FRECUENCIA DE INVENTARIO Y FRECUENCIA DE USO EN LOS ELEMENTOS GRAMATICALES*

GUILLERMO ROJO
Universidade de Santiago de Compostela

RESUMEN

En los últimos años han aparecido numerosas investigaciones que han puesto de relieve la importancia que tiene la consideración de la frecuencia en los diversos componentes lingüísticos. El presente trabajo, que se inscribe en la misma dirección, pretende poner de relieve los ajustes conceptuales necesarios para que su aplicación al componente gramatical resulte adecuada. Se propone aquí utilizar como marco genérico la distinción entre frecuencia de inventario y frecuencia de uso, cuya utilidad se muestra al aplicarla a varios fenómenos gramaticales con perspectiva sincrónica y diacrónica.

Palabras clave: Frecuencias en gramática, frecuencia de esquemas sintácticos, frecuencia y cambio lingüístico.

ABSTRACT

In recent years, various research works have highlighted the importance of taking the frequency of the different linguistic components into consideration. This paper, which supports the above view, emphasizes the need to make some conceptual adjustments so that frequency can be applied to the grammatical component. The paper establishes as its general frame a distinction between frequency of inventory and frequency of use. Such distinction will prove highly advantageous when applied to various grammatical phenomena both synchronically and diachronically.

Key Words: Frequency in grammar, frequency of syntactic constructions, frequency and linguistic change.

RECIBIDO: 19/09/2010 APROBADO: 18/10/2010

* Este artículo constituye la revisión y ampliación de la ponencia plenaria que presenté en el XXXIX Simposio de la Sociedad Española de Lingüística (Santiago de Compostela, febrero de 2010). Reitero aquí mi agradecimiento a la Junta directiva de la SEL y al Comité organizador del congreso por haberme hecho tan honroso encargo. Debo expresar también mi gratitud a Mercedes Sedano (UCV) y Victoria Vázquez Rozas (USC), que leyeron una versión previa de este trabajo y me hicieron importantes observaciones y sugerencias. Naturalmente, la responsabilidad de los defectos que todavía contenga es exclusivamente mía.

1. Introducción

Como todo el mundo sabe, la consideración de la frecuencia en los estudios lingüísticos ha sido muy cambiante. Poco atendida en la lingüística tradicional, por razones perfectamente comprensibles, duramente atacada en las fases inicial y central de la lingüística de orientación chomskiana, la frecuencia es hoy centro de atención en todos los componentes lingüísticos y en todas las perspectivas de análisis, desde la sociolingüística hasta la lingüística histórica, pasando, por supuesto, por la fonología, la morfología o la sintaxis. La importancia que tiene la frecuencia en los fenómenos lingüísticos es, como tantos otros que hemos visto en la historia de nuestra disciplina, un enfoque al que en principio nadie o casi nadie presta atención, es rechazado en una segunda fase y termina finalmente siendo una verdad aceptada, sin discusión, por todo el mundo. Bybee lo ha expresado magníficamente al escribir, en el primer párrafo de la presentación del libro en el que reúne sus trabajos sobre la frecuencia, lo siguiente:

> A newcomer to the field of linguistics might be surprised to learn that for most of the twentieth century facts about the frequency of use of particular words, phrases, or constructions were considered irrelevant to the study of linguistic structure. 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. So how is it that so many professional linguists for so many decades, maybe even centuries, have missed (or perhaps avoided) this basic point? (Bybee 2007, p. 5).

La causa de esa sorprendente falta de atención radica fundamentalmente, según Bybee, en el hecho de que las frecuencias pueden observarse bien en los niveles individuales (los de un cierto sonido, una palabra concreta, una determinada construcción sintáctica), mientras que los intereses básicos de los lingüistas se han movido casi siempre en la zona de los fenómenos generales, las estructuras abstractas, las líneas evolutivas genéricas y regulares (el cambio lingüístico en la concepción de los neogramáticos, por ejemplo). En una dirección similar apunta el predominio del interés por la cara más abstracta en las dicotomías del tipo lengua/habla o competencia/actuación. En los últimos años, sin embargo, se han dado varios factores que han alterado el panorama de forma considerable. El primero de estos factores, de acuerdo con Bybee, es la consideración de la gramática y las estructuras gramaticales como algo que emerge del propio discurso; en términos más generales, en la consideración del conocimiento gramatical como uno más de los procesos cognitivos, lo cual significa que la repetición, la regularidad, la convencionalización y todos los procesos semejantes son factores fundamentales, procedentes todos ellos de la producción lin-

güística, del discurso, con lo que se anula la separación preexistente entre el conocimiento lingüístico y su puesta en práctica. A este importante cambio teórico se añaden circunstancias adicionales que refuerzan el papel de las repeticiones en los procesos de adquisición del lenguaje y, de especial interés para nosotros, la posibilidad de trabajar con corpus lingüísticos de gran tamaño, también en los estudios diacrónicos, en los que se puede analizar el comportamiento real de los elementos lingüísticos, en muchas ocasiones notablemente alejado de lo que puede producir la introspección. Todos esos elementos han puesto de relieve la importancia de considerar las estructuras lingüísticas como sistemas complejos, que son aquellos en los que «a small number of mechanisms operate in real time and with repetition lead to the emergence of what appears to be an organized structure such as a sand dune» (Bybee 2007, p. 8).

Son todos ellos ajustes que han tenido lugar a lo largo de los últimos veinticinco años, de modo que la falta de consideración de las características y efectos de la frecuencia es propio de lo que podemos llamar ya la época clásica de la lingüística estructural, en todas sus variantes, y también la de raíz chomskyana. En efecto, la vinculación entre la frecuencia de los elementos y su comportamiento ante los cambios lingüísticos, su longitud media y otros rasgos que se han destacado en los últimos años estaban presentes de forma explícita en autores de finales del XIX y comienzos del XX, como han señalado Krug 2003 o la propia Bybee, entre otros. No parece, sin embargo, que esas conexiones vayan habitualmente más allá de una formulación general. Creo que la razón de ello está en que, además de los factores apuntados por Bybee, hay otros, que han tenido un papel de cierta importancia en la falta de atención a la frecuencia y sus efectos, y un peso considerable hasta hace muy poco tiempo. El más destacado de ellos es, a mi modo de ver, la propia dificultad asociada a los estudios estadísticos en la era pre-electrónica, es decir, en un mundo sin computadoras y, antes, sin calculadoras eléctricas. Por razones que me parecen perfectamente comprensibles dadas las dificultades existentes, la mayor parte de los estudios de frecuencias que se han llevado a cabo hasta hace muy pocos años se han realizado sobre elementos léxicos y, en un alto porcentaje de los casos, con el objetivo de contribuir a la mejora de la enseñanza y aprendizaje de segundas lenguas. Esas mismas limitaciones explican que el conjunto del que se obtienen los datos sea habitualmente pequeño o incluso muy pequeño, y no solo para los tamaños que estamos acostumbrados a manejar en la actualidad. Por recordar un caso bien conocido, los textos de los que proceden las estadísticas del *FDSW* de Juilland y Chang 1964 suman un total de 500.000 formas, que fueron fichadas y lematizadas por procedimientos manuales, aunque en los cálculos estadísticos y la configuración final fue ya posible usar una computadora. Para situar

un término de comparación cómodo, los diccionarios de frecuencias derivados del corpus *CUMBRE* (Almela y otros 2005) y una parte de los textos correspondientes al siglo XX del *Corpus del español* (Davies 2006) proceden del análisis de 20 millones de formas.

Las mismas razones explican que esas estadísticas, casi siempre léxicas como he dicho, se queden con mucha frecuencia en la perspectiva más general, sin entrar en el análisis detallado de lo que sucede en diferentes tipos de textos. El *FDSW*, que da las frecuencias de lemas y formas en cada uno de los cinco «mundos» diferentes que reconoce, es, en este punto, más la excepción que la regla y no me parece arriesgado pensar que la posibilidad de llevar a cabo los recuentos finales usando una computadora constituyó un factor decisivo para ello. Lo que acabo de apuntar tiene mayor importancia si se tiene en cuenta que los otros dos diccionarios mencionados, con muchos más recursos electrónicos, no pasan de la frecuencia y los índices generales, puesto que no creo que deban tenerse en cuenta en ese punto las indicaciones en las entradas de Davies acerca de la especial predilección o repugnancia de algunos lemas a aparecer en textos orales, de ficción o de no-ficción[1].

Por último, en parte por los objetivos con que se llevan a cabo, las dificultades apuntadas y, sobre todo, por los costes de las ediciones en papel, los inventarios de frecuencias léxicas no son completos casi nunca. Quiero decir que no contienen ni siquiera todo lo que aparece en el corpus limitado sobre el que trabajan. El *FDSW* contiene únicamente los 5024 lemas con mayor índice de frecuencia, que son la quinta parte de los lemas que aparecen en el corpus utilizado[2]. Siguiendo ese discutible modelo, los de Almela y otros 2005, y Davies 2006 tienen también los 5000 lemas más frecuentes de cada uno de los corpus manejados.

A partir de lo anterior se puede entender mejor la casi total ausencia de estadísticas sobre fenómenos gramaticales. Son más complejos que los elementos léxicos y más difíciles de reducir a unidades contables. De ahí que los no demasiado numerosos trabajos que incluían estadísticas gramaticales utilizaran habitualmente o bien recuentos totales de un número muy reducido de textos –casi siempre literarios, pero ese es otro problema– o bien recuentos parciales de conjuntos más amplios de textos. Por todo ello siguen resultando realmente asombrosas las listas de Keniston, a las que es necesario acudir todavía.

---

[1] Las indicaciones que aparecen en algunos lemas «show that the word in question has a high (+) or low (–) score (a combination of frequency and range) in the indicated register (oral, fiction, non-fiction). These symbols appear only when the word is in the top 8-9 percent or the bottom 8-9 percent of the words in that register, in terms of its relative frequency to the other two registers» (Davies 2006, p. 9).

[2] Incluyendo en el inventario nombres propios y extranjerismos. Sin estas dos clases de elementos, los lemas seleccionados se reducen a una cuarta parte de los registrados.

Este panorama, cargado de dificultades, cambia radicalmente con la aparición de las computadoras y la difusión de su uso en todos los campos científicos. En el nuestro, ese hecho tiene especial importancia, lo cual me ha llevado a caracterizarla como una auténtica revolución instrumental en el sentido de Dyson (cf. Rojo 2010). Como sucede en todos los ámbitos, la realización de estadísticas complejas resulta ahora infinitamente más sencilla que hace unos años, pero, sobre todo, tenemos la posibilidad de tratar y consultar con rapidez y comodidad enormes cantidades de textos, con cientos o incluso miles de millones de formas, a los que podemos acceder seleccionando los parámetros oportunos y adecuados en cada caso (tipo de texto, procedencia, época, etc.).

Por supuesto, no estoy describiendo ingenuamente un panorama en el que un trabajo plagado de dificultades sin cuento se haya convertido en una tarea que se pueda llevar a cabo de forma totalmente automática. Por ir a la zona aparentemente menos conflictiva, las estadísticas léxicas siguen encontrándose con los problemas de siempre a la hora de decidir cómo se estructura la lematización, pero presentan además todas las dificultades derivadas de la necesidad de resolver automáticamente cuál es el lema adecuado (según lo que se haya decidido previamente) en los numerosísimos casos de homografía. Este problema, que no se daba en los procedimientos «manuales», es solo la primera manifestación de las múltiples dificultades existentes para lograr la anotación sintáctica de corpus integrados por millones o cientos de millones de formas.

2. Los primeros años

A las dificultades internas, inherentes a la propia naturaleza del tratamiento computacional de los textos lingüísticos, se unió el rechazo del primer Chomsky tanto a la utilización de corpus lingüísticos como al papel de las consideraciones estadísticas en el análisis gramatical. Son dos aspectos distintos, pero vinculados en lo que aquí nos interesa, en tanto que la utilización de corpus lingüísticos es la única forma de obtener estadísticas realmente bien fundadas. Por tanto, hay que pensar que ese rechazo, formulado en el marco de una aproximación teórica que estaba entonces en plena línea ascendente y se caracterizaba por un fuerte sentido unitario en torno a las ideas de su representante máximo, retrasó durante al menos un par de décadas, la utilización amplia de los aspectos estadísticos en la gramática.

Los textos chomskianos son bien conocidos y no merece la pena insistir aquí en lo que ya se ha dicho en múltiples ocasiones. Me interesa en cambio señalar algo que no se destaca habitualmente, no al menos con la importancia que en mi opinión tiene. El hecho que estimo rele-

vante es que la invectiva de Chomsky tanto sobre los corpus como sobre la utilización de la estadística en gramática estaba referida a lo que se hacía en aquel momento desde la orientación distribucionalista, a lo que se había formulado como objetivo más o menos lejano o bien a lo que el propio Chomsky consideraba, con razón o sin ella, que era esperable desde esa aproximación. Está perfectamente claro en lo que se refiere a la utilización de corpus lingüísticos. La consideración de Chomsky 1962, p. 159 según la cual

> [a]ny natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list[3],

fue formulada como respuesta a una observación de A. G. Hatcher acerca de la posibilidad de extraer de un corpus el conjunto de las oraciones nucleares («kernel sentences») e intentar luego generar la totalidad del corpus a partir de ellas. No es difícil apreciar la importancia de tener en cuenta el contexto teórico en el que se formula esta observación y cuál es el objetivo del ataque.

Algo no muy distinto, *mutatis mutandis*, ocurre con la utilización de los elementos estadísticos en los estudios gramaticales. Ya en *Syntactic Structures*, al fijar la distinción entre oraciones gramaticales y agramaticales como objetivo fundamental de la teoría gramatical, señala Chomsky que para identificar el conjunto de secuencias gramaticales no sirve el corpus de oraciones que ha podido obtener un lingüista en su trabajo de campo, ni cabe identificar la noción de «gramatical» con «significativo». Por fin, utilizando las famosas secuencias *colorless green ideas sleep furiously* [=1] y *furioulsy sleep ideas green colorless* [=2], señala Chomsky 1957, pp. 15-16:

> [t]he notion «grammatical in English» cannot be identified in any way with the notion «high order of statistical approximation to English». It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sen-

---

[3] El texto, en lo que conozco, fue mencionado inicialmente por Leech 1991, p. 18 y de ahí pasó a muchos otros trabajos. Aunque no suele indicarse ese detalle, es crucial tener en cuenta que este fragmento no se encuentra en la versión publicada de la ponencia de Chomsky en la *3rd Texas Conference on problems of linguistic analysis*, celebrada en 1958, sino en la reproducción de la discusión posterior. Eso explica el carácter de réplica que he tratado de recoger aquí y, en otro sentido, el hecho de que no se pueda localizar el texto en la más conocida reimpresión de la ponencia que figura en Fodor y Katz 1964, pp. 211-245. En la continuación de la discusión, Hatcher opina que el posible sesgo de un corpus depende de su tamaño, Chomsky replica que incluso un corpus muy grande (todos los libros de la Biblioteca del Congreso, dice) tendría lagunas y estaría sesgado; Hatcher insiste en que su corpus de 125.000 oraciones no lo está, a lo cual Chomsky contesta que no puede haber una máquina que genere todas esas oraciones.

> tences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as being equally «remote» to English. Yet (1), though nonsensical, is grammatical, while (2) is not. [...] To choose another example, in the context *I saw a fragile* __, the words *whale* and *of* may have equal (e.g., zero) frequency in the past linguistic experience of the speaker who will immediately recognize that one of these substitutions, but not the other, gives a grammatical sentence.

De lo que concluye Chomsky 1957, p. 16:

> Evidently, one's ability to produce and recognize grammatical utterances is not based on notions of statistical approximation and the like.

Y más tarde, en las conclusiones, insiste en que «[t]he notion of grammaticalness cannot be identified with meaningfulness (nor does it have any special relation, even approximate, to the notion of statistical order of approximation)» (Chomsky 1957, p. 106).

Aunque, como ha señalado Ellis 2002, p. 163, las notas que acompañan a los textos reproducidos son un tanto más abiertas, especialmente en la referencia al empleo de modelos estadísticos diferentes a los utilizados o propuestos habitualmente en lingüística por aquellos años, la condena es clara: la frecuencia, la estadística, no tiene nada que ver con la gramaticalidad, que es, después de todo, el objetivo último de la teoría lingüística. Después de tantos años, no será necesario insistir en que es muy discutible que la probabilidad de aparición de esas dos palabras en el contexto mencionado sea la misma ni tampoco en que la consideración chomskyana de los factores estadísticos en gramática está desenfocada: no tiene el menor sentido trabajar con palabras concretas y oponerlas en la forma sugerida por Chomsky

Mucho más citado es el texto que procede de la ponencia de Chomsky en la conferencia de Texas a la que acabo de aludir. Introduce –recuérdese que estamos en 1958– la noción de regla de selección (es decir, una regla que permite diferentes reescrituras del mismo elemento en función del contexto en que aparezca). En nota, compara el mecanismo de descripción gramatical que se obtiene mediante la adición de reglas de este tipo al que se basa en el modelo de estados finitos, ejemplificado aquí en la forma en que lo había expuesto Hockett. Es evidente que este otro modelo resulta muy limitado, así que es necesario complementarlo. En ese momento, señala Chomsky:

> Hockett modifies this picture by assigning probabilities to the particular rules. However, this is an irrelevant complication. It seems that probabilistic considerations 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, p. 215, n. 10)[4].

Tener en cuenta estas circunstancias modifica considerablemente lo que se afirma en general con relación a este punto. Creo que, en el contexto concreto en el que se hace esa afirmación, Chomsky está en el camino adecuado al considerar que la vía correcta no es la de la simple adición de un componente probabilístico, que solo aplique el resultado del análisis general de las frecuencias observadas, pero sea ciego al contexto concreto en que se ha de aplicar la regla en cuestión. Ese modelo es muy insuficiente y la adición de las probabilidades generales es un recurso torpe, que no mejora su calidad.

Ya en general, el error de Chomsky está en la idea de que las consideraciones probabilísticas en gramática se refieren a si secuencias como *vivo/he nacido/soy de... Nueva York* son más o menos frecuentes que *vivo/he nacido/soy de... Nevada.* También en 1958 era claro que la estadística gramatical no se refiere a secuencias concretas, sino a estructuras lingüísticas. Para decirlo con palabras de Stefanowitsch, «corpus grammarians are not –and never have been– concerned with the frequency of individual sentences, but rather with the frequency of sentence *patterns*» (Stefanowitsch 2005, p. 295) y, por tanto, el dato que aduce Chomsky en ese fragmento está desviado y es irrelevante[5].

En resumen, es cierto que las opiniones de Chomsky acerca del papel de la frecuencia en la gramática son negativas y que una buena parte de su argumentación resulta inadecuada. Sin embargo, conviene tener en cuenta que fueron formuladas en un contexto concreto, refe-

---

[4] La referencia habitual a este planteamiento no opone Nueva York a Nevada, sino a Dayton (Ohio) (cf., por ejemplo, McEnery y Wilson 1996, p. 8), pero el único texto que yo he podido encontrar es el que reproduzco aquí. Cf. Stefanowitsch 2005, nota 1, para más detalles sobre estas discrepancias textuales, que tienen interés no solo filológico.

[5] No creo, en cambio, que Stefanowitsch tenga razón en la otra cuestión que plantea en esta nota. En su opinión, los lingüistas que trabajan con corpus no deberían, como hacen algunos, derivar de la afirmación de Chomsky la necesidad de complementar con datos procedentes de la introspección lo que resulta del análisis de los textos comprendidos en un corpus. Lo realmente interesante no es la frecuencia absoluta, sino la relativa, pensamiento indiscutible que lo lleva a estudiar si la frecuencia observada de estas secuencias en textos existentes en la red difiere de la esperada a partir de las frecuencias individuales de los elementos que las componen (*I live in* seguido de *New York* o *Dayton* y *I live in* con otros complementos) y una operación similar con los tamaños de las poblaciones respectivas. El hecho de que las frecuencias observadas difieran muy poco de las esperadas no nos dice nada acerca de la gramática, ni siquiera de la bondad del corpus manejado. Si el resultado hubiera sido el contrario, tampoco tendría incidencia sobre los aspectos gramaticales, pero habría que pensar que, por alguna razón, la composición del corpus utilizado resulta incongruente con los datos de población o con los derivados de su propia composición lingüística, pero sin incidencia real sobre el papel de las consideraciones estadísticas en el análisis gramatical ni en la teoría gramatical.

ridas a una cierta aproximación, y que tanto los partidarios de la orientación chomskyana como sus opositores han olvidado ese factor. Creo que es un capítulo más en la larga serie de incomprensiones derivadas del desconocimiento que los generativistas –con Chomsky a su cabeza– han tenido del estructuralismo europeo y de la obcecación de muchos lingüistas en considerar como genéricas o incluso específicamente referidas al estructuralismo europeo las observaciones del Chomsky de los primeros tiempos con relación al distribucionalismo.

3. La frecuencia en el componente gramatical

En cualquier caso, es cierto que los factores internos y la mala consideración que la frecuencia tuvo entre los generativistas impidieron que su estudio pudiera experimentar todo el desarrollo que habría sido posible durante los primeros veinte años de difusión de la lingüística de corpus. Afortunadamente, esa etapa ha sido superada ya y, como hemos visto en el primer apartado, hoy son mayoría los lingüistas convencidos de que la frecuencia de los elementos y estructuras es un factor del que no se puede prescindir, que explica los diferentes comportamientos que se observan en un período determinado y resulta decisivo en la comprensión del cambio lingüístico.

El estudio de la frecuencia de los tipos y subtipos de unidades sintácticas constituye una de las parcelas más complicadas (y también más apasionantes) de todo este territorio. Las razones de ello proceden, en primer lugar, de la propia complejidad interna de las entidades abstractas con las que hemos de enfrentarnos. En efecto, llegar a la determinación de una entidad como «esquema sintáctico constituido por una construcción activa con sujeto, predicado, complemento directo y complemento de régimen» y la posterior identificación de las cláusulas que en un determinado conjunto de textos responden a este esquema supone, como mínimo, la caracterización de las diferentes construcciones y funciones sintácticas, la delimitación entre ellas y la determinación general de funciones argumentales frente a las no argumentales, todo ello tanto en términos genéricos, teóricos, como en su aplicación a secuencias concretas. De otro lado, puesto que la mayor parte de los estudios acerca de frecuencias se han hecho sobre elementos léxicos, es inevitable que una buena parte de la perspectiva general con que se llevan a cabo estos trabajos y también de las herramientas metodológicas se construyan sobre lo que sucede en este componente, lo cual puede dar lugar a un cierto desajuste cuando se trabaja en sintaxis.

Comencemos por este último punto. Aunque no se da ha dado un gran número de reflexiones teóricas sobre esta distinción, se han manejado, de modo más o menos formalizado, caracterizaciones que pue-

den ser reducidas a las que voy a llamar «frecuencia en el inventario» y «frecuencia en el uso». Como ha señalado Bakker 1968, su establecimiento y utilización puede remontarse ya a Mathesius, que, en una fecha tan temprana como 1929, postulaba la necesidad de estos dos enfoques para lograr la comprensión total de las características de un sistema fonológico determinado. Con sus propias palabras, se puede

> étudier un système phonologique dans la composition et les rapports réciproques de ses termes. Mais on peu aussi étudier le répertoire des éléments phonologiques à titre de matériaux fonctionnels et se préoccuper de l'emploi particulier qui en est fait dans le courant du discours ou dans le lexique. Les résultats de ces différentes méthodes se compléteront, confirmeront ou corrigeront mutuellement. C'est de leur emploi combiné seulement que résultera la caractérisation phonologique complète d'une langue étudiée (Mathesius 1929, p. 67).

Con ejemplos triviales, se obtiene un cierto panorama comparando los porcentajes relativos de, por ejemplo, vocales y consonantes en una lengua determinada o poniendo en relación los de varias lenguas. Surge otra perspectiva, que puede resultar considerablemente distinta, observando la frecuencia con que esos elementos aparecen en el conjunto de formas de cada una de esas lenguas o en una serie de textos formulados en ellas (cf. infra). Lo mismo sucede, claro está, con la estructura silábica (tipos de sílaba posibles frente a frecuencia con que aparecen), etcétera.

Por su parte, Joan Bybee, que, como es sabido, ha dedicado gran atención al estudio de la frecuencia y sus implicaciones en los más diversos terrenos, pero con especial interés por la gramática, ha popularizado en estos últimos años la distinción entre las que llama *token frequency* y *type frequency*. La *token frequency* «counts the number of times a unit appears in running text. Any specific unit, such as particular consonant [s], a syllable [ba], a word *dog* or *the*, a phrase *take a break* or even a sentence such as *Your toast popped up* can have a token frequency» (Bybee 2007, p. 9). La *type frequency*, en cambio,

> is a very different sort of count. Only patterns of language have type frequency, because this refers to how many distinct items are represented by the pattern. Type frequency may apply to phonotactic sequences; it would be the count of how many words of the language begin with [sp] versus how many begin with [sf]. It may apply to morphological patterns, such as stem + affix combinations. For instance, the English past tense pattern exemplified by *know, knew, blow, blew* has a lower type frequency than the regular pattern of adding the *-ed* suffix. Syntactic patterns or constructions also have type frequencies: the ditransitive pattern in English, exemplified by *He gave me the change*, is used with only a small set of verbs, while the alternate pattern *He gave the change to me* is possible with a large class of verbs (Bybee 2007, pp. 9-10).

En otras palabras, la *token frequency* se refiere a la frecuencia de un cierto elemento en los textos. Aunque volveré más tarde sobre este tema, me interesa resaltar que, en esta presentación, se puede apreciar ya que las unidades a las que Bybee hace referencia son elementos de un nivel de abstracción relativamente bajo: un fonema, una sílaba determinada, una secuencia específica concreta, etc. La *type frequency*, en cambio, se refiere al número de elementos que forman o constituyen una clase determinada: número de fonemas vocálicos de una lengua, número de palabras distintas que tienen una determinada configuración morfológica, número de verbos que aceptan un determinado esquema sintáctico, etc. Para este último caso, indica que «[t]ype frequency in syntactic constructions would count how many distinct items of a particular lexical or grammatical class (e.g. verbs) can be used in the construction» (Bybee y Thompson 1997, p. 269).

Las dos distinciones presentadas son interesantes, pero necesitan, para que puedan resultar realmente útiles en el componente sintáctico, algunas adaptaciones o, cuando menos, alguna aclaración acerca de cómo deben ser entendidas cuando se utilizan en este terreno. La formulación de Mathesius no puede ser aplicada sin más a, por ejemplo, el componente léxico, puesto que, como ha señalado Bakker 1968, p. 13, «[i]t is obvious that word-frequencies in running texts and in the lexicon cannot be compared, since in the latter every word occurs with the persistent frequency of 1»[6]. Eso es cierto, pero reconvertirla y reinterpretarla en la distinción entre frecuencia en el inventario y frecuencia en el uso permite su empleo sin restricciones aparentes en el componente gramatical.

Veámoslo con un ejemplo claro. El recuento de los lemas de un diccionario (o un lemario obtenido del procesamiento de un corpus) proporciona una distribución determinada de, por ejemplo, las clases de palabras que constituyen el componente léxico de una lengua o de sus manifestaciones en un conjunto más o menos amplio de textos. De ahí obtenemos que el léxico de esa lengua está constituido por un cierto porcentaje de sustantivos, otro porcentaje de adjetivos, verbos, preposiciones, etc. Una visión totalmente distinta produce el análisis de las frecuencias de las diferentes clases de palabras en los textos. La razón es clara: la frecuencia de todos los elementos en un lemario es igual a 1, mientras que la que los lemas presentan en los textos es la suma de las apariciones de todas las formas que pueden ser adscritas a ese lema. El perfil resultante es siempre la aparición de unos pocos elementos (le-

[6] Y continúa: «Comparable are only the frequencies of those units that occur both in running texts and in the lexicon at a specific frequency, such as phonemes, phoneme-combinations, structures of syllables and morphemes, in other words: units that are characteristic for the system of a language».

mas en este caso) con un alto índice de frecuencia y muchos elementos con índices de frecuencia muy bajos, en gran parte con frecuencias igual a 1. Véase, como simple muestra de lo que puede obtenerse por cualquier parte, para cualquier lengua y cualquier conjunto de textos, lo que resulta del análisis de algunas clases de palabras en las aproximadamente 265.000 unidades que constituían la primera versión pública del *CORGA* etiquetado (abril de 2009)[7]:

Clase de palabras	En el inventario		En los textos	
	Frecuencia	Porcentaje	Frecuencia	Porcentaje
Sustantivos (comunes)	5046	48,70%	58.249	23,68%
Adjetivos	2900	27,99%	20.593	8,37%
Verbos	1754	16,93%	29.823	12,13%
Adverbios	419	4,04%	10.163	4,13%
Preposiciones	40	0,39%	48.468	19,71%
Conjunciones	31	0,30%	12.352	5,02%
Pron. personales	19	0,18%	5119	2,08%
Otras clases	153	1,48%	6193	24,88%
Totales	10.362	100%	245.960	100%

CUADRO 1: Distribución de algunas clases de palabras en la versión del *CORGA* etiquetado cerrada en abril de 2009[8].

Fuente: *CORGA* <http://corpus.cirp.es/corgaetq/>. Elaboración propia.

Dada la finalidad con que lo incluyo aquí, he configurado el cuadro anterior de modo que ni la multiplicidad de clases ni su inevitable dependencia del modo en que se agrupen (esto es, de la forma en que se establezcan las clases, los lemas, etc.) pueda distraer del objetivo princi-

[7] Esta primera versión pública (cerrada en abril de 2009), constaba de unas 250.000 formas gráficas, procedentes de noticias de prensa sobre economía (cf. <http://corpus.cirp.es/corgaetq/>). Es un subconjunto del *Corpus de referencia do galego actual* (*CORGA*), que comprende algo más de 25 millones de formas correspondientes a textos comprendidos entre 1975 y 2009. Cf. <http://corpus.cirp.es/corga/>.

[8] No se han considerado signos de puntuación, cifras, locuciones ni categorías periféricas (abreviaturas, siglas, etc.). Además, dado que el tipo de textos incluidos en este corpus parcial (noticias de prensa) podría distorsionar los datos, se han excluido también los nombres propios.

pal, que es la comprobación de los diferentes panoramas que se obtienen al considerar la distribución de las clases de palabras de una lengua según se trabaje sobre el inventario o sobre los textos reales. Teniendo todo esto en cuenta, creo que la tabla es suficientemente expresiva de la diferencia que existe entre lo que una determinada clase de palabras supone en el inventario de elementos léxicos de un conjunto de textos (es decir, los *types* de la visión habitual en lingüística de corpus) y lo que se aprecia si se analiza eso mismo tal como aparece en los textos (esto es, la suma de las apariciones de las distintas clases de los que llamamos normalmente *tokens*).

La distinción quedará todavía más clara si observamos que, siguiendo con este ejemplo concreto, cada cálculo de frecuencias responde a dos preguntas diferentes. En la primera opción, la pregunta se refiere al número de sustantivos, adjetivos, etc., que hay en el inventario léxico de una lengua, concretado en un diccionario determinado, el lemario obtenido de un conjunto de textos, etc. En el segundo, la pregunta es acerca del número total de (casos de) sustantivos, adjetivos, etc., que existe en un conjunto de textos, en un corpus. Las dos preguntas son interesantes, pero, como es evidente, son también radicalmente distintas.

No estoy seguro de que se pueda dar cuenta de algo tan sencillo como lo anterior utilizando la distinción entre *token frequency* y *type frequency* propuesta y repetidamente aplicada por Bybee. La *type frequency* es el número de elementos que poseen una característica lingüística determinada, como (para usar los ejemplos preferidos de esta autora) el número de verbos que presentan las características morfológicas del tipo *know/knew* o el de los que entran en el esquema sintáctico ditransitivo *(he gave me the change)*. Aunque habría que aceptar que los rasgos que establecen la pertenencia a una clase determinada son en muchos casos menos claros y discutibles, no parece que pudiera haber problemas para considerar un caso de *type frequency* también el de la pertenencia a una cierta clase de palabras. En ese caso, la *type frequency* sería equivalente a la frecuencia en el inventario. No se ve con la misma claridad, en cambio, el salto que va desde la *token frequency*, definida como la que presenta en los textos una unidad determinada (una consonante, una sílaba, una palabra) hasta llegar a lo que se necesita en el caso anterior, que es la suma de todos los casos que presenta, en todas sus formas posibles, el conjunto integrado por todos los elementos de cada una de las clases definidas. Este salto constituye, sin embargo, una generalización imprescindible si queremos aplicar la distinción con comodidad a lo largo y ancho del componente gramatical.

Veamos otro ejemplo ilustrativo de la utilidad de esta ampliación. Aunque no se ha hecho muchas veces, la pregunta acerca de cuál es la frecuencia de las tres conjugaciones del español (o su equivalente para otras lenguas) oculta una notable ambigüedad, puesto que no queda

claro si hablamos del número de verbos que pertenecen a cada una de las conjugaciones (frecuencia en el inventario o *type frequency* de cada una de ellas) o a la frecuencia general de los verbos de cada una de las conjugaciones en los textos (frecuencia en el uso, pero no, al menos en interpretación estricta del concepto, *token frequency*). De nuevo con la técnica de antes, la primera opción responde a una pregunta acerca de cuántos verbos de cada una de las tres conjugaciones hay en español (si tal pregunta tiene sentido), en un determinado diccionario, en el lemario obtenido de un cierto corpus, etc. La segunda, en cambio, se refiere a la frecuencia con la que se pueden encontrar formas pertenecientes a los verbos de cada una de las tres conjugaciones en los textos. También aquí hay diferencias interesantes, que expongo sucintamente para poder referirme más tarde a este mismo problema en una dimensión distinta.

Tal como expuse en Rojo 2006, las frecuencias de inventario que aparecen al analizar las conjugaciones en diccionarios y en lemarios procedentes de corpus son bastante congruentes, a pesar de las esperables diferencias entre los dos grandes tipos de fuentes de datos. Puede verse una ilustración de este punto en el cuadro siguiente[9]:

	DRAE_2001	GDLE	LexEsp	BDS	CREA_P1
-ar	85,43%	86,21%	84,84%	81,46%	81,21%
-er	7,88%	7,70%	6,93%	8,61%	8,41%
-ir	6,69%	6,09%	8,23%	9,92%	10,37%
Totales	100% (N = 11.249)	100% (N = 9398)	100% (N = 5298)	99,99% (N = 3437)	100% (N = 3268)

CUADRO 2: Distribución entre las tres conjugaciones de los verbos contenidos en dos diccionarios y tres corpus españoles.

Elaboración propia.

Es evidente que la primera conjugación es absolutamente mayoritaria, puesto que, tanto en los diccionarios como en los lemarios obtenidos de los corpus, pertenece a este grupo entre el 81% y el 86% de los

[9] Reproduzo aquí una parte de los datos que aparecen en Rojo 2006, p. 319, cuadro 5. En ese lugar figura también la referencia detallada a las fuentes. Señalo aquí simplemente que se trata de dos diccionarios (la edición del DRAE 2001 y la primera edición del *Gran diccionario de la lengua española Larousse)* y tres corpus de pequeño tamaño: *LexEsp* (unas 500.000 formas), un prototipo del *CREA* etiquetado que constaba de un millón de formas y la *BDS* (v. infra), integrada por algo menos de 1,5 millones de formas.

verbos existentes en los distintos inventarios, mientras que las otras dos conjugaciones se reparten el resto de forma bastante equitativa. Los datos de la tabla coinciden, me parece, con la que probablemente sería la respuesta de los hablantes de español a una pregunta acerca de este punto, aunque quizá no pensarían en una distancia tan marcada entre la mayoritaria y las otras dos. Los datos del uso, sin embargo, presentan un panorama radicalmente distinto, como muestra el cuadro (3): la conjugación más usada sigue siendo la primera, pero aquí no llega ni siquiera al 50% del total, y, además, va seguida a menos de 10 puntos porcentuales por la segunda, mientras que la tercera queda a mucha distancia de las otras. Lo que se está midiendo aquí es, evidentemente, la frecuencia de los verbos españoles en los textos y esas frecuencias se agrupan luego en función de la pertenencia de cada verbo a una determinada conjugación[10].

-ar	45,94%
-er	37,29%
-ir	16,77%
Totales	100% (N = 191.701)

CUADRO 3: Distribución porcentual del uso de las formas de los verbos de las tres conjugaciones en la *BDS*.

Fuente: *BDS*. Elaboración propia.

No parece necesario insistir en que las dos visiones son adecuadas y proporcionan datos reales y significativos sobre la configuración que presenta esta zona de la gramática del español. Lo más interesante, de todas formas, es la muy llamativa diferencia que existe entre ambas, ya que, siendo igualmente válidas, proporcionan una imagen muy diferente de este aspecto de las frecuencias verbales. La razón de ello es evidente: la frecuencia en el inventario supone que todos los elementos tienen el mismo peso; en la frecuencia de uso, en cambio, cada uno de los considerados (en este caso, cada una de las tres conjugaciones) tiene el peso que corresponde a la suma de las frecuencias individuales de las formas que lo integran.

[10] Para todos los detalles acerca de la *Base de datos sintácticos del español*, v. <http://www.bds.usc.es>. Debo señalar aquí, puesto que es elemento de importancia en la interpretación de los datos del cuadro, que los recuentos se refieren únicamente a verbos principales: *iban a volver* o *hemos vuelto* son, en la *BDS*, casos del verbo *volver* y no se contabilizan en *ir* ni en *haber*.

Del análisis de estos dos fenómenos, cuya única dificultad se reduce a disponer de los recursos adecuados para hacer los recuentos de forma cómoda, podemos obtener algunas conclusiones acerca de cómo debemos entender y manejar los conceptos relacionados con las frecuencias, no solo en fenómenos de este tipo, sino en otros considerablemente más abstractos y complejos, como, por ejemplo, la frecuencia de las oraciones condicionales o de sus diferentes tipos, de un esquema sintáctico del tipo «construcción activa formada por sujeto, predicado, complemento directo y complemento de régimen», etc. A mi modo de ver, alcanzar ese objetivo requiere llevar a cabo dos ajustes que afectan de modo ligeramente distinto a las dos distinciones habituales.

Tal como la define Bybee, la *token frequency* parece un caso particular de frecuencia en el uso. En efecto, la utilización de la noción de frecuencia en el uso no supone restricción alguna sobre el carácter del elemento considerado, aunque no puedo excluir que ello se deba a que no se ha planteado nunca desde un ángulo teórico y la mayor parte de sus aplicaciones se han dado en el componente léxico. En Bybee, en cambio, parece existir una limitación en el empleo de la *token frequency* a elementos de carácter relativamente concreto, como un morfema, una palabra o una secuencia determinada. En la revisión, amplia, aunque no exhaustiva, que he hecho de sus definiciones, solo he encontrado un caso en el que los elementos con los que ilustra la noción de *token frequency* muestra un carácter más amplio. En el capítulo que escribió para el *Handbook of Historical Linguistics* abre la posibilidad de aplicar la noción a la construcción *be going to* + *V*, esto es, a todos los casos que esa construcción presenta en los textos analizados:

> Token or text frequency is the frequency of occurrence of a unit, usually a word or morpheme, in running text. For instance, *broke* (the past tense of *break*) occurs sixty-six times per million in Francis and Kučera 1982, while the past-tense verb *damaged* occurs five times in the same corpus. The token frequency of *broke* is much higher than that of *damaged*. We can also count the token frequency of a grammaticizing construction, such as *be going to*, by counting just those occurrences of *be going to* that are used with a following verb (rather than a noun) (Bybee 2003, pp. 338-339).

No quedan claras las razones de un salto tan fuerte (de las formas *broke* o *damaged* a los casos de una construcción perifrástica como la mencionada) sin alusión a la gran cantidad de elementos que podríamos situar entre ambos y muchos otros tipos, como los que he citado ya en varias ocasiones (una clase de palabras, las oraciones condicionales, un determinado esquema sintáctico, etc.). Quizá el hecho de que el trabajo del que procede esta última cita se centre en los fenómenos de gramaticalización y el interés especial que presenta el juego entre el aumento de la frecuencia de casos de esta construcción y la ampliación

en el número de verbos que entran en ella puedan explicar esa incorporación. En cualquier caso, la cita queda aislada, sin más referencia a las construcciones sintácticas (en proceso de gramaticalización o no) y, además, las formulaciones más recientes que ha hecho Bybee de la distinción vuelven a referirse exclusivamente a palabras y secuencias concretas.

Parece claro, sin embargo, que la utilización rentable de esta noción exige una formulación en la que podamos considerar que cabe hablar de frecuencia de uso (o de *token frequency* en esta interpretación amplia que propongo) para hacer referencia a las que, en niveles cada vez más elevados, tienen en los textos la forma *salíamos*, el conjunto de las formas del pretérito imperfecto de indicativo del verbo *salir*, las formas de indicativo del verbo *salir*, el verbo *salir*, los verbos de la tercera conjugación o los verbos de movimiento por un lado y todas las formas de pretérito imperfecto de indicativo de todos los verbos, todas las formas de indicativo o todas las formas verbales por otro. Por la misma razón, ahora con elementos de otra serie, debe ser posible hacer referencia a la frecuencia de uso (o, de nuevo, la *token frequency* en interpretación amplia) de las oraciones condicionales irreales de pasado con el esquema *si hubiera tenido, habría dado*, de las oraciones condicionales irreales de pasado, de las oraciones condicionales o, más en general, de las oraciones bipolares.

El segundo ajuste que estimo necesario, muy relacionado con el anterior, consiste en la aceptación del bien conocido carácter relativo de las nociones de elemento y clase (o cualesquiera otras equivalentes): lo que es una clase formada por un cierto conjunto de elementos en un nivel puede ser considerado un elemento que forma parte de otra clase en el nivel inmediatamente superior y así sucesivamente. En esa línea ascendente, los elementos y clases son de carácter cada vez más abstracto y de ahí la conexión con el aspecto mencionado anteriormente. La serie que comienza en la frecuencia de la forma *salíamos* y termina en la frecuencia del conjunto de los verbos de la tercera conjugación es un buen ejemplo de lo que quiero decir. De otro componente: el esquema sintáctico «construcción activa constituida por sujeto, predicado y complemento directo» tiene una *type frequency* equivalente al número de verbos que pueden aparecer en este esquema y una *token frequency* en interpretación amplia (equivalente pues a la frecuencia de uso) que es el número de cláusulas con este esquema que podemos encontrar en un corpus determinado[11]. Ahora bien, podemos además considerar este esquema como un elemento que forma parte de las clases «construccio-

[11] Naturalmente, no equivale a la frecuencia conjunta de los verbos que presentan ese esquema, puesto que es habitual que el mismo verbo pueda aparecer en construcciones sintácticas diferentes.

nes activas» y «construcciones biargumentales». La *type frequency* del metaesquema «construcción biargumental» es el número de construcciones de este tipo que podamos documentar en una lengua o en un conjunto de textos (es decir, S+D, S+I, S+R, etc.) y su frecuencia en el uso será la suma de los casos de todos esos esquemas que aparecen en los textos. Lo que necesitamos, por tanto, es diferenciar entre la frecuencia de un elemento o una clase de elementos en el inventario correspondiente y la frecuencia de un elemento o una clase de elementos en los textos, teniendo en cuenta que, como hemos visto, elemento y clase son nociones relativas, que se concretan en cada nivel y se integran unas en otras. En ese sentido, la *token frequency* es un caso particular de la frecuencia en los textos y la *type frequency* un caso particular, de especial interés en sintaxis, de la frecuencia en el inventario[12].

4. Sobre los efectos de la frecuencia en sintaxis

En mi opinión, hay dos grandes esferas en las que debemos estudiar el efecto de la consideración de la frecuencia en los fenómenos gramaticales. Por una parte, existe la posibilidad de valorar si, en la línea de la afirmación de Bybee citada al comienzo, los elementos de alta frecuencia se comportan de modo distinto al que caracteriza a los que la tienen media o baja y, en caso positivo, en qué consiste y cómo se puede explicar esa diferencia. En segundo lugar está la utilización de la frecuencia, es decir, de las diferencias en la frecuencia relativa, como uno de los aspectos más reveladores del cambio lingüístico. Dedicaré esta última sección a mostrar algunos resultados generales en cada uno de estos aspectos.

Veamos la primera de ellas. La frecuencia de los elementos ha sido reiteradamente vinculada a fenómenos como la irregularidad y la analogía en el sentido de que las formas más frecuentes en una lengua son irregulares en una proporción muy alta y, al tiempo, son resistentes a los procesos analógicos, que suelen afectar más a las formas de frecuencia media o baja. Por otro lado, las formas más frecuentes de una lengua son más cortas, suelen ser más antiguas y presentan una alta tasa de elementos gramaticales. Por último, los planteamientos cognitivistas han destacado el vínculo de la frecuencia con la consolidación (*entrechment* es el término original propuesto por Langacker y utilizado también por

[12] En al menos un caso, Bybee 2003, p. 338 equipara *type frequency* con *dictionary frequency* para referirse a la de «a particular pattern, such as a stress pattern, an affix, and so on». El ejemplo que aporta, utilizado en varias ocasiones, se refiere al número de verbos que hacen el pasado con el tipo *damage/damaged* y los que lo hacen con el tipo *break/broke*. Aunque se puede entender, esa equiparación no funciona bien con otras posibilidades como, por ejemplo, el número de verbos que pueden entrar en la perífrasis *be going to + infinitivo*, ya que ahí la referencia al diccionario resulta poco comprensible.

Bybee y otros autores) (cf. Krug 2003, Diessel 2007)[13]. Es tentador tratar de establecer direccionalidad y vínculos causales entre la frecuencia y esos otros fenómenos, pero aquí nos basta con señalar la relación existente entre ellos. En el bien conocido trabajo de Bybee y Thompson 1997 sobre los efectos de la frecuencia en sintaxis se destacan tres: el efecto reductor, que en sintaxis explica los acortamientos y la pérdida de estructura interna; el efecto conservador, con el que cabe explicar, por ejemplo, la especial persistencia del subjuntivo en el francés canadiense con ciertos verbos muy frecuentes; por último, la *type frequency* puede dar cuenta de fenómenos como, por ejemplo, la conservación y extensión de la construcción ditransitiva en inglés.

A todo ello, que habrá que analizar con más profundidad y tratar de aplicar al español, me gustaría añadir otro factor que me parece relevante y que surge de la puesta en relación de dos aspectos bien conocidos. Por una parte, parece estar claro que los elementos más frecuentes en cualquiera de los componentes de una lengua tienen un comportamiento peculiar, característico. Por otra, es también evidente que los textos lingüísticos, como casi todo lo que puede ser contado, presentan una estructura estadística en la que unos pocos elementos suponen un alto porcentaje de lo que encontramos. Es sabido que con los diez elementos más frecuentes en los textos de una lengua (formas ortográficas, lemas, lemas de una cierta clase, etc., se puede dar cuenta de un porcentaje situado en torno al 25% de cuanto hay en los textos y que con unos 150 elementos se llega a una tasa próxima al 50%. Con estadísticas bastante menos conocidas, pero probablemente más interesantes, podemos añadir que de los 3437 verbos documentados en la *BDS*, los 32 más frecuentes (es decir, algo menos del 1% de los registrados) suponen un porcentaje conjunto ligeramente superior al 50% de todas las cláusulas analizadas, cifra especialmente importante si se piensa que los datos anteriores no incluyen los usos de los verbos como auxiliares de perífrasis[14].

La consecuencia de unir estos dos aspectos parece clara: la visión de cuáles son las características básicas de una lengua en un cierto aspecto

---

[13] Diessel 2007, pp. 123-124 hace un clarificador resumen de los efectos de la frecuencia (no solo en sintaxis):

> «First, the strengthening of linguistic representations. Frequency of use reinforces the representation of linguistic expresions in memory, which in turn influences their activation and interpretation in language use.
>
> »Second, the strengthening of linguistic expectations. Since linguistic expressions are arranged in recurrent orders, the language user develops expectations as to which linguistic expressions may occur after a particular word or a particular category, which influences the comprehension and production of linguistic units and can give rise to diachronic change.
>
> »Third, the development of automatized chunks. Linguistic expressions that are frequently combined may become automatized, i.e. they may develop into a processing unit in which the boundaries between linguistic elements are blurred and the whole chunk is compressed and reduced.»

[14] Cf. Rojo 2008 para más detalles sobre estos aspectos de la estructura estadística de los textos.

está determinada en buena parte por las que muestren los elementos más frecuentes de la zona estudiada, que pueden ser diferentes o incluso muy diferentes de las que realmente posea la totalidad del conjunto correspondiente. Dicho de otro modo, la generalización de lo obtenido en el análisis de los elementos filtrados para la confección de un diccionario de frecuencias léxicas resulta habitualmente inadecuada, puesto que las características de ese subconjunto pueden diferir considerablemente de las que muestra el conjunto al cual pertenecen.

Tenemos una buena prueba de todo ello en lo que sucede con el caso de la distribución de los verbos en las tres conjugaciones, al que ya he hecho alusión. Hace unos cuantos años, me sorprendió la más que notable diferencia existente entre los datos que, en un paciente y concienzudo trabajo sobre los que se pueden extraer del *FDSW*, había obtenido Dolores Corbella 1987 y los que yo encontraba al analizar el corpus que constituye la *BDS*. Los datos, tanto de las frecuencias de inventario como de las frecuencias de uso, son los que aparecen en el cuadro siguiente[15]:

	Porcentaje de verbos en el inventario		Porcentaje de uso en el corpus	
	FDSW	BDS	FDSW	BDS
-ar	68,55%	81,46%	37,59%	45,94%
-er	15,57%	8,61%	45,78%	37,29%
-ir	15,88%	9,92%	16,63%	16,77%
Totales	100% N = 957	99,99% N = 3437	100% N = 73.902	100% N = 191.701

CUADRO 4: Porcentaje de verbos en el inventario y en el corpus sobre los datos del *FDSW* (según Corbella 1987) y la *BDS*.

Elaboración propia.

Sorprende, en primer lugar, la diferencia en el número de verbos registrados en cada caso (algo menos de 1000 frente a unos 3450). El tamaño de los corpus utilizados (500.000 formas en el *FDSW* frente a 1,5 millones en la *BDS*) puede explicar una parte de esa diferencia, pero solo una parte. La causa real, a la que nunca se ha prestado atención, radica en el hecho de que Juilland y Chang-Rodríguez, en una línea de actuación seguida reiteradamente por los responsables de diccionarios de frecuencias, decidieron no publicar los datos de los aproximadamen-

[15] Los correspondientes a la *BDS* figuran también en los cuadros 2 (frecuencia de inventario) y 3 (frecuencia de uso).

te 25.000 lemas obtenidos en el análisis del corpus, sino únicamente los correspondientes a los 5024 que alcanzaron los índices más altos en frecuencia, dispersión y uso[16]. En otras palabras, los datos publicados (que son los que pudo analizar Corbella)[17] corresponden únicamente al 25% de los lemas documentados en los textos.

Ese filtro (comprensible en una obra de este tipo destinada a ser publicada en papel) explica también todo lo demás. Hay una evidente diferencia en la frecuencia del inventario de verbos correspondientes a las tres conjugaciones en los dos corpus. En ambos casos predominan los verbos de la primera conjugación, pero en el *FDSW* no llegan al 70%, mientras que rebasan el 80% en la *BDS*. Mucho más llamativa es la discrepancia que se ve en los porcentajes correspondientes al uso. Incluso se da el cambio de la conjugación más usada: la primera según la *BDS* y la segunda según el *FDSW*. Curiosamente, los porcentajes cruzados son muy semejantes.

Es claro que la configuración de esta zona de la gramática del español cambia de forma sorprendente según se manejen los datos del *FDSW* o de la *BDS*. Los datos son reales en ambos casos, pero en uno de ellos se ha aplicado un filtro que, de forma realmente inesperada, incluso para los estudiosos de las frecuencias, cambia radicalmente el panorama. La posibilidad de obtener datos de la *BDS* según diferentes tramos de frecuencias muestra el fenómeno con toda claridad. El cuadro siguiente muestra la comparación de los datos del *FDSW* con los 1000 verbos más frecuentes de la *BDS*:

	BDS	FDSW
-ar	69,5%	68,55%
-er	15%	15,57%
-ir	15,5%	15,98%
Totales	100% (N = 1000)	100% (N = 957)

CUADRO 5: 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.

[16] El análisis de los textos produjo unos 25.000 lemas diferentes que pasaron a 20.000 al eliminar extranjerismos y nombres propios. Quedaron unos 14.000 al ser eliminados todos los que tenían frecuencia inferior a 4,9000 al dejar solo los que estaban presentes en, al menos, tres de los cinco «mundos» y, finalmente, 5024 al exigir un índice de uso igual o superior a 3,08. Para detalles, véase Juilland y Chang-Rodríguez 1964, pp. LXXIV-LXXVI.

[17] Y Guiter 1971. Para más detalles, v. Rojo 2006.

La proximidad de los datos es realmente llamativa, a pesar de que los verbos concretos implicados en cada conjunto difieren aproximadamente en un 20% (cf. Rojo 2006, p. 321 para más detalles). El modo en que evoluciona la configuración aparece en los cuadros siguientes, en los que doy los datos que corresponden a diferentes cortes de frecuencia en los verbos de la *BDS* tanto en la frecuencia de inventario como en la frecuencia de uso:

	100 más frecuentes	250 más frecuentes	500 más frecuentes	1000 más frecuentes	2000 más frecuentes	*BDS* completa
-ar	46%	58%	63,8%	69,5%	75,90%	81,46%
-er	28%	22,8%	19,2%	15%	11,25%	8,61%
-ir	26%	19,2%	17%	15,5%	12,85%	9,92%
Totales	100%	100%	100%	100%	100%	99,99%

CUADRO 6: Distribución porcentual 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% (N=124.251)	100% (N=149.496)	100% (N=168.085)	100% (N=181.523)	100% (N=189.230)	100% (N=191.701)

CUADRO 7: Distribución porcentual del uso de los verbos de las tres conjugaciones en diferentes conjuntos de frecuencias de la *BDS*.

Fuente: *BDS*. Elaboración propia.

No será fácil encontrar un fenómeno en el cual la configuración obtenida vaya cambiando de forma tan constante y regular a medida que se amplía el rango de frecuencia de los elementos considerados. En la frecuencia de inventario, la primera conjugación es siempre la más nutrida, pero su porcentaje asciende con la ampliación de los elemen-

tos y las otras dos se van reduciendo de forma similar, de modo que se mantienen en unos pesos bastante parecidos entre sí. Más curiosa es la evolución en el uso. Con los 100 primeros verbos, la más usada es la segunda conjugación, que va descendiendo de forma constante con la consideración de frecuencias cada vez más amplias[18]. En contrapartida, la primera va aumentando, se iguala con la segunda al llegar a los 500 verbos más frecuentes y ocupa sistemáticamente la primera posición a partir de ese tramo. Esta es la explicación de los diferentes resultados que se han obtenido en diferentes estudios realizados. La tercera conjugación, en cambio, se mantiene en una frecuencia de uso parecida en todos los tramos de frecuencia considerados.

Las conclusiones me parecen claras: las relaciones entre las tres clases morfológicas del español tienen una configuración diferente según hagamos una consideración total, que comprende todos los verbos existentes en un corpus representativo, o bien tomemos en cuenta únicamente el subconjunto de los más frecuentes (incluso los 1000 más frecuentes). Por otro lado, la comprensión adecuada de lo que sucede solo se puede alcanzar si manejamos una visión de la frecuencia que distinga frecuencia de inventario y frecuencia de uso, al tiempo que aplique ambos conceptos a elementos y clases gramaticales de los más diversos niveles. La unión de todos estos factores nos da, en una zona totalmente distinta, la causa de la extrañeza que nos produce el cuadro que resulta del estudio de Patterson y Urrutibéheity 1975 sobre la estructura del léxico español. Al trabajar sobre los lemas del *FDSW*, obtienen una imagen que resulta absolutamente dispar de la que podemos obtener por otras vías o, simplemente, intuimos: que la primera mitad del siglo XX haya aportado únicamente el 0,3% de los lemas choca de frente con lo que sabemos, pero es un dato perfectamente integrable si, además de tener en cuenta la época a la que pertenecen los textos tomados en consideración, sabemos que la distribución es válida únicamente para los 5024 lemas con los índices de frecuencias, dispersión y uso más elevados (en el recuento del *FDSW*)[19].

---

[18] Recuérdese lo ya indicado acerca de la no consideración de los usos auxiliares de los verbos.

[19] Los autores son perfectamente conscientes de esa distorsión y defienden el enfoque adoptado para intentar reflejar la estructura del léxico español: «The Spanish vocabulary is represented by its manifestations in modern Spanish texts rather than by the inventory listed in a modern dictionary. In other words, we aim to account for, and make predictions about certain fundamental properties of Modern Spanish texts rather than dictionaries. While a standard dictionary lists some 100.000 entries, the first 100 most frequently used words account for more than 30% of the lexical materials that constitute any Spanish text; the first 1000 account for more than 50%; the first 5000 for more than 90%. In other words, the occurrences of most entries registered in a Spanish dictionary account for only a minute part of representative Spanish texts. Therefore, instead of basing our conclusions on a study of all lexical items, which would give a distorted picture of lexical structure by placing on the same footing the many words

	Elementos	Porcentaje	Porcentaje de uso
Orígenes	1176	23,50%	81,31%
Siglo X	92	1,88%	0,39%
Siglo XI	38	0,74%	0,27%
Siglo XII	220	4,40%	2,46%
Siglo XIII	783	15,66%	4,56%
Siglo XIV	221	4,42%	1,14%
Siglo XV	1164	23,28%	5,12%
Siglo XVI	475	9,50%	1,69%
Siglo XVII	403	8,06%	1,58%
Siglo XVIII	240	4,80%	0,97%
Siglo XIX	173	3,46%	0,47%
Siglo XX	15	0,30%	0,05%

CUADRO 8.

Fuente: Patterson y Urrutibéheity 1975, p. 38, tabla 14.

La importancia de la distinción entre frecuencia de inventario y frecuencia de uso se aprecia con gran claridad en el caso de los esquemas sintácticos, cuestión a la que he dedicado ya cierta atención (cf. Rojo 2003). En este punto, a los problemas metodológicos comunes al estudio de todas las frecuencias gramaticales se unen las dificultades especiales para encontrar corpus de cierta amplitud que tengan el detalle y la finura de anotación suficientes como para poder localizar con rapidez los esquemas en que se localiza un verbo, la frecuencia general y relativa con que lo hace en cada uno de ellos, los verbos que entran en un cierto esquema sintáctico, los que lo hacen con un porcentaje mínimo, los verbos que pueden alternar dos esquemas sintácticos con un cierto grado de frecuencia, etcétera.

Las dificultades de ambos tipos explican, me parece, la escasez de estudios sobre este punto. Y no se trata, como ocurre en tantos otros casos, de algo que suceda en español y no se dé en otras lenguas, porque aquí la situación es más o menos la misma para todas ellas. Afortunadamente, para el español disponemos de la *Base de datos sintácticos del*

---

likely to occur once every 1000 pages with the very few likely to occur ten times per page, it is preferable to consider only a few thousand words, taken from the top of a frequency hierarchy» (Patterson y Urrutibéheity 1975, p. 9).

español (BDS), que está a disposición de todos los interesados desde hace ya varios años <http://www.bds.usc.es>[20].

La primera cuestión, planteada en muy pocas ocasiones, consiste en saber cuántos esquemas sintácticos se pueden documentar en textos españoles contemporáneos y cuál es la frecuencia relativa de cada uno de ellos. En la organización de la *BDS*, un esquema sintáctico es una estructura abstracta formada por un tipo general de construcción (activa, media, pasiva perifrástica o pasiva pronominal) y una secuencia de funciones sintácticas argumentales. Aunque no puedo detenerme aquí en este punto, para que sea posible valorar las cifras que doy a continuación debo indicar que el elenco de funciones argumentales es, en la *BDS*, bastante elevado, porque tiene en cuenta diferencias con las que no todo el mundo está de acuerdo. Por citar el caso más claro, se distingue entre complemento de régimen (el suplemento de Alarcos en su sentido más estricto) y complementos adverbiales, lo cual duplica el número de esquemas que presentan una de estas dos funciones con respecto a aproximaciones que no establezcan esta distinción. Pues bien, teniendo eso en cuenta, en la *BDS* hemos documentado 158 esquemas, pero solo 15 de ellos (esto es, el 9,49%) tiene una frecuencia de uso igual o superior al 1%. La distribución general puede observarse en el cuadro siguiente:

Esquemas con	Frecuencia	% sobre total de esquemas	Porcentaje acumulado
F > = 1588	15	9,49%	9,49%
1588 > F > = 159	20	12,66%	22,15%
159 > F > = 100	7	4,43%	26,58%
100 > F > = 50	16	10,13%	36,71%
50 > F > = 25	7	4,43%	41,14%
25 > F > = 10	17	10,76%	51,90%
10 > F > = 5	11	6,96%	58,86%
5 > F > = 2	36	22,78%	81,64%
F = 1	29	18,35%	99,99%
Totales	158	99,99%	

CUADRO 9: Distribución de los esquemas sintácticos según su frecuencia de aparición en las 158.769 cláusulas contenidas en la *BDS*.

Fuente: *BDS*. Elaboración propia.

[20] En la página indicada está toda la información relevante, de modo que será suficiente con indicar aquí que es una base de datos que contiene el análisis manual de unas 160.000 cláusulas con la indicación detallada de la clase a la que pertenecen, la función que desempeñan, el verbo que funciona como predicado, la voz, los elementos argumentales que las componen y el tipo al que pertenece la secuencia que realiza esa función.

El cuadro muestra la apariencia típica en la distribución de las frecuencias: unos pocos elementos con frecuencia muy elevada y muchos elementos con frecuencias bajas o muy bajas. En números redondos, el 75% de los esquemas documentados presentan menos de 100 casos en un corpus de casi 160.000 cláusulas analizadas. Si, para no perdernos en un número excesivo de esquemas, nos quedamos únicamente con aquellos que tienen una frecuencia relativa igual o superior al 0,1% del total (es decir, 159 ejemplos o más en toda la *BDS*), obtenemos los resultados que pueden apreciarse en el cuadro siguiente (10). En conjunto, las cláusulas que presentan alguno de estos esquemas suponen el 98,26% del total de la *BDS*, de modo que los 122 restantes significan únicamente el 1,74% de las cláusulas analizadas[21].

Por supuesto, la distribución reflejada en el cuadro no descubre aspectos inesperados, pero permite cuantificar –y, por tanto, abre la posibilidad de confirmar o rechazar– las intuiciones que pudiéramos tener al respecto. Como no tiene sentido volver aquí sobre los esquemas en sí (cf. Rojo 2003), me limitaré a indicar que los datos del cuadro reafirman el interés de la distinción que he propuesto. Para el estudio del papel de los esquemas sintácticos importa tanto su frecuencia de inventario, estrictamente equivalente en este caso a la *type frequency* de Bybee, es decir, el número de verbos en los que se ha documentado, como su frecuencia de uso en los textos examinados. En la frecuencia de uso de un esquema influye el número de verbos que lo presenta, la frecuencia de uso de cada verbo y el peso relativo del esquema en cuestión en cada uno de los verbos en que aparece.

El interés de la distinción general se observa con mucha claridad en, por ejemplo, el esquema «activa con sujeto y predicativo de sujeto». Es el segundo esquema en frecuencia de uso y aparece en el 15,18% de las cláusulas de la *BDS* (frecuencia de uso corregida, v. supra), pero solo

---

[21] Este cuadro muestra ciertas diferencias con el que aparece en Rojo 2003, p. 419. Las discrepancias proceden del hecho de que en 2003 hice los recuentos de las cláusulas efectivamente incluidas en la *BDS*. Sin embargo, ahora he tenido en cuenta que el fichado manual de los textos no incluyó todos los casos de verbos muy frecuentes y con un abanico de esquemas relativamente reducido, de modo que he recalculado las frecuencias y los porcentajes correspondientes a cada esquema añadiendo a los casos efectivamente incluidos los resultados del cálculo adicional de los que teóricamente habrían aparecido si el proceso de fichado hubiera sido exhaustivo. Como aumenta el total de las cláusulas tomadas en consideración, los porcentajes descienden con respecto a los calculados en 2003, a excepción del esquema formado por sujeto y predicativo de sujeto en voz activa. El cambio, perfectamente explicable por otro lado, produce una reconfiguración de cierta importancia, que hace que este esquema, que antes aparecía en la tercera posición, pase ahora a la segunda. Hay algunos otros casos, de menor importancia, en los que se dan cambios en el rango del esquema.

Las claves utilizadas son: S = sujeto, D = comp. directo, I = comp. indirecto, SP = suplemento, AD = comp. adverbial, MD = comp. modal, PR = otro comp. preposicional argumental, A = agente, PS = predicativo de sujeto, PD = predicativo de comp. directo, PO = predicativo de otros complementos.

Construcción	Funciones	Porcentaje sobre el total de las cláusulas	Número de verbos que presentan el esquema	Porcentaje sobre el total de los verbos
Activa	S, D	34,95%	2421	70,44%
Activa	S, PS	15,18%	63	1,83%
Activa	S	11,43%	1176	34,22%
Activa	S, D, I	5,54%	624	18,16%
Activa	S, AD	4,19%	179	5,21%
Media	S	3,35%	816	23,74%
Activa	S, I	3,11%	222	6,46%
Activa	S, SP	2,65%	321	9,34%
Media	S, SP	2,2%	370	10,77%
Activa	S, D, PD	2,16%	95	2,76%
Activa	D	1,86%	3	0,09%
Activa	S, D, AD	1,68%	197	5,73%
Media	S, AD	1,42%	171	4,98%
Pas. ref.	S	1,41%	546	15,89%
Media	S, PS	1,13%	46	1,34%
Activa	S, D, SP	1,03%	289	8,41%
Activa	S, PR	0,79%	110	3,20%
Media	S, D	0,67%	130	3,78%
Pas. perif.	S, A	0,55%	473	13,76%
Activa	S, D, PR	0,45%	139	4,04%
Activa	S, I, PS	0,39%	21	0,61%
Media	S, I	0,32%	134	3,90%
Pas. perif.	S	0,29%	324	9,43%
Activa	S, D, PS	0,27%	41	1,19%
Media	S, PR	0,19%	84	2,44%
Activa	S, D, I, AD	0,15%	35	1,02%
Activa	S, I, SP	0,15%	27	0,79%
Pas. ref.	S, I	0,14%	90	2,62%
Activa	SP	0,13%	4	0,12%
Pas. ref.	S, PS	0,13%	58	1,69%
Pas. perif.	S, SP	0,13%	76	2,21%
Pas. ref.	S, SP	0,12%	65	1,89%
Media	S, D, SP	0,11%	6	0,17%
Activa	S, AD, PS	0,11%	13	0,38%

CUADRO 10: Distribución de los esquemas que poseen frecuencia igual o superior al 0,1% en la *BDS*.

Fuente: *BDS*. Elaboración propia.

se documenta en el 1,83% de los verbos (frecuencia en el inventario). Por otro lado, un análisis más detallado, que solo puedo insinuar en esta ocasión, muestra que este último porcentaje es un tanto engañoso: con los datos corregidos, el 93,09% de los casos de este esquema pertenecen únicamente a tres verbos (*ser, estar* y *resultar*, naturalmente). Los 60 verbos restantes que también presentan el esquema no pasan de un 7% en conjunto.

El dato anterior pone de relieve una cierta inconsistencia de la frecuencia de inventario cuando es utilizada sin relación con la frecuencia de uso y su distribución. En efecto, si lo que cuenta es que un verbo presente algún ejemplo en un cierto esquema, perdemos de vista la importantísima diferencia que existe entre la utilización de un esquema como el preferido de un verbo, su utilización en un porcentaje importante, y su uso marginal. Todos esos casos cuentan y pesan del mismo modo si no introducimos un filtro adicional, con unos mínimos cuantitativos que habrá que establecer y ajustar en cada caso. El cuadro siguiente (11) muestra la configuración que tienen los diez esquemas sintácticos mayoritarios del español con la exigencia de tener en cuenta únicamente aquellos verbos que presenten este esquema en más del 50% de los casos documentados. Además, para evitar la distorsión que supondrían los verbos de baja frecuencia, añado el requisito de un mínimo de 20 casos para la totalidad del verbo, esto es, una frecuencia relativa de aproximadamente 13 casos por millón de formas[22].

Los datos son, me parece, del mayor interés. Para destacar solo el más llamativo: el 35% de los verbos que tienen una frecuencia de uso que podemos considerar alta (estimada aquí como de 13 casos o más por millón de palabras) tienen el esquema «construcción activa + sujeto + predicado + complemento directo» como esquema básico de uso, estimado aquí como equivalente a aparición superior al 50% de los casos registrados. Lo que me interesa destacar especialmente es que esa visión de la relación entre verbos y esquemas surge solo si tenemos en cuenta la frecuencia del esquema sintáctico, el número de verbos que lo admite, la frecuencia de cada verbo y la frecuencia conjunta de todos aquellos que admiten ese esquema.

Son numerosos los trabajos que en estos últimos años han tratado de poner de relieve las relaciones existentes entre el aumento de la frecuencia de los elementos o construcciones y el avance de los procesos de gramaticalización. El tema es largo y complejo, dista de estar resuelto en términos estrictamente teóricos y necesita además de un número suficiente de estudios específicos, que, por otro lado, solo pueden hacerse con mucho trabajo adicional sobre corpus amplios, bien construi-

[22] Ese corte es equivalente a trabajar únicamente con los verbos que figuran entre los 4500 o 5000 lemas más frecuentes del español.

Construcción	Funciones	Número de verbos con F > = 20 que presentan más del 50 % de los ejemplos en este esquema	Porcentaje sobre los verbos que se documentan en este esquema	Porcentaje sobre los 864 verbos que en la BDS tienen frecuencia igual o superior a 20
Activa	S, D	375	15,48%	34,84%
Activa	S, PS	4	6,34%	0,35%
Activa	S	73	6,2%	7,52%
Activa	S, D, I	19	3,04%	1,74%
Activa	S, AD	15	8,37%	1,39%
Media	S	19	2,32%	1,50%
Activa	S, I	8	3,6%	0,81%
Activa	S, SP	29	9,03%	2,78%
Media	S, SP	31	8,37%	2,55%
Activa	S, D, PD	2	2,1%	0,12%

CUADRO 11: Número de verbos con frecuencia absoluta en la *BDS* superior a 20 que presentan en un esquema determinado más del 50% de sus casos.

dos y anotados. Aun con esas precauciones, parece claro que, en la mayor parte de los casos, el avance en el proceso de gramaticalización de una construcción debería ir acompañado del aumento en la frecuencia de uso y, sobre todo, del incremento de su frecuencia de inventario.

En esta línea, presento a continuación unos datos, superficiales, pero creo que suficientemente ilustrativos, de lo que sucede en español con la construcción *ir a* + infinitivo, que, como es bien sabido, evoluciona desde la fase en que es una construcción sintáctica formada por un verbo de movimiento seguido de la preposición *a* y un verbo en infinitivo que expresa el objetivo con que se realiza ese movimiento (*fui a buscar el libro*, por ejemplo, paralelo a *fui a la biblioteca a recoger el libro*) hasta convertirse en una construcción perifrástica plena con valor de posterioridad, pasando por una situación intermedia de expresión de la intencionalidad (cf. Rojo 1974). Un proceso de este tipo, con todo el complejo fenómeno de la expresión de la futuridad en las len-

guas románicas (y de otras familias) al fondo, debe tener reflejo cuantitativo en el aumento de la frecuencia de la construcción. Al tiempo, a medida que va perdiendo su valor estricto de movimiento, se ensancha la gama de los sujetos posibles y también de los verbos en infinitivo que pueden formar parte de la construcción (que son los que pasan a establecer las restricciones sobre el sujeto). Ejemplos del tipo *ir a ser*, *ir a haber*, *ir a llover*, *ir a ir*, etc., no pueden darse mientras el verbo *ir* tenga valor estricto de verbo de movimiento. En consecuencia, el aumento de la frecuencia de inventario de la construcción y, sobre todo, de verbos cuyo carácter no los hace fácilmente compatibles con expresión de movimiento será un indicio claro de la gramaticalización de la construcción, de la consolidación de una perífrasis verbal en el sentido más estricto posible.

El estudio de *ir a* + infinitivo en el *CORDE* tropieza con la imposibilidad de hacer consultas por lemas, de modo que, para poder presentar una visión provisional de lo que sucede en este punto, he recurrido a un procedimiento que combina la extracción de datos correspondientes a diferentes épocas usando la aplicación de la RAE con la obtención de concordancias para los casos de *va/iba a* + **[aei]r* y el posterior refinamiento de los datos[23]. La flexibilidad del sistema de búsqueda del *CORDE* me ha permitido, en cambio, recoger los datos correspondientes a tramos temporales de diferente extensión (más reducidos a medida que se acercan a la época actual), diferentes tipos de texto, etc. Algo parecido he hecho con el *CREA*, con lo que disponemos de una primera panorámica, sin duda muy parcial e insuficiente todavía, de la evolución de esta construcción a lo largo de toda la historia del español. Los resultados, que aparecen en el cuadro siguiente, reflejan, como es obligado, la frecuencia normalizada (por millón de formas ortográficas).

---

[23] El *CORDE* no está lematizado, como es bien sabido, de modo que he recurrido a la obtención de los casos de *va* e *iba* seguidos de *a*. La aplicación de la RAE no admite en este caso búsquedas que las combinen con expresiones como **r*, **[aei]r* o similares para captar los infinitivos, puesto que le resulta excesivamente compleja. En consecuencia, he exportado los datos obtenidos con la búsqueda más general (sorteando los límites en el número de casos que devuelve mediante la selección por tipos de texto, años, países, etc., según lo más adecuado en cada caso). Del conjunto de los textos resultantes he obtenido las concordancias correspondientes a *va / iba a *[aei]r* con el módulo de concordancias de WordSmith Tools. A partir de esas concordancias, he revisado y limpiado los casos obtenidos (eliminación de *va a Agadir*, etc.), aislado las construcciones y reducido las diferencias a esquemas del tipo *ir a comprar*, *ir a ser*, *ir a decir*, etc., mediante una serie de rutinas escritas en Perl. Por tanto, en los datos que siguen no hay análisis del carácter realmente perifrástico o no de las construcciones aisladas, solo aparecen los casos con *va* o *iba* como forma auxiliar y es muy probable que se hayan deslizado algunos errores tanto en la recogida como en el tratamiento de los datos. De todas formas, creo que proporcionan una visión general adecuada de la evolución de esta construcción. Cf. Schulte 2009 para una exposición clara de la necesidad de combinar procedimientos en el análisis de los corpus diacrónicos y Smith, Hoffmann y Rayson 2008 para la conveniencia de la anotación manual de los datos obtenidos en los corpus.

	Frec. normalizada (casos por millón de formas)	Núm. de verbos distintos
1100-1250	29,61%	2
1251-1300	2,52%	5
1351-1400	2,09%	5
1451-1500	6,28%	35
1551-1600	17,89%	99
1651-1700	20,57%	46
1701-1750	14,18%	54
1751-1800	20,44%	72
1801-1850	71,64%	208
1851-1900	88,56%	292
1901-1950	153,44%	532
1951-1974	288,64%	580
1975-2004 *(CREA)*	279,77%	1273

CUADRO 12: Frecuencia de las construcciones *va / iba a* + infinitivo en distintos tramos temporales del *CORDE* y el *CREA*.

Fuente: *CORDE* y *CREA* <http://www.rae.es>. Elaboración propia.

Para la correcta interpretación de los datos debe tenerse en cuenta que solo están las construcciones con *va* e *iba* y también que se ha hecho una búsqueda puramente formal, de modo que no están diferenciados los casos que realmente habría que considerar perifrásticos de todos los demás. En algún tramo, esta insuficiencia, que solo se supera con el análisis individual, provoca una visión distorsionada. El más claro es, sin duda, el primero, el de los documentos anteriores a 1250. Los 66 casos que aparecen en ese segmento proceden todos del mismo texto, con tan solo dos verbos distintos en infinitivo, 65 de ellos son del tipo *ir a dar* y se refieren a ríos, caminos, etc. Es, pues, una cifra engañosa que un análisis profundo debería reducir a cero casos de construcción perifrástica. No obstante, y siempre teniendo en cuenta que se trata de datos muy provisionales, que tendrán que ser completados y depurados, se observa una línea de incremento muy clara, con inflexiones marcadas en la segunda mitad del XVI y, sobre todo, en la primera parte del XIX. Si se prescinde de los primeros tramos, la línea que se mues-

tra en el gráfico (1) se aproxima bastante a la S típica de las frecuencias en los procesos de cambio lingüístico, con las frecuencias de nuevo estabilizadas en el tramo final (cf. Raumolin-Brunberg 2003).

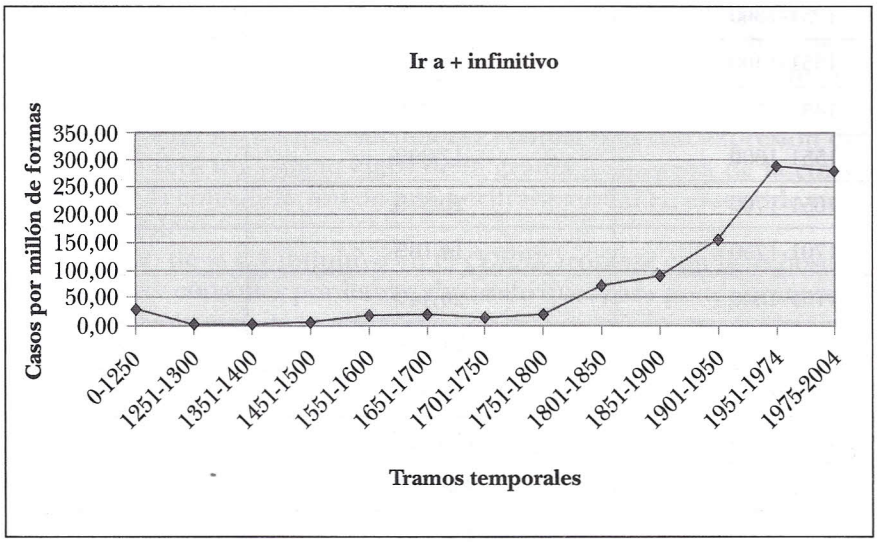

GRÁFICO 1: Frecuencia de la construcción *va / iba a* + infinitivo en *CORDE* y *CREA*

Fuente: *CORDE* y *CREA* <http://www.rae.es>. Elaboración propia.

Aunque se trata de resultados parciales obtenidos únicamente sobre dos formas de *ir* y sin eliminar los casos en los que no hay realmente construcción perifrástica, el panorama que dibujan los datos del cuadro (12) y el gráfico (1) parecen plenamente fiables. Una línea similar, pero mucho menos detallada, se puede obtener de la consulta del *Corpus del español*, construido por Mark Davies <http://www.corpusdelespanol.org>. Este corpus está parcialmente lematizado, de modo que es posible hacer una consulta referida a todos los casos de *ir a* + infinitivo. La respuesta facilita las frecuencias absolutas y normalizadas de la construcción segmentadas por siglos, que son las que figuran en el cuadro (13)[24].

[24] La lematización parcial constituye una ventaja innegable, pero se ve luego fuertemente disminuida por la rigidez en el establecimiento de los tramos temporales (solo admite segmentación por siglos) y la imposibilidad de obtener datos diferenciados por países, áreas temáticas o tipos de texto. Trabajar con la estructuración en siglos parece poco adecuado para captar las líneas evolutivas de los diferentes fenómenos. Véanse, por ejemplo, las fuertes diferencias que muestran las frecuencias correspondientes a los siglos XVII, XVIII y XIX, en claro contraste con la línea que se observa si los tramos acotados son más reducidos.

Siglos	Frec. normalizada (casos por millón de formas)
XIII	30,08
XIV	51,69
XV	72,54
XVI	245,62
XVII	298,99
XVIII	239,59
XIX	643,36
XX	1041,44

CUADRO 13: Frecuencias relativas de la construcción *ir a* + infinitivo en el *Corpus del español.*

Fuente: *Corpus del español* <http://www.corpusdelespanol.org/> [30/7/2010]. Elaboración propia.

Como era de esperar, puesto que están todos los casos de la construcción y no únicamente los que llevan *va* o *iba* en el auxiliar, los valores normalizados son más altos, pero, con pequeñas diferencias explicables sin duda por el diferente tamaño de los períodos considerados, la línea general es la misma, con los siglos XIX y XX claramente distanciados de los precedentes.

La *BDS,* que solo contiene textos posteriores a 1988, permite en cambio un análisis más seguro, puesto que es posible obtener los casos de construcción realmente perifrástica. Además, es fácil obtener los valores correspondientes a los diferentes tipos de textos representados en este corpus.

BDS	Frec. normalizada (casos por millón de formas)	Núm. de verbos distintos
Ensayo	272	56
Prensa	390	49
Narrativa	1316	284
Teatro	2682	232
Oral	1666	167
Total BDS	1290	499

CUADRO 14: Frecuencia de la perífrasis *ir a* + infinitivo en diferentes tipos de textos de la *BDS.*

Fuente: *BDS* <http://www.bds.usc.es>. Elaboración propia.

Las diferencias son evidentes. Los textos ensayísticos y periodísticos presentan unos valores considerablemente más bajos que los que se encuentran en los orales o de ficción. Una configuración semejante se puede observar en los textos del siglo XX integrados en el *Corpus del español* y también en el *CREA*.

	Frec. relativa (casos por millón de formas)
Académico	47
Prensa	488
Ficción	1473
Oral	2629
Total siglo XX	1157

CUADRO 15: Frecuencias relativas de la construcción *ir a* + infinitivo en diferentes tipos de textos del *Corpus del español* (siglo XX).

Fuente: *Corpus del español* <http://www.corpusdelespanol.org> [30/7/2010]. Elaboración propia.

CREA	Frec. normalizada (casos por millón de formas)
Ensayo (solo hipercampos 1 y 2)	138,92
Prensa (solo periódicos)	196,15
Ficción	460,90
Oral	2601,07

CUADRO 16: Frecuencia de la construcción *va/iba a* + infinitivo en diferentes tipos de texto del *CREA*.

Fuente: *CREA* <http://www.rae.es>. [Julio de 2010]. Elaboración propia.

Pero no se trata solo del incremento en el número de casos. A medida que el proceso de gramaticalización avanza e *ir* pierde su significado inicial de verbo de movimiento, son más los verbos que pueden figurar en esta construcción, como muestra con claridad el cuadro (16). Es decir, aumenta la frecuencia de uso de la construcción y se amplía su frecuencia de inventario (aquí estrictamente equivalente a la *type frequency*). Teniendo en cuenta el escaso tamaño del corpus analizado

para la *BDS*, es significativo que 499 verbos distintos (esto es, el 14,51% de los 3437 verbos documentados) presenten esta construcción.

Por fin, podemos lograr algo parecido a la integración de estas dos consideraciones estudiando los casos de la construcción *ir a ser*, que solo es posible cuando *ir* ha cubierto ya íntegramente o casi íntegramente el camino de su conversión en auxiliar de una perífrasis verbal. El cuadro siguiente muestra los valores obtenidos en los corpus *CORDE* y *CREA* y el *Corpus del español*. De nuevo puede observarse que las cifras son distintas (por las diferencias en lo que reflejan), pero siguen la misma línea: el salto fuerte se da en el siglo XIX y, como de nuevo muestran el último período del *CORDE* y el *CREA*, el proceso se estabiliza en la segunda mitad del XX. Puede añadirse a lo anterior que, con los datos del *CORDE*, la construcción *va/iba a ser* presenta los primeros casos en la segunda mitad del XVI y *ser* es el infinitivo más frecuente en esta construcción a partir del período 1751-1800. Dado que *ser* es el verbo más frecuente, eso es precisamente lo esperado desde el momento en que se ha consolidado el proceso de gramaticalización.

CORDE-CREA	Frec. normalizada (casos por millón de formas)	CORPUS DEL ESPAÑOL	Frec. normalizada (casos por millón de formas)
1100-1250	0,0		
1251-1300	0,0	XIII	0,1
1351-1400	0,0	XIV	
1451-1500	0,0	XV	0,6
1551-1600	0,2	XVI	2,3
1651-1700	0,2	XVII	4,5
1751-1800	1,9	XVIII	4,0
1801-1850	9,1		
1851-1900	16,0	XIX	31,0
1901-1950	23,9		
1951-1974	41,4	XX	75,1
1975-2004 *(CREA)*	41,7		

CUADRO 17: Frecuencia de la construcción *ir a ser* en *CORDE*, *CREA* y *Corpus del español*.

Fuentes: *CORDE* y *CREA* <http://www.rae.es> y *Corpus del español* <http://www.corpusdelespanol.org>. Elaboración propia.

5. Conclusiones

Es preciso reconocer que, por una amplia serie de factores a los que no he podido prestar más que una atención superficial, apenas estamos empezando a comprender las implicaciones que tiene la frecuencia en el comportamiento de los elementos sintácticos y en la configuración que muestran. Es necesario desarrollar todavía mucho trabajo para tratar de establecer las relaciones de la frecuencia con otros factores más allá de la apariencia que proporcionan unos cuantos casos especiales y, sobre todo, intentar reconocer, si los hay, los vínculos causales entre ellos. Por otro lado, el español, de todas las épocas, necesita investigaciones profundas para llegar al conocimiento adecuado de las implicaciones de la frecuencia, tanto en orientaciones sincrónicas como diacrónicas.

A pesar de ello y la prudencia que aconseja el carácter provisional de muchos de los datos manejados, parece claro que hay unos cuantos puntos en los que podemos estar de acuerdo y sentirnos seguros. En primer lugar, la frecuencia de los elementos lingüísticos en general, y de los sintácticos en particular es un aspecto relevante para el funcionamiento de las lenguas y nuestro conocimiento de ellas. Naturalmente, no se trata de la frecuencia sin más ni de la frecuencia de un fenómeno cualquiera, es decir, no del número de palabras que comienzan por un cierto fonema o de apariciones de la secuencia *Vivo en Santiago*, sino de la frecuencia de fenómenos relevantes para la estructura y funcionamiento de la lengua en el módulo correspondiente. Por otro lado, está claro también que hablamos no de frecuencias absolutas, siempre dependientes de la configuración del corpus que podemos manejar, sino de las frecuencias relativas.

La distinción entre frecuencia de elemento *(token frequency)* y frecuencia de tipo *(type frequency)*, propuesta por Bybee y muy extendida en los últimos años, supone un avance importante sobre las visiones tradicionales, muy poco fundamentadas, pero requiere también, para dar todo el juego necesario en gramática, algunos ajustes o, en una formulación más prudente, la seguridad de que es posible entenderla en los sentidos expuestos aquí. Bajo la consideración de *elemento* debería entrar cualquier unidad manejada en el análisis lingüístico, desde un fonema hasta un esquema sintáctico o un tipo de construcción, pasando por un lema, una clase de palabras, una conjugación verbal, etc. Al tiempo, la integración de unas unidades lingüísticas en otras proporciona el marco necesario entre elementos y clases, de modo que podamos ir ascendiendo en la escala y considerar como elemento en un nivel lo que ha sido visto como clase en el nivel inmediatamente inferior. Por otra parte, la comprensión de la distinción entre frecuencia en el inventario y frecuencia en el uso, en el sentido amplio que he propuesto aquí, tiene todavía mayor interés si logramos integrar ambas perspecti-

vas en la línea apuntada al analizar verbos que presentan un cierto esquema en un porcentaje importante, cuya entidad habrá que establecer. Lo aquí apuntado basta, sin embargo, para dejar claro que la simple indicación de cuántos verbos entran en un esquema es insuficiente, si no tenemos también en cuenta la frecuencia de uso de ese verbo en general y la frecuencia de uso de ese verbo con ese esquema.

Con estos ajustes, estrictamente necesarios, la distinción propuesta por Bybee se puede formular ya como la que existe entre frecuencia en el uso y frecuencia en el inventario, tal como ha sido expuesta aquí. Los ejemplos, parciales, que he manejado a lo largo de este trabajo muestran la capacidad y riqueza que, para el mejor conocimiento de la sintaxis de una lengua, tiene la distinción entendida de este modo.

Referencias bibliográficas

Almela Pérez, R., Cantos, R., Sánchez, A., Sarmiento, R. y Almela, M. (2005): *Frecuencias del español: diccionarios y estudios léxicos y morfológicos*, Madrid, Universitas.

Bakker, J. J. M. (1968): «Frequency in usage and in the lexicon», *Lingua* 21, pp. 13-22.

Bybee, J. (2003): «Mechanisms of change in grammaticization. The role of frequency», en Janda, R. y B. (eds.), *Handbook of historical Linguistics*, Oxford, Blackwell, pp. 602-623. (Cito por su reedición en Bybee 2007, pp. 336-357.)

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

— y Thompson, S. (1997): «Three frequency effects in syntax», *Berkeley Linguistics Society* 23, pp. 65-85. (Cito por su reedición en Bybee 2007, pp. 269-278.)

Chomsky, N. A. (1957): *Syntactic structures*, La Haya, Mouton (trad. esp. de Otero, C. P.: *Estructuras sintácticas*, México D.F., Siglo XXI editores).

— (1962): «A transformational approach to syntax», en Hill, A. A. (ed.), *Proceedings of the 3rd Texas Conference on Problems of Linguistic Analysis in English, 1958*, Austin, Univ. of Texas, pp. 124-158.

Corbella, D. (1987): «Algunos datos estadísticos del paradigma verbal español», en AA.VV., *In memoriam Inmaculada Corrales* 1, pp. 145-159, Universidad de La Laguna.

CORDE, Banco de datos en línea / *Corpus diacrónico del español*, <http://www.rae.es> (versión cerrada en abril de 2005).

CORGA_ETQ, Corpus de referencia do galego actual (versión etiquetada en pruebas). En línea: <http://www.corpus.cirp.es/corgaetq/>.

CORGA, Corpus de referencia do galego actual. En línea: <http://www.corpus.cirp.es/corga/>.

Corpus del español (construido por Mark Davies). En línea: <http://www.corpusdelespanol.org>.

CREA, Banco de datos en línea / *Corpus de referencia del español actual*, <http://www.rae.es> (versión cerrada en junio de 2008).

DAVIES, M. (2006): *A Frequency dictionary of Spanish. Core vocabulary for learners*, Nueva York, Routledge.

DIESSEL, H. (2007): «Frequency effects in language acquisition, language use, and diachronic change», *New Ideas in Psychology* 25, pp. 108-127.

ELLIS, N. C. (2002): «Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition», *Studies in Second Language Acquisition* 24, pp. 143-188.

FODOR, J. A. y KATZ, J. J. (eds.) (1964): *The structure of language. Readings in the philosophy of language*, Englewood Cliffs, Prentice-Hall.

GUITER, H. (1971): «Fréquences verbales dans les langues romanes», *RLR* 35, pp. 358-387.

JUILLAND, A. y CHANG-RODRÍGUEZ, E. (1964): *Frequency dictionary of Spanish words*, La Haya, Mouton.

KRUG, M. (2003): «Frequency as a determinant in grammatical variation and change», en Rohdenburg, G. y Mondorf, B. (eds.), *Determinants of grammatical variation in English*. Berlín-Nueva York, Mouton de Gruyter.

LEECH, G. (1991): «The state of the art in corpus linguistics», en Aijmer, K. y Altenberg, B. (eds.), *English corpus linguistics. Studies in honour of Jan Svartvik*, Londres, Longman, pp. 8-29.

MAIR, C. (2004): «Corpus linguistics and grammaticalisation theory. Statistics, frequencies and beyond», en Lindquist, H. y Mair, C. (eds.), *Corpus Approaches to Grammaticalization in English*, Ámsterdam, John Benjamins, pp. 121-150.

MATHESIUS, V. 1929: «La structure phonologique du lexique du tchèque moderne», *TCLP* 1, pp. 69-84.

MCENERY, T. y WILSON, A. (1996 [2001²]): *Corpus linguistics*, Edimburgo, Edinburgh Univ. Press.

PATTERSON, W. y URRUTIBEHEITY, H. (1975): *The lexical structure of Spanish*, La Haya, Mouton.

RAUMOLIN-BRUNBERG, H. (2003): «Temporal aspects of language change: what can we learn from the CEEC», en Wilson, A., Rayson, P. y McEnery, T. (eds.), *Corpus Linguistics by the Lune. A Festschrift for Geoffrey Leech*, Frankfurt, Peter Lang, pp. 139-156.

ROJO, G. (1974): *Perífrasis verbales en el gallego actual*, Universidad de Santiago de Compostela.

— (2003): «La frecuencia de los esquemas sintácticos clausales en español», en Moreno Fernández, F., Gimeno Menéndez, F., Samper, J. A., Gutiérrez Araus, M.ª L., Vaquero, M. y Hernández, C. (coords.), *Lengua, variación y contexto. Estudios dedicados a Humberto López Morales*, Vol. I, Madrid, Arco/Libros, pp. 413-424.

— (2006): «Sobre las frecuencias verbales en español» en Sedano, M., Bolívar, A. y Shiro, M. (comps.), *Haciendo lingüística. Homenaje a Paola Bentivoglio*, Universidad Central de Venezuela, pp. 309-324.

— (2008): «Lingüística de corpus y lingüística del español», *Actas del XV Congreso de la Asociación de Lingüística y Filología de América Latina*, Montevideo, Edición en CD. <http://gramatica.usc.es/~grojo/Publicaciones/Lgca_corpus_lgca_espanol.pdf>.

— (2010): «Aguja de navegar corpus», ponencia plenaria en el XII Congreso de la Sociedad Argentina de Lingüística (Mendoza, 6-9 de abril de 2010). En Castel, V. M. y Cubo de Severino, L. (eds.), *La renovación de la palabra en el bicentenario de la Argentina. Los colores de la mirada lingüística.* Mendoza, Ed. FFyL (Univ. Nacional de Cuyo), pp. 1151-1163.

— (en prensa): «El papel de los corpus en el estudio de la historia del español», *Actas del VIII Congreso internacional de historia de la lengua española, Universidad de Santiago de Compostela (14-18 de septiembre de 2009)*, <http://gramatica.usc.es/~grojo/En_prensa/Rojo_corpus_diacronicos.pdf>.

SCHULTE, K. (2009): «Using non annotated diachronic corpora: benefits, methods and limitations», en Enrique-Arias, A. (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*, Madrid-Frankfurt, Iberoamericana-Vervuert, pp. 167-180.

SMITH, N., HOFFMANN, S. y RAYSON, R. (2008): «Corpus tools and methods, today and tomorrow: Incorporating linguists' manual annotations», *Literary and Linguistic Computing* 23, 2, pp. 163-179.

STEFANOWITSCH, A. (2005): «New York, Dayton (Ohio), and the raw frequency fallacy», *Corpus linguistics and linguistic theory* 1, 2, pp. 295-301.