"C U M B R E":

CORPUS LINGÜÍSTICO DEL ESPAÑOL CONTEMPORÁNEO.
Fundamentos, metodología y aplicaciones de los corpus lingüísticos.

Aquilino Sánchez
editor

Aquilino Sánchez

Autor y Director del Corpus.

Director del área de Análisis Lexicográfico y Enseñanza de lenguas extranjeras.

Ramón Sarmiento

Director del área de Análisis Gramatical.

Pascual Cantos

Director del área de Tratamiento Informático del corpus.

José Simón

Diseño de aplicaciones informáticas

CONTENIDO:

Capítulo I

1. Qué es un corpus lingüístico.

Desde que Chomsky (1957:16) dijera que "la capacidad del hablante nativo para producir y reconocer oraciones gramaticales no se fundamenta sobre cuestiones de aproximación estadística o similares"; o desde que su énfasis en la **intuición** del hablante ahuyentara los intentos de los estructuralistas para basar sus observaciones sobre repertorios lingüísticos (es decir, sobre la **observación de datos lingüísticos tomados del uso**) y tal punto de vista se convirtiese en quasi-dogmático dentro del campo de la lingüística, ha habido importantes innovaciones y han tomado cuerpo nuevos enfoques que han enriquecido el análisis lingüístico.

La anécdota que comentan Biber/Finegan (Ajmer 1991:204), referida a dos importantes lingüistas, a principios de los años sesenta, R. B. Lees y W. Nelson Francis, es también reveladora e ilustrativa de los cambios que se han dado en sólo tres décadas. Cuando Lees preguntó a Nelson Francis sobre sus proyectos futuros y éste le informó que le había sido concedida una ayuda para preparar un corpus computarizado del inglés, la sorprendente respuesta de Lees no se hizo esperar: "¿Y para qué sirve eso?" A la réplica de Nelson Francis, "Para descubrir la realidad de la gramática inglesa", Lees le contestó maravillado: "Esto es una pérdida total de tiempo y de dinero público. Usted es un hablante nativo del inglés: puede producir en sólo diez minutos más ejemplos sobre cualquier punto gramatical del inglés que los que podría encontrar en muchos millones de palabras de textos elegidos al azar". Lo que interesa destacar aquí no es que uno u otro de los interlocutores sea o no acertado en sus juicios, sino la distancia que separa los puntos de vista sobre los que cada uno de los lingüistas se fundamenta en la apreciación de un hecho: Lees no ve ninguna utilidad en la investigación propuesta porque la intuición del hablante nativo es más que suficiente. En realidad -se viene a decir- el hablante nativo de una lengua es como un "corpus viviente", siempre potencialmente productivo e inagotable, mientras que un repertorio de textos, por muy amplio que éste sea, es una recopilación necesariamente reducida a la "actualización" (*performance)* lingüística y, en consecuencia, **cerrada**. Francis, por el contrario, considera que la gramática inglesa precisa de una base documental que la avale. Si bien es verdad que los hablantes nativos son "potencialmente" susceptibles de generar un sinfín de "producciones lingüísticas", se dan circunstancias que limitan esa producción individualizada, especialmente si la comparamos con el conjunto de la producción lingüística realmente generada por todos los hablantes de una lengua. El corpus, con la ayuda del ordenador, tiene precisamente la posibilidad de superar muchas de esas limitaciones, no solamente en amplitud y en aumento de representatividad, sino en la posibilidad de ofrecer acceso selectivo y abundante a aquellos ámbitos o elementos que deseemos o precisemos en un momento determinado.

No parece que deba darse, por tanto, oposición entre el cometido del hablante nativo y el corpus. A no ser que en el transfondo de todo ello alguien pueda percibir la eterna polémica entre quienes consideran que la lengua debe ser "como los lingüistas y gramáticos dicen que debe ser" y aquellos otros que afirman que la lengua es un contrato entre los hablantes de un mismo idioma y, por ende, es el uso que los hablantes hacen de ella lo que determina la adecuación o corrección del sistema comunicativo utilizado. En el primer caso, la necesidad de un corpus lingüístico es ociosa y hasta puede ser "perjudicial", ya que la norma la determinan unos pocos hablantes selectos y son únicamente éstos quienes deben ser propuestos como modelos. En el segundo caso, por el contrario, el corpus lingüístico es un utensilio imprescindible para llevar a cabo un análisis adecuado del lenguaje porque ofrece mayores garantías de representatividad respecto al uso que de él hacen los hablantes. Y tampoco parece existir ninguna razón de peso que impida la proyección de los resultados de un corpus hacia la globalidad de la lengua ("generalización"), como se alegaba en la escuela generativista. No debería inferirse de lo anterior que los corpus deben constituir la norma de referencia para los hablantes, entre otras razones, quizás porque el uso presenta tales variantes que harían difícil desembocar en una sola norma (piénsese, por ejemplo, en las "desviaciones" de todo tipo que se dan en el lenguaje oral, especialmente en el familiar o coloquial, frente al nivel de "corrección" que rige en la lengua escrita, especialmente en la de carácter literario). Pero tampoco cabe la menor duda de que un corpus representativo de cualquier idioma ofrece al lingüista un conjunto de datos lingüísticos ideal para el análisis gramatical, léxico o fonético, cual no se había tenido hasta el momento. Si los resultados del corpus serán o no tenidos en cuenta para establecer la norma y en qué medida, es un tema abierto. No obstante, nadie puede negar que el uso real siempre ha servido de fuente primera para los lexicógrafos "manuales" y ha protagonizado los cambios en fonología, morfología y sintaxis a lo largo de la historia de cualquier lengua.

) Qué es un corpus lingüístico?

Un corpus lingüístico es un conjunto de datos lingüísticos (pertenecientes al uso oral o escrito de la lengua, o a ambos), sistematizados según determinados criterios, suficientemente extensos en amplitud y profundidad de manera que sean representativos del total del uso lingüístico o de alguno de sus ámbitos y dispuestos de tal modo que puedan ser procesados mediante ordenador con el fin de obtener resultados varios y útiles para la descripción y el análisis.

Un corpus lingüístico no es, pues, un mero repertorio de textos, sino un repertorio de datos lingüísticos que pueden alcanzar extraordinaria variedad y amplitud, que se recopilan con fines de investigación y análisis y que precisan de la ayuda del ordenador para hacerlos accesibles al investigador en términos de practicidad y utilidad. Según esta definición, el tamaño por sí sólo no bastaría para constituir un corpus. Actualmente la edición electrónica de textos se ha extendido tanto en todos los ámbitos (no solamente en las editoriales e imprentas, sino también en las empresas y en el trabajo individual) que sería posible referirse a corpus con suma facilidad. El corpus, además del tamaño, debe estar adecuadamente sistematizado y ordenado en todos sus aspectos. Sin ese requisito sería imposible alcanzar un mínimo de **representatividad** respecto al ámbito de uso lingüístico que se

tome como referencia. No ha de dejarse de lado un hecho fundamental: si el corpus se utiliza con fines analíticos y de investigación, debe ser susceptible de **generalización** en cuanto a los resultados que de él puedan obtenerse.

Quizás un somero contraste de lo que puede ofrecer el corpus frente al procedimiento habitual seguido en la elaboración de los diccionarios ilustrará mejor las novedades que aporta el corpus. Tomemos como ejemplo el diccionario de la RAE, modelo indiscutible en la lexicografía española. Este diccionario se ha ido elaborando sobre la base de un "corte horizontal" del uso lingüístico a través de la historia, especialmente centrado en el uso que de la lengua han hecho los autores reconocidos o de prestigio, fundamentalmente literatos. La gran diferencia respecto al corpus estriba en el hecho de que los usos anotados en el diccionario de la RAE no solamente se restringen a un género (el literario), sino que, además, pasan siempre por un segundo filtro interpretativo, el de los académicos, quienes actúan con un notable sesgo purista y, por lo general, conservador. Este hecho no resta autoridad al diccionario, pero sí disminuye el índice de representatividad respecto al uso. El corpus no se somete a ninguna de estas restricciones: la sistematización debe garantizar una razonable representatividad de todos los ámbitos de uso lingüístico y el analista debe atenerse a los significados **reales** que aparecen en las muestras, siempre dentro de su contexto natural.

Sobre la base del corpus se sustenta **la lingüística del corpus**, disciplina que empieza actualmente a cultivarse y que, sin lugar a duda, en un próximo futuro no solamente experimentará un notable auge, sino que se expandirá en distintas direcciones, como ocurrió con la **Lingüística Aplicada** en las dos últimas décadas. Es previsible que algunas subdisciplinas de la lingüística encontrarán en los corpus el mejor aliado para llegar a conclusiones fiables. Tal es el caso, entre otros, de la pragmática, de la sociolingüística y de la lexicografía; pero no menos relevante será la utilidad de la lingüística del corpus para la sintaxis, la morfología y la fonética, así como para la gramática en general. Y a partir de ahí, se reconocerá su validez en prácticamente todas las áreas de la lingüística aplicada.

La utilidad de un corpus no debe llevar a ilusiones vanas. Un corpus no es la panacea de la lingüística: es una valiosa ayuda y un instrumento de trabajo que pronto será imprescindible en los estudios lingüísticos. Tiene, sin embargo, limitaciones que conviene señalar de inmediato:

a. Un idioma, cualquiera que éste sea, es un sistema de comunicación siempre **abierto** a nuevas modalidades y a nuevas adquisiciones. En tal sentido las lenguas son "sistemas estructurados de comunicación que nunca están cerrados ni pueden ser considerados como tales". Definir un sistema lingüístico como **cerrado** equivale a "certificar su defunción", porque ello equivaldría a afirmar que tal sistema ya no está en uso o carece de sujetos que se valgan de él como medio de comunicación. Esa es en realidad la causa que posibilita definirlo como "cerrado". El corpus tiene una limitación temporal incuestionable. Hacia atrás, porque la lengua oral es ya inaccesible (a excepción del período en que las grabaciones magnéticas empezaron a llenar los archivos de emisoras de radio y televisión, aunque debe tenerse en cuenta que este tipo de archivos es también limitado a un tipo de lenguaje oral). Hacia adelante, porque es evidente que el presente en que se recopila marca el punto final de cualquier repertorio. Esa limitación temporal convierte al corpus en un conjunto de datos lingüísticos

"encajonados por arriba y por abajo". Como tales, su vigencia está sujeta a los límites de supervivencia de las muestras recopiladas. En la medida en que la lengua cambia y evoluciona, en la medida en que incorporará nuevos elementos, en esa misma medida el corpus se irá distanciando de la realidad comunicativa del momento e irá perdiendo validez como instrumento representativo para el análisis o la investigación del presente.

b.) Existe algún diccionario que recoja todas las palabras de un idioma? La respuesta es, obviamente, negativa. Este es uno de los principales retos de cualquier obra lexicográfica y el origen de notables esfuerzos económicos y humanos por parte de un gran número de editoriales que pretenden mantenerse al día en la interminable carrera de la adecuación al uso lingüístico. A pesar de todo, hay que reconocer que los diccionarios cuentan ya en su haber con una trayectoria de siglos y suelen combinar y compaginar un buen trecho del pasado con el presente más inmediato. El corpus lingüístico puede incluir en sí el legado del pasado y acercarse todavía más que el diccionario al presente real. Pero es prácticamente imposible que recoja en un repertorio limitado -un repertorio ilimitado sería imposible- todas las palabras y todos los usos gramaticales que encierra un idioma. De la misma manera que no es posible que incluya todas las palabras, tampoco será posible que incluya todos los significados o acepciones de cada término. Un corpus no puede comprender el 100% del uso lingüístico.

c. También en el tema de la **representatividad** el corpus cuenta con limitaciones de importancia. La obra lexicográfica ni siquiera se plantea este problema: los diccionarios dan cuenta de las palabras y acepciones registradas por el lexicógrafo sin preocuparse de señalar la frecuencia de uso y a veces sin especificar si tal uso persiste o no en la actualidad y cómo. El corpus tiene precisamente como objetivo el suministro de información sobre el uso actual, de manera que el investigador o analista pueda emitir conclusiones al respecto. El corpus también informa sobre la frecuencia, aunque en términos **relativos,** ya que las conclusiones generales no pueden ser más que proyecciones *a partir de datos limitados en el tiempo y en el espacio*. En muchos casos sólo se certifica que una voz o una acepción aparece una vez o unas pocas veces, sin poder otorgar garantía más fiable sobre el uso real de tal voz o acepción en relación con el **total** de la lengua (que el mismo corpus no puede representar). En consecuencia, la fiabilidad de un corpus sobre la frecuencia debe entenderse también con limitaciones. Si el artículo "el" aparece 50.000 veces y la palabra "corporeidad" solamente una vez, el analista está autorizado a concluir que "el" aparece en **este** repertorio 50.000 veces más a menudo que "corporeidad". Es, por tanto, una conclusión en términos relativos, restringida al ámbito concreto del corpus y generalizable sólo en función del grado de representatividad de aquél. Adviértase, no obstante, que la relatividad de las conclusiones no debe inducir a restarles el valor y utilidad que les corresponde.

d. El análisis del corpus se sustenta de manera preponderante en los listados de frecuencias y en las *concordancias* o palabras en contexto. Ello es así porque el ordenador sólo es capaz de identificar letras o estructuras que previamente le han sido suministradas y definidas. Tanto las frecuencias como las concordancias se ofrecen de manera ordenada, pero el computador no diferencia,

por ejemplo, entre **corro** (nombre) o **corro** (primera persona del verbo *correr*), si bien es cierto que el desarrollo de los *analizadores morfológicos y sintácticos* contribuirá en el futuro a resolver algunas de estas ambigüedades, especialmente si los elementos lingüísticos se ajustan a procedimientos generales formulados con antelación. De igual manera que no es capaz de asociar las diferentes flexiones verbales a la forma de infinitivo (a no ser que se le indique mediante subrutinas especiales). El investigador tiene que intervenir personalmente para detectar estas peculiaridades y reordenar estos datos. La elaboración de las concordancias puede adoptar variantes muy diversas y útiles.

e. No menos importante es tener en cuenta otra realidad: el corpus presenta una ingente cantidad de datos. Es la gran ventaja que le caracteriza. Sobre una forma muy usada, como puede ser el artículo, es posible que tengamos miles o decenas de miles de ejemplos de uso, según sea el tamaño del corpus. En tal caso es también evidente que la ventaja resulta un inconveniente de cara a la manejabilidad de los datos ofrecidos. Hablar de cifras en términos globales es ciertamente impresionante: un corpus de ocho millones de palabras generará unas concordancias de, aproximadamente, 200.000 páginas. Sólo las concordancias correspondientes a las voces de la letra **A** ocuparán cerca de 20.000 páginas. La consulta de tales muestras o su análisis implican, por tanto, la dedicación de muchas horas de trabajo, algo que probablemente está vedado al esfuerzo individual y requiere la colaboración y trabajo conjunto de un equipo.

h. Finalmente, el corpus lingüístico no debe equipararse al concepto de "norma". Sería equivocado interpretar la lingüística del corpus como "lingüística normativa". Ese paso, si se da, debe darse por otras razones que salen ya del ámbito a que se cicumscribe un corpus: éste no hace sino ofrecernos una muestra de **cómo es y cómo se comporta la lengua en el uso que de ella hacen los hablantes nativos.** Ahí acaba su función.

De esta enumeración de limitaciones no tiene por qué surgir una actitud pesimista, sino más bien **realista**: ser consciente de los problemas a que nos enfrentamos a la hora de trabajar con un corpus es la condición necesaria para no forjarnos ilusiones vanas. Además, el lado menos halagüeño de un corpus no tiene por qué ocultar la otra vertiente, la vertiente positiva.

Ya apunté anteriormente que el corpus ha nacido ligado al ordenador. Con la ayuda de éste último, es posible obtener y clasificar los datos de tal manera que resulten no solamente útiles (los datos lingüísticos serían útiles al lingüista en cualquier forma en que fuesen presentados), sino también **accesibles de manera sistematizada** y en abundancia. Mediante el cuenteo de frecuencias podemos saber con rapidez cuántas veces aparece una palabra o una estructura determinada (la que queramos, mediante definición previa que debe suministrarse al ordenador), cuál es el contexto anterior y posterior. Si elegimos el verbo **depender** podremos comprobar enseguida cuántas veces se utiliza sin preposición y cuántas con la preposición **de**. En general, el régimen preposicional de los verbos quedará ilustrado en suficientes instancias para que el analista pueda llegar a conclusiones realmente representativas de la realidad lingüística. De igual manera será fácil detectar las variantes de uso de voces o expresiones en lenguajes específicos, en estratos diferentes, en regiones diferentes, etc.: bastará una codificación adecuada y sencilla de los textos recopilados para que el investigador

detecte con rapidez el ámbito de uso de cada voz o expresión. El desarrollo de los programas de ordenador, siempre a la zaga de los avances que estas máquinas han logrado en cuanto a la velocidad de proceso, irá ampliando notablemente las posibilidades de manipulación de los datos, de manera que éstos puedan presentarse al interesado de modo más accesible y con opciones más personalizadas, según los intereses del investigador. Pero ya en la actualidad un buen gestor informático puede lograr resultados finales altamente útiles y reveladores, además de los listados de frecuencia y concordancias a que se ha hecho referencia.

La tradición lexicográfica, sin duda la más afín al tema del corpus y a su primer aprovechamiento en la investigación, se ha caracterizado por la paciente recopilación de ejemplos tomados de fuentes escritas, antes eminentemente literarias, ahora más conectadas con la prensa diaria y periódica. En general, el recopilador recoge uno o dos párrafos, en los cuales se inserta la voz. A veces la extensión se reduce a una frase con sentido pleno, aunque ortográficamente quede incompleta. La cita así recopilada o bien es autoexplicativa en lo referido al significado de la voz subrayada o bien se acompaña de una explicación que el recopilador formula. Lo más interesante y menos frecuentemente comentado respecto a esta tradición manual es la razón que empuja a seleccionar o no una determinada voz. En efecto, el recopilador, que suele ser siempre también lector, siguiendo una regla subconsciente de la mente humana, tenderá a fijarse no en lo normal y habitual, sino más bien en **lo nuevo, en lo exótico o en lo poco usual**. Este proceder es extremadamente adecuado para detectar nuevas voces o nuevas acepciones. Y en esta "batalla" están inmersas casi todas las editoriales a la hora de informar o hacer publicidad sobre sus obras lexicográficas: se destaca el número de voces nuevas, la adecuación a los tiempos, la puesta al día de significados. Lo demás queda en un segundo plano. Sin embargo, se da la circunstancia de que "lo demás" es lo más substantivo, ya que lo nuevo constituye para el sistema lingüístico en su totalidad solamente una mínima parte, a veces una parte insignificante. El problema en este caso deriva de **la selección realizada por el recopilador** y de los condicionantes a que éste está sometido. En contraste con este proceder sesgado, el corpus no selecciona voces o acepciones sino que **ofrece muestras**, sin más, tal cual aparecen en los contextos de uso. Naturalmente, y se ha de especificar una vez más, la evidencia que aportan tales muestras será fiable sólo si el corpus en su conjunto es fiable, es decir, si es representativo de la realidad lingüística, lo cual exige un mínimo en tamaño y amplitud y una adecuada sistematización de los datos.

Se ha mencionado varias veces el término **representatividad**. El corpus no debe carecer de esta cualidad, ya que en tal caso no sería fiable para ser utilizado como objeto de análisis, tanto en lexicografía como en otras áreas de la lingüística. Los objetivos que se pretende conseguir requieren que el corpus sea representativo de la totalidad de la lengua usada y, por ende, no debe restringirse ni a una variedad geográfica, ni a un registro, ni a un estrato social, ni a un área específica (ciencia, literatura, periodismo, etc.). Puesto que es imposible reunir en un solo corpus **toda la lengua**, se hace imprescindible seleccionar lo que es típico y central en relación con cada región, con cada variedad, con cada estrato, con cada género y con cada área temática. A ese criterio responde la selección de muestras de lengua del corpus "Cumbre" (véase el Capítulo 2).

El lingüista que se enfrenta a un corpus puede tener la sensación de que se enfrenta a un mundo nuevo. Y, al menos en parte, esta sensación está justificada: el volumen de "evidencia" a que debe hacer frente es tal que puede desbordarle. En todo caso, lo cierto es que tiene ante sí un conjunto ordenado de muestras que hasta ahora no había sido accesible a nadie, no porque faltara voluntad para ello, sino porque se trataba de una empresa irrealizable sin la cooperación de las computadoras. Este ingente volumen de datos es una fuente de información cuya transcendencia probablemente todavía no apreciamos en su justa medida. Pero no cabe duda que la existencia de los corpus lingüísticos supondrá un hito decisivo en los estudios sobre las lenguas. Tampoco conviene olvidar la realidad y la fuerza de los hábitos. Hasta ahora era habitual inventar ejemplos para ilustrar un determinado punto de vista: la gramática "tradicional" es un buen punto de referencia, como también lo ha sido la teoría generativo-transformacional propiciada por Noam Chomsky. Tales ejemplos no son desechables "per se", ya que, en el último de los casos, son fruto de la "intuición" del hablante nativo y producto de su creatividad lingüística. Aunque tampoco hay que olvidar que dichos "productos" (= oraciones) nacen para ilustrar algo ya existente (reglas, teorías ...), **pero no surgen de la necesidad de utilizar la lengua como instrumento normal de comunicación.** Hay una notable diferencia, por tanto, en las razones que dan origen a los ejemplos "expresamente inventados" por los lingüistas y los ejemplos ofrecidos por el corpus. A lo dicho se añade aún otra dimensión que debe tenerse en cuenta: el contexto. Los ejemplos creados "ad hoc" se caracterizan, entre otras cosas, por carecer de contexto extra-oracional. El corpus, por el contrario, ofrece al analista un contexto amplio, dentro del cual las palabras y las frases cobran su significado pleno. Es una importante ventaja no solamente en el campo de la lexicografía, sino también en el de la gramática en general.

Trabajar con los datos que aporta el corpus no es una tarea sencilla: tiene ventajas e inconvenientes, si bien estos últimos no afectan tanto a la calidad de lo ofrecido cuanto a la dificultad de acceder a todos ellos por parte del investigador, que es necesariamente un ser limitado, especialmente en cuanto al tiempo se refiere. Pero es preciso subrayar que quien empieza a trabajar con los datos de un corpus tarda muy poco en convencerse de la abundancia de la información que tiene ante sí, de la riqueza y variedad de esa información y de las insospechadas posibilidades que se abren ante él. Para aceptar lo que el corpus ofrece es necesaria una posición de relativa "humildad" y atreverse a aceptar algunos riesgos, especialmente cuando los datos pongan en evidencia ciertos planteamientos o creencias aceptados que no encuentren justificación en el corpus. Lo que no debería constituir ningún obstáculo ni generar "complejos" dentro de la lingüística del corpus es la renuencia a la innovación que cabe esperar por parte de algunos estudiosos. En ocasiones puede resultar más divertido viajar en carro que en avión. Pero el hecho no anula las ventajas del avión para trasladarse de un lugar a otro con mayor rapidez. La lingüística del corpus exige el trabajo conjunto de hombre y "máquina". El rechazo de la "máquina" en este caso equivaldría a preferir "el carro frente al avión".

2. Breve historia de los corpus lingüísticos.

Aunque sea posible hablar del interés de los estructuralistas americanos por la recopilación de "estructuras" que avalasen las conclusiones lingüísticas, la realidad del corpus, tal cual en estos

últimos años se está desarrollando y configurando, no puede ser disociada de otra realidad: el ordenador. Aún más, el corpus de la década de los noventa es radicalmente distinto al de las décadas anteriores en relación con una característica tan fundamental como es la del tamaño y las posibilidades de procesarlo mediante ordenador. Si los corpus de varios cientos de miles de palabras eran considerados como adecuados en un principio, actualmente se proyectan corpus de 100 millones de palabras o más y los corpus representativos se deben situar en no menos de 8 millones de palabras. El corpus "cobuild" anuncia que su banco de datos lingüísticos se sustenta ya sobre un total de 200 millones de palabras. Las posibilidades de procesamiento de los ordenadores han sido decisivas a la hora de fijar algunos de los parámetros que distinguen a un corpus. Las computadoras han pasado por "varias generaciones", hecho que tiene que ver tanto con su potencia y velocidad de proceso como con sus posibilidades de almacenamiento. Esto origina cambios decisivos en la disponibilidad de los datos lingüísticos que pueden ponerse a disposición del lingüista. Uno de los rasgos típicos del lingüista o gramático investigador ha sido el de la individualidad: ésta venía condicionada no solamente por el hecho físico de que los estudios los realizaba una sola persona, sino también porque el ámbito de tales estudios se restringía, por necesidad, a la experiencia vital de un individuo. En buena medida, los distintos ámbitos del análisis se circumscribían a la "memoria" de quien llevaba a cabo el análisis, memoria que, como mucho, era capaz de proyectarse algunos años, o en el mejor de los casos algunas décadas, hacia atrás, siempre limitada a la vida personal del analista y a su entorno.

De no menor importancia e incidencia era también el valor de la **intuición** para el análisis. Es difícil comprender algunas teorías lingüísticas sin la presencia de este ingrediente y sin reconocer el carácter decisivo del mismo en la concepción teórica y en los resultados derivados de tal concepción. Frente a este sistema de análisis, los datos objetivos desempeñaban un cometido o secundario o "despreciable". Es notorio que los datos objetivos, dependiendo de dónde procediesen, eran incluso rechazados. Sobre tales bases, la función comunicativa del lenguaje no podía ser reconocida en su plenitud y, en todo caso, esta función, esencial en toda lengua, era siempre "filtrada" por tamices subjetivos derivados de la intuición o de lo establecido como "normativo" por criterios no siempre ligados al lenguaje.

Frente a esta situación, el corpus ofrece no solamente una alternativa, sino también una nueva visión del lenguaje y, en consecuencia, nuevas perspectivas para los estudios lingüísticos. Es cierto que hace sólo unas décadas, este tipo de trabajos era imposible y quizás utópico. En la actualidad ya no lo es.

El nacimiento de los corpus se sitúa en la década de los sesenta. Pero no faltan casos anteriores que pueden ser considerados como precursores de este tipo de recopilaciones. Atkins y Zampolli (1994:21) mencionan a F. W. Kaeding, quien en 1898 publicó un listado de frecuencias basado en un corpus de once millones de palabras (*Häufigkeitswörterbuch der deutschen Sprache*), fruto de la recopilación y análisis manual. Poco después, en 1907, J. B. Estoup (*Gammes sténographiques)* utilizó esos datos para realizar algunos cálculos estadísticos sobre la frecuencia de palabras y formas en el texto. No menos conocida es la moda o la pasión por los listados de frecuencias léxicas en la década de los años treinta y posteriormente en la década de los años sesenta (propiciada por la metodología audio-oral en la enseñanza de lenguas extranjeras). Destacan, por ejemplo, Thorndike y Lorge en 1944 (*The Teacher's workbook of 30.000 words*) y la *General Service*

List of English Words de Michael West (1953), punto de referencia obligada para muchos autores de manuales y profesores de inglés como segunda lengua. Se trata, sin embargo, de esfuerzos aislados que aún no han encontrado ni el caldo de cultivo adecuado ni los medios técnicos que eviten tan ingente esfuerzo humano. Probablemente haya que afirmar que es la aparición de los ordenadores lo que da un impulso decisivo a este tipo de estudios.

Entre los estudiosos de la lengua inglesa, dentro de la escuela de lingüistas londinenses, es bien conocido el nombre de R. Quirk. En 1959 este lingüista hizo público el proyecto para recopilar un corpus del inglés británico hablado y escrito (*Survey of English Usage, SEU*). Sin embargo, este corpus no se concibió en su momento para ser tratado por ordenador. El hecho de que estuviese constituido en un 50% por datos lingüísticos orales y en otro 50% por datos de la lengua escrita no facilitaba esta tarea: la digitalización del lenguaje oral presentaba importantes problemas debido a su dificultad para ser introducido en el ordenador. Muy poco después, en los inicios de la década de los 60, otros dos lingüistas americanos, Nelson Francis y H. Ku□era promovieron la elaboración del *Brown Corpus,* que pretendía ser representativo del inglés americano escrito y se planteaba ya "su procesamiento mediante computadores digitales". También hay que reconocer que en el *Brown Corpus* el lenguaje oral no tiene cabida, con lo cual se obviaba una gran dificultad, al mismo tiempo que se restringía el índice de representatividad de la muestra desde el punto de vista de la globalidad de la comunicación lingüística.

Pocos años transcurrirían hasta que esta restricción fuese superada. En 1975, otro bien conocida lingüista, Jan Svartvik, en Lund, se propuso como objetivo transcribir y digitalizar el corpus oral recogido en el *Survey of English Usage*. La transcripción era meticulosa y detallada y dio origen a un corpus de reconocido prestigio: el *London-Lund Corpus*. Estos ejemplos fueron ciertamente pioneros y podrían denominarse "corpus de primera generación". Debe destacarse que el volumen de los datos recogidos no sobrepasaba el millón de palabras (en 500 textos de 2.000 palabras cada uno).
En 1978 se completó en Inglaterra la contrarréplica del "Brown Corpus" americano: el *Lancaster-Oslo-Bergen corpus (LOB)*, que posteriormente ha sido anotado con fines de análisis.

En la década de los 80 los corpus entran en una segunda fase: tanto los adelantos en la potencia de los ordenadores como la posibilidad de captar ópticamente textos escritos por medios mecánicos permiten aumentar considerablemente el tamaño o volumen de los datos susceptibles de ser procesados automáticamente. Nace así una segunda generación de corpus, iniciada con el proyecto "Cobuild", liderado por el lingüista J. Sinclair, en la Universidad de Birmingham (Sinclair 1987) y seguida por el *Longman/Lancaster English Language Corpus*, el primero de unos 7 millones de palabras en su formato básico y el segundo de más de 20 millones de palabras.

El proyecto Cobuild se llevó a cabo en colaboración con una empresa editorial (Collins). Sin lugar a duda fue este hecho el que propició una notable popularización de los corpus. Hasta entonces estos instrumentos habían servido para el estudio de sólo unos pocos y su divulgación ni siquiera era significativa entre los mismos lingüistas. Con el proyecto Cobuild se añade una dimensión nueva: la recopilación no solamente ha de servir para estudios científicos sino que debe dar origen a aplicaciones prácticas y útiles, es decir, debe servir de base a publicaciones que puedan llegar a todo tipo de públicos, no solamente a especialistas en lingüística. Por parte de la editorial se intentan recuperar las inversiones realizadas. Para los autores del proyecto este hecho implica analizar los

materiales lingüísticos de manera que los resultados sean accesibles al público no especializado. El primer producto de estos trabajos se concretó en un diccionario de la lengua inglesa (*Collins-Cobuild Dictionary of the English Language*, 1987), elaborado con algunos criterios novedosos y fundamentado en ejemplos tomados del corpus. A esta obra se añadió pronto una gramática del inglés, un curso de inglés para extranjeros y, recientemente, una variada colección de publicaciones fundamentadas todas ellas en el corpus. El hecho más sobresaliente fue, sin duda, que la publicidad dada al proyecto Cobuild popularizó la idea del corpus en los estudios lingüísticos. Desde entonces no solamente ha aumentado el número de recopilaciones en varias lenguas, sino también los estudios lingüísticos basados en los datos aportados por los corpus. De otra parte, el imparable avance en el campo de los ordenadores está haciendo posible ya la aparición de corpus "de tercera generación", cual es el caso -entre otros- del *British National Corpus Initiative*, liderado por Oxford University Press, en colaboración con Longman Group, Chambers Pubishers, las Universidades de Lancaster y Oxford y la British Library, que se propone la recopilación de un corpus de 100 millones de palabras, incluyendo en ambos casos muestras orales y escritas. El fervor por los corpus se está dejando sentir ya en muchos otros países. Zampolli (Aijmer 1991:21), informaba en 1990 sobre la elaboración de distintos proyectos de corpus en 16 lenguas europeas. En ellos destacan los 190 millones de palabras (fundamentalmente extraídas de textos literarios) recopiladas para el francés (*Trésor de la langue française*, por B. Quemada, hoy día accesible en CD-ROM y "on-line", Quemada (1983)) y los 60 millones para el holandés. Los datos en este ámbito quedan, no obstante, muy rápidamente desfasados: en 1993 se mencionabann ya 26 corpus o proyectos de corpus en el ámbito de la lengua inglesa; 4 en el de la lengua francesa; 7 referidos al alemán; dos de italiano, etc. (Edwards 1993:272ss).

En cuanto al español, se han llevado a cabo recopilaciones parciales (especialmente basadas en el lenguaje literario) en Holanda y se han anunciado algunos otros proyectos de "repertorios textuales". Marcos Marín (1994) y el reciente "Informe sobre Recursos Lingüísticos para el Español", del Instituto Cervantes (1995), aportan informaciones diversas sobre el desarrollo de corpus en España. El Informe del Instituto Cervantes concluye que son más los proyectos de repertorios textuales que los corpus disponibles (el autor cifra en no más de tres millones de palabras los datos disponibles). La conclusión es correcta solamente si se refiere a lo realizado en España, ya que en la Universidad Autónoma de Madrid se puede acceder a un corpus argentino de 2 millones de palabras, a otro corpus chileno, también de 2 millones de palabras y al Corpus Oral de Referencia del Español Contemporáneo, de un millón cien mil palabras. El acceso puede realizarse por **ftp**, mediante conexión a una red informática. El proyecto centrado en la Universidad de Santiago de Compostela, que parece el más adelantado (con más de 2 millones de palabras en 1993), no es aún accesible. También se ha mencionado en la prensa un macroproyecto de cien millones de palabras por parte de la Real Academia de la Lengua. Sería injusto dejar de lado el primer repertorio del español, realizado hace más de diez años en Méjico y dirigido por Luis Fernando Lara, corpus de casi dos millones de palabras, que ha servido de base para la publicación del *Diccionario fundamental del español de México* (1982, 1993). A todo lo reseñado debe añadirse el corpus general de mayor importancia, tanto en tamaño como en extensión, realizado hasta el momento: el corpus "Cumbre", patrocinado por la editorial SGEL. "Cumbre" consta inicialmente de 8 millones de palabras, incluye muestras

representativas del español oral y escrito de España e Hispanoamérica y seguirá ampliándose en el futuro. Este repertorio, fruto de la iniciativa y financiación privada, no es accesible al público, aunque sus resultados en distintas áreas se irán conociendo a través de publicaciones diversas.

Es importante tener presente que los repertorios de varios millones de palabras suponen el almacenamiento de grandes cantidades de datos, lo cual ya constituye por sí solo una importante dificultad para quien desee acceder a ellos de manera individual mediante un ordenador personal. En todo caso, también es preciso tener en cuenta que algunos de los más sobresalientes repertorios lingüísticos desarrollados hasta el momento para otras lenguas y mejor estructurados de cara al análisis se deben a la iniciativa privada (editoriales o empresas ligadas a la informática y los ordenadores). Este hecho limita la accesibilidad y convierte en inocuos los juicios de quienes afirman que "ya existen suficientes corpus lingüísticos". Además, los corpus lingüísticos, en cuanto anclados en un momento determinado del tiempo, deben tener continuidad si pretenden ser representativos en cada momento de la historia de una lengua. En cualquier caso, el análisis, investigación o explotación comercial de un corpus requiere un trabajo y esfuerzo ingentes y una no desdeñable inversión en tiempo.

Probablemente ya puede afirmarse que la "era de los corpus" no solamente se va consolidando, sino que empieza a ser tomada en serio con carácter general. El corpus, como conjunto de datos objetivos que el lingüista tiene a su alcance, ordenado en múltiples modalidades, constituye en estos momentos un paradigma de la investigación lingüística. Mas no solamente es útil para los estudiosos del lenguaje en las universidades: es una necesidad para las empresas dedicadas a la informática y, en general, para todos aquellos que se ocupan del procesamiento de los lenguajes naturales. En el momento en que la lengua oral pueda ser digitalizada de manera automática, los corpus ganarán, además, en fiabilidad y representatividad con la inclusión más equilibrada del lenguaje hablado.

A los corpus les queda aún un largo camino por recorrer. Para progresar y avanzar adecuadamente en su análisis deben desarrollarse y perfeccionarse primero las herramientas de "ordenación y presentación de los datos", es decir, los programas de ordenador que permitan ampliar la presentación actual, centrada de modo prioritario en los listados léxicos, en las concordancias (palabras en contexto), en los resultados estadísticos y en los listados de estructuras especiales, de índole variada. No hay que minimizar la utilidad de todo lo conseguido hasta el momento. Pero su limitación es evidente si pensamos, por ejemplo, en formatos más complejos, como sería la presentación no lineal de la información. La obtención de estos resultados implicaría la búsqueda y ordenación no lineal de datos, lo cual permitiría acceder a elementos relacionados entre sí mediante conexiones multidireccionales de carácter semántico, gramatical e incluso pragmático. Naturalmente, este tipo de resultados requieren una adecuada anotación de los corpus, estadio en el que todavía se ha avanzado poco. Es preciso mejorar los analizadores y marcadores sintácticos y morfológicos, para llegar luego a los analizadores y marcadores semánticos que permitan una anotación automática o semi-automática y, posteriormente, una agrupación automática de los datos requeridos con un razonable índice de fiabilidad. No estamos haciendo ciencia ficción. Nos referimos a algo que la potencia de los ordenadores ya hacen posible en muestras reducidas. La posibilidad de relacionar diferentes corpus, el acceso simultáneo a bases de datos lexicográficos o la utilización de analizadores

morfológicos y sintácticos más sofisticados permitirán en un futuro no lejano obtener resultados útiles en el campo de la lexicografía, en el campo de la gramática (piénsese, por ejemplo en la "gramática de probabilidades"), en el campo de los patrones estilísticos y en un largo etcétera que quizás todavía hoy ni siquiera pensamos como posible.

El corpus es una novedad de corta edad en el campo de la lingüística. Quizás nos ha sorprendido a todos tanto por la celeridad con que se ha impuesto en los estudios lingüísticos como por lo insospechado de sus posibilidades. Quienes han trabajado con los corpus lingüísticos se sienten pronto inmersos en "un mundo en expansión", con posibilidades sin fin en cuanto al aprovechamiento de los datos que se ofrecen al investigador. Probablemente los lingüistas ya no pueden permitirse el lujo de prescindir de estas ayudas.

3. De la utilidad del corpus a su necesidad. Características distintivas.

Al aproximarnos por vez primera a un conjunto de textos con el fin de aplicarles programas informáticos, pensamos de inmediato en el ordenador como un instrumento capaz de llevar a cabo un trabajo puramente mecánico, ordenando los textos, elaborando listados por orden alfabético, poniendo en archivos prefijados determinadas estructuras o palabras. De esta manera, en un corto período de tiempo, se facilita al analista un conjunto de datos dispuestos para su valoración. Y ciertamente esta descripción es fiel reflejo de la realidad en una primera fase. Quienes iniciaron el trabajo sobre los corpus quizás no pensaban en repercusiones de otra índole y quizás tampoco pensaban en llegar a conclusiones verdaderamente innovadoras. La realidad está demostrando que los objetivos iniciales se ven ampliamente desbordados por la realidad.

En primer lugar, si disponemos de la programación adecuada, el trabajo del ordenador, aunque mecánico en su esencia, es capaz de ofrecernos muchos más datos de los que podríamos esperar al inicio. De hecho, lo que es un mero listado por orden alfabético puede presentar variantes múltiples: podemos listar las palabras tal cual aparecen en un diccionario, pero también podemos listarlas según las terminaciones; podemos agrupar las palabras que contengan "x" número de vocales o de consonantes, listar las palabras que contengan un determinado número de sílabas o letras, o en relación con la voz que sigue o precede (una fuente inagotable de información sobre las relaciones, por ejemplo, entre el sufijo femenino español y su dependencia de la terminación del nombre al que se adhiere o del cual depende)... Lo que es un listado, por tanto, presenta tal variedad de posibilidades que sobrepasa con creces el concepto más elemental de "listado".

Otro de los grandes pilares sobre los que se ha justificado la utilidad de un corpus es el de las concordancias léxicas o la presentación de cada palabra con un determinado contexto, antes y después de la palabra en cuestión. Estos datos son realmente útiles para el lexicógrafo o para el gramático que, además del significado léxico, quiere indagar sobre la colocación de los términos dentro de la oración. Las concordancias no deben entenderse como una simple secuencia de formas. Una adecuada programación permite la obtención de variantes de gran utilidad para el investigador del lenguaje: cabe la posibilidad de seleccionar sintagmas, grupos de palabras encadenadas, grupos de palabras con sufijos o prefijos predeterminados de acuerdo con los parámetros que nos interese conocer para un fin concreto. Este hecho puede ser de capital importancia para el estudio gramatical de una lengua, para

el estudio estilístico, para el análisis de los textos poéticos y literarios, para averiguar cómo se logran determinados efectos poéticos a través del ritmo y la rima... De nuevo, hay que reconocer que la "concordancia" puede tornarse más compleja y útil de lo que podría suponerse en un principio. Piénsese que este tipo de muestras se obtienen sin que el ordenador "reconozca" las palabras (objetivo que podría lograrse, al menos parcialmente, con la ayuda de analizadores morfológicos o sintácticos). Si en un futuro, como se espera, el ordenador adquiere alguna capacidad de discriminar las formas (mediante marcadores), los datos que llegarán al investigador habrán sido filtrados previamente y evitarán un trabajo suplementario, pudiéndose aquél concentrar mejor en un tipo de análisis más refinado.

Es evidente la potencia de un ordenador para contar y llegar al cálculo de valores estadísticos. En este campo la ayuda de este instrumento para el investigador es significativa. Con el cuenteo que el computador hace de letras, palabras, oraciones y párrafos y con la posibilidad que tiene para relacionar todos estos datos en pocos segundos, es fácil comprender cuán lejos cabe llegar en la valoración de resultados relativos a la frecuencia, porcentajes de frecuencias, usos de palabras y su incidencia porcentual en la configuración de un texto, etc. A través de los valores estadísticos cobran pleno sentido los "vocabularios fundamentales", que adquieren fiabilidad y validez por fundamentarse en muestras representativas. Cobran sentido los vocabularios especiales propios de cada área, que también pueden derivar de muestreos representativos. No es irrelevante para un lingüista saber, por ejemplo, que de una muestra de 300.000 palabras del lenguaje oral, sólo 21.310 son términos diferentes (incluidas diversas flexiones de verbos, nombres y adjetivos); que de esas 21.310 formas lingüísticas, 10.134 (un 47,56%) aparecen una vez en el texto y constituyen sólo el 3,39% del total; más de 16.000 palabras aparecen de 1 a 4 veces (el 76,6% de palabras diferentes, que, sin embargo, constituyen sólo el 8,8% del total textual), mientras que solamente 15 voces se repiten 49 veces. Las que se repiten más de 110 veces apenas pasan de 150 y sólo 4 palabras se utilizan más de 900 veces en el total de las muestras. En resumen, con las palabras diferentes (incluidas flexiones de una misma voz) se llega a constituir un 57,7% del total del texto. Dicho de otra manera (como el reverso de una moneda), casi la mitad del texto (42.3%) es fruto de la repetición de palabras, conclusión a la que también llegó Sinclair en el análisis estadístico del corpus "Cobuild" (Sinclair 1991:35; Philips M. 1985). De estos datos es posible concluir que el hablante de español no se diferencia de los hablantes de otras lenguas en la variedad léxica utilizada: con pocas palabras llega a constituir alrededor de un 10% del discurso oral. Si bien sería erróneo finalizar el análisis en estas cifras, no es menos cierto que tales datos estadísticos pueden servir de base excelente para el estudio, por ejemplo, de los vocablos de alta incidencia y de su distribución a lo largo del discurso.

Lo que más sorprende al estudioso del corpus es que las muestras revelan con nitidez que el uso lingüístico no siempre se atiene a lo esperado o a lo descrito y prescrito en gramáticas y manuales. Hay con frecuencia una clara disociación entre el lenguaje real y el "oficial". De tal manera que no carece de sentido plantearse con seriedad algunos interrogantes como:) Hasta qué punto las palabras tienen significado autónomo y hasta qué punto el significado depende del contexto?) Hasta qué punto las unidades de significado sobrepasan el umbral de la palabra, del sintagma e incluso de la oración?) Hasta qué punto la gramática se puede separar de la sintaxis o de la morfología y viceversa?) Hasta qué punto reflejan los estudios sobre el lenguaje el alto nivel de repetición y redundancia que se

detecta en los corpus?) Hasta qué punto y en qué medida es fiable la intuición individual en el estudio de la lengua?) Cuál es la relación entre forma y significado o entre forma y gramática? Por otro lado,) acaso no viene siendo demasiado "fácil" el salto que da la lingüística de la realidad a la teoría? El corpus obliga a un análisis más riguroso porque hace imposible huir con rapidez de la realidad usual, la cual se interpone tercamente con la fuerza irrefutable de los hechos. Si contrastamos el fruto del análisis individual con la evidencia "bruta" pero objetiva que presenta el corpus es fácil comprobar cómo el estudioso filtra a veces en exceso las muestras estudiadas, dejando de lado lo obvio o no relevante para su interés y destacando sin mesura lo que afecta a su particular punto de vista. Es difícil, quizás imposible, escapar totalmente a esa limitación. Y lo es todavía más renunciar a la creación de ejemplos que se ajustan a una tesis determinada, substituyendo este procedimiento por la subordinación y sometimiento de cualquier hipótesis a la realidad de los ejemplos derivados del uso real de la lengua.

Se ha hablado reiteradamente de la **utilidad** del corpus para los estudios sobre el lenguaje. En realidad, conocida la potencialidad de este instrumento, es preciso ir un poco más lejos y hablar no sólo de la utilidad sino también de la **necesidad de los corpus**. Consideremos algunas razones en favor de este "salto cualitativo" en la apreciación de los corpus.

Si contrastamos el corpus y los diccionarios, no parece que pueda establecerse comparación de igualdad entre ambas realidades. El diccionario puede y debe ser complementado de varias maneras y esto lo hace el corpus, ofreciendo:

- actualidad de usos y acepciones en cuanto al tiempo, lo cual evita que el diccionario se convierta en un "cementerio de voces usadas",
- contexto adecuado, que explicita el valor exacto de cada uso y
- evidencia objetiva, que no permite caer en la subjetividad interpretativa (véase más adelante el capítulo dedicado a la lexicografía en relación con el corpus).

Si contrastamos el corpus y las gramáticas, de nuevo surgen importantes ventajas en favor del corpus. Las gramáticas suelen combinar los usos lingüísticos (especialmente los del propio autor) con las abstracciones de dichos usos en orden a la formulación de generalizaciones o reglas. El gran enemigo de los gramáticos es la intuición personal, que puede llevar a conclusiones escasamente sustentadas por la generalidad de los usuarios de un determinado sistema lingüístico. De nuevo debe afirmarse que el corpus hace prohibitivo este inconveniente, ya que obliga al lingüista a fundamentar sus generalizaciones en evidencia suficiente que garantiza la formulación de las reglas. De manera similar sería posible concretar la necesidad del corpus en prácticamente todos los ámbitos de los estudios lingüísticos.

Ya se habló en el apartado 1. de algunas de las características que definen el corpus. Algunos comentarios adicionales ayudarán a comprender con mayor profundidad qué cualidades deben concurrir en un corpus para que éste cumpla fielmente con las funciones y finalidades descritas hasta el momento.

Al hablar de la "historia de los corpus lingüísticos" se mencionaron los realizados en la

década de los sesenta, entonces considerados como excelentes por contar con un millón de palabras. Desde la perspectiva actual, sería adecuado referirnos a aquellas recopilaciones como "muestras de corpus", más que como "corpus". Si un "corpus" debe constituir un repertorio representativo del lenguaje usado, en tal caso aquellos dos proyectos se quedan cortos en el índice de **representatividad**: un millón de palabras es una cantidad baja en relación con la totalidad de una lengua. En consecuencia, podría decirse que aquellas primeras recopilaciones fueron y siguen siendo actualmente un buen repertorio de textos, orales o escritos, pero sin gozar del grado de representatividad necesario para ser denominados con el término de "corpus".

De la importancia de la representatividad como fundamento principal del corpus dan fe tanto la realidad léxica como la gramatical. Se ha mencionado el componente léxico en varias ocasiones. Y probablemente el léxico es el primer objetivo perceptible en un corpus o, al menos, el más perceptible. Pero no es menos importante la incidencia del corpus en la gramática, incluso en campos menos conocidos o populares, cual sería el de la **gramática probabilística**. Nuestra manera de aprender la primera lengua, de niños, está basada en gran medida en un "cálculo de probabilidades", el cual nos orienta en la selección de elementos que configuran el código lingüístico para lograr establecer la comunicación adecuada. Quizás debería ser entendida así la "actualización" del uso lingüístico por parte de los hablantes nativos en los primeros estadios del aprendizaje. Pues bien, un corpus es susceptible de ofrecernos información fiable y objetiva capaz de ayudarnos en la búsqueda de esos parámetros de probabilidad en que se basan los hablantes de un idioma en el proceso de adquisición de éste. Si esto es así, podríamos adelantar con mayor eficacia en la comprensión de los mecanismos de adquisición lingüística. Es evidente que para que este objetivo sea alcanzable se precisan cotas adecuadas de representatividad de las muestras lingüísticas analizadas.

Es difícil, o quizás más arriesgado que difícil, fijar los parámetros de la representatividad de un corpus. Carecemos aún de estudios suficientes para poder dictaminar cuándo los resultados obtenidos empiezan a ser fiables de manera satisfactoria. Será necesario todavía comparar los resultados obtenidos con corpus de diversos tamaños para poder llegar a conclusiones fiables. De momento se baraja la cifra de 8 ó 10 millones de palabras como punto de partida. Pero lo que la tecnología de los ordenadores ha permitido hacer en cada momento puede que haya sido la razón para llegar a establecer dichas cifras. Esto no sería adecuado. Si en la década de los 60 se hablaba de un millón de palabras, al menos en parte ello era debido a que las posibilidades que ofrecían los ordenadores tenían límites muy bajos. Cuando en la década de los ochenta se iniciaron los corpus de varios millones de palabras, los ordenadores ya permitían procesar tales volúmenes de información. En la actualidad se proyectan y elaboran corpus de 100 millones de palabras. Esta cifra ya podría incluso incrementarse a varios cientos de millones de palabras... Sin embargo, la potencialidad tecnológica de los medios utilizados para ordenar los datos no debe erigirse en razón decisiva para determinar la representatividad lingüística. Esto no obstante, siempre será inamovible el principio estadístico de que "cuanto mayor sea el número de elementos tomados como base para los cálculos, mayor será el grado de fiabilidad de la muestra". Sin menoscabo de este principio, los lingüistas deberían establecer un mínimo a partir del cual sea posible afirmar que un determinado corpus es fiable. En el presente caso se ha trabajado con la hipótesis de 8 millones de palabras. Creemos que es un número razonable, pero habría que validarlo.

CAPÍTULO II

Organización del corpus "Cumbre"

El diseño de un corpus admite varias posibilidades. En cada caso se elegirá aquella que mejor se ajuste a dos parámetros esenciales:

- los objetivos que se pretende lograr y
- los medios de que se dispone para llevar a cabo el proyecto.

Los objetivos pueden enunciarse de modos muy diversos o apuntar hacia finalidades bien diferenciadas. A manera de ejemplo, mediante un corpus es posible que se pretenda:

- recopilar "al azar" muestras lingüísticas, las suficientes para configurar el total de un corpus. Si en este procedimiento tenemos en cuenta los condicionantes que exigiría elegir muestras por azar con el fin de lograr un todo representativo, la recopilación sería poco rentable: precisaríamos un muestreo muy extenso, ya que la representatividad exigida implicaría la existencia de un alto número de casos que garantizasen la variedad tanto en extensión como en profundidad.

- recopilar muestras que se rijan por el criterio de "calidad literaria o académica", partiendo de la convicción de que esta variedad del lenguaje es la que más influencia ejerce sobre los hablantes. Este proceder volvería a enfrentarnos al criterio de "autoridad", poco hermanado con el de representatividad del uso. Con este procedimiento obtendríamos un corpus "sectorial".

- recopilar muestras del lenguaje fundamentadas en la demografía. En este caso se perseguiría la representatividad geográfica (regiones) y social (estratos sociales).

- hacer una recopilación basada en el tipo de lengua "usada" mayoritariamente. Sin lugar a duda el lenguaje coloquial, la radio, la televisión, los diarios y las revistas ocuparían, en este caso, una clara preeminencia.

Podríamos seguir enumerando otras variantes u objetivos, todos ellos legítimos. Nuestro corpus, sin oponerse radicalmente a ninguno de los procedimientos apuntados, se propone ser representativo del lenguaje usado **en su globalidad**. La selección de las muestras se ajusta, por tanto, a ese objetivo y se ha decidido sobre criterios de variedad (modalidad oral y escrita), criterios que, a su vez, están modulados y condicionados por rasgos fácilmente objetivables, como son, por ejemplo, la región geográfica, el estrato social, el medio de transmisión, el género, el tiempo en que se han producido las muestras o el presumible grado de influencia de éstas sobre los usuarios.

La representatividad de una parte en relación con un todo nunca alcanza la perfección, ya que aquélla se basa en la proyección de las partes sobre el todo. De ahí que nuestra selección haya seguido un procedimiento que podríamos denominar "multidireccional y arbóreo". Cada uno de los criterios moduladores se subdivide en otros subconjuntos, hasta alcanzar los límites impuestos por el corpus diseñado.

El criterio temporal, por ejemplo, ha de definir el **cuándo** y el **hasta cuándo.** Nos decidimos por el español contemporáneo, en vez del español "actual". El tratamiento no es idéntico, sin embargo, en el lenguaje oral y en el escrito. Las muestras orales de "Cumbre" se refieren en su totalidad a dos años: 1993 y 1994. La decisión no fue gratuita: la recopilación de datos orales imponía en parte referirnos al presente en que estábamos situados. La sola excepción a esta regla se aplica a películas y series de televisión, que en su origen pueden datar de años anteriores. Pero el lenguaje escrito se extiende a la segunda mitad del siglo XX , aunque, siempre que era posible, se ha preferido el más actual o cercano al presente. Así se ha hecho, por ejemplo, en el caso de la prensa, las revistas y los folletos de divulgación. Estos sectores tienen "vigencia", de manera especial y a menudo casi de manera exclusiva, en el "día a día". El alcance de este tipo de lenguaje es reducido en el tiempo, si bien es verdad que el grado de influencia sobre los usuarios o su capacidad de reflejar el uso espontáneo de la lengua por parte de los hablantes son **intensos**. Podríamos decir que esta variedad lingüística gana en intensidad lo que pierde en extensión o amplitud temporal. Las muestras recogidas de libros (literatura, lenguajes sectoriales, manuales educativos, etc.) gozan de una perdurabilidad mayor, tanto hacia el pasado como hacia el futuro. Basándonos en este hecho, no ha parecido oportuno limitar la selección a aquellas publicaciones realizadas en los dos o tres últimos años, por ejemplo.

Otra decisión de importancia se refiere al ámbito geográfico. Era obvio que debía estar representado en nuestro corpus tanto el español de España como el de Hispanoamérica. La decisión más problemática se planteaba a la hora de definir qué porcentaje del total se reservaba a una u otra área geográfica. Las fórmulas posibles eran muchas y cada una dependía de los criterios que se tomasen en consideración. Finalmente se decidió que el corpus sería mayoritariamente "de España", pero sin que tal "mayoría" alcanzase cotas tan altas que acabasen imponiendo de hecho la variedad española. En favor de tal decisión se consideraron válidos los siguientes argumentos:

a. El corpus se llevaba a cabo en España y los recursos disponibles no permitían, de momento, equilibrar las muestras al 50%.

b. La representatividad de las partes respecto al todo no debía medirse exclusivamente en razón de la cantidad. En el presente caso concurren otros factores importantes. Entre ellos destaca el que podría denominarse de "prestigio". Actualmente la lengua hablada en la Península Ibérica todavía ejerce mayor influencia sobre el uso general del idioma, porque los hablantes así lo perciben (las razones y motivos que puedan inducir a ello admiten gran variedad de justificaciones y explicaciones) y porque las publicaciones impresas en España y de autores españoles son más abundantes y más utilizadas en Hispanoamérica que al revés.

c. A pesar de lo anterior, no debe dejarse de lado otra realidad, de gran importancia en muchos países europeos y actualmente en los Estados Unidos: las relaciones comerciales con los países hispanoamericanos y las fuertes corrientes migratorias otorgan un peso nada despreciable a la modalidad lingüística de Hispanoamérica. El hecho se percibe claramente en el estudio de las motivaciones que impulsan al aprendizaje del español como lengua extranjera.

d. A estas tres razones se unían dos elementos más: por un lado la intencionalidad del patrocinador (la editorial SGEL), claramente inclinada hacia una mayor presencia de las variantes hispanoamericanas, y, por otro lado, las expectativas de muchos posibles usuarios extranjeros, para

quienes el peso demográfico de los hablantes se percibe con gran fuerza, debiendo traducirse esta realidad en una mayor presencia de Hispanoamérica a la hora de tratar asuntos relacionados con la lengua española.

Como consecuencia del análisis de todos estos factores, se asignó a la variedad lingüística de España el 65% del total y el 35% se reservó a la variedad lingüística de los países de habla hispana en el continente americano.

Una tercera decisión de especial relieve se refiere a las dos modalidades de uso lingüístico: la lengua oral y la lengua escrita. Se han aplicado dos criterios ligeramente diferentes en cada ámbito geográfico. En España se ha asignado a la modalidad escrita un 70% y a la oral un 30%, mientras que en Hispanoamérica el porcentaje ha sido del 60% para la lengua escrita y del 40% para la lengua oral. La razón para esta diferencia en cada uno de los ámbitos geográficos se justifica en el mayor peso del uso escrito en España de cara a la fijación del estándar o de la norma y en la abundante exportación de libros a Hispanoamérica. En el continente americano se ha aumentado un diez por ciento el peso de la modalidad oral precisamente para "equilibrar" la balanza frente a la mayor presencia de la lengua escrita en el ámbito geográfico de España. Parece razonable admitir, además, que la influencia de la lengua escrita u oral en una sociedad de hablantes está relacionada con la mayor o menor producción en cada una de esas modalidades.

Esta distribución en dos macroáreas geográficas va seguida luego de una más detallada definición de regiones y grandes capitales en España (Asturias-Cantabria, Galicia, Castilla-León, Aragón, Cataluña-Levante-Murcia, Andalucía, Extremadura, Canarias; Madrid, Barcelona, Sevilla, Las Palmas) y de grandes capitales y grupos de naciones en Hispanoamérica. La América hispana se dividió en cinco grandes zonas (América Central, México, Venezuela-Colombia-Ecuador, América Andina (Perú-Bolivia-Chile), Argentina-Sur (Argentina-Paraguay-Uruguay) y se tomaron algunas grandes ciudades como puntos de referencia para la recogida de muestras -especialmente prensa y conversaciones cara a cara- (La Habana, México DF, Caracas-Bogotá, Lima, Santiago de Chile, Buenos Aires). La organización en zonas se decidió sobre bases lingüísticas, en razón de las diferencias o afinidades más sobresalientes. Debieron dejarse de lado, necesariamente, multitud de variantes propias de cada uno de los países. Las limitaciones de la recopilación obligaban a estas decisiones. Para alcanzar un mayor grado de representatividad lo ideal sería incluir todas las variantes que se dan en el uso lingüístico, atendiendo a todo tipo de parámetros, horizontal y verticalmente. Este objetivo es inalcanzable en un corpus de 8 millones de palabras y exigiría, además, una aportación muy elevada de recursos. No obstante, es válido el principio de que cuantas más variantes lingüísticas se incluyan en el total, más representatividad se alcanzará en el producto final. De ahí que en "Cumbre" se haya buscado siempre la meta de la variedad, dentro de las disponibilidades permitidas por el diseño general. El tamaño de las muestras afecta también a la variedad. Un modelo utilizado en varias recopilaciones es el texto de 2.000 palabras (*Brown Corpus*). En nuestro caso no solamente hemos puesto límites en la extensión, sino que hemos buscado la variedad temática o de secciones en la prensa, en la radio, en la televisión, en las revistas o en algunos libros o manuales. Un periódico, por ejemplo, no se ha recopilado en su totalidad durante varios días, sino que se ha ampliado el periodo temporal seleccionando un extracto de cada una de las secciones que normalmente lo integran. Se gana así en variedad. Algo semejante se ha hecho con los manuales. En cuanto a las obras

literarias se ha preferido aumentar el número de títulos o de autores limitando la muestra seleccionada en cada caso. No podía ser de otro modo, si se tiene en cuenta que la proporción de palabras diferentes disminuye en relación directa con la extensión de los textos.

Este ha sido el esquema general sobre el cual se sustenta la selección de muestras para nuestro corpus:

I. DISTRIBUCIÓN GENERAL DEL CORPUS:

1. ESPAÑA (total): 65%: 70% escrito.
30% oral :

2. HISPANPOAMERICA (total): 35 %: 60% escrito.
40% oral:

II. DISTRIBUCIÓN DEL CONTENIDO:

A. ESPAÑA

I. Lenguaje escrito:

1. Libros:

1.1. lenguajes literarios y formales (novelística):
1.2. poesía
1.3. cuentos
1.4. relatos cortos
1.5. historia
1.6. economía
1.7. arte
1.8. arquitectura
1.9. técnica
1.10. sociedad
1.11. sicología
1.12. filosofía
1.13. informática
1.14. cine
1.15. derecho
1.16. viajes
1.17. biografía
1.18. entretenimiento
1.19. deporte
1.20. medicina/salud
1.21. enciclopedias
1.22. cocina
1.23. música

2. Revistas:

2.1. de información general.
2.2. revistas de la mujer.

2.3. revistas de la casa y cocina.
2.4. revistas del corazón.
2.5. revistas técnicas/especializadas.

3. Prensa diaria:
3.1. nacional.
3.2. regional.

4. Manuales educativos:
4.1. Universidad
4.2. educación primaria
4.3. educación secundaria
4.4. formación profesional
4.5. enseñanzas no regladas

5. Folletos de información/divulgación:
5.1. administración pública:
legislación varia
vida en sociedad
casa
derechos ciudadanos
información varia
5.2. anuncios/propaganda:
nuevas tecnologías
deporte
turismo
consejos
avisos
venta de productos

6. Antologías:
Textos literarios

7. Humor, entretenimiento:
7.1. escritos humorísticos cortos (varios)
7.2. TBOs
7.3. chistes

8. Correspondencia escrita:
8.1. formal

8.2. no formal (amigos/familia)

9. Lenguajes sectoriales:

9.1. ancianos
9.2. adultos
9.3. jóvenes
9.4. niños
9.5. hombres
9.6. mujeres
9.7. modas
9.8. política
9.9. manuales de español para extranjeros

II. LENGUAJE ORAL:

NOTA IMPORTANTE:

- Por "lenguaje oral" se entiende toda comunicación lingüística entre personas, que tiene lugar sin lectura de guión o manuscrito (= producción "espontánea").

A. Radio/TV (ámbito nacional): 50%
B. Zonas y ciudades: 50%.

1. Muestras de Radio y TV

1.1. RADIO (emisoras de ámbito nacional)

1.1.1. Conversación, estrato medio/alto, registros formal y no formal:

i sociedad
ii cultura
iii ciencia
iv educación
v ciencias humanas
vi historia
vii religión
viii economía
ix medio ambiente
x política
xi otros

1.1.2. Conversación, estrato medio/bajo, registros formal y no formal.

1.1.3. Debate, en ambos estratos sociales y registros.

1.1.4. Discusión en grupo, con participación de ambos estratos y registros.

1.2. Muestras grabadas de programas de TV:

1.2.1. Conversación, estrato medio/alto, registros formal y no formal:

i sociedad
ii cultura
iii ciencia
iv educación
v ciencias humanas
vi historia
vii religión
viii economía
ix medio ambiente.
x política
xi otros

1.2.2. Conversación, estrato medio/bajo, registros formal y no formal.

1.2.3. Debate, en ambos estratos sociales y registros.

1.2.4. Discusión en grupo, con participación de ambos estratos y registros.

2. Conversaciones cara a cara (de la vida diaria) :

saludos
salud
el tiempo
dinero
compras
viajes
casa, etc.

3. clases:

Universidad
Bachillerato
Enseñanza primaria

4. Charla entre amigos/familia:

salud
el tiempo
familia
hechos habituales
compras
viajes
planes, etc.

5. Conversaciones por teléfono:

relaciones profesionales
relaciones habituales

6. Narración de hechos, etc.:

temas varios (hechos pasados ocurridos en la vida normal y diaria).

7. Situaciones reales de la vida diaria:

i despacho médico
ii despacho abogado
iii despacho profesor
iv trabajo
v agencia viajes
vi encuestas en la calle/domicilio
vii en el banco
viii en la tienda
ix en el restaurante
x en el bar
xi en el taller/garaje
xii en una sala de espera
xiii en un medio de transporte
xiv en el taxi
xv con la policía (preguntando, comisaría...)
xvi en la recepción del hotel
xvii en una fiesta
xviii en el mercado
xix otros

B. HISPANOAMÉRICA

I. Lenguaje escrito.

Este apartado se distribuye exactamente igual que el correspondiente al lenguaje escrito en España.

II. LENGUAJE ORAL

La distribución se ajusta a las mismas proporciones que las fijadas para el lenguaje oral en España, teniendo en cuenta las variantes referidas al área geográfica **(Zonas -25%-, Ciudades -25%)**. Por ejemplo:

1. Radio y TV (emisoras de ámbito nacional o internacional):

1.1. RADIO:

1.1.1. Conversación, estrato medio/alto, registros formal y no formal:

i	sociedad	México
ii	cultura	México
iii	ciencia	Venezuela
iv	educación	América Andina
v	ciencias humanas	Argentina-Sur
vi	historia	Lima/La Habana
vii	religión	México D.F.
viii	economía	Caracas
ix	medio ambiente	Santiago de Chile
x	política	Buenos Aires
xi	otros	Venezuela, Argentina-Sur, México.
etc.		

Establecido este marco de selección, deben determinarse con exactitud los porcentajes que corresponden a cada subapartado, así como el número de páginas estándar que ha de ser seleccionado. Así, por ejemplo, refiriéndonos al ámbito del lenguaje escrito, en España, se especifican los siguientes porcentajes, número de páginas y muestras concretas:

A. ESPAÑA:

I. Lenguaje escrito: 70% (equivalente a 8.190 páginas estándar):

1. Libros: 30% (que equivale a 2.457 páginas estándar).

1.1. Lenguajes literarios y formales (novelística): 31 % (del total reservado a libros. Equivale 765 págs. estándar)

1.1.1. Ficción en general: 50% (que equivale a 383 páginas estándar, seleccionadas, en proporciones iguales de 35 páginas, entre:)

R. J. Sender, *Requiem por un campesino*
C. J. Cela, *La familia de Pascual Duarte*
Juan Benet, *Volverás a región*
J. Martín Santos, *Tiempo de silencio*
J. M. Gironella, *Los cipreses creen en Dios*
Sánchez Ferlosio, *El Jarama*
Ana M0 Matute, *Primera memoria*
Juan Marsé, *Ultimas tardes con Teresa*
Torrente Ballester, *Los gozos y las sombras*
Delibes, *Cinco horas con Mario*

1.1.2. Novela histórica: 10% (del total reservado a libros, que equivale a 77 páginas, seleccionadas entre las siguientes obras, en proporciones iguales):

Manuel Villar Raso, *Las Españas perdidas*
Sender, *Mr. Witt en el cantón*
J. Lozano, *Germania*
Goytisolo, *La traición del conde Don Julián*
etc.
(...)

2.4. revistas del corazón: 20% (del total del apartado 2., que equivale a 132 páginas estándar, tomadas de)

Hola
Lecturas
Pronto
Semana
(...)

B. HISPANOAMÉRICA

CÓDIGOS DE ZONAS:
AC = América Central
M = México
V = Venezuela/Colombia/Ecuador
AA = América Andina
AS = Argentina-Sur

I. LENGUAJE ESCRITO: 35% (equivalente: 3.780 páginas estándar):
1. Libros: 30% (que equivale a 1.134 páginas estándar):
1.1. Lenguajes literarios y formales (novelística): 31 % (= 352 páginas estándar, seleccionadas a partes iguales entre:)
1.1.1. Ficción en general: 50% (que equivale a 176 páginas)
García Márquez, *Cien años de soledad*
J. L. Borges, *El aleph*
Julio Cortázar, *Rayuela*
Juan Rulfo, *Pedro Páramo*
etc.

1.1.2. Novela histórica: 10% (= 35 páginas estándar)
Carlos Fuentes, *La muerte de Artemio Cruz*
Augusto Roa Bastos, *Yo el supremo*

Miguel Angel Asturias, *El señor presidente*

1.2. poesía: 3% (= 34 páginas estándar)
Octavio Paz, *Pasado en claro*
Mario Benedetti, *Inventario*
Neruda, *Antología poética*
César Vallejo, *Poemas humanos*

II. LENGUAJE ORAL: 35%

(equivalente: 210 horas de grabación; 105 horas de la Radio y televisión, de las cuales 53 h. corresponderán a Zonas y 52 h. a Ciudades, según los códigos predeterminados).

1. Radio y TV (emisoras de ámbito nacional o internacional):
1.1. Radio = 53 h.
1.1.1. Conversación, estrato medio/alto, registros formal y no formal:

i	sociedad	1 h.	M
ii	cultura	1 h.	M
iii	ciencia	1 h.	V
iv	educación	1. h.	AA
v	ciencias humanas	1 h.	AS
vi	historia	1 h.	LH
vii	religión	1. h.	MJ
viii	economía	1 h.	Ca
ix	medio ambiente	1 h.	S
x	política	1 h.	BA
xi	otros	3,5 h.	V, AS, M
etc.			

La planificación anterior requiere, además, muchas otras decisiones en lo relacionado con la aplicación del diseño global a cada caso en particular. Así, por ejemplo, el apartado relativo a la variedad en la prensa y en las revistas, la identificación de las novelas más leídas (a tal fin se han tenido en cuenta los índices de libros más vendidos en la actualidad, además de otros criterios), la selección de manuales más utilizados... todos los capítulos requieren decisiones individualizadas. La selección de muestras relativas al lenguaje escrito es relativamente sencilla, si la comparamos con la complejidad que encierra la recopilación de muestras del lenguaje oral. En el caso de "Cumbre" la complejidad aumenta considerablemente, ya que las muestras orales deben ser representativas de un área geográfica muy extensa y de un abundante número de "medios" diferentes. Para las grabaciones se recurrió a personas situadas en cada uno de los entornos, quienes recibieron las directrices precisas para llevar a cabo su trabajo, de acuerdo con las especificaciones del corpus.

En cuanto a los estratos sociales, se llegó a la conclusión de diferenciar sólo dos: el que denominamos **estrato medio/alto** (que cubre el nivel de la población culta, con estudios universitarios

o equivalentes, pero con la inclusión ocasional de profesiones o personas consideradas cultas, aunque no las avalase una titulación académica) y el denominado **estrato medio/bajo** (que incluye al resto de la población). A lo largo de las grabaciones pronto llegamos al convencimiento de que esta diferenciación era más bien orientativa, puesto que la realidad comunicativa en algunos medios, radio y televisión por ejemplo, incluía personas de ambos estratos en el debate, en la conversación o en la discusión. La diferenciación parece más clara en la selección del registro **(formal - no formal**), si bien también se dio la concurrencia de ambos en algunos casos. De todos modos, los límites de tales registros en el lenguaje oral más habitual no siempre son fácilmente definibles: en los debates televisivos o por radio, se dan frecuentes cruces y mezclas e incluso es preciso reconocer que los debates aparentemente formales conllevan tantas distorsiones del discurso en los campos del léxico, de la sintaxis, de la fonología e incluso de la morfología que todo ello hace que esta variedad llegue a aproximarse a lo que podría denominarse registro no-formal. Sólo en contextos como la clase, debates entre profesionales y sobre temas profesionales, los registros formales se mantienen con mayor "pureza".

A estos problemas se añaden las dificultades para realizar grabaciones en algunos contextos, especialmente en los que se dan cara a cara, en familia, entre amigos, etc. La transcripción de las grabaciones y su digitalización no es tanto difícil cuanto trabajosa y costosa en términos de tiempo y dinero (el corpus no prevé de momento la anotación fonológica de las grabaciones, que habría implicado un alto coste económico y un dilatado período en el tiempo).

Trabajar con un volumen de datos considerable requiere una adecuada previsión en relación con la identificación de las fuentes sobre las cuales se realiza el análisis. Este requisito implica la codificación de cada texto de manera tal que el analista pueda recurrir al original con facilidad. Tal necesidad se hace especialmente urgente en las concordancias. Cada uno de los ejemplos de palabras en contexto viene siempre precedido del código correspondiente. Este código contiene toda la información necesaria para identificar con rapidez:

- el área geográfica
- la modalidad de lengua a que pertenece (escrita / oral)
- la variedad o ámbito
- el registro (y en su caso la zona o ciudad).
- el número de la línea del texto a que pertenece.

Esta información aparece, siempre que lo deseemos, al inicio de la línea de cada concordancia (véase el capítulo sobre "Tratamiento informático del corpus").

CAPÍTULO III

Tratamiento informático y obtención de resultados

> *"Since computers will not go away and are, quite obviously, here to stay, it makes no sense to renounce their application in such an important area of human behaviour as language output. You don't throw away any important tool; you use it." (Josselson 1971:51).*

Como bien se ha apuntado ya con en otros capítulos, la incorporación del ordenador para el análisis y procesamiento informático de datos lingüísticos no fue inmediata. La rotundidad de Chomsky al afirmar que la tarea de recopilar corpus lingüísticos no era propia de un lingüista, sino más bien tarea digna de los documentalistas, no puso más que cortapisas y provocó reticencias entre los estudiosos, quienes no 'osaban' oponerse a las ideas y métodos de trabajo del gran maestro. Chomsky tampoco mostró interés por el estudio de la traducción automática, ni siquiera en el periodo inicial, cuando, burocráticamente al menos, pertenecía a un grupo consagrado oficialmente a ella (Hill 1960:161). El peso predominante y la influencia de la figura de Chomsky sobre otros lingüistas, no obstante, empezaron a debilitarse a principios de los años 80.

Por otro lado, avanzan notoriamente los estudios en microelectrónica gracias al archiconocido >chip'. Este componente diminuto, capaz de albergar miles de circuitos integrados, abrió una nueva dimensión en el mundo de la informática, transformando las potentes y costosas computadoras en ordenadores de sobremesa, más veloces, más asequibles, de mayor fiabilidad y con posibilidades de almacenamiento impensables hasta entonces. Estos adelantos en el campo de la arquitectura de los ordenadores no hicieron más que alentar a los lingüistas que veían, más claro que nunca, la posibilidad de contar con el soporte tecnológico propicio para poner en práctica sus propósitos. Surgen nuevas teorías lingüísticas, sobre todo de orden sintáctico, que desafían los postulados transformacional-generativistas de Chomsky. Nos estamos refiriendo a teorías generativistas como la *>Lexical Functional Grammar'* (Kaplan y Bresnan 1982) y la *>Generalized Phrase Structure Grammar'* (Gazdar y Pullum 1982). La pujanza y frescura de estas teorías llevó a Chomsky a formular su *>Goverment and Binding Theory'* (Chomsky 1982). El punto de mira de estas tres teorías, incluso la de Chomsky, no es otro que ofrecer una base lingüístico-teórica que permita crear y diseñar formalismos específicos para el procesamiento de lenguas naturales (*>Functional Unification Grammar'*, Kay 1983, y *>PATR-II'*, Shieber 1984).

Como puede apreciarse, la batalla aún está muy próxima en el tiempo aunque, eso sí, parece ya definitivamente ganada.

Otra cuestión, menor sin duda, se refiere a la taxonomía de algunos autores respecto a las diversas disciplinas que hacen uso del ordenador para el análisis o procesamiento lingüístico-computacional. La manía por establecer líneas divisorias y clasificatorias no lleva más que a atomizar aún más el saber, privándolo de su carácter interdisciplinario, amputando las ligaduras que interconectan la lexicografía histórica con la lexicografía computacional o la lexicografía basada en el corpus, por ejemplo. Butler (1985:12-35) reconoce cinco disciplinas distintas que se valen del ordenador para el tratamiento lingüístico: (1) la lingüística computacional -procesamiento de lenguas naturales-, (2) la lexicografía, (3) la enseñanza de lenguas asistida por ordenador, (4) el tratamiento y la edición de textos, y (5) la lingüística del corpus o computación lingüística -término este último acuñado por algunos autores (Gazdar y Mellish 1989:3) para disociar la lingüística computacional de la lingüística del corpus. De nuevo observamos ese ánimo de purificar las disciplinas, despojándolas de todo lo que no esté directamente conexo a las mismas. Paradójicamente, es la lingüística computacional la que más uso hace de las recopilaciones de textos para confirmar sus teorías y postulados (Souter y Atwell 1993), ya que son estas colecciones textuales las que contienen más datos lingüísticos acerca del uso real de la lengua (registros, lenguajes sectoriales, variedades de la lengua, estratos sociales, etc.). No cabe duda que por mucho que se esfuercen estos autores en disociar la lingüística computacional de la lingüística del corpus, el grado de interdependencia de estas disciplinas es muy alto, con un punto de partida común: la lengua. Y son dichos elencos de textos, cuidadosamente seleccionados según unos índices de proporcionalidad y debidamente codificados, los que convierten el corpus en la materia prima e indispensable para cualquier investigación lingüística futura que quiera presumir de ser fiable y válida.

Adecuación de un corpus para el tratamiento informático

En esencia, y sin hacer referencia al diseño, cabe definir un corpus lingüístico como una colección de textos anexados uno a otro de forma secuencial (t_1, t_2, t_3 ... t_n). **Es importante diferenciar claramente entre corpus y base de datos. La utilización indiscriminada de ambos términos, incluso por algunos pioneros en la lingüística del corpus, puede conllevar alguna que otra confusión. Una base de datos, en el ámbito que nos ocupa, es un corpus específico compuesto por ficheros de distinta naturaleza (texto y audio) pero que contienen el mismo extracto lingüístico, estando estos archivos enlazados entre sí. De forma que manipulando el texto escrito podemos a la vez**

comprobar y escuchar la secuencia oral correspondiente (véase, por ejemplo Knowles 1993). Los avances en esta parcela de la lingüística del corpus están aún en una fase embrionaria y está pendiente la obtención de algunos resultados concretos. Pero volvamos al corpus y retomemos la idea del mismo tal y como ha quedado definida previamente: secuenciación de textos. Un corpus con dicha disposición secuencial de los fragmentos lingüísticos sin más, reduciría mucho el potencial que éste encierra, y no sería suficiente para sacar el máximo provecho a los sofisticados filtros selectivos propios de los aplicaciones informáticas, privándonos de cruces y análisis más complejos, tales como, listados de las voces utilizadas en el lenguaje oral de Méjico referido al área de la economía con una frecuencia de uso superior a 100 (MeOEco(fre>100)) y su posterior cruce con las voces utilizadas para la comunicación económica en Venezuela con una frecuencia de uso superior a 100 (VeOEco(fre>100)), para conocer luego qué palabras son comunes en ambas variedades lingüísticas ([MeOEco(fre>100)]1[VeOEco(fre>100]), y cuáles son exclusivas de cada variedad ([MeOEco(fre>100)]-[VeOEco(fre>100] y [VeOEco(fre>100)]-[MeOEco(fre>100]).

El corpus precisa, por lo tanto, de una adaptación o adecuación tanto en lo que se refiere al formato como a la disposición del contenido. En esta fase de adecuación de los textos para su posterior tratamiento informático convergen diversos aspectos que, si no han sido previstos con antelación, pueden hacer tambalear tanto la operatividad y la representatividad del corpus como su grado de fiabilidad y validez. Nos estamos refiriendo a aspectos como: (1) el diseño de las partes de las que consta el corpus (proporcionalidad y porcentajes), (2) la elección de la aplicación o aplicaciones informáticas que se van a utilizar, así como su capacidad de gestión, limitaciones, velocidad de cálculo, fiabilidad, posibilidad de modificar los filtros selectivos, etc, (3) la adecuación de los textos en tamaño y formato para que puedan ser capturados sin problemas por las diversas aplicaciones informáticas (compatibilidad), (4) la disponibilidad de un soporte físico adecuado (ordenador(es) con procesadores rápidos, con gran capacidad de almacenamiento y, si es posible, de uso exclusivo para el tratamiento y gestión del corpus, además de impresora(s) y escáner, etc.). Por muy obvios que resulten estos cuatro aspectos es necesario preverlos de antemano. Un proyecto de corpus lingüístico puede no llegar a un final feliz, haciéndose imposible la obtención de resultados, bien porque: (1) el corpus no está debidamente codificado, (2) el programa de gestión es muy limitado, (3) el formato del corpus no es compatible con el exigido por el programa de concordancias o, peor aún, (4) el ordenador es incapaz de gestionar el volumen de trabajo del corpus. Estos fallos y errores de previsión en el diseño del corpus no son apreciables en su fase inicial, pero se convierten muy pronto en una auténtica pesadilla cuando se procede a la fase de informatización, codificación y

análisis de los datos. A ello debe añadirse la previsión de futuro del corpus, es decir, la posibilidad de seguir ampliándolo, lo cual implica que el formato de archivo sea, sino compatible al 100%, sí fácilmente adaptable a los formatos de futuras aplicaciones lingüístico-informáticas.

En este sentido "Cumbre" -Corpus Lingüístico del Español Contemporáneo- parte de un estudio detallado respecto al grado de representatividad que debe tener un corpus, precisando cómo dicha representatividad se distribuye porcentualmente. Parelela a esta labor logística, examinamos varias aplicaciones informáticas diseñadas para este tipo de trabajos, con el fin de conocer la potencialidad, limitaciones y el grado de manipulación interna de los programas para intentar adecuarlos a nuestras necesidades y expectativas, al menos en un principio. Posteriormente se tomó la decisión de diseñar un tratamiento específico para Cumbre, teniendo en cuenta de manera muy especial los objetivos futuros. En cuanto al hardware (ordenadores, impresoras y escáner), ha sido necesario adecuarlo al ingente caudal de datos que se iban acumulando.

La preparación de los textos para su posterior procesamiento informático implica requisitos muy variados, según la información que se pretenda extraer del corpus. Estos requisitos pueden conllevar cambios en: (1) los códigos de indexación, (2) la numeración de líneas del texto (inserción del número de línea al inicio de cada una de ellas) y (3) el formato de archivo (ASCII -*American Standard Code for Information Interchange*-, SGML -*Standard Generalized Markup Language*- y TEI -*Text Encoding Iniciative*-, para compatibilizar la captura y lectura de los textos por las herramientas informáticas).

E = España
H = Hispanoamérica (América Central, México, Venezuela, etc.)

o = oral (radio, televisión, cara a cara)

A = estrato social medio/alto
B = estrato social medio/bajo
F = registro formal
C = registro no formal
i-xi = tema (sociedad, economía, política, educación, deporte, etc.)

e = escrito (libros, revistas, prensa diaria, folletos, correspondencia, etc.)

Cuadro-resumen de los códigos de etiquetación básica del Corpus

Las modificaciones en la etiquetación fueron necesarias para adecuar los textos al formato de referencia COCOA (*>COunt and COncordance on Atlas'*). Este formato permite que mediante la etiquetación inicial que encabeza cada uno de los textos, el programa sepa dónde encontrar la información requerida en la búsqueda y el análisis. Por ejemplo, si deseamos obtener un listado de las voces que aparecen en los textos médicos, acotaremos la búsqueda mediante el parámetro ><C MED>>, de manera que solamente se seleccionarán los textos con dicha etiqueta, ignorándose todos los demás extractos lingüísticos.

La codificación se ajustó finalmente, al siguiente formato (las tres primeras etiquetas encabezan el texto, mientras que la última precede a cada línea):

- Etiqueta A: <A Ho111>; nombre y nE de la muestra.
- Etiqueta B: <B MAF-viii>; país, estrato social, registro y tema genérico.
- Etiqueta C: <C ECO>; tema explícito.
- Etiqueta L: <L 1964>; nE de línea.

```
<A Ho123>
<B VABF-viii>
<C Economía>
<L 1> (A) Muy buenas noches. Bienvenidos a >Puerta Cerrada'. El mes de
<L 2> abril ha concentrado un número de paros inusual en la historia
<L 3> de los paros del país. Después de un año violento como mil
<L 4> novecientos noventa y dos en el que se produjeron mil
<L 5> cuatrocientas protestas, sin olvidar los dos intentos de golpe. <L 6>
Diferentes sectores demandan reivindicaciones salariales y
<L 7> mejoras económicas que en el fondo son un reclamo por una mejor
<L 8> vida como seres humanos y como venezolanos. Los trabajadores del ...
```

Fragmento de un texto codificado

No menos importantes son las anotaciones o apuntes que consideramos relevantes (por ejemplo, para el lenguaje escrito: título, autor(es), capítulos, secciones, distintos tipos de letra, etc.; y para el lenguaje oral: variantes fonológicas, variedades dialectales, omisión de sonidos, pausas/silencios de los interlocutores, entonación, etc.). Dicha información queda anotada entre paréntesis. Este símbolo grafológico funciona como elemento discriminativo. El programa, al encontrarse con dicho símbolo, incorporará o ignorará la información que guarda encerrada, dependiendo de los

parámetros que hayamos seleccionado para nuestro análisis.

Tratamiento informático

Las aplicaciones para la computación de datos lingüísticos, indistintamente de la complejidad y versatilidad individual de cada una de ellas, basan el procesamiento en la reorganización de los resultados e ítems lingüísticos, depojándolos del contexto innecesario o irrelavante, con el fin de resaltar las voces, locuciones, sintagmas, etc., de interés para cada estudio. Las técnicas de procesamiento están bastante estandarizadas en este tipo de *software* específico y son fruto de varios trabajos de investigación en el área de la lexicografía, llevados a cabo hace casi ya veinte años (Jones y Sinclair 1974: Reed 1977). A pesar del tiempo transcurrido, estas técnicas están presentes en todas las aplicaciones lingüístico-computacionales actuales, siendo su popularización entre la comunidad científica bastante reciente. Las causas de este retraso se concentran en tres factores: (1) la falta de infraestructura en cuanto a medios informáticos en muchos de los centros de investigación (la mayoría de los programas creados necesitaban grandes ordenadores -'*mainframes*'- para su funcionamiento); (2) la renuencia o rechazo del ordenador en la investigación lingüística; y (3) la escasa popularidad de la lingüística del corpus.

Todas estas técnicas de procesamiento parten de la palabra como objeto básico para el análisis -sin olvidar algunos estudios que se basan en el cuenteo de matrices (número medio de caracteres por palabra, palabras más largas/cortas, etc.). Una palabra, en lo que se refiere al lenguaje escrito, es una secuencia de matrices concatenadas una a otra sin espacios en blanco intercalados. El espacio en blanco que precede y el que sigue a cada cadena de caracteres delimita cada palabra y la diferencia de las demás.

Antes de proseguir es importante establecer lo que aquí entendemos por *palabra*, *forma* y *lema*. En el párrafo anterior nos hemos referido a la *palabra*, sin precisar su definición desde el punto de vista de la computación. El objetivo era sencillamente explicar el procedimiento discriminatorio que utilizan las técnicas de procesamiento para reconcer, extraer y luego analizar las secuencias de caracteres. Precisando más, cabría definir *palabra* como una secuencia de letras sin espacios en medio, siendo el primer signo grafológico de la misma el que sigue de inmediato a un espacio en blanco o el que marca el inicio de una nueva línea textual (oración, párrafo, capítulo, etc.) y siendo el final el último grafema de la cadena inmediatamente anterior al primer espacio en blanco o signo de puntuación que encuentra a su derecha.

Para aclarar lo que se entiende por *forma*, prestemos atención a las siguientes siete secuencias de caracteres: *comemos*, *comido*, *comía*, *comemos*, *comiendo*, *comido* y *comería*. Las siete secuencias constituyen siete palabras, de las cuales dos (*comemos*, *comido*) están repetidas. Tenemos, por lo tanto, cinco palabras distintas, todas ellas derivadas del verbo *comer*. Estas cinco palabras diferentes en cuanto a su grafía, se corresponden, a la vez, a cinco flexiones verbales distintas, o *formas* diferentes, derivadas todas ellas de un mismo verbo: *comer*. La forma del infinitivo *comer* es el término que aglutina las cinco formas diferentes anteriores: es el *lema*. El concepto de *lema* es equiparable al de "entrada" o "voz" en un diccionario, es decir, la raíz o forma básica -sin morfemas derivativos, flexiones verbales, etc.-, a la que recurrimos en la búsqueda de una palabra. Resultaría imposible encontrar la forma o palabra *comíamos*, si no lo hacemos bajo el lema o entrada *comer*, recurriendo a su forma original o Apura@, no flexionada. *Lema* es, por tanto, el término potencialmente más amplio, ya que engloba todas las posibles *formas* que de él pueden derivarse. Por su parte, *forma* abarca todas las posibles *palabras* -representaciones gráficas- que pueden darse en un texto (*LemaεFormaεPalabra*).

Esta distinción entre *palabra*, *forma* y *lema* no está, sin embargo, exenta de problemas.) Qué sucede cuando nos encontramos con dos palabras que tienen la misma ortografía, diferenciándose únicamente por el grafema inicial -en un caso mayúscula y en el otro minúscula? Es posible que se trate de la misma forma, debiéndose la diferencia ortográfica exclusivamente a la puntuación. Pero no podemos descartar la posibilidad de encontrar dos palabras distintas que, a su vez, se corresponden con dos formas distintas, siendo precisamente la grafía del primer carácter el elemento discriminatorio entre ambas formas, por ejemplo, *Ángel* (nombre propio) y *ángel* (nombre común). El problema tiene difícil solución, más aún si tenemos en cuenta que este tipo de decisiones hay que tomarlas *ad hoc* y antes de empezar el procesamiento, es decir, a la hora de definir los filtros de selección. Para este tipo de Aencrucijada@ en concreto existe, no obstante, una convención o acuerdo generalizado, que es el de aceptar todas las palabras -indistintamente de la grafía del carácter inicial (mayúscula/minúscula) o de si toda la palabra va en mayúsculas o minúsculas-, como palabras que corresponden a una misma forma. Los estudios estadísticos realizados al respecto corroboran este hecho, apreciándose una incidencia insignificante en los resultados finales. Por el contrario, el hecho de discriminar todas las palabras cuya grafía inicial se diferencia por ser mayúscula o minúscula, nos llevaría a una Aexplosión@ en los cuenteos finales (total de palabras y formas), con cifras demasiado alejadas de la realidad.

Otra alternativa es ir añadiendo símbolos tipográficos a las voces homógrafas en

el texto (aquellas que se diferencian en la letra inicial -mayúscula/minúscula- o las que, compartiendo la misma ortografía, pertenecen a distintas categorías gramaticales, por ejemplo, *vino* -nombre común o pretérito indefinido del verbo *venir*). Estas anotaciones nos servirán de claves de identificación y discriminación para los filtros de selección, permitiéndonos afinar en la búsqueda y selección de los datos lingüísticos. No olvidemos, sin embargo, los problemas potenciales que encierra esta solución, como son el posible aumento desmesurado de códigos y símbolos tipográficos que, en vez de clarificar, ofuscan el texto, convirtiéndolo en ilegible e impenetrable para otros usuarios. Y por otro lado, está el tremendo esfuerzo y retraso que ello supone, primero, logístico y luego de codificación e informatización. Probablemente, lo más aconsejable siga siendo conservar, en la medida de lo posible, el formato inicial de los textos, con los defectos y virtudes que ello conlleva. Las posibles soluciones son, como acabamos de ver, bastante parciales.

Listados

Veamos ahora las diversas técnicas de procesamiento utilizadas en la lingüística del corpus. Para ilustrar con mayor claridad las diversas herramientas, nos basaremos, en la mayoría de los casos y a modo de ejemplo, en un breve extracto lingüístico (*texto modelo*). Éste facilitará la comprensión de cómo operan y proceden estas herramientas informáticas, ofreciendo al lector una panorámica clara acerca del valor que estas técnicas tienen cuando se aplican a un corpus lingüístico representativo -como es el caso de ACumbre@.

El Consejo de Universidades ha dejado claro que los préstamos avalados por el Estado para que los estudiantes puedan cursar carreras universitarias, son incompatibles con el disfrute de cualquier beca. Según este organismo, los créditos se conciben como una retribución a los estudiantes con buenas calificaciones académicas, pero con un nivel de renta familiar que les impide acceder a una beca.

Por lo tanto, se estima que en el periodo comprendido entre los años 1995 y 2004, unos 250.000 estudiantes pueden recibir un crédito avalado por el sector público. Con el préstamo concedido, estas personas podrían disponer anualmente de 500.000 pesetas para sufragar los gastos que ocasiona estudiar una carrera universitaria.

Los mismos prestatarios dispondrán de un periodo de carencia para devolver el

crédito, asociado al momento en que logren encontrar empleo; en este sentido, las administraciones públicas destinarán casi 43 millones de pesetas para la subvención de intereses y avales.

Texto modelo

Una primera toma de contacto con el texto podría consistir en contar el número de palabras que lo componen, y comprobar cómo dichas palabras se distribuyen a lo largo del texto, según su orden de aparición en el mismo. Pero veámoslo antes en la siguiente oración: *el niño se comió el bocadillo.* Esta oración simple nos da el siguiente listado organizado según orden de occurencia, indicando a la vez el número de veces que cada palabra sale en el texto:

```
el          2
niño        1
se          1
comió       1
bocadillo   1
```

y aplicada esta técnica al texto modelo, obtenemos (*Ejemplo 1*):

El	7	Estado	1	beca	2
Consejo	1	para	4	Según	1
de	8	estudiantes	3	este	2
Universidades	1	puedan	1	organismo	1
ha	1	cursar	1	créditos	1
dejado	1	carreras	1	se	2
claro	1	universitarias	1	conciben	1
que	6	son	1	como	1
los	7	incompatibles	1	una	3
préstamos	1	con	4	retribución	1
avalados	1	disfrute	1	a	2
por	3	cualquier	1	...	

Ejemplo 1. Listado por orden de aparición

Como puede apreciarse, las palabras se listan solamente en una ocasión: la primera vez que salen en el texto. En lo sucesivo, únicamente se van contabilizando.

Otra posible presentación de la misma información anterior es la distribución por orden alfabético, bien de forma ascendente (a-z; *Ejemplo 2*) o descendente (z-a), según interese.

a	2	buenas	1	Consejo	1
académicas	1	calificaciones	1	crédito	2
acceder	1	carencia	1	créditos	1

administraciones	1	carrera	1	cualquier	1
al	1	carreras	1	cursar	1
anualmente	1	casi	1	de	8
años	1	claro	1	dejado	1
asociado	1	como	1	destinarán	1
avalado	1	comprendido	1	devolver	1
avalados	1	con	4	disfrute	1
avales	1	concedido	1	dispondrán	1
beca	2	conciben	1	...	

Ejemplo 2. Listado alfabético ascendente

La organización afabética no tiene por qué estar ordenada según la primera letra de la palabra. Es posible que nos interese un listado ordenado según el grafema final (listado reverso), ascendente (*Ejemplo 3*) o descendente.

a	2	de	8	puedan	1
beca	2	impide	1	podrían	1
ha	1	entre	1	destinarán	1
carencia	1	se	2	dispondrán	1
universitaria	1	anualmente	1	en	3
la	1	este	2	conciben	1
estima	1	disfrute	1	pueden	1
ocasiona	1	que	6	logren	1
una	3	casi	1	Con	4
para	4	al	1	subvención	1
carrera	1	el	7	retribución	1
renta	1	nivel	1	...	

Ejemplo 3. Listado alfabético reverso

Ahora bien, si en vez de organizar y presentar los datos atendiendo a las diversas formas, tomamos como parámetros de distribución los índices de frecuencia, podemos listarlos según el número de veces que cada palabra o forma está presente en el texto, también de forma ascendente (de menos a más frecuentes) o descedente (de más a menos frecuentes; *Ejemplo 4*).

de	8	beca	2	años	1
el	7	crédito	2	asociado	1
Los	7	este	2	avalado	1
que	6	periodo	2	avalados	1
con	4	pesetas	2	avales	1
para	4	se	2	buenas	1
en	3	y	2	calificaciones	1
estudiantes	3	académicas	1	carencia	1
por	3	acceder	1	carrera	1

un	3	administraciones	1	carreras	1
una	3	al	1	casi	1
a	2	anualmente	1	...	

Ejemplo 4. Listado de frecuencias

Para dar una visión más amplia sobre el potencial que encierran estos listados organizados por índices de frecuencia, podemos aplicar esta técnica a un corpus lingüístico mayor. Este listado nos permite conocer las formas que más frecuentemente utilizan los hablantes de dicha lengua, para, así, tenerlas en cuenta a la hora de, por ejemplo, decidir qué vocabulario básico debe incluir un método de enseñanza para ese idioma en concreto. Más aún, podemos, mediante los diversos filtros selectivos aplicados, acotar los listados a un determinado número de voces o extraer bandas de frecuencia muy concretas. Por ejemplo, obtener un listado ordenado alfabéticamente con las formas comprendidas entre las bandas de frecuencia 40 y 1.000 (*Ejemplo 5*; muestra extraída de un fragmento de *Cumbre*).

a	425	la	403	un	195
ahora	44	las	98	usted	80
al	74	le	73	vamos	47
años	71	lo	274	y	356
bueno	65	los	242	ya	52
como	81	más	83	Yo	159
con	120	me	111		
cuando	47	menos	43		
de	671	muy	56		
decir	63	no	358		
del	108	o	100		
dicho	44	Para	95		
el	278	pensiones	45		
en	298	pero	96		
es	308	por	156		
ese	43	porque	88		
eso	58	pues	57		
está	60	que	970		
este	51	qué	45		
esto	41	se	204		
gente	43	señor	40		
ha	126	si	89		
han	50	sí	66		
hay	77	su	54		

Ejemplo 5. Listado de frecuencias ordenado alfabéticamente (frec>40 y <1.000)

Si disponemos de un corpus debidamente codificado y etiquetado, podemos ser más selectivos todavía, precisando y afinando mucho más en la búsqueda (lenguajes sectoriales, estratos sociales, países, lenguaje oral/escrito, etc.). El ejemplo que sigue ofrece un listado especial. La muestra se corresponde con las formas (frec>49 y <250)

utilizadas en Venezuela para la comuncación oral de ámbito económico.

más	246	Ahora	97	caso	60
Hay	243	estado	95	fue	60
está	232	ser	95	mucho	60
Eso	230	mil	94	bancos	58
si	223	esto	92	cómo	58
Entonces	219	tener	92	financiero	57
pero	203	sea	91	parte	57
mercado	178	bien	88	sector	57
tiene	177	estos	88	importante	56
muy	175	dos	87	te	56
economía	170	esa	87	van	56
su	157	realmente	84	menos	54
vamos	157	sobre	84	tienen	54
Venezuela	148	también	83	uno	54
qué	147	decir	82	años	53
nos	144	tu	81	Así	53
va	144	hacer	80	cosas	53
cuando	143	gobierno	76	Estados	53
precios	142	hoy	76	trabajadores	53
ha	141	problema	75	otra	52
nosotros	141	año	74	paro	52
ese	137	dinero	73	cuatro	51
son	137	Sí	72	ellos	50
bolívares	135	programa	71	nacional	50
ya	134	todos	71		
tenemos	129	inflación	70		
país	128	donde	69		
dólares	123	momento	69		
me	113	paros	69		
todo	113	Pues	69		
estamos	112	sus	68		
le	112	económico	66		
bueno	110	sin	65		
aquí	109	ejemplo	63		
Creo	107	verdad	63		
usted	106	ahorro	62		
gente	105	Banco	62		
están	104	esos	61		
Puede	99	han	61		
esta	98	presidente	61		

Ejemplo 6. Listado de frecuencias (Venezuela, economía, frec>49 y <250)

Todos los listados presentados anteriormente ofrecen exactamente la misma información: formas e índices de frecuencia. La única diferencia estriba en la organización y presentación de los resultados, resaltando en cada caso diferentes patrones de comportamiento lingüístico, patrones que resultarían muy difíciles, sino imposibles, de detectar con métodos no-computacionales.

Otros listados de interés

Es imposible dar cabida a todo el amplio abanico de posibles listados que pueden obtenerse con estas herramientas lingüístico-computacionales. La variedad y combinatoria de los diversos filtros selectivos permiten obtener resultados realmente sofisticados.

A continuación presentamos sucintamente algunas ideas adicionales que pueden resultar de interés para incorporarlas y combinarlas con los ejemplos anteriores o, simplemente, para aplicarlas de forma autónoma.

Los listados mencionados hasta ahora se refieren al texto en su totalidad, y no resulta difícil imaginarse que si aplicásemos dichas técnicas sin más a un corpus como *Cumbre*, el listado de formas llegaría a ocupar varias cientos de páginas. Además, una vez que tenemos el listado completo, constatamos que no toda la información resulta relevante. De ahí la importancia de acotar la búsqueda a aquellos datos que realmente nos interesan del corpus en cada momento. Esto conlleva un notable ahorrro en el tiempo de procesamiento, además de ahorro en material (papel, impresora, etc.). Uno de los procedimientos más comunes y sencillos para la selección de formas de grandes colecciones textuales, al margen de las ya expuestas (frecuencia, etc.), es estabalecer acotaciones mediante letras iniciales. Por ejemplo, extraer todas aquellas formas o palabras que empiezan por *A*, o las que están incluídas entre dos secuencias iniciales (por ejemplo, A*ma*-@ y A*min*-@), o, como muestra el ejemplo siguiente (*Ejemplo* 7), las que están comprendidas alfabéticamente entre dos palabras (*debía* - *desfase*).

debía	2	definitiva	3	demográficas	1
debido	1	definitivamente	1	demuestre	1
debimos	1	defraudado	1	den	1
decepcionar	1	defraudando	1	dentro	4
decía	6	deja	4	depende	2
decían	1	Déjame	6	depravación	2
decías	1	dejar	6	deprecia	1
decide	1	dejarse	1	derechitos	1
décima	1	deje	1	derechos	1
decir	63	déjeme	1	deriva	1
decirle	5	dejen	2	des	1
decirlo	1	dejo	2	desaparecer	1
decirme	2	del	108	desarrollar	1
decisión	1	delante	1	Desarrollo	3
decrece	1	Dele	1	desastre	1
dedica	3	delfín	1	desbarata	1
dedicado	1	Delgado	1	desbaratarme	1
dedicados	1	delito	1	descontado	1
dedicaron	1	demagogia	2	descontrol	1
dedicarse	1	demanda	1	descubran	1
dedos	1	demás	6	desde	25
defender	1	demasiado	1	deseo	1
defensa	2	democracia	3	desfase	1

Ejemplo 7. Listado alfabético (debía-desfase)

De igual modo, podemos centrar la búsqueda en formas que contienen determinados prefijos, sufijos, desinencias, morfemas, etc. Por ejemplo, formas acabadas con el morfema femenino plural A-*as*@, desinencia verbal (segunda persona del singular) o cualquier otra forma acabada en esta secuencia de caracteres, o un listado con todas las formas adverbiales acabadas en A-*mente*@.

Como puede observarse, cualquier característica ortográfica o morfológica es suficiente para ser utilizada como parámetro de búsqueda y selección. Igualmente útil puede resultar el número de letras que componen cada forma o palabra. Dependiendo del estudio, es posible que deseemos conocer todas las formas compuestas por un determinado número de letras (*Ejemplo 8*) o puede que nos interese un listado ordenado según el número de grafemas de las diversas formas que componen el texto o extracto lingüístico (*Ejemplo 9*).

alternativas	2	desbaratarme	1	exgobernante	1
brillantísima	1	desprestigia	1	expectativas	1
complementen	1	desrremunerar	2	extensamente	1
comprensible	2	desvergüenza	1	florecientes	1
comunicación	1	determinados	1	forzosamente	1
Constitución	1	directamente	2	garantizando	1
consumidores	1	documentarme	1	hipotecarios	1
contestación	1	enfermedades	2	hospitalario	1
continuación	1	especialidad	1	impetulancia	1
cotizaciones	2	especulación	2	imposiciones	1
culpabilidad	1	estadísticas	1	incentivando	1
demográficas	1	ex-gobernador	1	...	

Ejemplo 8. Listado alfabético (formas con 12 caracteres)

Cabe advertir que en estos dos ejemplos, el programa interpreta las duplicaciones consonánticas A*ll*@ y A*rr*@ con un único cáracter, a la vez que tampoco contabiliza el guión (*ex-gobernador*).

a	2	las	1	para	4
y	2	les	1	pero	1
al	1	los	7	unos	1
de	8	Por	3	claro	1
el	7	que	6	entre	1
en	3	son	1	estas	1
ha	1	una	3	nivel	1
la	1	años	1	renta	1
lo	1	beca	2	Según	1
se	2	casi	1	tanto	1
un	3	como	1	avales	1
con	4	este	2	...	

Ejemplo 9. Listado ordenado por número de caracteres

Los ejemplos expuestos no son en absoluto los únicos posibles, pero sí los más utilizados. No obstante, y sin haber pretendido ser exhaustivos, consideramos suficientes las muestras presentadas para dar constancia de la diversidad de datos que estas técnicas permiten obtener y la relevancia que estos datos obtenidos mediante procedimientos informáticos pueden tener para investigaciones lingüísticas.

Análisis estadísticos

El valor de los listados, sobre todo los de frecuencias, será más fiable y válido cuanto mayor y más representativo sea el corpus y, más, si lo combinamos con un soporte estadístico que nos proporcione datos y valores referentes a: (1) bandas de frecuencia, (2) frecuencias relativas, e (3) índices y porcentajes individuales y parciales con respecto a los totales de palabras y formas aparecidas en el texto.

Para entender mejor los diversos índices estadísticos que obtenemos en este tipo de análisis, examinemos en el cuadro que sigue (*Datos estadísticos*), correspondiente al texto modelo. Este análisis ofrece diez parámetros: nueve de los cuales aparecen agrupados por formas con el mismo patrón de comportamiento en cuanto a su frecuencia de aparación en el texto, mientras que el restante parámetro es un índice global, que resulta de dividir el total de formas por el total de palabras presentes en la muestra textual.

Elijamos, por ejemplo, los datos que conforman la primera línea del análisis estadístico, referidos a las formas que aparecen una sola vez en el texto (columna 1). Las columnas 3 y 4 nos revelan el número de formas y de palabras, respectivamente, que conforman cada banda de frecuencia. En este caso tenemos el mismo valor en ambas columnas, ya que al ocurrir cada palabra en una sola ocasión a lo largo del texto, cada una de ellas se corresponde, a su vez, con una forma distinta, estableciéndose una relación de correspondencia entre palabra y forma del tipo 1:1. En cambio, en la segunda línea (formas con frecuencia 2) podemos observar cómo las 8 formas que se repiten 2 veces en el texto (columnas 3 y 1, respectivamente), suman un total de 16 palabras (8*2; columna 4).

Mientras que en las columnas 3 y 4 obtenemos cuenteos individuales, es decir, restringidos a cada banda de frecuencia concreta, las columnas 5 y 6 muestran estos mismos índices, pero acumulativos, es decir, sumando las formas y palabras nuevas a

las formas y palabras ya contabilizadas en bandas de frecuencia anteriores. Así, las columnas 5 y 6 son las sumas o totales de todas las formas y palabras, respectivamente, contabilizadas hasta ese momento del análisis. Puesto que nos referimos a datos con frecuencia de aparición 1, no tenemos evidencia de otras formas o palabras que hayan salido con anterioridad, con un índice de frecuencia inferior a uno; de ahí que las columnas 5 y 6 tengan los mismos valores que sus homólogas 3 y 4 en este caso. No obstante, si nos detenemos de nuevo en la segunda línea del análisis, observamos cómo la columna 5 (formas contabilizadas hasta ese momento) se ve incrementada de 79 a 87, es decir, en 8 formas (ver columna 3), al igual que la columna 6, que hasta entonces contabilizaba 79 palabras, aumenta en 16 (columna 4), con un total, ahora, de 95 palabras.

La columna 7 muestra el índice porcentual que tiene el total de las formas aparecidas hasta ese momento del análisis con respecto al total de formas o vocabulario que constituyen el texto en su totalidad. Así, las 79 formas con frecuencia 1 suponen un 80,61% sobre el total de las 98 formas del texto. De modo similar, la columna 8 revela el porcentaje de las 79 palabras -*no* formas- sobre el total de las 146 de que se compone el extracto lingüístico. Es importante anotar que los datos de las columnas 7 y 8 se calculan sobre los totales de las columnas 5 y 6, respectivamente (totales acumulados).

La columna final se refiere al porcentaje de texto que suponen todas las formas correspondientes a una misma banda de frecuencia en conjunto. Con estos datos podemos concluir que, por ejemplo, las 79 formas con frecuencia 1 son suficientes para redactar el 54,11% del total del texto.

Mientras que la columna última nos proporciona índices porcentuales referidos a la colectividad de formas con el mismo patrón de frecuencia, la frecuencia relativa (columna 2) es el índice individual de cada una de las formas y su peso porcentual con respecto al texto. De modo que si las 79 formas son suficientes para reconstruir más de la mitad del texto (54,11%), cada una de estas formas con frecuencia 1, escasamente llegaría a configurar el 0,7% del mismo.

Finalmente, el parámetro "ratio formas/palabras", revela el índice global que resulta al dividir el total de formas por el de palabras aparecidas en un mismo extracto lingüístico. Esta ratio da idea de la variedad o riqueza léxica del texto. Los índices oscilan entre 1 y 0, siendo 1 el caso de máxima variedad léxica posible (lo que supondría un texto con tantas formas como palabras), y 0, o mejor dicho muy próximo a 0 (ya que siempre tendremos como mínimo una forma), para el caso opuesto: variedad léxica Anula@, en donde la única forma de la que consta el texto se repite tantas veces como palabras lo conforman.

1	2	3	4	5	6	7	8	9
FRECUENCIA	FRECUENCIA RELATIVA	FORMAS	PALABRAS	TOTAL FORMAS	TOTAL PALABRAS	% FORMAS T.FORMAS	% PALABRAS T.PALABRAS	% TEXTO
1	0.68493	79	79	79	79	80.61	54.11	54.11
2	1.36986	8	16	87	95	88.78	65.07	10.96
3	2.05479	5	15	92	110	93.88	75.34	10.27
4	2.73973	2	8	94	118	95.92	80.82	5.48
6	4.10959	1	6	95	124	96.94	84.93	4.11
7	4.79452	2	14	97	138	98.98	94.52	9.59
8	5.47945	1	8	98	146	100.00	100.00	5.48

RATIO FORMAS/PALABRAS: 0.67123

Datos estadísticos

No es nuestro propósito confundir al lector no experto con fórmulas y operaciones matemático-estadísticas complejas, pero no podemos negar el enorme valor y, sobre todo, la objetividad y neutralidad de estos datos para reflexionar, plantearse hipótesis y extraer conclusiones de interés para la investigación lingüística. Meros indicios y/o hipótesis acerca de ciertos fenómenos y comportamientos lingüísticos o aspectos sobre corrección gramatical podrán confirmarse o rechazarse con la inclusión de estos datos para el estudio y el análisis, pudiéndose, en algunos casos, llegar a detectar errores e inconsistencias referidas a cómo dichos fenómenos o comportamientos lingüísticos aparecen descritos en las gramáticas y cómo se Acomportan@ en la realidad. Los índices estadísticos son fiel imagen de hechos lingüísticos tal y como estos aparecen en un corpus, y si, además, se trata de un corpus lingüístico representativo -muestra inequívoca del uso real de la lengua por sus hablantes nativos- no nos debería sorprender encontrar y comprobar cómo la utilización real de la lengua difiere en algunos aspectos, de forma notable, respecto a la prescripción normativa reflejada en las gramáticas.

Otros datos de interés

Al margen de los datos más completos y, sobre todo, complejos, que proporcionan el análisis estadístico anterior, todos los listados o técnicas, sean de la naturaleza que sean, pueden acompañarse con información breve referida a los totales de palabras y formas. Estos datos meramente informativos y descriptivos indican: (1) el total de palabras reconocidas o leídas, (2) el total de palabras seleccionadas, y (3) el total de formas o vocabulario total. La diferencia entre el primer índice y los dos restantes estriba en que mientras el primero contabiliza todas las palabras del texto, los dos restantes solamente consideran las palabras y formas que se ajustan a las acotaciones

marcadas (por ejemplo, frecuencia, número de caracteres, etc.). Los cuenteos referidos a nuestro texto modelo nos dan:

```
TOTAL PALABRAS RECONOCIDAS:          146
TOTAL PALABRAS SELECCIONADAS:        146
TOTAL VOCABULARIO:                    98
```

Si, por el contrario, seleccionásemos exclusivamente las formas cuyo primer carácter está comprendido entre las letras *a* y *c*, obtendríamos el siguiente informe:

```
TOTAL PALABRAS RECONOCIDAS:          146
TOTAL PALABRAS SELECCIONADAS:         34
TOTAL VOCABULARIO:                    28
```

Los datos pueden ampliarse no sólo a las palabras y formas encontradas en el texto, sino también a los periodos oracionales de los que consta el mismo. Esta información permite estudiar las características propias -estilísticas- de los distintos textos (poéticos, narrativos, periodísticos, etc.), para luego compararlos entre sí. El análisis completo de nuestro texto modelo contiene, a pesar de su brevedad, datos realmente interesantes desde un punto de vista estilístico y merece la pena detenerse en ellos un instante.

```
TOTAL PALABRAS RECONOCIDAS:          146
TOTAL PALABRAS SELECCIONADAS:        146
TOTAL VOCABULARIO:                    98
PALABRA MÁS LARGA (EN LETRAS):        16
PALABRA MÁS CORTA (EN LETARS):         1
N1 MEDIO DE LETRAS POR PALABRA:        5,49
N1 MEDIO DE LETRAS POR FORMA:          6,5

TOTAL ORACIONES:                       5
ORACIÓN MÁS LARGA (EN PALABRAS):      39
ORACIÓN MÁS CORTA (EN PALABRAS):      21
N1 MEDIO DE PALABRAS POR ORACIÓN:     29,2
ORACIONES CON > 50 PALABRAS:           0          0%
ORACIONES CON 40-49 PALABRAS:          0          0%
ORACIONES CON 30-39 PALABRAS:          3         60%
ORACIONES CON 20-29 PALABRAS:          2         40%
ORACIONES CON 10-19 PALABRAS:          0          0%
ORACIONES CON < 9 PALABRAS:            0          0%
```

Totales palabras/formas y datos estilísticos

Concordancias

En los apartados anteriores han quedado claras las virtudes de los diversos listados. Los datos que estas técnicas nos facilitan (formas e índices de frecuencia) se refieren siempre a la totalidad del corpus, del texto o de los fragmentos que hayamos seleccionado o acotado (tema, país, banda de frecuencia, etc.). Sin embargo, en ningún momento del análisis se especifica el lugar concreto de la ocurrencia (texto, línea, etc.). La localización de la ocurrencia puede ser de gran utilidad para conocer el contexto en el que aparece una palabra o forma y verificar su distribución en el texto. Para suplir esta falta de información en los listados, se recurre a la técnica conocida en la computación lingüística como *concordancia*. Concordancia es el conjunto de ocurrencias de una forma, con referencia explícita al contexto en que dicha forma aparece. La técnica de concordancia más sencilla es la que proporciona, además de un listado de formas con sus respectivos índices de frecuencia, la referencia del lugar exacto mediante los códigos de etiquetación. Este tipo de concordancia se conoce también como listado indexado. El ejemplo a continuación (*Ejemplo 10*) ofrece un listado de concordancias indexado de la forma A*económica*@.

```
económica   21
Ho111 VAF-viii Eco 126, Ho111 VAF-viii Eco 127, Ho111 VAF-viii Eco 130,
Ho111 VAF-viii Eco 227, Ho111 VAF-viii Eco 411, Ho111 VAF-viii Eco 523,
Ho112 VABF-viii Eco 533, Ho112 VABF-viii Eco 534, Ho112 VABF-viii Eco 550,
Ho122 VAF-viii Eco 343, Ho122 VAF-viii Eco 641, Ho122 VAF-viii Eco 932,
Ho122 VAF-viii Eco 1199, Ho122 VAF-viii Eco 1228, Ho122 VAF-viii Eco 1363,
Ho122 VAF-viii Eco 1426, Ho123 VABF-viii Eco 251, Ho123 VABF-viii Eco 266,
Ho123 VABF-viii Eco 482, Ho123 VABF-viii Eco 1049, Ho124 VBC-viii Eco 460
```

Ejemplo 10. Listado indexado

Como puede apreciarse, la indexación se basa en los diversos códigos de etiquetación que encabezan e identifican a cada uno de los extractos lingüísticos (ver *Cuadro resumen de los códigos de etiquetación del corpus* - *Ho111*: español oral, grabación radiofónica; *VAF-viii*: Venezuela, estrato social medio/alto, registro formal; *ECO*: ámbito econónico; *126*: línea 126 del texto). Al igual que los demás listados anteriores, las concordancias indexadas se pueden ordenar según el orden de aparición de las formas en el texto, alfabéticamente o por orden de frecuencia, y acompañarse con datos estadísticos, además de la combinación de todos los filtros selectivos disponibles en este tipo de herramientas informáticas. Valga como muestra el siguiente ejemplo (*Ejemplo 11*): listado indexado por orden de frecuencias de todas las formas que contienen la secuencia inicial A*económic*-@.

```
económico   66
```

```
Ho111 VAF-viii Eco 3, Ho111 VAF-viii Eco 17, Ho111 VAF-viii Eco 62,
Ho111 VAF-viii Eco 63, Ho111 VAF-viii Eco 68, Ho111 VAF-viii Eco 74,
Ho111 VAF-viii Eco 84, Ho111 VAF-viii Eco 86, Ho111 VAF-viii Eco 97,
Ho111 VAF-viii Eco 107, Ho111 VAF-viii Eco 124, Ho111 VAF-viii Eco 141,
Ho111 VAF-viii Eco 146, Ho111 VAF-viii Eco 148, Ho111 VAF-viii Eco 151,
Ho111 VAF-viii Eco 265, Ho111 VAF-viii Eco 349, Ho111 VAF-viii Eco 373,
Ho111 VAF-viii Eco 394, Ho111 VAF-viii Eco 499, Ho111 VAF-viii Eco 640,
Ho111 VAF-viii Eco 641, Ho111 VAF-viii Eco 647, Ho111 VAF-viii Eco 697,
Ho111 VAF-viii Eco 700, Ho111 VAF-viii Eco 735, Ho112 VABF-viii Eco 3,
Ho112 VABF-viii Eco 6, Ho112 VABF-viii Eco 7, Ho112 VABF-viii Eco 18,
Ho112 VABF-viii Eco 19, Ho112 VABF-viii Eco 19, Ho112 VABF-viii Eco 219,
Ho112 VABF-viii Eco 248, Ho112 VABF-viii Eco 251, Ho112 VABF-viii Eco 252,
Ho112 VABF-viii Eco 265, Ho112 VABF-viii Eco 271, Ho112 VABF-viii Eco 481,
Ho112 VABF-viii Eco 484, Ho112 VABF-viii Eco 486, Ho112 VABF-viii Eco 525,
Ho112 VABF-viii Eco 554, Ho112 VABF-viii Eco 626, Ho112 VABF-viii Eco 636,
Ho113 VBC-viii Eco 343, Ho113 VBC-viii Eco 347, Ho113 VBC-viii Eco 515,
Ho113 VBC-viii Eco 528, Ho113 VBC-viii Eco 592, Ho113 VBC-viii Eco 594,
Ho113 VBC-viii Eco 598, Ho122 VAF-viii Eco 353, Ho122 VAF-viii Eco 780,
Ho122 VAF-viii Eco 1175, Ho122 VAF-viii Eco 1236, Ho122 VAF-viii Eco 1281,
Ho123 VABF-viii Eco 18, Ho123 VABF-viii Eco 271, Ho123 VABF-viii Eco 295,
Ho123 VABF-viii Eco 594, Ho123 VABF-viii Eco 608, Ho123 VABF-viii Eco 739,
Ho124 VBC-viii Eco 40, Ho124 VBC-viii Eco 193, Ho124 VBC-viii Eco 385

económica   21
Ho111 VAF-viii Eco 126, Ho111 VAF-viii Eco 127, Ho111 VAF-viii Eco 130,
Ho111 VAF-viii Eco 227, Ho111 VAF-viii Eco 411, Ho111 VAF-viii Eco 523,
Ho112 VABF-viii Eco 533, Ho112 VABF-viii Eco 534, Ho112 VABF-viii Eco 550,
Ho122 VAF-viii Eco 343, Ho122 VAF-viii Eco 641, Ho122 VAF-viii Eco 932,
Ho122 VAF-viii Eco 1199, Ho122 VAF-viii Eco 1228, Ho122 VAF-viii Eco 1363,
Ho122 VAF-viii Eco 1426, Ho123 VABF-viii Eco 251, Ho123 VABF-viii Eco 266,
Ho123 VABF-viii Eco 482, Ho123 VABF-viii Eco 1049, Ho124 VBC-viii Eco 460

económicos    9
Ho111 VAF-viii Eco 16, Ho111 VAF-viii Eco 317, Ho113 VBC-viii Eco 516,
Ho122 VAF-viii Eco 411, Ho122 VAF-viii Eco 423, Ho122 VAF-viii Eco 896,
Ho122 VAF-viii Eco 1163, Ho123 VABF-viii Eco 311, Ho124 VBC-viii Eco 5

económicas    8
Ho111 VAF-viii Eco 157, Ho113 VBC-viii Eco 341, Ho122 VAF-viii Eco 1376,
Ho122 VAF-viii Eco 1396, Ho122 VAF-viii Eco 1413, Ho122 VAF-viii Eco 1491,
Ho123 VABF-viii Eco 7, Ho123 VABF-viii Eco 789
```

Ejemplo 11. Listado indexado (Aeconómic-@, ordenado según índices de frecuencia)

El valor de estos listados de concordancias indexados está fuera de duda, pero son insuficientes para algunos análisis y estudios de orden léxico, morfológico o sintáctico, por ejemplo, en donde lo que interesa es tener constancia del Acomportamiento@ de los diversos ítems lingüísticos con respecto a los demás componentes del periodo oracional. Es decir, interesa también comprobar el lugar que ocupa cada forma en la cadena textual en relación con las demás formas o palabras: categoría gramatical, función sintáctica, régimen preposicional -si lo hubiera-, usos lingüísticos, giros, locuciones, etc. Estos datos únicamente se pueden extraer si conocemos el contexto que precede y sigue a cada forma. El contexto aportará la información necesaria para descubrir cómo Aactúa@ y cómo se Acomporta@ cada forma

cuando ésta es utilizada por los hablantes nativos de una lengua. La técnica que nos permite obtener los contextos literales que envuelven a una forma o palabra se denomina *listado de concordancias* o, simplemente, *concordancias*. El listado de concordancias es el procedimiento más completo y, a la vez, el más complejo, ya que reúne todas las características expuestas hasta ahora: (1) índice de formas que aparecen en un texto, (2) número de veces que dichas formas ocurren (frecuencia), (3) indexación con respecto al lugar concreto de ocurrencia, y (4) contextualización anterior y posterior de cada forma (formato KWIC -*Key Word In Context*). Para entender mejor esta técnica, veámos lo que sucede cuando se aplica a una oración simple como: *el niño se comió el bocadillo*:

```
                bocadillo    1
1      comió el  bocadillo.
                comió  1
1       niño se  comió  el bocadillo.
                el     2
1               el  niño se
1      se comió  el  bocadillo.
                niño   1
1            el  niño  se comió
                se     1
1       el niño  se  comió el
```

Las formas y sus sucesivas representaciones aparecen alineadas y centradas con respecto al contexto anterior y posterior. En el ejemplo hemos reducido el contexto a dos palabras por razones meramente prácticas, pero puede ampliarse a una o varias líneas o, incluso, si se desea, a la oración completa en donde ocurre la forma en cuestión. De modo similar, el orden alfabético puede alterarse o sustituirse por la distribución que más convenga (frecuencia, orden de ocurrencia, número de caracteres, etc.). Los dígitos que acompañan a los datos son comunes a los otros procedimentos ya descritos y se refieren a la frecuencia de aparición (a la derecha de cada forma), y al lugar en el cual ocurren las palabras (indexación -en este caso solamente el número de línea- a la izquierda de las concordancias). El que sigue es un ejemplo más completo (*Ejemplo 12*) y corresponde al listado de ocurrencias ya indexado anteriormente (*Ejemplo 10*). En este caso hemos ampliado el contexto a toda la línea.

```
                                                                económica    21
Ho111 VAF-viii  Eco  126  mercado es, hoy por hoy, la única concepción  económica en el mundo no sólo en Venezuela ni en América
Ho111 VAF-viii  Eco  127  rica Latina,  en el mundo la única concepción económica que está impulsando las  economías en el mund
Ho111 VAF-viii  Eco  130  país comunista que ha emprendido una  reforma económica a fondo y que está dentro de una estructura
Ho111 VAF-viii  Eco  227  la gestión gubernamental que con la  posición económica propiamente dicha. Entonces, vamos a retomar
Ho111 VAF-viii  Eco  411  conomía de  mercado es simplemente la sanidad económica, por qué?, porque  economía de mercado signif
Ho111 VAF-viii  Eco  523   gran problema  que confrontaría una política económica centrada en la  revaluación del bolívar. Una
HO112 VABF-viii Eco  533  ados del Caribe para fomentar  la integración económica de la región. Esta asociación tendría  como o
HO112 VABF-viii Eco  534  dría  como objetivo avanzar en la integración económica y la  cooperación entre los países de la cuen
```

```
Ho112 VABF-viii Eco  550  re comercio de América del Norte. En la parte económica en esta reunión del grupo de los tres CONCRIC
Ho122 VAF-viii  Eco  343  reocupados con los subsidios  de la Comunidad Económica Europea y particularmente los quesos  holande
Ho122 VAF-viii  Eco  641  española para otros capitales de la Comunidad Económica  Europea. O sea que esto pudiera inclusive pe
Ho122 VAF-viii  Eco  932   Venezuela tenemos una economía y una cultura económica que por  años ha estado acostumbrado por un M
Ho122 VAF-viii  Eco 1199  Caldera conciba así la economía, la  política económica y Fernández tampoco. Aquí no se ha planteado
Ho122 VAF-viii  Eco 1228   más inteligentes como  economía, la política económica. Hoy se manejan en los países de  forma mucho
Ho122 VAF-viii  Eco 1363  ue por favor los responsables de la política  económica por la vía fiscal, por la vía comercial, por
Ho122 VAF-viii  Eco 1426   xxx el  gobierno ha fracasado en su política económica y el mismo tiene  que rectificar. Ahora, sí ha
Ho123 VABF-viii Eco  251  is. La crisis está originada por  la política económica que ha llevado este gobierno y se ve pues  el
Ho123 VABF-viii Eco  266  isis. La crisis está generada por la política económica de este  gobierno.  - El paro, por ejemplo, d
Ho123 VABF-viii Eco  482   es el pueblo, no es la gente, es la política económica  neoliberal. Enfrentémosla juntos. Vamos a en
Ho123 VABF-viii Eco 1049  nsideran  que su salario, en la circunstancia económica en que estamos  viviendo de una inflación gal
Ho124 VBC-viii  Eco  460  ado, un pasado aunque en esta nueva política  económica se tuvieron a esta gente muy controlada y  ar
```

Ejemplo 12. Listado de concordancias

Otra posibilidad, ya mencionada, es ampliar el contexto a la oración completa (*Ejemplo 13*; solamente las seis primeras ocurrencias). Esto puede resultar especialmente útil, por ejemplo, para estudios lexicográficos, ya que permite extraer y reconocer los diferentes significados de determinadas formas en contextos diversos.

```
                                                                económica    21
                                                 La economía de
Ho111 VAF-viii  Eco  126  mercado es, hoy por hoy, la única concepción  económica en el mundo no sólo en Venezuela ni en América
                                                                           Latina, en el mundo la única concepción
                                                                           económica que está impulsando las economías en
                                                                           el mundo.

                                              La economía de mercado
                         es, hoy por, la  única  concepción  económica
                         el  mundo  no sólo en Venezuela ni en América
Ho111 VAF-viii  Eco  127 Latina,  en  el  mundo  la  única  concepción económica que está impulsando las economías en el mundo.

                                    El mejor ejemplo de esto lo
                         tienen ustedes en China, en China que  es  un
Ho111 VAF-viii  Eco  130 país comunista que ha emprendido una  reforma económica a fondo y que está dentro de una estructura
                                                                           política muy rígida, es cierto, dentro de una
                                                                           estructura política muy rígida esa economía
                                                                           china está produciendo una transformación
                                                                           bárbara dentro precisamente centrada en la
                                                                           economía de mercado, en la liberación de
                                                                           mercados.

                                  Entonces, lo que nosotros recordamos
                         de que el paquete de la economía  de  mercado
                         es realmente las cosas  negativas  y  sabemos
                         que  no  es  así,  sabemos  que   simplemente
                         lo   que   se  ha...,  el  mal  entendimiento
                         que  existe  en   términos  de   economía
                         de mercado, está más relacionado con una mala
Ho111 VAF-viii  Eco  227 gestión  gubernamental  que  con la  posición económica propiamente dicha.

                                     Yo no entiendo por qué  Carlos
                         Rafael Silva dice que no cree en la  economía
Ho111 VAF-viii  Eco  411 de   mercado   es   simplemente  la   sanidad económica, por qué?, porque  economía de mercado
                                                                            significa que quien demande bienes es
                                                                            porque ha aportado al proceso productivo.

                                                 Te agradezco
                         la   pregunta  porque  ese  es  realmente  el
Ho111 VAF-viii  Eco  523 gran  problema  que confrontaría una política económica centrada en la  revaluación del bolívar.
                                                                        .
                                                                        .
                                                                        .
```

Ejemplo 13. Listado de concordancias (contexto ampliado a toda la oración)

El hecho de contar con información textual adicional en los listados de concordancias, amplía el abanico de posibilidades en cuanto al procesamiento de los datos lingüísticos, además de permitir ciertas sutilizas. Por ejemplo, ordenar las ocurrencias de cada forma, no sólo por orden de aparición, sino alfabéticamente, pero atendiendo a la primera letra de la palabra que sigue inmediatamente a la forma en cuestión (*Ejemplo 14*); u ordenar las ocurrencias según el último signo de

la palabra que precede a la forma alineada (orden alfabético reverso; *Ejemplo 15*).

```
                                                                económica    21
Ho111 VAF-viii  Eco  130  país comunista que ha emprendido una  reforma económica a fondo y que está dentro de una estructura
Ho111 VAF-viii  Eco  523   gran problema  que confrontaría una política económica centrada en la  revaluación del bolívar. Una
Ho123 VABF-viii Eco  266  isis. La crisis está generada por la política económica de este  gobierno.  - El paro, por ejemplo, d
HO112 VABF-viii Eco  533  ados del Caribe para fomentar  la integración económica de la región. Esta asociación tendría  como o
Ho111 VAF-viii  Eco  126  mercado es, hoy por hoy, la única concepción  económica en el mundo no sólo en Venezuela ni en mérica
HO112 VABF-viii Eco  550  re comercio de América del Norte. En la parte económica en esta reunión del grupo de los tres CONCRIC
Ho123 VABF-viii Eco 1049  nsideran  que su salario, en la circunstancia económica en que estamos  viviendo de una inflación gal
Ho122 VAF-viii  Eco  641  española para otros capitales de la Comunidad Económica  Europea. O sea que esto pudiera inclusive pe
Ho122 VAF-viii  Eco  343  reocupados con los subsidios  de la Comunidad Económica Europea y particularmente los quesos  holande
Ho122 VAF-viii  Eco 1228   más inteligentes como  economía, la política económica. Hoy se manejan en los países de  forma mucho
Ho123 VABF-viii Eco  482   es el pueblo, no es la gente, es la política económica  neoliberal. Enfrentémosla juntos. Vamos a en
Ho122 VAF-viii  Eco 1363  ue por favor los responsables de la política  económica por la vía fiscal, por la vía comercial, por
Ho111 VAF-viii  Eco  411  conomía de  mercado es simplemente la sanidad económica, por qué?, porque  economía de mercado signif
Ho111 VAF-viii  Eco  227  la gestión gubernamental que con la  posición económica propiamente dicha. Entonces, vamos a retomar
Ho111 VAF-viii  Eco  127  rica Latina,  en el mundo la única concepción económica que está impulsando las  economías en el mund
Ho123 VABF-viii Eco  251  is. La crisis está originada por  la política económica que ha llevado este gobierno y se ve pues  el
Ho122 VAF-viii  Eco  932   Venezuela tenemos una economía y una cultura económica que por  años ha estado acostumbrado por un M
Ho124 VBC-viii  Eco  460  ado, un pasado aunque en esta nueva política  económica se tuvieron a esta gente muy controlada y  ar
Ho122 VAF-viii  Eco 1426   xxx el  gobierno ha fracasado en su política económica y el mismo tiene que rectificar. Ahora, sí ha
Ho122 VAF-viii  Eco 1199  Caldera conciba así la economía, la  política económica y Fernández tampoco. Aquí no se ha planteado
HO112 VABF-viii Eco  534  dría  como objetivo avanzar en la integración económica y la  cooperación entre los países de la cuen
```

Ejemplo 14. Listado de concordancias (occurencias ordenadas alfabéticamente según la palabra que sigue)

Una rápida ojeada a estos datos revela el alto índice de preposiciones -*a, de, en* y *por*- , de la conjunción copulativa -*y*- y del nexo -*que*- que acompañan a la forma A*economía*@, entre otras posibles observaciones.

```
                                                                económica    21
Ho111 VAF-viii  Eco  523   gran problema  que confrontaría una política económica centrada en la  revaluación del bolívar. Una
Ho123 VABF-viii Eco  266  isis. La crisis está generada por la política económica de este  gobierno.  - El paro, por ejemplo, d
Ho122 VAF-viii  Eco 1228   más inteligentes como  economía, la política económica. Hoy se manejan en los países de  forma mucho
Ho123 VABF-viii Eco  482   es el pueblo, no es la gente, es la política económica  neoliberal. Enfrentémosla juntos. Vamos a en
Ho122 VAF-viii  Eco 1363  ue por favor los responsables de la política  económica por la vía fiscal, por la vía comercial, por
Ho123 VABF-viii Eco  251  is. La crisis está originada por  la política económica que ha llevado este gobierno y se ve pues  el
Ho124 VBC-viii  Eco  460  ado, un pasado aunque en esta nueva política  económica se tuvieron a esta gente muy controlada y  ar
Ho122 VAF-viii  Eco 1426   xxx el  gobierno ha fracasado en su política económica y el mismo tiene que rectificar. Ahora, sí ha
Ho122 VAF-viii  Eco 1199  Caldera conciba así la economía, la  política económica y Fernández tampoco. Aquí no se ha planteado
Ho123 VABF-viii Eco 1049  nsideran  que su salario, en la circunstancia económica en que estamos  viviendo de una inflación gal
Ho111 VAF-viii  Eco  130  país comunista que ha emprendido una  reforma económica a fondo y que está dentro de una estructura
Ho122 VAF-viii  Eco  932   Venezuela tenemos una economía y una cultura económica que por  años ha estado acostumbrado por un M
Ho111 VAF-viii  Eco  411  conomía de  mercado es simplemente la sanidad económica, por qué?, porque  economía de mercado signif
Ho122 VAF-viii  Eco  641  española para otros capitales de la Comunidad Económica  Europea. O sea que esto pudiera inclusive pe
Ho122 VAF-viii  Eco  343  reocupados con los subsidios  de la Comunidad Económica Europea y particularmente los quesos  holande
HO112 VABF-viii Eco  550  re comercio de América del Norte. En la parte económica en esta reunión del grupo de los tres CONCRIC
HO112 VABF-viii Eco  533  ados del Caribe para fomentar  la integración económica de la región. Esta asociación tendría  como o
HO112 VABF-viii Eco  534  dría  como objetivo avanzar en la integración económica y la  cooperación entre los países de la cuen
Ho111 VAF-viii  Eco  227  la gestión gubernamental que con la  posición económica propiamente dicha. Entonces, vamos a retomar
Ho111 VAF-viii  Eco  126  mercado es, hoy por hoy, la única concepción  económica en el mundo no sólo en Venezuela ni en mérica
Ho111 VAF-viii  Eco  127  rica Latina,  en el mundo la única concepción económica que está impulsando las  economías en el mund
```

Ejemplo 15. Listado de concordancias (ocurrencias ordenadas alfabéticamente -reverso- según la palabra que precede)

Es importante resaltar la claridad y sencillez con que esta organización permite reconocer los compuestos o locuciones más frecuentes de una forma. En este caso observamos el alto índice de ocurrencias de la secuencia A*política económica*@ (9 veces), seguida de otras como: *Comunidad Económica Europea, integración económica* y *concepción económica*, entre otras.

Todas las ejemplificaciones presentadas hasta ahora, son exclusivas de formas o palabras aisladas, es decir, selección y búsqueda de una única secuencia

específica de letras. Sin embargo, también es posible obtener resultados y datos sobre determinados grupos de palabras: sintagmas, locuciones, giros, correlaciones, etc. A continuación citamos ejemplos basados en combinados de dos formas o palabras.

```
                                                                algunas bibliotecas     1
Eo124 EAF-vii  Rel  101  toral, en aquel momento, y quedan ejemplares en  algunas bibliotecas, y él defiende allí, defiende muy clar

                                                                antropomórficas formas     1
Eo124 EAF-vii  Rel  692   generalmente nunca más de varias  horas, formas antropomórficas, formas animaloides, generalmente  aparece

                                                                aquellas cosas     1
Eo124 EAF-vii  Rel 1132  ún fenómeno sobrenatural, cuando  aparecen todas aquellas cosas que, efectivamente, no nos  explicamos con

                                                                cosas extraordinarias     1
Eo124 EAF-vii  Rel 1284  o en el Xxx y en Xxx.  (F) Yo he visto, digamos, cosas extraordinarias, digamos, la  Virgen me ha dejado ve

                                                                demoníacas lanzadas     1
Eo124 EAF-vii  Rel  727  nipulaciones, más o  menos malignas, más o menos demoníacas, lanzadas desde el bajo  astral para confundir

                                                                determinadas grutas     1
Eo124 EAF-vii  Rel  124  , estas  apariciones al lado de las encinas y en determinadas grutas, yo  soy más partidario que lo milagro

                                                                devotas activas     1
Eo124 EAF-vii  Rel  554  engo, además, seis personas que son todas ellas, devotas  activas, de lugares donde se dice que ha habido a

                                                                ellas devotas     1
Eo124 EAF-vii  Rel  554  ero,  tengo, además, seis personas que son todas ellas, devotas  activas, de lugares donde se dice que ha h

                                                                encinas mágicas     1
Eo124 EAF-vii  Rel  763  ros, en arroyos, en encinas, en esos bosques de  encinas mágicas a los que se refería Moncho Alpuente, en g

                                                                enrevesadas complicadas     1
Eo124 EAF-vii  Rel  961  simos pero, que las explicaciones eran  igual de enrevesadas, complicadas que eso. Me veía la señora Xxx,

                                                                esas fotografías     1
Eo124 EAF-vii  Rel  717   por allá. Las fotografías no son un fraude,  en esas fotografías no hay truco, han sido estudiadas. )Qué p

                                                                esas personas     2
Eo124 EAF-vii  Rel  881  las apariciones, y otra cosa es el hecho de que  esas personas que participan de ese fenómeno, sean gente
Eo124 EAF-vii  Rel 1377  ue, de alguna manera, han aportado su ayuda para esas  personas y, cómo no, también, para todos nosotros.
```

Ejemplo 16. Listado de concordancias (dos palabras consecutivas acabadas en A-as@)

Este ejemplo es un fragmento de un listado de concordancias mayor y da cuenta, por orden alfabético, de las ocurrencias de dos palabras consecutivas acabadas en A-*as*@. En este caso hemos decidido ignorar la coma, con el fin de obtener ejemplificaciones de posibles enumeraciones encontradas en el texto (la secuencia A*xxx*@ sustituye a palabras irreconocibles o inaudibles en las grabaciones). Este listado puede resultar útil para observar la concordancia gramatical entre artículos, adjetivos y nombres/adjetivos, y/o comprobar el orden sintáctico entre nombres y adjetivos calificativos, por ejemplo.

Otro caso de interés puede ser conocer los contextos de ocurrencia de determinadas locuciones o perífrasis verbales (por ejemplo, *tener que* + *infinitivo*) y estudiar posibles diferencias en su uso o en cuanto a su comportamiento sintáctico (*Ejemplo 17*).

```
                                                                tener que     4
Eo113 EABF-viii Eco  61   ha crecido,  por fin, la condición de que van a tener que vender alguno de sus  compañeros de viaje para p
Eo113 EABF-viii Eco 451  s no lo ha dicho así, dice, 'poco menos vamos a  tener que reunirnos un día, los que estamos en comisión y
Eo113 EABF-viii Eco 454  dos a comisiones e investigaciones, un día van a tener que  acordar, celebrar un pleno del Congreso para le
Eo124 EAF-vii   Rel 317  o voces, por aquí, a ver.  (C) Bueno, yo lamento tener que quebrar este diálogo entre  pastores y ovejas, q
```

Ejemplo 17. Listade de concordancias (A tener que@)

Podemos, también, combinar varios parámetros a la vez y extraer, por ejemplo, todas las ocurrencias que contienen adverbios acabados en A-*mente*@, seguidos del nexo A*que*@ y que aparecen en el lenguaje sectorial económico de Venezuela con una frecuencia de uso inferior a 3 (*Ejemplo 18*).

```
                                                                  aparentemente que     1
Ho123 VABF-viii  Eco  302  o para  todo el mundo menos para el gobierno, aparentemente que en  Venezuela venía un problema camb

                                                                  concretamente que     1
Ho112 VABF-viii  Eco  362  par de instituciones, no tres  instituciones, concretamente, que están pagando los intereses de  esa

                                                                  constantemente que     1
Ho123 VABF-viii  Eco  653  aro se necesita dinero y  el gobierno anuncia constantemente que no hay dinero. Mi pregunta  es )rea

                                                                  definitivamente que     1
Ho112 VABF-viii  Eco  450   de tasas de interés más  razonable.  - Mira, definitivamente que no, )no?. La ley de bancos, pues,

                                                                  eficientemente Que     1
Ho113 VBC-viii   Eco  476  amos a  explicar que es eso de que se asignen eficientemente. Que se  asignen eficientemente signifi

                                                                  honestamente que     1
Ho123 VABF-viii  Eco  216  uvimos a un gran líder,  que Dios lo bendiga, honestamente, que fue Lech Walesa, un hombre  que se l

                                                                  indudablemente que     1
Ho123 VABF-viii  Eco  750  y una respuesta  positiva en el día de mañana indudablemente que el magisterio no  le quedará otro r

                                                                  positivamente que     1
Ho113 VBC-viii   Eco  196  e, esto está grabado en  Venezuela, que yo sé positivamente, que hay firmas, que hay gente  que trab

                                                                  precisamente que     1
Ho111 VAF-viii   Eco   90   insistiendo, y esa es la tesis de mi libro,  precisamente, que acabamos de bautizar, mi libro Salid

                                                                  públicamente que     1
Ho112 VABF-viii  Eco   91  mpresario y banquero Orlando  Castro denunció públicamente que era necesario investigar el  entorno

                                                                  realmente que     2
Ho122 VAF-viii   Eco  870  e  responde a la xxx quiénes pueden averiguar realmente que los  procesos de costo son así y no son
Ho122 VAF-viii   Eco 1013   calle sobre el control de  precios?. )Desean realmente que se estabilicen los precios en el  país?.

                                                                  sencillamente que     1
Ho122 VAF-viii   Eco 1166  ..  - Hacia arriba.  - Pero pudiéramos pensar sencillamente que manteniendo ganancias  aceptables, q

                                                                  solamente que     1
Ho123 VABF-viii  Eco  449   defienden como  empleado, le dan el lomito y solamente que tienen el lomito para  cobrar sino que v

                                                                  tremendamente que     1
Ho123 VABF-viii  Eco  983  yo quiero expresar lo siguiente, me preocupa  tremendamente que la sociedad civil tenga un concepto
```

Ejemplo 18. Listado de concordancias (A-mente que@, Venezuela, economía, frec<3)

El análisis se puede extender no solamente a palabras consecutivas, sino también a partículas correlativas e interdependientes separadas por otras palabras. A menudo la distancia que aleja a los dos ítems lingüísticos no es predecible y puede variar en cada caso (por ejemplo, *por un lado ... por otro lado, tanto ... como*). El ejemplo que sigue corresponde a la correlación A*no* ... *ni*@, para los casos en que la separación máxima entre ambas partículas no excede de cuatro palabras.

```
                                                                  no ni     8
Eo113 EABF-viiiEco  101  les servidores del amo  y no de quien los elige, no tendremos aquí, ni democracia, ni  justicia, ni una eco
Eo113 EABF-viiiEco  578  ez años, no sólo no tenemos pensiones sino,  que no tendremos, ni siquiera, sueldo.  (D) Fernando, no seas
Eo113 EABF-viiiEco  958  yentes que lo han hecho posible  porque, si esto no hubiera funcionado, ni antena ni nada. O sea,  hay tres
Eo124 EAF-vii  Rel  215  los sitios de apariciones, que  en aquel momento no había videntes, ni había, digamos,  apariciones, sino q
Eo124 EAF-vii  Rel  319  o un diálogo muy interesante, sobre todo, cuando no has  sido ni pastor, ni oveja. Porque, uno ve el mundo
Eo124 EAF-vii  Rel  628  ro, que el espectador no quede desorientado. Eso no es,  ni como ha dicho el señor Sánchez Dragó, ni como h
Eo124 EAF-vii  Rel 1004  talmente inválida y unos  dolores espantosos que no se pueden ni contar, )no? Al momento  yo xxx las aparic
Eo124 EAF-vii  Rel 1009  aba curada, y hasta hoy. Yo llevaba plantillas,  no podía andar ni un paso, incluso me querían poner en un
```

Ejemplo 19. Listado de concordancias (Ano ... ni@, separación máxima 4 palabras)

Esta sencilla muestra revela, por ejemplo, que, a pesar de limitar la distancia máxima a cuatro palabras, en las ocho ocurrencias la separación no va en ningún caso más allá de dos palabras.

Listados de formas a partir de análisis de concordancias

Si el texto que contextualiza a una forma en un listado de concordancias es importante para estudiar el comportamiento y significado de dicha forma, igualmente relevante puede resultar procesar dicho contexto -como si se tratase de un texto corriente- y conocer qué formas son las que más frecuentemente ocurren en estos contextos. Es decir, qué asociaciones libres de palabras establecen los hablantes de una lengua cuando hacen uso de voces concretas para comunicarse. Estos datos pueden dar luz sobre, por ejemplo, qué formas configuran los campos semánticos asociados a una forma. Veámos el siguiente listado de frecuencias obtenido a partir de las concordancias de A*economía*@.

de	201	mundo	10	Hoy	5
economía	197	estado	9	presidente	5
la	159	este	9	puede	5
que	122	esto	9	son	5
en	110	nuestra	9	año	4
mercado	85	cómo	8	bienes	4
una	72	cree	8	cosas	4
el	62	Cuando	8	cual	4
es	52	estamos	8	dinero	4
no	48	las	8	diversificar	4
y	48	tenemos	8	económica	4
a	42	cosa	7	eso	4
con	35	creo	7	está	4
lo	26	si	7	gasto	4
se	22	su	7	mucho	4
por	21	va	7	Pérez	4
del	20	al	6	petrolera	4
un	19	donde	6	programa	4
los	18	ese	6	público	4
o	18	ha	6	Rafael	4
más	14	ninguna	6	Realmente	4
venezolana	14	nos	6	regulada	4
Pero	13	nosotros	6	rentista	4
Porque	13	país	6	sí	4
qué	13	sobre	6	Silva	4
tiene	13	ya	6	tener	4
hay	12	acuerdo	5	términos	4
para	12	años	5	uno	4
Yo	12	decir	5	vamos	4
esa	11	económico	5	Venezuela	4

precios	11	forma	5	vía	4
como	10	hacer	5		
Entonces	10	hemos	5		

Ejemplo 20. Listado de frecuencias (formas que ocurren en torno a la forma Aeconomía@, frec>3)

Este listado nos permite extraer algunos datos interesantes y, sobre todo, conocer las asociaciones de ideas y de palabras en torno a la forma: *economía* (sustantivos: *mercado, precios, acuerdo, presidente, bienes, dinero, gasto, programa, rentista, términos, vía*; verbos: *diversificar, regular* (*regulada*); adjetivos: *petrolera, público*; formas derivadas de *economía*: *económico y económica*). Por razones de espacio, hemos reducido la lista a las formas con una ocurrencia superior a tres. No obstante, los datos del resto del listado son igualmente relevantes y dejan constancia de otras muchas formas asociadas a *economía* (*ahorro, bolívares, petróleo, política, crecimiento, demanda, desarrollo, dólares, esquema, exportaciones, fabricante, gobierno, inflación, negocios, sector, tasa, tipos, agentes, bancos, cambio, competencia, costos, déficit, desabastecimiento, descuento, desestabilización, empobrecimiento, empresas, estabilización, hacienda, industria, ingreso,* etc.).

Etiquetadores morfológicos y analizadores sintácticos

La longevidad de un corpus depende -al margen de su constante ampliación y actualización- de la facilidad en poder reutilizar los materiales para otros proyectos. Un factor que puede contribuir en gran medida a ello es un buen etiquetado y una anotación completa. Un corpus anotado, además de contar con la codificación y organización interna antes aludidas, incorpora etiquetas que van desde sencillos símbolos para marcar rasgos fonéticos/fonémicos o prosódicos, hasta etiquetas morfológicas complejas con la categoría gramatical anexada a cada palabra, marcadores sintácticos (componentes estructurales) y, en algunos casos, hasta anotaciones semánticas, pragmáticas o discursivas.

El ejemplo que sigue (*Ejemplo 21*) muestra el texto modelo etiquetado. Como puede apreciarse, cada palabra aparece unida mediante un guión a su etiqueta morfológica. Las abreviaturas de las categorías gramaticales están bastante estandarizadas -la única duda puede surgir con *CNTR* (contracto de *a* + *el*, por ejemplo)-, pudiéndose modificar si así se desea.

```
El_ART Consejo_NP de_PREP Universidades_NP ha_VHABER dejado_PART claro_ADV que_CONJ
los_ART préstamos_N avalados_ADJ por_PREP el_ART Estado_NP para que_CONJ los_ART
estudiantes_N puedan_V cursar_VINF carreras_N universitarias_ADJ ,_PUNT son_VSER
incompatibles_ADJ con_PREP el_ART disfrute_N de_PREP cualquier_QUANT beca_N ._PUNT
```

```
Según_PREP este_DEM organismo_N ,_PUNT los_ART créditos_N se_PRON conciben_V
como_CONJ una_ART retribución_N a_PREP los_ART estudiantes_N con_PREP buenas_ADJ
calificaciones_N académicas_ADJ ,_PUNT pero_CONJ con_PREP un_ART nivel_N de_PREP
renta_N familiar_ADJ que_REL les_PRON impide_V acceder_VINF a_PREP una_ART beca_N
._PUNT
Por lo tanto_CONJ ,_PUNT se_PRON estima_V que_CONJ en_PREP el_ART periodo_N
comprendido_ADJ entre_PREP los_ART años_N 1995_NUM y_CONJ 2004_NUM ,_PUNT unos_ART
250000_NUM estudiantes_N pueden_V recibir_VINF un_ART crédito_N avalado_ADJ
por_PREP el_ART sector_N público_ADJ ._PUNT Con_PREP el_ART préstamo_N
concedido_ADJ ,_PUNT estas_DEM personas_N podrían_V disponer_VINF anualmente_ADV
de_PREP 500000_NUM pesetas_N para_PREP sufragar_VINF los_ART gastos_N que_REL
ocasiona_V estudiar_VINF una_ART carrera_N universitaria_ADJ ._PUNT Los_ART
mismos_ADJ prestatarios_N dispondrán_V de_PREP un_ART periodo_N de_PREP carencia_N
para_PREP devolver_VINF el_ART crédito_N ,_PUNT asociado_ADJ al_CNTR momento_N
en_PREP que_REL logren_V encontrar_VINF empleo_N ;_PUNT en_PREP este_DEM sentido_N
,_PUNT las_ART administraciones_N públicas_ADJ destinarán_V casi_ADV 43_NUM
millones_N de_PREP pesetas_N para_PREP la_ART subvención_N de_PREP intereses_N
y_CONJ avales_N ._PUNT
```

Ejemplo 21. Texto etiquetado morfológicamente

La posibilidad de disponer de un corpus etiquetado, multiplica el potencial de los resultados que podemos extraer. Por ejemplo, listados de todas las formas sustantivas, listados de concordancias con secuencias de dos adjetivos consecutivos, etc.

Otra herramienta complementaria y próxima a los etiquetadores, es el lematizador. Esta aplicación anexa a las diversas formas o palabras el lema del que derivan (*Ejemplo 22*; correspondiente al texto modelo).

```
a_PREP (a)                          impide_V (impedir)
académicas_ADJ (académico)          incompatibles_ADJ (incompatible)
acceder_VINF (acceder)              intereses_N (interés)
administraciones_N (administración) la_ART (la)
al_CNTR (al)                        las_ART (la)
anualmente_ADV (anualmente)         les_PRON (le)
años_N (año)                        logren_V (lograr)
asociado_ADJ (asociado)             los_ART (el)
avalado_ADJ (avalado)               millones_N (millón)
avalados_ADJ (avalado)              mismos_ADJ (mismo)
avales_N (aval)                     momento_N (momento)
beca_N (beca)                       nivel_N (nivel)
buenas_ADJ (bueno)                  ocasiona_V (ocasionar)
calificaciones_N (calificación)     organismo_N (organismo)
carencia_N (carencia)               para_PREP (para)
carrera_N (carrera)                 para que_CONJ (para que)
carreras_N (carrera)                periodo_N (periodo)
casi_ADV (casi)                     pero_CONJ (pero)
claro_ADV (claro)                   personas_N (persona)
como_CONJ (como)                    pesetas_N (peseta)
comprendido_ADJ (comprendido)       podrían_V (poder)
con_PREP (con)                      por_PREP (por)
concedido_ADJ (concedido)           Por lo tanto_CONJ (Por lo tanto)
conciben_V (concebir)               préstamo_N (préstamo)
Consejo_NP (Consejo)                préstamos_N (préstamo)
crédito_N (crédito)                 prestatarios_N (prestatario)
créditos_N (crédito)                públicas_ADJ (público)
cualquier_QUANT (cualquier)         público_ADJ (público)
```

```
cursar_VFIN (cursar)                puedan_V (poder)
de_PREP (de)                        pueden_V (poder)
dejado_PART (dejar)                 que_CONJ (que)
destinarán_V (destinar)             que_REL (que)
devolver_VINF (devolver)            recibir_VINF (recibir)
disfrute_N (disfrute)               renta_N (renta)
dispondrán_V (disponer)             retribución_N (retribución)
disponer_VINF (disponer)            se_PRON (se)
el_ART (el)                         sector_N (sector)
empleo_N (empleo)                   según_PREP (según)
en_PREP (en)                        sentido_N (sentido)
encontrar_VINF (encontrar)          son_VSER (ser)
entre_PREP (entre)                  subvención_N (subvención)
Estado_NP (Estado)                  sufragar_VINF (sufragar)
estas_DEM (este)                    un_ART (un)
este_DEM (este)                     una_ART (un)
estima_V (estimar)                  Universidades_NP (Universidad)
estudiantes_N (estudiante)          universitaria_ADJ (universitario)
estudiar_VINF (estudiar)            universitarias_ADJ (universitario)
familiar_ADJ (familiar)             unos_ART (uno)
gastos_N (gasto)                    y_CONJ (y)
ha_VHABER (haber)
```

Ejemplo 22. Listados de formas con etiquetas morfológicas y lemas

Tanto los etiquetadores como los lematizadores operan sobre un lexicón ya definido que contiene los lemas y sus correspondientes categorías gramaticales. Además, de alguna información de índole sintáctica y semántica necesarias para casos ambiguos. Mediante un proceso de reconocimiento de las formas irregulares, flexiones y enclíticos, entre otros, estas aplicaciones van reduciendo las secuencias de caracteres hasta dejarlas en su forma básica o lema, para luego pasarlas a cotejar e identificar en el lexicón.

Estas complejas herramientas lingüístico-computacionales (etiquetadores y lematizadores) permiten un análisis lingüístico superior -morfológico. Pero un corpus etiquetado tiene otra aplicación añadida: la posibilidad de ampliarlo y completarlo a otro nivel de anotaciones, el sintáctico (componentes estructurales). La secuenciación de palabras con su etiquetado morfológico permite agrupar a éstas mediante reglas sintácticas, hasta formar constituyentes sintácticos superiores -sintagmas, frases y oraciones. Así, por ejemplo, la secuencia `A`***El Consejo de Universidades***`@` **forma un sintagma nominal, constituído a partir de un artículo (*el*), un nombre (*Consejo*) y un sintagma preposicional (*de Universidades*), siendo éste último el resultado de anexar una preposición (*de*) a otro sintagma nominal (*Universidades*):**

O

SN SV

Art NP SP

Prep SN

NP

El Consejo de Universidades

Prolog **es un lenguaje de programación que resulta especialmente apropiado y útil para desarrollar gramáticas mediante las cuales analizar, representar y hasta generar secuencias lingüísticas. La representación sintáctica de los diversos componentes estructurales que constituyen la primera oración de nuestro texto modelo ha sido obtenida con una gramática prototipo desarrollado en Prolog. Por razones prácticas y para ajustarnos a la sintaxis de Prolog, han sido necesarias algunas modificaciones ortográficas (texto en minúsculas) y cambios en el etiquetado (etiquetas morfológicas precediendo a cada palabra y agrupación, mediante un guión, de las formas verbales compuestas).**

```
[o,[sn,[art,el],[np,consejo],[sp,[prep,de],[sn,[np,universidades]]]],[sv,[v,ha_dejado
],[adv,claro],[o1,[conj,que1],[o,[sn,[art,los],[n,prestamos],[adj,avalados],[sp,[prep
,por],[sn,[art,el],[np,estado]]],[o1,[conj,para_que],[o,[sn,[art,los],[n,estudiantes]
],[sv,[v,puedan_cursar],[sn,[n,carreras],[adj,universitarias]]]]]],[sv,[v,son],[adj,i
ncompatibles],[sp,[prep,con],[sn,[art,el],[n,disfrute],[sp,[prep,de],[sn,[quant,cualq
uier],[n,beca]]]]]]]]]
```

Ejemplo 23. Análisis sintáctico.

La presentación de estos datos sin más y con esta disposición, resulta bastante farragosa, sobre todo, para quienes no están familiarizados con el análisis de lenguas naturales, y, en particular, con DCG (*>Definite Clause Grammars'*) o PATR. Una solución bastante sencilla y práctica sería la de mostrar los diversos componentes estructurales de forma Asangrada@ según su nivel de dependencia -lo que no viene a ser más que una simplificación de las conocidas representaciones arbóreas.

```
[o,
     [sn,
          [art,el],
          [np,consejo],
          [sp,
               [prep,de],
               [sn,[np,universidades]]]],
     [sv,
          [v,ha_dejado],
          [adv,claro],
          [o1,
               [conj,que1],
               [o,
```

```
[sn,
     [art,los],
     [n,prestamos],
     [adj,avalados],
     [sp,
          [prep,por],
          [sn,
               [art,el],
               [np,estado]]],
     [o1,
          [conj,para_que],
          [o,
               [sn,
                    [art,los],
                    [n,estudiantes]],
               [sv,
                    [v,puedan_cursar],
                    [sn,
                         [n,carreras],
                         [adj,universitarias]]]]]],
[sv,
     [v,son],
     [adj,incompatibles],
     [sp,
          [prep,con],
          [sn,
               [art,el],
               [n,disfrute],
               [sp,
                    [prep,de],
                    [sn,
                         [quant,cualquier],
                         [n,beca]]]]]]]]]
```

Ejemplo 24. Análisis sintáctico (representación Asangrada@).

Gracias a las anotaciones sintácticas podemos extraer listados o concordancias, atendiendo a descriptores ya no solamente ortográficos o morfológicos, sino también sintácticos, y conocer todos los sintagmas nominales presentes en un determinado texto (*Ejemplo 25*). Estos datos pueden dar luz sobre la constitución y complejidad morfo-sintáctica de los diversos sintagmas (elementos constituyentes, nivel de Aanidamiento@ de los sintagmas, orden sintáctico, sustantivación, etc.).

```
[sn,[art,el],[np,consejo],[sp,[prep,de],[sn,[np,universidades]]]]

[sn,[np,universidades]]

[sn,[art,los],[n,prestamos],[adj,avalados],[sp,[prep,por],[sn,[art,el],[np,estado]]],
[o1,[conj,para_que],[o,[sn,[art,los],[n,estudiantes]],[sv,[v,puedan_cursar],[sn,[n,ca
rreras],[adj,universitarias]]]]]

[sn,[art,el],[np,estado]]

[sn,[art,los],[n,estudiantes]]
```

```
[sn,[n,carreras],[adj,universitarias]]

[sn,[art,el],[n,disfrute],[sp,[prep,de],[sn,[quant,cualquier],[n,beca]]]]

[sn,[quant,cualquier],[n,beca]]
```

Ejemplo 25. Sintagmas nominales (primera oración del texto modelo)

Conclusiones

Como hemos ido comprobando a lo largo de este capítulo, cada nivel de anotaciones (texto Allano@, texto con etiquetas morfológicas o con demarcaciones sintácticas) abre nuevas posibilidades y expectativas a la hora de extraer datos, siendo éstos cada vez más selectivos y complejos, y de mayor interés lingüístico. De ahí nuestra intención de presentar y ejemplificar las diversas técnicas para la computación de datos lingüísticos de forma escalonada, yendo de las más comunes y extentidas hacia las más sofisticadas (etiquetadores y lematizadores).

La característica más sobresaliente y que, a la vez, diferencia a la lingüística del corpus de las demás disciplinas lingüísticas o filológicas es su convergencia con la informática, creándose una dependencia interdisciplinaria entre ambas que desvían aquélla de los paradigmas y métodos de trabajo habituales hasta el presente. El carácter de innovación y acercamiento a los medios tecnológicos produjo un rechazo frontal en gran parte de la comunidad científica. Las consecuencias son evidentes, especialmente en España. Mientras que los tamaños de los corpus del español no superan los tres millones de palabras -a excepción de Cumbre con 8 millones de palabras en la actualidad- para otras lenguas se están creando corpus cuya magnitud oscila entre las decenas y las centenas de millones de palabras.

No olvidemos que estamos en los inicios de la lingüística del corpus. Es una disciplina en auge que no ha hecho más que despertar el interés académico, además del interés comercial. Pero quizás lo más negativo del interés comercial sea la aparición incesante de herramientas y productos informáticos, creados por personas ajenas a la materia. Dichos artículos no van más allá de simples rutinas de búsqueda y reconocimiento textual -que pueden encontrarse en cualquier procesador de texto potente-, siendo meras caricaturas de lo que en realidad es capaz de procesar un programa de concordancias. Una aplicación que no permita obtener gran parte de los datos presentados en este capítulo, difícilmente podrá satisfacer las necesidades de un investigador, y tendrá serios efectos contraproducentes: surgirá el desencanto, y la pérdida de confianza y credibilidad en la lingüística del corpus.

La virtud del ordenador para almacenar, buscar, clasificar, entresacar y en

general, manipular miles de datos lingüísticos a la vez, según los parámetros y patrones que le hayamos marcado de antemano, no hacen más que convertirlo en el mejor peón y aliado de un(a) lingüista. Tanto es así, que la Aindustria del lenguaje@ se está apresurando en desarrollar bases y recursos lexicográficos sobre los cuales compilar sus propios diccionarios, materiales didácticos, gramáticas, etc. La lingüística del corpus es una nueva forma y visión de hacer lingüística basada en el uso real de la lengua.) Existe mejor garante que éste para el trabajo lingüístico que quiera presumir de fiable y válido?

> A*... a linguistic description which is not supported by the evidence of the language has no credibility.@* **(Sinclair 1991:36)**

CAPITULO IV

EL CORPUS "CUMBRE" Y LA LEXICOGRAFÍA

El primer proyecto de corpus, el SEU, se proponía el estudio y análisis gramatical, más que la elaboración de un diccionario. Por el contrario, el corpus "Cobuild" fue aplicado en primera instancia y aprovechado comercialmente para la publicación de un diccionario. El repertorio de Luis Fernando Lara (1982), referido al español de México, también ha sido aprovechado para la publicación de un *Diccionario Fundamental del Español de México.* Realmente trabajar con un caudal de palabras tan grande como el que proporciona un corpus hace pensar de inmediato en las obras que históricamente han tenido como objetivo la definición del significado de las palabras. Y aunque es verdad que este tipo de aplicación lingüística no es el único, sí que constituye uno de los principales. Frente a la recopilación manual y personal de citas (que equivale a la recopilación de palabras en contexto), el ordenador ofrece una insuperable ventaja, tanto en extensión como en velocidad. Además, el significado de la palabra que aparece en las concordancias es quizás lo más sobresaliente y lo que más llama la atención, ayudando a ello la misma presentación formal de los listados, en los cuales siempre sobresale una voz dentro de cada contexto. No es de extrañar, por tanto, que frecuentemente el corpus se asocie de manera casi automática al léxico o a los diccionarios.

Los listados de frecuencias constituyen uno de los "frutos lexicográficos" del corpus más accesibles, sin que su obtención suponga en la actualidad dificultad alguna: pueden obtenerse en pocos minutos, aunque el tamaño del corpus sea de varios millones de palabras. Un corpus ideal, que abarcase la totalidad de la lengua en un determinado período de tiempo, nos proporcionaría un completo listado de palabras, exactamente las palabras o voces que servirían de base para un diccionario actualizado en cualquier época concreta. Para algunos, este corpus ideal no es posible, ni lo será en un futuro próximo. Sin embargo, listar las voces usadas mediante un ordenador sería viable siempre que nos refiriéramos a un período concreto, desde un punto inicial en el pasado hasta un punto terminal en el presente, restringiéndonos a la modalidad escrita y siempre que el computador pudiese procesar la enorme cantidad de millones de palabras a que daría lugar la recopilación de todo el material escrito de una lengua. La tarea sería factible, pero quizás no rentable en cuanto a tiempo y dinero. Lo que sí estaría al alcance de nuestras posibilidades actuales sería fundir en uno todos los diccionarios elaborados hasta el momento y obtener un listado de voces diferentes... De todos modos, no tendría mucho

sentido detenernos en consideraciones utópicas. Sin embargo cabe preguntarnos:) Acaso los diccionarios no vienen a ser como un libro que pretende encerrar en sí toda la riqueza léxica de una lengua?) Acaso todos los diccionarios de un idioma no podrían ser razonablemente representativos de las voces y acepciones de ese mismo idioma? Aún más:) Acaso los diccionarios no se han hecho recopilando el uso año tras año, década tras década, siglo tras siglo?) Y acaso no se ha consolidado como el mejor sistema para elaborar un diccionario la práctica de anotar en fichas citas, frases, etc. como método para ilustrar el significado de las palabras? Si esto es así, el corpus, con la ayuda de un ordenador, ofrece todo lo anterior, sin menoscabo de sus ventajas, pero incrementando notablemente éstas en lo relativo a tiempo, esfuerzo personal, fiabilidad y representatividad de los resultados y extensión de las fuentes a que podemos tener acceso.

Los listados de frecuencias proporcionados por el corpus son una referencia ideal para el trabajo de los lexicógrafos; no solamente son útiles para tomar decisiones sobre qué voces incluir en un momento determinado de la historia de una lengua, sino también para determinar la importancia de una voz frente a otra y sobre todo de una acepción frente a otras pertenecientes a la misma voz. En este campo un corpus obligaría a hacer numerosas correcciones en los diccionarios al uso, si decidiésemos que el orden de aparición de las acepciones se ajustase, por ejemplo, al orden de frecuencia de éstas, de mayor a menor o viceversa.

Las frecuencias de las palabras pueden sorprender al investigador. Obsérvese la rápida disminución de frecuencias a partir de las siete primeras palabras (*de, la, que, y, el, en, a*), entre las 38 más frecuentes de nuestro corpus:

de	385.255	es	66.280
la	250.306	por	65.553
que	242.689	con	62.224
y	183.837	una	61.929
el	183.536	lo	51.241
en	169.117	para	44.212
a	167.112	su	41.665
los	104.273	al	37.894
se	89.690	como	33.045
no	82.947	más	30.916
un	70.503	o	27.852
las	68.019	pero	26.445
del	66.814	me	25.670

le	22.210
si	20.586
ha	18.717
sus	17.775
ya	16.076
porque	15.516
yo	15.423
muy	14.448
este	14.356
hay	13.490
todo	13.452
está	13.071

También pueden interesar las diferencias que se detectan entre el uso lingüístico total (oral y escrito) y sólo el uso oral. En el Apéndice I se ofrece un cuadro ilustrativo de las voces correspondientes a parte de la letra "A" (de "a" a "ac") en el total del corpus (en la columna de la izquierda) y en la muestra del español oral de España (en la columna de la derecha).

En los listados de un corpus limitado a varios millones de palabras no es probable que aparezcan todas las voces contenidas en un gran diccionario de la lengua. El hecho no constituye ninguna deficiencia: cabría incluso preguntarse qué relieve ha de tener en un diccionario una palabra o acepción que no es usada a lo largo de un corpus de 8, 15, 20 ó 100 millones de palabras, por ejemplo. La respuesta invita a la reflexión, especialmente en el caso del lexicógrafo. Un diccionario que busque economizar espacio deberá dejar de lado tales voces. Un diccionario que pretenda ser reflejo del uso actual ha de cuestionarse, cuando menos, la inclusión de tales palabras. No puede negarse que la rentabilidad comunicativa de una palabra de frecuencia "0" en un corpus extenso es precaria o, en el mejor de los casos, ínfima. Tales voces pertenecen más a un diccionario con connotaciones de tipo histórico o a repertorios léxicos dilatados en el tiempo y en los usos pasados del idioma. Debe tenerse, además, en cuenta que las palabras de frecuencia escasa no abundan en acepciones diversas; por el contrario, suelen restringirse casi siempre a sólo un significado. De hecho, si los significados fuesen más numerosos, este mismo hecho ya llevaría consigo una mayor frecuencia de uso. Lo cual equivale a decir que la alta funcionalidad comunicativa de una voz incide en proporción directa en una mayor variedad de matices de significado y, en consecuencia, en una mayor frecuencia de uso.

Los listados de frecuencias también pueden ser parciales. Si el corpus está

diseñado para estos fines, es posible obtener listados léxicos limitados a determinadas áreas del saber, a determinados ámbitos geográficos, a determinados estratos sociales, etc. La utilidad de estos resultados es evidente para el lexicógrafo interesado en el uso especial del idioma (medicina, ciencia en general, política, religión, etc.).

De igual manera, los listados cobran, con la ayuda del ordenador, una "flexibilidad" extraordinaria: es posible obtenerlos en orden alfabético, en orden alfabético inverso (por el final de cada palabra), por tramos de frecuencia, por sufijos o prefijos determinados, etc.

Existen muchos diccionarios que tratan de recoger las voces propias de los países hispanoamericanos en general, o de algún país en particular. Las carencias de tales obras son patentes, no solamente porque una obra escrita queda pronto obsoleta en cuanto a las novedades que genera el idioma en cada país o región y en campos variados, sino porque es extremadamente difícil reunir en un libro las diferencias existentes en tan amplia extensión geográfica y demográfica con los métodos lexicográficos habitualmente puestos en práctica. En estas mismas obras es más que frecuente la presencia de voces y acepciones cuya diferenciación frente al uso de la lengua en otras áreas geográficas ofrece serias dudas. Este tipo de problemas encontrará en el corpus la solución más adecuada: el cruce de listados hace posible la fijación de las diferencias con precisión y rapidez, mientras que la consulta de cada palabra en su contexto facilitará al lexicógrafo la definición de las voces seleccionadas en razón de tales diferencias. Este proceso es fácilmente extensible a áreas geográficas amplias, a países en relación con un conjunto, a países en relación con grupos de países, etc.

También en este campo pueden surgir "sorpresas" relativas al uso. Es bien conocido el significado de "coger" en varios países de habla hispana. En nuestro corpus, las diferentes formas del verbo "coger" aparecen poco más de 100 veces en los textos hispanoamericanos (83 veces en la variedad escrita y 29 veces en lavariedad oral). La cifra sorprende por su escasez. Pero todavía es más sorprendente que en todos esos ejemplos nunca se ilustra el significado relativo al acto sexual, tan peculiar y propio del uso en varios países de Hispanoamérica. Precisamente esta escasa frecuencia es quizás lo que permite inferir que se trata de una palabra altamente tabú.

La voz "carro" substituye a la voz "coche" en el uso hispanoamericano. Así se lee en muchos manuales y gramáticas. La muestra de nuestro corpus avala esta afirmación pero con matices: "coche(s)" aparece en 1.091 ocasiones, "carro(s)" solamene es usada 249 veces. De tales ocurrencias, "carro(s)", en la variedad hispanoamericana, se usa 135 veces con el significado general de "coche", mientras que en el ámbito de uso de España aparece 114 veces, con significados de otra índole. A ello ha de añadirse que esta voz aparece en el uso hispanoamericano con una frecuencia dos veces mayor en el el lenguaje

oral (86 veces) que en el escrito (49). En contraste con lo anterior, la voz "coche(s) aparece solamente 57 veces en el uso hispanoamericano, puntualizando que de ellas 36 ocurrencias se dan en el lenguaje oral.

La voz "lindo/a", muy propia del uso hispanoamericano, demuestra también claramente que su incidencia en el uso oral y escrito de España es escasa (31 y 27 veces, respectivamente, a menudo en la expresión "de lo lindo"), mientras que en la lengua oral de Hispanoamérica goza de mayores preferencias, ya que aparece en 168 ocasiones (a pesar de que la muestra de la variedad hispanoamericana es menor en cantidad que la relativa al español de España). El uso justifica, en este caso, las apreciaciones o afirmaciones de los especialistas.

Consultando un diccionario, queda a menudo la impresión de que en él los significados de las palabras son autónomos e independientes, como inherentes a esa misma palabra. Las explicaciones relacionadas con el contexto se añaden sólo en ocasiones y las relacionadas con la estructura en la cual se dan tales significados o valores son todavía más escasas. La "conciencia lingüística" de los hablantes es, por lo general, poco explícita. Los hablantes no suelen saber o conocer las definiciones de las palabras tal cual aquéllas aparecen en los diccionarios. Pero a pesar de ello, saben usar tales palabras adecuadamente. Por otra parte, los lexicógrafos llegan a enunciar definiciones a partir del uso que registran para cada una de las voces definidas. Aunque parezca una obviedad, conviene recordar que las palabras no se definen primero y luego se usan de acuerdo con la definición establecida. El proceso es el inverso. En muchas ocasiones la palabra sigue también al concepto formado sobre algo o al utensilio que se crea o inventa para determinadas funciones. De la constatación de tal realidad sale reforzada la importancia del contexto.

El corpus tiene la posibilidad de suplir con creces carencias de índole contextual. Las concordancias generadas por el ordenador delimitan con precisión el significado porque éste siempre aparece en su contexto de uso. Al mismo tiempo las concordancias ofrecen también el ejemplo que puede ayudar a ilustrar el significado de la voz en cuestión. No debería olvidarse que en un alto porcentaje de casos, especialmente los más relacionados con el uso diario de las palabras, el significado exacto de una voz está íntimamente ligado al contexto, de manera que la ausencia de éste sería decisiva para provocar ambigüedad o indefinición. La palabra "hoja" queda perfectamente definida en cada uno de sus dos posibles significados en contextos como

"*Las hojas caían de los árboles presagiando la cercanía de la primavera*" o
"*Gastó más de mil hojas en el borrador de su novela*".

De igual manera es el contexto el que define el valor de "cociendo" en frases como:

"Hay una máquina monstruosa cociendo asfalto y una guardia permanente de fuego".

"... o es tonto, porque se estaba cociendo debajo de usted y no se enteró".

) Hasta dónde llega el poder decisivo del contexto para definir el significado real que queremos dar a las palabras? Aunque sea difícil determinar el tope, no cabe duda de que su función desambiguadora es capital. Es obligado afirmar que, en el peor de los casos, si el significado explicitado en un diccionario debe responder a la realidad del uso, el corpus es una garantía para que el lexicógrafo no se desvíe de tal realidad.

Estrechamente ligadas a la utilidad de los listados de frecuencias están las concordancias obtenidas para el estudio y análisis de objetivos especiales. En este campo, tanto la potencia de los ordenadores (que actualmente ya es más que suficiente para esta finalidad) como la intuición del lingüista son factores decisivos para la obtención de resultados adecuados y útiles.

Si nos centramos en el análisis de las frecuencias y concordancias de "tener", es importante conocer no solamente las veces que aparece este verbo, sino también la frecuencia de sus diferentes formas verbales, especialmente las irregulares:

En nuestro corpus, del total de 38.090 veces que aparece "tener" (en sus diferentes formas y flexiones), "tenemos" se encuentra 2.946 veces, mientras "tenéis" sólo 186 y "teníais" o "tuviérais" solamente se utilizan tres vez; "tengo" la encontramos en 3.201 ocasiones, "tenía" en 3.612, "tenías" sólo en 102, y las formas compuestas ("... tenido") en 1.453 ocasiones. Sobresale de manera muy especial "tiene", que aparece en 10.284 oraciones, "tienen" en 3.789 y "tienes" en 1.323. En consecuencia, el uso de las formas de presente de indicativo desplaza numéricamente a todas las demás constituyendo por sí solas el 57% (21.729) (aunque sin olvidar que un 27% corresponde a "tiene").

Analizados los resultados desde otra perspectiva, las flexiones de "tener" con irregularidad en la vocal radical (e > ie) suman ellas solas 15.396, es decir, un 40% del total de las formas de "tener". Las irregularidades del futuro ("tendré", etc.) suman 1.172 veces, la forma "tengo" aparece en 3.201 ocasiones y las equivalentes del subjuntivo presente ("tenga", "tengas"...) 1.632. Frente a estos números, las formas regulares de presente e imperfecto de indicativo (incluido el infinitivo, participio pasado en tiempos compuestos y gerundio), suman 13.420 veces. En conjunto, las formas irregulares tienen un mayor protagonismo que las regulares. La relación entre uso y cambios morfológicos puede quizás ilustrarse en otros muchos ejemplos. En tal caso un corpus podría muy bien avalar datos históricos sobre la evolución de una lengua o sobre cambios determinados

que se han dado en ella.) Sería posible relacionar el uso más o menos frecuente con la evolución más o menos intensa de las formas?

Resultados de interés sobre los usos de las diferentes formas pueden obtenerse también del análisis de frecuencias relativas a los verbos "ser", "estar" y "tener" , como se puede apreciar en la siguiente tabla:

TENER	frec.	ESTAR	frec.	SER	frec.
tiene	10284	está	13067	es	66036
tienen	3789	están	5056	son	12542
tener	3633	estaba	4888	ser	10750
tenía	3612	estado	4524	era	10528
tengo	3201	estamos	2891	fue	7906
Tenemos	2946	estoy	2728	sea	5549
tenido	1453	estar	2703	sido	4442
tienes	1323	estaban	1264	fuera	2545
tuvo	1233	estás	1039	eran	2075
tenga	918	esté	794	será	2067
tenían	771	estuvo	754	soy	1974
tendrá	589	estábamos	397	sería	1877
tuve	455	estará	391	fueron	1816
tengan	449	estén	360	siendo	1224
tendría	434	estaría	286	somos	973
teniendo	423	estuviera	281	sean	902
tuvieron	393	estuve	258	eres	860
tuviera	308	estuvieron	190	fui	572
teníamos	291	estarán	137	serán	569
tendrán	239	estáis	132	fuese	345
tenéis	186	estabas	115	fueran	301
Tendremos	186	estuvimos	111	fuimos	250
tuvimos	163	estaremos	87	serían	247
tengamos	144	estuvieran	81	sois	184
tendrían	114	estarían	75	seas	146
tengas	106	estemos	69	era	142
tenías	102	estaré	62	serlo	140
tendríamos	98	estuviese	50	fuesen	85
tuvieran	85	estaríamos	48	eras	76
Tendrás	83	estuviste	37	seamos	76
ten	68	estarás	26	fuiste	67
tendré	64	estuviéramos	21	seré	54
tuviese	59	estéis	16	fueras	51
tuviste	34	estuviesen	14	fuere	43
tuviéramos	30	estuvieras	12	fuéramos	42
tendrías	27	estabais	9	serás	34
tuvieras	15	estaréis	8	seremos	27
tengáis	14	estarías	8	seríamos	21
tened	14	estuviésemos	7	éramos	18
tendréis	11	estáte	6	fuésemos	17
tuviesen	11	estuvieses	4	seáis	14
tuviésemos	7	estad	3	serías	12
tendríais	5	estuvisteis	1	erais	12
tuvisteis	4	estábais	1	fueses	10
tuviere	4	estaríais	1	seréis	8
teníais	3	estaos	1	érase	5
tuviérais	3	estuvierais	1	fueren	4
tuvieses	2	estuvieren	1	fuerais	4
tuvieren	2			fuisteis	2
tengámoslo	1			fueseis	2
tuvieseis	1			seríais	1

A este estudio cabe todavía añadir otros datos de interés relacionados con el contexto en que aparece cada una de las formas:

El sintagma "tener que" (en sus diversas formas) aparece en 7.035 ocasiones, lo cual equivale a decir que de todas las oraciones en que aparece "tener", el 18,5% cobra el valor que le añade la adición de "que" (expresión de obligatoriedad).

Otros sintagmas presentan frecuencias muy inferiores, como "tener en cuenta". El uso de "tener" con complemento directo sobresale también en frecuencia sobre otras formaciones sintácticas.

Habituados a valernos del idioma cada día, cada hora, cada momento de nuestras vidas, lo más normal es que los hechos obvios pasen desapercibidos. El corpus es un instrumento excelente para enfrentarnos a este problema, precisamente porque nos obliga a enfrentarnos a datos objetivos que de otra manera escaparían a nuestra atención inmediata.

Los usos de "tener" con matices especiales en cuanto al significado, quedan ilustrados fehacientemente en múltiples ocasiones:

... tenían su siestecita...

... tengo acceso ...

... tengo ánimo de variar...

...) Tengo bien la pastilla debajo de los labios?

... hay una que tengo empezada desde hace más de 2 años...

etc.

El régimen preposicional de los verbos, nombres o adjetivos carece aún de una investigación adecuada y profunda en español, como también ocurre con la modalidad verbal. El corpus constituye el instrumento ideal y necesario para este tipo de estudios. En listados como el que sigue a continuación el investigador tiene a su disposición información muy completa sobre los usos de "depender + de", o de "contar con", por ejemplo:

depende de 42

EO121	E-x	Pol	ta y cinco, por ejemplo?. B Yo no sé situarlo porque	depende de	cómo evolucionen los acontecimientos, pero un panor
EO123	E-x	Otr	quiera, digo yo o tú cómo lo ves. H Yo creo que sí,	depende de	cómo te cases. Si te casas por la iglesia pensarás
EO114	E-x	Otr	firiendo al cutis. B Eso es el vello. C No, oye, y,	depende de	donde sea el vello, porque me estáis diciendo... B
EO112	E-i	Edu	tificadas, que realmente su mejora en la sociedad no	depende de	ellos mismos, que depende de otros, que la única ví
EO114	E-i	Cul	del Rey. Esto, reunirlos en el Hospital Távera, que	depende de	la Fundación Medinaceli, por vías más o menos de co
EO121	E-x	Otr	leyes dicen, la responsabilidad política es algo que	depende de	la interpretación de las actuaciones de cada cual,
Ho113	MC-	Soc	uado y con bolas de leche, qué le recomienda? A No,	depende de	la leche que le esté dando. Yo le recomiendo mucho
HO003	MJ	C-B	aico y caemos en los momentos xxx, en la vulgaridad,	depende de	la manera que tratemos los elementos. Si nosotros q
EO123	E-x	Otr	nidad, a mí me ha ido muy bien con la primera. M Eso	depende de	la mujer. N Yo llevo veintitrés años casado y me va
EO124	E-x	Pol	petía en distintas películas con cierto ingenio, no?	depende de	la película, de aventuras, que ocurrían en las mont
EO124	E-x	Pol	elícula, de aventuras, que ocurrían en las montañas,	depende de	la película, si urbana. Entonces era difícil, a pri
EO121	E-x	Otr	de la responsabilidad de los estados y luego también	depende de	la responsabilidad de la Unión Europea. Tiene usted
EO111	E-v	Eco	Bueno, la verdad es que yo tendría que ... A Claro,	depende de	la situación. B ... tendría que conocer la regulaci
Ho124	V-v	Eco	dónde van a estar los mayores índices de rendimiento	depende de	la situación de cada país, si ha habido dinero, si
EO114	E-v	Eco	claro, es por el dinero, más o menos, porque, claro,	depende de	las dos pensiones muy pequeñas y, no sé. A Es que,
EO114	E-v	Eco	as pensiones que las personas mayores tenga, eh? Eso	depende de	las necesidades que tenga en ese momento. In situ,
HO122	MBF	Pol	fases en este asunto de la basura ... C O cuatro. D	Depende de	las que queramos. La recolección, la transportación
EO113	E-i	Edu	consumo externo, claro que, son alumnos xxx del o ,	depende de	las sedes y los sitios, son gentes de la empresa, f
EO111	E-i	MeA	uien es comisionista, no estamos diciendo nada malo,	depende de	lo que cobre por esa comisión. La economía se mueve
Ho124	V-v	Eco	ciero mucho más desarrollado si lo hay, pero esto ya	depende de	lo que usted quiera. A Y aquella gente simplemente
EO124	E-x	Pol	Dudar. K Pero, él, bueno, o sea, en los matrimonios	depende de	los acuerdos, quiero decir, él era muy libre de hac
EO114	E-i	Cul	bién, para vidas y personas y, es decir, pero eso no	depende de	los españoles ni de los franceses ni de los inglese

etc.

contar con 13

EO114	E-i	Soc	ilde.Grabación nueva A Para eso tenemos la suerte de	contar con	compañeros en todos los centros de Radio Nacional d
EO111	E-i	MeA	o la fortuna desde mil novecientos ochenta y ocho de	contar con	el respaldo de prácticamente todas las fuerzas polí
HO113	MAF	Cul	colega, es cierto. A mí, lo que me gustaría es poder	contar con	el tiempo para ver, por un lado las zonas de contro
EO111	E-i	MeA	sunción lo que dijo en su momento de que le gustaría	contar con	él pero que naturalmente tenía que respetar una dec
HO121	MAF	Pol	én. A Qué posibilidades habrá, señor, de que podamos	contar con	esto dentro de poco tiempo?. D Bueno, yo creo que
Ho114	MF-	Pol	se estuvieron meses, yo creo que es suficiente para	contar con	estos comprobantes. B Bien, Salvador, cómo las pas
EO113	E-v	Rel	Inglaterra no tenía autoridad para dar este paso sin	contar con	la aprobación de Roma. Una segunda razón, recuerdan
EO111	E-i	MeA	a del partido, pues, siempre es un orgullo, verdad?,	contar con	la confianza de los compañeros de partido, pero no
EO111	E-i	MeA	de tu vida, en realidad, son muy pocos que se pueden	contar con	los dedos de una mano y quizá sobre alguno. B Ese y
HO123	MF-	Cul	cias al profesor Villegas, gracias a usted. Esperamos	contar	con su presencia en la próxima ocasión. Hasta entonces.
EO111	E-i	MeA	e comprometió a estar aquí el mes que viene. Vamos a	contar con	su presencia, Sr Alvero? A Con mucho gusto, sí. B
HO114	MC-	Soc	ida política de este conflicto es indispensable para	contar con	un clima político favorable, durante el año y sólo
HO124	MC-	Cul	iquera, pero sin grupo, pero en el futuro, esperamos	contar con	un grupo que nos apoye, no? Ahorita estoy muy conte

EO124	E-x	Pol	dado. E No, eso es absolutamente falso. Yo tenía una	cuenta con	despacho xxx de la Concha xxx una cuenta absolutame
Ho123	V-v	Eco	. Según el directivo de CANTEBE, dicho contrato, que	cuenta con	el aval de la federación de trabajadores de telecom
EO113	E-v	Eco	una cuenta ya no secreta, simplemente que tenía una	cuenta con	el señor de la Concha, si no recordaba, tampoco que
EO111	E-v	Eco	n. Ahora bien, que el Sr Martín no va a cenar por su	cuenta con	el Sr Conde. B Pero a mí, que cenen me parece bien
HO111	MAF	Otr	hasta qué punto contamos con la iglesia y la iglesia	cuenta con	nosotros, como ciudadanos, incluso, este tema, que
EO113	E-v	CiH	por un pequeño lapsus de tiempo. Por eso, la Iglesia	cuenta con	toda una batería de elementos que son los sacrament

EO114 E-x Otr pasa es que hay un miedo exacerbado. C Y por qué no cuentan con el apoyo de sus partidos?, y por qué no cuentan co
EO114 E-x Otr cuentan con el apoyo de sus partidos?, y por qué no cuentan con el apoyo de sus partidos si tan enterados están?
HO121 MAF Pol de cuarenta millones de ciudadanos mejicanos que ya cuentan con su credencial de elector podrán acudir a las urnas
HO122 MBC Otr ento en la productividad nacional. Instituciones que cuentan con un presupuesto estatal controlado por la secretarí
etc.

La premodificación y posmodificación del nombre por el adjetivo nos ofrece otro ejemplo ideal para el estudio de la sintaxis y del significado: el corpus puede proporcionarnos las secuencias de concordancias nombre-adjetivo o adjetivo-nombre con gran facilidad, poniendo a nuestro servicio el volumen necesario de datos, debidamente ordenados, para proceder a un análisis adecuado.

agencias publicitarias 1
HE006 MEX ANT;A 1797 vernáculo o en los departamentos creativos de las agencias publicitarias. Palinuro de México es, entre otras

agrias protestas 1
HE006 MEX ANT;A 4416 r y la amistad, la de la ira resignada, la de las agrias protestas soterradas; sus versos son las uvas exacta

aguadas amargas 1
HE013 HE CUE;C 312 ontrabandista. Los campos estaban empastados; las aguadas, amargas; el Inglés, para corregir esas deficienci

aguas desatadas 1
He12 HIS POE; 2091 obo; en mi costado lates y tu latir me anega: las aguas desatadas del bautismo remoto mi sueño mojan, nombra

aguas desnudas 1
He12 HIS POE; 2069 y desnuda, piedra. lV Vivimos sepultados en tus aguas desnudas, noche, gran marejada, vapor o lengua lenta

aguas estancadas 1
He112 NHI FIC; 1449 ostro al tuyo, por qué, como tú, lo hundes en las aguas estancadas y te repite, ahora que no la escuchas, Me

aguas hondas 1
He113 He NBI;N 1424 ñador, salteador de caravanas, cantor, catador de aguas hondas y de metales. Padecí cautiverio durante un añ

aguas horas 1
He12 HIS POE; 2048 tu corazón, alza tus pechos, y arrastra entre sus aguas horas, memorias, días, despojos de Li misma. Entre r

aguas mudas 1
He12 HIS POE; 2021 as entregadas. Viento parado en una apenas rama; aguas mudas, sonámbulas, sin freno; tierra henchida soñand

aguas negras 2
HE031 VEN PRN;P 7849 an otras y hasta trataron de inundar el lugar con aguas negras. Hoy se apretujaban en la colina miles de cru
HE031 VEN PRN;P 2086 an otras y hasta trataron de inundar el lugar con aguas negras. Hoy se apretujaban en la colina miles de cr

aguas tibias 1
HE013 HE CUE;C 1574 o de la ducha, Joaquín Bermejo decidió dejarse de aguas tibias, y empezó a cerrar el caño de agua caliente m

aguas turbias 1
He112 NHI FIC; 1304 eléctrica que a mí me mataría. Necesito navegar en aguas turbias, comunicarme a largas distancias, repelar a

alabadas barbas 1
He12 HIS POE; 2722 s, funcionan los labriegos a tiro de neblina, con alabadas barbas, pie práctico y reginas sinceras de los va

alas blancas 1
He112 NHI FIC; 2258 arde, chasqueé al porteño. Al punto la garza verde alas blancas de veinte metros de eslora cubrió las setenta

Es muy difícil que un diccionario pueda recoger sistemáticamente la dependencia que se da entre el significado y el contexto estructural de una voz. También el corpus constituye en este caso una ayuda ideal, como puede apreciarse en estos ejemplos, que ilustran eficazmente no sólo el significado de "**(en) contacto con**", sino también el contexto sintáctico que precede y sigue:

Ho111 MAF Cul Media. Hasta dónde los indígenas en estos ir a tener contacto con distintas tecnologías, con distintos grupos están
EO111 E-i MeA esa de Alba. Me lo ha proporcionado. Me ha puesto en contacto con el Club Antares y allí, a través de Don Ignacio R
EO113 E-i Edu nos se matriculan en cursos de verano para tener ese contacto con el departamento, con el conjunto de profesores qu
EO112 E-v Rel os dicho subjetivo hemos añadido luego la pérdida de contacto con el mundo de afuera, es decir, llega una persona x
HO122 MBC Otr us actividades al presidente manteniendo un estrecho contacto con el y con todo su gabinete. Esta ley también señal
EO111 E-v Eco no sólo del gobierno, del régimen que han estado en contacto con él. B De ahí la doctrina de la salida digna. A N
Ho112 MBF Pol e demos respuesta a su situación o podamos entrar en contacto con él, para recomendarle algún tipo de asesoría prof
EO113 E-v CiH aje muy bonito a El Escorial, y después he estado en contacto con él y me ha enseñado cosas muy interesantes, y pie
EO114 E-x Otr desde luego la que sí lo tiene que saber, póngase en contacto con Encarna Sánchez. Me la comí, me la comí, porque e
HO005 MJ Tel rego. C Exacto y a ver si yo en la noche me pongo en contacto con esta señorita. B Muy bien. Bueno. C Gracias Arm
Ho112 MBF Pol ontagiar el sida. D Sí. A Sí, también. E Cualquier contacto con fluidos sea en semen, vagina, a través del ano, c
HO113 MAF Cul y en otros más difícil. El contacto, hasta ahora, el contacto con la guerrilla ha sido mínima, los hemos visto en l

EO114 E-x Otr que está estudiando otra carrera... C Por eso tiene contacto con la juventud. A Vas a hacer más carrera que Curro
EO123 E-x Otr natural de vez en cuando porque tienes que entrar en contacto con la naturales para no perder tu salud mental, pero
EO114 E-v Eco nes, a partir de mañana lunes puede usted ponerse en contacto con la residencia Santa Eugenia, es el , o bien el te
Ho113 MC- Soc llos de barandilla, que son los que tienen el primer contacto con los homosexuales a los que se reprimen de primera
EO113 E-v Eco lez, se habló, incluso, de que anoche mismo, mantuvo contacto con Manuel Chaves y pocas horas antes con Alfonso Gue
EO114 E-i Soc s cierto. A Bueno, estamos esperando que se ponga en contacto con nosotros Pilar Socorro, que está en Tenerife. Nue
Ho112 MBF Pol se lo comentamos en alguna otra ocasión, que entró en contacto con nosotros, su problema requiere de una asesoría pr
Ho123 V-v Eco n acerca de los paros existentes en el país, hacemos contacto con nuestro compañero Orlando Martínez. - De las och
EO114 E-v CiH nen una razón suficiente. A El otro día, en nuestro contacto con oyentes en la xxx de Madrid nos pedían algunos de

EO113 E-i Edu cuota fija que es corregir, atender llamadas o tener contacto con tutores que es lo mismo, pero posiblemente, el de
EO113 E-i Cul cía falta ningún apartado de correos para ponerse en contacto con usted, que con llamar al gabinete de radio de la
EO114 E-v CiH uellos que por la mañana los llamamos, nos ponemos en contacto con ustedes y les damos información. B Primero que h
EO113 E-i Edu utos nada más nos dice un teléfono donde ponerse en contacto con ustedes por ejemplo. A Bueno, pues aquí, con el
HO113 MAF Cul al inicio del conflicto, tampoco era muy fácil hacer contacto con xxx y entrar con algunas de las organizaciones. P

En el desarrollo del trabajo lexicográfico fundamentado en el corpus no habría que caer en simplificaciones conceptuales. El gramático tradicional, y en buena medida también el lingüista moderno, han consolidado la tendencia a pensar que la labor del lexicográfico es "práctica", sin más connotaciones. En consonancia con ello, el lingüista ha marginado con demasiada frecuencia la lexicografía y, si se ha acercado a ella, ha sido desde el podio de quien gusta "vigilar sin inmiscuirse" en el acto mismo de la vigilancia o sin ofrecer mejores estrategias o técnicas para vigilar. El lexicográfico de hoy en día no puede compartir, ni siquiera tolerar, este planteamiento. Lo que un lexicógrafo ha de hacer es desarrollar una teoría substantiva que le permita desarrollar "la capacidad del lenguaje para hablar de sí mismo". Ese es precisamente el reto y el objetivo del diccionario: propiciar la comprensión de los elementos lingüísticos mediante el uso de otros elementos lingüísticos. Eso implica desarrollar procedimientos y técnicas eficaces y claras. Y a ese objetivo solamente puede llegarse mediante la reflexión lingüística para descubrir los mecanismos que propician la generación de la claridad léxica. De hecho la explicitación léxica es habitual entre los hablantes y especialmente para informar a los nuevos miembros de una comunidad lingüística sobre el valor de los vocablos que usan o quieren usar (desde los niños hasta los adultos que aprenden un segundo idioma, sin dejar de lado a los nativos que amplían o desean ampliar su conciencia lingüística). Para decidir si la definición más adecuada de "cocción" es

Operación y resultado de cocer(se).

convendría preguntarnos sobre la potencialidad de esta definición para clarificar esta voz entre los hablantes.) O quizás sería más claro definir la voz de otra manera? Por ejemplo:

La cocción de algo tiene lugar cuando el líquido alcanza una temperatura que lo cambia del estado líquido al gaseoso y provoca la evaporación.

En este caso se tomaría como referente el uso real de esta voz en ejemplos concretos, en los cuales la "cocción" no se representa como una abstracción relativa a la acción de "cocer" (al igual que se hace en las definiciones "tradicionales"), sino como **un proceso real durante el cual ocurre algo**. Por otro lado, esta definición se aproxima más al tipo de definiciones que suelen hacer los hablantes nativos cuando los niños preguntan por el significado de palabras desconocidas (*"Es cuando..."*).

La realidad del uso avala este procedimiento. En nuestro corpus aparecen 294 veces los términos relacionados con "cocer" (**cocer, cociendo, cocción, cuece, cocía, cocido).** Pues bien: en casi todos los casos la referencia es concreta (... *he cocido castañas..., ... leche sin cocer..., ... cociendo los chapulines ...).*

En la comunicación lingüística las definiciones suelen hacer referencia a la realidad, especialmente prestando atención a la funcionalidad del significado. Las definiciones abstractas son menos frecuentes. Es verdad que las definiciones de ese tipo exigen más espacio, son más redundantes, pero también más claras y diáfanas. El acopio de usos propiciados por el corpus es la mejor guía y ayuda para seguir en esta dirección y propiciar cambios útiles y "frescos" en lexicografía.