

Lo dicho anteriormente fija el propósito de este número: presentar de manera conjunta algunos de los corpus de español hablado de los que se puede ya disponer. Más concretamente, se informa de las características de cada uno de estos corpus, de sus objetivos, de su metodología (selección de informantes, variables sociolingüísticas, tamaño de la muestra), de la calidad de los datos, del sistema de transcripción y, en su caso, etiquetado, de su explotación, análisis y aplicación, esto es, de los resultados y frutos de las investigaciones.

Estoy seguro de que falta más información, de que algunos corpus ya elaborados o en fase de desarrollo avanzada solo han sido mencionados de pasada, de modo indirecto o superficial, ya sea en el estado de la cuestión que presentan algunos autores o dentro de los proyectos en que se han integrado: *Corpus de referencia del español actual* (CREA oral), *Macrocorpus para el estudio de la norma lingüística culta* (MC-NC, dentro del proyecto PILEI) o el *Proyecto para el estudio sociolingüístico del español de España y de América* (PRESEEA). Incluso, puede que de alguno ni siquiera haya referencia por olvido o por ignorancia de quien ha coordinado este volumen. Mientras redacto estas líneas de presentación me vienen a mi mala memoria algunos corpus como el utilizado para el estudio del habla de Tucumán, coordinado por Elena Rojas, el de Mérida (Venezuela), coordinado por Carmen Luisa Domínguez y Elsa Mora, o algunos proyectos en marcha, como el corpus COLA, para el estudio del lenguaje adolescente en distintas ciudades de España y de América, dirigido por Annete Myre Jørgensen (Universidad de Bergen), que podrá consultarse muy pronto en Internet. Pido excusas por ello y solicito esos otros datos para que puedan ser añadidos en otro momento. Se trata de sumar esfuerzos y nunca de excluir. A más cantidad y calidad de información sobre estos corpus, mayor beneficio para los analistas del español hablado.

El volumen se inicia con un repaso de los sistemas de transcripción de la lengua hablada –hay quien prefiere hablar de *transliteración* para referirse a la reproducción ortográfica de lo oral o no solo ortográfica (se añaden signos y convenciones ajenos al sistema ortográfico) y reserva el de *transcripción* exclusivamente para la que es fonética–, y de las ventajas e inconvenientes de cada uno de estos sistemas de convenciones.

En mi opinión, que es también la de los autores de este primer trabajo, 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 representa un único fenómeno y que cada uno de los fenómenos aparezca codificado mediante una única convención o, lo que es lo mismo, que exista una relación unívoca entre signo y realidad representada. Con-

sitúa el sistema de transcripción del corpus Val.Es.Co. (2002) (§ 2.4.). Es un método semiestrecho, que permite que el lector pueda reproducir con bastante fidelidad la conversación original y, a la vez, no dificulta su lectura fluida. Los signos son claros, sencillos y económicos y carecen de posible ambigüedad.

Para el profesor que decida trabajar con corpus orales del español, puede resultar útil la siguiente selección y descripción de corpus orales del español. Todos ellos pueden emplearse como materiales en la enseñanza de E/LE. En el cuadro I (pág. 52) se describen algunos corpus de conversaciones y entrevistas de los que se pueden extraer fragmentos para el aprovechamiento didáctico. Junto a corpus que recogen la variedad coloquial, presentamos algunos de carácter formal, que también pueden ser útiles para el trabajo en el aula de E/LE. Se señalarán en cada uno los siguientes parámetros:

- tipo de discurso que recoge el corpus. La mayor parte de los corpus que se describe aquí son conversaciones informales o semidirigidas y entrevistas semidirigidas
- características de los interlocutores participantes: edad, sexo, nivel sociocultural
- sistema de representación de lo oral que emplea
- fecha de grabación
- área geográfica del español.

Los corpus presentados se encuentran publicados en soporte de papel, electrónico o son accesibles en la red virtual. Son los siguientes:

1. Val.Es.Co. (2002), *Valencia Español Coloquial* (<<http://www.uv.es/valesco>>), dirigido por Antonio Briz Gómez. Publicado en papel.
2. GRIESBA (2001), *Grupo de Español de Barcelona*, dirigido por Rosa Vila Pujol. Publicado en papel.
3. CREA oral, finalizado en 2004, *Corpus de Referencia del Español*, dirigido por Francisco Marcos Marín. Acceso electrónico: <<http://www.crea.es>>.
4. COLA, *Corpus Oral de Lenguaje Adolescente*, dirigido por Annette Myre Jørgensen. Acceso electrónico: <<http://www.colam.org>>.

realization of these pragmatic functions. This study illustrates the suitability of spoken corpora as a resource to study language variation.

Spoken corpora are being used to study how language varies depending on the situation, on the text type or domain, on the region where it is spoken, or on variables such as social class, gender, age, etc. Some pieces of research draw on spoken corpora to explore variational pragmatics, in spite of the difficulties involved in working comparatively with spoken corpora. Studies of variational pragmatics may involve the use of two different spoken corpora to compare specific pragmatic features in two different languages or in two different varieties. The *COLA* corpus (*Corpus Oral de Lenguaje Adolescente* – Oral corpus of teenage language) is being used for this purpose. This is a corpus of informal language spoken by teenagers in Madrid and other Spanish speaking cities (e.g., Santiago de Chile, Buenos Aires, La Habana). The *COLA* corpus follows the same pattern as the *COLT* (*Bergen Corpus of London Teenage Language*) and *UNO* (*Språkkontakt och ungdomsspråk i Norden*) corpora, which makes it possible to carry out comparative analyses between the language spoken by Spanish, English and Nordic teenagers. For example, Stenström (2005, 2006) compares the use of some prominent features of teenage language (e.g., intensifiers, pragmatic markers, taboo words) in *COLT* and *COLAm* (*COLA* Madrid). In general, intensifiers are more frequent in the Madrid girls' conversations (Stenström 2005), but taboo words are more often used by the English middle/upper class girls than by the Spanish ones (Stenström 2006). The most popular ones were *fuck* and *joder*, which have the same meaning.

Other researchers have used other spoken corpora for similar studies. Müller (2004) investigates the use of the discourse marker *well* by German EFL speakers and compares the results with the use of this marker by American native speakers (NS). Of the twelve functions of *well* found in the data, nine were used more by the EFL than by the native speakers. Adolphs and O'Keeffe (2005) use the *Cambridge and Nottingham Corpus of Discourse in English* (*CANCODE*) and the *Limerick Corpus of Irish English* (*LCIE*) to compare listenership response tokens (i.e., mm, yeah, umhum) in British and Irish English. They examine the data at the level of