El inglés y el español desde una perspectiva cuantitativa y distributiva: equivalencias y contrastes[1]

Pascual CANTOS
Aquilino SÁNCHEZ

Universidad de Murcia
pcantos@um.es
asanchez@um.es

Recibido: 7/01/2011
Aceptado: 21/03/2011

RESUMEN
Las similitudes o diferencias que percibimos al comparar dos o más lenguas son siempre relativas. El análisis de estos contrastes lingüísticos se puede formalizar tanto desde diversos ángulos y/o niveles lingüísticos (fonético-fonológico, morfológico, sintáctico, semántico, pragmático, diacrónico, etc.), como mediante diferentes metodologías (inductiva, deductiva, descriptiva, etc.). La finalidad de este estudio comparativo entre el español y el inglés es ofrecer, por primera vez, datos cuantitativos objetivos y fiables sobre la estructura y vertebración del inventario léxico en ambas lenguas (riqueza léxica, crecimiento léxico: formas, tramos de frecuencias, formas léxicas y funcionales, formas más frecuentes, longitud de formas, etc.), y poder aproximarnos a las diferentes propiedades matemáticas que subyacen en las lenguas española e inglesa. Todos los datos cuantitativos se han obtenido partiendo de dos corpus lingüísticos equivalentes (en estructura, composición y tamaño: 20 millones de palabras): *Cumbre* (para el español) y *Lacell* (para el inglés).

Palabras clave: Propiedades cuantitativas, distribución del léxico, frecuencias, español, inglés.

English and Spanish from a distributional and quantitative perspective: Equivalences and contrasts

ABSTRACT
The similarities or differences people perceive while comparing two or more languages are always relative. However, the analysis of these linguistic contrasts can be formalized either from different angles and/or linguistic levels (phonetic/phonological, morphological, syntactic, semantic, pragmatic, diachronic, etc.), or from different research methodologies (inductive, deductive, descriptive, etc.).
The purpose of this comparative study between Spanish and English is to offer, for the first time, objective and reliable quantitative data on the structure and structuring of the lexical inventory in both languages (lexical richness, lexical growth, types, frequency bands, lexical and functional types, most common types, type-length, etc.); and to approach the different underlying mathematical properties of

[1] Esta investigación deriva de otras afines relacionadas con el proyecto financiado por la Fundación SENECA, de la Comunidad Autónoma de la Región de Murcia, proyecto 00481/PI/04.

Spanish and English. All the quantitative data were obtained from two equivalent linguistic corpora (in structure, composition and size: 20 million words): *Cumbre* (for Spanish) and *Lacell* (for English).

Keywords: Quantitative properties, lexical distribution, frequencies, Spanish, English

1. INGLÉS Y ESPAÑOL: DOS LENGUAS DIFERENTES A PARTIR DE UN TRONCO COMÚN

Para la mayor parte de los hablantes de inglés y de español, la percepción más generalizada es que ambas lenguas son muy diferentes. Las diferencias se suelen reducir a afirmaciones como 'el inglés es una lengua muy difícil de pronunciar', 'el español tiene muchas formas verbales', 'el inglés tiene una gramática más fácil que el español', 'el español tiene algunos sonidos muy difíciles, como la *z, j*', etc.[2] Son afirmaciones muy genéricas, pero están basadas en la percepción que los hablantes de una de las dos lenguas tienen sobre la otra, especialmente cuando inician su aprendizaje. Para el lingüista, la gramática del inglés es tan complicada o compleja como la del español, si bien es cierto que en algunos aspectos, como el de las flexiones, el inglés ofrece un cuadro de formas más simplificado, debido al hecho de que no existe en esta lengua, por ejemplo, la flexión de género y, en consecuencia, disminuye también la cantidad de flexiones relativas al número, que sí está marcado en inglés. Por razones similares, el hablante inglés suele afirmar que el subjuntivo español, o las formas del pasado verbal, son muy difíciles, hecho evidente si se tiene en cuenta que el inglés ha desterrado prácticamente las formas del subjuntivo y ha simplificado notoriamente las de pasado. En realidad, la percepción de facilidad o dificultad de una lengua suele basarse en lo que podría denominarse 'su estructura superficial o formal', que es lo primero que el hablante percibe, dejando de lado aspectos quizás más fundamentales relativos a otros niveles de la lengua (el semántico, la organización de la oración o del discurso, etc.), menos perceptibles a primera vista.

La similitud o diferencia entre lenguas es siempre relativa. Si la comparación se estableciera entre el español y el chino, por ejemplo, las diferencias entre ambos idiomas serían muchas y profundas, ciertamente muchas más que las que separan el español y el inglés. Y si la comparación se llevara a cabo entre el español y el catalán, las coincidencias o semejanzas probablemente superarían a las diferencias.

Los lingüistas acostumbran a describir la historia de las lenguas utilizando el símil del árbol. La figura de un árbol refleja con fidelidad el desarrollo de las diferentes lenguas a partir de un tronco común, que luego se divide o subdivide en ramas, atendiendo al aislamiento o separación –y consiguiente diferenciación– de los grupos humanos que se valen de cada sistema lingüístico como medio de comunicación. El tronco común entre el inglés y el español se remonta al indo-indoeuropeo, lengua que

---

[2] Véase al respecto algunos foros de discusión, por ejemplo *¿Qué idioma crees que es más difícil de aprender: inglés o español?* En http://es.answers.yahoo.com/question/index?qid=20090817200500AAZiP6H

supuestamente predominaba en la zona poblada de Europa central y en una franja que se prolongaba desde los Balcanes hasta Irán, y que luego empezó a diferenciarse en varias lenguas y dialectos, en torno al tercer milenio antes de Cristo. El inglés y el español surgen de ramas y rutas diferentes: el inglés deriva de la rama germánica –con importantes aportaciones posteriores del latín y del francés–, mientras que el español surge del latín, otra de las grandes divisiones en que se ramificó el indoeuropeo. El nacimiento y desarrollo de ambas lenguas no coincide plenamente en el tiempo, pero se lleva a cabo dentro de la relativa unidad geográfica y cultural de Europa. Este hecho –junto con otros– contribuye a la configuración de los dos idiomas 'limando' diferencias y acercando procedimientos y estructuras, como podrá comprobarse en los apartados siguientes.

2. CUANTIFICACIÓN DE LOS HABLANTES DE INGLÉS Y ESPAÑOL

El español y el inglés están prácticamente igualados en número de hablantes nativos, si bien es verdad que las estadísticas no coinciden plenamente, o bien porque los datos exactos son difíciles de computar, o bien porque los criterios de recopilación difieren en algunos parámetros. Globalmente, los datos sobre los hablantes de español e inglés como lengua materna pueden resumirse en la siguiente tabla (*Tabla 1*):

	Estimación de *The Ethnologue*[3]	
Inglés (2008)	328 Millones	4,89%
Español (2008)	329 Millones	4,90%
Población mundial estimada (2008)	6.706 Millones[4]	100,00%

Tabla 1. Hablantes nativos de español e inglés (2008) y porcentaje sobre población mundial

Los datos de hablantes nativos, tomados aisladamente, no bastarían sin embargo para dar una información ajustada sobre el uso de cada una de las dos lenguas. La lengua inglesa es actualmente lengua 'franca' en todo el mundo, y cuenta con ciertos privilegios en más de 75 países. Este hecho por sí solo hace que la cifra de hablantes de inglés como primera o segunda lengua supere con creces los mil millones de personas. *The Internet World Stats*[5] eleva esta cifra a 1.247 millones (en 2008). Dentro de los mismos parámetros, el español alcanza los 408 millones. Si nos centramos en Europa, se calcula que la mitad de sus habitantes puede comunicarse en inglés, pero no en español. Algunos datos más son también significativos al respecto:

1. Una gran mayoría de científicos lee los textos de su especialidad en inglés.
2. Más del 75% de la correspondencia en el mundo se lleva a cabo en inglés.

[3] En http://www.ethnologue.com
[4] En http://es.wikipedia.org/wiki/Poblacion_mundial
[5] En http://www.internetworldstats.com/languages.htm

3. En cualquier momento del día, hay unos 120.000 alumnos estudiando inglés en centros del *British Council* desparramados por todo el mundo.
4. El número de alumnos que estudia inglés cada año asciende a más de mil millones.
5. Cada año viajan a Inglaterra más de 600.000 alumnos para estudiar inglés, generando unos ingresos de más de mil millones de libras anuales.
6. Más de un 80% de la información electrónica se almacena en inglés.
7. Y si nos centramos en Internet, los datos revelan de nuevo el incuestionable predominio del inglés como lengua vehicular, que copa casi un tercio de todos los usuarios (*Tabla* 2):

	Usuarios de Internet	Crecimiento en Internet (2000 - 2008)	Usuarios de Internet (% sobre el total)
Inglés	463 Millones	226,70 %	29,10 %
Español	130 Millones	619,30 %	8,20 %
Total mundial	1.596 Millones	342,20 %	100,00 %

Tabla 2. El inglés y el español en Internet *(Fuente: Internet World Stats[6], 2009)*

Así pues, desde el punto de vista del número de hablantes nativos, ambas lenguas están equiparadas, pero en cuanto al uso de cada una de ellas en el contexto mundial, el inglés ocupa una preeminencia y liderazgo indiscutibles.

3. EL LÉXICO INGLÉS Y ESPAÑOL

3.1. PALABRAS Y CONCEPTOS

Uno de los aspectos que suelen llamar poderosamente la atención es la cuestión de cuántas son las palabras de que dispone una determinada lengua. De ahí que muchos hablantes se hagan la pregunta sobre 'cuál es la lengua más rica en palabras, la inglesa o la española'. El tema no es trivial, ya que si las palabras reflejan la conceptualización que los hablantes tienen de la realidad, en tal caso una mayor riqueza léxica equivaldría a una más rica conceptualización de aquella y, en consecuencia, a una más rica y extensa capacidad comunicativa. Que una lengua disponga de más palabras que otra, significa que sus hablantes manejarán más conceptos.

En consecuencia, el tema de si una lengua cuenta con más o menos palabras respecto a otra u otras no es baladí. 'Ir de compras' implica una realidad subyacente y propia de una sociedad consumista, en la que abundan los productos y la gente dispone de dinero suficiente para a 'practicar' esa actividad. Difícilmente podríamos

[6] En http://www.internetworldstats.com

imaginar esa actividad en una sociedad primitiva, o incluso en un país en el que el desarrollo económico no hubiera alcanzado ciertos mínimos. Y si no existe esa realidad, tampoco existe su conceptualización entre los hablantes de la comunidad. De manera similar, el término 'paella' no existe en muchas lenguas, sencillamente porque este plato tampoco existe en las sociedades que las hablan. Los conceptos, objetos o realidades dan origen a palabras para referirnos a ellos, y las palabras, a su vez, son reflejo de aquellos.

En razón de lo dicho, sería poco realista pretender definir con exactitud el número de palabras de una lengua: esto equivaldría a poder conocer con precisión el número de conceptos que los hablantes manejamos en la comunicación. Y no lo sería menos afirmar con ligereza que un idioma cuenta con más de un millón de palabras (el inglés), mientras otras (como el alemán, el francés o el español) cuentan sólo con cien o doscientas mil. ¿Podría darse tal diferencia en la conceptualización de la realidad en comunidades culturalmente tan afines como la inglesa, la alemana, la francesa o la española?

Además, el número exacto de palabras usadas en una lengua es imposible de determinar con precisión porque las lenguas están en continua evolución y no sería viable disponer de una lista completa y actualizada de todos los términos y conceptos relacionados y utilizados por los hablantes en cada momento. No obstante, es razonable atenerse a determinados criterios objetivos, en la medida en que tales criterios pueden ser tomados como representativos del uso que los hablantes hacen de una lengua. Hasta el presente, los diccionarios han sido aceptados y utilizados como instrumentos fiables respecto a los recursos léxicos de un idioma en cada momento histórico. Los recursos que actualmente ofrece Internet, así como la capacidad y velocidad de procesamiento que propician los ordenadores, mejoran la fiabilidad y precisión de las obras lexicográficas, pero aún no han dado origen a sustitutos fiables de las obras lexicográficas tradicionales. Para ello sería necesario recopilar fiablemente la producción lingüística dispersa en la red, sistematizarla, analizarla y definir los términos realmente utilizados y los sentidos atribuidos a cada uno de ellos. De momento, este objetivo aún debe ser alcanzado.

En cuanto a diccionarios, el *Diccionario de la Real Academia de la Lengua Española* (diccionario oficial de la lengua española) incluye unas 93.000 entradas en su última edición. El *Gran Diccionario de Uso de la Lengua Española* (SGEL, 2001), fundamentado en un corpus del español actual (*Cumbre*[7]) y más centrado en el uso real y actual del español, incluye poco más de 70.000. Así pues, podría afirmarse que el número de lemas de la lengua española, si nos atenemos a lo recogido en las fuentes lexicográficas, podría aproximarse, en el mejor de los casos, a los cien mil.

El inglés no cuenta con un ente oficial que regule el uso y registro de la lengua. Los intentos llevados a cabo por Jonathan Swift o su sucesor americano John Adams,

---

[7] *Cumbre* es un corpus de 20 millones de palabras, equilibrado en su diseño y referido al español oral y escrito usado en todos los países de habla hispana, entre 1950 y 1995. *Cumbre* es propiedad de la editorial SGEL S.A. y fue utilizado en la elaboración del *Gran Diccionario de uso del español actual*. Madrid: SGEL, 2001.

en el siglo XVIII, no llegaron a buen fin y la Academia de la Lengua Inglesa nunca llegó a materializarse. Quizás este simple hecho ha propiciado el afán de los hablantes del inglés por cuantificar la riqueza léxica del idioma. McCrum et al. (1992), en su amena descripción de la historia de la lengua inglesa, afirman que esta lengua cuenta actualmente con el vocabulario más rico y extenso de todas las lenguas existentes (que suelen cifrarse en entre 3.000 y 6.000). Y en favor de tal afirmación aducen datos del *Oxford English Dictionary,* que contiene cerca de medio millón de palabras. A ellas hay que añadir las que no recoge esta obra, términos de carácter técnico y científico, dialectales, etc. Los autores no dudan en concluir que si se registraran todas las palabras existentes, se rondaría el millón de palabras. La *Encyclopedia Americana* (2006), de otra parte, afirma que el inglés ha pasado de las aproximadamente 50.000 palabras del inglés antiguo a las más de 650.000 de la actualidad. En términos aún más concretos, el *Webster's Third New International Dictionary* (2010), afirma que incluye más de 470.000 entradas. La segunda edición del *Oxford English Dictionary* (1989) contenía poco más de 200.000 palabras. Estos diccionarios buscan todos ellos, de alguna manera, la exhaustividad, meta ciertamente difícil de lograr. En conclusión, comparando los datos de las obras mencionadas, el inglés multiplicaría hasta por cinco el volumen léxico del español: 650.000 frente a menos de 100.000. ¿Puede ser tomado este dato como realista? Es decir, ¿cabe pensar que la comunidad de hablantes del inglés supere en cinco veces la capacidad expresiva de la comunidad de hablantes de español, o que su riqueza conceptual sea cinco veces mayor? Si esta conclusión es difícil de imaginar, las diferencias detectadas en las obras lexicográficas deben responder a otros factores. Uno de ellos podría ser la menor atención prestada al estudio del léxico en la comunidad de hispanohablantes. Este hecho explicaría que los diccionarios del español suelan reducirse a recopilar las palabras más habituales en la comunicación, incluso poniendo el énfasis en la producción literaria y humanística, pero dejando de lado, entre otros, el léxico técnico-científico, las novedades léxicas más recientes, o el vocabulario especializado, dialectal o regional. Y el segundo factor que merece una atención especial es la *metodología* usada en la elaboración de los diccionarios. Si se unificase el método (criterios y recursos) sobre el cual se asentase la preparación de las obras lexicográficas, los resultados obtenidos alcanzarían un grado de fiabilidad mucho más alto que el actual.

Lo apuntado en el párrafo anterior se ve corroborado por el análisis de algunas obras lexicográficas generales del inglés. En efecto, los diccionarios generales del inglés elaborados con criterios similares a los habituales entre la comunidad hispanohablante no se diferencian tanto de los diccionarios del español o de otras lenguas del entorno (como el francés o el alemán). *The New Oxford Dictionary of English* (*NODE*) (1998) recoge unas 90.000 entradas (según la proyección estadística de una muestra parcial de la obra). En la solapa, los autores mencionan la cifra de '350.000 voces, frases y definiciones', pero no deben confundirse las voces o lemas con las formas (palabras diferentes en su forma) o con los significados o acepciones de las voces registradas.

Los datos léxicos incluidos en los diccionarios, aunque sean reales, deben ser analizados con atención para entenderlos en su justa medida. El número de voces seleccionadas puede variar:

1. Según el criterio que se aplique para definir lo que se entiende por 'palabra' en el diccionario;
2. Según los criterios seguidos para incluir o excluir determinados términos;
3. Según el tratamiento que se dé a los términos técnicos y especializados y a los nombres propios;
4. Según el criterio aplicado respecto al grado de exhaustividad que se pretende alcanzar. En la tradición lexicográfica inglesa y norteamericana se ha buscado con frecuencia la exhaustividad, sobre todo a partir del *Oxford English Dictionary*, finalizado en 1928, y a raíz de la 'guerra de los diccionarios' desatada en los Estados Unidos en la segunda mitad del s. XIX, en la que competían los diccionarios de Webster –más innovador– y de Worcester –más clásico y conservador. De ahí la tendencia en la lexicografía inglesa a incluir todo tipo de voces (derivados, nombres propios, voces coloquiales, obsoletas, *slang*, etc.).

Por otro lado, la propaganda de las obras lexicográficas se basa con excesiva frecuencia en la cantidad del caudal léxico incluido. Y en este aspecto, al lector se le confunde fácilmente apelando al número de voces, voces derivadas, voces compuestas, voces de igual forma y con más de una categoría gramatical, abreviaciones, nombres propios y número de significados, pero sin aclarar lo que cada cosa supone respecto al número de voces reales. El lector no versado tiende a asociar la cifra destacada con el número de palabras diferentes ofrecidas por el diccionario.

3.2. CORPUS LINGÜÍSTICOS, FORMAS Y PALABRAS

Los diccionarios suelen recoger el uso lingüístico actual y heredado. Es decir, son en gran medida sincrónicos y diacrónicos, y resultan de la conjunción de lo actual con lo heredado. Además, a lo heredado o transmitido por generaciones pasadas, se le suele atribuir más peso que a lo actual. Hay razones para que esto sea así (la estabilidad del sistema comunicativo, por ejemplo), pero hasta ahora contribuía a mantener esta situación una razón cualitativamente inocua: recopilar el uso lingüístico del presente con las herramientas de la lexicografía tradicional (ficha y lápiz, anotaciones ocasionales, obras literarias –sobre todo de autores ya consagrados) era una tarea difícil y poco eficiente. Desde finales del siglo XX, no obstante, la lexicografía cuenta ya con herramientas mucho más eficaces y fiables: los corpus lingüísticos, o grandes recopilaciones de muestras lingüísticas, orales y/o escritas, susceptibles de ser tomadas de cualquier ámbito comunicativo.

Los corpus adecuadamente diseñados reflejan con gran fidelidad cuál es el uso real de una lengua en el periodo en el cual se ha hecho la recopilación. Además, si el corpus es representativo del conjunto de un idioma, ofrecerá una radiografía fiel de cuáles y cuántas palabras 'están en activo' –se usan– en un determinado momento histórico. Los corpus *Cumbre* (del español contemporáneo) y *Lacell* (del inglés

contemporáneo) creemos que responden a estas características y por esa razón serán tomados en este estudio como referencia para la proyección de datos comparativos relativos a la cantidad de palabras totales y palabras diferentes en las lenguas aquí comparadas. Es evidente que la estructura o características de un corpus condicionan de alguna manera la variedad de palabras contenidas en él. Es importante, por tanto, tener en cuenta que la proyección que hacemos a continuación debe tomarse como una proyección relativa, aplicable a textos o recopilaciones textuales de estructura y composición similar.

Nuestros corpus se ajustan a los siguientes parámetros:

1. Se refieren a la lengua española o inglesa de los últimos cincuenta años.
2. Recogen el habla de todos los países en los cuales tanto el español como el inglés son lenguas oficiales, aunque en cantidades notablemente sesgadas a favor de España y Reino Unido, con un porcentaje medio del 65% sobre el total, mientras que los países hispanoamericanos y los de habla inglesa diferentes del Reino Unido sólo cuentan con un 35 % de las muestras.
3. Respecto a la modalidad de lengua, la lengua hablada recibe un 35% del peso total. A la lengua escrita se le asignó el 65%. En ambas modalidades se buscó una amplia variedad respecto a los ámbitos de uso, tanto vertical como horizontalmente (la variedad de textos en su conjunto consta de varios miles de muestras diferentes).
4. Los listados de frecuencia fueron depurados de todos los elementos no estrictamente léxicos: se eliminaron, por ejemplo, todas las secuencias numéricas y de signos, aunque no las palabras con errores ortográficos o los términos extranjeros.

Los resultados obtenidos quedan reflejados en la Tabla 3:

		Total de palabras (*tokens*) diferentes	Total de formas (*types*) diferentes
Inglés	*Corpus LACELL*[8]	21.427.248	173.391
Español	*Corpus Cumbre*[8]	21.785.202	218.643

Tabla 3. Corpus del inglés y del español: palabras (*tokens)* y formas diferentes (*types*)

Cabe destacar que el número de palabras diferentes está en relación directa con el tamaño del corpus, es decir, a mayor número de palabras (*tokens*) recopiladas corresponde mayor número de formas diferentes (*types*)[9]. Puede observarse en los datos de la Tabla 3 que el corpus *Cumbre* (español) y el corpus *Lacell* –de tamaño y estructura similares– presentan algunos rasgos diferenciadores en cuanto a la cantidad

[8] En http://www.um.es/grupos/grupo-lacell/quees.php

[9] No obstante, véase Sánchez y Cantos (1997, 1998) en relación con la naturaleza de la curva hiperbólica del incremento.

de formas. Importa tener en cuenta que las palabras diferentes (*types*) no son necesariamente lemas. Además, los corpus suelen incluir un alto número de formas provenientes de otras lenguas, nombres propios, e incluso errores ortográficos que en cuanto 'diferentes' son contabilizados por el ordenador como formas distintas. Pero dado que la metodología aplicada en la recopilación es la misma para ambas lenguas, puede concluirse que los resultados son razonablemente válidos y homogéneos.

Los datos confirman algunas diferencias esperadas. El corpus español contiene 45.252 *types* más que su homólogo inglés. Lo cual hace que la relación *type/token* sea de 0,010 para el español, mientras para el inglés es de 0,008; o lo que es lo mismo, como media, en español la ratio de formas distintas es de aproximadamente 10 por cada 1.000 palabras de texto, mientras que en inglés esta cifra asciende solamente a 8 palabras. El español es por tanto más prolijo en formas que el inglés, conclusión ya previsible, dado que el español es una lengua con mayor carga flexiva.

$$Ratio\_type\_token_{Inglés} = \frac{Types}{Tokens} = \frac{173391}{21427428} = 0{,}008$$

$$Ratio\_type\_token_{Español} = \frac{Types}{Tokens} = \frac{218643}{21785302} = 0{,}010$$

De forma análoga, si se compara la repetición media de esas mismas formas en inglés y en español, los datos muestran diferencias significativas:

$$Rp\_formas_{Inglés} = \frac{21427428}{173391} = 123{,}5786$$

$$Rp\_formas_{Español} = \frac{21785302}{218643} = 99{,}638$$

La repetición media, por cada 21 millones de palabras, de una forma en inglés es de casi 124 veces, mientras que en español se llega sólo a 100. En consecuencia, en inglés las formas se repiten, como media, casi un 24% más que en español.

3.3. EL RITMO INCREMENTAL DE LAS FORMAS

Analicemos ahora las formas con mayor detalle. Para homogeneizar los resultados, es preciso determinar la relación entre palabras y formas mediante una medida estándar que nos permita establecer comparaciones directas entre ambas lenguas. Al tratarse de corpus con tamaños ligeramente divergentes, no utilizaremos la ratio entre las palabras (*token*) y las formas (*types*)[10]. En su lugar utilizaremos una fórmula para

[10] Medida conocida como *ratio type-token*.

conocer la dependencia entre las palabras y las formas, independientemente del tamaño desigual de los corpus, como es la propuesta por Sánchez y Cantos (1997)[11]:

$$Formas = K\sqrt{Palabras}$$

La medida *K* determina el ritmo incremental de nuevas formas dentro de un texto. Esta medida *K* es distinta para cada lengua, texto o autor, y actúa a modo de seña de identidad exclusiva, una especie de 'ADN'[12] personal e intransferible. Así, para calcular la medida *K* para el español e inglés bastará con transformar la fórmula original de Herdan en una expresión logarítmica, resultando:

$$K = \frac{Formas}{\sqrt{Palabras}}$$

Y obteniendo para el español el siguiente valor *K*:

$$K = \frac{218643}{\sqrt{21785302}} = \frac{218643}{4667,47} = 46,84$$

Y para el inglés:

$$K = \frac{173391}{\sqrt{21427248}} = \frac{173391}{4628,95} = 37,45$$

Estos valores confirman que el ritmo incremental de formas del español es mayor que el del inglés. Además, estos valores nos permiten comparar ambas lenguas en cuanto a la relación que guardan entre palabras y formas, y hacen posible proyectar los datos y apreciar mejor el ritmo incremental de formas en español e inglés. En el *Gráfico 1* se muestra el comportamiento y relación entre palabras y formas en español e inglés, proyectando los datos hasta 100 millones. Se aprecia cómo el español, por ser lengua más flexiva, contiene un repertorio de formas (*types*) más amplio que el inglés (ver también *Tabla 4*). El español cuenta con unas 21.000 formas de media más en un volumen de texto de 5 millones de palabras, cantidad que llega hasta casi 94.000 en 100 millones.

[11] Otros métodos alternativos son la *ratio estandarizada type-token* (Scott 1999), o los métodos propuestos por Tuldava (1995); Yang, Cantos y Song (2000); Chipere, Malvern, Richards y Duran (2001), entre otros.

[12] En Cantos (2000) se emplea este índice para la discriminación automática de textos dependiendo de la temática de los mismos.

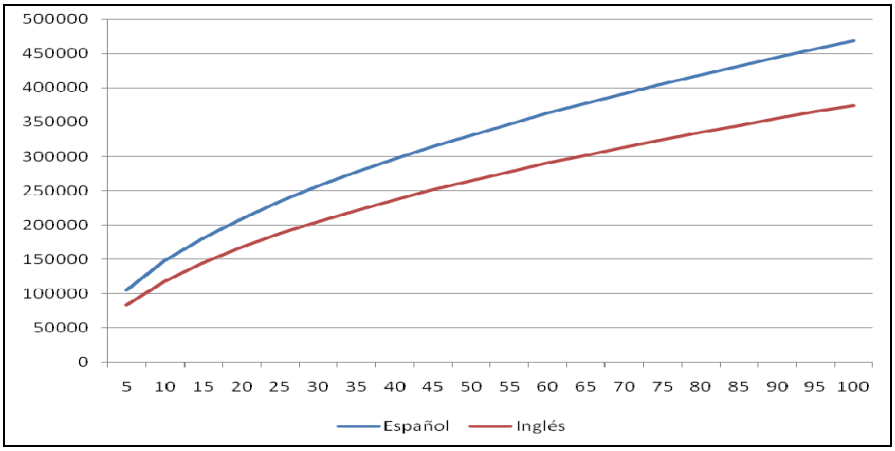

Gráfico 1. Ritmo incremental de formas en español e inglés

Palabras	Formas			
	Español	Inglés	Diferencia (E-I)	Incremento de la diferencia (E-I)
5000000	104737	83741	20997	20997
10000000	148121	118427	29694	8697
15000000	181411	145043	36367	6674
20000000	209475	167481	41993	5626
25000000	234200	187250	46950	4957
30000000	256553	205122	51431	4481
35000000	277109	221557	55552	4121
40000000	296242	236855	59388	3836
45000000	314212	251222	62990	3602
50000000	331209	264811	66397	3407
55000000	347375	277737	69638	3241
60000000	362821	290086	72735	3097
65000000	377636	301932	75705	2970
70000000	391892	313329	78562	2858
75000000	405646	324327	81320	2757
80000000	418950	334963	83987	2667
85000000	431843	345272	86572	2585
90000000	444363	355282	89081	2510
95000000	456540	365017	91522	2441
100000000	468400	374500	93900	2378

Tabla 4. Ritmo incremental de las formas en español e inglés

Es interesante observar que el aumento de formas no es lineal, sino de tipo polinómico, lo que justifica que dicho ritmo vaya decreciendo poco a poco. Esta realidad es lógica, ya que cuanto más volumen de texto tenemos, más improbable resulta encontrar formas nuevas (las formas ya han aparecido previamente a lo largo de los textos precedentes).

Este mismo fenómeno –aumento no lineal, sino polinómico de las formas– explica también cómo la diferencia incremental de formas en español es superior al inglés y sigue a su vez un patrón exponencial (*Gráfico 2*).

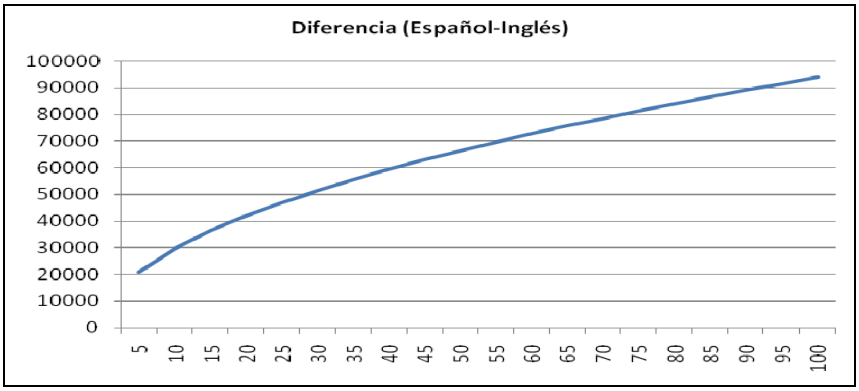

Gráfico 2. Diferencia del ritmo incremental de formas entre el español y el inglés

En consecuencia, es posible modelar el incremento de formas para el español y el inglés tomando como base las medidas *K* para ambas lenguas, 46,84 y 37,45 respectivamente. Así para calcular el hipotético volumen de formas de un corpus español de 1.000 millones de formas, aplicaremos la siguiente fórmula:

$$Formas = K\sqrt{Palabras}$$

Asumiendo que $K = 46{,}84$ y $Palabras = 1.000.000.000$, se concluye que un corpus español de 1.000 millones de palabras contendría un volumen de 1.481.210 formas:

$$Formas = 46{,}84\sqrt{1000000000}$$
$$Formas = 46{,}84 * 31622{,}77$$
$$Formas = 1481210{,}5468$$

Y un corpus de inglés de idéntica magnitud, llegaría a 1.184.272 formas, 296.938 menos que un corpus español de igual tamaño.

$$Formas = 37{,}45\sqrt{1000000000}$$
$$Formas = 37{,}45 * 31622{,}77$$
$$Formas = 1184272{,}73$$

Llevando la hipótesis a cantidades extremas, un corpus del español de 1 trillón[13] de palabras y otro de 1 trillón + 1 millón de palabras (es decir, un millón de palabras más que el anterior) contendrían la misma cantidad de formas diferentes (*types*). Es decir, no se incrementarían las formas nuevas, puesto que todas ellas estarían ya contenidas en el corpus. Habríamos llegado a un punto de inflexión en el que no valdría la pena añadir más muestras textuales, puesto que tendríamos recogida prácticamente 'toda la lengua'.

Llegados a este punto, es importante advertir que la proyección de datos a 1.000 millones de palabras ha sido hecha tomando como referencias las medidas *K* obtenidas a partir de los cálculos de los corpus *Cumbre* y *Lacell* para 21 millones de palabras. Lo cual significa que las proyecciones son únicamente válidas si los corpus de 1.000 millones de palabras a los cuales hemos hecho referencia mantienen la misma estructura y proporción en cuanto al contenido (textos de España e Hispanoamérica, -o del Reino Unido y demás países de habla inglesa)-, o porcentajes similares en las muestras orales y escritas, temática, etc. La proyección de datos de cualquier otro corpus del español y/o del inglés, con temáticas y estructuras distintas a las de *Cumbre* y *Lacell*, requeriría calcular previamente las constantes *K* en cada caso, constantes que podrían variar, incluso de forma sustancial, con respecto a los obtenidos para los corpus *Cumbre* y *Lacell*.

3.4. MÁS DATOS DISTRIBUTIVOS DEL ESPAÑOL Y DEL INGLÉS

La frecuencia del léxico es variada. ¿Lo es también su distribución atendiendo a criterios de grupos o bandas de frecuencia? Es decir, ¿Tienen una distribución homogénea las palabras muy raras o muy poco frecuentes, las palabras de frecuencia baja media, etc.? Y esa distribución, ¿es similar en inglés y en español? La *Tabla 5* ofrece datos clarificadores sobre la distribución de las palabras en tramos de frecuencia (según los corpus *Cumbre* y *Lacell*), tanto en español como en inglés:

Tramos de frecuencias	Frecuencia total por tramos y porcentajes			
	Español		Inglés	
1	85.350	39,04%	60.231	34,74%
2-5	53.331	24,39%	45.052	25,98%
6-20	44.046	20,15%	35.088	20,24%
21-100	23.200	10,61%	20.699	11,94%
101-1000	10.940	5,00%	10.040	5,79%
1001-	1.776	0,81%	2.281	1,32%

Tabla 5. Frecuencias léxicas y su distribución, según el Corpus *Cumbre* y *Lacell*

Según estos datos, el español cuenta, en términos absolutos, con casi 25.000 palabras más que el inglés usadas solamente una vez; lo que corresponde a un 5%

[13] Un millón de billones, o lo que es lo mismo, un uno seguido de 18 ceros.

más de *hápax legomena*, una vez normalizados los datos para su mejor contraste. En los restantes tramos de frecuencia, las diferencias son menores (frecuencia de 6-20, por ejemplo; *Gráfico 3*).

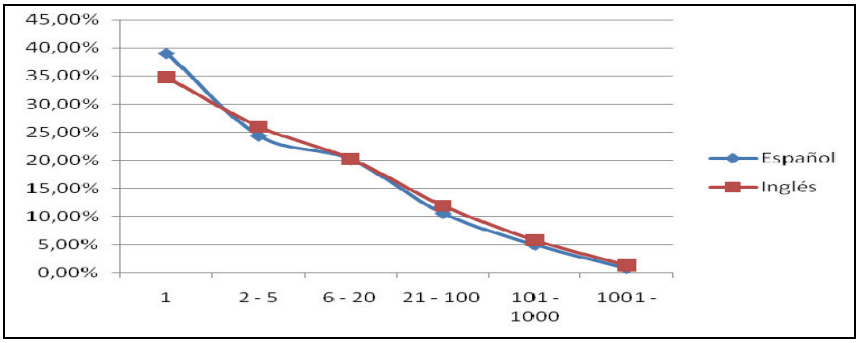

Gráfico 3. Frecuencias léxicas y su distribución en español e inglés

Si bien el número de formas (*types*) es mayor en español que en inglés, probablemente en razón de la mayor carga flexiva de aquel, el número de lemas no tiene por qué ajustarse exactamente al mismo parámetro, si bien tampoco cabe esperar desviaciones de importancia. Es difícil precisar el número de voces de una y otra lengua para aseverar sin reservas que una u otra cuenta con un acervo léxico más prolijo que la otra. Lo que sí cabe afirmar sin duda alguna es que el español es más flexivo que el inglés, especialmente en lo que respecta a las formas verbales y adjetivas. Ello hace que el español cuente con un inventario de formas (no necesariamente lemas) más elevado que el inglés. Al no disponer de ambos corpus lematizados, no es posible determinar con exactitud el número de lemas de cada corpus. Sí cabe identificar, no obstante, la cantidad de palabras léxicas y funcionales. La *Tabla 6* muestra los datos de los corpus *Cumbre* y *Lacell*, referidos a los tres parámetros especificados:

	Palabras	Palabras léxicas	Palabras funcionales
CUMBRE (español)	21.785.302	16.916.177	4.869.125
LACELL (inglés)	21.427.248	18.053.547	3.373.701

Tabla 6. Datos léxicos de los *Cumbre* y *Lacell*

Los datos reflejan que el uso de palabras léxicas (sustantivos, verbos, adjetivos y adverbios) es más frecuente entre angloparlantes que entre hispanohablantes. Los cálculos que siguen detallan las diferencias:

$$Palabras\_léxicas_{Inglés} = \frac{18.053.547}{21.427.248} = 0{,}8425$$

$$Palabras\_léxicas_{Español} = \frac{16.916.177}{21.785.302} = 0{,}7764$$

Mientras los primeros se valen, de media, de 84 palabras léxicas por cada cien palabras, los segundos reducen esta cifra a 78. Este hecho revela, en contrapartida, que el español abunda más que el inglés en el uso de palabras funcionales (preposiciones, artículos, pronombres, etc.).

$$Palabras\_funcionales_{Inglés} = \frac{3.373.701}{21.427.248} = 0{,}1574$$

$$Palabras\_funcionales_{Español} = \frac{4.869.125}{21.785.302} = 0{,}2235$$

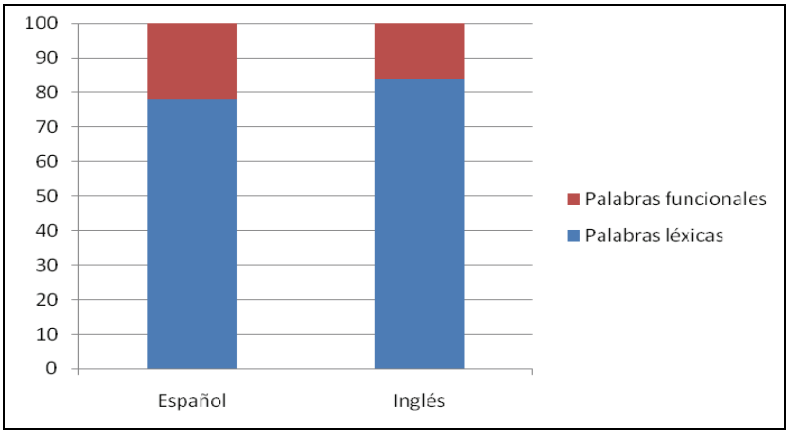

Gráfico 4. Distribución de palabras léxicas y funcionales en español e inglés

Se suele afirmar que el inglés es una lengua más sobria y concisa, menos proclive a la ornamentación léxica. Un reflejo de este supuesto podría detectarse en el número de oraciones usadas en cada lengua y en su longitud, asumiendo que frases más cortas implican el uso de menos recursos sintácticos y léxicos, y tienden a la concisión expresiva. El análisis de los dos corpus equivalentes, *Cumbre* y *Lacell*, demuestra en efecto, que las oraciones del español son, de media, más largas que las del inglés. Los datos de ambos corpus muestran que el corpus inglés contiene más frases (1.138.760) que el corpus español (1.042.417). La media de palabras por oración es consecuencia de estas cifras:

$$Ratio\_palabras/oración_{Inglés} = \frac{21.427.248}{1.138.760} = 18{,}8162$$

$$Ratio\_palabras/oración_{Español} = \frac{21.785.302}{1.042.417} = 20{,}8988$$

A raíz de tales datos, podemos definir cómo sería un texto tipo de 1.000 palabras en inglés y en español:

1. En inglés constaría de 53 oraciones (con 19 palabras por oración); 843 palabras léxicas; 157 palabras funcionales.
2. En español constaría de 48 oraciones (21 palabras por oración); 776 palabras léxicas; 224 palabras funcionales.

Gráficamente, una oración tipo para el inglés y español se ajustaría al siguiente esquema gráfico[14]:

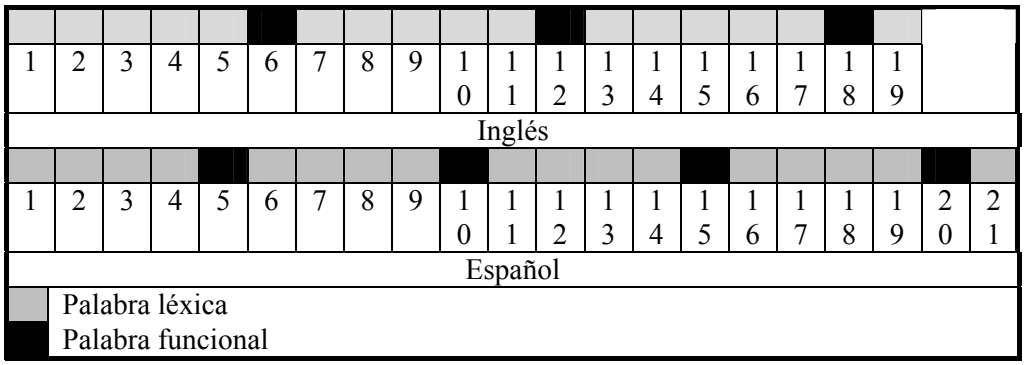

Gráfico 5. Oración tipo: español *versus* inglés

Es frecuente hacer referencia a otro tópico en torno a la longitud de las palabras: las palabras en inglés –se dice– son más cortas que en español. De ahí –se arguye– la tendencia 'monosilábica' de la lengua anglosajona, menos perceptible en español debido a la mayor extensión de las palabras. El cómputo de las palabras de 1 a 10 caracteres en ambas lenguas confirma la tendencia del inglés hacia la 'concisión ortográfica'. Los datos globales son los siguientes (*Tabla 7*):

	Español	*Inglés*
Media de letras por palabra	4,423	3,992

Tabla 7. Letras por palabras

La media general en la longitud de las palabras no presenta discrepancias de relieve entre ambos idiomas. Pero el contraste del número de palabras con igual cantidad de letras sí es bastante más significativo en algunos casos. Destacan especialmente (i) las palabras de una sola letra, que son más de tres veces superiores en número en inglés que en español (aunque no debe olvidarse la escasa importancia de este tipo de palabras en inglés: se reducen al artículo 'a' –con casi medio millón de

[14] Cada celda equivale a una palabra.

casos– y a las letras del alfabeto usadas para abrir párrafos o similares); (ii) las palabras de dos letras, cuyo número en inglés supera al del español en casi dos millones; o (iii) las palabras de 4 letras, que en inglés casi duplican el volumen del español. La *Tabla 8* y el *Gráfico 6* recogen los datos referidos a palabras de 1 a 10 letras en cada idioma:

Letras por palabra	Español		Inglés	
	Frecuencias	*%*	*Frecuencias*	*%*
1 letra	816.275	3,75	2.656.653	12,40
2 letras	3.642.015	16,72	5.335.686	24,90
3 letras	4.366.524	20,04	3.335.324	15,57
4 letras	3.637.621	16,70	1.939.735	9,05
5 letras	2.408.555	11,06	2.174.200	10,15
6 letras	1.804.230	8,28	1.579.171	7,37
7 letras	1.655.522	7,60	1.439.280	6,72
8 letras	1.150.188	5,28	1.162.258	5,42
9 letras	830.798	3,81	788.230	3,68
10 letras	533.148	2,45	595.246	2,78
Más de 10 letras	940.426	4,32	421.465	1,97

Tabla 8. Cantidad de palabras según el número de letras que las integran

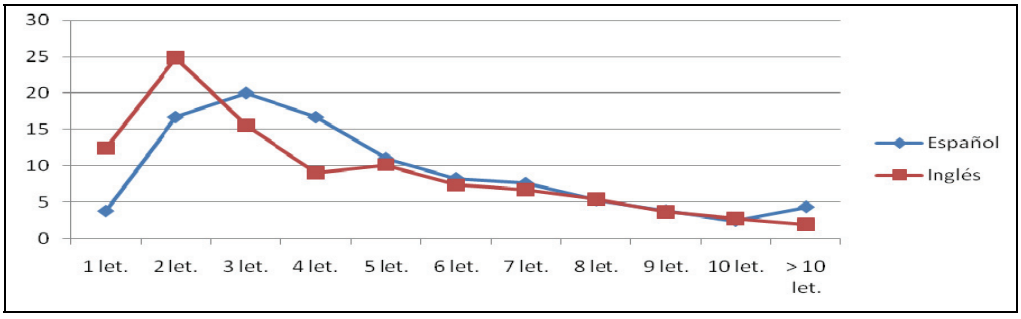

Gráfico 6. Distribución comparativa (español-inglés) de palabras según el número de letras

3.5. USO DE VOCALES Y CONSONANTES

En las aplicaciones del lenguaje a diferentes ámbitos industriales y computacionales es muy importante conocer determinados aspectos cuantitativos básicos, como son la frecuencia de vocales y de consonantes y la tipología y frecuencia de determinadas secuencias de letras. Estos hechos son decisivos a la hora de elaborar estadísticas para la corrección ortográfica, por ejemplo, o para la síntesis del habla. Respecto al uso de cada letra en inglés y en español, hay diferencias significativas tanto en lo referido a las vocales como a las consonantes. Llama la atención la diferencia de frecuencia en

el uso de la vocal 'a' en cada idioma; pero el resto de vocales no presenta contrastes notables (*Tabla 9* y *Gráfico 7*). También existen diferencias relevantes en la distribución de las vocales a final de palabra, especialmente si se trata de la 'a' y la 'o', tal y como reflejan los datos de la *Tabla 9* y el *Gráfico 8*.

Vocales	Inglés		Español	
	%	% (posición final)	%	% (posición final)
E	11,8	3,53	12	2,78
A	7,7	0,57	11,5	2,98
O	7,2	0,82	8,1	1,95
I	7,3	0,12	7	0,07
U	2,6	0,06	3,6	0,09
TOTAL	36,6	5,1	42,1	7,87

Tabla 9. Uso y distribución de las vocales en inglés y en español

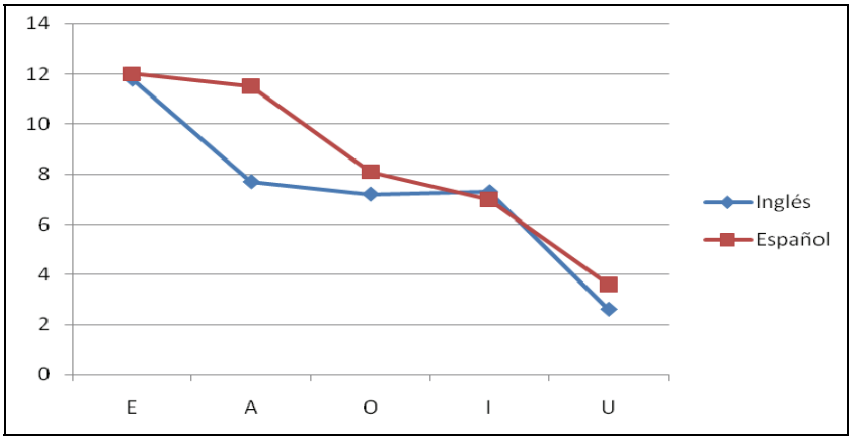

Gráfico 7. Distribución comparativa (español-inglés) del uso de las vocales

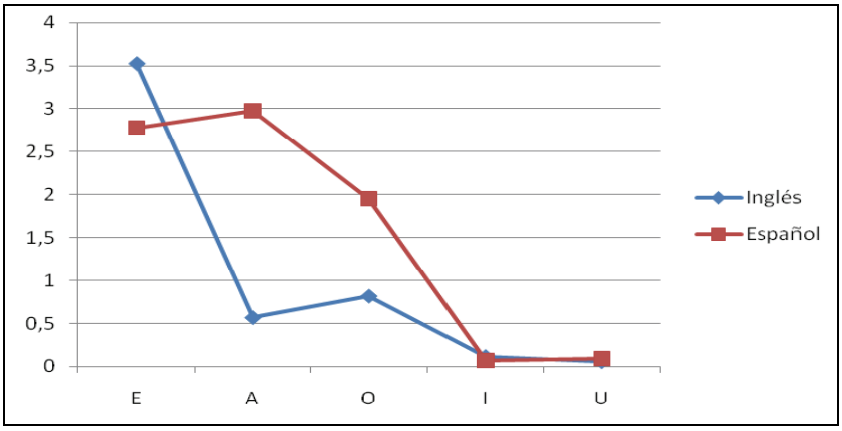

Gráfico 8. Distribución comparativa (español-inglés) del uso de las vocales en posición final

Respecto a las consonantes, los corpus analizados reflejan los siguientes datos (en orden de frecuencia):

Consonantes	*Inglés (%)*	*Español (%)*
T	8,9	4,3
N	6,7	6,7
S	6,5	7
R	5,9	6,2
H	5	0,6
D	3,5	5
P	1,9	2,6
F	2,1	0,7
M	2,3	2,5
W	1,7	0,006
Y	1,6	0,8
G	1,9	0,99
V	0,9	0,93
K	0,5	0,01
Q	0,09	0,8
X	0,19	0,14
J	0,2	0,44
Z	0,07	0,35
Ñ	0	0,19
B	1,4	1,2

Tabla 10. Frecuencia de consonantes en inglés y en español

Obsérvese que en la tabla anterior se han sombreado los casos con las diferencias más significativas. Especialmente revelador resulta la diferencia de uso en las consonantes *t, h, f* y *w*. En conjunto, sin embargo, predominan las semejanzas entre ambas lenguas (Gráfico 10).

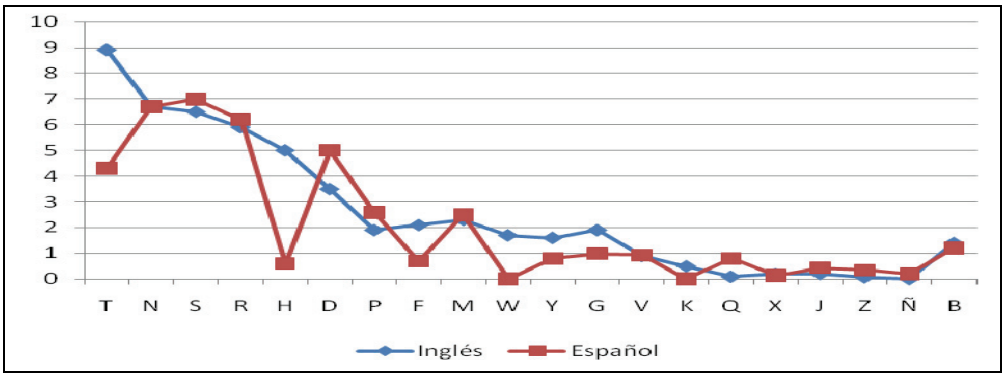

Gráfico 9. Distribución comparativa (español-inglés) del uso de las consonantes

En lo que se refiere a dígrafos consonánticos, dos de ellos contrastan en ambos idiomas, la *ch* (con notables diferencias) y la *sp*:

Dígrafos	*Inglés (%)*	*Español (%)*
CH	0,5	0,15
SP	0,15	0,2

Tabla 11. Dígrafos: *ch* y *sp*

3.5. NATURALEZA DE LAS PALABRAS MÁS FRECUENTES Y SU DISTRIBUCIÓN

Aparte del estudio comparativo de las frecuencias léxicas, cabe también preguntarse si esas mismas frecuencias se distribuyen homogéneamente en ambos idiomas. A tal fin, es preciso seleccionar un determinado tramo de frecuencia en cada lengua y analizar las palabras que lo integran. En la *Tabla 12* se ofrece la lista de las 100 formas más frecuentes en inglés y en español, especificando la frecuencia de cada forma en su respectivo corpus:

	Inglés		Español			Inglés		Español	
	Frecuencia	Forma	Frecuencia	Forma		Frecuencia	Forma	Frecuencia	Forma
1	1263621	The	1216593	De	51	44243	Your	28013	Había
2	620841	Of	770656	La	52	44169	Do	27671	Nos
3	584526	And	673494	Que	53	43763	Up	27604	Años
4	565223	To	563139	Y	54	43620	No	26605	Tiene
5	480032	A	557608	El	55	42778	Out	26062	Hasta

6	398530	In	522624	En	56	40807	Her	25753	Desde
7	264186	That	429826	A	57	39222	Some	25473	Te
8	234752	Is	327038	Los	58	39011	Said	25374	Eso
9	227414	It	265864	Se	59	36647	Them	24739	Fue
10	190559	For	233880	No	60	35753	Time	24580	Todos
11	172430	You	222746	Un	61	34890	People	23953	Puede
12	154364	On	215618	Las	62	34832	Other	23722	Pues
13	147579	Was	211604	Del	63	34319	Er	23513	Han
14	141415	With	202115	Por	64	34030	Two	23160	Así
15	138725	Be	183781	Con	65	33465	Like	23031	Bien
16	133777	As	181191	Una	66	33431	My	21350	Vez
17	115053	Are	176187	Es	67	33309	Into	21134	Ni
18	110872	He	137537	Lo	68	32539	Then	20690	Sólo
19	109124	At	135083	Para	69	31961	Well	20130	Ahora
20	104301	This	127901	Su	70	31770	Very	19975	Él
21	103977	Have	113580	Al	71	31522	Its	19047	Uno
22	103853	We	102577	Como	72	31353	Now	18868	Parte
23	100022	By	93074	Más	73	30823	Than	18736	Ese
24	99428	They	72694	Pero	74	30335	Just	18128	Tiempo
25	97517	But	69189	Me	75	30102	Only	18086	Vida
26	90445	Or	61247	Le	76	29335	New	17886	Mismo
27	90188	From	55331	Ha	77	29321	Think	17405	Otro
28	89997	Not	54089	Sus	78	28725	Any	17357	Día
29	73227	There	53544	Si	79	28616	Know	17225	Cada
30	72231	An	47042	Yo	80	27700	These	17146	Hacer
31	71331	His	45649	Ya	81	27575	May	17000	Siempre
32	71104	Which	42028	Este	82	27526	Could	16891	Entonces
33	67826	One	38804	Porque	83	26846	Our	16844	Nada
34	64639	Will	38398	Muy	84	26760	Also	16768	Donde
35	63656	All	38351	Todo	85	26693	First	16764	Esa
36	62922	Had	37518	Cuando	86	26178	Me	16658	Hace
37	60724	If	37472	Qué	87	25206	How	16574	Bueno
38	59822	Can	36372	Sin	88	25144	Over	16543	Decir
39	57838	Has	36067	Son	89	25092	Because	16504	Tan
40	57311	So	35467	Sobre	90	25002	Him	16485	Otra
41	57144	What	34596	Está	91	24326	Re	16388	Esto
42	55547	Their	33752	También	92	24308	Get	15823	Después
43	54671	Were	33333	Esta	93	23554	See	15733	Ella
44	50627	About	33141	Hay	94	22873	After	15282	Menos
45	47932	Been	33067	Sí	95	22668	Most	15269	Tanto
46	47928	More	31747	Entre	96	22593	Don't	15232	Otros
47	47604	Would	31317	Ser	97	22585	Where	15203	Mundo
48	47299	When	31293	Era	98	22230	Should	15182	Aquí
49	46355	Who	30871	Mi	99	21642	Many	15163	Va
50	44866	She	29334	Dos	100	21470	Way	14904	Poco

Tabla 12. Las 100 formas más frecuentes en inglés y en español

Una primera valoración comparativa muestra un comportamiento aparentemente similar, dibujando una curva de más a menos y a un ritmo exponencial negativo; es decir, una función hiperbólica, similar a una cotangente hiperbólica (*Gráfico 10*).

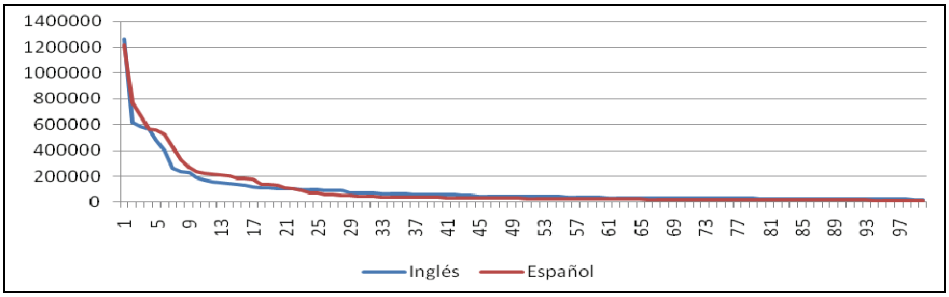

Gráfico 10. Distribución comparativa (español-inglés) de las frecuencias de las palabras más usadas

Un análisis más detallado, centrado en las diferencias entre frecuencias, revela diferentes patrones de comportamiento entre el inglés y el español:

1. Hay oscilaciones importantes entre las aproximadamente 21 formas más usadas (salvo la primera y la cuarta), a favor del español; es decir, el español hace un uso más extensivo de las palabras más usadas.
2. Esta tendencia se torna a favor del inglés de la palabra 22 en adelante, o lo que es lo mismo, con palabras cuyas frecuencias absolutas están en torno a 100.000 (≈ 0,5% de frecuencia relativa). A partir de este punto el inglés se vuelve más repetitivo que el español (*Gráfico 11*).

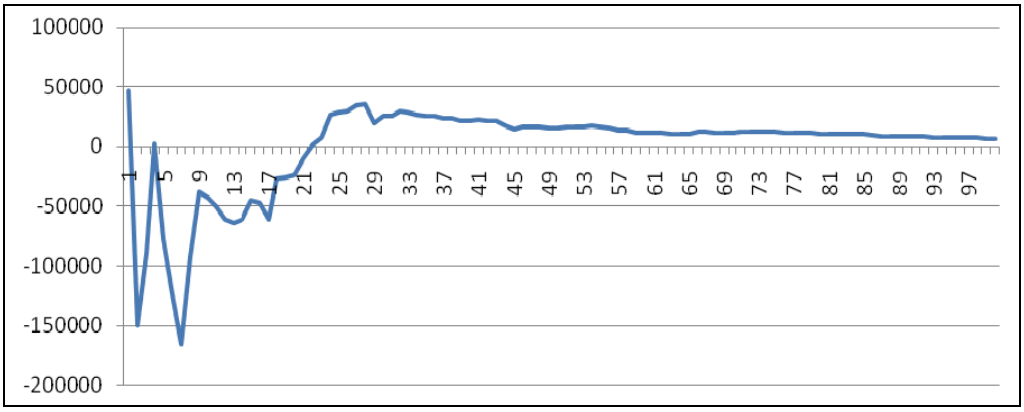

Gráfico 11. Distribución comparativa y diferencial de las frecuencias de las palabras más usadas (inglés y español)

En cuanto a los términos de mayor frecuencia, las palabras funcionales ocupan los primeros puestos en ambas lenguas, si bien es verdad que su orden en la lista no es plenamente coincidente. El término inglés *the* (artículo) es la palabra más frecuente, que se corresponde en español con *de* (preposición). Aunque en realidad esta diferencia en categoría gramatical se debe a razones flexivas. Si sumamos todas las

formas flexivas del artículo determinado en español, en singular y en plural (*el, la, los, las*), el resultado es superior al *the* inglés: la suma asciende a 1.870.920 veces (un 8,58% del total de palabras usadas en los 21 millones), muy por encima de la preposición *de* (5,53%) o del artículo *the* (5,89%). A su vez, la segunda palabra más frecuente en inglés, *of* (preposición), se usa casi tres veces por cada cien palabras, mientras que su equivalente española, *de*, casi la duplica en el uso (*Tabla 13*).

	Inglés	Español
The frente a *El/la/los/las*	5,89%	8,58%
Of frente a *De*	2,89%	5,53%

Tabla 13. Voces más frecuentes en inglés y español: *the* y *of*; *de* y *el/la/los/las*

También destaca el hecho de que en inglés las formas verbales más frecuentes son, y en este orden, *was, be, are, have*, mientras que el español prefiere *es, ha, hay, había*, mostrando una inversión parcial respecto a la preeminencia de (*to) be,* seguido de *(to) have*.

Respecto a los sustantivos, es interesante advertir que estos aparecen en posiciones sorprendentemente retrasadas: *time* ocupa la posición 60 en la lista, y *años* la posición 53. La preeminencia de las voces funcionales y de las formas verbales es evidente en ambos idiomas.

La comparación de la frecuencia de los lemas (*Tabla 14* y *Gráfico 11*) no varía en algunos de los rasgos más sobresalientes. Destaca el hecho de que, globalmente, las frecuencias son muy similares en ambos idiomas y que las palabras funcionales siguen en cabeza:

1. El lema menos frecuente (de los 100 primeros) presenta una diferencia de unas 2.000 ocurrencias a favor del inglés.
2. El verbo *to be* dobla con creces la frecuencia de *ser* y *estar* juntos.
3. *To have* casi dobla la frecuencia de *haber*, mientras que *to do* y *hacer* son muy similares en este parámetro.
4. En las cien primeras voces, el español incluye 40 términos léxicos, mientras que el inglés sólo cuenta con 29.

	Inglés		Español			Inglés		Español	
	Frecuencia	Lema	Frecuencia	Lema		Frecuencia	Lema	Frecuencia	Lema
1	1263621	the	1870920	el	51	38332	see	30482	mismo
2	847926	be	1216684	de	52	38553	know	30463	dar
3	618888	of	710981	que	53	36685	time	29720	parir
4	537572	and	563142	y	54	35844	take	29225	parar
5	437273	a	522628	en	55	34682	them	28937	bueno
6	384863	in	429846	a	56	34234	some	28889	querer
7	324170	to	319626	ser	57	33677	could	28159	día

8	275127	have	316814	un	58	33464	so	27673	nos
9	218037	it	301055	lo	59	33002	him	27666	ha
10	205866	to	233886	no	60	32786	year	26062	hasta
11	177575	for	211605	del	61	32693	into	25752	desde
12	176819	I	202117	por	62	32616	its	25481	te
13	152079	that	183787	con	63	32130	then	25461	deber
14	139099	you	165596	saber	64	30776	think	24747	bien
15	136251	he	139412	haber	65	30523	my	23724	pues
16	136147	on	132957	se	66	30374	come	23655	entrar
17	135005	with	113593	al	67	29523	than	22848	alguno
18	111919	do	111401	estar	68	29205	more	22824	ello
19	106832	at	104468	tener	69	28910	about	22301	nuestro
20	103434	by	98029	hacer	70	28760	now	21881	esa
21	93097	not	93082	más	71	28012	last	21558	llegar
22	92389	this	89433	este	72	27666	your	21136	ni
23	91924	but	86639	si	73	27630	me	21102	partir
24	86906	from	85501	todo	74	27405	no	20690	sólo
25	86688	they	82940	o	75	27037	other	20677	donde
26	85379	his	82280	como	76	26283	give	20592	pasar
27	76862	that	81111	poder	77	25703	just	20523	tiempo
28	76051	she	81056	para	78	25678	should	20359	político
29	74761	or	79162	ir	79	25088	these	20132	ahora
30	74406	which	76696	decir	80	25086	people	19981	él
31	72832	as	72713	pero	81	24976	also	19844	primero
32	71607	we	71898	uno	82	24890	well	19741	tanto
33	68612	an	70274	me	83	24731	any	19733	poner
34	66703	say	68077	le	84	24425	only	19628	poco
35	59456	will	58391	unir	85	23104	new	19569	ese
36	54469	would	50682	otro	86	22982	very	19428	país
37	53223	can	47054	yo	87	22731	when	18721	vida
38	52217	if	45656	ya	88	22604	may	18518	usted
39	52183	their	40492	año	89	22527	way	18284	hombre
40	49908	go	39674	porque	90	22211	look	18243	sobrar
41	49893	what	39033	comer	91	22018	like	18077	son
42	47892	there	38854	cuando	92	21764	use	17839	parecer
43	46147	all	38759	ver	93	21742	her	17733	sobre
44	44188	get	38399	muy	94	21704	such	17626	quien
45	43651	her	38016	mi	95	20701	how	17336	nuevo
46	43453	make	36372	sin	96	20600	because	17226	cada
47	41086	who	34593	mucho	97	20524	when	17015	siempre
48	40393	as	33765	también	98	20316	as	16964	llevar
49	40363	out	31468	vez	99	20130	good	16893	entonces
50	39085	up	31399	eso	100	19779	find	16842	hablar

Tabla 14. Listado de los 100 lemas más frecuentes

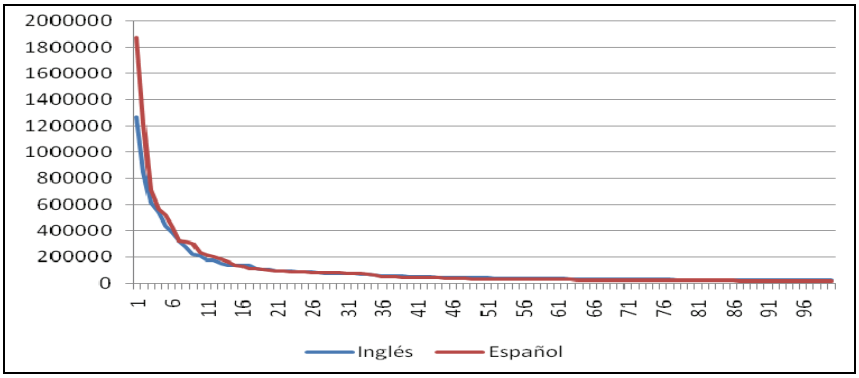

Gráfico 11. Distribución comparativa (español-inglés) de las frecuencias de los lemas más usados

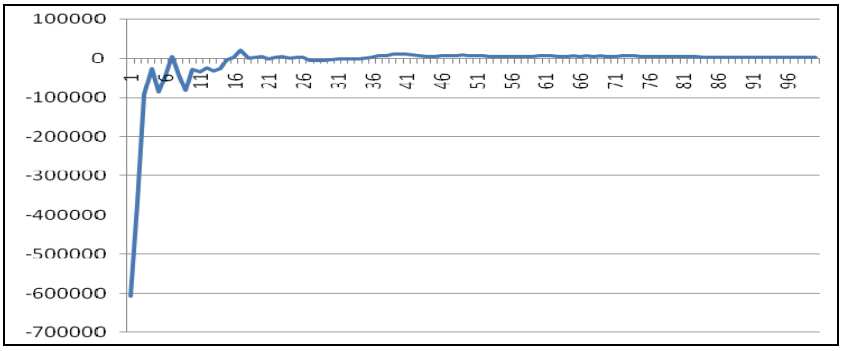

Gráfico 12. Distribución comparativa y diferencial de las frecuencias de los lemas más usadas

4. TIPIFICACIÓN PROTOTÍPICA DE UN TEXTO EN ESPAÑOL Y EN INGLÉS

A tenor de los datos obtenidos de las dos lenguas, español e inglés, es posible mostrar el perfil de un texto prototipo en cada una de las lenguas mencionadas. Así un texto de unas 25.000 palabras en español y otro de igual tamaño en inglés, reflejarían los siguientes rasgos y composición léxica interna:

	Español	**Inglés**	**Diferencia en %**
Palabras (tokens)	25.000	25.000	-
Formas (types)	7.406	5.921	25,08[15]

[15] Los valores positivos señalan índices a favor del español y los negativos a favor del inglés.

Ratio de formas nuevas[16]		29	23	26,09
Ratio de repetición de formas		3,37	4,22	-20,14
Distribución de formas por tramos de frecuencia	1	9.760	8.685	12,38
	2-5	6.098	6.495	-6,11
	6-20	5.038	5.060	-0,43
	21-100	2.653	2.985	-11,12
	101-1000	1.250	1.448	-13,67
Palabras léxicas		19.413	21.065	-7,84
Ratio de palabras léxicas		77,65	84,26	-7,84
Palabras funcionales		5.588	3.935	42,01
Ratio de palabras funcionales		22,35	15,74	41,99
Oraciones		1.196	1.329	-10,01
Ratio de palabras por oraciones		20,90	18,81	11,11
Cantidad de palabras según el número de letras	1 letra	938	3.100	-69,74
	2 letras	4.180	6.225	-32,85
	3 letras	5.010	3.893	28,69
	4 letras	4.175	2.263	84,49
	5 letras	2.765	2.538	8,94
	6 letras	2.070	1.843	12,32
	7 letras	1.900	1.680	13,10
	8 letras	1.320	1.355	-2,58
	9 letras	953	920	3,59
	10 letras	613	695	-11,80
	> 10 letras	1.080	493	119,07
Uso de vocales en palabras	A	3.000	2.950	1,69
	E	2.875	1.925	49,35
	I	2.025	1.800	12,50
	O	1.750	1.825	-4,11
	U	900	650	38,46
Uso de vocales en palabras (posición final)	A	695	883	-21,29
	E	745	143	420,98
	I	488	205	138,05
	O	18	30	-40,00
	U	23	15	53,33
Uso de consonantes en palabras	T	1075	2225	-51,69
	N	1675	1675	0,00
	S	1750	1625	7,69
	R	1550	1475	5,08
	H	150	1250	-88,00
	D	1250	875	42,86
	P	650	475	36,84
	F	175	525	-66,67

[16] Las ratios se expresan por cada 100 ítems.

	M	625	575	8,70
	W	2	425	-99,53
	Y	200	400	-50,00
	G	248	475	-47,79
	V	233	225	3,56
	K	3	125	-97,60
	Q	200	23	769,57
	X	35	48	-27,08
	J	110	50	120,00
	Z	88	18	388,89
	Ñ	48	0	-
	B	300	350	-14,29
Uso de dígrafos	CH	38	125	-69,60
	SP	50	38	31,58

Tabla 15. Resumen cuantitativo del español *versus* inglés

En resumen, el español es más prolijo que el inglés en formas distintas (*types*) y tiene, por tanto, una mayor ratio de variedad de formas. No obstante y a pesar de esta mayor diversidad de formas, el español usa un repertorio léxico menos variado respecto al inglés, y tiende a repetir con mayor frecuencia las palabras léxicas que utiliza.

La distribución de grupos de palabras por frecuencias también tiene patrones diferenciados en ambas lenguas. Las diferencias se hacen especialmente patentes en los extremos de la banda de frecuencia: en las palabras 'raras' (palabras que ocurren una sola vez: *hapax legomena*), más frecuentes en español, y en las palabras más comunes, más frecuentes en inglés, especialmente a partir de la frecuencia de 8 palabras por cada 10.000[17] (en muestras de 25.000). Todo ello indica que el español hace un mayor acopio de palabras 'raras' y repite comparativamente menos las más frecuentes.

Aunque la sintaxis y el discurso no son el objeto de nuestro estudio, la longitud media de las oraciones en ambos idiomas apunta a que el español formula y construye periodos oracionales más largos que el inglés, a la vez que contiene, tal y como se vio con anterioridad, una mayor concentración de palabras funcionales.

En cuanto al patrón distributivo de las palabras atendiendo a su magnitud gráfica, el comportamiento en ambas lenguas es bastante dispar: el uso de palabras de uno o dos caracteres es más frecuente en inglés que en español; la tendencia se torna a favor del español en palabras con tres y cuatro letras y se agudiza en el uso de palabras polisilábicas con más de diez caracteres, en cuyo ámbito la lengua española duplica a la inglesa.

[17] Este dato se ha obtenido dividiendo 21 por 25.000; 21 : 25.000 = 0,00084.

5. ALGUNAS CONSIDERACIONES FINALES

El análisis llevado a cabo en este artículo ha sido de carácter cuantitativo, basado en dos corpus equivalentes de poco más de 20 millones de palabras cada uno. Las proyecciones realizadas deben ser entendidas, por tanto, en su justa dimensión, con las limitaciones y bondades propias de las muestras utilizadas. Cabe afirmar, sin embargo, que aunque un corpus más amplio habría aportado una mayor precisión y refinamiento en algunos de los datos extraídos, lo esencial de lo expresado a lo largo de estas páginas no habría cambiado.

De otra parte, no cabe la menor duda de que los datos cuantitativos inciden directamente en los datos cualitativos que podrían extraerse. La frecuencia de determinadas palabras, por ejemplo, no es inocua: refleja el índice de uso de los conceptos que aquellas conllevan y, por ende, el grado de importancia que dichos conceptos tienen en la sociedad de hablantes nativos.

Con frecuencia se enfatizan las diferencias entre lenguas. Pero nuestro estudio demuestra que también habría que poner de relieve lo que comparten o tienen en común. Las lenguas tienden a diferenciarse en aspectos formales, pero mantienen muchos otros elementos básicos, de carácter estructural. La comparación del español y del inglés no deja lugar a dudas sobre los rasgos que ambas lenguas comparten en aspectos léxicos y estructurales, aspectos que no deberían dejarse de lado en estudios contrastivos. Las semejanzas o diferencias pueden referirse a la cuantía léxica, a la distribución léxica, a la longitud de las oraciones, a la longitud de las palabras, al uso de vocales y consonantes, al peso de las palabras funcionales y léxicas, a los sonidos, y en general, a la simetría en parámetros tanto ortográficos como morfológicos y sintácticos. Existen también semejanzas en la organización conceptual que subyace en las palabras, pero este tema requeriría disponer de datos que aún no son fáciles de obtener automáticamente con las herramientas de análisis con que contamos.

Los estudios contrastivos entre lenguas no gozan actualmente de la popularidad que gozaron en la década de los sesenta, pero las recopilaciones textuales que actualmente están a nuestro alcance permiten alcanzar resultados no solamente más ambiciosos, sino también más fiables. En esa línea de futuro pretendemos situar nuestro trabajo, invitando a retomar una línea de investigación que la disponibilidad de los corpus augura como de gran utilidad, además de prometedora.

REFERENCIAS

Barnbrook, G. (1996). *Language and Computers*. Edinburgh: Edinburgh University Press.

Cantos, P. (1995). Tratamiento informático y obtención de resultados. In Sánchez, ed., 39-70.

Cantos, P. (2000). Investigating Type-token Regression and its Potential for Automated Text Discrimination. *Cuadernos de Filología Inglesa* 9(2)*,* 71-91.

Cantos, P. and A. Sánchez (2001). Lexical Constellations: What Collocates Fail to Tell. *International Journal of Corpus Linguistics* 6(2), 199-228.

Castagno, J. P., ed. (2006). *Encyclopedia Americana*. Connecticut: Grolier Incorporated.

Chipere, N., D. Malvern, B. Richards and P. Duran (2001). Using a Corpus of School Children's Writing to Investigate the Development of Vocabulary Diversity. In Rayson, Wilson, McEnery, Hardie and Khoja, eds., 126-133.

Church, K.W., W. Gale, P. Hanks and D. Hindle (1991). Using Statistics in Lexical Analysis. In Zernik ed., 115-164.

Hoover, D. L. (2003). Another Perspective on Vocabulary Richness. *Computers and the Humanities* 37(2), 151-178.

Hunston, S. and G. Francis (1999). *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English.* Amsterdam: John Benjamins.

Mcenery, T. and A. Wilson (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McCrum, R., W. Cran and R. MacNeil. (1992). *The Story of English*. New York: Penguin.

Rayson, P., A. Wilson, T. McEnery, A. Hardie and S. Khoja, eds. (2001). *Technical Papers. Volume 13. Special Issue. Proceedings of the Corpus Linguistics 2001 Conference.*

Sánchez, A., ed. (1995). *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*. Madrid: Sociedad General Española de Librería.

Sánchez, A. and P. Cantos (1997). Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-Million-Word Corpus of Contemporary Spanish'. *International Journal of Corpus Linguistics* 2(2): 259-280.

Sánchez, A., and P. Cantos (1998). El ritmo incremental de palabras nuevas en los repertorios de textos. Estudio experimental y comparativo basado en dos corpus lingüísticos equivalentes de cuatro millones de palabras, de las lenguas inglesa y española y en cinco autores de ambas lenguas. *ATLANTIS* XIX(2): 205-223.

Sánchez, A., ed. (2001). *Gran diccionario de uso del español actual*. Madrid: Sociedad General Española de Librería.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

The New Oxford Dictionary of English (1998). Oxford: Oxford University Press.

Tuldava, J. (1995). *Methods in Quantitative Linguistics*. Trier: Wissenschaftlicher Verlag Trier.

Webster's Third New International Dictionary (2010). Springfield: Merriam Webster Inc.

Woods, A., P. Fletcher and A. Hughes (1986). *Statistics in Language Studies*. Cambridge: Cambridge University Press.

Yang, D. H., P. Cantos, M. Song (2000). An Algorithm for Predicting the Relationship between Lemmas and Corpus Size. *International Journal of the Electronics and Telecommunications Research Institute (ETRI)* 22: 20-31.
Zernik, U. ed. (1991). *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Zipf, G. (1935). *The Psycho-Biology of Language*. Boston: Houghton Mifflin.