNEDY, G. 1987. "Expressing temporal frequency in academic English". *TESOL Quarterly*, 21: 69-86.

KENNEDY, G. 1998. *An introduction to corpus linguistics*. Londres: Longman.

KENNEDY, G. 2007. "An under-exploited resource: Using the BNC for exploring the nature of language learning". En M. Hundt, N. Nesselhauf y C. Biewer (eds.), *Corpus linguistics and the web*. (pp. 151-166). Ámsterdam/Nueva York: Rodopi.

KENNEDY, G. 2008. "Phraseology and language pedagogy. Semantic preference associated with English verbs in the British National Corpus". En F. Meunier y S. Granger (eds.), *Phraseology in foreign language learning and teaching*. (pp. 21-41). Ámsterdam/Filadelfia: John Benjamins.

KESHAVARZ, M. H. Y SALIMI, H. 2007. "Collocational competence and cloze test performance: A study of Iranian EFL learners". *International Journal of Applied Linguistics*, 17, 1: 81-92.

KILGARRIFF, A. 1995. "BNC Database and Word Frequency Lists". http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html [Última consulta: 21.02.2009]

KILGARRIFF, A. 1997. "Putting frequencies in the dictionary". *International Journal of Lexicography*, 10, 2: 135-155.

KILGARRIFF, A., HUSÁK, M., MCADAM, K., RUNDELL, M. Y RYCHLÝ, P. 2008. "GDEX: Automatically finding good dictionary examples in a corpus". En *Proceedings of the XIII EURALEX International Congress*. Vol. 1. (pp. 425-432). Barcelona, julio 2008. Disponible en http://www.kilgarriff.co.uk/Publications/2008-KilgEtAl-euralex-gdex.doc [Última consulta: 27.02.2009]

KILGARRIFF, A. Y TUGWELL, D. 2001. "WORD SKETCH: Extraction and display of significant collocations for lexicography". En *Proceedings of the Collocations Workshop of the 39th Annual Meeting of the Association for Computational Linguistics*. (pp. 32-38). Toulouse, 2001. Disponible en: ftp://ftp.itri.bton.ac.uk/reports/ITRI-01-12.pdf [Último acceso: 04.01.2009]

KILGARRIFF, A., RYCHLÝ, P., SMRZ, P. Y TUGWELL, D. 2004. "The Sketch Engine". En *Proceedings of the Eleventh EURALEX International Congress*. (pp. 105-116). Lorient, julio 2004.

KJELLMER, G. 1984. "Some thoughts on collocational distinctiveness". En J. Aarts y W. Meijs (eds.), *Corpus linguistics*. (pp. 163-171). Ámsterdam: Rodopi.

KJELLMER, G. 1994. *A dictionary of English collocations: Based on the Brown Corpus*. Oxford: Clarendon Press.

KLEIN-BRALEY, C. Y RAATZ, U. 1984. "A survey of research on the C-test". *Language Testing*, 1: 134-146.

KOIKE, K. 2001. *Colocaciones léxicas en el español actual: Estudio formal y léxico-semántico*. Alcalá de Henares: Servicio de Publicaciones de la Universidad de Alcalá.

LADO, R. 1961. *Language testing: The construction and use of foreign language tests*. Londres: Longman.

LAUFER, B. 1989. "What percentage of text-lexis is essential for comprehension?". En C. Lauren y M. Nordman (eds.), *Special language: From humans thinking to thinking machines*. (pp. 316-323). Clevedon: Multilingual Matters.

LAUFER, B. 1992. "How much lexis is necessary for reading comprehension?" En P.J. Arnaud y H. Béjoint (eds.), *Vocabulary and applied linguistics*. (pp. 126-132). Londres: Macmillan.

LAUFER, B., ELDER, C., HILL, K. Y CONGDON, P. 2004. "Size and strength: do we need both to measure vocabulary?". *Language Testing*, 21: 202-226.

LAUFER, B. Y NATION, P. 1995. "Vocabulary size and use: Lexical richness in L2 written production". *Applied Linguistics*, 16, 3: 307-322.

LAUFER, B. Y NATION, P. 1999. "A vocabulary-size test of controlled productive ability". *Language Testing*, 16, 1: 33-51.

LEECH, G. 1993. "100 million words of English". *English Today*, 33, 9: 9-15.

LEECH, G. 1997. "Teaching and language corpora: A convergence". En A. Wichmann, S. Fligelstone, A. McEnery y G. Knowles (eds.), *Teaching and language corpora*. (pp. 1-23). Londres: Longman.

LEECH, G., RAYSON, P. Y WILSON, A. 2001. *Word frequencies in written and spoken English based on the British National Corpus*. Londres: Longman Pearson Education Limited.

LEHMANN, H. M., HOFFMANN, S. Y SCHNEIDER, P. 2002. *BNCweb*. Información disponible en: http://www.bncweb.info/ [Último acceso: 05.01.2009]

LEWIS, M. 1993. *The lexical approach*. Hove: Language Teaching Publications.

LEWIS, M. 2000a. "Introduction". En M. Lewis (ed.), *Teaching collocation. Further developments in the lexical approach*. (pp. 8-9). Hove: Language Teaching Publications.

LEWIS, M. 2000b. "Language in the lexical approach". En M. Lewis (ed.), *Teaching collocation: Further developments in the lexical approach*. (pp. 126-154). Hove: Language Teaching Publications.

LEWIS, M. (ED.). 2000c. *Teaching collocation: Further developments in the lexical approach*. Hove: Language Teaching Publications.

LONG, M. 1991. "Focus on form: A design feature in language teaching methodology". En K. de Bot, R. Ginsberg y C. Kramsch (eds.), *Foreign language research in cross-cultural perspective*. (pp. 39-52). Ámsterdam/Filadelfia: John Benjamins.

LÓPEZ-MEZQUITA MOLINA, M. T. 2005. *La evaluación de la competencia léxica: Tests de vocabulario. Su fiabilidad y validez*. Tesis doctoral. Universidad de Granada.

LÓPEZ-MEZQUITA MOLINA, M. T. 2007. *La evaluación de la competencia léxica: Tests de vocabulario. Su fiabilidad y validez*. Madrid: Ministerio de Educación y Ciencia.

LÓPEZ MORALES, H. 1993. *Sociolingüística*. Madrid: Gredos.

LORENZ, G. 1999. *Adjective intensification – learners versus native speakers: A corpus study of argumentative writing*. Ámsterdam: Rodopi.

LORGE, I. 1949. *The semantic count of the 570 commonest English words*. Nueva York: Columbia University Press.

LORGE, I. 1953. "The semantic count". En M. West (comp. y ed.), *A general service list of English words*. (pp. xi-xiii). Londres: Longmans, Green and Co.

LOUW, W. 1993. "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies". En M. Baker, G. Francis y E. Tognini-

Bonelli (eds.), *Text and technology: In honour of John Sinclair*. (pp. 157-176). Ámsterdam: John Benjamins.

LUQUE DURÁN, J. DE D. Y PAMIES BERTRÁN, A. 2005. *La creatividad en el lenguaje. Colocaciones idiomáticas y fraseología*. Granada: Método.

MADSEN, H. S. 1983. *Techniques in testing*. Oxford: Oxford University Press.

MALMKJÆR, K. (ed.). 2004. *The linguistic encyclopedia*. Londres: Routledge.

MANNING, C. Y SCHÜTZE, H. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

MARTIN, W., AL, B. Y VAN STERKENBURG, P. 1983. "On the processing of a text corpus: From textual data to lexicographical information". En R. Hartman (ed.), Lexicography, principles and practice. (pp. 77-87). Londres: Academic Press.

MARTON, W. 1977. "Foreign vocabulary learning as problem no. 1 of language teaching at the advanced level". *Interlanguage Studies Bulletin*, 2, 1: 33-57.

MCCARTHY, M. 1990. *Vocabulary*. Cambridge: Cambridge University Press.

MCCARTHY, M. Y O'DELL, F. 2005. *English collocations in use*. Cambridge: Cambridge University Press.

MCDONOUGH, J. Y MCDONOUGH, S. 1997. *Research methods for English language teachers*. Londres: Arnold.

MCENERY, A. 2006. *Swearing in English: Bad language, purity and power from 1586 to the present*. Londres: Routledge.

MCENERY, A., XIAO, R. Y TONO, Y. 2006. *Corpus-based language studies: An advanced resource book*. Londres/Nueva York: Routledge.

MCINTOSH, A. 1961. "Patterns and ranges". *Language*, 37: 325-337.

MCNAMARA, T. 2000. *Language testing*. Oxford: Oxford University Press.

MEARA, P. 1996. "The dimensions of lexical competence". En G. Brown, K. Malmkjær y J. Williams (eds.), *Performance and competence in second language acquisition*. (pp. 35-53). Cambridge: Cambridge University Press.

MEARA, P. Y BUXTON, P. 1987. "An alternative to multiple choice vocabulary tests". *Language Testing*, 4: 142-151.

MEL'ČUK, I. 1996. "Lexical functions: A tool for the description of lexical relations in a lexicon". En L. Wanner (ed.), *Lexical functions in lexicography and natural language processing*. (pp. 37-102). Ámsterdam/Filadelfia: John Benjamins.

MEL'ČUK, I. 1998. "Collocations and lexical functions". En A. P. Cowie (ed.), *Phraseology. Theory, analysis, and applications*. (pp. 23-53). Oxford: Oxford University Press.

MITCHELL, T. F. 1966. "Some English phrasal types". En C. Bazell, C. Catford, M. Halliday y R. Robins (eds.), *In memory of J. R. Firth*. (pp. 335-358). Londres: Longmans.

MITCHELL, T. F. 1971. "Linguistic 'goings-on': Collocations and other lexical matters arising on the syntactic record". *Archivum Linguisticum*, 2: 35-69.

MOCHIZUKI, M. 2002. "Exploration of two aspects of vocabulary knowledge: Paradigmatic and collocational". *Annual Review of English Language Education in Japan*, 13: 121-129.

MOON, R. 1998a. *Fixed expressions and idioms in English. A corpus-based approach*. Oxford: Oxford University Press.

MOON, R. 1998b. "Frequencies and forms of phrasal lexemes in English". En A. Cowie (ed.), *Phraseology. Theory, analysis, and applications*. (pp. 79-100). Oxford: Oxford University Press.

MORENO JAÉN, M. 2006. "Measuring lexical difficulty in L2 reading by means of a new electronic device: ADA (ADELEX ANALYSER)". *International Journal of Technology, Knowledge and Society*, 2, 5: 19-32.

MORGAN, B. 1930. "A German frequency word book". *Publications of the American and Canadian Committees on Modern Languages*, vol. IX. Nueva York: Macmillan Company.

MURRAY, J. 2002. "Creating placement tests". Disponible en http://www.eslmag.com/novdec02art.html [Última consulta: 23.03.2009]

NATION, P. 1983. "Testing and teaching vocabulary". *Guidelines*, 5: 12-25.

NATION, P. 1990. *Teaching and learning vocabulary*. Rowley, MA: Newbury House.

NATION, P. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

NATION, P. Y WARING, R. 1997. "Vocabulary size, text coverage and word lists". En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, acquisition and pedagogy*. (pp. 6-19). Cambridge: Cambridge University Press.

NESSELHAUF, N. 2004. "What are collocations?". En D. Allerton, N. Nesselhauf y P. Skandera (eds.), *Phraseological units: Basic concepts and their application*. (pp. 1-29). Basel: Schwabe.

NESSELHAUF, N. 2005. *Collocations in a learner corpus*. Ámsterdam/Filadelfia: John Benjamins.

NESSELHAUF, N. Y TSCHICHOLD, C. 2002. "Collocations in CALL: An investigation of vocabulary-building software for EFL." *Computer Assisted Language Learning*, 15, 3: 251-279.

NUCCORINI, S. 2003. "Towards an "ideal" dictionary of English collocations". En P. van Sterkenburg (ed.), *A practical guide to lexicography*. (pp. 366-387). Ámsterdam/Filadelfia: John Benjamins.

OAKES, M. 1998. *Statistics for corpus linguistics*. Edimburgo: Edinburgh University Press.

OPPENHEIM, H. 2000. "The importance of recurrent sequences for nonnative speaker fluency and cognition". En H. Riggenbach (ed.), *Perspectives on fluency*. (pp. 220-240). Ann Arbor: University of Michigan Press.

PALLANT, J. 2007. *SPSS survival manual: A step-by-step guide to data analysis with SPSS*. Crows Nest, NSW: Allen & Unwin.

PALMER, H. 1917. *The scientific study and teaching of languages*. Londres/Nueva York: Word Book Company.

PAQUOT, M. 2007. *EAP vocabulary in EFL learner writing: From extraction to analysis: A phraseology-oriented approach*. Tesis doctoral no publicada. Université Catholique de Louvain.

PARTINGTON, A. 1998. *Patterns and meanings*. Ámsterdam/Filadelfia: John Benjamins.

PARTINGTON, A. 2004. "'Utterly content in each other's company': semantic prosody and semantic preference". *International Journal of Corpus Linguistics*, 9, 1: 131-156.

PARTINGTON, A. 2006. "Aims, tools and practices of corpus linguistics". *IntUne Papers*, ME-06-05. Disponible en: www.intune.it/file_download/23 [Último acceso: 04.01.2009]

PARTINGTON, A. 2007. "The armchair and the machine: Corpus-assisted discourse studies". Ponencia presentada en el Seminario Internacional "Corpora: Seminar and Workshops", Padua, marzo 2007.

PAWLEY, A. Y SYDER, F. 1983. "Two puzzles for linguistic theory: nativelike selection and nativelike fluency". En J. Richards y R. Schmidt (eds.), *Language and communication*. (pp. 191-226). Londres: Longman.

PAZOS BRETAÑA, J. M. 2005. *Detección automatizada de fraseologismos*. Tesis doctoral no publicada. Universidad de Granada.

PEARCE, M. 2008. "Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine". *Corpora*, 3, 1: 1-29.

PÉREZ BASANTA, C. 1995. "Coming to grips with progress testing: some guidelines for its design". *English Teaching Forum*, l33, 3: 55-58.

PÉREZ BASANTA, C. 1999. "La enseñanza del vocabulario desde una perspectiva lingüística y pedagógica". En S. Salaberri (ed.), *Lingüística aplicada a la enseñanza de lenguas extranjeras*. (pp. 262-306). Almería: Universidad de Almería.

PÉREZ BASANTA, C. 2005. "Assessing the receptive vocabulary of Spanish students of English philology: An empirical investigation". En J.M. Martínez-Dueñas Espejo, N. McLaren, C. Pérez Basanta y L. Quereda Rodríguez-Navarro (eds.), *Towards an understanding of the English language: Studies in honour of Fernando Serrano*. (pp. 456-477). Granada: Universidad de Granada.

PÉREZ BASANTA, C., PINILLA PEINADO, A. Y GARCÍA JIMÉNEZ, J. 1992. "Fundamentaciones teóricas del *testing*". En J. A. Martínez López (ed.), *Actas de las VIII Jornadas Pedagógicas para la Enseñanza del Inglés*. Granada: Greta.

PÉREZ FERNÁNDEZ, A. 2002. *Las colocaciones léxicas de los adverbios intensificadores ingleses. Propuesta de un diccionario*. Tesis doctoral no publicada. Universidad de Jaén.

PHILIP, G. 2007a. "'...and I dropped my jaw with fear' – The role of corpora in teaching phraseology". Comunicación presentada en la 7th Teaching and Language Corpora Conference (TALC7), París, julio 2006. Disponible en: http://amsacta.cib.unibo.it/archive/00002361/01/TaLC06.pdf [Último acceso: 18.12.2008]

PHILIP, G. 2007b. "Decomposition and delexicalisation in learners' collocational (mis)behaviour". En *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham, julio 2007. (pp. 1-9) Disponible en http://www.corpus.bham.ac.uk/corplingproceedings07/paper/170_Paper.pdf [Último acceso: 11.01.2009]

PRENDERGAST, T. 1864. *The mastery of languages; or, the art of speaking foreign tongues idiomatically*. Londres: Richard Bentley.

RADFORD, A. 1997. *Syntax: A minimalist introduction*. Cambridge: Cambridge University Press.

READ, J. 1993. "The development of a new measure of L2 vocabulary knowledge". *Language Testing*, 10, 3: 355-371.

READ, J. 1997. "Vocabulary and testing". En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, acquisition and pedagogy*. (pp. 303-320). Cambridge: Cambridge University Press.

READ, J. 2000. *Assessing vocabulary*. Cambridge: Cambridge University Press.

READ, J. Y CHAPELLE, C. 2001. "A framework for second language vocabulary assessment". *Language Testing*, 18, 1: 1-32.

REAL ACADEMIA ESPAÑOLA. 2001. *Diccionario de la lengua española*, 22ª edición. Madrid: Espasa Calpe. Disponible en: www.rae.es [Último acceso: 29.01.2009]

REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual*. Disponible en: http://www.rae.es [Último acceso: 28.02.2009]

RENOUF, A. 1987. "Corpus development". En J. Sinclair (ed.), *Looking up. An account of the Cobuild Project in lexical computing*. (pp. 1-40). Londres: Collins ELT.

RICHARDS, J. 1976. "The role of vocabulary teaching". *TESOL Quarterly*, 10, 1: 77- 89.

RICHARDS, J. Y RODGERS, T. 1986. *Approaches and methods in language teaching: A description and analysis*. Nueva York: Cambridge University Press.

RIVERS, W. 1983. *Communicating naturally in a second language: Theory and practice in language teaching*. Nueva York: Cambridge University Press.

ROMAINE, S. 2000. *Language in society: An introduction to sociolinguistics* (2ª edición). Oxford: Oxford University Press.

ROOS, E 1976. "Contrastive collocational analysis". *Papers and Studies in Contrastive Linguistics,* 5: 65-75.

RUNCIE, M. 2002. *Oxford collocations dictionary of students of English*. Oxford: Oxford University Press.

SÁNCHEZ, A. 2001. *Gran diccionario de uso del español actual*. Madrid: Sociedad General Española de Librería.

SCHMID, H. J. 2003. "Collocation: hard to pin down, but bloody useful". *Zeitschrift für Anglistik und Amerikanistik*, 51, 3: 235-258.

SCHMITT, N. 1999. "The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge". *Language Testing*, 16, 2: 189-216.

SCHMITT, N. 2000. *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

SCHMITT, N. (ed.). 2004. *Formulaic sequences: Acquisition, processing and use.* Ámsterdam/Filadelfia: John Benjamins.

SCHMITT, N., SCHMITT, D. Y CLAPHAM, C. 2001. "Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test". *Language Testing*, 18, 1: 55-88.

SCHMITT, N. Y UNDERWOOD, G. 2004. "Exploring the processing of formulaic sequences through a self-paced reading task". In N. Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use.* (173-190). Ámsterdam/Filadelfia: John Benjamins.

SCOTT, M. 1999. *WordSmith Tools*. Información disponible en: http://www.lexically.net/wordsmith/ [Último acceso: 30.12.2008]

SECO, M. 1978. "Problemas formales de la definición lexicográfica". En M. V. Conde et al. (eds.), *Estudios ofrecidos a Emilio Alarcos Llorach*. (pp. 217-239). Oviedo: Universidad de Oviedo.

SECO, M., ANDRÉS, O. Y RAMOS, G. 2005. *Diccionario del español actual*, 2 vol. Madrid: Aguilar.

SERETAN, V. 2008. *Collocation extraction based on syntactic parsing*. Tesis doctoral. Université de Génevè. Disponible en: http://doc.rero.ch/lm.php?url=1000,40,3,20080918145728-JU/these.pdf [Último acceso: 06.01.2009]

SHANNON, C. Y WEAVER, W. 1949/1972. *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

SHIN, D. Y NATION, P. 2008. "Beyond single words: The most frequent collocations in spoken English". *ELT Journal*, 62, 4: 339-348.

SIEVERS, M. 2008. *Day, money, way – A corpus-based investigation of the phraseology of three high frequency nouns and its implications for the design of TEFL materials*. Hannover: Grin.

SINCLAIR, J. 1966/1996. "Beginning the study of lexis". En J. Foley (ed.), J.M. Sinclair on lexis and lexicography. (1-20). Singapur: Unipress.

SINCLAIR, J. 1987. *Looking up: An account of the COBUILD project in lexical computing*. Londres: Collins.

SINCLAIR, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

SINCLAIR, J. 1996. "The search for units of meaning". *Textus*, IX: 75-106.

SINCLAIR, J. 1998. "The lexical item". En E. Weigand (ed.), *Contrastive lexical semantics*. (pp. 1-24). Ámsterdam: John Benjamins.

SINCLAIR, J. (ed.). 2001. *Collins COBUILD English Dictionary for Advanced Learners*. (Tercera edición). HarperCollins Publishers.

SINCLAIR, J. 2003. *Reading concordances*. Londres: Pearson/Longman.

SINCLAIR, J. 2004. *Trust the text: Language, corpus and discourse*. Londres: Routledge.

SINCLAIR, J. 2008. "Preface". En S. Granger y F. Meunier (eds.), *Phraseology. An interdisciplinary perspective*. (pp. xv-xviii). Ámsterdam/Filadelfia: John Benjamins.

SINCLAIR, J., JONES, S. Y DALEY, R. 1970/2004. *English collocations studies: The OSTI report*. 2ª edición, R. Krishnamurthy (ed.). Londres/Nueva York: Continuum.

SINCLAIR, J. Y RENOUF, A. 1988. "A lexical syllabus for language learning". En R. Carter y M. McCarthy (eds.), *Vocabulary and language teaching*. (pp. 140-160). Londres: Longman.

SMADJA, F. 1993. "Retrieving collocations from text: Xtract". *Computational Linguistics*, 19, 1: 143-177.

SMITH, R. (ed.). 2003. *Teaching English as a foreign language, 1912-1936: Pioneers of ELT*. Londres: Routledge.

STUBBS, M. 1995. "Collocations and semantic profiles: On the cause of the trouble with quantitative studies". *Functions of Language*, 2, 1: 23-55.

STUBBS, M. 2001. *Words and phrases. Corpus studies of lexical semantics*. Oxford: Blackwell Publishers.

SUMMERS, D. 1991. *Longman/Lancaster English Language Corpus: Criteria and design*. Harlow: Longman.

SWEET, H. 1899/1964. *The practical study of languages: A guide for teachers and learners*. Londres: Oxford University Press.

TEJADA FERNÁNDEZ, J. 1997. *El proceso de investigación científica*. Barcelona: Fundación La Caixa.

THORNDIKE, R. L. Y HAGEN, E. 1980. *Tests y técnicas de medición en psicología y educación*. Méjico: Editorial Trillas.

THORNDIKE, E. Y LORGE, I. 1944. *The teacher's word book of 30,000 words*. Nueva York: Teachers College, Columbia University.

TOGNINI-BONELLI, E. 2001. *Corpus linguistics at work*. Ámsterdam/Filadelfia: John Benjamins.

TOGNINI-BONELLI, E. 2008. "Phraseology as the starting point in a corpus-driven perspective". Conferencia plenaria ofrecida en el Seminario Internacional "New Trends in Corpus Linguistics for Language Teaching and Translation Studies. In Honour of John Sinclair". Granada, septiembre 2008.

TRIBBLE, C. 1997. "Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching". En J. Melia y B. Lewandowska-Tomaszczyk (eds.), *PALC'97 proceedings*. Lodz: Lodz University Press.

Disponible en http://www.ctribble.co.uk/text/Palc.htm [Último acceso: 22.02.2009]

VAN DER WOUDEN, T. 2002. "Particle research meets Corpus Linguistics: On the collocational behaviour of particles". *Belgian Journal of Linguistics*, 16, 1: 151-174.

VAN ROEY, J. 1978. "Collocation in lexical analysis". En J. Lerot y R. Kern (eds.), *Mélanges de linguistique et de littérature*. (pp. 155-162). Universidad de Lovaina: Recueil de Travaux d´Histoire et de Philosophie, 14.

WEINREICH, U. 1963. "Lexicology". En T. A. Sebeok (ed.), *Current trends in linguistics, Vol. I*. (pp. 60-93). La Haya: Mouton de Gruyter.

WEIR, C. 1990. *Communicative language testing*. Nueva Jersey: Prentice Hall.

WESCHE, M. Y PARIBAKHT, S. 1996. "Assessing L2 vocabulary knowledge: Depth versus breadth". *Canadian Modern Language Review*, 53: 13-40.

WEST, M. 1930. "Speaking-vocabulary in a foreign language". *The Modern Language Journal*, 14, 7: 509-521.

WEST, M. 1937. "The present position in vocabulary selection for foreign language teaching". *The Modern Language Journal*, 21, 6: 433-437.

WEST, M. (comp. y ed.). 1953. *A general service list of English words*. Londres: Longmans, Green and Co.

WIDDOWSON, H. 2000. "On the limitations of linguistics applied". *Applied Linguistics*, 21, 1: 3-25.

WILLIS, D. 1990. *The lexical syllabus: A new approach to language teaching*. Londres: Collins.

WOOLARD, G. 2000. "Collocation: Encouraging learner independence". En M. Lewis (ed.), *Teaching collocation. Further developments in the lexical approach*. (pp. 28-46). Hove: Language Teaching Publications.

WOOLARD, G. 2005a. *Key words for fluency. Learning and practising the most useful words of English*. Londres: Thomson.

WOOLARD, G. 2005b. "Noticing and learning collocation". *English Teaching Professional*, 40: 46-48.

WRAY, A. 1999. "Formulaic language in learners and native speakers". *Language Teaching*, 32, 4: 213-231.

WRAY, A. 2000. "Formulaic sequences in second language teaching: principle and practice". *Applied Linguistics*, 21, 4: 463-489.

WRAY, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

XIAO, R. Y MCENERY, A. 2006. "Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective". *Applied Linguistics*, 27, 1: 103-129.

YORIO, C. A. 1989. "Idiomaticity as an indicator of second language proficiency". En K. Hyltenstam y L. K. Obler (eds.), *Bilingualism across the lifespan*. (pp. 55-72). Cambridge: Cambridge University Press.

ZANETTIN, F. 1998. "Bilingual comparable corpora and the training of translators". *Meta*, 43, 4: 616-630. Disponible en: http://www.erudit.org/revue/meta/1998/v43/n4/004638ar.pdf [Último acceso: 25.01.2009]

ZIMMERMAN, C. 1997. "Historical trends in second language vocabulary instruction". En J. Coady y T. Huckin (eds.), *Second language vocabulary acquisition*. (pp. 5-19). Cambridge: Cambridge University Press.

ZIPF, G. 1932. *Selected studies of the principle of relative frequency in language*. Cambridge (Mass.): MIT Press.

ZIPF, G. 1935. *The psycho-biology of language*. Cambridge (Mass.): MIT Press.

APÉNDICES

Apéndice 3: Listado de frecuencias de colocaciones definitivo

Apéndice 4: Test de Colocaciones ADELEX v.1

ANEXO 3

LISTADO DE FRECUENCIAS DE COLOCACIONES DEFINITIVO

1 time
- long
- spend
- short
- waste
- take
- spare
- give
- present
- working

2 year
- past
- coming
- full
- start
- take
- spent

3 people
- ordinary
- working
- elderly
- disabled

4 way
- long
- go
- find
- get
- make (someone's N)
- come
- easy

5 ways
- separate
- find
- alternative
- explore
- develop
- subtle

6 man
- big
- tall
- small
- dark

7 day
- all (N long)
- present
- spent
- final
- full
- working
- whole
- start
- (have a) nice

8 days
- early
- past (few/number)
- old
- spent
- working
- take

9 thing
- whole
- (do the) right
- real
- (do the) wrong

10 child
- only
- expecting

11 government
- coalition
- interim
- elected

12 part
- take (part)
- play (the/a)
- integral
- large
- major
- essential
- small

vital
form
significant
full

13 parts

spare
component
constituent
played
separate

14 life

real
family
everyday
whole
private
normal
live
saved
personal
public
early
spent
daily
lead

15 lives

save
past
live
lost
private
claimed (a number of)
risk
lead
cost
daily
spend
everyday
separate
ordinary
normal

16 case

court
present
particular
(a N) adjourned
prosecution
made
put
take
clear
strong
extreme

17 cases

severe
court
extreme
serious
criminal
injury

18 woman

beautiful
attractive
pretty
tall
(good) looking

19 work

hard
do
started
voluntary
charity
carried (out)
find
paid
major
community

20 system

welfare

21 group

small
ethnic (minority N)
largest
formed
minority
leading
single
research

22 number

large
small
increasing

growing
limited
total
increased
significant
greater
reduce
considerable
substantial
high

23 world
real
whole
modern
developing

24 area
particular
play
surrounding
residential
wide
central
major

25 course
(IN) due
training
main
designed
first
take
completed
ordinary
followed
full
start

26 company
owned

27 problem
solve
real
major
serious
big
particular
lies
tackle
overcome
common
faced
arises
cause
resolve
sort (out)
address

28 service
free
telephone
provide
memorial
intelligence
advisory
delivery
full
postal
security
regular

29 hand
held
took
shook
wash
raised
waved
lifted
helping
stretched out (his/her…+N)
extended
strong

30 hands
shook
holding
clasped
wash
raised
rubbed
folded

31 party
ruling
opposition
led

32 school
- left
- medical
- went
- whole

33 place
- take
- wrong
- (final) resting (place)
- safe
- right
- put
- secure

34 point
- made
- reached
- whole
- prove
- raised

35 house
- detached
- family
- moved +N

36 country
- whole
- leave
- foreign

37 week
- past
- spent
- previous
- whole
- coming

38 member
- family
- former
- senior
- leading
- active
- elected
- union
- full

39 end
- west
- east
- came
- put
- far
- brought (to an N)
- reached
- bottom

40 ends
- loose
- achieve
- opposite

41 word
- use
- say
- spread
- key
- final

42 example
- set
- give
- classic
- perfect
- take (the)
- typical
- prime
- follow
- provides
- simple
- obvious
- recent
- extreme
- striking
- outstanding
- notable
- specific

43 family
- whole
- extended
- close
- entire
- large
- nuclear
- big

44 families
- (lone/one/ single) parent
- one-parent
- lone
- single-parent
- low-income

45 fact
- actual
- simple
- N + remains
- mere
- face

46 facts
- relevant
- stated
- known
- hard
- basic
- true
- face
- full

47 state
- mental
- declared
- announced

48 home
- come
- family
- stay
- return
- left
- ideal
- get
- made
- feel
- matrimonial

49 month
- past
- spent

50 side
- (left-right) hand
- wrong
- dark

51 sides
- opposite
- N + agreed
- take

52 night
- spent
- stay
- opening
- dark
- early

53 nights
- sleepless
- spent

54 eye
- keep
- public
- watchful
- naked
- cast
- look (in the N)
- twinkle
- critical

55 eyes
- wide
- fixed
- bright
- rolled
- turned
- filled (with)
- pale
- narrowed
- raised
- held

56 head
- shook
- turned
- bowed
- raised
- bent
- lifted
- held (N high)
- jerked
- lowered
- nodded
- clear
- pulled (N back)

threw (back)
tilted

57 information
further
provide
detailed
additional
relevant
obtained
confidential
exchange
collected
supplied
contain
background
valuable
gathering

58 question
answer
ask
simple
raised
N + remains
whole
posed
real
address
begs
crucial
central
relevant
fundamental

59 business
small
doing
whole
family
running
serious
private
conduct
successful

60 power
use
labour
real
seized
exercise
increase
vested
full
held

61 money
spend
save
get
make
raise
waste
earn
give
(pay) good

62 change
significant
radical
major
global
make
dramatic
fundamental
rapid
sudden
real
subject to
N + occurred

63 changes
make
major
proposed
dramatic
occur
radical
significant
sweeping
necessary
introduced
minor
rapid

64 interest
public
special
lower
high
mortgage

real
fixed
particular
shown
take (an N in)
expressed
vested
common
strong

65 interests
(in the) best
protect
outside
vested
varied
serve
pursue
defend

66 order
make
keep
maintain
standing
restore
pecking
rigid
working
good
put
established
take

67 book
latest
illustrated
reference

68 development
child
sustainable
curriculum
career
further
rapid

69 room
double
single
leave
main
TV

70 water
fresh
deep
running
clean
salt
still
shallow

71 form
application
order
booking
complete (verb)
claim
fill
registration
take
(in) good

72 car
park (verb)
hire
drive
police
racing
crashed

73 level
high
low
top
reached
intermediate
current
water
achieved

74 policy
foreign
monetary
common
made
adopted

75 line
- direct
- railway
- draw (verb)
- straight
- crossed
- fine
- take
- dividing

76 need
- urgent
- feel
- desperate
- meet
- stressed
- N + arises

77 needs
- meet (verb)
- individual
- suit (verb)
- satisfy
- strong

78 effect
- have
- cause+ N
- take
- knock-on
- direct
- immediate
- dramatic
- desired
- beneficial
- profound
- devastating
- opposite
- overall
- adverse
- detrimental

79 effects
- long-term
- harmful
- adverse
- direct
- beneficial
- damaging
- profound

80 use
- drug
- make
- good (make good/full N of)
- widespread
- full
- extensive

81 uses
- common

82 idea
- whole
- give (an)
- get
- rejected
- original
- faintest
- simple
- wrong
- put (forward/aside the + N)
- form
- developed

83 ideas
- exchange
- original
- fresh
- bright

84 study
- present + N
- N + found
- recent
- pilot
- showed
- carried
- conducted
- research
- major
- reported
- undertaken
- designed
- significant

85 lot
- awful
- whole

86 job
- do
- good
- get
- part-time
- lost
- find
- quit
- full-time
- finish
- leave
- proper

87 jobs
- create
- lost
- temporary
- do
- odd
- save
- shed (number)
- secure
- permanent
- unskilled
- full-time

88 name
- real
- given
- put
- full
- proper
- second
- last

89 names
- give
- call (someone's N)

90 result
- direct
- final
- net
- achieve
- inevitable
- produce
- overall
- further
- poor
- significant
- desired

91 body
- upper
- dead
- N + lay
- naked
- lower

92 friend
- best
- close
- old
- family
- made N in

93 friends
- family
- close
- old
- make
- became
- real

94 right
- vote + N

95 rights
- pension
- full
- minority

96 authority
- given (give + N)
- exercise + N or someone's + N
- legitimate

97 view
- take (the + N)
- shared
- full
- expressed
- support
- give
- clear
- held
- (give) a true (and fair)
- accept

98 report
- full
- final

N+ released
issued

99 bit
wee (Scottish)

100 face
pale
slap + (someone's) + N
pretty
turned
round
flushed
beautiful
smiling
handsome

101 market
free
housing
major

102 hour
spent
took
late
early
N + passed
finest

103 hours
spent
early (N of the morning)
working
long
took
permitted
small
normal
daylight
school
put in
late
flying

104 rate
fixed
Inflation
low

105 law
breaking
passed
enforce

106 door
knock
slammed
N + swung (open)
pushed
pulled
flung (open) N

107 court
told (the N)
heard
ruled
appear (in)
held
take obj to

108 office
took
held

109 war
end
declared
N+ broke (out)
fought

110 reason
(for the) simple
(there is) real
(for no) apparent
given a reason
major
sufficient
(for some) unknown

111 minister
appointed
junior
elected

112 person
missing
average
single
disabled
elderly

authorised

113 terms
practical
general
agreed (to the N)
broad
(on) equal
(in) absolute
accept
relative

114 sorts
all

115 period
minimum
extended
trial
limited
prolonged
considerable
maximum

116 society
modern
whole
classless
high

117 process
whole
involved
decision-making
continuous
complete
gradual
lengthy

118 mother
single
foster

119 voice
low
loud
soft
deep
quiet
high
lowered
raised
clear
husky
harsh
hoarse
cracked
sharp

120 police
called
questioned

121 price
high
retail
average
low
increase
expected
full
fixed
reasonable
fair
put
reduced
competitive

122 action
take
further
call for N
immediate
strike
direct
appropriate
urgent
drastic
libel
joint
remedial

123 issue
big
major
raised (an N)
key
sensitive
central
take N
crucial
fundamental

controversial
serious
complex
N was resolved

124 position
strong
hold
difficult
marginal
dominant
placed (in)
maintain
privileged
competitive
leading
strengthen
reached

125 cost
low
effective
extra
total
high
cover
estimated
additional
meet
reduced
reasonable

126 matter
simple
further
serious
referred
raised
resolved

127 figure
leading
tall
central
high
public
major
low
increase

128 figures
released
official
leading
public

129 type
particular
certain
common
media

130 research
shows
done
suggests
further
carried
conducted
future

131 education
further
public

132 programme
major
full
launched
extensive
current
completed
run
ambitious

133 minute
hang
take
hold
full

134 moment
right
long
paused
took
brief
precise

135 girl
- beautiful
- pretty

136 age
- early
- golden
- retirement
- maternal
- reached

137 ages
- dark

138 centre
- entertainment
- (day) care

139 control
- take
- lost
- have
- complete
- total
- gain
- regain
- full
- keep
- effective
- direct
- exercise
- tight
- retain
- strict
- overall
- maintain

140 value
- good N (for money)
- great
- added
- total
- real
- high
- equal
- increase
- full
- add
- true
- estimated
- critical
- fair

141 health
- mental
- poor
- medical
- improve

142 decision
- made
- final
- taken
- court
- announced
- conscious
- unanimous
- follows
- reached
- support
- came
- reversed

143 force
- use
- military
- driving
- peacekeeping
- excessive
- full
- dominant
- powerful
- strong
- armed
- brute

144 forces
- security
- joined
- opposition

145 condition
- good
- serious
- medical
- critical
- stable
- excellent
- poor
- immaculate
- sufficient
- deteriorated

perfect

146 conditions
- weather
- living
- working
- improve
- appalling
- poor
- satisfied (verb)
- favourable
- impose
- prevailing
- adverse
- meet
- fulfilled

147 paper
- printed

148 section
- previous
- final
- separate
- further

149 patients
- treated

150 road
- main
- cross
- busy
- major
- line (the N with people)

151 table
- league
- lay the N

152 church
- whole
- established

153 mind
- bear
- keep
- open (adj)
- had
- crossed
- conscious
- came (to/into N)

154 team
- dream
- national
- winning
- N + won
- led
- beat
- support + N

155 experience
- previous
- practical
- gain
- learning
- relevant
- working
- extensive
- past
- direct
- valuable

156 death
- sudden
- accidental
- caused
- tragic

157 act
- criminal
- passed
- amended

158 sense
- make
- strong
- real
- false
- true

159 staff
- trained
- support
- extra
- permanent
- qualified
- experienced

160 student
- former

161 language
- foreign
- use
- speak
- modern
- native
- official

162 morning
- early
- late
- spent
- bright

163 plan
- peace
- action
- radical
- approved
- strategic
- drawn up
- adopted
- ambitious
- prepare

164 product
- final
- make
- launch+ N
- competitive
- created

165 city
- capital
- ancient
- beautiful

166 ground
- common
- lost
- open
- gaining

167 letter
- quick
- post
- business
- resignation

168 evidence
- give
- N + suggests
- presented
- hard
- provide
- strong
- N + shows
- sufficient
- conclusive
- considerable
- supporting
- ample
- circumstantial
- insufficient
- growing
- indicates

169 foot
- set

170 feet
- bare
- rose (to her N)

171 boy
- bad
- small
- good
- big

172 game
- win
- final

173 food
- fresh
- delicious
- diet

174 role
- play
- key
- leading
- major
- active
- central
- vital
- starring
- crucial

take
traditional
supporting
prominent
passive
essential
minor

175 practice
common
private (doctors)
medical
classroom

176 bank
collapsed

177 support
strong
public
provide
received
family
technical
giving
emotional
substantial

178 event
major
unlikely
sporting
big
held
annual
history
final

179 building
main
tall
house

180 range
wide
long
full
offers
free
broad
vast
complete

181 stage
early
final
reached
retirement
further

182 meeting
annual
held
summit
emergency
called
arranged
adjourned

183 art
modern
contemporary
great
ancient

184 club
join
exclusive

185 arm
put (someone's N around)
upper
grabbed
broken
broke
raised
grasped

186 history
family
medical

187 parent
single
elderly

188 parents
foster
adoptive
working
single

189 land
- housing
- held
- open
- work (verb)
- native
- farming
- waste

190 trade
- world
- foreign
- retail

191 situation
- current
- present
- improve
- desperate
- remedy
- impossible
- intolerable

192 teacher
- former
- primary (school)

193 record
- world
- set the record
- breaking
- unbeaten
- criminal
- keeping
- poor
- impressive
- held
- achieved

194 relation
- bear
- close

195 field
- visual

196 window
- front
- back
- smashed
- rear

197 account
- taken (into)
- opened

198 difference
- make
- big
- tell
- little
- see
- significant
- potential
- real
- fundamental
- marked (adj)
- showed

199 material
- raw

200 marerials
- building

201 air
- fresh
- open
- clean
- clear
- breathe
- thick
- heavy

202 wife
- second
- left + N
- former
- estranged

203 project
- whole
- completed
- complete
- ambitious
- run

204 sales
- retail
- increased
- record
- rose

annual
boost
total
direct
potential
expected

205 relationship
close
loving
special
working
lasting
love
serious
direct
strong
establish
developed
inverse
causal
intimate
to end
see
public
maintain

206 light
bright
shed
natural
low
throw
switch (on/off)
turned (on/off)
strong
cast
flashing
soft
cold
shining
pale
full
fading
flickering
clear

207 care
take
medical
primary
private
home
nursing
provide
family
domiciliary

208 rule
(as a) general
golden
established

209 story
tell
true
full
main
real
big

210 quality
high
good
top
poor
improve
excellent
air
superb
low
total

211 tax
increases
gross

212 nature
human
N+ calls

213 structure
family

214 data
collected
suggest
support
present

215 method
use

payment
effective
simple
alternative
traditional
adopted
applied
preferred
provide

216 unit
family
hospital

217 bed
get (into/out of/off the N)
lying
double
(four) poster
go
single
make

218 union
join

219 movement
rights N

220 detail
greater
(in) fine
considerable
sufficient
further

221 details
further
full
give
provide

222 model
following (verb)
developed
top

223 computer
running
using

224 hospital
general
(to be) admitted (to N)
mental
go
transferred

225 chapter
discussed
previous
contains
examines
preceding
covered

226 scheme
training
launched
successful
designed
devised

227 theory
critical

228 property
private
commercial
land
owned
(to) purchase

229 officer
chief
army
prison

230 charge
free (of)
take
extra
additional
initial
minimum
made
put (in/on)
fixed

231 charges
bank

232 director
general
film
appointed
theatre

233 approach
cautious
adopted
taken
alternative
made
fresh
radical
effective

234 chance
have
get
given
stand a N
take
great
second
little
missed
offered
big
fair
came
reasonable
jumped (at the N)

235 application
complete
made

236 amount
certain
small
enormous
huge
fair
tremendous
large
considerable
minimum
full
reasonable
maximum
substantial
increased
outstanding

237 son
eldest
baby
oldest
illegitimate

238 operation
rescue
police
major
successful
performed
underwent
military
emergency

239 opportunity
given
equal
take
provide
unique
golden
missed
perfect
offered
rare
ideal
seize
the N presented itself
excellent
ample
real
arose

240 leader
opposition

241 look
take
have
quick
closer
hard
gave
shot
(take a) fresh N (at)

242 production
- increase
- stage
- theatre

243 picture
- exclusive
- complete
- taken
- shows
- clear
- overall
- N + emerges
- accurate
- mental
- presented
- similar
- full

244 source
- major
- single
- primary
- valuable
- constant
- principal

245 security
- top
- maximum
- increased

246 contract
- signed
- expires
- won
- awarded
- agreed
- binding
- terminate

247 agreement
- reached
- signed
- peace
- (free) trade
- co-operation
- nodded (in) N
- concluded
- ceasefire
- entered (into)
- come
- draft
- mutual
- final
- N + provided (for)
- broad
- failed (to reach N)
- end
- make
- (in) full
- voluntary
- binding
- common
- complete

248 site
- (a+ number) acre + N
- caravan
- archaeological
- suitable
- construction

249 labour
- female
- temporary

250 test
- fitness
- passed
- final
- failed
- have
- put (to the N)
- take
- showed
- pilot
- standard

251 loss
- weight
- memory
- suffered
- hair
- caused
- total
- made N
- reported
- blood
- hearing
- great

job
lead (to N)
consequential
felt
considerable
cover (the N)
substantial
serious
severe
incurred
major
complete

252 colour
full
add
light
dark
bright
pale
give
high
strong

253 shop
gift
pet
run
antique
video
High (Street)
souvenir

254 benefit
housing
get
maximum
marginal
claim
receive
invalidity (pension…)

255 benefits
enjoy

256 animal
wild

257 heart
broken
beating
pounding (adj)
suffered
stopping
lies (at the N of)
thumping
heavy
thudding

258 election
general
won
held
following
lost
fought
forthcoming

259 elections
general
national
municipal

260 purpose
serve
sole
general
specific
primary
useful
real
special
intended
dual

261 standard
high
poor
set
required
double
reasonable
maintain
meet
improve
low
minimum
acceptable

262 standards
safety
meet

raise
minimum

263 secretary
former
private
appointed

264 date
closing
blind
set
effective
due
fixed
target
early
agreed

265 music
played
classical
live
popular
traditional
loud
contemporary
make
written
composed

266 hair
blonde
blond
curly
thick
dyed
facial
fair
cropped
dye
straight
wash
brushed
soft
permed
ginger
auburn

267 factor
major
key
risk
significant
crucial
determining
contributory
deciding
limiting
time
common
critical
contributing
decisive
vital
N + influencing
cost
human
additional

268 pattern
follow
set
established
regular
shows
complex
common
repeated

269 front
house

270 evening
early
late
spent

271 population
civilian
large
entire
total
whole
overall

272 plant
processing
reprocessing

273 pressure
- put N on
- high
- increasing
- (peer) group
- intense
- enormous
- growing
- low
- applied
- came (under N)
- considerable
- exerted
- take N off
- strong
- constant
- heavy
- severe

274 response
- immediate
- positive
- sexual
- public
- initial
- poor
- emotional
- given
- rapid
- quick

275 street
- side
- walking (up/down the N)
- busy

276 performance
- strong
- poor
- improve
- outstanding
- live
- impressive
- give

277 knowledge
- common
- prior
- acquired
- required
- background
- expert
- N + gained
- full
- working
- special
- true
- professional
- intimate
- specific
- wide

278 design
- interior
- garden
- original
- art
- research

279 page
- blank
- turn
- printed
- (on the) opposite
- previous

280 rest
- spend (the)
- have
- get
- take
- give

281 basis
- regular
- daily
- (on a) day(-to-day)
- form
- weekly
- individual
- provide
- permanent
- annual
- temporary
- voluntary
- monthly
- firm
- casual

282 size
small
full
large
fits
increase
half (the)
standard
average
medium
reduce

283 environment
natural
protect
safe
work
home
damage
improve
stable
hostile
healthy

284 fire
opened N on
set N to / obj on N
caught
friendly
started
(fire) burning
N + broke (out)
light (verb)
caused
make
spread
control

285 series
whole
drama
comedy
documentary
produced
launched
final
successful
present
runs

286 success
huge
achieved
enjoyed
make
considerable
major
ensure
early
guarantee
relative
continuing
outstanding
likely
immediate

287 thought
spare (a N)
given
bear
(a N)struck
(a N) occurred
deep
careful
N came (to me/mind)
serious
conscious
modern

288 thoughts
on second
interrupted

289 list
long
full
make
complete
drawn (up)
compiled

290 future
near
foreseeable
bright
uncertain
secure (verb)
immediate
(N+) holds
discuss
distant

(N+) lies
long-term
face
bleak

291 analysis
detailed
data N
showed
provide
performed
carried
suggests
careful
full
(a) developed

292 space
storage
open
parking
limited
living
extra
empty
little
use
free

293 tv
watching
live
colour

294 demand
meet
increase
huge
consumer's
strong
high
growing
made
satisfy + N
reduce

295 statement
issued
N+ said
made
joint
monthly
released

296 attention
pay
draw
turned
focused
attract
media
special
medical
public
particular
close
brought
given
caught
divert
received
N is/should be directed
full
careful
considerable
serious
concentrated
distract
hold
constant
urgent

297 love
making
fall
true
real
young
passionate
wonderful

298 principle
general
basic
fundamental
established
underlying

299 principles
basic
general
fundamental

apply
underlying
established
guiding
broad

300 set
complete
full
whole
standard
single

301 doctor
see
hospital
consult
medical
went
called
local
visit
examined

302 choice
have
make
wide
given
offer
obvious
personal
free
consumer
limited
possible
natural
informed

303 feature
regular
striking
distinctive
major
particular
prominent
attractive
significant
characteristic
notable
unusual
unique
dominant

304 couple
happy
elderly

305 step
take
further
big
major
unusual
positive
necessary
significant
essential

306 machine
use

307 income
low
annual
high
current
disposable
family
taxable
average
guaranteed
fixed
regular
national
paid
gross
extra
additional
(to) supplement

308 training
job
youth
teacher
vocational
professional
staff
military
management
provide
formal

intensive
further
college
do (some)
in-service
initial
given
quality
received
business
full

309 film
making
directed

310 effort
make
put
relief
concerted
required
considerable
involved
conscious
major
ensure
extra
joint
minimum
tremendous

311 player
former
outstanding

312 award
won
presented
received
given
prestigious

313 village
fishing
pretty
situated
picturesque
quiet
deserted
attractive

314 news
good
radio
national
world
bad
told
daily
brief
today
latest
broke (the N to)
television
heard
tonight
N + reached
spread

315 difficulty
have
considerable
overcome
increasing

316 cell
prison

317 energy
saving
low
high
renewable
waste
supply
alternative
reduce

318 degree
(to a) greater
high
(to a) certain
joint
considerable
lesser
honorary
awarded
remarkable
obtained

319 means
- provide
- effective
- use
- possible
- alternative
- private
- necessary

320 growth
- rapid
- strong
- potential
- employment
- personal
- high
- slow
- continued
- major
- overall
- unemployment

321 treatment
- have
- medical
- hospital
- further
- require
- receiving
- drug
- given
- emergency
- effective
- preferential
- prescribed
- surgical
- undergoing
- following

322 sound
- heard
- poor
- made
- loud
- distant
- turned (down/up/off)
- soft
- deep
- produce

323 task
- impossible
- daunting
- simple
- tough
- set (himself)
- perform
- take (on/over)
- carry out
- completed (V)
- faced
- undertake
- formidable
- hard

324 provision
- made
- adequate

325 behaviour
- bad
- good
- learned
- responsible
- disruptive

326 function
- perform
- primary
- brain
- fulfil

327 resource
- human
- valuable
- natural

328 resources
- natural
- human
- available
- limited
- energy
- required
- existing
- increase

329 defence
- made

330 garden
- back
- vegetable
- secret
- front
- walled
- herb
- flower
- rose
- do
- rear
- pretty
- attractive

331 floor
- top
- upper
- reach

332 technology
- new
- latest
- advanced
- using
- developed

333 style
- traditional
- classic
- unique
- adopt

334 feeling
- get
- strong
- general
- strange
- gave
- bad
- sinking
- deep
- gut
- growing
- experienced

335 doubt
- cast
- reasonable
- put (in/beyond)
- serious
- leave (no/little)

336 horse
- riding

337 answer
- simple
- give
- lies
- get
- right
- provide
- correct
- obvious
- clear
- wrong
- straight
- satisfactory

338 user
- allows

339 funds
- raise

340 character
- central
- main
- great
- strong
- distinctive

341 risk
- high
- reduce
- take
- increased
- run
- put
- serious
- potential
- involved
- significant
- minimise
- grave

342 dog
- walking (a N)
- pet

343 army
- joined

344 station
- radio
- tv

345 glass
- broken

346 glasses
- wearing

347 cup
- win
- make (a N of)
- played
- nice N of tea

348 husband
- estranged
- N + left
- former
- lost
- second

349 capital
- foreign
- current
- major

350 note
- take
- made (a mental)
- left
- mental
- delivery

351 season
- start

352 argument
- put (forward)
- further
- cited
- additional
- strong
- made
- rejected
- support
- accept
- logical
- presented
- convincing
- powerful

353 show
- put (on a N)
- live
- quiz
- open
- go (on N)
- held
- run
- annual

354 responsibility
- take
- accept
- personal
- full
- moral
- assume
- given

355 deal
- great
- good
- peace
- struck
- signed
- fair
- agreed
- get
- make

356 economy
- slowing
- planned
- boost
- strong

357 element
- key
- essential
- strong
- major
- significant

358 duty
- heavy

359 attempt
made
serious
desperate
failed (+ N)
deliberate
unsuccessful
vain

360 investment
foreign
capital
business
made
increase
major
direct
attract
massive

361 brother
younger
older
elder
big
little

362 title
world
won
national
defend
retain
full
took

363 hotel
luxury
stay
(family-) run
comfortable
family
elegant
booked

364 aspect
particular
further

365 increase
huge
significant
substantial
marked
large
real
dramatic
massive
small
further
rapid
sharp
considerable
caused
major
modest
big

366 help
seek
professional
great
provide
give
offers
medical
receive
call (for)
appealed (for)
domestic
require

367 summer
late
spent
early
dry
full

368 daughter
eldest
youngest
only

369 baby
expecting
unborn
lost

370 sea
open
calm

371 skill
- required
- considerable

372 claim
- make
- rejected
- compensation (claim)
- unfair

373 concern
- expressed
- cause
- growing
- major
- public
- deep
- real
- increasing
- considerable
- voiced

374 discussion
- open
- further
- considerable
- major

375 customer
- particular

376 box
- gift
- (tick the) appropriate N

377 conference
- peace
- party
- told
- annual
- called
- planned
- hold
- major
- opened

378 profit
- net
- make
- increase
- gross
- small

379 division
- top
- made
- major

380 procedure
- followed
- standard
- using
- normal
- adopted
- complaints + N

381 king
- crowned
- late
- appointed

382 image
- public
- improve
- clean (up)
- presented

383 oil
- virgin
- cooking

384 circumstance
- causal

385 circumstances
- certain
- (in/under) normal
- surrounding
- exceptional
- changing
- personal
- suspicious
- unusual
- causal
- extreme
- favourable

386 proposal
- rejected
- made
- accepted
- original

approved

387 sector
voluntary

388 direction
(in the) right
opposite
(in the) wrong
future
reverse

389 sign
(a) good
shows
little
gave
(a) sure
visible
made (a/no N)

390 measure
(in) large
temporary
(in) small
(in) equal
give
provide
(in) full

391 attitude
adopted
right

392 disease
chronic
cause
transmitted
severe
serious

393 commission
set (up) + N
appointed + N
N + established
take

394 seat
back
front
passenger
take
rear
driver's
driving
parliamentary
marginal

395 president
former
elected
re-elected

396 goal
scored
(the) winning N
early
ultimate
achieve
N + disallowed

397 affair
whole

398 affairs
state
current

399 technique
using
developed
useful
applied
simple
effective
employed

400 respect
mutual
great
due
earned

401 respects
pay

402 item
news

403 version
film
original

final
current
launched

404 ability
develop
limited

405 good
do
public
common

406 goods
manufactured
deliver

407 campaign
election
launched
bombing
run
media
mounted
fought
waged
conducted

408 advice
give
seek
expert
free
offer
professional
practical
take
provide
sound
further
followed

409 pupil
former

410 advantage
take
full
great
(to the) best
added
gain
unfair
give
big
major
significant
distinct
special
additional

411 memory
short-term
stored
long
used
working
short
visual
long-term
recent

412 memories
childhood
brought (back)
happy
past
fond
vivid
old
long

413 culture
popular
western
dominant
traditional

414 blood
infected
give

ANEXO 4

ADELEX COLLOCATION TEST v.1

TASK 1

Fill in the blanks in each sentence by adding only one word. The word you need to add can only be an adjective, a noun or a verb. The first letter of each word is provided to help you.

Example:

0. **N_____ light** is preferable to artificial light.

Answer:

0. **N**atural **light** is preferable to artificial light.

1. Applications are particularly welcome from women, people from ethnic minority backgrounds and **d_____ people**.
2. Use a map to help you **f_____ your way** along the trail.
3. Bob started painting over 10 years ago in his **s_____ time**.
4. Everything I read and studied for many years of very **h_____ work** explicitly rejected God as an explanation of anything.
5. Many of the **r_____ areas** which surround the city centre have inadequate car parking and garaging facilities for residents.
6. The programme gives students the possibility to **s_____ a problem** step by step.
7. Rates are expected to lower below 5 per cent over the **c_____ weeks.**
8. We were lucky, as none of our **c_____ family** or friends died in the disaster.
9. Carmen, showing at Royal Opera House. **O_____ night**: 8 December, 2006.
10. He **r_____ a question** for debate.
11. A newer tradition is the New Year's Day Parade which **r_____ money** for charity.
12. After viewing a story, you'll see four boxes to rate your interest in the story from no interest to **h_____ interest.**
13. When we are thirsty, we turn on the tap and get clean, **f_____ water**.
14. They lost their lives when their **car c_____** into a tree.

15. Feel free to visit our website to find the holiday that **m_____ your needs**.

16. I was so depressed about the comments from my boss that I **q_____ my job.**
17. We may release details of **m_____ persons** to the media if we consider this would help to find them.
18. When my house finally reaches a **r_____ price** I'll be able to sell it and leave you forever!
19. We needed to **t_____ action** in Iraq for humanitarian reasons.
20. I was so confused, it didn't even **c_____ my mind** to get the attention of other boats.
21. The **s_____** and unexpected **death** of the director this week has left the Government trying to rebuild public confidence.
22. You may feel depressed, tearful, angry or anxious **for no a_____ reason.**
23. The government, acting in **f_____ agreement** with the organisation, made the following declaration.
24. This website provides a **w_____ range** of information.
25. Windows at the school were kept open so pupils can **b_____ fresh air** at all times.
26. Bookstores have substantially **i_____ sales** using the Internet.
27. Use a **s_____ light** to create a relaxing atmosphere.
28. Whether this is a **t_____ story** or mere speculation we have yet to find out.
29. He intended to be the first **o_____ leader** to confront the government openly from inside Malawi.
30. First of all, I would like to give you an **o_____ picture** of the discussions today.
31. Britain has been the **m_____ source** of software in Europe for some time.
32. We were able to chat and exchange information while enjoying a **n_____ cup of tea** and a sandwich.
33. Why would you need a licence to keep dangerous **w_____ animals**?
34. Political parties compete to **w_____ elections** by submitting distinct programmes from which the electorate can choose.
35. The **c_____ date** for applications will be October 29, 2004, but submissions sent well in advance of this date would be appreciated.
36. **R_____ operations** ended Thursday when the bodies of the nine were recovered.
37. She **c_____ music** for the film.
38. Leg injuries were common among players in those early days and contributed to the team's **p_____ performances.**

39. It is **c_____ knowledge** today that chewing gum between meals helps to reduce the build-up of plaque.
40. Please note that prices can change on a **d_____ basis.**
41. There is **p_____ space** for a maximum of two cars outside the front garden gate.
42. Consumers, especially those on **l_____ incomes**, or those living in areas of economic and social deprivation are often the worst affected.
43. Why are violence and **d_____ behaviour** in schools a growing problem in Britain?
44. In this book the **m_____ character** is sir Charles Baskervilles.
45. It's particularly important to **s_____ help** from your doctor.
46. If you **m_____ a claim**, you need to give evidence.
47. They **l_____ a campaign** for an independent, unitary Bosnia.
48. Customers can **t_____ advantage** of several services.
49. While the **f_____ seats** are comfortable, it's a different story in the back.
50. On arrival the guest must complete a **r_____ form**.

TASK 2

Translate the following collocations into English. Add <u>either one single word or one hyphenated word</u> in each case.

Example:

0. Prestar atención: To __pay___ attention

1. Empezar el día: To _____ the day
2. La vida cotidiana: _____ life
3. Arte antiguo: _____ art
4. Trabajo de voluntariado: _____ work
5. Un número creciente: A/An _____ number
6. El mundo en vías de desarrollo: The _____ world

7. Un problema grave: A/An _____ problem
8. Una casa no adosada (separada): A/An _____ house
9. Dar ejemplo/Servir como ejemplo (a los demás): To _____ an example
10. Familias monoparentales: _____ families
11. Irse a la cama temprano: To have a/an _____ night
12. Negar con la cabeza: To _____ your head
13. Llevar/dirigir un negocio: To _____ a business
14. Despilfarrar el dinero: To _____ money
15. Un cambio repentino: A/An _____ change
16. El lugar equivocado: The _____ place
17. Desarrollo sostenible: _____ development
18. Agua corriente: _____ water
19. Hacer un examen: To _____ a test
20. Política exterior: _____ policy
21. Hacer efecto: To _____ effect
22. Un cuerpo desnudo: A/An _____ body
23. Darte una bofetada: To _____ your face
24. Tardar una hora: To _____ an hour
25. Llamar a la puerta: To _____ on the door
26. Lo que nos depara el futuro: What the future _____
27. Ampliar un periodo (o plazo): To _____ a period
28. Una voz áspera: A/An _____ voice
29. Cita a ciegas: _____ date
30. Condiciones climáticas: _____ conditions
31. Futuro no muy lejano: Not too _____ future

32. Un proyecto ambicioso: A/An _____ project

33. Papel estelar (en cine, teatro, …): _____ role

34. Una estrecha relación: A/An _____ relation

35. Marcar/Establecer un récord: To _____ a record

36. Materia prima: _____ material

37. Trabajo fijo: A/An _____ job

38. Surgir la oportunidad: The opportunity _____

39. Superar una dificultad: To _____ a difficulty

40. Suspender un examen: To _____ a test

41. Tienda de regalos: _____ shop

42. Pelo rizado: _____ hair

43. Televisión en directo: _____ tv

44. Una extensa variedad: A/An _____ choice

45. Perfeccionamiento /capacitación docente: _____ training

46. Un sonido lejano: A/An _____ sound

47. Reunir fondos: To _____ funds

48. Trato/Acuerdo justo: _____ deal

49. Hermano mayor: _____ brother

50. Reservar un hotel: To _____ a hotel

TASK 3

Choose the word which <u>does not</u> collocate with the noun given (there is <u>only one wrong collocation</u> in each case). If all of them are correct, tick the option "no wrong collocation".

Example:

0. To ___ a list

☐ compile ☒ compose ☐ make ☐ no wrong collocation

1. To___ your hand

☐ hold ☐ rise ☐ shake ☐ no wrong collocation

2. ___ place

☐ safe ☐ secure ☐ sure ☐ no wrong collocation

3. ___ quality

☐ big ☐ high ☐ superb ☐ no wrong collocation

4. To___ the example

☐ adopt ☐ follow ☐ take ☐ no wrong collocation

5. ___ fact

☐ easy ☐ mere ☐ simple ☐ no wrong collocation

6. To___ an eye

☐ bear ☐ cast ☐ keep ☐ no wrong collocation

7. To___ your head

☐ raise ☐ turn ☐ wave ☐ no wrong collocation

8. To___ money

☐ cast ☐ earn ☐ spend ☐ no wrong collocation

9. To ___ to an end

☐ bring ☐ come ☐ get ☐ no wrong collocation

10. ___ need

☐ desperate ☐ imperious ☐ urgent ☐ no wrong collocation

11. ___ price

☐ full ☐ high ☐ sharp ☐ no wrong collocation

12. ____ garden

☐ herb ☐ plant ☐ rose ☐ no wrong collocation

13. ___ matter

☐ serious ☐ severe ☐ simple ☐ no wrong collocation

14. ___ style

☐ classic ☐ exceptional ☐ unique ☐ no wrong collocation

15. To ___ in mind

☐ bear ☐ have ☐ keep ☐ no wrong collocation

16. ___ experience

☐ durable ☐ relevant ☐ valuable ☐ no wrong collocation

17. ___ staff

☐ prepared ☐ qualified ☐ trained ☐ no wrong collocation

18. ___ letter

☐ business ☐ reception ☐ resignation ☐ no wrong collocation

19. To ___ evidence

☐ give ☐ grant ☐ provide ☐ no wrong collocation

20. To___ the doctor

☐ consult ☐ see ☐ visit ☐ no wrong collocation

21. To ___ a meeting

☐ call ☐ hold ☐ put ☐ no wrong collocation

22. ___ moment

☐ adequate ☐ precise ☐ right ☐ no wrong collocation

23. ____ officer

☐ army ☐ prison ☐ soldier ☐ no wrong collocation

24. To ___ sales

☐ boost ☐ foster ☐ increase ☐ no wrong collocation

25. ___ light

☐ diminishing ☐ fading ☐ soft ☐ no wrong collocation

26. ___ loss

☐ memory ☐ weight ☐ work ☐ no wrong collocation

27. ___ colour

☐ flashy ☐ full ☐ light ☐ no wrong collocation

28. ___ interest

☐ common ☐ concrete ☐ particular ☐ no wrong collocation

29. To ___ a date

☐ fix ☐ grant ☐ set ☐ no wrong collocation

30. ___ hair

☐ auburn ☐ dyed ☐ tinged ☐ no wrong collocation

31. ___ future

☐ foreseeable ☐ imprecise ☐ long-term ☐ no wrong collocation

32. ___ demand

☐ emergent ☐ growing ☐ huge ☐ no wrong collocation

33. ___ feature

☐ distinctive ☐ striking ☐ unique ☐ no wrong collocation

34. ___ effort

☐ considerable ☐ drastic ☐ tremendous ☐ no wrong collocation

35. To ___ the news

☐ break ☐ expand ☐ spread ☐ no wrong collocation

36. ___ degree

☐ bigger ☐ greater ☐ lesser ☐ no wrong collocation

37. ___ village

☐ attractive ☐ deserted ☐ picturesque ☐ no wrong collocation

38. ___ sound

☐ deep ☐ shallow ☐ soft ☐ no wrong collocation

39. To ___ a task

☐ make ☐ perform ☐ undertake ☐ no wrong collocation

40. ___ position

☐ dominant ☐ leading ☐ ruling ☐ no wrong collocation

41. ___ sea

☐ calm ☐ open ☐ unsettled ☐ no wrong collocation

42. ___ circumstances

☐ exceptional ☐ extreme ☐ undesirable ☐ no wrong collocation

43. In the ___ direction

☐ opposite ☐ reverse ☐ wrong ☐ no wrong collocation

44. ___ memory

☐ incredible ☐ recent ☐ short ☐ no wrong collocation

45. ___ situation

☐ current ☐ intolerable ☐ unsolved ☐ no wrong collocation

46. ___ difference

☐ elemental ☐ fundamental ☐ marked ☐ no wrong collocation

47. A fire ___

☐ breaks out ☐ extends ☐ spreads ☐ no wrong collocation

48. To ___ responsibility

☐ accept ☐ adopt ☐ assume ☐ no wrong collocation

49. To ___ a disease

☐ cause ☐ provoke ☐ transmit ☐ no wrong collocation

50. ___ advantage

☐ full ☐ significant ☐ total ☐ no wrong collocation

TASK 4

Choose the correct collocation and tick the appropriate box. There is only one correct collocation in each case. If none of the 3 first options is correct, tick the option "none of these".

Example:

0. A place devoted to entertainment is called **a/an ____ area**.

☐ break ☐ game ☒ play ☐ none of these

1. If you say **in ____ fact**, you indicate that you are giving more detailed information about what you have just said.

☐ actual ☐ real ☐ true ☐ none of these

2. If you **____ your way** somewhere, you walk or travel there.

☐ begin ☐ make ☐ reach ☐ none of these

3. A day on which people go to work is called **a/an ____ day**.

☐ job ☐ labour ☐ working ☐ none of these

4. A child who has no brothers or sisters is called **a/an ____ child**.

☐ alone ☐ only ☐ unique ☐ none of these

5. When you **____ a report**, you make it publicly available.

☐ announce ☐ issue ☐ proclaim ☐ none of these

6. **A/an ____ process** lasts for a long time.

☐ durable ☐ lengthy ☐ stretched ☐ none of these

7. When you refer to the world and life in general, in contrast to a particular person's own life, experience and ideas, which may seem untypical, you talk about **the____ world**.

☐ authentic ☐ real ☐ true ☐ none of these

8. The most important part of a meal is called **the ____ course**.

☐ full ☐ principal ☐ strong ☐ none of these

9. If you are in control or in charge of a political party, you **____ the party**.

☐ direct ☐ govern ☐ lead ☐ none of these

10. A person who belongs to a workers' organisation is called **a/an____ member**.

☐ collective ☐ syndicate ☐ union ☐ none of these

11. A family group which includes relatives such as uncles, aunts, and grandparents as well as parents, children, brothers and sisters is referred to as ____ **family**.

☐ broad ☐ extended ☐ inclusive ☐ none of these

12. Nights during which you don't sleep are ____ **nights**.

☐ awake ☐ sleepless ☐ wakeful ☐ none of these

13. When you gain money by selling things, you ____ **profits**.

☐ gather ☐ make ☐ win ☐ none of these

14. When someone is in **a/an____ position**, s/he has a special advantage or has the chance to do something that most people cannot do.

☐ honoured ☐ outstanding ☐ privileged ☐ none of these

15. If you ____ **an argument**, you agree with it.

☐ advocate ☐ maintain ☐ support ☐ none of these

16. If you ____ **authority**, you use it, or put it into effect.

☐ employ ☐ execute ☐ exercise ☐ none of these

17. The water from the sea is called ____ **water**.

☐ salt ☐ salted ☐ salty ☐ none of these

18. Rich, powerful and fashionable people are called ____ **society**.

☐ high ☐ peak ☐ top ☐ none of these

19. If you wait until you know all the facts before forming an opinion you have **a/an ____ mind**.

☐ ample ☐ open ☐ vast ☐ none of these

20. If you are not wearing anything on your feet you have ____ **feet**.

☐ bare ☐ nude ☐ undressed ☐ none of these

21. When you have good reasons to show that something is true or untrue, right or wrong, you have **a/an____ argument**.

☐ convincing ☐ sure ☐ true ☐ none of these

22. When you are really worried about something you have **a/an ____ concern**.

☐ deep ☐ hard ☐ inner ☐ none of these

23. If you say that a woman is ____ **a baby**, you mean that she is pregnant.

☐ expecting ☐ hoping ☐ waiting ☐ none of these

24. If you ____ **an important part** in an event, you are very involved in it and have an important effect on what happens.

☐ have ☐ play ☐ take ☐ none of these

25. If you are not sure about your future, you talk about **a/an ____ future**.

☐ doubtful ☐ unlikely ☐ unstable ☐ none of these

26. If you ____ **a course** you do all of it.

☐ complete ☐ conclude ☐ end ☐ none of these

27. A question which is essential or very important is **a/an ____ question**.

☐ considerable ☐ fundamental ☐ remarkable ☐ none of these

28. A business which makes lots of money is called **a/an ____ business**.

☐ perfect ☐ successful ☐ winning ☐ none of these

29. If somebody ____ **your needs**, s/he gives you enough of what you need.

☐ contents ☐ pleases ☐ satisfies ☐ none of these

30. If you say you don't have **the ____ idea**, you emphasize you don't know something.

☐ faintest ☐ lightest ☐ smallest ☐ none of these

31. If your **face** ____, it goes red because you are feeling a strong emotion such as embarrassment or anger.

☐ flushes ☐ heats ☐ illuminates ☐ none of these

32. **The ____ cost** is an approximate judgement or calculation of a value or price.

☐ guessed ☐ speculated ☐ vague ☐ none of these

33. **A/an ____ person** is typical or normal and has qualities that most people have.

☐ average ☐ mean ☐ standard ☐ none of these

34. If someone damages your reputation, you can pursue ____ **action**.

☐ denigration ☐ libel ☐ prosecution ☐ none of these

35. ____ **age** is the period when you stop working, usually because of your age.

☐ retirement ☐ retreat ☐ withdrawal ☐ none of these

36. If you successfully agree on something with other people, you ____ **a decision**.

☐ accomplish ☐ achieve ☐ grasp ☐ none of these

37. **A/an ____ road** is a place full of vehicles and movement.

☐ busy ☐ heavy ☐ packed ☐ none of these

38. If a person goes to live in another house, s/he ____ **house**.

☐ changes ☐ moves ☐ transfers ☐ none of these

39. When someone's death has not been deliberately intended or planned it is called **a/an ____ death**.

☐ accidental ☐ casual ☐ involuntary ☐ none of these

40. Foods which contain less sugar or fat than ordinary are called ____ **food**.

☐ diet ☐ fit ☐ soft ☐ none of these

41. If you get someone's respect because you deserve it, you ____ **his/her respect**.

☐ earn ☐ gather ☐ take ☐ none of these

42. **A/an ____ club** is limited to people who have a lot of money or who belong to a high social class.

☐ exclusive ☐ luxurious ☐ selective ☐ none of these

43. When something happens normally or is usually true, it happens **as a/an ____ rule**.

☐ average ☐ extended ☐ standard ☐ none of these

44. If an event is not expected to happen it is **a/an ____ event**.

☐ dubious ☐ unlikely ☐ unpredictable ☐ none of these

45. **To____ treatment** is to say what medicine or treatment a sick person should have.

☐ order ☐ prescribe ☐ stipulate ☐ none of these

46. A feeling that you are sure is right, although you cannot give a reason for it, is called **a/an ____ feeling**.

☐ gut ☐ impulse ☐ instinct ☐ none of these

47. When you quickly and eagerly do something when you have the chance to do it, you ____ **the opportunity**

☐ catch ☐ hold ☐ seize ☐ none of these

48. If you give someone **a/an ____answer**, you answer him/her without any delay.

☐ fast ☐ hasty ☐ immediate ☐ none of these

49. If you ____ **the risk** of something, you make it safer.

☐ diminish ☐ reduce ☐ weaken ☐ none of these

50. An effect badly damaging something is called **a/an ____ effect**.

☐ appalling ☐ devastating ☐ overwhelming ☐ none of these

DATOS PERSONALES

Apellidos, Nombre:
Edad:
Nacionalidad:
Lengua materna:
Licenciatura que cursa:
¿Qué curso de la licenciatura realiza?

¿Alguno de sus padres es nativo de habla inglesa? SÍ/NO

¿Ha recibido alguna noción sobre las colocaciones (definición teórica, actividades de clase, etc.) en sus años de aprendizaje del inglés? Explíquela brevemente.

_____

_____

¿Desea añadir alguna información que considere relevante para este test?

_____

_____

MUCHAS GRACIAS

FACULTAD DE FILOSOFÍA Y LETRAS
DEPARTAMENTO DE FILOLOGÍAS INGLESA Y ALEMANA

Tesis Doctoral

"Recopilación, desarrollo pedagógico y evaluación de un banco de colocaciones frecuentes de la lengua inglesa a través de la lingüística de corpus y computacional"

La Directora de la tesis, La Doctoranda,

Dra. Carmen Pérez Basanta María Moreno Jaén

Granada, 2009