Frecuencia y distintividad en el uso lingüístico: casos tomados de la lematización verbal de corpus de distintos registros

Antonio Moreno Sandoval
Universidad Autónoma de Madrid
José María Guirao Miras
Universidad de Granada

ABSTRACT
El objeto de esta comunicación es establecer una comparación entre el concepto de frecuencia de uso y el de distintividad. Para ello utilizaremos el caso de los lemas verbales del español en diferentes registros y variedades del español. La metodología consistirá en emplear tres corpus lematizados automáticamente en el Laboratorio de Lingüística Informática de la UAM. Los tres corpus representan el habla espontánea de la variedad peninsular (C-ORAL-ROM), el habla infantil espontánea (CHIEDE) y un corpus de noticias de la Agencia EFE. Para realizar la extracción de unidades distintivas, aplicamos el test de Dunning a cada corpus, confeccionando una lista de lemas verbales distintivos del registro. El resultado final son dos listas de verbos para cada registro. En el Apéndice se muestran los 100 primeros casos de cada tipo textual, propuesta que puede ser de interés para la enseñanza de segundas lenguas y la lexicografía.

PALABRAS CLAVE: corpus orales y escritos; variedades del español; lematización; frecuencia de uso,

I. FRECUENCIA DE USO Y DISTINTIVIDAD

El concepto de frecuencia de uso es recurrente y básico en gran parte de la investigación en Lingüística de Corpus, así como en otras áreas como Análisis del Discurso, Sociolingüística, Fonología o Lingüística Histórica. Joan Bybee (2007) es un buen ejemplo de la argumentación a favor de la frecuencia de uso como factor esencial (funcional) en el análisis y explicación de la estructura de las lenguas naturales. Bybee, representante del funcionalismo, destaca alguna de las características más importantes de la frecuencia de uso:

> To the uninitiated, it does not seem unreasonable at all to suppose that high-frequency words and expressions might have one set of properties and low-frequency words and expressions another. (Bybee 2007: 5)

Precisamente lo que Bybee trata de demostrar en su libro es cómo la frecuencia y la repetición de formas y unidades lingüísticas influyen en fenómenos como la gramaticalización (cambio y creación de estructuras) o la marcación (la formas o unidades no marcadas son más frecuentes que las marcadas). La hipótesis básica de bastantes corrientes empiricistas y funcionalistas es que "specific instances of experiences give rise to generalizations, and they can do so without being swallowed up themselves by the general pattern.[…] The reason frequency or repetition plays a role in

grammar formation is that the mind is sensitive to repetition" (Bybee 2007: 7-8).

Como es bien sabido, en el recuento de frecuencias se puede distinguir entre ejemplos (*tokens*) y tipos (*types*). En nuestro caso de estudio, cada forma verbal de un verbo es un ejemplo de dicho verbo (tipo)[1]. Así, *amo, amas, ama*… son ejemplos de *AMAR*. De esta manera, el recuento de todas las formas que aparecen en un corpus y que pertenecen al mismo lema supone una abstracción de los ejemplos al concepto más abstracto, la unidad léxica verbal.

En nuestro experimento vamos a utilizar la lematización y la frecuencia de tipo como elemento de discusión. Queremos analizar la posible influencia de la repetición de la unidad léxica verbal en la determinación de las características de una variedad o registro lingüísticos.

En contraposición a la frecuencia oponemos el concepto de distintividad, entendido como unidad léxica que es característica y distintiva de una variedad porque aparece típicamente en dicho registro y no tanto en otros. Informativamente, una unidad distintiva destaca porque proporciona más información sobre el contenido y contexto. Igualmente, la probabilidad de que dicho lema aparezca en un determinado registro es más alta que en otros contextos. Esta idea ha sido expuesta por Dunning (1993): se observa que las palabras más representativas de un texto o de un registro tienen una frecuencia de aparición muy baja (dos o tres veces) con respecto a otras más generales en cualquier corpus. Sin embargo, dichas palabras identifican muy claramente el contenido o la tipología del texto.

Por tanto, *palabra más frecuente* y *palabra más significativa* son dos conceptos que no deben confundirse y no necesariamente equivalentes. El objeto de este artículo es mostrar las correspondientes listas extraídas de diferentes corpus para ver hasta qué punto la distinción es productiva en aplicaciones de la Lingüística, como la enseñaza de lenguas o la terminología.

II. METODOLOGÍA

II.1. Los corpus empleados

Para nuestro experimento emplearemos tres corpus que han sido anotados automáticamente por medio un analizador morfosintáctico, GRAMPAL (Moreno 1991, Moreno y Guirao 2006). En el caso de los dos corpus de habla espontánea, los resultados de la lematización han sido revisados a mano por lingüistas. En el caso del corpus escrito, dado el tamaño, no se ha procedido todavía a la revisión.

II.1.1. C-ORAL-ROM

El corpus C-ORAL-ROM está compuesto por cuatro subcorpus, comparables en tamaño y distribución, en cuatro lenguas romances (francés, italiano, portugués y español). Se trata de un corpus de habla espontánea, de carácter general. El subcorpus español contiene 180 grabaciones transcritas, anotadas y alineadas (cada *utterance* con su correspondiente señal acústica). El corpus se describe en detalle en Cresti y Moneglia (eds.) (2005) y se pueden consultar tanto las transcripciones como las anotaciones y el

[1] El trabajar con lemas supone una abstracción con respecto a los recuentos de frecuencias de palabras. En este último caso, cada palabra es el *tipo* y las veces que aparece son los *ejemplos*.

audio, que acompañan al libro con un DVD. En Moreno y Urresti (2006) se da una exposición pormenorizada de los estudios que se han realizado sobre el subcorpus español.

El corpus se divide en tres grandes secciones: el registro informal (alrededor de 150.000 palabras), el registro formal (unas 80.000 palabras) y la de los medios de comunicación (70.000 palabras). La temática es muy variada y el número de hablantes diferentes supera los 500, siguiendo una distribución equitativa de hombres y mujeres, aunque no se tuvo en cuenta su distribución en los registros. Mayoritariamente, los hablantes son de la variedad centro-peninsular, aunque hay locutores de muy variadas procedencias.

Lo pertinente para este artículo es lo referente a la lematización. Efectivamente, cada uno de los subcorpus fue anotado morfosintácticamente y se crearon listas de frecuencias con las formas y los lemas de cada lengua. Nosotros tomaremos directamente la lista de los 100 lemas verbales más frecuentes en este corpus.

II.1.2. CHIEDE

CHIEDE (Garrote 2008) es un corpus de habla espontánea infantil. Toma como modelo la metodología empleada en C-ORAL-ROM, y lo aplica a la variedad infantil entre los 3 y los 5 años. Las grabaciones fueron realizadas en un colegio de Educación Infantil de Castilla-La Mancha. Está compuesto por dos tipos de grabaciones: las de asamblea, en las que todos los niños intervienen guiados por su profesora; y las de entrevista entre un niño y la investigadora. En total se recogen unas 60.000 palabras y varias horas de grabación, distribuidas proporcionalmente entre los tres años de la muestra.

El corpus ha sido anotado morfosintácticamente con la misma herramienta (GRAMPAL) y luego su resultado ha sido revisado y corregido manualmente por la investigadora. La lista de los 100 lemas verbales más frecuentes ha sido tomada de la tesis de M. Garrote.

II.1.3. Corpus de la Agencia EFE

Este corpus ha sido recogido y anotado por María Cristina Tovar como trabajo de investigación para la obtención del DEA en el programa de doctorado "El lenguaje humano: su origen, uso y aplicaciones", de la UAM. El corpus se va a utilizar en la tesis que ella está desarrollando en el LLI-UAM sobre las características del registro escrito periodístico en diferentes variedades geográficas del español.

Como se trata de una investigación en marcha, no disponemos de publicaciones pero en estos momentos ha pasado la etapa de revisión y recategorización de los textos en función de su tipología y se ha comenzado la primera fase de anotación morfosintáctica. El corpus está compuesto por más de 15 millones de palabras y nos parece impracticable su revisión manual completa, como se ha hecho con los corpus orales. Por tanto, procederemos a una revisión de una muestra aleatoria, aunque de momento para el experimento se han utilizados los resultados de la lematización automática. Por lo tanto, la lista de los 100 verbos más frecuentes no es más que una primera aproximación, aunque creemos que será bastante parecida a la definitiva.

II.2. Lematización automática

GRAMPAL es un analizador morfosintáctico del español que asigna la etiqueta más probable para cada palabra o unidad de palabras (*multiwords*). Esta etiqueta contiene información sobre la categoría sintáctica, su lema y rasgos morfosintácticos (persona, número, tiempo, aspecto y forma no personal en el caso de los verbos). GRAMPAL fue diseñado originalmente para analizar textos escritos y dar todos los análisis posibles para una forma dada. Así, por ejemplo, para la forma *bajo* debe proporcionar el análisis como preposición, verbo, adjetivo y nombre. Obviamente, hay muchas formas que no son ambiguas en el español, es decir, que sólo tienen un análisis morfosintáctico, pero también es cierto que formas muy frecuentes como *que, la, las* o *los* tienen dos análisis categoriales, al menos.

Para dar una idea de la ambigüedad morfosintáctica del español, en Moreno y Guirao (2006) damos una evaluación con corpus escritos y orales. La distribución entre palabras no ambiguas y ambiguas en el corpus escrito es de 65% a 35%, respectivamente. Sin embargo, la relación de ambigüedad está prácticamente al 50% en el corpus oral.

Como originariamente GRAMPAL no estaba diseñado para desambiguar, hubo que incorporar un módulo de desambiguación estadístico, basado en un corpus de entrenamiento formado por textos revisados a mano. Nuestra experiencia ha sido que en cada cambio de registro o variedad, se ha tenido que corregir entre un 5 y un 10 % los resultados, ya que la categorización morfosintáctica es sensible al tipo de texto. Otra innovación que hemos introducido ha sido el tratamiento de las unidades multipalabra, como *por ejemplo* o *en lugar de.* En el caso de los verbos, lo más relevante es que se ha incluido un módulo de reconocimiento de verbos que no están en el lexicón, de manera que si tiene forma analizada presenta una terminación propia de los verbos españoles, se le asigna provisionalmente la etiqueta de verbo. Para nuestro experimento, los casos de verbos que no estaban en el lexicón han sido eliminados del recuento, hasta que no se realice una verificación manual.

El grado de precisión de nuestro programa está en torno al 95%, que es la cifra típica de los etiquetadores avanzados, en español y en otras lenguas. Mejorar dicha precisión es difícil, dada la ambigüedad inherente en las lenguas, que hace complicada tomar una decisión incluso a lingüistas expertos.

II.4. El test de Dunning

Para identificar los lemas distintivos de cada subcorpus de nuestro experimento hemos empleado el test de razón de verosimilitud (*log-likelihood ratio test*) propuesto por Dunning (1993). Este método no asume distribuciones estadísticas normales de las unidades de un corpus. Por el contrario, la ratio de probabilidad (logarítmica) asume una distribución binomial más apropiada para palabras poco comunes pero significativas. Una ventaja adicional de este test es que no que necesita que los subcorpus estén equilibrados para llevar a cabo la comparación. Este método se ha aplicado con éxito para hallar colocaciones (Dunning 1993) y términos (Daille 1994). Para probar el método con la intención de encontrar unidades distintivas en dominios específicos, podemos trabajar con dos hipótesis:

i. Dos registros (o subcorpus) no muestran ninguna diferencia en unidades distintivas (*Hipótesis nula*).

ii. Para un subcorpus dado, podemos hallar unidades distintivas (*Hipótesis alternativa*).

Para comprobar cuál de las dos hipótesis es la correcta aplicamos el test a dos subcorpus bien definidos: lenguaje adulto e infantil. La manera de comprobarlo es ver la distribución de las unidades que han obtenido mayor puntuación en la razón de verosimilitud. Por ejemplo, las palabras más significativas de los adultos en el corpus CHIEDE fueron:

FORMAS	**ADULTOS (36.905)**	**NIÑOS (21.080)**	**TEST de DUNNING**
qué	1.123	108	510.29
te	743	59	373.43
a ver	371	23	207.58
bien	304	14	189.00
ah	270	18	146.32
claro	231	15	126.53
tú	264	27	113.88

Tabla 1: La palabras más características de los adultos

La fórmula estadística es:

$$-2 \log \lambda = 2 [\log L (p1, k 1, n1) + \log L (p2, k 2, n2) - \log L (p , k 1, n1) - \log L (p , k 2, n2)]$$

Las cifras de esta tabla de contingencias deben entenderse de la siguiente manera. Se forman dos conjuntos, el que se analiza para encontrar unidades distintivas y su conjunto complementario. En nuestro ejemplo, el conjunto principal es el formado por las palabras emitidas por los adultos y el conjunto complementario es el de los niños. Como se dijo anteriormente, este test no exige que el tamaño de los conjuntos sea equilibrado. El número de palabras emitidas por los adultos es de 36905 ($n1$), mientras que el de los niños es de 21080($n2$), para dar un total de 57985 en el corpus. Para cada palabra se proporciona las ocurrencias en adultos ($k1$), niños ($k2$) y el valor que proporciona el test de Dunning (resultado final de la fórmula). Cuanto mayor es el valor de la razón de verosimilitud, más característica es la palabra para el conjunto principal.

En nuestro ejemplo, el pronombre interrogativo *qué* aparece proporcionalmente muchas más veces (1123 entre 36905) en los adultos que en los niños (108 entre 21080). Eso le asigna una ratio de 510,29[2].

En la tabla podemos comprobar que una frecuencia de aparición mayor no necesariamente proporciona mayor razón de verosimilitud. La palabra *tú* aparece más veces (264) que la palabra *claro* (231) y sin embargo la segunda obtiene una razón

[2] La aplicación de la fórmula es como sigue: $n1$ y $n2$ son el número total de ejemplos de los conjuntos 1 y 2. $k1$ y $k2$ son el número de veces que aparece una determinada unidad (sea palabra, fonema, lema, categoría sintáctica, etc.). $p1$ es la probabilidad del primer conjunto y se calcula mediante $p1 = k1 / n1$. Análogamente, $p2 = k2 / n2$. La probabilidad del total, p, se calcula $p = (k1 + k2) / (n1 + n2)$. Finalmente, se aplica una razón de logaritmos, en el numerador está el caso específico: log L($p1$, $k1$, $n1$) + log L($p2$, $k2$, $n2$); y en el denominador se calcula la del total: log L(p, $k1$, $n1$) + log L(p, $k2$, $n2$). Como se puede apreciar, lo crucial es la razón entre los ejemplos concretos de la unidad ($k1$) en relación con el tamaño del conjunto ($n1$) y la misma relación en el conjunto complementario.

mayor (126,53) frente a la primera (113,88). Esto es debido a que *tú* aparece proporcionalmente más veces en el corpus complementario, el infantil, que *claro*. El test de la razón de verosimilitud favorece los casos que son más frecuentes (en comparación con el número total de ejemplos) en el conjunto principal que en el conjunto complementario.

En general, toda ratio que supera el valor de 8 es considerada como indicación de que la unidad es significativa para el conjunto en cuestión. Como se puede comprobar en la tabla, todas las palabras son relativas a la interacción del adulto con el niño, ya sea para preguntar (*qué, a ver*), como para asentir (*bien, ah, claro*) o para dirigirse a él o ella (*te, tú*).

Las 5 palabras que salieron con mayor puntuación en el conjunto infantil fueron:

mi	524.66
yo	300.54
sí	255.77
me	198.53
tengo	141.16

Todas ellas reflejan el uso característico de los pronombres y la primera persona, como habitualmente se describe en los estudios de lenguaje infantil. Garrote et al (2008) presentan más evidencias (entre ellas, fonemas y categorías) a favor de la fiabilidad de esta técnica estadística para encontrar unidades características de un conjunto frente a su complementario.

En este artículo aplicaremos la misma técnica para extraer los lemas verbales más significativos de los distintos registros que analizamos en el siguiente apartado.

III. COMPARACIÓN Y DISCUSIÓN DE LOS RESULTADOS

Vamos a considerar tres registros:

1. Habla espontánea adulta
2. Habla espontánea infantil
3. Texto escrito periodístico

Para realizar el cálculo de la razón de verosimilitud, enfrentaremos entre sí el habla espontánea adulta e infantil, y los textos periodísticos con el corpus de habla espontánea adulta.

Para cada registro se proporcionan dos listas, ordenadas por mayor frecuencia y mayor valor de razón de verosimilitud. En este apartado solo discutiremos los resultados más relevantes. Los datos completos se presentan en el Apéndice.

III. 1. Habla espontánea adulta

Los 10 verbos más frecuentes (sobre un total de 50.122 formas verbales) en C-ORAL-ROM se muestran en la tabla siguiente:

HABLA ADULTA			
Puesto	Verbo	Frecuencia Absoluta	Frecuencia Relativa
1	SER	7404	14.77%
2	DECIR	2652	5.29%

3	ESTAR	2404	4.79%
4	TENER	2388	4.76%
5	HACER	2220	4.42%
6	HABER	1456	2.90%
7	IR	1392	2.77%
8	VER	964	1.92%
9	DAR	886	1.76%
10	SABER	865	1.72%

Tabla 2: Los 10 verbos más frecuentes en C-ORAL-ROM

Los 10 verbos más significativos de C-ORAL-ROM, con su valor del test de Dunning calculado en oposición al conjunto de lemas verbales del corpus de la Agencia EFE son:

	HABLA ADULTA	
puesto	verbo	Dunning
1	SER	4.806,5
2	IR	3.052,8
3	CREER	2.693,4
4	ESTAR	2.465,4
5	DECIR	2.087,0
6	VER	1.691,0
7	SABER	1.690,3
8	VENIR	1.557,6
9	PASAR	1.084,1
10	LLAMAR	1.080,0

Tabla 3: Los 10 verbos más significativos en C-ORAL-ROM

De la comparación de los datos, se puede observar que un buen porcentaje de verbos coincide en ambas listas (SER, ESTAR, IR, DECIR, VER y SABER), lo que indicaría que frecuencia y distintividad en este caso irían bastante parejas. Destaca la presencia de verbos de movimiento (IR, VENIR) y los verbos de interacción conversacional como DECIR y LLAMAR. Ambos fenómenos se podrían asociar a las características propias de la oralidad, donde se describen eventos en una situación dialógica.

III. 2. Habla espontánea infantil

Los diez lemas verbales más frecuentes en CHIEDE son los que aparecen en la Tabla 4. Lo más llamativo es que coinciden con los del corpus de habla adulta salvo en el verbo JUGAR, que en los adultos es DAR. En cuanto al orden en la posición, en el léxico infantil TENER ocupa la segunda posición, en contraposición con DECIR, que es el segundo verbo en frecuencia de uso en los adultos (probablemente por su importancia como verbo *dicendi* en el registro oral).

HABLA INFANTIL			
Puesto	Verbo	Frecuencia Absoluta	Frecuencia Relativa
1	SER	509	12.5
2	TENER	330	8.1
3	ESTAR	193	4.7
4	SABER	176	4.3
5	HACER	172	4.2
6	IR	129	3.1
7	DECIR	118	2.8
8	HABER	93	2.2
9	VER	89	2.1
10	LLAMAR	88	2.1

Tabla 4: Los 10 verbos más frecuentes en CHIEDE

Los 10 verbos más significativos de CHIEDE, en contraposición con el corpus adulto de C-ORAL-ROM se muestran en la Tabla 5.

HABLA INFANTIL		
puesto	verbo	Dunning
1	JUGAR	200,9
2	SABER	102,9
3	CAER(SE)	97,8
4	TENER	76,2
5	PORTAR(SE)	71,7
6	REGALAR	64,3
7	PICAR	53,1
8	PINTAR	46,7
9	COMPRAR	41,5
10	CANTAR	40,2

Tabla 5: Los 10 verbos más significativos en CHIEDE

Al comparar las dos listas de léxico verbal de los niños, lo primero que llama la atención es que sólo coinciden tres verbos: JUGAR, SABER y TENER. Los otros siete verbos característicos se refieren o bien a actividades típicas de la infancia: CAER(SE), PORTAR(SE), PINTAR y CANTAR; o bien a actividades propias de los adultos en su interrelación con los niños: REGALAR y COMPRAR. El caso de PICAR es muy ilustrativo. Aparece sólo 18 veces (de un total de 4070 formas verbales empleadas por los niños). El uso más habitual de este verbo en CHIEDE es “me pica…” Este verbo ocupa la posición séptima, antes que verbos más frecuentes en el léxico infantil como PINTAR, porque PICAR aparece sólo 13 veces en el corpus C-ORAL-ROM (que tiene 50119 formas verbales).

Es bien conocido en la lingüística de corpus que los resultados son muy dependientes del tamaño del corpus y los corpus empleados en nuestro estudio no tienen un número suficiente (especialmente el infantil) de palabras para extraer conclusiones.

Sin embargo, los datos ofrecidos por el test de Dunning son compatibles con la bibliografía en psicolingüística infantil y coherentes con nuestra experiencia.

Hemos extraído la lista de verbos significativos para los adultos de C-ORAL-ROM, en situación complementaria con los verbos empleados por los niños, y entre los primeros 20 verbos distintivos encontramos 11 que no aparecen ninguna vez en CHIEDE: UNIR, RECORDAR, SUPONER, SOBRAR, TRATAR, EXPLICAR, CONSIDERAR, MANTENER, PERMITIR, CONSEGUIR y REALIZAR. Esto es un indicio de que estos verbos, aunque muy generales y habituales en la actuación lingüística adulta, no forman parte del léxico activo de los niños de entre 2 y 5 años. Habría que confirmar esta conjetura con estudios experimentales psicolingüísticos. Esta forma de extraer diferencias léxicas entre adultos y niños puede ser no sólo de inspiración para nuevos estudios experimentales sino que también podría ser empleada para diseñar estrategias pedagógicas de enseñanza del léxico.

III.3. Registro periodístico

Las Tablas 6 y 7 muestran los primeros lemas verbales en este registro. En primer lugar, destaca la presencia de los verbos *dicendi* propios de un registro informativo: DECIR, SEÑALAR, ASEGURAR, INFORMAR. Sin embargo, mientras que en la frecuencia de uso nos seguimos encontrando con los verbos generales (SER, TENER, HACER, ESTAR, HABER), en la razón de verosimilitud todos los ejemplos son de verbos de comunicación o declarativos. Esto nos confirma la utilidad del test de Dunning para extraer elementos característicos en registros especializados.

TEXTOS PERIODÍSTICOS			
Puesto	verbo	Frecuencia Absoluta	Frecuencia Relativa
1	SER	98694	5,98%
2	TENER	44153	2,67%
3	HACER	33509	2,03%
4	DECIR	30579	1,85%
5	HABER	23843	1,44%
6	ESTAR	23038	1,40%
7	SEÑALAR	14098	0,85%
8	DAR	13927	0,84%
9	ASEGURAR	12992	0,79%
10	INFORMAR	12275	0,74%

Tabla 6: Los 10 verbos más frecuentes en el corpus EFE

	TEXTOS PERIODÍSTICOS	
puesto	verbo	Dunning
1	SEÑALAR	691,2
2	ASEGURAR	610,0
3	AFIRMAR	594,2
4	INFORMAR	592,6
5	DESTACAR	450,9
6	INDICAR	438,8
7	PRESENTAR	400,6
8	AGREGAR	332,6
9	CONSIDERAR	317,9
10	CELEBRAR	300,4

Tabla 7: Los 10 verbos más significativos en el corpus EFE

III.4. Conclusiones y trabajo futuro

Como reflexión final, podemos sacar algunas conclusiones a partir de los datos. En primer lugar, hay que destacar que los conceptos de frecuencia de uso y distintividad son coincidentes en cierta medida en el habla espontánea adulta, como una prueba más del carácter básico de la oralidad en las lenguas humanas.

En segundo lugar, los datos nos sugieren que los verbos más frecuentes en el habla espontánea son los mismos en adultos y niños, con cierta variación en el orden. Sin embargo, en cuanto a distintividad, la mayoría de los verbos significativos están relacionados con las actividades propias de unos y otros.

Por otra parte, al analizar un registro especializado, como es el periodístico, comprobamos que los verbos característicos no coinciden con los más frecuentes, que suelen ser los generales de la lengua.

Finalmente, si comparamos la frecuencia relativa en el uso de verbos (ver Apéndice 1) se observa una significativa desproporción en el uso del verbo SER en el habla espontánea (tanto adulta como infantil), donde la tasa está en el 12-14% frente al 6 % en el registro periodístico. Esta relación también se produce con otros verbos muy frecuentes. Dicho de otra manera, la diversidad de lemas es mucho mayor en el corpus escrito que en el oral, situación que es conocida y esperada, y que nuestro recuento ha cuantificado.

Retomando las palabras iniciales de Bybee en este artículo, las propiedades de los lexemas verbales más frecuentes suelen ser muy relevantes para la oralidad. Los datos de nuestro análisis apoyan la hipótesis funcionalista de la importancia de la repetición en la conformación de estructuras lingüísticas básicas. Sin embargo, los dominios y registros especializados muestran la relevancia de las unidades distintivas, que no son muy frecuentes pero son muy informativas. En este caso, como afirmaba Dunning: "Unfortunately rare events *do* make up a large fraction of real text."

Las aplicaciones de la frecuencia de uso y de la distintividad en un registro dado son muy sugerentes para la lexicografía, terminología y didáctica de lenguas, ya que permiten diferenciar lo general y frecuente de lo particular y característico. En cualquier caso, nos parece que este tipo de listados como el que ofrecemos en los Apéndices sirven de base para el conocimiento general sobre las lenguas.

AGRADECIMIENTOS

Esta investigación ha sido parcialmente financiada por el proyecto BRAVO-RL del MEC-CICYT (TIN2007-67407-C03-02) y por la Comunidad de Madrid en el marco del convenio MAVIR (S-0505/TIC/0267).

REFERENCIAS BIBLIOGRÁFICAS

Bybee, Joan (2007): *Frequency of use and the organization of language*. Oxford, Oxford University Press.

Cresti y Moneglia (eds.) (2005) *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam, John Benjamins.

Dunning, (1993): Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19 (1): 61-74.

Garrote, M. (2008): *CHIEDE: corpus de habla infantil espontánea del español*. Tesis doctoral. Universidad Autónoma de Madrid.

Garrote, M, Guirao, J.M. y Moreno, A. (2008): Extracción de unidades distintivas en adultos y niños de un corpus de lengua oral espontánea. En *Actas del 8º Congreso de Lingüística General*. Madrid, Universidad Autónoma de Madrid.

Moreno, A. (1991): *Un modelo basado en la unificación para el análisis y generación de la morfología en español*. Tesis doctoral. Universidad Autónoma de Madrid

Moreno y Guirao (2006): Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation. In Spoken Language Corpus and Linguistic Informatics. Amsterdam, John Benjamins.

Moreno y Urresti (2006): El proyecto C-ORAL-ROM y su aplicación a la enseñanza de español. *Oralia*, 8.

APÉNDICE 1: Los 100 verbos más frecuentes en los tres corpus

HABLA ADULTA				HABLA INFANTIL				TEXTOS PERIODÍSTICOS			
Puesto	Verbo	Frecuencia Absoluta	Frecuencia Relativa	Puesto	Verbo	Frecuencia Absoluta	Frecuencia Relativa	Puesto	verbo	Frecuencia Absoluta	Frecuencia Relativa
1	SER	7404	14.77%	1	SER	509	12.5	1	SER	98694	5,98%
2	DECIR	2652	5.29%	2	TENER	330	8.1	2	TENER	44153	2,67%
3	ESTAR	2404	4.79%	3	ESTAR	193	4.7	3	HACER	33509	2,03%
4	TENER	2388	4.76%	4	SABER	176	4.3	4	DECIR	30579	1,85%
5	HACER	2220	4.42%	5	HACER	172	4.2	5	HABER	23843	1,44%
6	HABER	1456	2.90%	6	IR	129	3.1	6	ESTAR	23038	1,40%
7	IR	1392	2.77%	7	DECIR	118	2.8	7	SEÑALAR	14098	0,85%
8	VER	964	1.92%	8	HABER	93	2.2	8	DAR	13927	0,84%
9	DAR	886	1.76%	9	VER	89	2.1	9	ASEGURAR	12992	0,79%
10	SABER	865	1.72%	10	LLAMAR	88	2.1	10	INFORMAR	12275	0,74%
11	PASAR	731	1.45%	11	PONER	86	2.1	11	PRESENTAR	11993	0,73%
12	PONER	645	1.28%	12	JUGAR	77	1.8	12	CONSIDERAR	11742	0,71%
13	CREER	624	1.24%	13	PASAR	65	1.5	13	EXPLICAR	11400	0,69%
14	VENIR	591	1.17%	14	DAR	56	1.3	14	AFIRMAR	11160	0,68%
15	LLAMAR	557	1.11%	15	VENIR	53	1.3	15	LLEGAR	10341	0,63%
16	LLEVAR	526	1.04%	16	COMER	49	1.2	16	MANTENER	10002	0,61%
17	HABLAR	511	1.01%	17	CAER	48	1.1	17	INDICAR	9859	0,60%
18	QUEDAR	489	0.97%	18	COMPRAR	42	1	18	ENCONTRAR	9532	0,58%
19	QUERER	459	0.91%	19	CONTAR	42	1	19	REALIZAR	9530	0,58%
20	LLEGAR	388	0.77%	20	LLEVAR	39	0.9	20	DESTACAR	9167	0,56%
21	DEJAR	318	0.63%	21	METER	37	0.9	21	PEDIR	8574	0,52%
22	SALIR	308	0.61%	22	GUSTAR	37	0.9	22	LLEVAR	8068	0,49%
23	PARECER	302	0.60%	23	COGER	36	0.8	23	RECIBIR	8008	0,49%
24	GUSTAR	294	0.58%	24	TRABAJAR	32	0.7	24	CELEBRAR	7933	0,48%
25	PENSAR	270	0.53%	25	SALIR	31	0.7	25	PRODUCIR	7573	0,46%
26	COMER	251	0.50%	26	MIRAR	30	0.7	26	JUGAR	7087	0,43%
27	TRABAJAR	250	0.49%	27	ECHAR	30	0.7	27	ANUNCIAR	7015	0,42%
28	CONTAR	234	0.46%	28	TIRAR	29	0.7	28	RECORDAR	6881	0,42%
29	COGER	221	0.44%	29	DEJAR	26	0.6	29	PERMITIR	6865	0,42%
30	UNIR	200	0.39%	30	REGALAR	26	0.6	30	CONTAR	6856	0,42%
31	VALER	198	0.39%	31	SUBIR	25	0.6	31	CONOCER	6627	0,40%
32	ENCONTRAR	194	0.38%	32	QUEDAR	25	0.6	32	AÑADIR	6617	0,40%
33	METER	181	0.36%	33	QUERER	25	0.6	33	PONER	6607	0,40%
34	EMPEZAR	172	0.34%	34	PINTAR	24	0.5	34	ESPERAR	6589	0,40%
35	CONOCER	171	0.34%	35	LEER	23	0.5	35	PARTICIPAR	6398	0,39%
36	PODER	169	0.33%	36	TRAER	22	0.5	36	CONSEGUIR	6373	0,39%
37	MIRAR	169	0.33%	37	TOCAR	22	0.5	37	LOGRAR	6261	0,38%
38	PEDIR	163	0.32%	38	PORTAR	22	0.5	38	TRATAR	6143	0,37%
39	ENTENDER	160	0.31%	39	CORRER	21	0.5	39	GANAR	6079	0,37%
40	VIVIR	156	0.31%	40	PARECER	21	0.5	40	MOSTRAR	6019	0,36%
41	ENTRAR	156	0.31%	41	DORMIR	21	0.5	41	IR	5987	0,36%
42	SEGUIR	155	0.30%	42	HABLAR	20	0.4	42	DECLARAR	5946	0,36%
43	BUSCAR	151	0.30%	43	CUMPLIR	19	0.4	43	MANIFESTAR	5864	0,36%
44	SACAR	150	0.29%	44	VIVIR	18	0.4	44	PREVER	5811	0,35%
45	VOLVER	147	0.29%	45	ENSEÑAR	18	0.4	45	AGREGAR	5552	0,34%
46	COMPRAR	144	0.28%	46	PICAR	18	0.4	46	ABRIR	5518	0,33%

47	PAGAR	144	0.28%	47	QUITAR	18	0.4	47	SEGUIR	5515	0,33%
48	PREGUNTAR	144	0.28%	48	LEVANTAR	17	0.4	48	VER	5502	0,33%
49	TOMAR	142	0.28%	49	PILLAR	17	0.4	49	DEJAR	5435	0,33%
50	CAMBIAR	137	0.27%	50	ACORDAR	17	0.4	50	TRABAJAR	5426	0,33%
51	SUBIR	133	0.26%	51	SENTAR	16	0.3	51	PARTIR	5341	0,32%
52	PERDER	125	0.24%	52	ROMPER	16	0.3	52	QUEDAR	5164	0,31%
53	ESPERAR	125	0.24%	53	PERDER	15	0.3	53	PASAR	4997	0,30%
54	ECHAR	124	0.24%	54	CREER	15	0.3	54	INCLUIR	4926	0,30%
55	ACORDAR	123	0.24%	55	SACAR	15	0.3	55	TOMAR	4911	0,30%
56	GANAR	122	0.24%	56	LLEGAR	15	0.3	56	OFRECER	4890	0,30%
57	TRAER	121	0.24%	57	MATAR	13	0.3	57	DECIDIR	4886	0,30%
58	ABRIR	117	0.23%	58	PEGAR	13	0.3	58	SUPONER	4837	0,29%
59	MANDAR	112	0.22%	59	ACOSTAR	13	0.3	59	RECONOCER	4814	0,29%
60	RECORDAR	111	0.22%	60	ESCRIBIR	13	0.3	60	FORMAR	4811	0,29%
61	SUPONER	109	0.21%	61	CANTAR	12	0.2	61	DIRIGIR	4670	0,28%
62	QUITAR	108	0.21%	62	EMPEZAR	12	0.2	62	APROBAR	4643	0,28%
63	SOBRAR	101	0.20%	63	MORIR	12	0.2	63	EXISTIR	4640	0,28%
64	ACABAR	100	0.19%	64	LAVAR	12	0.2	64	AFECTAR	4570	0,28%
65	LEER	100	0.19%	65	LLORAR	12	0.2	65	ACUSAR	4550	0,28%
66	IMAGINAR	99	0.19%	66	BAJAR	11	0.2	66	REUNIR	4539	0,27%
67	TRATAR	94	0.18%	67	PARAR	11	0.2	67	INICIAR	4446	0,27%
68	ESTUDIAR	93	0.18%	68	OLVIDAR	11	0.2	68	CUMPLIR	4371	0,26%
69	INTENTAR	92	0.18%	69	MONTAR	11	0.2	69	OBTENER	4351	0,26%
70	OCURRIR	91	0.18%	70	GUARDAR	10	0.2	70	REGISTRAR	4337	0,26%
71	ESCUCHAR	91	0.18%	71	PODER	10	0.2	71	SABER	4324	0,26%
72	SENTAR	87	0.17%	72	CORTAR	10	0.2	72	SUFRIR	4323	0,26%
73	TOCAR	84	0.16%	73	PLANCHAR	10	0.2	73	HABLAR	4272	0,26%
74	CASAR	82	0.16%	74	ENCONTRAR	10	0.2	74	ENTRAR	4269	0,26%
75	EXPLICAR	82	0.16%	75	ACABAR	10	0.2	75	PERDER	4267	0,26%
76	UTILIZAR	81	0.16%	76	CERRAR	10	0.2	76	COMENZAR	4250	0,26%
77	TIRAR	77	0.15%	77	ABRIR	10	0.2	77	FIRMAR	4159	0,25%
78	CONSIDERAR	77	0.15%	78	BEBER	9	0.2	78	SUBRAYAR	4153	0,25%
79	JUGAR	75	0.14%	79	MORDER	9	0.2	79	ESTABLECER	4143	0,25%
80	DORMIR	75	0.14%	80	FALTAR	9	0.2	80	DISPUTAR	4141	0,25%
81	MANTENER	74	0.14%	81	SALTAR	9	0.2	81	CONFIRMAR	4096	0,25%
82	LEVANTAR	74	0.14%	82	GANAR	9	0.2	82	PRECISAR	4029	0,24%
83	MORIR	74	0.14%	83	PINCHAR	8	0.1	83	ALCANZAR	3991	0,24%
84	TERMINAR	74	0.14%	84	OÍR	8	0.1	84	EXPRESAR	3984	0,24%
85	CAER	73	0.14%	85	VESTIR	8	0.1	85	DETENER	3919	0,24%
86	PERMITIR	70	0.13%	86	ESCUCHAR	8	0.1	86	SALIR	3888	0,24%
87	MOVER	70	0.13%	87	TOMAR	8	0.1	87	BUSCAR	3869	0,23%
88	NECESITAR	70	0.13%	88	ANDAR	8	0.1	88	SITUAR	3843	0,23%
89	SALAR	69	0.13%	89	VOLAR	7	0.1	89	INTENTAR	3820	0,23%
90	CONSEGUIR	69	0.13%	90	SEGUIR	7	0.1	90	CERRAR	3704	0,22%
91	FIJAR	69	0.13%	91	CONOCER	7	0.1	91	VIVIR	3678	0,22%
92	SERVIR	69	0.13%	92	SOÑAR	7	0.1	92	CREAR	3658	0,22%
93	APARECER	69	0.13%	93	CONVERTIR	7	0.1	93	CONVERTIR	3627	0,22%
94	BAJAR	68	0.13%	94	ESCONDER	7	0.1	94	UNIR	3619	0,22%
95	REALIZAR	68	0.13%	95	BUSCAR	7	0.1	95	OCURRIR	3567	0,22%
96	COSTAR	67	0.13%	96	ENCENDER	7	0.1	96	CONCLUIR	3529	0,21%
97	REFERIR	66	0.13%	97	TERMINAR	7	0.1	97	DENUNCIAR	3472	0,21%

98	INTERESAR	66	0.13%	98	PISAR	7	0.1	98	UTILIZAR	3429	0,21%
99	APRENDER	66	0.13%	99	BAÑAR	6	0.1	99	INSISTIR	3410	0,21%
100	ANDAR	65	0.12%	100	ESPERAR	6	0.1	100	AYUDAR	3404	0,21%

APÉNDICE 2: Los 100 verbos más significativos en los tres corpus

HABLA ADULTA			HABLA INFANTIL			TEXTOS PERIODÍSTICOS		
puesto	verbo	Dunning	puesto	verbo	Dunning	puesto	verbo	Dunning
1	SER	4.806,5	1	JUGAR	200,9	1	SEÑALAR	691,2
2	IR	3.052,8	2	SABER	102,9	2	ASEGURAR	610,0
3	CREER	2.693,4	3	CAER	97,8	3	AFIRMAR	594,2
4	ESTAR	2.465,4	4	TENER	76,2	4	INFORMAR	592,6
5	DECIR	2.087,0	5	PORTAR	71,7	5	DESTACAR	450,9
6	VER	1.691,0	6	REGALAR	64,3	6	INDICAR	438,8
7	SABER	1.690,3	7	PICAR	53,1	7	PRESENTAR	400,6
8	VENIR	1.557,6	8	PINTAR	46,7	8	AGREGAR	332,6
9	PASAR	1.084,1	9	COMPRAR	41,5	9	CONSIDERAR	317,9
10	LLAMAR	1.080,0	10	CANTAR	40,2	10	CELEBRAR	300,4
11	HACER	1.046,9	11	CORRER	39,4	11	EXPLICAR	287,4
12	COGER	821,2	12	TIRAR	37,9	12	ANUNCIAR	286,8
13	QUERER	706,9	13	ROMPER	33,7	13	DECLARAR	284,8
14	TENER	661,0	14	MORDER	33,6	14	MANIFESTAR	274,3
15	HABLAR	609,5	15	LLAMAR	29,1	15	LOGRAR	269,8
16	GUSTAR	603,7	16	PINCHAR	29,0	16	PREVER	260,3
17	PONER	593,4	17	CUMPLIR	28,7	17	RECIBIR	250,6
18	VALER	573,2	18	PLANCHAR	28,3	18	AÑADIR	247,7
19	HABER	559,9	19	COMER	26,0	19	MANTENER	246,2
20	METER	444,2	20	RAPAR	25,9	20	PARTICIPAR	244,7
21	QUEDAR	429,9	21	LAVAR	25,5	21	REALIZAR	241,6
22	SOBRAR	415,8	22	LLORAR	23,7	22	PRECISAR	241,2
23	DAR	374,0	23	ECHAR	22,9	23	DISPUTAR	227,5
24	MIRAR	368,6	24	SOÑAR	22,5	24	MOSTRAR	220,5
25	IMAGINAR	333,3	25	CAZAR	22,4	25	ALCANZAR	203,9
26	MANDAR	308,0	26	METER	21,4	26	ACUSAR	190,5
27	ECHAR	282,3	27	PILLAR	21,3	27	INICIAR	184,8
28	PARECER	268,5	28	COLUMPIAR	20,7	28	SUBRAYAR	183,7
29	EMPEZAR	257,3	29	ACOSTAR	20,5	29	REUNIR	176,1
30	PENSAR	247,7	30	DORMIR	19,6	30	DETENER	170,7
31	LLEVAR	234,1	31	TOCAR	18,8	31	PRODUCIR	167,6
32	LECHAR	232,8	32	CONTAR	18,7	32	CALIFICAR	166,2
33	LIAR	226,5	33	ENSEÑAR	16,9	33	ASISTIR	159,1
34	PILLAR	215,2	34	ESCONDER	16,8	34	REGISTRAR	157,0
35	QUITAR	213,2	35	PONER	16,7	35	CONCLUIR	154,2
36	TRAER	205,6	36	LEER	16,2	36	CONDENAR	154,1
37	SALIR	202,7	37	BORRAR	15,8	37	OBTENER	149,9
38	COMER	197,7	38	ENCENDER	15,7	38	SOLICITAR	141,3
39	ANDAR	193,7	39	PISAR	15,7	39	CONFIRMAR	140,6
40	DORMIR	175,8	40	GATEAR	15,5	40	COMENZAR	137,3
41	COSER	173,6	41	BOTAR	15,5	41	PARTIR	135,5
42	JODER	172,9	42	GUIÑAR	15,5	42	OFRECER	135,3
43	ENSEÑAR	137,4	43	BAÑAR	15,4	43	INCLUIR	133,9

44	REÍR	128,8	44	PEGAR	14,9	44	AFECTAR	129,5
45	ACOSTAR	128,6	45	PARAR	14,4	45	EXPRESAR	127,6
46	SENTAR	128,1	46	VOLAR	13,9	46	SUFRIR	127,2
47	ENROLLAR	123,5	47	HUNDIR	13,4	47	DIRIGIR	125,0
48	SACAR	122,7	48	SALTAR	13,2	48	PERMITIR	121,7
49	TIRAR	121,9	49	MIRAR	13,0	49	ADVERTIR	120,5
50	MOLAR	119,1	50	DESPERTAR	12,8	50	JUGAR	119,9
51	ENTERAR	118,7	51	COGER	12,8	51	CONVOCAR	118,1
52	CAGAR	116,6	52	SUBIR	12,3	52	APROBAR	117,8
53	CENAR	111,4	53	LEVANTAR	12,0	53	REITERAR	113,0
54	DEJAR	107,6	54	MATAR	11,9	54	DESTINAR	110,1
55	LEER	103,0	55	NADAR	11,7	55	RECHAZAR	109,0
56	CASAR	101,8	56	TIRITAR	11,2	56	FORMAR	109,0
57	ENTENDER	101,1	57	CERRAR	11,1	57	GARANTIZAR	108,1
58	APRENDER	100,2	58	MONTAR	11,1	58	EVITAR	106,5
59	CURRAR	95,3	59	OLVIDAR	11,1	59	CONTINUAR	106,1
60	TOCAR	85,6	60	VESTIR	11,1	60	PRESIDIR	105,7
61	APETECER	85,3	61	FALTAR	10,8	61	DEMOSTRAR	105,3
62	OÍR	78,0	62	RASCAR	10,3	62	CONSEGUIR	104,6
63	ENCANTAR	75,8	63	COLOREAR	10,3	63	DENUNCIAR	102,4
64	FLIPAR	73,2	64	RULAR	10,3	64	NEGOCIAR	100,8
65	SALAR	72,9	65	BEBER	10,3	65	TRASLADAR	100,1
66	ESCUCHAR	70,3	66	TRAER	10,1	66	CAUSAR	99,3
67	COMPRAR	67,8	67	GUARDAR	9,6	67	RECONOCER	99,1
68	LEVANTAR	67,6	68	QUEMAR	9,4	68	SITUAR	95,6
69	PERDONAR	64,7	69	MERENDAR	9,1	69	RETIRAR	95,3
70	PREGUNTAR	64,0	70	DISFRAZAR	8,2	70	FIGURAR	94,3
71	SUBIR	64,0	71	PEINAR	8,2	71	REPRESENTAR	93,5
72	ENAMORAR	62,5	72	SENTAR	7,5	72	ACUDIR	92,0
73	MOVER	59,8	73	ESCRIBIR	7,2	73	AUMENTAR	91,7
74	VOLVER	57,4	74	QUITAR	6,7	74	EMITIR	91,3
75	UNIR	57,4	75	BARRER	6,7	75	INTEGRAR	91,2
76	SOLAPAR	55,1	76	CHILLAR	6,7	76	INAUGURAR	91,2
77	AGOBIAR	54,3	77	SALUDAR	6,7	77	FALLECER	90,0
78	DESAYUNAR	53,6	78	EMPUJAR	6,6	78	ABANDONAR	87,0
79	DOLER	51,2	79	INVENTAR	5,7	79	CONCEDER	86,5
80	AMAR	50,8	80	CASTIGAR	5,7	80	SOSTENER	86,0
81	PINTAR	50,1	81	GUSTAR	5,7	81	PRETENDER	84,1
82	ABURRIR	49,7	82	TRABAJAR	5,3	82	PUBLICAR	83,8
83	EXPERIENCIAR	49,3	83	SUJETAR	5,1	83	CRITICAR	83,7
84	CUIDAR	48,4	84	ESCAPAR	5,0	84	MARCAR	83,6
85	PODER	46,9	85	CHOCAR	5,0	85	PRENSAR	82,0
86	PEGAR	44,4	86	CORTAR	4,7	86	DECIDIR	82,0
87	ADELGAZAR	43,6	87	MANCHAR	4,4	87	CERRAR	81,4
88	MOSQUEAR	43,4	88	MORIR	4,2	88	PROVOCAR	81,0
89	CABREAR	43,0	89	CANSAR	3,8	89	VIAJAR	80,9
90	COSTAR	42,7	90	BAJAR	3,8	90	CUMPLIR	80,8
91	DIBUJAR	41,9	91	ACORDAR	3,7	91	EFECTUAR	79,4
92	PLANCHAR	39,4	92	FREGAR	3,3	92	REGRESAR	79,2
93	REGAR	39,3	93	ENTERRAR	3,3	93	IMPULSAR	78,5
94	CAMBIAR	39,2	94	ESTIRAR	3,3	94	VOTAR	77,4

95	ACABAR	37,9	95	ALCANZAR	3,3	95	AFRONTAR	76,9
96	MAMAR	37,8	96	DESAYUNAR	3,2	96	FIRMAR	76,7
97	TRABAJAR	36,9	97	GRITAR	2,8	97	ADMITIR	75,1
98	OLER	36,1	98	CHUPAR	2,8	98	RECUPERAR	74,4
99	ENTRABAR	35,2	99	FELICITAR	2,8	99	DEFENDER	74,4
100	CLOCAR	35,2	100	DISPARAR	2,8	100	PROPONER	74,2