Sobre las frecuencias verbales en español

Guillermo Rojo
Universidade de Santiago de Compostela

en Sedano, Mercedes, Adriana Bolívar y Martha Shiro (comps.): *Haciendo Lingüística. Homenaje a Paola Bentivoglio*, Universidad Central de Venezuela, 2006, 309-324.

En los últimos años se ha ido poniendo de relieve, cada vez con mayor intensidad, la importancia de los aspectos estadísticos para la comprensión de la estructura y funcionamiento de las lenguas. Cuando menos, para buscar una formulación más prudente, se ha destacado el peso que algunos factores relacionados con la frecuencia de elementos y sus combinaciones tienen en la conformación de algunos aspectos estructurales. Sin negar, por supuesto, la importancia que este enfoque tiene en el estudio de los componentes fónico y léxico, la nueva perspectiva resulta de especial interés cuando se aplica a la gramática, probablemente por el hecho de que es el terreno en que la estadística se ha practicado con menos frecuencia y su utilidad resulta menos evidente. Afortunadamente, queda ya muy lejos la afirmación de Chomsky según la cual los aspectos estadísticos son totalmente irrelevantes para la teoría gramatical y también para el conocimiento de la gramática de una lengua particular.[1] Por el contrario, nociones como gramaticalización y emergencia parecen estar fuertemente relacionadas con la frecuencia con que los elementos y sus diferentes combinaciones se manifiestan en los textos lingüísticos reales (cf. Bybee y Hopper 2001).

[1] En efecto, en un texto fuertemente crítico hacia los hábitos científicos de los distribucionalistas, Chomsky juzgaba totalmente inadecuado considerar que el objetivo del trabajo realizado por el lingüista pudiera consistir en dar cuenta de las secuencias que integran un corpus. El objetivo debe ser la lengua, esto es, las secuencias posibles en la lengua, las reglas que hacen posibles esas secuencias, la gramática, el conocimiento lingüístico que existe en el hablante-oyente ideal. Y la forma adecuada de acceder a ese conocimiento es, por supuesto, la introspección. Además, al negar la validez del modelo de estados finitos como forma de construir la gramática, incluso con el refinamiento de introducir en él las probabilidades de los diferentes elementos en cada estado, rechaza Chomsky la utilidad de los aspectos estadísticos que, según dice, "have nothing to do with grammar, e.g. surely is not a matter of concern for the grammar of English that 'New York' is more probable than 'Nevada' in the context 'I come from __.' In general, the importance of probabilistic considerations seems to me to have been highly overrated in recent discussions of linguistic theory" (Chomsky 1962:215, nota).

Es de todo punto evidente la necesidad de diferenciar entre los aspectos estadísticos de un inventario de elementos (fonemas o palabras, por ejemplo) y la frecuencia que esos mismos elementos presentan en los textos reales. Los diccionarios habituales contienen el léxico de una lengua o de alguna de sus variedades de un modo en el que la frecuencia de todas las palabras que forman la macroestructura es igual a 1. Por supuesto, hay diccionarios que incorporan ciertas indicaciones acerca de la mayor o menor tasa de aparición de los elementos que contienen, pero no se trata realmente en esos casos de hacer un recuento exacto –lo cual sería un tanto absurdo–, sino de adscribir las palabras a uno de los tres o cuatro grandes bloques que se establecen habitualmente. Tampoco va en contra del principio general el hecho de que muchos diccionarios contengan o no las palabras en función de la frecuencia –real o supuesta– que presentan en los textos. Una vez se ha decidido la incorporación de un elemento al inventario, su frecuencia queda igualada con la de todas las seleccionadas y, como consecuencia de ello, todas tienen el mismo peso. Algo parecido podría decirse, *mutatis mutandis*, de los inventarios de elementos fónicos, morfológicos, etc.

Los textos reales muestran, en cambio, una estructura estadística sistemáticamente descompensada. Para decirlo rápidamente, un texto, cualquier texto, está formado, en todos sus niveles, por unos pocos elementos que muestran frecuencias altas o muy altas y muchos elementos que aparecen poco o muy poco. La existencia de un grupo, reducido, de formas de frecuencia muy elevada explica, por ejemplo, el hecho de que la suma de las apariciones de las diez formas más frecuentes de un texto o un conjunto de textos suponga normalmente un porcentaje próximo al 30% del total o que, para ir a un terreno más concreto, con solo los 32 verbos más frecuentes del español se construya algo más del 50% de todas las cláusulas (cf. Rojo 2001). En el otro extremo, el enorme número de elementos que presentan frecuencias muy bajas supone la necesidad de construir corpus de gran tamaño para conseguir documentar adecuadamente aquellos que aparecen, por término medio, una vez cada diez o veinte millones de palabras.

La comparación de estos dos bloques produce un fuerte contraste al poner en paralelo el inventario y el uso: el grupo de formas muy frecuentes supone un escasísimo porcentaje con respecto al total del inventario, pero da cuenta de una buena parte de lo que se encuentra en los textos; el grupo de formas de baja frecuencia, en cambio, constituye la mayoría del inventario, pero el conjunto de sus utilizaciones en los textos supone un porcentaje muy reducido.

En los párrafos anteriores, he utilizado el concepto de inventario sin aludir al modo en que se ha originado, lo cual produce, por lo menos, dos

tipos diferentes. El inventario que encontramos en un diccionario general –no, claro, en los diccionarios basados exclusivamente en corpus– deriva de la actuación conjunta de una serie compleja de factores entre los que no suele figurar la mayor o menor frecuencia de uso de un elemento. El inventario que procede del análisis de un corpus presenta, en cambio, únicamente los elementos presentes en ese corpus. Son bien conocidas las diferencias que este doble origen produce en la comparación de lo obtenido en cada caso: hay muchas palabras contenidas en los diccionarios que no se documentan en los corpus y muchas más palabras presentes en los textos que no figuran en los diccionarios.

Como he insinuado en algún otro lugar, el español es una lengua no bien conocida en los aspectos estadísticos y mucho menos, como es lógico, por las dificultades que presentan las tareas previas, en lo referente a fenómenos gramaticales. El *Frequency Dictionary of Spanish Words* (Juilland y Chang 1964 = FDSW) es una obra valiosísima en la cual está contenida información que permite ir mucho más allá de los aspectos puramente léxicos y entrar en los gramaticales. En tanto que está –no podría ser de otro modo– basada en el análisis de un corpus (de aproximadamente medio millón de formas), la obra proporciona un inventario del segundo tipo de los mencionados más arriba: el derivado de un corpus y precisamente por ello ha sido utilizada para estudiar, por ejemplo, la estructura del léxico español o la frecuencia de las formas verbales. Sin embargo, no se ha tenido suficientemente en cuenta el hecho de que el FDSW no presenta en realidad el inventario de las formas y lemas contenidos en el corpus estudiado, sino el subconjunto de los 5024 lemas más "frecuentes" según el conjunto de factores utilizado por los autores (frecuencia, dispersión y uso) de un total de unos 20 000 obtenidos del corpus.[2] En otras palabras, contiene aproximadamente las formas vinculadas al 25% más frecuente de los lemas, lo cual produce una situación peculiar, que hace interesante la comparación de los resultados obtenidos en este elenco con el que resulta de los materiales extraídos de otros conjuntos. En este trabajo me propongo llevar a cabo esa labor en lo referente a la distribución de elementos y usos de las tres conjugaciones verbales del español.

---

[2] El corpus utilizado contenía algo menos de 50 000 formas, que fueron reducidas a unos 25 000 lemas. La eliminación de extranjerismos y nombres propios dejó alrededor de 20 000 lemas. Posteriormente, la decisión de cortar en los de frecuencia inferior a 4 redujo el inventario a 14 000, que pasó a 9000 al prescindir de todos los que no estaban presentes en, al menos, tres de los "mundos" establecidos. Por fin, se llegó a los 5024 lemas considerados al establecer el límite inferior de uso en 3,08. *Vid.* detalles en Juilland y Chang (1964:LXXIV-LXXVI).

Preguntarse sin más por la frecuencia de las tres conjugaciones oculta las diferencias entre una posible consulta acerca del número de verbos de cada una de ellas y otra en torno al grado de utilización en la lengua de formas pertenecientes a distintos verbos, cada uno de ellos vinculado a un cierto modelo general de conjugación. En el primer caso, se está tratando como un único caso el conjunto de formas que integran el paradigma de cada verbo, y la pregunta se refiere al número mayor o menor de verbos que pertenecen a cada uno de los diferentes modelos conjugacionales. En la segunda interpretación, se está haciendo referencia a la mayor o menor frecuencia de todas y cada una de las formas pertenecientes al paradigma de cada verbo y se pretende que la respuesta trate en un bloque al conjunto de todas las formas registradas de todos los verbos que pertenecen a la misma conjugación. En otras palabras, por la primera vía se trabaja con el inventario de lemas verbales y se evalúa la cantidad de ellos que hay que adscribir a cada conjugación. En la segunda, se cuantifica el número de veces que se utiliza cada forma verbal y se suman las frecuencias de todas aquellas que pertenecen a cada una de las tres conjugaciones. Inventario de lemas frente a uso de formas, en definitiva.

Naturalmente, no se trata de considerar que solo una de las dos interpretaciones sea correcta. Las dos preguntas son válidas, pero las respuestas tienen que tomar en consideración aspectos distintos. Las diferencias entre ambas posibilidades quedaban ya perfectamente claras en un concienzudo trabajo de Dolores Corbella (1987) en el que utiliza, tras larga y paciente reelaboración manual, los datos procedentes del FDSW. Reproduzco en mi cuadro 1, con una ligera adaptación, el que incluye Corbella (1987:148) como resumen de los datos obtenidos:

Cuadro 1. Distribución de formas verbales y verbos en el FDSW según su pertenencia a las diferentes conjugaciones. Fuente: Corbella (1987:148 y sigs.). Reelaboración propia

	Formas verbales		Verbos	
	Frecuencia	Porcentaje	Frecuencia	Porcentaje
-ar	27 777	37,59	656	68,55
-er	33 834	45,78	149	15,57
-ir	12 291	16,63	152	15,88
TOTALES	73 902	100,00	957	100,00

Las columnas de la izquierda muestran la frecuencia absoluta y relativa de las formas verbales en el corpus manejado para la elaboración del FDSW. Las de la derecha reflejan el número (y el porcentaje que le corresponde) de los

verbos pertenecientes a cada una de esas tres conjugaciones registrados en el FDSW. Lo primero que salta a la vista es, por supuesto, la fuerte diferencia existente entre ambos recuentos. Los verbos en *-ar* suponen el 68,55% de los verbos reflejados en el FDSW, pero sus formas alcanzan solo el 37,59% de la frecuencia total de las formas verbales. En sentido contrario, los verbos en *-er*, que son muy poco más del 15% de la lista de verbos, tienen en conjunto una frecuencia tan alta que la utilización de las formas pertenecientes a sus paradigmas se sitúa en el 45,78% del total. Para Corbella, estos resultados refuerzan los obtenidos previamente por H. Guiter para varias lenguas románicas[3] y derivan fundamentalmente de "la elevada cantidad de verbos polisémicos y formadores de perífrasis que se incluyen en la segunda conjugación" (Corbella 1987:149). Siempre según Corbella, es necesario por tanto matizar las afirmaciones del *Esbozo* para el español y de Menéndez Pidal para el latín y las lenguas románicas acerca de la prioridad de la primera conjugación, puesto que "si sólo consideramos el valor de 'V' [los verbos, GR], el grupo formado por los infinitivos en *-ar* es el que presenta mayor frecuencia, pero no es el que posee más variantes ni el que tiene mayor uso" (*ibídem*).

Parece existir en las afirmaciones precedentes una cierta confusión entre número de verbos, frecuencia de verbos, número de formas verbales y frecuencia de formas verbales[4] que se resuelven diferenciando con claridad, como se hace aquí en los cuadros 1 y 2, entre el número de verbos perteneciente a cada una de las conjugaciones y las frecuencias de uso de las formas que integran los paradigmas de esos verbos. En otras palabras, es la aplicación a lo obtenido en el análisis de un corpus concreto –el que está detrás del FDSW– de la diferencia entre la estadística del uso y la estadística del inventario.

Con esos dos aspectos bien diferenciados, las divergencias entre las tres conjugaciones se presentan con toda claridad si ponemos en relación el nú-

---

[3] Para Guiter (1971:359) "la conjugaison en *a* se caractérise par un nombre très important de vocables; la conjugaison en *e* par une fréquence d'emploi souvent élevée. Ces deux éléments font simultanément défaut à la conjugaison en *i*, remarquable tant par le petit nombre des vocables que par celui des ocurrences".

[4] En efecto, el *Esbozo* (268) se limita a indicar que "de los tres grupos en que pueden clasificarse los verbos españoles según la conjugación a que pertenecen, el primero es con mucha diferencia el más numeroso. Es también el más estable y productivo". La misma ausencia de delimitación clara entre número de elementos pertenecientes al modelo y mayor o menor frecuencia de uso se manifiesta también, quizá con mayor claridad, en el cuadro de la página 149, en el que, con datos del *Dictionnaire fréquentiel et Index inverse de la langue latine* (Delatte 1981), refleja frecuencias de uso de formas de las cuatro conjugaciones latinas y del mayor peso porcentual de las correspondientes a empleos de verbos de la tercera conjugación deduce que "tampoco en latín la llamada primera conjugación era la más numerosa" (Corbella 1987:149).

mero de verbos de cada una con la frecuencia de uso del conjunto de sus formas, es decir, si calculamos la cantidad de veces que, por término medio, figuran en el corpus del FDSW los verbos de cada conjugación, que es lo que muestra el cuadro número 2:

Cuadro 2. Frecuencia media de utilización de los verbos de cada una de las tres conjugaciones. Fuente: Corbella (1987:148). Reelaboración propia

	Frecuencia de las formas	Número de verbos	Utilización media de los verbos de cada conjugación
-ar	27 777	656	42,34
-er	33 834	149	227,07
-ir	12 291	152	80,86
TOTALES	73 902	957	72,22

La proporción global obtenida ilustra con toda claridad los datos contenidos en el cuadro 1: los verbos en *-ir* presentan una utilización media bastante próxima a la general, que supone casi el doble de la que tienen los verbos en *-ar*, pero esa diferencia, sin duda importante, resulta muy inferior a la que encontramos en los verbos en *-er*, que casi sextuplican en media de uso a los de la primera conjugación.

La discrepancia resulta muy llamativa, como ya señalaron Guiter (1971) y Corbella (1987), y su gran entidad hace suponer que refleja un fenómeno general, por lo que es esperable que se presente más o menos de la misma forma en el análisis de cualquier otro corpus español. Para confirmar esa posibilidad, he obtenido los resultados que arroja el análisis de la *Base de datos sintácticos del español actual* (BDS),[5] que son los que figuran en el cuadro número 3:

[5] Disponible en http://www.bds.usc.es. El proyecto fue financiado en su primera fase (1988 a 1991) por la *Dirección Xeral de Ordenación Universitaria e Política Científica* de la *Consellería de Educación* de la *Xunta de Galicia* (referencia XUGA 82710088) y por la Dirección general de Investigación científica y técnica del Ministerio de Educación y Ciencia entre 1991y 1994 (ref. PB90 0376). A lo largo de los quince años transcurridos desde su comienzo, han sido muchas las personas que, en diferentes grados y con distintas tareas, han colaborado en él. En el momento de escribir este trabajo (diciembre de 2002) siguen vinculados a la BDS los siguientes miembros del equipo inicial, todos ellos adscritos a la Universidad de Santiago de Compostela salvo indicación en contrario: Francisco García Gondar, José María García-Miguel (Univ. de Vigo), Belén López Meirama, Inmaculada Mas Álvarez, María José Rodríguez Espiñeira, Guillermo Rojo y Victoria Vázquez Rozas. Además, colaboran activamente en la corrección final y la aplicación de los resultados a diferentes fenómenos gramaticales: Cristina Blanco Canosa, Fernando Castro Paredes, Eva Muñiz Álvarez, Marta Rebolledo Lemus, María Paula Santalla del Río y Susana Sotelo Docío. Para detalles acerca de las características y posibilidades de explotación de la BDS, puede verse, además de la documentación existente en la propia página web, Rojo (2001). En la actualidad, además de continuar la revisión de los datos existentes, el grupo ha emprendido la tarea de refinar los análisis en un proyecto financiado por la *Secretaría Xeral de Investigación e Desenvolvemento* de la *Xunta de Galicia* (referencia PGIDT00PXI20410PR).

Cuadro 3. Distribución de formas verbales y verbos según las tres conjugaciones. Fuente: *BDS*. Elaboración propia

	Formas verbales		Verbos		Media de
	Frecuencia	Porcentaje	Frecuencia	Porcentaje	uso
-ar	88 058	45,94	2800	81,46	31,45
-er	71 495	37,29	296	8,61	241,54
-ir	32 148	16,77	341	9,92	94,28
TOTALES	191 701	100,00	3437	99,99	55,78

Las diferencias con respecto a lo que se observa en los cuadros 1 y 2 son importantes. En primer lugar, en la BDS, que tiene un número de verbos más de tres veces mayor que el que figura en el FDSW, el porcentaje de los que pertenecen a la primera conjugación es unos trece puntos superior al obtenido en el primer caso (81,46% frente al 68,55%). La reducción que experimentan las otras dos conjugaciones es similar (7 y 8 puntos porcentuales, respectivamente). Como consecuencia, la imagen que obtenemos en los dos corpus presenta un perfil parecido: la primera conjugación tiene una mayoría muy marcada y las otras dos, a mucha distancia de ella, presentan porcentajes muy próximos, pero la diferencia resulta mucho más clara en la BDS, donde ocho de cada diez verbos documentados en este corpus pertenecen a la conjugación en *-ar*, mientras que solo uno de cada diez corresponde a cada una de las otras.

La similaridad relativa observada en el número de verbos desaparece por completo si pasamos a comparar la frecuencia de las formas. Según los datos obtenidos por Corbella del FDSW (cf. *supra*, cuadro 1), casi una de cada dos formas empleadas en los textos pertenece a verbos en *-er*. Pues bien, eso es lo que ocurre en la BDS con respecto a los verbos en *-ar*. Como deja ver la comparación de los cuadros 1 y 3, los porcentajes de uso son muy semejantes, pero están cruzados, de modo que el que corresponde a formas de verbos de la primera conjugación en el recuento del FDSW está muy próximo al que encontramos en la BDS para los verbos en *-er* y el porcentaje que presentan los verbos de la segunda en el FDSW es cercano al que la BDS arroja para los verbos en *-ar*. Los papeles, pues, están invertidos, idea reforzada por el hecho de que, frente a lo que ocurre con los otros dos modelos, el porcentaje de los verbos en *-ir* es muy parecido en los dos recuentos.

¿A qué pueden deberse esas diferencias? Como ya he señalado en otros lugares (cf., p.e., Rojo 2001:264), la finalidad con que fue construida la BDS, el estudio de las capacidades combinatorias de los verbos, obliga a referir cada cláusula fichada al verbo que desempeña la función de predicado. Como consecuencia de ello, los casos de formas compuestas y otras perífrasis verbales han sido sistemáticamente atribuidos al verbo que constituye

la forma auxiliada y no figuran, en cambio, entre los correspondientes al que funciona en ese caso como forma auxiliar. Eso supone una considerable diferencia en el número de casos de verbos como *haber*, *estar*, *ir* y algunos otros muy frecuentes que presentan usos independientes y también empleos como auxiliares, además de repercutir en la frecuencia total de los verbos.[6] El sistema empleado en la BDS, que es el único utilizable para estudios relacionados con las estructuras clausales y los predicados que aparecen en ellas, es también el más adecuado para el cómputo de las frecuencias léxicas, pero resulta mucho más complicado de llevar a cabo por la necesidad de añadir a los recuentos automatizados el conocimiento lingüístico necesario para saber diferenciar los casos de perífrasis verbal de los de aquellas construcciones verbales que no lo son.

La entidad cuantitativa de la diferencia entre las dos aproximaciones es irrelevante si pensamos en el número de verbos, pero tiene gran importancia, en cambio, si nos referimos a su frecuencia. Supone, por ejemplo, que el verbo *haber*, que aparece en el segundo puesto de la lista de frecuencias de esta clase de palabras en el FDSW y otros corpus, queda relegado al décimo lugar, con el 1,5% del total, en la BDS (cf. Rojo 2001:265). Las cifras resultan perfectamente claras. El FDSW, que diferencia en las listas de frecuencia de formas entre los usos de *haber* como auxiliar y como verbo independiente, contiene 5085 casos de formas compuestas. Si las eliminamos de los recuentos, los datos referidos a uso del cuadro número 1 se alteran de forma considerable. Si, por el contrario, en las cifras obtenidas en la BDS añadimos los 10 969 casos de formas compuestas entre los correspondientes a la segunda conjugación, dejando las otras dos con las cantidades previas, el cuadro resultante se modifica también para acercarse al perfil general que muestran los procedentes del FDSW. Como muestra el cuadro número 4, la regularización del tratamiento de las formas compuestas (en cualquiera de los dos sentidos posibles) acerca considerablemente los resultados obtenidos en los dos corpus, aunque persisten ciertas diferencias que, por ejemplo, implican que las formas mayoritarias son siempre las de la primera conjugación en la BDS y las de la segunda en el FDSW. La distancia, sin embargo, no es ya tan grande como muestran las columnas extremas del cuadro 4, que son las que presentan los tratamientos heterogéneos de este factor. Si no tenemos en cuenta las formas compuestas, las medias de uso de la segunda

[6] Efectivamente, se produce una importante discrepancia en el número de casos contabilizados. Una secuencia como *hemos estado paseando*, por ejemplo, supone un caso para cada uno de los verbos *haber*, *estar* y *pasear* en los recuentos del FDSW y otros corpus, mientras que la BDS registra únicamente un caso de *pasear*, aunque indica también que se trata de una forma compuesta y utilizada en una cierta perífrasis.

conjugación son 192,95 en el FDSW y 241,53 en la BDS (cf. cuadro 3). Si las consideramos, las medias son 227,07 en el FDSW (cf. cuadro 2) y 278,59 en la BDS.

Cuadro 4. Porcentajes de uso de las formas verbales en BDS y FDSW. Fuentes: BDS, Corbella (1987) y Juilland y Chang (1964). Elaboración propia

	Sin *haber* auxiliar		Con *haber* auxiliar	
	BDS	FDSW	BDS	FDSW
-ar	45,94	40,36	43,45	37,59
-er	37,29	41,78	40,69	45,78
-ir	16,77	17,86	15,86	16,63
TOTALES	100,00 (N = 191 701)	100,00 (N = 68 817)	100,00 (N = 202 670)	100,00 (N = 73 902)

Podría suceder que las divergencias fuesen el resultado de las distintas configuraciones de ambos corpus. En efecto, la BDS está formada por textos escritos y orales editados o producidos con posterioridad a 1980, mientras que los que constituyen el FDSW son solo escritos y fueron publicados entre 1920 y 1940 (cf. Juilland y Chang 1964:XVI-XXI). Tales diferencias –cabe suponer– pueden dar lugar a que cada uno de estos corpus contenga inventarios léxicos notablemente distintos y de ahí deriven las diferencias en la distribución. La idea, sin embargo, no resiste el primer análisis: si excluimos dos variantes,[7] *soler* es el único verbo del FDSW que no figura en la BDS. Dado que esa ausencia está justificada por lo ya indicado acerca del distinto modo de asignar el predicado a cada cláusula en la BDS, se puede concluir que todos los verbos del FDSW están contenidos en el inventario de la BDS y que, en consecuencia, no cabe pensar en la existencia de diferentes inventarios léxicos: el del FDSW es un subconjunto del documentado en la BDS.

Podría pensarse también que, al lado de ese núcleo común, el corpus de la BDS contuviera elementos peculiares, elementos que, en definitiva, hicieran que una parte de su inventario, precisamente el que no coincide con el del FDSW, resultase escasamente representativo y, en consecuencia, desviado de lo habitual en español. Un modo de comprobarlo consiste en comparar los dos corpus que estamos contraponiendo a otros de características semejantes, o incluso a algunos diccionarios. En efecto, en los diccionarios están neutralizadas las diferencias de frecuencia en tanto que todos los ele-

[7] El FDSW incluye *desparecer* (también *desaparecer*) y *transponer*. El *concedir* que figura en la relación de la segunda parte es una errata; en las listas aparece *conceder*.

mentos tienen frecuencia igual a uno, pero pueden, en cambio, proporcionar un índice útil para juzgar la congruencia de la distribución de los elementos documentados en un corpus con la que se da en el propio inventario. En ese sentido, el contraste de la distribución de los verbos entre las tres conjugaciones en diccionarios y corpus puede proporcionar una impresión basada en datos acerca de la mayor o menor adecuación de un corpus como muestra representativa del inventario léxico de una lengua en un momento determinado de su historia.

El cuadro número 5 muestra la distribución porcentual de los verbos entre las tres conjugaciones en dos diccionarios –DRAE y GDLE–[8] y cinco corpus de características y tamaños diferentes.[9] Los datos resultan realmente nítidos: los dos diccionarios y los tres corpus de mayor tamaño muestran una semejanza poco habitual, puesto que las diferencias máximas que se dan entre un diccionario y un corpus no llegan a alcanzar los cinco puntos porcentuales. Los dos corpus más pequeños, en cambio, resultan relativamente próximos entre sí, pero están muy alejados de los otros conjuntos. En realidad, más que la semejanza entre los resultados de estos dos corpus sorprende la distancia que presentan, puesto que Guiter obtuvo los datos de un subconjunto de los publicados en el FDSW,[10] lo cual podría explicar su posición extrema. La impresión general es clara: en el FDSW las dos conjugaciones minoritarias están sobrerrepresentadas y ese rasgo está especialmente marcado en el subconjunto utilizado por Guiter, en el que la segunda conjugación contiene un porcentaje de verbos que resulta casi tres veces superior al que podemos encontrar en los diccionarios y en otros corpus.

---

[8] Utilizo las versiones electrónicas de la vigésima primera edición del Diccionario de la Real Academia Española (DRAE) y de la primera del Gran Diccionario de la Lengua Española (GDLE) *Larousse*.

[9] LexEsp es un corpus del español actual constituido por fragmentos de textos tomados de noticias de diferentes temas, editoriales, ensayos, novelas, etc. que contiene unos cinco millones y medio de formas. Es el resultado del proyecto LexEsp APC 96-0125, financiado por la CAICYT, dirigido por Nuria Sebastián Gallés. Los datos que aparecen aquí proceden de un recuento realizado por mí sobre el resultado de la lematización automática realizada en junio de 1998 por el grupo de la Universidad de Barcelona dirigido por Maria Antònia Martí. Agradezco a la Dra. Martí y a todo el equipo que participa en el proyecto la generosa cesión del corpus lematizado y anotado morfosintácticamente. Para más datos sobre las características del corpus y el proceso de anotación y desambiguación llevado a cabo, vid. Atserias *et alii.* (1998) y Carmona *et alii.* (1998). Para los textos que componen el corpus, consúltese http://www.uniovi.es/UniOvi/Apartados/Departamento/Psicologia/metodos/soft/corpus.
El que denomino aquí CREA_P1 es el primer prototipo anotado, desambiguado automáticamente y revisado manualmente del *Corpus de Referencia del Español Actual* desarrollado por la Real Academia Española. Para detalles, *vid.* http://www.rae.es.

[10] Después de indicar que ha tomado los datos del FDSW, señala Guiter (1969:132): "Pour l'espagnol, nous nous sommes limités au vocabulaire emprunté aux romans et aux essais, afin d'obtenir des résultats aussi comparables que posible avec ceux des autres langues". Su muestra, pues, consta de 200 000 formas (cf. Juilland y Chang 1964:xxvi).

Cuadro 5. Distribución entre las tres conjugaciones de los verbos contenidos en dos diccionarios y cinco corpus españoles

	DRAE[21]	GDLE	LexEsp	BDS	CREA_P1	FDSW	Corpus de Guiter
-ar	85,43	86,21	84,84	81,46	81,21	68,55	60,38
-er	7,88	7,70	6,93	8,61	8,41	15,57	21,70
-ir	6,69	6,09	8,23	9,92	10,37	15,92	17,92
TOTALES	100,00 (N=11 249)	100,00 (N=9398)	100,00 (N=5298)	99,99 (N=3437)	100,00 (N=3268)	100,00 (N=957)	100,00 (N= ?)

La descompensación que muestra el cuadro 5 se hará más evidente si, usando de nuevo la comparación de corpus y diccionarios, calculamos el porcentaje de verbos de cada conjugación presentes en el DRAE que son registrados en cada corpus. Los datos generales están en el cuadro 6, en el que se puede observar con claridad el descenso de los porcentajes a medida que lo va haciendo también el tamaño del corpus, desde los cinco millones de formas de LexEsp hasta el medio millón sobre el que se elaboró el FDSW.

Cuadro 6. Porcentaje de representación de verbos de las tres conjugaciones contenidos en el DRAE en diferentes corpus y el FDSW. Fuentes: Versión electrónica del DRAE[21], LexEsp, BDS, prototipo del CREA y Corbella (1987). Elaboración propia

	DRAE[21]	LexEsp	BDS	CREA_P1	FDSW
-ar	9610	46,77	29,14	27,61	6,82
-er	886	41,42	33,30	31,04	16,70
-ir	753	57,90	45,29	45,02	20,19
TOTALES	(N = 11 249)	(N = 5298)	(N = 3437)	(N = 3268)	(N = 957)

Es, por supuesto, el efecto esperable, ya que lo lógico es que el número de elementos diferentes vaya aumentando con la extensión del corpus. Sin embargo, el cuadro 6 muestra un rasgo adicional que no parece ser simplemente el efecto automático de las diferencias de tamaño. Si, tomándolo como elemento de comparación, hacemos igual a 1 el porcentaje de verbos de la primera conjugación contenidos en el DRAE y documentados en cada corpus y luego calculamos la proporción que suponen los otros dos, obtenemos los datos que figuran en el cuadro número 7. Se observa en él con toda claridad el descenso relativo de las conjugaciones segunda y tercera según aumenta el tamaño del corpus. Es de especial interés el caso de la segunda, que empieza siendo dos veces y media el porcentaje de la primera y termina –en los cinco millones de formas de LexEsp– por debajo de ella.

Cuadro 7. Proporciones de representación de las conjugaciones segunda y tercera sobre la primera en diferentes corpus y el FDSW. Fuentes: LexEsp, BDS, prototipo del CREA y Corbella (1987). Elaboración propia

	LexEsp	BDS	CREA_P1	FDSW
-ar	1	1	1	1
-er	0,88	1,14	1,12	2,45
-ir	1,24	1,55	1,63	2,96

No es un simple efecto del aumento el tamaño de los corpus. El problema de fondo radica en que el corte realizado en el FDSW distorsiona de modo importante la distribución de los verbos. Como ya he indicado anteriormente, del corpus utilizado para elaborar el FDSW resultaron –una vez eliminados nombres propios y extranjerismos– unos 20 000 lemas, que fueron reducidos a 5024 mediante diferentes criterios vinculados a la frecuencia, la dispersión y el uso (cf. Juilland y Chang 1964:LXXIV-LXXV). Los datos del FDSW no contienen, pues, los lemas correspondientes a un corpus de medio millón de formas, sino, en números redondos, el 25% más "frecuente" de esos lemas. Esa selección, como estamos viendo, produce un efecto deformador sobre la distribución de las tres conjugaciones, puesto que los verbos en *-ir* y, sobre todo, los verbos en *-er* son mucho más abundantes entre los verbos más frecuentes que en la distribución general.

No es difícil realizar comprobaciones adicionales. Si, como simple ejercicio de acercamiento, eliminamos de los datos de la BDS todos aquellos verbos que tienen frecuencia inferior a 5 –esto es, lo mismo que se hizo para el FDSW a pesar de las diferencias de tamaño entre los corpus de base–, los porcentajes de registro de verbos de cada modelo sobre los que hay en el DRAE pasan a ser del 14,26%, 24,37% y 31,87%, respectivamente. Ocupan, pues, un lugar intermedio con respecto a los del FDSW y la totalidad de los documentados en la BDS.

Una operación distinta, pero de sentido semejante, consiste en buscar la igualación de ambos conjuntos de verbos a base de tomar de la BDS únicamente los 1000 más frecuentes, con lo que se trabaja con un número similar al que las diversas operaciones de reducción han dejado en el FDSW.

Cuadro 8. Distribución porcentual de las tres conjugaciones en los 1000 verbos más frecuentes de la BDS y el FDSW. Fuentes: BDS y Corbella (1987). Elaboración propia

	BDS	FDSW
-ar	69,5	68,55
-er	15,0	15,57
-ir	15,5	15,92
TOTALES	100,0 (N = 1000)	100,00 (N = 957)

Los resultados, que aparecen en el cuadro 8, muestran una proximidad realmente notable, que supone una nueva prueba de lo que estamos manteniendo. Para valorar adecuadamente la importancia de esta semejanza en la distribución debe tenerse en cuenta que la coincidencia en los verbos seleccionados dista bastante de ser total. Como ya he indicado, todos los verbos que aparecen en el FDSW están también en la BDS –salvo tres casos justificados–, pero solo 783, es decir, el 82,08% de los 954 que figuran en las listas, están entre los 1000 más frecuentes de la BDS. La coincidencia en la distribución se da, pues, a pesar de que los lemas más frecuentes en cada conjunto difieren en un 20% aproximadamente.

La conclusión de todo lo anterior es clara: Corbella y Guiter han trabajado sobre un conjunto de datos bastante reducido y, sobre todo, parcial, en tanto que contiene únicamente el 25% de los lemas realmente documentados en el corpus de base. Esa selección, que probablemente carece de efectos importantes en otros fenómenos, resulta ser, sin embargo, especialmente fuerte en lo que nos ha ocupado aquí: el número y frecuencia de uso de verbos de las tres conjugaciones. La razón de ello está en el hecho de que, en los verbos más frecuentes, las conjugaciones segunda y tercera suponen un porcentaje de elementos muy superior al que se aprecia en una consideración más amplia. Dicho de otro modo, el porcentaje correspondiente a los verbos en *-ar* va aumentando a medida que lo hace el inventario. El proceso aparece con toda claridad en el cuadro número 9, en el que se observa la evolución del peso relativo de las tres conjugaciones en la BDS según va ampliándose el rango de frecuencias permitidas. Si entre los cien primeros la distribución es 46, 28, 26, cuando se considera la totalidad de los documentados se alcanza una relación del tipo 80, 10, 10. Los datos procedentes de los diccionarios (cf. *supra*, cuadro 5) muestran que el proceso continúa en la misma dirección.

Cuadro 9. Distribución porcentual de los verbos de las tres conjugaciones en diferentes conjuntos de frecuencias de la BDS. Fuente: BDS. Elaboración propia

-ar	46	58,0	63,8	69,5	75,90	81,46
-er	28	22,8	19,2	15,0	11,25	8,61
-ir	26	19,2	17,0	15,5	12,85	9,92
TOTALES	100 (N = 100)	100,0 (N = 250)	100,0 (N = 500)	100,0 (N = 1000)	100,00 (N = 2000)	99,99 (N = 3437)

Los datos del cuadro 9 proceden del mismo corpus y juegan con el rango de frecuencias de los elementos, pero parece claro que lo mismo sucede con la consideración de todos los verbos registrados en corpus que fuesen aumentando en tamaño: aparecerán más verbos, cada vez de menor frecuencia, con lo que el aumento se dará fundamentalmente en los verbos en *-ar*.

Debido al corte repetidamente mencionado, los datos manejados por Dolores Corbella presentan una distribución bastante diferente de la que, como hemos visto, resulta del análisis de conjuntos más amplios y representativos. Y, como es de esperar a partir de lo señalado, las diferencias en la frecuencia de uso son muy superiores a las que se pueden observar en el número de elementos. Con los datos de la BDS no se puede afirmar, como tiene que concluir Corbella con los procedentes del FDSW, que la frecuencia conjunta de los verbos en *-er* es superior a la que presentan los de la primera conjugación. El cuadro 10, en el que reflejo los porcentajes correspondientes a los usos de los verbos de las tres conjugaciones en los grupos de frecuencia manejados para el número de elementos muestra, de nuevo con toda claridad, lo que sucede realmente. Como antes, el peso de los verbos en *-ar* va ascendiendo y el correspondiente a los verbos en *-er* va descendiendo, pero ahora el proceso tiene lugar con un perfil tal que los porcentajes se cruzan y se invierten, de modo que de una relación aproximada del tipo 30, 50, 20 en los cien primeros se llega a otra del estilo 45, 37, 16 cuando los consideramos todos.

Cuadro 10. Distribución porcentual del uso de los verbos de las tres conjugaciones en diferentes conjuntos de frecuencias de la BDS. Fuente: BDS. Elaboración propia

	100 más frecuentes	250 más frecuentes	500 más frecuentes	1000 más frecuentes	2000 más frecuentes	BDS completa
-ar	32,21	38,04	41,43	43,85	45,38	45,94
-er	49,27	44,28	41,18	38,98	37,71	37,29
-ir	18,52	17,68	17,39	17,17	16,91	16,77
TOTALES	100,00 (N=124 251)	100,0 (N=149 496)	100,00 (N=168 085)	100,00 (N=181 523)	100,00 (N=189 230)	100,00 (N=191 701)

La evolución, pues, tiene el mismo perfil en el número de elementos y en su frecuencia de uso, pero el efecto resulta más fuerte en el caso de la frecuencia debido a que se produce incluso la alteración de la posición relativa. Los gráficos 1 y 2 muestran los datos contenidos en los cuadros 9 y 10, y explican, al mismo tiempo, la causa de las conclusiones erróneas alcanzadas por Corbella: el FDSW contiene datos que quizá son válidos para el estudio de las frecuencias léxicas, pero la selección llevada a cabo obliga a tomar muchas precauciones cuando se trata de utilizarlos en terrenos diferentes. Aunque no puedo demostrarlo, lo establecido aquí para el español es muy probablemente de aplicación a otras lenguas, con lo que las conclusiones obtenidas por Guiter para unas cuantas lenguas románicas deberían ser revisadas.

Gráfico 1. Distribución (en porcentajes) del número de verbos de las tres conjugaciones en diferentes conjuntos de frecuencias de la BDS

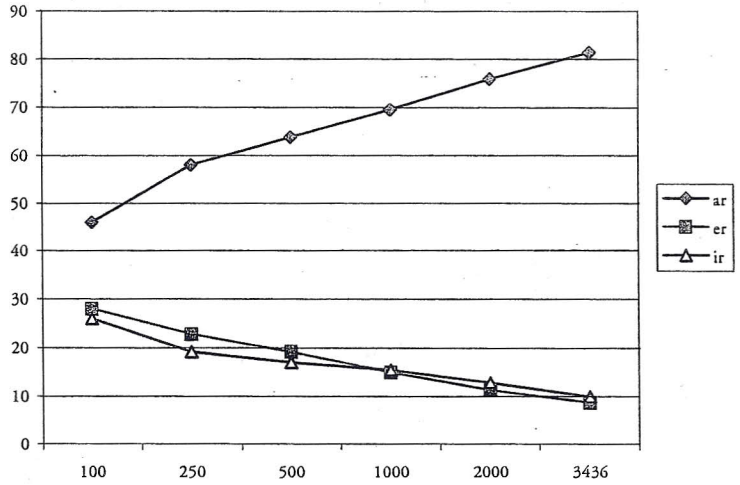
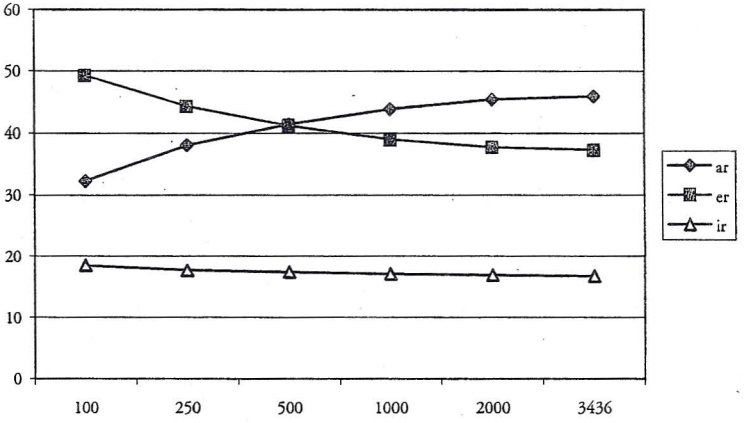


Gráfico 2. Distribución (en porcentajes) del uso de verbos de las tres conjugaciones en diferentes conjuntos de frecuencias de la BDS



Referencias

Academia, Real ___ Española. 1972. *Esbozo de una nueva gramática de la lengua española*. Madrid: Espasa-Calpe.

Atserias, J.; J. Carmona; I. Castellón *et alii*. 1998. Morphosyntactic analysis and parsing of unrestricted Spanish text. En *Proceedings of the 1st International Conference on Language Resources and Evaluation* (LREC'98), Granada.

Bybee, J. y P. Hopper (eds.). 2001. *Frequency and the emergence of linguistic structure*. Amsterdam / Philadelphia: John Benjamins.

Carmona, J.; S. Cervell; L. Márquez *et alii*. 1998. An environment for morphosyntactic processing of unrestricted Spanish text. En *Proceedings of the 1st International Conference on Language Resources and Evaluation* (LREC'98), Granada.

Chomsky, N. A. 1962. Comunicación presentada en la *3rd Texas Conference on Problems of Linguistic Analysis in English*, University of Texas, Austin. Cito por su reedición en J. Fodor y Katz (eds.), 1964. *The structure of language. Readings in the philosophy of language*, 211-245. Englewood Cliffs: Prentice-Hall.

Corbella, D. 1987. Algunos datos estadísticos del paradigma verbal español. En AA. VV. *In Memoriam Inmaculada Corrales*, tomo I, 145-159. Santa Cruz de Tenerife: Universidad de La Laguna.

Delatte, L. *et alii*. 1981. *Dictionnaire fréquentiel et index inverse de la langue Latine*. Liège: Lasla.

Guiter, H. 1969. Corrélations de signifiants et de signifiés dans les langues romanes. *Travaux de Linguistique et Litteratures (TraLiLi)* VII,1.131-159.

Guiter, H. 1971. Fréquences verbales dans les langues romanes. *RLR* 35.358-387.

Juilland, A. y E. Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words*. The Hague: Mouton.

Rojo, G. 2001. La explotación de la *Base de datos sintácticos del español actual (BDS)*. En J. de Kock (ed.), *Lingüística con corpus. Catorce aplicaciones sobre el español*, 255-286. Salamanca: Universidad de Salamanca. Disponible también en http://www.bds.usc.es.

Rojo, G. 2003. La frecuencia de los esquemas sintácticos clausales en español. En F. Moreno Fernández; F. Gimeno Menéndez; J. A. Samper; M.ª L. Gutiérrez Araus; M. Vaquero y C. Hernández (coords.), *Lengua, variación y contexto. Estudios dedicados a Humberto López Morales*, vol. I, 413-424. Madrid: Arco Libros. Disponible también en http://www.bds.usc.es.