Corpus lingüísticos de habla infantil y representatividad: el valor de los datos en repertorios de habla en desarrollo*

Linguistic Corpus and Representativeness: The Usefulness of Data in Child Language Corpus

Milagros Fernández Pérez
Depto. Lengua y Literatura españolas, Tª de la Literatura y Lingüística general
Universidad de Santiago de Compostela
Avda. de Castelao, s/n. Santiago de Compostela, 15782
magos.fernandez.perez@usc.es
Orcid ID 0000-0001-8296-4417

RECIBIDO: 22 DE ENERO DE 2019
ACEPTADO: 14 DE MARZO DE 2019

Resumen: Este trabajo destaca la importancia de la composición sobre la cantidad en los inventarios de datos de habla infantil. Las garantías de representatividad exigidas a catálogos de muestras verbales suelen ceñirse a la dimensión cuantitativa, de modo que las propiedades cualitativas ligadas a la naturaleza del propio repertorio no siempre parecen bien definidas y, de manera particular, apenas se contemplan en fuentes de habla en desarrollo. Nuestra contribución, de orden teórico-metodológico, justifica la necesidad de atribuir relevancia a las muestras de habla infantil sobre criterios cualitativos que alcanzan a características genuinas de la lengua-en-proceso. El lenguaje de los niños no está suficientemente documentado, así que antes que "corpus de referencia" con garantías de representatividad cuantitativa, los inventarios de habla infantil sustentan su valor en el significado de las muestras por sus propiedades idiosincrásicas. En concreto, defendemos tres dimensiones requeridas para la pertinencia de los datos en un corpus de adquisición de la lengua: (a) que contengan registros evolutivos de datos longitudinales; (b) que incluyan variables de contexto idiomático y de entorno habitual que canalizan el input; y (c) que se trate de compilaciones densas de muestras, o de compilaciones con diversidad de sujetos, para que en todo caso revelen los patrones interesantes y no solo los frecuentes.

Palabras clave: Inventarios de adquisición de la lengua. Datos significativos de habla infantil. Registros verbales evolutivos. Composición de corpus de habla en desarrollo. Corpus-driven vs. Corpus-oriented. Lenguaje infantil.

* Este trabajo se inscribe en el proyecto financiado por FEDER/Ministerio de Ciencia, Innovación y Universidades-Agencia Estatal de Investigación, *Adquisición fónica y corpus. Tratamiento en PHON del corpus koiné de habla infantil* (FFI2017-82752-P).

Abstract: This paper emphasizes the importance of composition over quantity in child language corpora. The 'representativeness' guaranteed in corpora of spoken language usually concerns only the quantitative aspects of the data, the qualitative properties associated to the nature of those corpora being not always well defined. Particularly, they are barely considered in language development corpora. The present, theoretical-methodological contribution explains the need of attributing relevance to child language samples, by using qualitative criteria related to the peculiar characteristics of the language-in-process. Child language is not documented enough, so rather than "reference corpora", guaranteed to be quantitatively representative, child language corpora are valuable due to the meaning of the samples and their peculiar properties. More concretely, three aspects are argued to be of relevance in the data of a language acquisition corpus: (a) evolutionary records of longitudinal data; (b) variables of idiomatic context and usual environment, responsible of the input; and (c) dense sampling, or with a diversity of individuals, that can reveal relevant patterns and not only the most frequent ones.

Keywords: Language Acquisition Inventories. Child Language Meaningful Data. Evolutionary Speech Records. Composition of Language Development Corpora. Corpus-driven vs. Corpus-oriented. Child Language.

El objetivo de este trabajo es subrayar la importancia de aquellas variables cualitativas asociadas a corpus de habla infantil que son imprescindibles para dar soporte al valor representativo del inventario de producciones verbales. Defendemos que la relevancia de los datos en los catálogos de desarrollo lingüístico no radica tanto en el tamaño de la muestra cuanto en propiedades de composición que deben definirse y que dan razón de ser al catálogo para convertirlo en espejo de la dinámica adquisitiva a la que aluden. Dada la naturaleza de lengua-en-desarrollo, las fuentes de habla infantil debieran surtir construcciones, procesos y categorías propias de la emergencia verbal en las primeras etapas. No obstante, las rutas y los procedimientos que se han aplicado en la confección de buena parte de los corpus de adquisición han seguido pautas de lengua-producto contemplada en el modelo de norma escrita. Se hace así necesario formular preguntas que garanticen la relevancia de las muestras genuinas en la elaboración de repertorios de habla especial en proceso. Cuestiones como: ¿en qué márgenes, con qué cadencia y cuánto se compila para conformar un corpus de habla infantil?; ¿qué dimensiones psicolingüísticas y contextuales han de examinarse para que los materiales contengan los atributos privativos de la dinámica?; ¿en qué propiedades lingüísticas radican el valor de las muestras y el significado de los datos que contiene el registro? Nuestra contribución busca dar respuesta a estas preguntas en el marco de la Lingüística de corpus, que distingue entre *corpus-driven* y *corpus-oriented* (Tognini-Bonelli; Hunston 2002),

y al abrigo de la Lingüística cognitiva y de construcciones, que aprecia las singularidades evolutivas de la emergencia verbal (Tomasello/Stahl; Lieven/Behrens; Diessel 2013; Gries 2013).

El artículo se ordena alrededor de tres apartados. El primero (Los conceptos de "representatividad" y "valor" en muestras y repertorios) se dedica a precisar en qué consiste la representatividad basada en la cuantificación y por qué el valor de los datos derivado de sus características pertinentes. El segundo apartado (Las peculiaridades del habla infantil: ¿qué características?, ¿qué representatividad?) está centrado en el lenguaje infantil y en la necesidad de considerar sus propiedades genuinas como lengua-en-proceso. El tercero (Garantías y valor de los datos en el desarrollo verbal: requisitos en corpus de habla infantil) contiene la aportación nuclear y se destina al examen y a la estimación de las exigencias que debe cumplir un inventario de habla infantil para que los datos resulten significativos.

Los conceptos de "representatividad" y "valor" en muestras y repertorios

Las investigaciones de corte observacional, basadas en materiales comunicativos que se compilan, se someten comúnmente a ciertas condiciones y con objeto de que sus resultados permitan elaborar modelos reflejo de la realidad verbal que denotan. Por tanto, los ejemplares de producciones rastreados han de ser "representativos" de la comunidad hablante cuyos modos lingüísticos se estudian. Precisamente así se definen en Sociolingüística los criterios de diseño de poblaciones para acometer el registro de muestras verbales y, sobre su estudio, componer patrones de uso verbal en comunidades de habla: la población que provee los datos no solo ha de tener suficiente tamaño, sino que sobre todo ha de contener aquellas características determinantes de las prácticas verbales significativas y relevantes. La importancia concedida a las variables sociolingüísticas como guía para definir la comunidad y establecer el equilibrio entre significado o calidad de la muestra y tamaño de la población no se ha tomado, sin embargo, como norte en todos los casos. En la esfera concreta de la llamada Lingüística de corpus se ha asociado la representatividad con el tamaño de la población o con la cantidad de datos de la selección. El criterio ha sido predominantemente cuantitativo. Los trabajos de, entre otros, D. Biber subrayan la relevancia estadística para que los datos del repertorio tengan garantías. Los cometidos de investigación con corpus demandan que

se cumplan exigencias de representatividad.[1] Con palabras de Biber y Jones (1287), si las pesquisas pretenden "to describe and interpret generalizable patterns of language use", para lo que han de servirse de "quantitative studies with the goal and generalizable findings representing some domain of use", entonces el conjunto de datos ha de ser representativo:

> Representativeness is determined by two considerations: composition and size. The composition of a corpus refers to the text categories included in the design of the corpus. The size of a corpus refers to the number of words or number of texts in the corpus. (Biber/Jones 1288)

Las condiciones de composición y de tamaño se definen habitualmente en términos de frecuencia. La composición del corpus ha de contener bien una muestra proporcional, bien una muestra estratificada en variables (lingüísticas, sociales, contextuales...) significativas. El tamaño ha de estimar sea el número de palabras, sea el número de textos, sea el número de palabras por texto, o el número de palabras por muestra. En cualquier caso, han de operarse comprobaciones estadísticas que verifiquen los ajustes de representatividad cuantitativa.

En el ámbito global de la Lingüística de corpus y del cómputo de registros, hay aún así apreciaciones interesantes en torno al peso relativo de la cantidad –la que se recomienda ponderar respecto de la calidad de los datos y con ello determinar en sentido apropiado si un corpus es representativo, si los datos de un repertorio tienen valor y si son significativos–. Autores como McEnery, Xiao y Tono tildan el concepto de representatividad de "fluido", ya que su soporte radica en el tema mismo o en los objetivos de la investigación que se desarrolle; incluso los propios estudiosos que acuden a corpus para sus pesquisas disponen de criterios para valorar su adecuación y representatividad.

> The research question one has in mind when building (or thinking of using) a corpus defines representativeness. If one wants a corpus which is representative of general English, a corpus representative of newspapers will not do. If one wants a corpus representative of newspapers, a corpus

1. El trabajo de Biber se ha convertido ya en un clásico cuya referencia se aduce siempre que se examina la cuestión de la representatividad. Conviene observar el planteamiento que el autor formula sobre la cuestión es de máximos: "Representativeness refers to the extent to which a sample includes *the full range of variability in a population*" (Biber 243; destacado nuestro); "*The design of a representative corpus is not truly finalized until the corpus is completed*, an analyses of the parameters of variation are required throughout the process of corpus development in order to fine-tune the representativeness of the resulting collection of texts" (Biber 256).

> representative of *The Times* will not do. Representativeness is a fluid concept. (McEnery/Xiao/Tono 18)

En esta misma línea se inscriben las reflexiones de Caravedo, quien señalaba:

> el factor más importante para resolver el problema de la representatividad no es propiamente cuantitativo, y tiene que ver con el ordenamiento interno de los datos en la estratificación de la muestra. Estratificar una muestra supone asignar individuos o entidades particulares a un conjunto, esto es, categorizar o clasificar la representatividad de lo observado depende más de la unidad de referencia de la estratificación (de los requisitos indispensables de pertenencia a un conjunto: esto es, del contenido de la estratificación) que del monto cuantitativo de entidades o de individuos, o de la mera cuantificación. Diría más bien que *una acertada combinación de lo cuantitativo con lo cualitativo en la formación del corpus es requisito indispensable para alcanzar la representatividad*. (69, destacado nuestro)

En resumen, una compilación de datos lingüísticos será significativa siempre y cuando provenga de muestras suficientes que reúnan aquellas propiedades definitorias de la población que quiere revelarse en la muestra (las condiciones de "composición" a las que aludía Biber). De algún modo, la cuestión del valor de los datos de un repertorio depende de su propio fundamento y de las metas que se atribuyan al corpus. No extraña que autores como Berber Sardinha se pregunten si realmente se pueden establecer criterios objetivos para determinar la representatividad. En su opinión,

> A representatividade está ligada à questão da probabilidade. A linguagem é de caráter probabilístico, conforme dito, havendo a possibilidade de estabelecer uma relação entre traços que son máis comuns e menos comuns em determinado contexto. O conhecemento da probabilidade de ocorrência de traços lexicais, estruturais, pragmáticos e discursivos está no cerne da Lingüística de Corpus e, portanto, o conhecimento acerca da probabilidade de ocorrência da maioria dos traços lingüísticos em varios contextos ainda está sendo adquirido. (23-24)

Aún más, la representatividad del corpus debiera abordarse desde la pregunta ¿representativo, para quién?, habida cuenta de que los repertorios como *artefactos* son estimados por los usuarios, quienes acuden a las fuentes de datos como garantes de datos empíricos y como sustento de generalizaciones (ver Hunston 2008). Berber Sardinha (25) es a este respecto rotundo cuando seña-

la que "*O ônus de demonstrar a representatividade* da amostra e de ser cuidadoso em relação à generalização dos seus achados para uma população inteira [...] *é dos usuarios*" (destacado nuestro).

En línea con estas consideraciones, conviene no olvidar que las fuentes de datos que buscan ser "de referencia" (se trate de una comunidad lingüística –el español de Buenos Aires, el inglés de Ciudad del Cabo–, de una variedad idiomática extensa –el español peninsular, el inglés americano– o de toda una población) han de cumplir los requisitos de representatividad en sentido continuado (como señalaba Biber), de modo que las posibles características relevantes vayan incorporándose al repertorio conforme se amplía la cantidad de datos. En una palabra, los llamados corpus "abiertos" y corpus "cerrados" cifran sus diferencias en, precisamente, tales requisitos de extensión: expedita en el primer caso y acotada en el segundo.

Con objeto de acondicionar no solo en máximos la exigencia de representatividad estadística, y con el fin de evitar su proyección excesivamente genérica, sin duda puede resultar útil distinguir entre "corpus con representatividad referencial" frente a "corpus con datos representativos o con propiedades significativas" por su valor en el repertorio. La tipología de inventarios comúnmente establecida de "corpus de referencia" (la exigencia extrema en términos estadísticos que plantea Biber), "corpus especializado" (las condiciones definidas desde su naturaleza y objetivos), "corpus piloto" (los requisitos básicos que se consideran significativos para dar entidad a la muestra), "corpus paralelo" y "corpus comparable" (ver Sinclair) constata el carácter cualitativo y flexible de la noción de representatividad.[2] La cuestión debe formularse no solo en términos de "qué representatividad" (sea por tamaño y también por composición), sino también de para qué es útil el repertorio

2. Donde el "corpus de referencia" se acoge a máximos de exigencia de representatividad de tamaño y composición, el "corpus piloto" responde a requisitos flexibles de composición, mientras que los inventarios "especializados"/"paralelos"/"comparados" son significativos porque cumplen ciertos niveles de lo que Gries (2011b, 84) denomina *granularity*: "a corpus can be somewhat representative *on some level* largely by virtue of the design [...] as with all work corpus, there are innumerable small number of dimensions can be chosen". En todos ellos, su valor descansa en las dimensiones y características que se destacan como prioritarias. Lo que por otra parte es, según Gries, el modo apropiado de abordar la representatividad: un corpus es representativo en alguna de sus dimensiones o en determinadas características (*granularity)*. Con sus palabras: "I think it is posible to achieve some degree of representativeness and balancedness when compiling a general corpus, but only on some level(s) of corpus granularity. A corpus that is perfectly representative and balanced on one level can be completely unrepresentative in terms of the frequency distribution of some specific pattern" (2011b, 86).

y cuál es el potencial de explotación que ofrece: patrones que contiene, características relevantes que facilita, probabilidades de construcciones y ocurrencias que sugiere. No en vano el sentido de la llamada Lingüística de datos reside en su aporte para hallar probabilidades de aparición de patrones sistemáticos (sean construcciones, distribuciones, colocaciones, colo-construcciones) y, por tanto, codificados (ver Halliday 2005; Teubert).

Las peculiaridades del habla infantil: ¿qué características?, ¿qué representatividad?

El habla infantil, como el habla disfuncional y como en su momento las variedades sociolingüísticas, no se ha reconocido hasta hace poco en su interés lingüístico por ser sistema comunicativo atípico y especial. La preponderancia del habla estándar abordada en su horma de lengua escrita y estructuralmente codificada ha sido el norte habitual en la descripción lingüística. Las exigencias de representatividad se traducen casi siempre en términos del patrón referencial, que es un molde común previamente reconocido. Aún más, muy frecuentemente tales exigencias se trasladan de modo reduccionista a todos los casos de repertorios de datos fundamentados en la observación de muestras comunicativas. Como si todas las "comunidades de habla" fuesen similares, y como si las características de los usos verbales en intercambio real estuviesen nítidamente establecidas. Conviene no olvidar que en las prácticas comunicativas se hallan propiedades antes nunca contempladas. Y conviene, asimismo, admitir que hay producciones lingüísticas que se inscriben en patrones distintos al de la lengua común. En todas esas circunstancias toman prioridad las características peculiares, los aspectos cualitativos, sobre los criterios exclusivamente cuantitativos del tamaño de la muestra. Ciertamente, el cómputo de frecuencias en repertorios verbales especiales reclama de modo decisivo que el investigador interprete en términos funcionales los aspectos cualitativos de las probabilidades. Con palabras de Gries (2009, 11): "It is up to the researcher to interpret frequencies of occurrence and co-occurrence in meaningful or functional terms".

El lenguaje infantil muestra singularidades exclusivas, propiedades genuinas que resulta imprescindible delimitar y describir. El habla de los niños en edades tempranas no puede abordarse desde la horma de la lengua-producto, sus unidades y características no son las de la lengua adulta. Investigaciones como la de Peters o las que actualmente encajan con planteamientos teóricos procesuales y cognitivos, como la llamada "gramática cognitiva" y la

"gramática de construcciones" (Tomasello; Bybee 2010; 2013; Croft; Diessel 2008; 2013), encaran las producciones infantiles en su naturaleza consustancial. El carácter evolutivo con cambios sistemáticos del habla infantil exige atención cuidadosa a lo que son códigos regulares por etapas (si se defienden prismas de la "teoría de la continuidad"), o a lo que son fases evolutivas de emergencia y de gestación de habilidades verbales y de estructuras (si la concepción encaja con presupuestos constructivistas) (ver Fletcher; Slobin 2014). Sea como fuere, en lo que es la dinámica y la práctica de la lengua en las primeras etapas, no se cuenta todavía con producciones suficientes de todas las fases de desarrollo que puedan tomarse como soporte referencial. No se dispone, por decirlo de algún modo, de material verbal efectivo suficiente que pudiera servir de horizonte para repertorios de datos con pretensiones de representatividad en máximos. Tratándose de lenguaje infantil quizás el calificativo de "representativo" aplicado a un repertorio de datos debiera sustituirse por el de "significativo" o datos con valor, por las propiedades que incluye. Sin duda, la pretensión de ser representativo en el sentido de ser reflejo de las unidades y propiedades adquisitivas "de referencia" en una lengua particular es, por ahora, excesivamente ambiciosa en la esfera del lenguaje infantil.

Parece claro que los repertorios de habla-en-desarrollo no se acogen a los requisitos de significación y de representatividad que rigen los corpus de la lengua-producto-definitivamente conformada. Son inventarios especiales de datos verbales.[3] Así que la pregunta crucial para abordar el valor de un depósito de datos de habla infantil ha de formularse alrededor de, cuando menos, las características estructurales y psicolingüísticas de las producciones de los niños. ¿Qué propiedades lingüísticas –evolutivas pero sistemáticas– verifican la significación de los datos del repertorio? ¿Qué variables contextuales y psicolingüísticas se contemplan para que los materiales muestren los atributos genuinos y propios del habla en desarrollo? ¿De qué modo, desde qué enfoque se identifican 'procesos', 'construcciones' y categorías en el habla infantil? ¿cómo han de describirse y etiquetarse?

3. Por ser bancos verbales especiales, en mayor medida ha de tomarse en cuenta el principio señalado por Hunston (2008, 156) de que todo corpus es un compromiso entre lo que sería deseable y lo que es posible: lo que el investigador ha planificado en la recolección de muestras suele verse limitado por razones prácticas de distinto cariz. En el lenguaje infantil son sobre todo factores contextuales los que condicionan la recogida de las producciones. Como se verá en el siguiente epígrafe, en estos últimos años se han perfilado dos tendencias: (a) compilarlo todo (*Not Sampling, Getting It All*, Roy), y (b) recoger muestras con cierta incidencia regular (*Dense Sampling*, Tomasello/Stahl; Lieven/Behrens).

Efectivamente, la cuestión capital sobre la pertinencia y el valor de un corpus de habla infantil hay que remitirla sin duda a las propiedades cualitativas que están todavía por determinar o que pueden emerger sin haberlo previsto. La dimensión cuantitativa del tamaño de las muestras o de la población se vuelve subsidiaria de las exigencias propias de rasgos genuinos y peculiares. Pero las rutas y los procedimientos generales que se han aplicado en la confección de buena parte de los corpus de habla infantil han seguido pautas de lengua-producto contemplada en la norma escrita. Hay incluso análisis que se han operado tomando unidades de la lengua adulta y obviando que el lenguaje infantil es una dinámica en proceso y los datos en desarrollo apenas están establecidos. Las cuestiones primordiales que se vuelven imperiosas cuando se trata de elaborar repertorios de habla especial giran en torno a qué parámetros determinan los aspectos cualitativos peculiares y cuáles serán los factores de evolución que canalizan cadencias sistemáticas. Preguntas como ¿cuánto y en qué márgenes temporales se compila para conformar una fuente de datos de habla infantil?, ¿qué etiquetado y qué codificación se sigue en el tratamiento de los datos?, ¿dónde debe radicar el valor de las muestras y el significado de los datos que contiene el repertorio?, son las que debieran orientar el trabajo en la esfera de elaboración de corpus de habla infantil.

Por consiguiente, antes que la cantidad como criterio principal y en exclusiva, ha de estar la composición: son "datos con valor", no corpus de referencia. Además, se trata de inventarios especiales, abiertos, multimodales, que han de dar cabida a propiedades aún por determinar. En otras palabras, la naturaleza y los cometidos de explotación de los repertorios de habla en desarrollo descansan en la idiosincrasia de las muestras, lo que sin duda subraya el carácter genuino y evolutivo de los datos que tales fuentes contienen.

Garantías y valor de los datos en el desarrollo verbal: requisitos en corpus de habla infantil

Aunque comparativamente con otros inventarios de lengua los fondos de datos verbales en edad infantil sean todavía reducidos, no obstante en esta última década se ha producido un notable incremento de repertorios de habla en desarrollo en diferentes entornos idiomáticos (constatación que se hace notar en el catálogo de archivos sobre muestras de adquisición en más de 30 lenguas tipológicamente variadas que se ofrecen en el *Talk Bank-CHILDES* <https://childes.talkbank.org/>), así como trabajos de corte metodológico y epistemológi-

co al respecto (Ambridge/Lieven; Behrens; Blume/Lust; Diessel 2007; 2008; 2013; Gries 2011a; 2011b; 2013; Hoff 2012; Naigles). En la línea, todos ellos, de antecedentes que fueron pioneros sobresalientes por su interés en el lenguaje espontáneo contemplado desde prismas observacionales. Los nombres de Claire y William Stern, Martin Braine, Roger Brown, Michael Halliday (1975), entre otros, resultan señeros por su labor en la recolección de datos de lenguaje en desarrollo. Las muestras compiladas proceden casi siempre de las producciones de los propios hijos, procedimiento este que se mantiene en instrumentos formales como los Inventarios MacArthur: padres que anotan a modo de diario los progresos de desarrollo verbal de sus hijos (ver Fenson y otros). Hay sin embargo alguna excepción a todos estos corpus precursores de caso único: Brown recogió muestras de tres niños, Eve, Adam y Sarah, y, con fundamento en los datos registrados, formuló cinco estadios de desarrollo y estableció distinciones cruciales como la que enfrenta edad cronológica y edad lingüística.

La patente visibilidad de los corpus de habla infantil en las últimas décadas ha repercutido en la prelación de los enfoques observacionales sobre los experimentales (Gilquin/Gries; Gries 2013), o al menos ha promovido la importancia de disponer de evidencia convergente (Penke/Rosenbach; Schönefeld) y, paralelamente, ha destapado la cuestión del valor científico de los datos en la investigación de la adquisición de la lengua. Preguntas sobre cuántos niños, cuántas producciones, qué niños, en qué condiciones, con qué cadencia de seguimiento, se han vuelto comunes por imprescindibles en la esfera de las compilaciones de desarrollo verbal. Se asumen ya dos tendencias:

a) La de recoger todas las actividades comunicativas (*Not Sampling*, *Getting It All*: "se registra todo lo que se produce porque lo más es lo mejor"), factible gracias a avances tecnológicos que posibilitan la grabación cotidiana de producciones verbales en márgenes temporales amplios: el equipo de Deb Roy en el MIT (*Human Speech Home Project*, <http://www.ted.com/talks/deb_roy_the_birth_of_a_word>) ha diseñado cámaras sofisticadas y robustas con autonomía de hasta seis meses en la recolección y depósito de muestras (ver Roy); la fundación LENA (*Language Environment Analysis*, <https://www.lena.org/>), orientada hacia la detección precoz de trastornos comunicativos (en la horquilla de los dos meses a los 4 años), dispone de medios técnicos para el registro cotidiano de producciones sonoras (con autonomía de hasta 16 horas) y de sistemas automáticos para el procesamiento y el análisis de dichas muestras, que suministran informes sobre aspectos comunicativos interesantes (Naigles).

b) La de grabar con cadencia suficiente para registrar muestras densas *(Dense Sampling)*, con datos significativos ("lo más no siempre es lo mejor, conviene hacer matizaciones"), que se ha instaurado recientemente entre los investigadores de la corriente cognitiva-construccionista que opera según el modelo *Item-Based Constructions* (en el *Max Planck Institute for Evolutionary Anthropology* y alrededor de las investigaciones de M. Tomasello y E. Lieven en el *Departament of Developmental and Comparative Psychology*, ver Lieven/Behrens). El trabajo de Tomasello y Stahl expone las ventajas de los repertorios densos y que se valoran especialmente por el incremento en el índice de probabilidad para incorporar aspectos genuinos relevantes: frente a los corpus que recogen muestras con cadencia quincenal, las compilaciones densas de muestras diarias de una hora aumentan de modo considerable las probabilidades de precisar la emergencia de estructuras de frecuencia relativamente baja, como es el caso de la sobrerregularización.[4]

¿Significa esto que el grueso de los repertorios incluidos en *Talk Bank-CHILDES* basados en caso único no son fiables? ¿Hay que entender que los corpus con densidad menor no incorporan propiedades suficientes para resultar útiles? ¿Cómo compatibilizar la variedad de técnicas y rutas de registro y la multiplicidad de pautas de observación con los requisitos metodológicos que se han ido definiendo? Las consideraciones y los argumentos que se exponen a continuación tienen la finalidad de canalizar algunas respuestas a las preguntas formuladas.

Requisitos de composición *de los datos*

En primer lugar, se hace imprescindible reorientar ciertos principios de la Lingüística de corpus en su proyección al habla en desarrollo. Como se viene adelantando, una de las exigencias es la de representatividad: más allá del número de palabras (criterio de Biber para corpus textuales) e incluso por encima del número de informantes, está la riqueza de las muestras, el valor de los datos por las propiedades significativas que contienen. El requisito de la "composi-

4. Si se parte del supuesto de que los niños ofrecen producciones a lo largo de 10 horas al día, los registros quincenales comunes de en torno a una hora no compilan más que el 1.5 % de las producciones totales. Las muestras densas, con registros diarios de una o dos horas, acrecientan el porcentaje hasta niveles del 7 % al 15 %, lo que facilita la captura de patrones menos habituales. En opinión de Tomasello y Stahl (118), "frequency with which certain structures are produced and the precise timing of ontogenetic emergence become crucially important".

ción" adquiere protagonismo sobre las condiciones de "tamaño". Sin duda, la relevancia y el valor han de derivarse de la productividad del repertorio para proveer patrones sistemáticos propios de la comunidad de habla que muestran (Caravedo; Berber Sardinha). En este sentido, la fertilidad del corpus va de la mano de sus posibilidades de explotación (lo que equivale a decir que son los usuarios los que dan soporte al "para qué los datos" del inventario, en la línea de Berber Sardinha (25) y Hunston (2008, 162). Una compilación amplia y suficiente de producciones verbales, así como la inclusión de diversidad de etapas de desarrollo son parámetros esenciales para hallar en las fuentes de datos variedad de características del habla infantil. De modo que, debido a la misma naturaleza evolutiva del proceso de adquisición, el carácter longitudinal de las muestras ha de ser requisito indispensable para que el corpus facilite la comparación entre patrones sistemáticos. Puede decirse que el norte y también el cometido de todo repertorio de habla infantil es proveer datos evolutivos longitudinales susceptibles de explotación para investigar la emergencia de la lengua. Cierto que los inventarios divergen en número de informantes, en número de muestras, en horquillas de edad o en técnicas de tratamiento de los datos, pero en cualquier caso sus bases de elaboración y su productividad serán consecuencia del alcance de explotación de los datos que contengan.[5] La dimensión de la magnitud detallada de los datos es capital para ponderar el valor en potencia del corpus: precisamente, la necesidad de "densidad" deriva de ausencia de *granularity* en las fuentes disponibles, o, lo que es lo mismo, de la necesidad de tomar en cuenta el principio de Zipf (1935; 1949): "unas cuantas

5. Lamentablemente, la explicitud sobre los fundamentos y el contenido de los corpus no siempre es un hecho. No es habitual que se constaten las bases epistemológicas, por lo cual Hunston (2008, 160) subraya la importancia de definir los cimientos del repertorio (según la autora, no hay corpus "inservible") con objeto de clarificar su alcance y sus utilidades: "A corpus is usually intended to be a microcosm of a larger phenomenon [...] the value of the corpus lies in being able to make somewhat more tentative statements about the body of language as a whole". Como muestra de la situación en cierto modo opaca y hasta caótica, los catálogos de datos sobre adquisición del español en CHILDES responden a técnicas variadas de compilación y de enfoques metodológicos diversos en su confección. Hay muestras que resultan de aplicar tareas verbales y pruebas de estimulación (es lo propio en los corpus ColMex, Hess, y Shiro), mientras en otros casos las producciones son espontáneas en contextos de naturalidad conversacional (lo que caracteriza a los corpus Diez-Itza, Koiné y Marrero). El seguimiento longitudinal con cadencia semanal/quincenal en la recogida de materiales y la consiguiente posibilidad de establecer módulos-perfil cada tres meses es ruta metodológica común en algunos de los catálogos, aunque no en todos. Hay una clara ventaja del planteamiento longitudinal que se contempla en al menos once corpus, siendo el prisma transversal la opción mantenida en seis casos (corpus Becacesno, ColMex, Diez-Itza, Hess, Jackson-Thal, Shiro), y sin que se constate método evolutivo definido en el resto de los inventarios.

formas son las más frecuentes", pero "muchas formas son las menos frecuentes". De manera que para hacer visibles aquellos aspectos no tan reiterados, aunque igualmente relevantes, se demanda acentuar la compacidad y la consistencia de las muestras. La cuestión que se plantea es si los objetivos de *granularity* solo es posible cumplirlos mediante densidad derivada de registros verbales diarios *(dense sample)* o si dichos cometidos no serían también factibles acudiendo a fuentes de datos diversas con muestras longitudinales de muchos niños. Con seguridad, el valor por solidez y compacidad del conjunto de repertorios de adquisición para cada lengua ubicados en el *CHILDES* requiere un examen detenido (Fernández Pérez, en prensa).

Valor de los corpus para el usuario

En segundo lugar, y al hilo de la necesidad de replantear algunos aspectos operativos de la Lingüística de corpus, hay que advertir de las diferencias entre "corpus como tarea" y "corpus como artefacto", según la feliz distinción de Hunston 2008. El corpus como artefacto es un depósito ya configurado, un recurso instrumental que suministra datos y evidencias para conducir o dar soporte a una investigación particular. Este repertorio-fuente proporciona ejemplos para analizar, contiene evidencias para probar hipótesis o principios, da soporte empírico a una indagación concreta, traza senderos e ilumina aspectos en dimensiones nuevas. El investigador centrado en el desarrollo de la lengua acude a las fuentes de datos y explota sus posibilidades, no sopesa sus garantías más allá de que contengan propiedades que interesan o de que se acomoden a variables consideradas pertinentes. Es el "valor para el usuario" que destaca Berber Sardinha el que resulta crucial para que la compilación resulte significativa. Así, el estudioso puede destacar en un inventario la presencia de "marcadores del discurso" en ciertos márgenes de edad y otorgarle categoría de variable pertinente, y sin embargo no considerar el parámetro del "sexo" aun cuando haya sido esta variable priorizada en la elaboración del corpus. En una línea similar, un repertorio amplio puede tomarse como surtidor de datos sobre el que el investigador selecciona una porción de muestras/sujetos que se justifican en sus garantías por los objetivos concretos de la pesquisa: puede destacar una serie de niños concretos por razón de su fertilidad verbal (abundantes producciones, frecuentes turnos de palabra, introducción de hilos de conversación, mantenimiento de hilos iniciados); o bien puede dar relieve a un conjunto determinado de muestras que verifican la presencia man-

tenida de conectores ("entonces", "porque", "pues", "pero") en una horquilla longitudinal; o incluso puede escoger fragmentos de producciones en distintos estadios temporales que muestran cambios interesantes para delinear perfiles verbales evolutivos (como el seguimiento de "procesos fónicos" hasta la emergencia de los sonidos propios del sistema). Hay, pues, argumentos específicos para atribuir relevancia particular a ciertos parámetros del corpus-artefacto, su significado deriva de la importancia de su *granularity* en la investigación definida que se está programando.

En cualquier caso, los corpus-artefactos ya disponibles como depósitos de datos pueden convertirse en motor para redefinir el diseño y la elaboración de nuevos repertorios, es decir, para abordar los corpus como tarea (Tognini-Bonelli; Teubert). Esto sin duda sucede en buena parte de las dinámicas de construcción de fuentes nuevas de datos: cada corpus trata de cubrir, ampliar o complementar aspectos y dimensiones que los anteriores no contemplan. Los proyectos más recientes orientados al registro del mayor número de muestras comunicativas en las primeras etapas (los previamente mencionados, el de la fundación LENA, *Language Environment Analysis*, el de D. Roy, *Human Speech Home Project*, y los repertorios densos de Tomasello y Lieven, *Dense Sampling*) responden precisamente a tal necesidad. La precisión sobre la emergencia de una construcción lingüística (por ejemplo, en qué momento se constatan las cláusulas de relativo, o cuándo se afianzan las oraciones condicionales), el rigor metodológico para hallar frecuencias regulares que puedan compararse (como puede ser el caso de las frecuencias en la presencia de marcadores como "pues", "como", "porque"), y también el requisito de comprobación objetiva sobre datos observacionales fiables, están todos ellos en el origen de las compilaciones compactas. Consideraciones como las siguientes así lo atestiguan:

> How much data do we need to investigate the development of a particular phenomenon? Specifically, how much data do we need to determine the age of appearance, the order of acquisition, and the development pathway? (Diessel 2008, 1198)

> The existence of changes with age in children's linguistic productions is not disputed; however, what is not obvious is what the changes indicate about the child's knowledge state, and the extent to which the changes are discrete and qualitative changes as opposed to more continuous. Because numerous theories of children's language acquisition depend on

> characteristics of the nature and trajectory of the child's linguistic knowledge state [...], it is critical to have at hand accurate and precise descriptions of these changes. (Naigles 241)

Propiedades relevantes en los repertorios de habla infantil

En tercer lugar, y como corolario de las advertencias anteriores, se vuelve decisivo determinar qué elementos confieren valor y significado principal a los datos de una fuente de producciones de habla infantil. Sin duda, serán criterio para estimar los inventarios ya disponibles y, primordialmente, para concebir nuevos diseños de catálogo. La variable fundamental que se considera imprescindible en la configuración de todo corpus de habla infantil es la inclusión de datos evolutivos; las muestras compiladas han de corresponder a fases de desarrollo distintas. Todo repertorio de habla infantil disfruta por esta razón de naturaleza dinámica, su carácter es abierto, los datos que contiene son producciones comunicativas y su soporte es casi siempre multimodal. Lo que convierte a estos depósitos de datos en especiales por dos motivos: en primer lugar, por su fértil disparidad (tanto por el variado número de niños en las muestras, como por las diferencias individuales en las producciones y por los márgenes temporales contemplados); y, en segundo lugar, por su *naturaleza documental* antes que referencial. Diessel (2008, 1198-99) destaca estas dos dimensiones de variación en el proceso de adquisición como vertientes de *granularity* que sin duda deben definir los inventarios. El habla cambia con la edad, así que el corpus ha de determinar qué márgenes temporales incluye y por qué. Asimismo, hay diferencias notables en el desarrollo verbal individual consecuencia del peso de factores contextuales y de entorno familiar (Hoff 2006; 2010) y que, por tanto, deben considerarse en el diseño del corpus y con objeto de minimizar la engañosa pero frecuente homogeneización: "most child language studies draw on data from *several children*" (Diessel 2008, 1198, destacado nuestro). Por otra parte, el valor de la diversidad consustancial a los repertorios de habla infantil se incrementa debido a las producciones significativas por parte de niños y de niñas, siendo la *variable sexo* pertinente quizás no tanto por razones biológicas de desarrollo cuanto por motivos de contexto social y familiar que intervienen en todo proceso de adquisición de lenguas incluso tipológicamente distintas. Los estilos comunicativos asociados a las reglas sociales que corresponden a los roles de sexo se constatan en los procesos de desarrollo de la lengua: incluso en entornos idiomáticos en los que no son

habituales sistemas de interacción como el *motherese* o el *Child-Directed-Speech*, los niños y las niñas reconocen comportamientos interactivos diferentes según los roles y los siguen (Ochs/Schieffelin; Snow 1995).

Así que, junto a la variable de datos evolutivos, los parámetros de (i) contexto idiomático (tipo de lengua, si son varias las lenguas en desarrollo simultáneo), (ii) entorno familiar (si es único, si son varios hermanos y orden entre ellos, si asiste a la escuela infantil), y (iii) características individuales (sexo, carácter y disposición) suelen tomarse como capitales en el diseño del corpus. Estas tres vertientes de diversidad y variación en la dinámica adquisitiva se han destacado una vez que se ha abordado el proceso de desarrollo como problema empírico antes que racional (Tomasello). Los corpus interlingüísticos y los planteamientos tipológicos de la adquisición predominan en la actualidad como se hace patente en el *Talk-Bank* CHILDES, y son enfoques comunes en los trabajos de estudiosos como Slobin (1985-1997; 2014), Bowerman/Levinson o Bowerman/Brown. La importancia del contexto sociofamiliar en que viven, así como el peso del *input* comunicativo que los niños reciben, se viene confirmando a raíz de las investigaciones de, entre otros, Snow (1977; 2014), Karrass y otros y Hoff (2003; 2006). Tanto es así que buena parte de los inventarios incluyen hoy procedencias sociofamiliares distintas y destacan su incidencia en la diversidad del desarrollo. Finalmente, las diferencias individuales se contemplan desde prismas cognitivos que subrayan las capacidades particulares de cada niño, pero también como consecuencia de entornos de desarrollo diferentes (Hoff 2003), lo que sin duda conduce a relativizar los repertorios de caso único, y ello aunque los inventarios de caso único puedan, como señalan Lieven y Behrens (234), confrontarse entre sí: "The pitfall of case studies –lack of inter-individual reliability checking– can be overcome if the language development of the child in question is compared to the development of other children's data".

Los repertorios de habla infantil tienen todos ellos la misión propia de la esfera de investigación en corpus: compilar y sistematizar datos procedentes de producciones que documentan el desarrollo verbal. El valor de estos inventarios pasa por su composición antes que por el tamaño de la muestra, de ahí que la pauta cuantitativa de la representatividad –comúnmente aplicada a catálogos que buscan ser fuentes referenciales– haya de redefinirse en términos de cualidades que deban figurar en la compilación y que son decisivas para disponer de datos relevantes y significativos. Parece necesario y sería suficiente con que el corpus explicite sus fundamentos de márgenes evolutivos, deter-

mine su carácter longitudinal, defina la cadencia de registro de las muestras y concrete los parámetros de variación, para que las producciones de emergencia verbal que contenga resulten interesantes para abordar dimensiones descriptivas del habla en las primeras etapas. Ahora bien, si se trata de investigar con precisión la emergencia de ciertos aspectos lingüísticos (como, por ejemplo, ¿en qué momento del desarrollo las concordancias de género en español figuran ya como formas gramaticales?, o ¿cuál es, en la dinámica adquisitiva, la secuencia de aparición de los morfemas en el verbo?, o ¿en qué momento y a través de qué fases los marcadores de discurso pasan a ejercer función de conectores?), quizás el alcance de los datos contenidos en las fuentes y repertorios no siempre sea el esperado. Como señalan Tomasello y Stahl,

> Being practical, we cannot simply ignore the immensely useful data already collected by many dedicated researchers, and, to repeat, the existing data are invaluable for answering many basic questions. But what we must do is to become more self-critical about the sampling process. For example, researchers should always take into account frequency when making age of emergence estimates, especially when comparing structures that occur at different frequencies (or children that use a given item or structure with different frequencies). (118)

Ha sido el interés por el detalle de propiedades específicas y por datar el instante de su emergencia lo que ha promovido registros densos que garanticen datos con proporciones notables de *granularity*. Con el fin de que también las características menos frecuentes puedan manifestarse en las producciones, se incrementan tanto el ritmo del seguimiento como los registros, si bien a costa de reducir el número de sujetos (las muestras densas lo son de caso único) lo que sin duda priva a estos inventarios del componente capital de la Lingüística de corpus, la pluralidad. La cuestión crucial de cómo proceder con rutas de observación científica basada en registros regulares que aseguren frecuencias reales (Tomasello/Stahl) debería de algún modo poder conjugarse con la diversidad: Biber y Hunston (2002, 28-29) insisten en las garantías que comportan los datos procedentes de un abanico de variación amplio.

El recurso que ofrecen los archivos CHILDES con el monto de compilaciones de datos verbales en procesos de desarrollo puede quizás facilitar las exigencias de *granularity* y densidad que se piden, y sin menoscabo de la diversidad. Para el español, por ejemplo, los 20 corpus que se hallan en el *TalkBank-CHILDES* contienen datos de alrededor de 567 niños en diferentes entor-

nos de *input*, con inclusión de variedades de Latinoamérica y peninsulares. Son más de 500 horas de conversación y en torno a 20 000 turnos de habla, que prometen fertilidad en propiedades idiosincráticas y expectativas en la cobertura amplia de su emergencia (Fernández Pérez, en prensa).

Discusión y conclusiones

En Lingüística de corpus el requisito de la representatividad se proyecta en sentidos diferentes dependiendo de la naturaleza de los datos que integran los repertorios. En cualquier caso, siempre ha de sostenerse sobre el equilibrio entre el tamaño de la muestra y la presencia de características pertinentes que dan razón de ser a la fuente documental como reflejo de lo que quiere representar (Caravedo). La composición es decisiva para determinar la cantidad o el tamaño (Biber/Jones). En corpus abiertos y especiales en los que no hay limitación de registros, como es el caso de los inventarios de habla infantil, las condiciones exigidas atañen principalmente a las propiedades de los datos antes que a la cantidad de producciones (Diessel 2008; 2013). Así que no se trata tanto de ser fuentes representativas –dado que no caben por ahora catálogos de referencia–, sino más bien de corpus que contienen datos significativos y con valor que definen precisamente las garantías del inventario por su composición.

La propia naturaleza del habla infantil determina las condiciones que demanda la elaboración de todo banco de datos de adquisición de la lengua para ser relevante en su composición (Fernández Pérez 2011). Efectivamente, en labores de *corpus-driven* los indicadores siguientes regulan la pertinencia de las muestras que se compilan: (a) se han de registrar producciones en márgenes evolutivos, los datos han de ser longitudinales y con ritmo regular en su registro; (b) se ha de contemplar la diversidad natural de la adquisición por razón del contexto idiomático (según el lema ya instaurado por Slobin "From *thought and language* to *thinking for speaking*"), del entorno habitual y familiar que canaliza el input (Snow 2014), y de las capacidades cognitivas individuales (Hoff 2010); y (c) se han de compilar muestras en cadencia o en pluralidad suficiente como para que se revelen no solo los patrones comunes por ser muy frecuentes, sino también aquellos otros que son significativos del desarrollo verbal pero que no resultan tan reiterados (según la famosa ley de Zipf [1949]).

Como es natural, ha sido la importancia de la tarea de abarcar probabilidades de patrones genuinos significativos la que viene marcando pautas en el

ritmo de acopio de ejemplares. Con objeto de lograr índices altos de expectativa en la captación de hormas propias del desarrollo, se ha optado por aumentar las horas de seguimiento de casos individuales (*Not Sampling, Getting It All*, es la ruta del *Human Speech Home Project* del equipo de D. Roy, frente a la tendencia de *Dense Sampling* del grupo de M. Tomasello y E. Lieven), si bien a costa de disipar la diversidad consustancial a la dinámica adquisitiva. Las cotas de *granularity* (Gries 2011a; 2013) que de este modo se logran, y que de modo patente precisan el detalle de propiedades del habla en desarrollo, se verían también cubiertas si se da cabida al abanico de variación en los datos (Hunston 2008), que con holgura y riqueza se hallan integrados en el depósito *Talk-Bank* CHILDES. Es ya habitual que las investigaciones sobre habla infantil *(corpus-oriented)* se sirvan de las muestras que ofrecen distintos repertorios y con objeto de cubrir requisitos de composición que posibiliten el descubrimiento de las construcciones sometidas a pesquisa.

Obras citadas

Ambridge, Ben, y Elena Lieven. *Child Language Acquisition: Contrasting Theoretical Approaches*. Cambridge: Cambridge UP, 2011.

Behrens, Heike. "Corpora in Language Acquisition Research: History, Methods, perspectives". *Corpora in Language Acquisition Research*. Ed. Heike Behrens. Amsterdam: Benjamins, 2008. xi-xxx.

Berber Sardinha, Tony. *Lingüística de corpus*. Barueri: Manole, 2004.

Biber, David. "Representativeness in Corpus Design". *Literary and Linguistic Computing* 8.4 (1993): 243-57.

Biber, David, y James K. Jones. "Quantitative Methods in Corpus Linguistics". *Corpus linguistics: An International Handbook*. Eds. Anke Lüdeling y Merja Kytö. Berlin: Walter de Gruyter, 2008. 1286-1304.

Blume, María, y Barbara Lust. *Research Methods in Language Acquisition: Principles, Procedures and Practices*. Berlin: De Gruyter Mouton, 2016.

Bowerman, Melissa, y Stephen Levinson. "Introduction". *Language Acquisition and Conceptual Development*. Eds. Melissa Bowerman y Stephen Levinson. Cambridge: Cambridge UP, 2001. 1-16.

Bowerman, Melissa, y Penelope Brown. "Introduction". *Crosslinguistic Perspectives on Argument Structure: Implications for Learnability*. Eds. Melissa Bowerman y Penelope Brown. Mahwah, NJ: Erlbaum, 2008. 1-26.

Braine, Martin. "The Ontogeny of English Phrase Structure". *Language* 39 (1963): 1-13.

Brown, Roger. *A First Language: The Early Stages*. Cambridge, MA: Harvard UP, 1973.

Bybee, Joan. *Language, Usage and Cognition*. Cambridge: Cambridge UP, 2010.

Bybee, Joan. "Usage-based Theory and Exemplar Representations of Constructions". *The Oxford Handbook of Construction Grammar*. Eds. Graeme Trousdale y Thomas Hoffman. Oxford: Oxford UP, 2013. 49-69.

Caravedo, Rocío. *Lingüística del corpus: cuestiones teórico-metodológicas aplicadas al español*. Salamanca: Universidad de Salamanca, 1999.

Croft, William. "Radical Construction Grammar". *The Oxford Handbook of Construction Grammar*. Eds. Graeme Trousdale y Thomas Hoffman. Oxford: Oxford UP, 2013. 211-32.

Diessel, Holger. "Frequency Effects in Language Acquisition, Language Use, and Diachronic Change". *New Ideas in Psychology* 25 (2007): 108-27.

Diessel, Holger. "Corpus Linguistics and First Language Acquisition". *Corpus Linguistics: An International Handbook*. Eds. Anke Lüdeling y Merja Kytö. Berlin: Walter de Gruyter, 2008. 1197-212.

Diessel, Holger. "Construction Grammar and First Language Acquisition". *The Oxford Handbook of Construction Grammar*. Eds. Graeme Trousdale y Thomas Hoffman. Oxford: Oxford UP, 2013. 347-64.

Fenson, Larry, Virginia Marchman, Donna Thal, Philip Dale, Elisabeth Bates y Steven Reznik. *MacArthur-Bates Communicative Development Inventoires*. 2.ª ed. Baltimore: Brookes Pub, 2007.

Fernández Pérez, Milagros. "El corpus koiné de habla infantil: líneas maestras". *Lingüística de corpus y adquisición de la lengua*. Ed. Milagros Fernández Pérez. Madrid: Arco Libros, 2011. 11-36.

Fernández Pérez, Milagros. "Hacia un repertorio de datos de adquisición del español: relevancia y significado de los corpus del CHILDES" (en prensa).

Fletcher, Paul. "Data and Beyond". *Journal of Child Language* 41 supplement 1 (2014): 18-25.

Gilquin, Gaëtanelle, y Stefan Gries. "Corpora and Experimental Methods: A State-of-the-art review". *Corpus Linguistics and Linguistic Theory* 5.1 (2009): 1-26.

Gries, Stefan. *Quantitative Corpus Linguistics with R: A Practical Introduction*. London/New York: Routledge, 2009.

Gries, Stefan. "Corpus Data in Usage-based Linguistics: What's the Right Degree of Granularity for the Analysis of Argument Structure Constructions?". *Cognitive Linguistics: Convergence and Expansion*. Eds. Mario

Brdar, Stefan Gries y Milena Žic Fuchs. Amsterdam/Philadelphia: John Benjamins, 2011a. 237-56.

Gries, Stefan. "Methodological and Interdisciplinary Stance in Corpus Linguistics". *Perspectives on Corpus Linguistics: Connections and Controversies*. Eds. Geoffrey Barnbrook, Vander Viana y Sonia Zyngier. Amsterdam/Philadelphia: John Benjamins, 2011b. 81-98.

Gries, Stefan. "Data in Construction Grammar". *The Oxford Handbook of Construction Grammar*. Eds. Graeme Trousdale y Thomas Hoffman. Oxford: Oxford UP, 2013. 93-108.

Halliday, Michael A. K. *Learning How to Mean: Explorations in the Development of Language*. London: Edward Arnold, 1975.

Halliday, Michael A. K. *Computational and Quantitative Studies*. New York: Continuum, 2005.

Hoff, Erika. "The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development via Maternal Speech". *Child Development* 74 (2003): 1368-78.

Hoff, Erika. "How Social Contexts Support and Shape Language Development". *Developmental Review* 26 (2006): 55-88.

Hoff, Erika. "Context Effects on Young Children's Language Use: Effects of Conversational Setting and Partner". *First Language* 30 (2010): 461-72.

Hoff, Erika, ed. *Research Methods in Child Language: A Practical Guide*. Sussex/Oxford: Wiley-Blackwell, 2012.

Hunston, Susan. *Corpora in Applied Linguistics*. Cambridge: Cambridge UP, 2002.

Hunston, Susan. "Collection Strategies and Design Decisions". *Corpus Linguistics: An International Handbook*. Eds. Anke Lüdeling y Merja Kytö. Berlin: Walter de Gruyter, 2008. 154-68.

Karrass, Jan, Julia Braungart-Rieker, Jennifer Mullins y Jennifer Lefever. "Processes in Language Acquisition: The Roles of Gender, Attention, and Maternal Encouragement of Attention over Time". *Journal of Child Language* 29 (2002): 519-43.

Lieven, Elena, y Heike Behrens. "Dense Sampling". *Research Methods in Child Language: A Practical Guide*. Ed. Erika Hoff. Sussex/Oxford: Wiley-Blackwell, 2012. 226-39.

McEnery, Tony, Richard Xiao y Yukio Tono. *Corpus-Based Language Studies: An Advanced Resource Book*. London/New York: Routledge, 2006.

Naigles, Letitia. "Not Sampling, Getting It All". *Research Methods in Child Language: A Practical Guide*. Ed. Erika Hoff. Sussex/Oxford: Wiley-Blackwell, 2012. 240-53.

Ochs, Elinor, y Bambi Schieffelin. "The Impact of Language Socialization on Grammatical Development". *The Handbook of Child Language*. Eds. Paul Fletcher y Brian MacWhinney. Oxford: Blackwell, 1995. 73-94.

Penke, Martina, y Anette Rosenbach. "What Counts as Evidence in Linguistics? An Introduction". *What Counts as Evidence in Linguistics? The Case of Innateness*. Eds. Martina Penke y Anette Rosenbach. Amsterdam: John Benjamins, 2007. 1-49.

Peters, Ann. *The Units of Language Acquisition*. Cambridge: Cambridge UP, 1983.

Roy, Deb. "New Horizons in the Study of Child Language Acquisition". *Proceedings of Interspeech 2009*. Vol. 1. Brighton, UK: Curran, 2009. 13-20.

Schönefeld, Doris, ed. *Converging Evidence: Methodological and Theoretical Issues for Linguistic Research*. Amsterdam/Philadelphia: John Benjamins, 2011.

Sinclair, John. "EAGLES Preliminary Recommendations on Corpus Typology". *EAG-TCWG-CTYP/P*. Birmingham: University of Birmingham, 1996. 22 de enero de 2019. <http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html>.

Slobin, Dan. *The Crosslinguistic Study of Language Acquisition*. 5 vols. Hillsdale, NJ, etc.: Lawrence Erlbaum, 1985-1997.

Slobin, Dan. "Before the Beginning: The Development of Tools of the Trade". *Journal of Child Language* 41 supplement 1 (2014): 1-17.

Snow, Catherine. "Mother's Speech Research: From Input to Interaction". *Talking to Children: Language Input and Acquisition*. Eds. Catherine Snow y Charles Ferguson. Cambridge: Cambridge UP, 1977. 31-49.

Snow, Catherine. "Issues in the Study of Input: Fineturing, Universality, Individual and Developmental Differences, and Necessary Causes". *The Handbook of Child Language*. Eds. Paul Fletcher y Brian MacWhinney. Oxford: Blackwell, 1995. 257-76.

Snow, Catherine. "Input to Interaction to Instruction: Three Key Shifts in the History of Child Language Research". *Journal of Child Language* 41 supplement 1 (2014): 117-23.

Stern, Claire, y William Stern. *Die Kindersprache: Eine psychologische und sprachtheoretische Untersuchung*. Leipzig: Barth, 1907.

Teubert, Wolgang. "My Version of Corpus Linguistics". *International Journal of Corpus Linguistics* 10.1 (2005): 1-13.

Tognini-Bonelli, Elena. *Corpus Linguistics at Work*. Amsterdam: John Benjamins, 2001.

Tomasello, Michael. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Cambridge, MA: Harvard UP, 2003.

Tomasello, Michael, y Daniel Stahl. "Sampling Children's Spontaneuos Speech: How Much Is Enough?". *Journal of Child Language* 31 (2004): 101-21.

Zipf, George. *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Boston: Houghton Mifflin, 1935.

Zipf, George. *Human Behavior and the Principle of the Least Effort*. Reading, MA: Addison-Wesley, 1949.