Los déficits de los corpus orales del español (y de algunos análisis)

Antonio BRIZ
Universitat de València. Grupo Val.Es.Co. IULMA

A partir de la descripción de los corpus orales y escritos del español que publicábamos en el *Anuario del Instituto Cervantes 2009* (Briz & Albelda 2009) se puede entender la imagen rica y variada, además de precisa, segura y sistemática que proporciona ya la lingüística con corpus orales en español.

Ahora bien, es preciso reconocer que se trata todavía de una imagen imperfecta. Por eso y para el avance de esta lingüística y de esta metodología, en este homenaje a nuestro admirado colega y amigo Guillermo Rojo, uno de los investigadores que más han colaborado en el desarrollo de la lingüística de corpus, nos proponemos reflexionar sobre algunas cuestiones y problemas sin resolver relacionados con la cantidad o calidad de los datos, la suficiencia de los corpus (las grandes bases de datos o los corpus con objetivos concretos), los accesos a la información, la digitalización, los sistemas de marcación y de transcripción, la explotación, así como el trabajo de análisis y de abstracción.

0. INTRODUCCIÓN

El corpus lingüístico (oral o escrito) es el banco de pruebas más eficaz y natural para analizar el lenguaje y, más aún, la actuación lingüística. Hoy pocos dudan (aunque los hay dentro de los denominados "lingüistas de sillón") que la lingüística científica ha de ser una lingüística de corpus, es decir, que ha de incorporar un conjunto *amplio* y *definido* (*suficiente* y *representativo*) de materiales que le proporcione datos *fiables*.

Amplitud, definición, suficiencia, representatividad y, como consecuencia de lo anterior, *fiabilidad*, son las características que ha de tener un corpus lingüístico para ser válido o, más exactamente, para que permita validar y comprobar empíricamente las hipótesis ya formuladas o que se llegan a formular a partir de la experimentación y la observación de dicho corpus.

De acuerdo con lo anterior, se entenderá que nuestra visión sobre la lingüística de corpus es que se trata de un método de explicación potente de la lengua y el uso, no sujeto a una disciplina ni a un enfoque teórico, y que se transforma por momentos y en algunos casos en teoría, sobre todo para quienes trabajamos desde el corpus y para el corpus[1].

En los últimos quince años ha habido un desarrollo en todo el dominio hispánico de la Lingüística de corpus. Este auge entronca con el desarrollo de los estudios sociolingüísticos, pragmáticos y del análisis del discurso. Ciertamente, el desarrollo imparable y el grado de perfeccionamiento de estos corpus ha ido unido al avance informático y de la tecnología digital.

[1] Contrasta nuestra visión de la Lingüística de corpus como método con la de quienes la consideran propiamente una teoría, un nuevo paradigma capaz de explicar al menos una parte del funcionamiento de la mente, y que acepta que el procesamiento estadístico del lenguaje natural es un modo de operar de la mente. Es decir, el lenguaje humano entendido como un mecanismo computacional. Así, por ejemplo, Chafe (1994) afirma que la tarea de la Lingüística de corpus es estudiar el lenguaje y, a través de este, llegar a la mente humana.

Cum corde et in nova grammatica, 115-137.

Pero también la lingüística actual y, en concreto, aquellas disciplinas que propiciaron su desarrollo son de ello claras beneficiarias. Es un hecho el avance que ha experimentado el estudio de la lengua hablada gracias a la elaboración y al análisis de estos corpus, sin contar con la cantidad y la calidad de los datos que han proporcionado para su análisis, el ahorro de tiempo en obtenerlos y la posibilidad de tratarlos estadísticamente. Asimismo, los recientes desarrollos de la lingüística aplicada, de la lingüística clínica o de la lingüística forense deben mucho a esta lingüística de corpus.

Ahora bien, el avance de esta lingüística de corpus pasa, entre otras cosas, por corregir algunos déficits de las muestras y del tratamiento de datos, por mejorar el método de selección, la cantidad y la calidad de estos corpus, así como por revisar el uso que hacemos de los corpus en los análisis. De acuerdo con estas cuestiones, este trabajo plantea un conjunto de reflexiones sobre lo que se ha hecho, incluido el uso que hacemos de los corpus orales, y sobre lo que, en nuestra opinión, queda por hacer, los retos futuros: ¿son suficientes los corpus orales de que actualmente disponemos? ¿cuáles son las carencias más notables? ¿cómo hay que proyectar los nuevos corpus, qué falta y qué conviene corregir? Y algo fundamental, ¿ayudan nuestros análisis a partir de corpus orales a construir, confirmar, destruir o desconfirmar teorías sobre la lengua hablada?

Todo lo cual puede ayudarnos a decidir con criterio hacia dónde debe ir la lingüística de corpus orales.

1. ESTADO ACTUAL DE LOS CORPUS ORALES DEL ESPAÑOL

¿Qué tenemos? ¿Cuál es el estado actual de los corpus orales del español?

Como ya señalábamos al inicio de este trabajo, una presentación de los corpus orales hispánicos o panhispánicos actuales ya concluidos, de sus características generales, sus objetivos, su metodología (selección de informantes, variables sociolingüísticas, tamaño de la muestra), la cualidad de los datos, el sistema de transcripción y, en su caso, etiquetado, su explotación, análisis y aplicación, esto es, los resultados y frutos de las investigaciones, puede encontrarse en Briz & Albelda (2009). Dicha información se sustenta en la ya aparecida en el anejo 8 de la revista *Oralia*, coordinado por A. Briz (2005), en los dos volúmenes de *Lingüística con corpus*, editados por R. Caravedo (1999) y J. De Kock (2001) y, por supuesto, a partir de la colaboración de los autores de muchos de esos corpus que, generosamente, completaron una ficha informativa que se diseñó para recopilar los datos actualizados sobre el estado de los corpus en español.

Es obvio que tenemos que felicitarnos, pues son ya 47 los corpus orales del español (o que integran muestras orales) de los que disponemos para impulsar definitivamente la lingüística de corpus (que son 63, si les sumamos los 16 del proyecto *PRESEEA*[2]). Es claro que a más cantidad y calidad de información sobre estos corpus, mayor beneficio para los analistas del español hablado.

No todos están concebidos del mismo modo, ni presentan las mismas características. De hecho, unos son en su concepción *macrocorpus*, denominados así por la amplia posibilidad

[2] *Proyecto para el estudio sociolingüístico del Español de España y de América* (*PRESEEA*), dirigido por Francisco Moreno, <http://www.linguas.net/portalpreseea>.

de contrastar resultados entre distintas normas regionales o por constituirse en grandes bases de datos. Otros, por su ámbito de acción y sus dimensiones más reducidas, así como por sus objetivos, en principio, más concretos, pueden denominarse *microcorpus*; los hay que reúnen diversas áreas geográficas bajo un mismo proyecto o que se dedican a una de esas áreas. Existen corpus con variedad de géneros discursivos orales, otros solo se centran en un género. La mayoría de los corpus orales están formados por entrevistas o conversaciones semidirigidas y hay muy pocos corpus de conversaciones coloquiales.

Por otro lado, hay corpus de acceso a través de concordancias (electrónicos), ya tengan fines generales o específicos. Y, en fin, hay también corpus dedicados a lenguajes técnicos, a la adquisición y desarrollo del lenguaje y al desarrollo de las tecnologías del habla.

Por lo general, los macrocorpus orales constituyen grandes bases datos con objetivos amplios, como el de servir a la investigación lingüística en general y al estudio y análisis del español hablado en su conjunto (por ejemplo, el *Corpus de referencia del español actual*, el *CREA*-oral[3], el *C-ORAL-ROM*[4], este último, incluso, es un corpus de habla espontánea en español, italiano, francés y portugués), lo cual no impide que, a partir de éstos, se puedan realizar también trabajos particulares o se apliquen a determinadas disciplinas, sea el caso de los frutos obtenidos del primero para la elaboración de diccionarios y gramáticas o las posibilidades experimentadas, por ejemplo, del segundo en relación con la enseñanza y aprendizaje de lenguas (Instituto Cervantes). No obstante, hay macro-corpus con objetivos más precisos: así, el *MC-NC*[5] y el *PRESEEA*, además de dos grandes bases de datos que surgen de integrar corpus más pequeños, sirven al análisis de la variación geográfica y social en España y América, el primero más centrado en el estudio de la norma culta.

2. REFLEXIONES Y VALORACIONES SOBRE LOS CORPUS ORALES DEL ESPAÑOL

Cuando elaboramos este tipo de trabajos de carácter informativo sobre los corpus se ha de decidir a qué se da cabida, y ello nos lleva de nuevo a replantearnos cuestiones que parecían resueltas: qué es un corpus, a qué llamar corpus, qué es lo que cabe identificar como corpus y quién hace lingüística de corpus.

2.1. Sobre el concepto de corpus textual y sobre quienes lo elaboran

Ciertamente, este desarrollo "corporal" actual requiere, en mi opinión, que nos volvamos a plantear como hace algunos años qué consideramos o a qué llamamos corpus textual e, incluso, quién hace lingüística "con cuerpo" y quién usa el "cuerpo" para lograr determinados fines, que, aunque igual de legítimo, no es lo mismo.

¿Quién hace lingüística de corpus

Seguramente hay varias lingüísticas de corpus y varios modos de hacerla, lo que ya se corrobora con la tipología diversa señalada. Basta con leer el programa de comunicaciones, pa-

[3] Una información detallada sobre el *CREA* oral, en Sánchez Sánchez (2005).
[4] *Vid.* Moreno Sandoval & Urresti (2005) y Cresti & Moneglia (2005).
[5] *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico,* coordinado por J. A. Samper Padilla, C. Hernández Cabrera & M. Troya Déniz (1998).

neles, talleres, etc., de algunos de los congresos dedicados a esta lingüística de corpus para darse cuenta de los modos diferentes de hacerla y de entenderla. Por un lado, no está haciendo la misma lingüística de corpus, al menos de momento, quien trabaja solo con una base de datos, como la del *CREA*, que quien trabaja con corpus con objetivos precisos, como *PRESEEA* o como *Val.Es.Co.*[6], dado que los resultados sociolingüísticos y pragmáticos obtenidos, respectivamente, del análisis de estos dos últimos corpus se refieren al propio corpus y son más precisos en cuanto que se plantean dentro de un género discursivo, un área geográfica, una situación de comunicación concreta, etc.

Por otro lado, aunque la confluencia entre lingüística, informática y estadística es inevitable en el momento actual, esta relación se aplica de forma diferente según los análisis y, asimismo, las prioridades son distintas en cada corpus.

Nuestra visión, una más de las varias que existen, es que hace lingüística de corpus quien carga a las espaldas con él, quien indaga con él (y no con datos fragmentados, inconexos o incompletos), quien estudia los hechos lingüísticos vinculados a la situación comunicativa, nota su variabilidad, esto es, analiza la lengua tal y como se produce y se recibe en unas circunstancias determinadas de comunicación.

Todo trabajo que usa un corpus para experimentar no necesariamente ha de ubicarse en esta lingüística de corpus. Cuando este es meramente un medio, no se está haciendo estrictamente lingüística de corpus; se está haciendo también una lingüística experimental, sin duda, pero no la que nosotros entendemos como lingüística de corpus. Un solo ejemplo: algunos gramáticos o pragmatistas han usado el corpus *Val.Es.Co.* de conversaciones y han obtenido una serie de datos y de resultados sobre el uso de un determinado marcador discursivo. Sin duda, han trabajado sobre corpus, sus trabajos pueden ser resultado de usar un corpus, pero estrictamente no entrarían dentro de nuestra visión de la lingüística de corpus.

No entendemos bien a quienes, trabajando desde esta lingüística, desligan el objeto lingüístico de análisis de sus hablantes y oyentes (*cfr.* la opinión de Teubert 2005: 6). Además, hay peligro de caer en un empirismo extremo, tan desaconsejable como el innatismo, que es el otro polo.

Tampoco entendemos la lingüística de corpus como una mera tecnología, aunque necesita de ella. No nos adscribimos a la *lingüística computacional de corpus*, según denominación de Parodi (2010: 18), sino a la que él llama *lingüística de corpus computacional* y, más exactamente, a esa lingüística que en el almacenamiento, soporte y digitalización emplea o puede llegar a emplear programas computacionales.

En relación con los límites de lo que llamamos corpus, cabe preguntarse si se puede considerar un corpus un conjunto de obras de teatro, películas, discursos o debates parlamentarios, etc., escaneados y colocados sin más en un CD o volcados en un servidor de Internet. Si la respuesta es sí, cualquier conjunto de textos almacenados en un ordenador o digitalizados tendrá ese carácter. Todo será "cuerpo". Por eso, nuestra respuesta es no. El corpus textual de la llamada lingüística de corpus lo entendemos como un conjunto de textos recogidos con el fin

[6] El corpus *Val.Es.Co.* está formado por conversaciones coloquiales (Briz & Val.Es.Co. 2002), en su mayor parte obtenidas de forma secreta, y por entrevistas semidirigidas (Gómez Molina 2001, 2005, 2007), estas últimas integradas en el corpus *PRESEEA*.

de estudiar el lenguaje en uso (no hay corpus sin objetivo, por amplio que este sea). En el caso de los corpus orales, el fin es la lengua hablada y, para su estudio y análisis, se ha recogido de la realidad del habla y de contextos naturales un conjunto *ordenado* y *definido* de textos más o menos *amplio* a partir de unos criterios de *selección* y *representatividad*. Y ese conjunto textual es *fiable* como punto de partida del análisis de lo oral (o de un objetivo más preciso), para crear teoría sobre lo oral, y/o sirve como punto de llegada, esto es, para verificar las hipótesis previas (*cfr*. Sinclair 1991: 171, Stubbs 1996).

Una muestra de oraciones recogidas de la radio para el estudio de la categoría preposición es un corpus de ejemplos, incluso estos puedan estar contextualizados, pero tampoco es un corpus textual. Nuestra percepción de la lingüística de corpus es la de aquella lingüística que se hace desde y con el corpus y no solo a través de este. Así pues, el corpus en esta LC no es solo un medio o instrumento, sino que además es un método, es una perspectiva metodológica, una opción de hacer un determinado tipo de ciencia lingüística.

Por otra parte, la representatividad es clave para ayudarnos a decidir sobre la frontera entre un corpus textual y un corpus de muestras o ejemplos. Los corpus han de ser representativos de aquello para lo que están pensados y los buenos resultados dependen de la constitución del corpus, de esta representatividad. Es cierto que la lengua es variada y heterogénea, y lo es también que un corpus no puede contenerlo todo, pero sí representarlo.

Un corpus textual ha de tener un plan previo y de selección de las muestras y de los hablantes. Luego, una restricción por arriba: no cualquier recopilación de textos es un corpus textual; y otra por abajo: no lo es tampoco un conjunto de muestras o ejemplos extraídos de diferentes materiales.

2.2. Dos modos de obtener y usar el corpus a la vista de lo ya hecho

Al revisar el estado de los corpus orales, puede notarse la existencia, en general, de dos tipos de corpus textuales según el modo de obtención de las muestras y el modo como los recibe el usuario (el tipo de herramienta con que este se encuentra):

a) la *base de datos* o *corpus indirecto*,

b) el *corpus oral propiamente dicho* o *corpus directo*.

Adelantamos nuestra preferencia por el segundo, quizás más representativo, para formular hipótesis y hacer teoría, y el de datos para confirmar, buscar y revalidar. Y nuestra decisión se basa en las cualidades que entendemos ha de tener un corpus oral, como se mostrará enseguida.

2.2.1. Las bases de datos o corpus orales indirectos

Los que hemos denominado *corpus-base de datos* o *corpus indirectos* se caracterizan por los rasgos siguientes:

— Son de grandes dimensiones.

— Se obtienen a través de otros medios, no en su contexto natural (por ejemplo, de los parlamentos, de los medios de comunicación, etc.).

— Tanto es así que el acceso a algunas de estas muestras podría hacerse por vías distintas a las del corpus.

— Son el instrumento para acceder a diferentes fines.

— Son herramientas de consulta.

— En su recogida puede no intervenir el investigador.

— Se presentan en formato de motor electrónico de búsqueda y no suelen permitir el acceso directo a los textos. Tienen siempre acceso electrónico, por ejemplo, a frecuencia de ocurrencias y a concordancias, pues quien accede a estos suele perseguir informaciones puntuales (léxicas, gramaticales, por ejemplo) o bien cómputos cuantitativos respecto a un fenómeno.

— Tienen objetivos menos precisos y, por tanto, límites menos definidos.

— Son muestras de tipología más heterogénea (lo que también es consecuencia de lo anterior).

— Los frutos obtenidos de su explotación son, asimismo, heterogéneos. Esto es, sirven a muchos y para mucho.

— Están siempre informatizados.

— No se presentan en papel.

— No parten de hipótesis previas sobre la lengua hablada.

Sirva de ejemplo de corpus-base de datos indirecto el *CREA* de la Real Academia Española, coordinado por G. Rojo, con la estrecha colaboración de M. Sánchez y R. Pino. Este pretende ser un corpus de referencia del español actual, lo que significa, de acuerdo con esta institución, que "ha de ser lo suficientemente extenso para representar todas la variedades relevantes de la lengua en cuestión", pues su objetivo es "proporcionar información exhaustiva acerca de una lengua en un momento determinado de su historia" (www.rae.es). En la actualidad el *CREA* cuenta con 160 millones de formas procedentes de textos tanto escritos como orales de los diversos países de habla hispana (50% de España y 50% de América). Este macrocorpus es un proyecto constantemente abierto a la actualización y crecimiento de la base de datos (de hecho, actualmente se desarrolla el proyecto *CORPES*: el *Corpus del Español del Siglo XXI*, en colaboración con todas las instituciones que forman la Asociación de Academias de la Lengua Española, cuyo objetivo es disponer de 25 millones de formas léxicas para cada uno de los años comprendidos entre 2000 y 2011).

A través de la página electrónica del *CREA* (http://corpus.rae.es/creanet.html) se puede acceder al motor de búsqueda y a la recuperación de concordancias, con diversas posibilidades de filtrado (cronológico, geográfico, oral/escrito, etc.). No forma parte de los objetivos de este tipo de bases de datos el acceder a los propios textos de modo completo, sino que su pretensión es la de ser una herramienta de consulta y obtención de frecuencias, pequeños párrafos contextualizados de los ejemplos y listados de concordancias.

La parte oral del *CREA* supone el 10% del total de registros y está compuesta por dos tipos de material: transcripciones propias de documentos sonoros extraídos de medios de comunicación y la incorporación de diversas transcripciones de corpus orales cedidos a la RAE y recodificados de acuerdo con el sistema de etiquetado del *CREA*. Los documentos orales que abarcan el período 2000-2004, a diferencia de los anteriores, permitirán, en muy poco tiempo, el acceso sonoro del texto de forma sincronizada[7].

[7] Para una mayor información sobre el *CREA*-oral, *vid.* Sánchez Sánchez (2005). Una comparación y valoración crítica de dos de las grandes bases de datos, el *CE* (Corpus del español), de M. Davies, y el *CREA* (Corpus de referencia del español actual) de la Real Academia Española, puede leerse en M. Davies (2008, 2009) y Rojo (2010).

2.2.2. Los corpus directos o corpus propiamente dichos

Hablamos de *corpus propiamente dichos, directos* cuando presentan los rasgos que siguen:

— Están formados por muestras de habla obtenidas o recogidas en su contexto natural de enunciación. "De la huerta a la mesa o al plato, sin pasar por intermediarios". Son auténticas muestras de habla directamente tomadas de la realidad.

— Son los propios investigadores, los grupos de trabajo o personal formado para tal fin quienes recolectan esas muestras.

— Son de acceso directo al objetivo.

— Se presentan directamente en formato textual. Por ello hablamos de corpus propiamente dichos, porque permiten siempre el acceso directo a los textos completos. "El cuerpo está presente". Este acceso directo es necesario cuando uno realiza estudios sociolingüísticos, sociopragmáticos o pragmalingüísticos.

— Tienen objetivos precisos y, por tanto, límites más definidos, lo cual no impide que puedan servir a otros fines (incluso, no previstos).

— Son muestras de tipología homogénea (lo que también es consecuencia de lo anterior).

— Los frutos principales obtenidos de su explotación están en relación con esos objetivos.

— Pueden estar informatizados o no. No siempre están marcados, ni tienen acceso electrónico por concordancias (aunque sería deseable y, sobre todo si son macro-corpus, muy necesario).

— Suele haber muestras impresas.

— Y suelen partir de hipótesis previas (otro límite del "cuerpo").

Dos ejemplos de *corpus directos* son, por un lado, el macro-corpus *PRESEEA*, corpus sociolingüístico que tiene como objetivo general el estudio de la variación geográfica y social en las distintas normas regionales del español en España y América, y que está integrado por corpus más pequeños recogidos en cada una de esas zonas, y, por otro lado, el micro-corpus *Val.Es.Co.*, de conversaciones coloquiales, que tiene como objetivo el estudio del español coloquial.

Tanto en un caso como en otro, los equipos trabajan con una metodología común en cuanto a la grabación y técnica de grabación: en el primer caso, se trata de entrevistas y, en el segundo, de conversaciones coloquiales. Los materiales están transcritos y digitalizados parcial o totalmente, incluso, como en los corpus-*PRESEEA*, etiquetados[8]; y parte de los materiales en ambos corpus se ha publicado en papel[9]. Los investigadores han sido en ocasiones observadores (participantes o no) de dichos materiales.

[8] *Vid.* el etiquetado del *PRESEEA* y los problemas de su conversión en documentos XML en Villena *et al.* (2010).

[9] En la actualidad el corpus *Val.Es.Co.* se ha digitalizado, se ha ampliado la cantidad de material transcrito y se pretende su etiquetado, todo ello dentro de un proyecto dirigido por Salvador Pons. La idea es poner el corpus a disposición pública en Internet a lo largo de 2012.

Los corpus-*PRESEEA* se estratifican en diversos niveles socioculturales (alto, medio y bajo), en grupos generacionales y se distribuyen igualitariamente en los dos sexos. En algunos casos se tienen en cuenta otras variables, como por ejemplo, el lugar de procedencia de los informantes. El corpus *Val.Es.Co.* de conversaciones tiene como objetivo, como se ha señalado, el análisis pragmalingüístico del español coloquial. Su singularidad consiste, sobre todo, en el método secreto de grabación y en un sistema de transcripción propio que intenta reflejar lo más fielmente posible la oralidad sin dificultar la lectura del texto[10]. Esta organizado a partir de dos criterios: el de la mayor o menor prototipicidad coloquial y el de estratificación sociocultural (estrato alto, medio, bajo).

Estos dos *corpus propiamente dichos*, *PRESEEA*-entrevistas y *Val.Es.Co.*-conversaciones coloquiales, además de poseer objetivos concretos, parten de hipótesis previas sobre los objetos de estudio establecidos (los límites de otros como los llamados *corpus-bases de datos* no están tanto en premisas teóricas cuanto en otros factores o aspectos como, por ejemplo, el interés por recoger muestras representantivas del español hablado de una época, de los varios registros, etc.).

Por ejemplo, la hipótesis inicial del grupo Val.Es.Co. era que el funcionamiento de la conversación coloquial podía explicarse no como transgresión de la gramática oracional, sino como conjunto de estructuras y estrategias, de base pragmática, constituidas en el proceso de interacción. Para comprobar dicha hipótesis era condición indispensable disponer de un corpus representativo de conversaciones, transcrito mediante un sistema de transcripción capaz de representar los hechos conversacionales objeto de nuestro estudio. Por su parte, en el *PRESEEA*, la recogida de numerosas muestras de habla de distintas comunidades del mundo hispánico, estratificadas según parámetros sociales y geográficos, tiene una finalidad sociolingüística y sociopragmática y tiende a la comparación de los resultados obtenidos por las investigaciones en España y América de los distintos grupos integrantes en el citado proyecto. La hipótesis previa deriva ya del propio método: la variación lingüística según variables socioculturales.

En suma, los *corpus directos* ejemplificados se elaboran en virtud de unos objetivos concretos y previos, aunque, ciertamente, su utilidad puede sobrepasar en muchos casos dichos objetivos.

2.2.3. A modo de propuesta

No debería ser así, pero de momento el investigador que acude a los *corpus de acceso directo al texto* persigue una finalidad distinta del que se aproxima a los corpus de concordancias. Los estudios pragmáticos, sociolingüísticos o socioculturales tienen apego a los corpus propiamente dichos, entre otras cosas porque permiten el acceso al texto completo y son corpus más controlados y, por ende, se entiende que son más fiables para el análisis de esos datos pragmáticos o sociolingüísticos. En cambio, la investigación gramatical y léxica tiene más apego a los *corpus-base de datos* por la cantidad de estos datos, unida a las herramientas de búsqueda que proporcionan, como por ejemplo el acceso electrónico por concordancias, por fre-

[10] Diversos investigadores de todo el ámbito hispánico han hecho uso del sistema de transcripción creado por el grupo Val.Es.Co. Muchos son trabajos individuales y otros institucionales. Sirvan solo de botón de muestra el corpus del español de Barcelona, dirigido por R. Vila (2001), o el corpus de conversaciones que Jorge Murillo ha recogido y se encuentra almacenado en la Universidad de Costa Rica.

cuencia de ocurrencias o por colocaciones. Es cierto que, por ejemplo, el *CREA* tiene también unas pocas muestras de habla conversacionales y, por tanto, extraídas de modo natural, pero son muchas más las obtenidas de forma indirecta. Y tiene, además, motores de búsqueda, pero esos motores, desgraciadamente, no detectan de momento los hechos pragmáticos.

Sin texto (transcripción, audio o, incluso, vídeo) no hay análisis pragmático o discursivo de lo oral que se sostenga. Luego, a lo que tenemos que tender es a aproximar ambos tipos de corpus y a que la diferencia quede en los grandes o más pequeños objetivos o fines y en las dimensiones.

Creemos que el avance más claro de la lingüística de corpus se verá cuando estas tipologías dejen de ser o de estar operativas, cuando los corpus sean capaces de contemplar, combinados, todos los subtipos y posibilidades que establecen los rasgos anteriores; cuando, en suma, combinen o sepan combinar cantidad y calidad, así como las cualidades que, sin duda, ambos tipos de corpus poseen.

En suma y a modo de propuesta: debería tenderse a elaborar corpus de auténticas muestras de habla directamente tomadas de la realidad, enriquecidas con muestras auténticas obtenidas a través de otros medios (los corpus artificiales han de ir pasando a mejor vida), con observación (incluso, participante) de los propios investigadores, que sean representativos de un objetivo más preciso, lo cual no impide que puedan servir a otros fines (incluso, no previstos), que permitan acceso directo al texto completo (necesario para estudios pragmáticos, sociolingüísticos y socioculturales), que los frutos principales obtenidos de su explotación estén en relación con esos objetivos. Se trataría, por tanto, de elaborar microcorpus precisos de objetivos y de hipótesis, que puedan formar parte de un macro o megacorpus, de grandes dimensiones, lo que aseguraría una cierta homogeneidad y control de los campos y tipos presentes y, sobre todo, de las ausencias, además de servir, por supuesto, a muchos y para mucho más de lo previsto. Y en cuanto al soporte y almacenaje, deberían estar informatizados y marcados y, asimismo, permitir accesos electrónicos a concordancias, a frecuencias de ocurrencias, etc. Además, una selección de las muestras podría estar impresa (volveremos sobre esta propuesta más tarde).

3. DÉFICIT, SESGOS Y MÁS PROBLEMAS

Sigamos con nuestro diagnóstico de lo hecho y con nuestras opiniones de lo que conviene hacer.

3.1. La falta de muestras orales

Según la revisión realizada en Briz & Albelda (2009), en la mayoría de los grandes corpus lo oral ocupa un espacio que no sobrepasa el 10%. Ello muestra que, si bien se ha avanzado mucho, es preciso seguir elaborando corpus orales por el bien de los análisis.

3.2. La falta de materiales de conversaciones

Dentro de estos corpus orales, sin duda, se observa un déficit importante de materiales de conversaciones, tanto formales como coloquiales. Solo un 10% de los materiales en el

PILEI es conversacional y no aparece en todas las ciudades grabadas. El *CREA* recoge 18 conversaciones reales (que proceden del corpus *ACUAH*[11], de Alcalá de Henares); el corpus publicado de Barcelona contiene 2 conversaciones telefónicas y 8 cara a cara[12]; asimismo, el *VUM*[13] de Málaga recoge unas pocas muestras de textos orales conversacionales. Solo el corpus *COLA*, dirigido desde la Universidad de Bergen por Annette Myre Jörgensen, empieza a tener una cantidad importante de muestras de conversaciones entre jóvenes en España e Hispanoamérica[14].

¿Hay una explicación para tan significativa escasez?

A pesar de que somos conscientes de que los corpus de conversaciones proporcionan una mayor espontaneidad y naturalidad en los interlocutores y más dinamismo y variedad situacional en cada una de las grabaciones (diverso número de hablantes, distintos tipos de relaciones entre los interlocutores, diversidad de temas, de espacios físicos, etc.), la dificultad de la grabación, de la transcripción, del etiquetado, de la menor flexibilidad de dichos materiales (en aras de la homogeneidad y de la reutilización; por ejemplo, no podemos corregir las agramaticalidades, recuperar los sonidos perdidos, etc.), los hacen menos apetecibles.

Desde los años sesenta, los investigadores han utilizado como método de grabación, por influencia de la dialectología y de la sociolingüística, la entrevista libre, sin cuestionario, entre un entrevistador y un informante, conscientes de serlo. Se trata de un método adecuado a los fines, pues el investigador necesita tener un control de las variables geográficas y sociales. La ventaja del método de extracción de datos a través de la entrevista es que ofrece mayor sistematicidad al lingüista, ya que permite controlar el equilibrio en los parámetros sociolingüísticos y asegurar la representatividad de los informantes en el total de la muestra.

Ni los avances técnicos ni los nuevos enfoques pragmáticos o de análisis del discurso, sea el caso del Análisis conversacional norteamericano, han podido desbancar el método citado, lo cual no es una crítica hacia el método, sino a quienes lo utilizan para después emprender estudios de algunos hechos pragmáticos como, por ejemplo, los relativos a la cortesía verbal (un tema estrella del Análisis del discurso). El resultado de un análisis que tiene como objetivo estudiar la cortesía en un corpus de entrevistas es parcial y, si se queda ahí, sería hasta poco adecuado. En una entrevista se vela siempre por las imágenes, luego, como ha demostrado Marta Albelda (2004), no se producen actos mitigadores de las amenazas a la imagen del otro y no existe la llamada anticortesía o descortesía fingida (a no ser que se trate de una entrevista polémica). Asimismo, poco rentable sería este tipo de muestras si el objetivo fuera el estudio de la conducta interaccional en relación con la toma libre de turnos o con el habla simultánea, por ejemplo.

---

[11] *Análisis de la conversación-Universidad de Alcalá de Henares*, recogido por Ana M.ª Cestero en 1991.

[12] *Corpus del español conversacional de Barcelona y su área metropolitana,* llevado a cabo por Rosa Vila y el Grupo GRIESBA en 2001.

[13] *Vernáculo Urbano Malagueño*, dirigido por Juan Andrés Villena Ponsoda.

[14] *Corpus Oral del Lenguaje Adolescente*, coordinado por Annettte Myre Jörgensen (Universidad de Bergen), recopila más de 300 conversaciones espontáneas entre jóvenes, de Madrid, Santiago de Chile, Buenos Aires, Guatemala, La Habana. Las conversaciones han sido recogidas en el período de 2002 a 2009. Es un proyecto todavía no finalizado; ver www.colam.org: acceso electrónico gratuito, previa solicitud de contraseña a través de la propia página electrónica.

La entrevista libre ha sido la base de muy valiosos trabajos sobre la lengua hablada, pero los resultados que se obtienen, relacionados ahora con el propio género, no siempre pueden extrapolarse a otros géneros más naturales, como el de la conversación. Y no por llamar a la entrevista conversación semidirigida es más conversación. Es cierto que estamos en un proceso de hibridación de géneros y que una entrevista puede acercarse a la conversación, pero insistimos, una entrevista se realiza entre dos personas, de las cuales una dirige el diálogo y, aunque llegan a ser espontáneas, prevalece en ellas una finalidad transaccional: el acuerdo previo de que el informante habla para que su testimonio lingüístico sea almacenado.

Sea por el motivo que sea, las conversaciones (coloquiales o formales) están representadas minoritariamente en los corpus actuales de lengua hablada en España e Hispanoamérica. Y más todavía, están prácticamente ausentes las conversaciones grabadas de forma secreta (y ahora más aún, por ley). En nuestra opinión, se pueden seguir obteniendo pequeñas muestras, al menos, en casa de familiares y amigos (y procurando posteriomente el método de "bola de nieve"), solicitando el permiso a posteriori y protegiendo siempre las identidades de los participantes, por ejemplo, alterando la antroponimia y la toponimia, si hiciera falta. No voy a entrar en la polémica, a mi modo de ver estéril, de si este procedimiento es ético o no. Seguramente, no lo es o no es muy ético, pero por eso mismo no se discute.

3.3. "Lo que no está… a veces sí existe". Otras manifstaciones necesarias de lo coloquial o de lo formal

Sobre el corpus se documentan y se extraen datos, los que hay y nos proporciona dicho corpus. Por supuesto no están todos los que son. Es en parte cierto, como dice J. Portolés (2007), que los corpus orales son una selección de la realidad y, en ocasiones sesgada, pues el investigador no puede grabar todas y cada una de las situaciones y algunos fenómenos pueden no estar documentados en dicho corpus oral. Señala el autor, buen conocedor del corpus *Val.Es.Co.*, que, por ejemplo, *pero siéntate por favor*, es un uso que se puede entender como frecuente en un corpus oral conversacional, pero en dicho corpus no aparece. Tampoco se localiza en los documentos orales del *CREA* y, sin embargo, no tuvo el citado autor problemas en hallar abundantes ejemplos en obras teatrales del mismo corpus.

En nuestra opinión, la razón es, simplemente, que hay pocas muestras tanto en uno como en otro corpus. Si hubiera más (y "más" es casi siempre "mejor") y si las muestras respondieran a un mapa de situaciones previamente establecido, estaría documentada esta o, al menos, otra construcción similar. Asimismo, podría pensarse en la posibilidad de extender las muestras orales a ciertos géneros escritos que reproducen o reflejan lo oral (cartas familiares, buena parte de la comunicación electrónica: chats, mensajes de correo y móvil, etc.) o que imitan lo oral (por ejemplo, los textos teatrales, de los que hablaba J. Portolés).

Los géneros discursivos formales tampoco están bien representados en los corpus orales. Por ejemplo, J. Portolés (2004) realizó el estudio de la partícula *antes por el contrario* para el *Diccionario de partículas discursivas del español* (*DPDE*, Briz, Pons & Portolés 2008). Escribe (2004: 43) que para su análisis hizo en el *CREA* la búsqueda de usos de dicha partícula y que todos los casos documentados provenían de Chile y, más exactamente, de documentos que provenían del Congreso de los diputados chileno. Esta documentación tan escasa puede ser real o, lo más seguro, es que se deba a la ausencia de textos prototípicos formales. Por tanto, si no

contamos con textos que aseguren una cierta formalidad, muchas constantes de dicho registro quedarán sin documentar.

Y en la escala de la formalidad, faltan, en concreto, corpus orales que recojan el lenguaje académico y profesional; es preciso incorporar los lenguajes orales de especialidad (transacciones comerciales, negociaciones, interacciones en vistas y comparecencias y en distintos órganos y jurisdicciones judiciales, interacciones abogado-cliente, etc.), a pesar de las dificultades para obtenerlos (*vid.* n. 17); añadamos tertulias, debates en los medios, debates en el Congreso de los diputados, telediarios, documentales. Todos estos textos serían prototipos para estudiar el grado alto de formalidad oral, a pesar de la coloquialización creciente de muchos de ellos.

Los sesgos y déficits que estamos notando en relación con las muestras orales afectan a la obtención de materiales y, por ende, perjudican el análisis.

3.4. El acceso

Si además de ser pocos los corpus orales, el acceso a estos es difícil, el problema aumenta considerablemente.

El acceso a los corpus orales existentes no siempre es sencillo. Obviamente, no parece tener mucho sentido construir corpus si no se prevé el fácil acceso a estos. Nótese que esta afirmación añade otra característica esencial de la lingüística de corpus, el altruismo. El corpus ha de ser para beneficio de muchos, aun a costa de la salud propia.

Como decíamos, es preciso cubrir las lagunas en las muestras orales, que son todavía muchas, pero quizás lo urgente sería reunir las existentes para poder utilizar los corpus de modo fácil, cada cual con sus ventajas e inconvenientes, cada cual con su método y objetivos, con su sistema de transcripción y de codificación propio. Al menos, podríamos disponer así y de momento, hasta que llegue la estandarización, de un efectivo y eficaz banco de pruebas común. En este asunto trabaja actualmente el grupo Val.Es.Co. de la Universidad de Valencia.

3.5. La heterogeneidad de métodos, características y fines

Que los planteamientos metodológicos, características y finalidades de los corpus sean diversos puede, en principio, ser enriquecedor, pero, en realidad, dificulta el uso de estos desde otra perspectiva que no sea aquella para la que fueron diseñados (Sinclair 1991, Alvar Ezquerra & Villena Ponsoda 1994, Alvar Ezquerra, Blanco Rodríguez & Pérez Lago 1994, Pons Bordería & Ruiz Gurillo 2005). Y tampoco ayudan las distintas técnicas utilizadas no solo en la extracción de datos, sino en la explotación de los mismos, los modos de acceso o consulta diferentes. Debemos tender a la homogeneidad de los corpus[15]. Ahora bien, ¿homogeneidad en qué sentido?, ¿es posible o es solo un deseo inalcanzable? Estamos, sin duda, de acuerdo en algunos de los criterios de armonización en la línea de *EAGLES* (1996a y 1996b)[16], algunos de los cuales se recogen en nuestra propuesta anterior (§ 2.2.3. *Vid.* también más adelante § 5).

[15] *EAGLES, Expert Advisory Group on Language Ingineering Standards*, es un proyecto en el marco de la Unión Europea que evalúa métodos y sistemas existentes, a partir de los que realiza sus propuestas, a modo de recomendaciones.

[16] *Vid.* n. anterior. *Cfr*. también Biber & Tracy-Ventura (2007).

Lo que parece claro es que hoy por hoy ningún corpus oral del español es autosuficiente, ni los macrocorpus, ni los microcorpus, ni los directos ni los indirectos. En ocasiones algunos de los que hemos llamado corpus propiamente dichos o directos pueden servir a investigaciones particulares o a fines más o menos precisos, pero no siempre pueden ir más allá. Por ejemplo, un corpus solo de entrevistas no puede ser suficiente para concluir de modo general sobre la lengua hablada, y otro de conversaciones coloquiales puede decir poco sobre el registro formal. Otro ejemplo: el control metodológico que ejerce el investigador sobre un corpus sociolingüístico puede restar espontaneidad en la muestra (¿de qué registro, entonces, podrían obtenerse resultados?)

Por otro lado, cabe preguntarse cuándo una gran base de datos tiene el número suficiente de textos para ser representativa y poder extraer resultados significativos. Y, más exactamente, los resultados obtenidos, ¿de qué son? ¿son extrapolables?… ¿Qué parámetros ha de tener la selección de estos textos y el sistema de búsquedas? ¿Son fáciles de exportar o no? ¿Qué filtros han de tener o pasar? ¿Qué finura han de tener, sobre todo, si hay anotación?

Seguramente, estas limitaciones serían menores si los corpus orales futuros incorporaran como criterio fundamental el de la situación, que incluiría a su vez el de los géneros y subgéneros discursivos más o menos formales o más o menos coloquiales. La situación, en nuestra opinión, es el centro de la variación y, por tanto, debería ser el criterio básico en la elaboración y desarrollo de esos corpus futuros.

Y ayudaría también a resolver algunas de las cuestiones antes planteadas el hecho de partir de la creación de microcorpus de las variedades diafásicas, de géneros discursivos orales o, al menos, de aquellos géneros prototípicos de lo oral que, luego, se volcarían a un macrocorpus. Esto es, junto a las entrevistas no podrían faltar muestras de conversaciones; más aún, la conversación debería ser el tipo más representado, pues es el modo más natural y espontáneo de hablar.

Por tanto, el reto, sin duda, en la confección de los nuevos corpus orales del español es incorporar la variación lingüística, ya no solo geográfica, sino de registros y de géneros más o menos coloquiales y formales (entre ellos, como decíamos más arriba, los géneros académicos, profesionales —los de las lenguas de especialidad[17]—, sociales, etc.), que permita comparaciones entre variedades de una lengua (o, incluso, entre lenguas), monológicos y dialógicos.

Una cosa conduce a la otra: un macrocorpus de estas dimensiones y características de diversificación tendría que usar los avances tecnológicos de soporte y almacenamiento, así como los sistemas de búsqueda y recuperación de datos. Y, como notaremos, precisaría de etiquetados.

3.6. Las muestras. Diseño y elaboración

La amplitud es un rasgo señalado repetidamente de los muestreos. Ya lo señalábamos antes: "más", en principio, es "mejor". Aunque la amplitud se mide, sobre todo, con la repre-

[17] Puede que el interés por las lenguas de especialidad favorezca la creación de corpus pequeños y con objetivos precisos —el español jurídico, comercial, turístico—, y de proporciones pequeñas, no tanto por un deseo de que sea así, sino por las dificultades de obtener muestras o, de nuevo, por cuestiones éticas (no se conoce a las personas ni hay afinidad o cercanía con ellas) y por la dificultad de obtener interacciones reales en empresas o agencias de turismo, por ejemplo, ante la falta de permisos de estas empresas.

sentatividad o suficiencia. Estas son las mejores varas de medir y calibrar la dimensión de un corpus. No requiere en este sentido la misma amplitud un corpus que pretenda estudiar el habla de Valencia que las diferentes normas regionales del español. Pero ambos deberían ser igualmente exhaustivos y representativos.

Esta dimensionalidad de los corpus nos lleva a otra consideración: los corpus son proyectos de grupo. La necesidad de elaborar, revisar (y validar) el corpus hace necesario el trabajo en grupo. No excluimos que una persona pueda elaborar una muestra, su propio corpus, pero es posible que al corpus le falte validación y, casi con toda seguridad, continuidad, reutilización, accesibilidad, etc.

Por otro lado, creemos, además, que los investigadores o personal instruido previamente tienen que implicarse en mayor o menor medida en la elaboración del corpus (grabación, transcripción, etc.). Así se logra que las distorsiones en todos los sentidos sean menores; más aún, cuando actualmente hay investigadores que están elaborando "metacorpus", en los que se graba a los grabadores, y, por tanto se impone la implicación directa de los analistas.

La formación en lingüística de corpus es absolutamente necesaria para elaborar un buen corpus. Hay que formar expertos, pues debemos tender a que la elaboración de esos corpus, el diseño, la extracción de datos y la transcripción o, al menos, la validación de esta, la hagan personas formadas en este método. Si no es así, corremos el peligro de alterar o falsear los resultados de los análisis posteriores.

3.7. Transcripción y codificación

Es preciso comenzar señalando que, en principio, cualquier sistema de transcripción es adecuado siempre que se ajuste al objeto de estudio y a la finalidad para la que se emplee y, por supuesto, cumpla los principios de exhaustividad y pertinencia de los signos, es decir, que cada signo represente un único fenómeno y que cada uno de los fenómenos aparezca codificado mediante una única convención. Conviene, no obstante, que el sistema presente otra característica, la adaptabilidad, esto es, que sea flexible en cuanto a su capacidad de estrecharse (introducir otros signos) o de ensancharse (eliminar algunos) en función de los objetivos más o menos concretos que puedan surgir. Y, finalmente, una vez aplicada, la transcripción ha de ser validada y revisada mediante filtrados de distinto tipo (por ejemplo, tanto en el Proyecto *C-ORAL ROM* como en el de la conversación coloquial del grupo Val.Es.Co. se insiste mucho en esta validación o verificación de lo transcrito). No se debe olvidar en este sentido que, por más externos y objetivos que sean los medios de manipulación y análisis de una transcripción, transcribir es un proceso humano y, por tanto, subjetivo y, por eso también nuestra insistencia en la validación.

Del texto y el sonido cada uno en su formato, es preciso ir pasando a oír y a ver al tiempo lo que se está oyendo. Por supuesto, se han recibido con entusiasmo estas nuevas herramientas informáticas que permiten el alineamiento perfecto del texto transcrito y del sonido, de las que ya hacen uso algunos corpus (el corpus *C-ORAL-ROM* y muy pronto el nuevo *CREA* oral). En cualquier caso, mientras llega definitivamente, nos conformamos con tener la transcripción, el audio (y, en su caso, el vídeo).

Sin duda, transcribir el corpus es la tarea más costosa y complicada, pero también una tarea necesaria. Y en esta es fundamental la mano del experto. ¿Existe un control férreo de las

transcripciones y codificaciones en todos los corpus? Desgraciadamente, nuestra experiencia nos dice que no. Tendremos que poner especial celo en el futuro.

Y lo mismo cabe decir de la codificación. Ha de estar realizada por personal formado en esta tarea y controlada por expertos.

Por supuesto, todo ello cuesta mucho tiempo y muchísimo dinero. La lingüística de corpus necesita de una fuerte financiación. Y uno de los argumentos que conviene esgrimir para la misma es esta externalización, es decir, que el corpus sirve para mucho y para muchos; y esta externalización, con más razón, cabe esgrimirla si se defiende que la lingüística de corpus ha de entenderse, como en nuestro caso, como un método disciplinar. No tiene sentido que todo el esfuerzo de elaborar un corpus, lo costoso en todos los sentidos que resulta, sirva solo a una investigación particular o a un grupo de personas.

Y ello apunta a otro debate, el de cómo se financian o deberían financiarse, que no vamos a tratar aquí.

3.7.1. El método de transcripción

En cuanto al cómo transcribir o qué sistema de transcripción emplear, cabe decir que no existe un modelo único de transcripción, cerrado y estático. Señalábamos que el mejor sistema de transcripción es el que se adapta, en cada caso, al objeto de estudio.

La transcripción de las muestras orales se debe hacer de acuerdo con un método que permita reproducir sobre el papel o la pantalla los aspectos más relevantes del acontecimiento comunicativo que se pretende reflejar. La distorsión inherente a toda transcripción (pensemos, por ejemplo, en la conversación), ineludible por el mero hecho de pasar del soporte oral a soporte escrito, no debe impedir, sin embargo, el estudio de los aspectos más interesantes de la misma. "¿Cómo se asegura que la distorsión entre el dibujo y la figura dibujada sea la menor posible?"

La mejor transcripción, en nuestra opinión, es la que concilia la precisión en la descripción del género discursivo que se esté transcribiendo con la comprensión de la lectura. Así, la transcripción, por ejemplo, de una conversación no puede ser solo ortográfica (o solo podría serlo para ciertos fines), ya que los signos ortográficos no pueden dar cuenta de muchos fenómenos de la lengua hablada (pensemos solo en el caso del habla simultánea, en las anotaciones prosódicas). Por otro lado, los datos obtenidos de ciertos corpus orales no se entienden en ocasiones por la falta de enriquecimiento contextual, lo que puede inducir a errores de interpretación. Como afirmábamos, un corpus oral transcrito sin prosodia, sin al menos algunas indicaciones prosódicas, es como un texto sin voz, lo cual es una paradoja difícil de superar. Claro que ello no puede suplir ni el audio ni el vídeo. Es obvio que el mejor sistema en este sentido será el que pueda ofrecer alineados, transcripción con audio (e, incluso, vídeo).

En el caso del sistema de transcripción Val.Es.Co. (Briz & Grupo Val.Es.Co. 2002: 29-31, Hidalgo & Sanmartín 2005, Pons Bordería & Ruiz Gurillo 2005), como la intención del grupo era combinar en la transcripción la precisión en la descripción de los fenómenos con la comprensibilidad de la lectura, optamos por este método que llamamos jeffersoniano, un sistema ortográfico enriquecido con signos que marcan fenómenos conversacionales. De ese modo, podíamos mostrar las superposiciones, las vacilaciones, la longitud de las pausas, la pronunciación marcada, la sucesión inmediata de intervenciones sin pausa, etc. fenómenos que traducen en una marca convencional un posible hecho pragmático (www.valesco.es).

3.7.2. Las marcas

Aunque el sistema Val.Es.Co. ha comprobado su utilidad en los trabajos llevados a cabo por numerosos investigadores, sigue planteando algunos problemas teóricos y prácticos, sobre todo en lo referente al grado de explicitud de la transcripción, al grado de reflejo de variedades dialectales o idiolectales, al sistema de transcripción prosódica, a las marcas pragmáticas, etc. De entre estas cuestiones nos gustaría señalar una, que se nos plantea con una cierta frecuencia, referida a por qué no se ha adoptado un método de transcripción informatizado, siguiendo alguno de los protocolos existentes (TEI, XML).

Como señalábamos, el sistema de transcripción ha de ser flexible y ha de poder estrecharse o ensancharse. Esta flexibilidad quiere decir, usando las palabras de Pons Bordería & Ruiz Gurillo (2005):

a) Que la cantidad de signos empleada y el carácter más o menos ancho o estrecho de la transcripción depende del objeto de estudio elegido. Así, las conversaciones semidirigidas del *PRESEEA*-Valencia (Gómez Molina 2001) han prescindido de indicar los solapamientos, por otro lado prácticamente inexistentes en este tipo de acontecimientos comunicativos. Por su parte, Hidalgo (1997), dedicado a la entonación coloquial, no solo utilizó el sistema completo de transcripción, sino que añadió datos sobre la frecuencia del fundamental de cada grupo de entonación. En definitiva, el sistema de transcripción está al servicio de las necesidades del investigador y no al revés; por ello, deberá estar sujeto a una revisión permanente que garantice su validez, pues, como técnica para facilitar y mejorar los análisis, cambiará según avance el instrumental que utiliza.

b) Que nada impide que una transcripción, por ejemplo, al modo Val.Es.Co, se vuelque también en un sistema de marcado, habida cuenta de que se establezca un sistema de marcas adecuado, que sirva a otros investigadores y a la investigación misma.

En nuestra opinión, los distintos sistemas no son incompatibles, más aún, creemos que se complementan y sería ideal que un corpus oral dispusiera de una transliteración con signos y convenciones específicas, junto con la codificación a través de marcas automáticas. Es decir, una doble transcripción, que permitiera un tratamiento informático a través de marcas y que tendiera también a la posibilidad de su lectura y fácil comprensión. Sin duda, de ese modo se aseguraría el análisis cuantitativo, una reutilización mayor y, a su vez, los análisis cualitativos saldrían beneficiados. Pensemos que una transcripción solo pensada para la recuperación de datos por ordenador a través de dichas etiquetas dificulta la lectura de los textos transcritos y, así pues, la descripción lingüística. Y no debe olvidarse que una buena descripción suele ser la base de una buena explicación.

Lo anterior nos conduce a otro problema general, el de los etiquetados. Dicho en pocas palabras, "tendrás lo que etiquetes". O, de otro modo, que la información que se puede obtener variará en virtud del tipo de etiquetas. Entonces, ¿qué marcas o etiquetas, de qué tipo y cuántas?

Los etiquetados presentan informaciones siempre útiles, pero no son siempre igual de necesarios, al menos tal y como se presentan actualmente en algunos corpus. Los beneficios de los textos etiquetados son evidentes en las búsquedas léxicas; son menores cuando se pretende una búsqueda sintáctica (a no ser que los enunciados estén analizados) y resultan mínimos, de momento, cuando lo que se pretende es buscar ironías, dobles sentidos, metáforas o relaciones de poder entre los hablantes. Estos objetivos más pragmáticos al estudiar, por ejemplo, la con-

versación coloquial, así como la búsqueda del principio de comprensibilidad de la transcripción, hizo que, en 1995 (y hasta la fecha), el grupo Val.Es.Co. optara por el método que, siendo máximamente informativo, dificultara menos la comprensión del texto. Análogamente, como han señalado Pons Bordería & Ruiz Gurillo (2005: n. 7), "si un investigador desea hacer concordancias de una obra literaria, buscar una determinada palabra o un personaje, recurrirá a una versión etiquetada de la obra que estudia. Si, por el contrario, estudia la evolución del protagonista, la influencia del paisaje o los diálogos entre dos personas, se servirá de una edición crítica". No se trata, pues, de que ambas representaciones sean incompatibles, sino de que sirven a fines distintos.

Así, las búsquedas de una partícula discursiva, una categoría más o menos cerrada y cuya forma es en general invariable, son sencillas en corpus sin etiquetar. Por otro lado, los etiquetados actuales de corpus orales no marcan tales partículas y, si solo se marcara el hecho de "ser partícula", esto es, con una sola marca de identificación de su categoría, no ganaríamos pragmáticamente mucho, ni siquiera tiempo (Portolés 2004).

Como señala este autor, los motores de búsqueda de algunos corpus no permiten algunas opciones elementales. Para estudiar la posición de las partículas en un miembro del discurso determinado, sería muy conveniente que un motor de búsqueda reconociera los signos de puntuación y, añadiríamos nosotros, las pausas e inflexiones finales. Existen partículas que se pueden situar en posición final de su miembro del discurso, pero se trata de una posición infrecuente y en ocasiones difícil de documentar. Si el motor admitiera búsquedas sencillas como "además.", "además?" o "además!" o, en su caso, "además↓" o "además/", se resolvería la cuestión de esta posición discursiva de los marcadores y sería de gran relevancia dicha marca para el análisis pragmático-discursivo.

Al hilo de lo anterior, se pone de manifiesto otro déficit de los corpus: la marcación o etiquetado pragmático. Sin más dilación, hay que enfrentarse a este reto, es preciso intentar proponer un sistema de marcas de esos hechos pragmáticos para la conversación, entre las que destacaríamos la de la "posición discursiva (inicial, intermedia o final)", vinculada, claro está, a una teoría de unidades y a las funciones pragmalingüísticas y sociopragmáticas que estas unidades realizan ("topicalización, ironía, deixis, atenuación, intensificación, conexión, regulación del contacto, actividades de imagen y cortesía", etc.).

Algunas de estas marcas deberían referirse a algunos de estos hechos: *superposiciones*, *reinicios (vacilaciones*, *cambios de plan)*, *estructuras suspendidas*, *orden de palabras*, *discurso directo*, *conexión*, *estructuración*, *modalización (intensificación*, *atenuación)*, *metaforización*, *ironía y humor*, *pronunciación marcada*, *velocidad de habla*, *alargamientos*, *inflexiones finales (descendente*, *ascendente*, *suspendida*, *circunfleja)*, *entonación expresiva*, *deixis (personal*, *espacial y temporal)*, etc.

Y las marcas y la marcación pragmáticas solo pueden establecerlas los especialistas. Otra cosa es que, después, los otros especialistas, los informáticos, las traduzcan a su lenguaje formal y pongan límites al campo.

3.8. Soporte

En la era de la información y ante los continuos avances de la electrónica y de la informática parece que plantear la duda sobre el soporte que han de tener los corpus es de necios. Ahora bien, que los datos se almacenen en soporte digital, electrónicamente (DVDs y discos

duros externos, en la propia red, etc.) —lo que parece necesario para lograr la armonización, la accesibilidad, etc. de los corpus— no significa negar la posiblidad de tener una muestra representativa en papel. Creemos que los volúmenes de corpus en papel (o en un libro digital) son necesarios, aunque, eso sí, sin marcas o etiquetas, pues establecen a las claras las diferencias que existen entre lo que es o se piensa solo como base de datos y lo que se considera una muestra de textos representativa del objeto de estudio. Por eso, algunos corpus tienen muestras publicadas. De ese modo nos podemos recrear en la lectura del texto y permitimos que el investigador pueda elegir el método que considere más oportuno para su análisis.

Pensamos, como ya se ha dicho, que el contacto con el corpus y la mayor contextualización que ello supone lo hace más rentable para la observación de los fenómenos lingüísticos, para dar con datos importantes.

4. ¿QUÉ USO HACEMOS DE LOS CORPUS? LOS PROBLEMAS DE ALGUNOS ANÁLISIS

Finalmente, algo de autocrítica sobre los análisis de corpus.

¿Se hace un buen uso de los corpus? ¿Sirven nuestros análisis para construir o destruir teorías? Y algo muy importante, ¿realmente han sido o son rentables todos los proyectos de corpus desarrollados? La respuesta a estas preguntas es "no siempre".

Algo que nos dice la experiencia de trabajar durante muchos años con corpus es que cada cual mira hacia su propio "cuerpo" y que en parte estamos dando la razón a quienes nos acusaban de descriptivistas, de quedarnos exclusivamente en mostrar el dato, o en cuantificaciones sin valoraciones.

Los frutos obtenidos a través de esta lingüística de corpus son también una vara de medir la eficacia o eficiencia del método y también del corpus concreto. Dos ejemplos. Hemos leído un par de tesis y en estas, como en otros muchos trabajos sobre corpus, es una práctica habitual cuantificar los datos estadísticamente o, al menos, trabajar con porcentajes (lo que en este método es necesario para validar los resultados de los análisis). Pues bien, en una se habla de mayor o menor frecuencia (estadística o porcentual), por ejemplo, de un hecho a partir de los datos y de su cuantificación, pero no se extraen consecuencias ni valoraciones, ni mucho menos se extraen resultados generalizadores, lo que significa coger el rábano solo por las hojas. Listar las frecuencias de un hecho sin más es como presentar en un balance comercial los gastos y las entradas, sin obtener la cuenta final de resultados, esto es los beneficios. Pues así ocurre con algunos análisis de datos. Y en la otra tesis, la cuantificación se ha convertido en un fin en sí mismo y no en el medio para verificar u objetivar los datos del análisis lingüístico, prácticamente inexistente. Parece que se nos olvida a veces, al menos en el ámbito de la pragmática y del análisis del discurso, que somos lingüistas.

Claro que, quizás, esta falta de método en la investigación lingüística no atañe solo a la lingüística de corpus, sino a la propia merma de la metodología de la investigación en general.

He leído últimamente numerosos trabajos sobre cortesía en español, y en uno de ellos se alude a la escasa deferencia de los españoles en la conversación. A esta hipótesis, sin demostrar, se suman algunas percepciones e intuiciones de lo descorteses que somos los españoles al hablar. Uno de los argumentos que se utilizan es que somos poco atenuados, que atenuamos muy poco nuestras acciones (por ejemplo, para pedir solemos usar el imperativo, *dame, trae*…,

hecho impensable por descortés en Hispanoamérica). Ahora bien, cuando se analiza un corpus real de conversaciones, se demuestra que esto es falso: que la ausencia o menor presencia de atenuación no significa descortesía y que, además, lo codificado y lo interpretado como cortés y descortés no siempre coinciden.

Los análisis nos dicen que, en efecto, en la conversación coloquial, en situaciones de inmediatez comunicativa, de coloquialidad, la atenuación es menor, pero de modo más particular, esos análisis precisan que la estrategia de atenuación aparece cuando existe un motivo concreto, por ejemplo, cuando existe un alejamiento ocasional del otro por una imagen comprometida, la problematicidad o el tema polémico de la conversación, el desacuerdo explícito o implicado. Y además, las características de los conversadores, su procedencia geográfica y social, tienen a veces incidencia clara sobre su frecuencia y uso.

Todas estas afirmaciones están empíricamente demostradas. La experimentación nos ha ayudado a construir nueva teoría y a invalidar intuiciones, percepciones e hipótesis sin demostrar.

Y la existencia de otros corpus geográficos y socioculturales nos permitiría la comparación en el uso de esta estrategia y de los modos de realización de la misma, las tácticas verbales y extraverbales.

Y la comparación con otros corpus de géneros orales enriquecería el análisis también.

Y comparar varios géneros nos permitiría mejorar no solo nuestro análisis sino también el método: por ejemplo, como ya decíamos, la entrevista no es el mejor género para estudiar cortesía, para construir la teoría de la cortesía.

Y todo ello podría verificarse aún más con un análisis cuantitativo.

Si las diferencias lingüísticas y culturales en el uso de un hecho pragmático como el de la atenuación o, en general, el de la actividad cortés en el ámbito hispánico se explican mejor atendiendo a los rasgos de situación, a la formalidad y cotidianidad del discurso, a los géneros discursivos y a los rasgos y papeles de los interlocutores en cada interacción, no cabe duda de que necesitamos corpus que atiendan a todos estos parámetros. Y, más aún, corpus, interculturales e interlingüísticos, que nos permitan extraer las diferencias, y las coincidencias, frente a otras lenguas y culturas como, por ejemplo, la inglesa, la holandesa o la sueca.

Quizás, la unión de objetivos, puede ser un modo de relanzar los corpus orales y de imprimir mayor homogeneidad al método y a los análisis.

5. A MODO DE CONCLUSIÓN. PROPUESTA PARA EL FUTURO INMEDIATO

La autora Tognini-Bonelli (2001) proponía distinguir entre una perspectiva *corpus-based*, en la que el corpus es método de investigación, indagación y corroboración de hipótesis, y otra *corpus-driven*, en la que el corpus es método, es parte de la investigación, es el origen o punto de partida para llegar a una teoría nueva o destruir otras previas. En este caso, la observación conduce a las hipótesis y la experimentación con ese mismo corpus sirve para la demostración.

Creemos que ambas perspectivas pueden combinarse en la Lingüística de corpus. Pero es cierto que el lingüista de corpus tiene, tal y como lo entendemos nosotros, el corpus como conductor.

Ha podido leerse nuestra visión, ni más ni menos cierta que la de otros, sobre los corpus y sobre la necesidad de cargar con ellos. Se ha expresado nuestro deseo, que es el de todos

los lingüistas de corpus, de tender a la estandarización de los corpus orales del español, más aún en esta época de culto al "cuerpo" lingüístico, pero no sin antes conciliar de forma estricta las metodologías y los objetivos concretos. Y lo mismo cabría decir de los métodos de transliteración y codificación. Supongo que llegará para bien de todos. Mientras tanto, corrijamos al menos los defectos y emprendamos nuevos retos para el futuro, en la línea apuntada en este trabajo.

En beneficio de la estandarización, valgan algunas recomendaciones, según nuestra forma de entender los corpus orales, las cuales combinan propiedades de los *corpus propiamente dichos o directos* y de los *corpus-bases de datos o indirectos*:

— Que los textos seleccionados se hayan producido en situaciones reales.

— Que estén representadas tanto las muestras obtenidas en su contexto natural de enunciación o a través de otros medios.

— Que estos materiales tengan procedencia clara.

— Que respondan a parámetros claros y explícitos de recolección.

— Que sean los propios investigadores, los grupos de trabajo o personal formado para tal fin quienes recolecten, transcriban y etiqueten esas muestras o que, al menos, sean estos quienes validen la transcripción o el etiquetado.

— Que permitan el acceso directo al objetivo. Si se desea estudiar la conversación coloquial, el corpus debe permitir el acceso directo a ese objetivo. Lo cual es fácil, pues bastaría que en las marcas de cabecera apareciera la etiqueta *conversación coloquial* (validada también por investigadores).

— Que se presenten también directamente en formato textual. Por ello hablamos de corpus propiamente dichos, porque permiten siempre el acceso directo a los textos completos (es la tendencia actual, incluso de los megacorpus). Este acceso directo contentaría a los estudiosos sociolingüistas, sociopragmáticos o pragmalingüistas.

— Que puedan tener objetivos precisos o más generales, y, por tanto, límites más o menos definidos, lo cual no impide que puedan servir a otros fines (incluso, no previstos). Y si, además de objetivos, existen hipótesis previas, se logrará poner más límites al campo. Todo ello ayudará a perfilar la representatividad (definición y selección de las muestras).

— Que sean muestras de tipología más homogénea (lo que también es consecuencia de lo anterior). La homogeneidad no está reñida con la variedad, siempre que esta esté representada de manera homogénea.

— Que estén informatizados, marcados y permitan el acceso electrónico por frecuencias, por concordancias y, al menos, un conjunto de etiquetas comunes, incluidas las de cabecera.

— Que una parte de las muestras directas esté impresa.

Y un reto más inmediato. Sin duda, como ya se observaba, es necesario elaborar un macrocorpus representativo de las variedades situacionales y de géneros discursivos orales o, al menos, de aquellos géneros protípicos de lo oral, lo cual permitiría superar ciertos límites y déficits que se plantean a la investigación, por ejemplo, aquellos relacionados con la variación pragmática, al poder comparar y contrastar registros y géneros coloquiales y formales (entre los que habría que dar cabida a los lenguajes de especialidad, académicos y profesionales, etc.).

En este megacorpus oral, las muestras de conversaciones, tanto formales como coloquiales, deberían ser las más representadas, en tanto prototípicas del habla.

La organización y metodología para la elaboración de ese corpus podría ser similar a la del proyecto *PRESEEA*: una base metodológica común, proyecto-macro, y un conjunto de proyectos-micro que se vayan sumando al general, con objetivos ahora más pragmáticos, pragmalingüísticos o sociopragmáticos, aunque sin renunciar a otros de carácter sociolingüístico y dialectal. Se lograría así un corpus ordenado, definido, suficiente y representativo, que proporcionaría muestras y datos fiables de las variedades de la lengua hablada, esto es, calidad del corpus, Y si conseguida la calidad, se añade ahora cantidad, pues "más" es "mejor".

Nosotros estamos ya poniendo nuestro granito de arena al intentar construir ese banco de conversaciones (coloquiales) en España y en América, aunque hay un hartazgo de quienes han elaborado corpus y de otros que no están por la labor, por lo ingrato, lo arduo y comprometido que resulta, sin olvidar la dedicación plena que necesitan, ya que requieren una revisión constante, una continua verificación, eso sin contar con la difícil tarea de la transcripción, etc.

Dicho corpus, dadas sus amplias dimensiones, estará informatizado, deberá combinar transcripción (ortográfica enriquecida) y etiquetado pragmático. Y su consulta combinará, asimismo, la consulta a través de concordancias o frecuencias y el acceso al texto completo. Y algo muy importante: se valdrá de un servidor que permita el acceso libre a todos o a cada uno de esos corpus.

Pero hasta que todo lo anterior sea una realidad**,** preferimos, usando una metáfora de Pons Bordería & Ruiz Gurillo (2005), la pesca lingüística de corpus orales con anzuelo a la pesca con redes. Los resultados que se obtienen son más modestos, pero muy preciados en la lonja de pescado.

Gracias, querido Guillermo, pues estas reflexiones se han fundamentado en todo el trabajo que llevas realizando de manera generosa durante años en la confección de un gran corpus que sirva para mucho y para muchos. Y gracias por tu lucha para darle un espacio cada vez más importante a lo oral.

REFERENCIAS BIBLIOGRÁFICAS

ALVAR EZQUERRA, M., M. J. BLANCO RODRÍGUEZ & F. PÉREZ LAGOS (1994): "Diseño de un corpus español en el marco de un corpus europeo". En ALVAR EZQUERRA & VILLENA PONSODA (1994: 8-29).

ALVAR EZQUERRA M. & J. A. VILLENA PONSODA (coords.) (1994): *Estudios para un corpus del español.* Málaga: Universidad de Málaga.

BIBER, D. & N. TRACY-VENTURA (2007): "Dimensions of register variation in Spanish". En G. PARODI (ed.): *Working with Spanish corpora.* London: Continuum, 54-89.

BRIZ, A. (2005): "Los corpus del español hablado. Presentación". *Oralia* 8, 7-12.

BRIZ, A. & M. ALBELDA (2009): "Estado actual de los corpus de lengua española hablada y escrita: I+D". En *Anuario del Instituto Cervantes. El español en el mundo*. Madrid: BOE / Instituto Cervantes, 165-226.

BRIZ, A. & GRUPO VAL.ES.CO. (2000): *¿Cómo se comenta un texto coloquial?.* Barcelona: Ariel-Practicum.

BRIZ, A. & GRUPO VAL.ES.CO. (2002): *Corpus de conversaciones coloquiales.* Anejo de la Revista *Oralia.* Madrid: Arco-Libros.

BRIZ, A., PONS, S. & J. PORTOLÉS (coords.) (2008): *Diccionario de partículas discursivas del español.* <www.dpde.es>

CARAVEDO, R. (1999): *Lingüística del corpus*. Salamanca: Ediciones Universidad de Salamanca.

CHAFE, W. (1994): *Discourse, consciousness and time*. Chicago: The University of Chicago Press.

CRESTI, E. & M. MONEGLIA (2005): *C- ORAL- ROM. Integrated Reference Corpora for Spoken Romance Languages.* Amsterdam: John Benjamins.

DAVIES, M. (2008): "Spanish and Portuguese Corpus Linguistics". *Studies in Hispanic and Lusophone Linguistics* 1, 149-186.

DAVIES, M. (2009): "Creating Useful Historical Corpora: a Comparison of *CORDE*, the *Corpus del español*, and the *Corpus do portugués*". En A. ENRIQUE-ARIAS: *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Madrid / Frankfurt am Main: Iberoamericana / Vervuert, 137-166.

DE KOCK, J. (ed.) (2001): *Lingüística con corpus. Catorce aplicaciones sobre el español*. Salamanca: Ediciones Universidad.

EAGLES (1996a): *Preliminary recommendations on subcategorisation.* Pisa: ILC-CNR. <http://www.ilc.cnr.it/EAGLES96/synlex/synlex.htm>.

EAGLES (1996b): *Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages*. Pisa: ILC-CNR. <http://www.ilc.cnr.it/EAGLES96/morphsyn/morphsyn.html>.

GÓMEZ MOLINA, J. R. (coord.) (2001): *El español hablado de Valencia. Materiales para el estudio sociolingüístico. Vol. I. Nivel sociocultural Alto*. Valencia: Universitat de València.

GÓMEZ MOLINA, J. R. (coord.) (2005): *El español hablado de Valencia. Materiales para el estudio sociolingüístico. Vol. II. Nivel sociocultural Medio*, Valencia, Universitat de València.

GÓMEZ MOLINA, J. R. (coord.) (2007): *El español hablado de Valencia. Materiales para el estudio sociolingüístico. Vol. III. Nivel sociocultural Bajo*. Valencia: Universitat de València.

HIDALGO, A. & J. SANMARTÍN (2005): "Los sistemas de trasncripción de la lengua hablada". *Oralia*, 8, 13-36.

HIDALGO NAVARRO, A. (1997): *La entonación coloquial. Función demarcativa y unidades de habla*. Valencia: Universidad de Valencia.

PARODI, G. (2010): *Lingüística de corpus: de la teoría a la empiria*. Madrid: Iberoamericana / Frankfurt am Main: Vervuert.

PONS BORDERÍA, S. & L. RUIZ GURILLO (2005): "Corpus para el estudio de la conversación coloquial: el corpus Val.Es.Co. (Valencia Español Coloquial)". *Oralia* 8, 243-264.

PORTOLÉS. J. (2004): "El Diccionario de partículas discursivas del español y las nuevas tecnologías". *Español Actual* 82, 37-44.

PORTOLÉS, J. (2007): "La información pragmática en un diccionario de marcadores discursivos". *10th International Pragmatics Conference en Goteborg* (Suecia), 10 de julio de 2007. Panel titulado "Corpus orales del español: aportaciones al análisis pragmático", coord. por A. Briz Gómez.

ROJO, G. (2010): "Sobre codificación y explotación de corpus textuales. Otra comparación del *Corpus del español* con el CORDE y el CREA". *Lingüística* 24, 11-50.

SAMPER PADILLA, J. A., C. HERNÁNDEZ CABRERA & M. TROYA DÉNIZ (1998): *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico (MC-NLCH).* Las Palmas de Gran Canaria: Servicio de Publicaciones de la Universidad de Las Palmas de Gran Canaria-ALFAL. CD-Rom.

SÁNCHEZ SÁNCHEZ, M. (2005): "Corpus de referencia del español actual (CREA). El CREA oral". *Oralia* 8, 37-56.

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

STUBBS, M. (1996): *Text and corpus analysis. Computer-assisted studies of language and culture*. Oxford: Blackwell.

TEUBERT, W. (2005): "My version of corpus linguistics". *International Journal of Corpus Linguistics* 10/1, 1-13.

TOGNINI-BONELLI, E. (2001): *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.

MORENO SANDOVAL, A. & J. URRESTI (2005): "El proyecto C-ORAL-ROM y su aplicación a la enseñanza del español". *Oralia* 8, 81-104.

VILA, R. & GRUPO GRIESBA (2001): *Corpus del español conversacional de Barcelona y su área metropolitana*. Barcelona: Edicions Universitat de Barcelona.

VILLENA PONSODA, J. A. *et al.* (2010): "Problemas de anotación e intercambio en los corpus orales. Estrategias para la transformación de textos etiquetados en documentos XML. El caso de los corpus PRESEEA". *Oralia* 13, 261-323.