PROBLEMAS METODOLÓGICOS EN LA FORMACIÓN DE CORPUS ORALES

Montserrat Recalde[*] (montserrat.recalde@usc.es)
Victoria Vázquez Rozas (victoria.vazquez@usc.es)
Universidad de Santiago

Resumen

El impulso creciente experimentado por la lingüística de corpus en sus variadas vertientes se manifiesta en una preocupación cada vez mayor por las condiciones de obtención y manejo de los datos recopilados. En la comunicación analizamos estas facetas metodológicas aplicadas a la formación de corpus orales con el objetivo de valorar las consecuencias de las decisiones que atañen a las técnicas de registro de muestras y al tratamiento y codificación de los datos grabados. Nos centraremos, por un lado, en el contraste existente entre la conversación espontánea y la entrevista semidirigida, y por otro en algunas implicaciones de la transcripción y codificación del material sonoro. En ambos casos observamos el peso de una tradición cultural grafocéntrica que las investigaciones sobre el discurso oral deben compensar abordándolo de forma integral y en toda su complejidad, y evitando la subordinación a los modelos escritos que dominan la historia de los estudios lingüísticos.

Palabras clave: corpus, lengua oral, metodología, conversación, entrevista semidirigida.

I. INTRODUCCIÓN: LA IMPORTANCIA DE LOS DATOS LINGÜÍSTICOS

A finales de la década de los 50 y principios de los 60 se produce un cambio en la orientación teórico-metodológica de la lingüística que genera la necesidad de usar datos reales y contextualizados como fundamento de la investigación. Esta nueva orientación ha venido impulsada por la convicción, cada vez más generalizada, de que para comprender el funcionamiento de las lenguas es necesario abandonar el idealismo en nuestro acercamiento a ellas, y desprenderse de posiciones en exceso reduccionistas. Esta convicción ha llevado a cuestionar la validez de los estudios lingüísticos fundamentados sobre ejemplos surgidos de la autoobservación e introspección del lingüista, y ha alentado, a ambos lados del Atlántico, el nacimiento de nuevas disciplinas. El contextualismo británico (Firth 1957; Halliday y Hasan 1985), la sociolingüística variacionista norteamericana (Labov 1972, 1981, 2001), la lingüística del texto o el análisis del la conversación —deudor de la etnografía de la comunicación de Hymes y Gumperz (1964), la etnometodología de Garfinkel (1967) o el interaccionismo simbólico de Goffman (1981)— son algunas muestras de tal transformación. Todas se caracterizan por haber abandonado el método hipotético-deductivo y la intuición del analista como evidencias probatorias de los postulados teóricos para basar sus afirmaciones en la observación y el análisis de datos del uso lingüístico real, extraído de contextos reales y representativo de hablantes reales. Desde esta nueva perspectiva, la lingüística pasará de ser una ciencia especulativa a ser una ciencia empírica.

La necesidad de contar con datos reales para el análisis lingüístico ha supuesto un fuerte impulso para la formación de corpus, convertidos ahora en la base de la descripción, explicación y teorización lingüística en cualquiera de sus múltiples perspectivas o enfoques

[*] Mi asistencia a este congreso ha sido financiada con cargo al proyecto *Dispoñibilidade léxica en Galicia* (Dispogal), referencia INCITE08PXIB204095PR, PGIDT/PGIDIT (Xunta de Galicia).

(desde la descripción fonético-fonológica, gramatical o léxico-semántica, hasta la investigación sociolingüística, etnográfica o pragmático-conversacional).

II. LOS DATOS PARA EL ANÁLISIS: LO ORAL Y LO ESCRITO EN LA LINGÜÍSTICA DE CORPUS

Una de las aportaciones más relevantes de la lingüística de corpus a la investigación sobre las lenguas es el haber incorporado la variación —social, geográfica y estilística— y el modo oral como objetos de pleno derecho del estudio y análisis lingüístico. Esto ha supuesto un cambio revolucionario que ha afectado no solo a la forma de hacer lingüística, sino también al fondo ideológico que impulsa este tipo de investigación. La consideración de las variedades no estándares, especialmente los dialectos sociales, y la conversación coloquial como objetivos legítimos de la lingüística ha implicado, en alguna medida, la superación de ciertos prejuicios en torno a la supremacía de la variedad estándar y del discurso escrito, así como el abandono de los planteamientos prescriptivos que solían guiar la teoría lingüística. En efecto, un buen número de los corpus más conocidos recogen muestras de la lengua oral —algunos en una pequeña proporción como el *CREA* o el *Birmingham Collection of English Text*, pero otros de forma exclusiva (*Santa Barbara Corpus of Spoken American English, CHILDES, PRESEEA, Val.Es.Co*). Para la formación de diferentes corpus también se ha atendido a la diversidad geográfica y social de las lenguas: algunos de ellos se han centrado en el lenguaje infantil y juvenil (*CHILDES*, *COLA*, *Bergen Corpus of London Teenage Language*), mientras otros han realizado un considerable esfuerzo por documentar la variación diatópica y diastrática, como es el caso de *PRESEEA*.

Pese a estos avances, todavía se detectan ciertos desequilibrios en la formación de los corpus que impiden alcanzar un mejor conocimiento de los sistemas lingüísticos y su funcionamiento. En primer lugar, la incorporación de la lengua hablada a los corpus no ha reequilibrado aún la desproporción del discurso escrito sobre el oral. El CREA, por ejemplo, incluye solo un diez por ciento de discurso oral, una descompensación especialmente acusada si pensamos que la proporción entre el número de interacciones orales y escritas en la vida diaria de los hablantes es seguramente la inversa. En segundo lugar, el encomiable esfuerzo realizado por recopilar muestras de uso diferentes a las del nivel culto tampoco ha impedido que, de momento, sea este aún el nivel mejor representado, situación en parte derivada del objetivo expreso de algunos corpus, interesados exclusivamente en la norma culta (*Macrocorpus de la Norma Lingüística Culta* del español) y, en parte, de la procedencia de los datos orales, recabados mayoritariamente del discurso institucional o mediático en el que la norma culta y el registro formal están sobrerrepresentados —sesiones parlamentarias, reuniones de negocios, clases en la universidad (*CREA, Knowles & Lawrence*)— o incluso de intercambios transaccionales en contextos socialmente marcados —como el *Pixi Corpora*, formado mediante grabaciones en librerías, a donde sin duda acuden sobre todo hablantes del nivel culto de la lengua.

Finalmente, no podemos dejar de señalar la escasa representación del registro oral espontáneo tal y como se produce en la conversación coloquial[1]. La conversación constituye, con diferencia, la forma más frecuente y espontánea de interacción humana. Es un género altamente funcional para los hablantes, dado que interviene en los procesos de construcción identitaria, en el establecimiento de las relaciones sociales y en la configuración de las instituciones colectivas. Por todo ello, existe un considerable acuerdo en que la investigación

1 Con algunas excepciones que hay que resaltar, como algunos de los materiales del *Proyecto de estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y Península Ibérica*; el corpus Valesco (cf. Briz et al 2002); parte del C-Oral-Rom; el corpus COLA.

lingüística basada en el uso no sólo no debe descuidar, sino que ha de privilegiar, los datos procedentes del discurso conversacional, en el que se realizan de forma más recurrente y general las funciones comunicativas primordiales para la vida diaria de los hablantes —entre las cuales la transmisión de información proposicional no tiene por qué ocupar el lugar más relevante, como ha constatado la pragmática lingüística. La conversación coloquial no sólo es el género en el que emergen de forma constante nuevas estructuras lingüísticas, sino también aquel en el que mejor se detectan los procesos de gramaticalización y lexicalización, y, en definitiva, el espacio donde se origina y se hace visible el cambio lingüístico.

Pero este notable consenso de la lingüística moderna sobre la importancia de analizar el funcionamiento de la conversación coloquial para avanzar en la descripción y conocimiento de la lengua, no ha podido imponerse a los obstáculos éticos y a las dificultades técnicas y metodológicas que supone recabar muestras de conversación natural, lo que ha impulsado la elaboración de métodos que se caracterizan por su distinto grado de artificiosidad y que han puesto "sobre el tapete" sus posibles y no deseables efectos sobre la validez y fiabilidad de los datos así obtenidos. Es decir, se trata de calibrar si los instrumentos utilizados para la formación de corpus lingüísticos sirven para obtener exactamente el tipo de datos que se desea obtener —por lo general muestras de la lengua oral tal y como aparecen en contextos no intervenidos por el lingüista—, y si las sucesivas aplicaciones de dichos instrumentos a los mismos hablantes darían exactamente los mismos resultados o, dicho de otro modo, si los datos obtenidos a través de métodos artificiales son un reflejo fiable de su distribución real en el universo poblacional.

III. LOS MÉTODOS DE RECOGIDA DE DATOS

Observar el uso lingüístico contextualizado requiere llevar a cabo una reflexión sobre los métodos de recogida de datos, dado que, como hemos dicho, del método dependen la validez y fiabilidad de los datos obtenidos y, más aún, el alcance de las conclusiones extraídas de su análisis. En general, los métodos de formación de corpus de lengua oral se encuadran en dos grandes grupos que examinaremos a continuación: los métodos no intrusivos y los métodos intrusivos.

III.1. Métodos no intrusivos

Existe una relación evidente entre las deficiencias anteriormente mencionadas de la lingüística de corpus y los problemas metodológicos de la obtención de los datos que conforman dichos corpus. Centrándonos específicamente en la formación de corpus orales, no cabe duda de que las mejores muestras de habla natural son aquellas que se recogen reduciendo lo máximo posible la intromisión del investigador de campo en la dinámica del evento comunicativo. Esta opinión la sustentan incluso los investigadores que han defendido los beneficios de los datos obtenidos artificialmente mediante procedimientos de simulación como la *representación de roles* ("role-play"):

> The results of the current study indicate that natural data represent the most valid way of observing different aspects of speech-act (verbal and non-verbal) behavior in social interaction, as there are various types of request forms that cannot be generated if one follows the role-play path. (Félix-Brasdefer 2007: 159)

El principio que preside el empleo de estas técnicas no intrusivas es que el investigador se limita a observar etnográficamente (con o sin participación personal) lo que sucede comunicativamente en situaciones sociales no controladas por él. Comparte sus bases epistemológicas con la antropología clásica y la etnografía de la comunicación: dado que el método y el investigador pueden interferir en los datos, es necesario controlar sus efectos reduciendo su "visibilidad" todo lo posible. No hay, pues, más método que la observación directa y el registro de la conducta lingüística de los hablantes. Beneficiándose de los avances tecnológicos, este sistema de recogida de datos usa sofisticados aparatos digitales para realizar grabaciones de audio o vídeo de gran calidad que han venido a reemplazar a las rudimentarias notas de campo —si bien es cierto que no las han eliminado totalmente[2]—, lo que no sólo garantiza la fidelidad de los datos obtenidos sino también su conservación, recuperación y tratamiento para el análisis. La no intromisión del investigador tiene importantes ventajas al permitir que sean los propios participantes los que establezcan el "contrato conversacional", esto es, los términos en que se desarrollará la interacción oral, incluidos sus roles y metas comunicativas. Son los participantes los que estructuran la interacción en función de sus propias necesidades comunicativa y no el investigador en función de sus intereses y propósitos personales.

El método ha sido ampliamente utilizado en lingüística de corpus para obtener datos de habla natural en diferentes tipos de intercambios comunicativos, transaccionales o no (cf. Chodorowska-Pilch 2002; Golato 2005; Márquez Reiter 2005), pese a lo cual tampoco está exento de críticas. Aunque la mayor dificultad sean las implicaciones éticas relativas al tratamiento de los datos personales de los informantes, su confidencialidad y la restricción de sus usos posteriores —dificultades que pueden solventarse con acuerdos de confidencialidad y anonimato—, probablemente sea el factor ideológico el que más pesa para disuadir al investigador de utilizar datos de conversaciones naturales. Nos referimos con ello a las representaciones de la conversación coloquial como género en alguna medida "defectuoso", dominantes dentro de la comunidad de habla y a las que sin duda han contribuido las caracterizaciones de los propios lingüistas, abundantes en atributos peyorativos como "irreflexivo", "dubitativo", "inconsistente", "erróneo", "degenerado", "improvisado" o "incompleto", por citar sólo algunos ejemplos recogidos por Warren (2006: 95 y ss). Así pues, la persistente desatención de la conversación coloquial no es ajena a las connotaciones de desprestigio asociadas, en el imaginario colectivo, a un género de uso común, no especializado y sentido como poco "distinguido" por los hablantes. Los analistas del discurso, han reforzado esta idea popular asociando el uso conversacional con funciones de carácter interactivo y expresivo frente al uso transaccional e informativo de otros géneros más especializados (cf. Brown y Yule 1983: 2-3). La subsiguiente asociación del uso representativo del lenguaje con el conocimiento, la cultura escrita y la escuela contribuyen a generar el estigma social en torno a la conversación coloquial, atribuyéndole un rango claramente inferior frente a otros usos más "elevados" de la lengua (algunos de ellos también orales):

(...) el registro coloquial no es sino el resultado de un proceso comunicativo ('coloquio') que tiene sus propios y peculiares condicionamientos pragmáticos, a los cuales debe precisamente sus características diferenciadoras. En tal proceso, el carácter abierto e

2 Las notas de campo se siguen utilizando en los estudios pragmáticos para recoger actos de habla caracterizados por su brevedad y su carácter ritualizado: la disculpa, el reproche, la invitación, etc. (Beebe y Cummings 1996; Kasper 2000). Naturalmente para tener acceso a eventos comunicativos más amplios y complejos, y poder analizar su estructura, organización y desarrollo (rasgos prosódicos, toma y alternancia de turnos, lenguaje gestual, etc.) se requiere el uso de métodos más sofisticados, como la grabación digital en vídeo.

irreflexivo (inmediatez, fugacidad, no trascendencia) de la comunicación establecida entre los interlocutores, la alternancia en el uso del canal y la combinación de medios lingüísticos y no lingüísticos en el contexto (compartido) de comunicación, configuran, por una parte, una modalidad en la que el entorno opera con un valor informativo decisorio y superior al de otras modalidades y, por otra, un mensaje caracterizado por la improvisación formal, en el que priman los valores que podríamos llamar "puramente comunicativos", como la fluidez de la transmisión y la expresión del sentido global subjetivo (más que de un exacto y preciso significado verbal). (Vigara Tauste 1992)

Moreno Sandoval y Urresti (2005: 99), tratando de explicar la menor fluidez de los discursos orales no conversacionales sobre los conversacionales, mencionan el deseo de los hablantes de "cuidar su expresión" en ese tipo de discursos, mientras que "en los textos informales parece más importante la necesidad de expresión y comunicación que la de emitir un discurso correcto y preciso". Como vemos, la destreza, rapidez, capacidad expresiva y comunicativa de los hablantes para interactuar verbalmente de una forma no planificada, sobre temas no previstos, con un número de interlocutores variable, sin una distribución rígida de turnos y roles, queda evaluada negativamente como un discurso 'no correcto e impreciso', cuando la complejidad de la tarea conversacional y la competencia necesaria para llevarla a cabo con éxito suponen un grado de sofisticación notablemente mayor que otras actuaciones lingüísticas más institucionalizadas y formales.

Al margen de los citados prejuicios, a los datos naturales se les han achacado, además, otras "debilidades", como su carácter asistemático, la dificultad de obtener con ellos la información necesaria sobre los hablantes para un adecuado análisis sociolingüístico, la falta de representatividad de las muestras y las escasas apariciones de ciertas variables pragmáticas (cf. Beebe y Cummings 1996). Estos autores también acusan al etnógrafo de utilizar frecuentemente datos procedentes de su red social, que no son representativos para caracterizar las normas sociolingüísticas de la comunidad de habla:

(...) the family, colleagues, friends, and acquaintances, not to mention the associated strangers, around a researcher are not necessarily a "speech community". (Beebe y Cummings 1996: 68)

La misma posición adoptan Kasper (2000) o Félix-Brasdefer (2007), quienes aluden a las dificultades de disponer de una cantidad suficiente de ciertos fenómenos pragmáticos a través de datos auténticos, situaciones en las que los datos artificiales provenientes de la entrevista semidirigida o las representaciones de rol se convierten no sólo en alternativas válidas a los datos naturales, sino incluso en alternativas "mejores" o en las "únicas posibles" (cursiva nuestra):

However, open role plays, if constructed with sufficient contextual information, *may offer some advantages* over natural data in that they have the potential of eliciting interactional data for research purposes while controlling for various sociolinguistic variables. (Félix-Brasdefer, 2007: 159)

In fact, *authentic data may just not be a viable option* when an essential component of the research goal is to compare the use of specific pragmatic features by different groups of speakers in a given context (e.g. pragmatic transfer studies which compare how native and non-native speakers respond to compliments under given circumnstances). (Kasper 2000: 320)

III.2. Métodos intrusivos

Las mencionadas dificultades para obtener la cantidad suficiente de datos naturales que permita el análisis lingüístico de fenómenos poco frecuentes, junto a los prejuicios existentes sobre el género conversacional, han impulsado el uso de métodos específicamente diseñados para recoger muestras sustitutivas de habla natural o de discurso auténtico (cf. Wolfson 1981, Kasper 2000), es decir, muestras de eventos comunicativos artificiales diseñados por el investigador para hacerlos semejantes a aquellos que surgen espontáneamente en contextos naturales (cf. Kasper 2000, Felix-Brasdefer 2007, Beebe y Cummings 1996). Existe un número no despreciable de diferentes técnicas de recolección de datos, entre los que destacan la *representación de rol* y la *entrevista semidirigida* —también llamada *entrevista sociolingüística* (Labov 1972, 1980)— que es probablemente el método más usado y elaborado.

La técnica de la entrevista semidirigida forma parte de lo que Kasper (2000: 322) denomina conversaciones "provocadas", dentro de las que incluye cualquier conversación establecida con el único propósito de recolectar datos para el análisis lingüístico[3]. Fue diseñada por Labov en los años sesenta para desarrollar su proyecto sobre la variación y el cambio lingüístico en inglés americano (Labov 1972, 1981, 1995-2001) y desde entonces ha sido ampliamente utilizada en la investigación sociolingüística. El origen de este instrumento de recolección de datos está en la dificultad de obtener muestras socialmente representativas del habla vernácula o espontánea (el estilo ideal para estudiar el cambio lingüístico), lo que animó a Labov a elaborar un sistema que permitiese obtener registros lingüísticos semejantes a los que aparecen en la conversación coloquial (*casual speech*).

Uno de los principales beneficios de la entrevista sociolingüística como método de recogida de datos es la posibilidad de recopilar muestras de habla amplias y representativas, socialmente estratificadas y de buena calidad. Se ha revelado como un método de incuestionable utilidad para la observación sincrónica del cambio lingüístico en marcha y también ha sido usado para recabar datos sobre expresiones referenciales, modalidad, estructuras narrativas, o determinados actos de habla (Kasper 2000: 322). Las aparentes bondades de la entrevista sociolingüística y la comodidad de su aplicación explican que buena parte de la descripción lingüística de la oralidad esté actualmente fundamentada sobre los datos obtenidos a través de este sistema (en el ámbito hispánico, ya hemos mencionado el corpus PRESEEA, o el *Macrocorpus de la norma culta* del español).

Sin embargo, la entrevista semidirigida plantea desde sus inicios problemas que aún no han sido superados, derivados en su mayor parte del carácter artificial e híbrido del método, a medio camino entre la entrevista convencional y la conversación espontánea (cf. Wolfson 1976; Wilson 1987; Briggs 1986; Milroy 1987; Milroy y Gordon 2003). A continuación enumeramos algunos de los más evidentes:

(1) *La "paradoja del observador",* o la dificultad de conseguir que los informantes cuya conducta lingüística se está observando, se comporten como lo harían cuando no son observados. Pese a los grandes esfuerzos invertidos en minimizar este efecto (Labov 1972, 1981, 2001), mediante la construcción de relaciones solidarias con el informante, o la "retirada" del investigador a un segundo plano interaccional, el hecho es que en la mayor parte de las aplicaciones de la entrevista, las relaciones de poder entre entrevistador y entrevistado no solo están presentes, sino que son difíciles de

3 Otro tipo de conversaciones provocadas son las *tareas conversacionales* ("conversation tasks"), en las que los participantes son requeridos para conversar sobre un tema o para alcanzar conjuntamente una meta a solicitud del investigador.

borrar. Como Kasper (2000) reconoce este método promueve una relación asimétrica entre el entrevistador y el entrevistado.

(2) Por otra parte, las estrategias ideadas para reducir los efectos de la paradoja del observador y provocar que la entrevista se asemeje en lo posible a una conversación —como el intercalado de comentarios y opiniones del investigador (Labov 1981: 12, 2001: 88)— tienen, en ocasiones, efectos contrarios a los perseguidos, al convertir la entrevista sociolingüística en un *género híbrido y ambiguo*, que no se corresponde con ningún género del discurso culturalmente reconocido (Wolfson 1976), lo que va en detrimento de la naturalidad de los datos lingüísticos obtenidos.

(3) *La artificiosidad del contexto de entrevista* en relación con los contextos de comunicación naturales deja sus huellas en el registro, más formal, del entrevistado, e incluso puede influir negativamente en los datos obtenidos el estrés, ansiedad o inseguridad que el informante sienta ante un intercambio comunicativo que no controla y que puede parecerle semejante a un examen.

(4) Es un método *poco apropiado para el estudio de la variación estilística.* Por una parte, en la variación estilística la invariante es el hablante y la variable de control la situación comunicativa. Sin embargo, los datos obtenidos mediante la entrevista semidirigida nos informan sobre cómo distintos hablantes se comportan en una misma situación comunicativa (la situación de entrevista)[4]. Por otra parte, las teorías de la "acomodación" (Giles *et al.* 1997[1975], 1991) o del "diseño para la audiencia" (*audience design*) (Bell 1984; 2001) sostienen que el uso lingüístico del hablante es especialmente sensible a las características del interlocutor, a la relación que existe entre los participantes (el "tenor interpersonal" sistémico) y a la dinámica particular de cada interacción (intenciones o propósitos comunicativos o "tenor funcional"). Por lo tanto, no cabe esperar que una situación de entrevista, por más que se apliquen estrategias coloquializadoras, permita obtener tanto muestras de un registro cuidado como del registro conversacional, cuando las condiciones contextuales no varían significativamente. Finalmente, el propio Labov tiene un especial cuidado en señalar que los estilos diferenciados por él en la entrevista, entre los que se encuentra el "estilo casual", no son "unidades naturales de la variación estilística" sino "divisiones formales de un continuum (...) que tiene como objeto la medición de la variación fonológica sobre el eje estilístico." (Labov 1972: 139). Es decir, el "estilo informal" es en la entrevista semidirigida el equivalente del estilo espontáneo en la conversación coloquial, lo que no significa que sean naturalmente intercambiables.

(5) Finalmente, la *variabilidad en los protocolos de aplicación* de la entrevista sociolingüística, con un grado muy diferente de estructuración entre unas y otras, introduce una gran heterogeneidad en las características y calidad de los datos obtenidos. Se puede prever que a mayor estructuración de la entrevista, más se alejará esta del estilo conversacional.

Con la finalidad de acreditar la validez del método, distintos autores se han esforzado en señalar las semejanzas entre la entrevista sociolingüística y el discurso natural. Kasper (2000: 317) resalta que ambos tienen en común el carácter interaccional y que permiten el examen de un amplio rango de rasgos discursivos, como la estructuración de los intercambios de habla, la distribución de los turnos, la secuenciación de las contribuciones conversacionales, la coordinación entre hablante y oyente y la consecución conjunta de las metas transaccionales e interpersonales. Sin embargo, el simple hecho de que dos géneros

4 Incluso puede ser discutible que la situación de entrevista sea la misma, al cambiar siempre al menos uno de sus componentes (el tenor interpersonal).

tengan carácter interaccional no los hace "parecidos", como se desprende de las ostensibles diferencias entre un proceso judicial, una sesión parlamentaria, una consulta médica o una conversación coloquial. A poco que se examinen ambos géneros, no puede más que reconocerse que las diferencias entre la entrevista semidirigida y la conversación coloquial (roles comunicativos, reglas interaccionales, tópicos discursivos, metas del intercambio, etc.) son, por lo menos, tantas como sus semejanzas. Por más que algunos autores hayan denominado a la entrevista semidirigida "conversación semidirigida" (Vázquez Veiga 1998 o Cortés Rodríguez 2004 *inter alia*) o "conversación grabada" (Silva-Corvalán 1989: 27), con el objetivo de resaltar sus puntos en común con el género conversacional, estas denominaciones resultan contradictorias o discutibles, en la medida en que la conversación, en cuanto a género de estructura abierta, que se negocia sobre la marcha y sin roles establecidos de antemano, si es semidirigida ya no es conversación, y en cuanto es "concertada con anterioridad" para informar al hablante de la futura grabación, pierde de forma drástica la espontaneidad e inmediatez conversacional.

Por lo que respecta a la *representación de rol*, es otro de los métodos intrusivos más frecuentemente utilizados para la obtención de datos, en especial de tipo pragmático (Bebee y Cummings 1996; Kasper 2000, Felix-Brasdefer 2007). La artificiosidad de este método es aún mayor que la de la entrevista semidirigida. En este caso, a los participantes se les da instrucciones por escrito para reaccionar ante una descripción situacional y responder oralmente tal como lo harían en una interacción conversacional (formulación de una solicitud, un reproche, un rechazo, etc.). Aquellos que pretenden ganar en naturalidad no ofrecen ninguna instrucción sobre el desarrollo de la "conversación", dejando libertad de acción a sus informantes. Las restricciones e inconvenientes de este método saltan a la vista. El *role-play* es algo parecido a un juego de rol comunicativo, que crea un contexto ficticio (que, además se presenta por escrito) incrustado dentro de un contexto real, que es la situación de investigación creada por el investigador. En circunstancias naturales, el hablante elabora una representación mental de una situación comunicativa compleja que le ofrece numerosos índices contextualizadores sobre lo que está sucediendo comunicativamente, los cuales, a su vez, lo orientan sobre cómo seguir la conversación y guían sus contribuciones. En la representación de rol la situación conversacional no existe y por tanto tampoco los diversos componentes situacionales que condicionan el *output* lingüístico, por lo que su éxito dependerá en gran medida de la capacidad que el hablante tenga de abstraerse de la realidad. Es muy probable que la imagen mental que el hablante construya sobre lo que está sucediendo no sea la que describe el experimento, sino la que le ofrecen los datos del contexto de investigación: un experimento en el que el investigador le requiere que actúe "como si" no estuviese participando en un experimento y "como si" estuviese participando en el intercambio descrito por el experimento. Por otra parte, tampoco aparece uno de los elementos más relevantes del género conversacional: la negociación sobre la marcha, lo que sin duda afectará al tipo de datos lingüísticos obtenidos con este sistema.

IV. NECESIDAD DE TRABAJAR CON DATOS NATURALES

A tenor de la artificiosidad de los métodos de recogida de lengua oral, el problema reside en saber si los corpus lingüísticos obtenidos con estos métodos sirven para extraer conclusiones sobre el funcionamiento de la interacción oral natural, tal y como esta se produce en situaciones no controladas por el investigador o, por el contrario, tienen un alcance mucho más limitado al aportar datos lingüísticos que inevitablemente están influidos por el propio método, y cuya validez y fiabilidad es, por tanto, limitada.

Diversos autores han mostrado los efectos del método de recogida sobre los datos. Wolfson (1976) ofrece datos del inglés sobre la diferente configuración temporal de las narrativas conversacionales y aquellas integradas en la entrevista. Más recientemente, Kasper (2000: 324) cita un estudio de Eisenstein y Bodman (1993) sobre las repercusiones del método en la longitud y complejidad de las expresiones de gratitud en inglés, siendo los datos auténticos los que daban lugar a muestras más largas y complejas. Félix-Brasdefer (2007) también ha subrayado los efectos de la metodología de obtención de datos, indicando la importancia del espacio social, el reconocimiento de los interlocutores y los índices no verbales en las interacciones naturales y su influencia en las características de los actos de habla analizados frente a los métodos no naturales, como la representación de roles. El trabajo de Félix-Brasdefer demuestra que hay diversos rasgos que aparecen con frecuencia en contextos naturales y son infrecuentes o están totalmente ausentes en las representaciones de rol (solicitudes elididas, usos de la partícula negativa, mitigadores, prefacios, anticipadores de respuesta...). También Beebe y Cummings (1996) reconocen que existen ciertos aspectos en los que los métodos artificiales no reflejan el habla natural (el número de palabras, las fórmulas y estrategias comunicativas, el número de tomas de turno, los niveles entonativos, las repeticiones, etc.). Y el propio Labov, durante la aplicación de la entrevista sociolingüística, aprovechaba los datos lingüísticos que surgían cuando la dinámica interaccional se alejaba del formato "entrevista", con la finalidad de obtener muestras más próximas al habla vernácula que las producidas en el seno de la entrevista propiamente dicha.

Así pues, pese a la reconocida utilidad de los métodos artificiales de formación de corpus, los estudiosos admiten su condicionamiento de los datos para el análisis. Todos coinciden en señalar que los datos artificiales "se aproximan" a los datos naturales, lo cual equivale a reconocer sus diferencias. También admiten todos las ventajas de utilizar datos naturales cuando se trata de caracterizar la conversación, lo que presupone que no es esto lo que habitualmente se hace —una apreciación semejante sobre el discurso jurídico, político, científico o periodístico, causaría por lo menos estupor. Siendo la conversación espontánea el único género para el que se utilizan sucedáneos, parece necesario plantear la necesidad metodológica que trabajar con datos naturales, lo que no podrá hacerse si no nos desprendemos de los prejuicios que envuelven a la conversación coloquial y reconocemos su enorme interés para profundizar en el conocimiento de la lengua. Frente a la opinión generalmente asumida de la conversación coloquial como género "menor" caracterizado por el error, la inconsistencia, el caos y todo tipo de deficiencias, algunos autores han constatado ciertos hechos que parecen contradecir el tópico de la simplicidad y el desorden. Warren (2006: 89-90), examinando la dicotomía función transaccional / función interactiva de los intercambios comunicativos, constata la simplicidad de los discursos transaccionales, caracterizados por su especialización, roles y metas prefijados y un "guión" establecido y aprendido de antemano. Todo ello se opone a la sofisticación de las conversaciones coloquiales, que tienen como objetivo "The complex and demanding business of building and maintaining social relationships" (*op. cit.*, 90). Observa también Warren cómo la aparición de discurso conversacional en interacciones especializadas y predominantemente transaccionales, que él denomina "pseudo-conversation", juega un papel comunicativo relevante en el logro de los objetivos del intercambio.

Moreno Sandoval y Urresti (2005: 97 y ss.), al referirse a los problemas que plantea la transcripción de conversaciones frente a textos más formales (previsiblemente más fáciles de transcribir por su mayor proximidad a la lengua escrita), observan sorprendentemente que los "rasgos de producción" que ralentizan la tarea de transcripción por alejarse del formato "acabado" de los textos escritos y que consisten en palabras fragmentadas, apoyos vocálicos, reinicios y autocorrecciones, son mucho más frecuentes en los discursos formales que en el diálogo espontáneo. Por otra parte, los mayores problemas de transcripción provienen de los

"rasgos de interacción (palabras por turno, solapamientos y tasa de velocidad)" (*op. cit.*, 99) propios del habla conversacional, caracterizada por turnos más cortos, más solapamientos y mayor rapidez de producción que el habla formal.

Finalmente, no se nos escapa que el terreno más difícil para un aprendiz de segunda lengua es adquirir fluidez en la conversación coloquial, especialmente si se lo compara con el éxito alcanzado en el discurso más formal o institucional, más prestigioso, sin embargo, que la conversación coloquial

V. EL TRATAMIENTO DE LOS DATOS

Otra faceta relevante de la formación de corpus orales es el sistema de codificación de los datos, que incluye decisiones previas sobre cómo registrarlos −grabación en vídeo y audio−, en parte restringidas por el tipo de datos seleccionado (por ejemplo, la videograbación no intrusiva es materialmente más difícil que el registro solo del audio), y opciones de presentación de los materiales a los usuarios del corpus, esto es, el acceso o no a los registros de vídeo y audio (o junto con la transcripción) y las convenciones de transcripción elegidas.

Las posibilidades técnicas actuales permiten la integración de los formatos de registro junto con las versiones escritas del componente verbal de la interacción, facilitando así la tarea a los estudiosos, que podrán tomar en consideración el carácter multimodal de la comunicación oral (no sólo la faceta estrictamente lingüística sino, conjuntamente con esta y en estrecha interrelación con ella, los aspectos no verbales y paraverbales del habla). Sin embargo, el hecho mismo de que sea el medio escrito el que permita acceder a la representación y al análisis de la oralidad supone un sesgo de partida que no puede minimizarse.

Incluso usando un sistema de transcripción que trate de reflejar ciertas peculiaridades del discurso oral, especialmente del conversacional, la fijación gráfica del habla implica transformar un proceso dinámico en un producto textual estático, implica atribuir secuencialidad a lo simultáneo (proxémico-gestual, paraverbal, suprasegmental), e inevitablemente conlleva perder de vista muchos de los elementos comunicativos presentes en el habla.

Como han señalado diversos autores, la transcripción supone siempre una selección y reducción de los datos, e inevitablemente refleja los intereses descriptivos y los presupuestos teóricos del transcriptor (Ochs 1979; Linell 2005: 32-34, 118, Hidalgo Navarro y Sanmartín Sáez 2005: 32), por lo cual la evaluación de los sistemas de transcripción se subordina al objeto y la finalidad del estudio (cf. Briz y grupo Val.Es.Co 2002: 28). Y si tenemos en cuenta que los estudios lingüísticos están aun hoy preferentemente orientados hacia los modelos escritos, no ha de extrañarnos que buena parte de los corpus orales disponibles se utilicen habitualmente solo a través de una versión gráfica adaptada a las convenciones de la escritura estándar. Hidalgo Navarro y Sanmartín Sáez (2005: 32) identifican dos tendencias en la codificación de la lengua hablada: o bien un sistemas de etiquetas, que facilita las búsquedas automáticas, los análisis cuantitativos y la aplicación de métodos estadísticos, y garantiza la comparabilidad (universal, en una situación ideal) de los datos, o bien una transcripción sin etiquetado pensada para favorecer la legibilidad. En cualquier caso, se empleen etiquetas u otro tipo de signos convencionales, el ideal es, como señalan por ejemplo Briz y grupo Val.Es.Co (2002: 28), un sistema exhaustivo y unívoco de marcas[5]. No obstante, quien se haya enfrentado alguna vez a "tarea ingrata y difícil" (ibíd.: 38) de transcribir el

5 A día de hoy la disyuntiva entre etiquetación y legibilidad puede resolverse gracias a la existencia de programas informáticos que permiten diferentes formatos de salida o visualización de las transcripciones, con más o menos marcas de codificación según las necesidades del analista.

habla habrá comprobado que la univocidad y discreción de los signos no deja de ser un desiderátum. La realidad del proceso de transcripción nos enfrenta con frecuencia a decisiones basadas en la subjetividad del analista, cuando no claramente arbitrarias, y en muchos casos insatisfactorias, que son una manifestación más de la resistencia del discurso oral a someterse al molde de la escritura.

Aquí haremos referencia, a título de ejemplo, a dos aspectos de la transcripción que revelan la influencia de los modelos escritos y normativos en la codificación del discurso oral. Nuestro objetivo es destacar una vez más la necesidad imperiosa de analizar, conjuntamente con las transcripciones, los registros de audio (y vídeo) en el análisis de las muestras de habla[6].

La transcripción convencional de las variaciones tonales y de intensidad reduce drásticamente la gama de matices significativos asociados al discurso oral. El empleo de los signos de interrogación y exclamación sigue las pautas de la escritura convencional, sometiendo a una categorización discreta lo que en la oralidad se presenta como un continuo, y aplicándose de modo general incluso cuando no existe un sustento fónico tras las decisiones de transcripción, como ocurre con los llamados marcadores de control de contacto o apéndices comprobativos −*¿no?, ¿eh?*, *¿sabes?,* etc.−, cuyos interrogantes no responden siempre a la entonación considerada interrogativa (cf. Briz y grupo Val.Es.Co 2002: 34). Asimismo, la indicación de las pausas mediante signos especiales, frecuentemente barras que reflejan la extensión de la pausa (/, //, ///), incluye un claro componente interpretativo cuando se indica que la pausa corta, en ocasiones "imperceptible" o "casi imperceptible" (cf. Briz y grupo Val.Es.Co 2002: 35; PRESEEA 2008: 8), debe marcarse para permitir al lector del texto una correcta interpretación. Tampoco el empleo de etiquetas complementarias a la escritura convencional, como la de "énfasis", facilita siempre la tarea, ya que puede responder a muy variados propósitos y formas.

Por otra parte, el empleo de etiquetas para "palabras cortadas" y la "reconstrucción" de formas cuya realización oral no se ajusta a los usos normativos (escritos) en aras de una mejor comprensión del texto es una práctica habitual en el proceso de transcripción. El componente interpretativo es de nuevo aquí muy evidente. La reconstrucción de, por ejemplo, *para* a partir de *pa*, incluso en casos en que la realización "normal" es la forma corta ("ni *pa* Dios"), es una muestra de la imposición del modelo escrito sobre la "norma" oral. Pero la alternativa de reproducir el uso hablado plantea otros problemas de difícil solución, ya que nos lleva a tomar decisiones sobre qué marcar y cómo hacerlo, es decir, nos obliga a diseñar un nuevo sistema de codificación, necesariamente convencional y discreto, y por tanto con limitaciones en esencia similares a las de la escritura estándar, sin poder garantizar tampoco que en la nueva convención estén previstas las diversas necesidades y objetivos de los posibles usuarios del corpus. Por otra parte, el sometimiento a la norma escrita tiene una faceta indudablemente ventajosa en los procesos de recuperación y búsqueda de formas lingüísticas, pues facilita un sistema unitario, universal y ampliamente conocido para la localización de las unidades de partida, que posteriormente serán analizadas incorporando la información proveniente del registro sonoro (y visual si es el caso). En el caso citado de la preposición *para*, el acceso inicial a los datos puede partir de las transcripciones convencionales, que darían entrada a las realizaciones vocales a través de un sistema de alineación texto-voz, como el que proporcionan algunos programas informáticos de transcripción.

6 Las propuestas de codificación tienen presente la parcialidad de las transcripciones y llaman la atención sobre ello. Por ejemplo, en el documento "de mínimos" del proyecto PRESEEA se dice explícitamente que "Las marcas y etiquetas no pretenden reflejar toda la información que ofrece un archivo sonoro, cuya audición será imprescindible para proceder al análisis de determinados aspectos lingüísticos y comunicativos" (PRESEEA 2008: 6).

El carácter arbitrario, convencional y categórico de las clasificaciones metalingüísticas tradicionales alcanza también a las propuestas de transcripción de los llamados "elementos cuasi-léxicos", parcialmente identificables con las interjecciones, así como ciertas vocalizaciones pertenecientes al componente paraverbal, como por ejemplo *ah, ay, aha, mmm, eeh, pff, bah* (vid. PRESEEA 2008: 7, 8). La existencia de convenciones de representación ortográfica también en estos casos supone acomodar las emisiones reales de los hablantes a unos moldes rígidos, casi siempre sesgados hacia alguna variedad o dialecto de prestigio, que reducen y simplifican la realidad del habla. En este ámbito es especialmente necesario un acercamiento no prejuicioso a datos originales emitidos en un contexto de uso natural.

VI. RECAPITULACIÓN

Los avances en los medios técnicos de investigación y en los conocimientos sobre el funcionamiento la lengua oral permiten tomar decisiones mejor fundamentadas acerca del tipo de datos adecuados a los fines perseguidos. Solo a través de un análisis empíricamente sustentado podremos conocer el funcionamiento real y las características de cada género discursivo y, por lo que atañe específicamente a las interacciones coloquiales, dejaremos de proyectar en el objeto de estudio las dificultades técnicas y los prejuicios teórico-descriptivos que lastran las investigaciones. El "problema" no está en los datos conversacionales, sino en lograr un acercamiento comprensivo que permita dar cuenta de su idiosincrasia sin evaluarlos con respecto a los modelos escritos tradicionales. Para ello es necesario incorporar al análisis los registros de audio y vídeo de las interacciones, así como toda la información contextual a la que podamos acceder.

Sin duda, el "written language bias" al que se refiere Linell (2005) deja su impronta en los métodos que usamos para analizar la oralidad. En esta presentación nos hemos referido únicamente, y de forma fragmentaria, a dos decisiones metodológicas: la técnica de obtención de datos y los formatos de presentación y codifición de esos datos. En ambos casos observamos el peso de una tradición cultural grafocéntrica, a la que no es posible (y sería absurdo) renunciar. Sin embargo, sí parece posible compensar el sesgo secular hacia el modo escrito otorgando a la comunicación oral la posibilidad de ser objeto del análisis lingüístico en su integridad y con toda su complejidad, sin subordinación a los modelos expresivos que han dominado la historia de los estudios lingüísticos.

Referencias bibliográficas

Bebee, L. M. & Cummings, M. C. (1996). Natural speech act data versus written questionnaire data: How data collection method affects speech act performance. In S. M.

Gass & J. Neu (Eds.), *Speech Acts Across Cultures. Challenges to Communication in a Second Language*. Berlín: Mouton de Gruyter, 65-88.

Bell, A. (1984). Language Style as Audience Design. *Language in Society,* 13, 145-204.
Bell, A. (2001). Back in style: reworking audience design. In P. Eckert & J. R. Rickford (Eds.), *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press, 139-169.

Briz, A. & grupo Val.Es.Co (2002). *Corpus de conversaciones coloquiales*. Madrid: Arco/Libros.

Briggs, Ch. L. (1986). *Learning how to ask: a sociolinguistic appraisal of the role of the interview in social science research.* Cambridge: Cambridge University Press.

Chodorowska-Pilch, M. (2002). Las ofertas y la cortesía en español peninsular. In M. E.

Placencia & D. Bravo (Eds.), *Actos de Habla y Cortesía en Español [Speech Acts and Politeness in Spanish]*. Londres: LINCOM, 21–36

Cortés Rodríguez, M. (2004). ¿Ser o estar? La variación lingüística y social de *estar* más adjetivo en el español de Cuernavaca, México. *Hispania*, 87:4, 788-796.

Eisenstein, M. & Bodman, J. W. (1993). Expressing gratitude in American English. In G.

Kasper & S. Blum-Kulka (Eds.), *Interlanguage Pragmatics*. Nueva York: Oxford University Press, 64-81.

Félix-Brasdefer, C. (2007). Natural speech vs. elicited data. A comparison of natural and role play requests in Mexican Spanish, *Spanish in Context*, 4:2, 159-185.

Firth, J. R. (1957). *Papers in Linguistics*. Londres: Oxford University Press.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Giles, H., & Powesland P., (1997[1975]). Accommodation Theory. In N. Coupland (Ed.), *Sociolinguistics: A Reader and Course-book*. Basingstoke: MacMillan, 232-239.

Giles, H., J. Coupland & Coupland, N. (Eds.) (1991). *Contexts of accommodation: developments in applied sociolinguistics*. Cambridge: Cambridge University Press.

Goffman, E. (1981). *Forms of Talk*. Philadelphia: University of Pennsylvania Press.

Golato, A. (2005). *Compliments and Compliment responses: Grammatical Structure and Sequential Organization*. Amsterdam/Philadelphia: John Benjamins.

Halliday, M.A.K. & R. Hasan (1985). *Language, context and text: a social semiotic perspective*, Deakin University Press.

Hidalgo Navarro, A. & Sanmartín Sáez, J. (2005). Los sistemas de transcripción de la lengua hablada, *Oralia*, 8, 13-36.

Hymes, D. & J. J. Gumperz (1964). *The Ethnography of Communication* [=*American Anthropologist* 66, 6].

Kasper, G. (2000). Data Collection in Pragmatics Research. In Helen Spencer-Oatey (Ed.), *Culturally Speaking. Managing Rapport through Talk across Cultures*. Londres: Continuum, 316-342.

Labov, William (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press. Citamos por la traducción española: *Modelos sociolingüísticos.* Madrid: Cátedra, 1983.

Labov, William (1981). Field methods of the project on linguistic change and variation. Sociolinguistic Working Paper nr. 81, Southwest Educational Development Laboratory: Austin, Texas.

Labov, W. (1995-2001). *Principles of Linguistic Change.* Oxford: Blackwell.

Labov, W. (2001). The anatomy of style-shifting. In P. Eckert y J. R. Rickford (Eds), *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press, 85-108.

Linell, P. (2005). *The written language bias in linguistics. Its nature, origins and transformations*. Londres: Routledge.

Márquez Reiter, R. (2005). Complaint calls to a caregiver service company: The case of *desahogo*. *Intercultural Pragmatics,* 2, 481–514.

Milroy, L. (1987). *Observing and analysing natural language: a critical account of sociolinguistic method.* Oxford: Basil Blackwell.

Milroy, L. & Gordon, M. (2003). *Sociolinguistics: Method and Interpretation*. Oxford: Basil Blackwell.

Moreno Sandoval, A. & J. Urresti (2005). El proyecto C-Oral-ROM y su aplicación a la enseñanza del español. *Oralia*, 8, 81-104.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. Schieffelin (Eds.), *Developmental Pragmatics*. New York: Academic Press, 43-72.

PRESEEA (2008). Marcas y etiquetas mínimas obligatorias. Vers. 1.2. 17□02□2008. <http://www.linguas.net/preseea>

Silva-Corvalán, C. (1989). *Sociolingüística. Teoría y Análisis*. Madrid: Alhambra.
Vazquez Veiga, N. (1998). *Marcadores discursivos: respostas mínimas reguladoras na conversación semidirixida. Marcadores discursivos: respuestas mínimas reguladoras en la conversación semidirigida*. Tesis Doctoral, Universidade de A Coruña.

Vigara Tauste, A. M. (1992). Economía y elipsis en el registro coloquial (español). Comunicación presentada al *Congreso Internacional de AESLA: Español 1492-1992*, Granada. <http://www.ucm.es/info/especulo/numero1/vigara.htm>

Warren, M. (2006). *Features of Naturalness in Conversation*. Amsterdam: John Benjamins.

Wilson, J. (1987). The Sociolinguistica Paradox: Data as a Methodological Product. *Language & Communication*, 7(2), 161-177.
Wolfson, N. (1976). Speech Events and Natural Speech: Some Implications for Sociolinguistic Methodology. *Language in Society,* 5, 189-209.

Wolfson, N. (1981). Invitations, compliments and the competence of the native speaker. *International Journal of Psycholinguistics,* 25, 7–22.