Revista de Investigación Lingüistica, n.º 21 (2018)
ISSN: 1139-1146
Universidad de Murcia

Lingüística de corpus y habla infantil: fundamentos para el diseño de una muestra de datos con valor

Pablo Figueiredo Palacios
Universidade de Santiago de Compostela
pablo.figueiredo.palacios@usc.es

Resumen: El presente artículo detalla la construcción de una muestra de datos con significación sobre la cual investigar, de manera comparada, la emergencia del componente fónico en el período de 1;6 a 3;6 en dos entornos idiomáticos (español peninsular e inglés americano). Se abordan dos cuestiones sobre la lingüística de corpus, como son la representatividad de los datos y el tratamiento de los mismos (se confrontan las perspectivas *corpus-based y corpus-driven*) en los repertorios de datos de habla infantil. Se analizan, asimismo, los proyectos CHILDES y PhonBank, que constituyen las fuentes de las que provienen los datos. Y se explican detalladamente los criterios psicolingüísticos manejados y su relevancia para el diseño de la muestra de datos.
Palabras clave: lenguaje infantil, fonología, lingüística de corpus.

Title: Corpus linguistics and child language: Foundations for the design of a significative sample.
Abstract: This paper deals with the elaboration of a significative sample which will be used to study, comparatively, the emergence of the phonological component in children aged 1;6 – 3;6 in American English and Peninsular Spanish. Two issues concerning corpus linguistics will be addressed, namely, the representativeness of data and its treatment (corpus-based and corpus-driven approaches shall be discussed) in child language corpora. The sources of our data, the projects CHILDES and PhonBank, will also be analyzed. Additionally, the psycholinguistic criteria which we shall employ will be closely scrutinized, as well as their relevance in the design of the data sample.
Key words: child language, phonology, corpus linguistics.

Trabajo realizado al amparo del proyecto Adquisición fónica y corpus. Tratamiento en PHON del corpus koiné de habla infantil (con soporte económico del Ministerio de Economía, Industria y Competitividad: FFI2017-82752-P)

1. INTRODUCCIÓN

Uno de los temas con mayor vigencia en la ciencia del lenguaje a día de hoy es el estudio del habla infantil. Si bien históricamente ha sido un asunto que ha despertado un gran interés, no es hasta las últimas décadas del siglo pasado cuando se aborda la cuestión desde una perspectiva verdaderamente científica. Sumado ello a la aparición y el desarrollo de las computadoras y, posteriormente, de los corpus electrónicos, en los últimos años se han producido enormes avances en este campo de investigación. Y es que no se concibe el estudio de la adquisición y el desarrollo de la lengua sin el apoyo de la lingüística de corpus. De todas formas, la investigación del habla infantil, por su carácter multidisciplinar dista mucho de ser un campo de estudio homogéneo, tanto en los enfoques teóricos como en los metodológicos.

En este artículo examinamos el diseño de una muestra de datos que pueda ser soporte con garantías para investigar la adquisición del componente fonético-fonológico en el período de 1;6 a 3;6. Más concretamente, la muestra debe acoger el cometido de comparar la emergencia de *procesos fónicos* en dos lenguas –castellano e inglés–, con el objeto último de observar similitudes y diferencias en estos dos entornos idiomáticos. El prisma adoptado marcará distancias con el marco teórico estructuralista clásico (más apropiado para el estudio del componente fónico del habla adulta), y se definirá en la teoría conocida como Fonología Natural (Stampe, 1969). Aquí los *procesos fónicos* son entendidos como "operaciones mentales innatas y universales"[1] que se realizan automáticamente en el momento de hablar. La finalidad de los procesos es doble: por un lado, obviar las posibles dificultades intrínsecas en ciertas combinaciones de sonidos, es decir, naturalizar la articulación del habla; por otro, que la cadena fónica sea más sencilla de decodificar por parte del oyente. Los *procesos* pueden ser, además, específicos de cada lengua, esto es, cada repertorio fonológico particular "escoge" qué *procesos* operan y cuáles no. En resumidas cuentas: los procesos no son más que soluciones fonológicas (específicas de cada lengua) a problemas fonéticos (intrínsecos a determinadas combinaciones de sonidos). Con todo, es precisamente su carácter de universal psicológico el que nos mueve a desarrollar una investigación interlingüística de esta índole.

El artículo se estructura en tres apartados. En el primer apartado se abordan

[1] No en el sentido generativista, sino innatos y universales debido a la configuración propia de los sistemas bucofonador y de percepción humanos. Los universales derivados son, pues, universales de tipo psicológico antes que lingüístico. La noción de proceso fónico es, precisamente, uno de estos universales psicológicos. Otros universales, también psicológicos, podrían ser las metáforas conceptuales o las metonimias, tal y como las entienden Lakoff y Johnson (1980).

algunos aspectos ineludibles en la investigación que hace uso de corpus, como son la representatividad de los datos, la distinción entre enfoques corpus-based y corpus-driven (Tognini-Bonelli, 2001), o los criterios de diseño de una muestra. El segundo apartado se centra en los corpus dedicados a la comparación interlingüística y los requisitos que han de cumplirse en una investigación que emplee tal planteamiento. Se incluye también una nómina de criterios de relevancia, acordes con esta investigación particular, y sobre los que descansa la configuración de la muestra de datos con valor para el estudio comparado de los procesos fónicos en dos lenguas. El tercer apartado contiene la descripción de la muestra perfilada, atendiendo a los criterios de relevancia anteriormente expuestos. Además, se explica la procedencia de los datos y se examinan las fuentes en los proyectos CHILDES (<http://childes.talkbank.org>) (MacWhinney & Snow, 1985) y PhonBank (<http://phonbank.talkbank.org>) (Rose & MacWhinney, 2014), los dos bancos de datos especializados de los que nos hemos servido para elaborar los dos subcorpus.

Finalmente, se concluye el trabajo con un apartado de discusión, en el que se comentan posibles aspectos de mejora de la muestra, como la posible ampliación del número de informantes, la adición de un tercer subcorpus para otra lengua, el cálculo de la edad lingüística de todos los participantes o la búsqueda de características fónicas más específicas. Se discute sobre la idoneidad y el rendimiento de tales implementaciones.

2. LINGÜÍSTICA DE CORPUS Y LENGUAJE INFANTIL

En este apartado se contemplan dos componentes de relativa importancia en la investigación que utiliza corpus lingüísticos de cualquier tipo, ya sean corpus de referencia, corpus comparados, corpus de habla peculiar, etc. El primero de ellos, la representatividad, es un asunto tratado con reiteración por los especialistas en el campo, derivando tres grandes posturas al respecto, que en seguida veremos. El segundo aspecto, de carácter metodológico, tiene que ver la función que detentan los repertorios en la investigación, cómo se rentabilizan y qué utilidades se extraen cuando se trata de dar respuesta a interrogantes. Se trata, en resumen, de la dicotomía corpus-based vs. corpus-driven, planteada, entre otros, por Elena Tognini-Bonelli o Susan Hunston, en los primeros años de este siglo.

2.1. La cuestión de la representatividad en los corpus de habla infantil

Si se tuviese que señalar el aspecto más importante a la hora de construir un

corpus, con total seguridad sería el criterio de la representatividad el primero que se tendría que destacar. Se trata de una noción ampliamente discutida y que ha derivado en distintas posturas.

Douglas Biber, autor con una larga trayectoria en los estudios de corpus, ha hecho especial hincapié en cuatro aspectos a la hora de determinar la *representatividad* de un corpus: (a) los criterios de diseño, (b) el muestreo (sampling), (c) el tamaño, o número de textos incluidos en el corpus) y (d) la composición del corpus, o las diferentes categorías textuales (Biber 1993, Biber et al.1998, Biber y Jones 2003). Si estos cuatro requisitos se contemplan al elaborar un corpus, el repertorio tiene todas las bazas para ser representativo.

Hay autores, sin embargo, que consideran que la representatividad es un objetivo sumamente difícil de alcanzar o inalcanzable. Como ejemplo de los primeros, mencionamos a Susan Hunston, quien afirma que garantizar la representatividad de un corpus se antoja complicado, debido principalmente a que desconocemos la población total de hablantes a los que un corpus aspira a representar (2002: 29). Autores como Tony Berber Sardinha van un paso más allá, y directamente sostienen que la representatividad es un objetivo inalcanzable:

> Para que qualquer amostra seja representativa, é necessário conhecer a população da qual ela provém. No caso da linguagem, a dimensão da população total é desconhecida, não sendo possível estimar qual seria uma amostra representativa. *Logo, estritamente falando, não se pode afirmar que um corpus qualquer seja representativo*. (2004: 23, cursiva nuestra).

Además, hay un tercer grupo de autores que consideran que no se debe hablar de *representatividad o no-representatividad* en términos absolutos. Sería esta postura la que defienden autores como Geoffrey Leech, cuando apunta que "it is best to recognize that these goals are not an all-or-nothing: there is a scale of representativity, of balancedness, of comparability" (2007: 144). También Tony McEnery y Andrew Hardie, quienes afirman que "the measures of balance and representativeness are matters of degree" (2012: 10). Esto es, no se considera la representatividad o la no-representatividad categorías discretas, sino grados en una escala. Con otras palabras, en cada investigación de corpus hay unos intereses particulares, los que, en primer lugar guían el diseño del repertorio de datos y, en última instancia, le dan sentido y representatividad en su selección. Un repertorio de datos enfocado al estudio del aprendizaje de una segunda lengua, por ejemplo, será totalmente distinto en su composición a otro dedicado al estudio de la adquisición de usos pragmáticos de una primera lengua, o a un corpus de referencia.

Pero si la representatividad resulta una cuestión delicada en los corpus de habla adulta, cuando se trata de repertorios de datos de habla infantil el asunto se

vuelve más complejo. Obviando que el habla infantil es un sistema evolutivo con enorme variabilidad interpersonal, el mayor obstáculo al que ha de hacérsele frente es la dificultad por la ausencia de documentación. No es posible, en la esfera del habla infantil, hablar de representatividad de los bancos de datos de habla infantil debido a que apenas se ha documentado el sistema que se procura modelizar. De este modo, no resulta adecuado referirnos a la "representatividad de corpus de habla infantil", sino que

> 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. (...) 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. (Fernández 2017: 7)

A día de hoy, las descripciones sobre el habla infantil están todavía en sus inicios; a diferencia de lo que sucede con el habla adulta, con fuentes documentales desde hace siglos, en el habla de los niños queda mucho camino por recorrer. Hasta entonces, es preciso compilar repertorios de datos de habla infantil para poder reconocer en ellos y sistematizar las características idiosincráticas en etapas de desarrollo.

Así, la muestra de datos que se va a perfilar más adelante podrá ser calificada de significativa en base a los criterios de composición empleados para su diseño. Dichos criterios son los que garantizan la significatividad de los datos, como adelantamos en §2 y veremos con mayor profundidad en §3.

2.2. La cuestión de la orientación: *corpus-based y corpus-driven* en fuentes de habla infantil

Una segunda cuestión de no menor importancia es la distinción establecida por Elena Tognini-Bonelli (2001, §4, §5) entre los enfoques *corpus-based y corpus-driven,* según los usos que se le dé a un corpus en la investigación lingüística. Esta distinción metodológica, que no se debe pasar por alto, se formula en los siguientes términos: un enfoque *corpus-based,* es aquel que emplea un corpus para comprobar determinadas hipótesis de una teoría previamente formulada (Tognini-Bonelli, 2001: 65) y en esta aproximación, además, los datos quedan relegados a un lugar secundario con respecto a dichas hipótesis (2001: 68). En el enfoque corpus-driven, por el contrario, existe un compromiso para con los datos, así que no cabe ignorar las evidencias empíricas que puedan salir a la luz, esto es, las conclusiones tienen que ser rechazadas con fundamento o aceptadas. Asimismo, un estudio que adopte este enfoque no deberá plantear hipótesis a priori, sino preguntas de

investigación, que serán las que guíen el curso de las indagaciones (2001: 84). Las investigaciones corpus-driven requieren la elaboración de un inventario de datos. Los trabajos corpus-based se basan en el rastreo de datos en corpus ya configurados.

En el terreno de la adquisición de la lengua, como se puede suponer, de poco vale acercarse al objeto de estudio con excesivo bagaje teórico, preconcepciones y lugares comunes. Principalmente por las características peculiares propias del habla de los niños, que no equivalen a las categorías de la lengua producto ya conformada. De hecho, la investigación de los últimos años aboga por todo lo contrario : es necesario dar valor a los usos lingüísticos de los niños en sí mismos, prescindir de enfoques prescriptivistas y dejar que los datos hablen por sí solos:

> lo prioritario son los DATOS DE ADQUISICIÓN. Son necesarias pues teorías instrumentales que faciliten el acceso a los materiales que se observan o con los que se experimenta, y que canalicen las descripciones necesarias en cada momento. Teorías analíticas suficientemente flexibles y versátiles que puedan dar cabida a materiales lingüísticos infantiles sin acomodarlos o someterlos al modelo adulto (Fernández, 2003: 274).

Esta perspectiva continúa la senda abierta por Ann M. Peters a partir de los años 70, sobre las unidades lingüísticas en el habla de los niños y la imposibilidad de establecer paralelismos con las unidades de la lengua producto adulta (cf. Peters 1980, 1983).

A la luz de todo ello, la orientación idónea no puede ser otra sino la de corpus-driven, en la que "the general methodological path is clear: observation leads to hypothesis leads to generalisation leads to unification in theoretical statement" (Tognini-Bonelli, 2001: 85). Deben hablar los datos observacionales del habla infantil para, sobre ellos, delinear procesos, categorías y funciones. En el análisis de esos datos se proyecta el método retroductivo –también llamado abductivo– muy en consonancia con el planteamiento corpus-driven, puesto que se trata de una ida de los datos a la teoría y con vuelta de la teoría a los datos, de tal modo que permite ir refinando y reformulando las hipótesis a medida que se analizan los materiales.

2. COMPARACIÓN DE CORPUS DE HABLA INFANTIL. ENFOQUE INTERLINGÜÍSTICO .

Un requisito fundamental que tienen que cumplir los corpus destinados a la comparación interlingüística es que sean tan similares como sea posible en lo que

[2] Para una ampliación sobre esta cuestión, vid. Fernández, 2003; 2005; 2006; 2007; 2015; Tomasello, 2003.

se refiere a la composición, de tal modo que la comparativa entre los repertorios de datos tenga sentido. Al referirnos a la composición de la muestra no nos limitamos únicamente al número de informantes, sino que es importante también que la calidad de los datos sea pareja. Además del entorno verbal de los informantes, las circunstancias de registro de las muestras (naturalidad de las conversaciones, participación o no de adultos, papel que estos juegan en la conversación, interacción con iguales), y los datos evolutivos en márgenes de edad son cruciales para hallar inventarios comparables.

Nuestro subcorpus ha sido diseñado a partir de dos corpus mayores. Para el castellano hemos tomado datos del corpus Koiné (DOI: 10.21415/T5SW39), un repertorio de habla infantil temprana elaborado por el grupo de investigación Koiné, dirigido por la Dra. Milagros Fernández Pérez y asociado al área de Lingüística General de la Universidade de Santiago de Compostela. Para el inglés nos hemos servido del corpus Providence (DOI: 10.21415/T5R30XC), resultado de la colaboración entre la Dra. Katherine Demuth (Macquarie University, Australia) y algunos colegas de la Brown University (EEUU).

Para obtener los dos subcorpus se han manejado cinco criterios de relevancia psicolingüística adecuados a los intereses de la investigación. Estos cinco parámetros, que se verán con mayor detalle en §3, son: (1) edad cronológica de los informantes; (2) seguimiento evolutivo; (3) cadencia, o permanencia evolutiva; (4) lengua que se

[3] Según Hanson (1958: 86), la retroducción es una inferencia con el siguiente aspecto:

1. Some surprising phenomenon P is observed.
2. P would be explicable as a matter of course if H were true.
3. Hence there is reason to think that H is true.

[4] Intentamos recoger el testigo del pionero Dan I. Slobin, quien hace más de tres décadas llevó a cabo un estudio interlingüístico en el que manejaba datos de hasta 14 lenguas (Slobin 1985a, 1985b, 1992, 1997a, 1997b)

[5] Esto, en la esfera del lenguaje infantil, quiere decir que tienen que ser similares en los aspectos relativos a la composición de la muestra (número de informantes, distribución por sexos, edades cronológicas y lingüísticas, seguimiento longitudinal, entorno de las grabaciones, etc.) pero también en las características "externas" del corpus (componentes lingüísticos atendidos, tratamiento y etiquetado de los datos).

[6] Lo integran transcripciones de conversaciones reales y espontáneas entre iguales en el ámbito de la escuela infantil. Está conformado por 71 informantes (34 niños y 37 niñas) con edades comprendidas entre los 18 y los 53 meses y procedentes de zonas urbanas céntricas y periféricas. La lengua inicial mayoritaria es el castellano, aunque la presencia del gallego (como L1 y en producciones ocasionales) también es notable.

El corpus está disponible en el sitio <http://childes.talkbank.org/access/Spanish/Koine.html> [fecha de última consulta: 18/09/2017]

está adquiriendo; y (5) sexo de los participantes.

Todas las intervenciones de los informantes de ambos inventarios que interesan en nuestra muestra –a excepción de una niña, DIA, del corpus Koiné– están transcritas y almacenadas en PhonBank, un macrocorpus que contiene datos de desarrollo fonológico en primeras y segundas lenguas (Rose 2014: 278). Una de las características más relevantes del sistema PhonBank es el software de construcción de corpus y de explotación de los datos, diseñado ad hoc: Phon. Se trata de un software de código abierto que permite tratar, analizar y compartir enormes cantidades de datos de desarrollo fonológico (Rose 2006; 2012; Rose et al. 2013). Phon permite gestionar y organizar en carpetas las transcripciones en archivos de texto, para posteriormente

- asociar las transcripciones con los intervalos de tiempo de sus grabaciones en audio o vídeo;
- pasar de transcripción ortográfica a transcripción fonética (AFI);
- validar las transcripciones de otros investigadores y que otros nos validen las nuestras con un sistema de múltiple ciego;
- dividir las transcripciones en frases, palabras, etc. según los intereses de cada investigador;
- hacer una división silábica, con un código de color y etiquetas descriptivas para cada posición dentro de la palabra;
- alinear, tras la división silábica, las formas fonológicas producidas (IPA Actual) con las formas adultas (IPA Target).
- hacer búsquedas en la base de datos;
- generar informes.

(Rose 2012: 368-72)

Phon, de este modo, se revela como una herramienta indispensable en los es-

[7] Participaron seis informantes (tres niños y tres niñas) de entre 11 meses y 4 años procedentes del sur de Nueva Inglaterra (EEUU), y el corpus está compuesto por las transcripciones de conversaciones espontáneas entre los niños y sus madres mientras jugaban en sus respectivos hogares. Todos presentaban una MLU (Mean Length of Utterance, o Longitud Media del Enunciado) dentro de los márgenes de lo esperable en sus respectivas edades cronológicas (Demuth et al., 2006: 143-44).

Está disponible en el sitio <http://phonbank.talkbank.org/access/Eng-NA/Providence.html> [fecha de última consulta: 18/09/2017]

[8] El PhonBank es fruto de la colaboración entre el psicolingüista norteamericano Brian MacWhinney (Carnegie Mellon University, EEUU) y el lingüista canadiense Yvan Rose (Memorial University, Canadá) y su objetivo es construir "a large database of accurately transcribed data on phonological development" para así "test alternative theories of phonology and phonological development" (Rose & MacWhinney 2014: 381).

[9] Disponible para descarga de forma gratuita en el sitio <https://www.phon.ca/>.

tudios de adquisición fónica. Mientras que los componentes sintácticos, léxicos o pragmáticos, por sus características materiales, son accesibles en mayor medida sin necesidad de usar una herramienta electrónica, los sonidos requieren una documentación y un análisis mucho más precisos; de ahí que, en proporción, los estudios de adquisición fónica sean mucho menos numerosos que los de adquisición de los demás componentes de la lengua:

> phonologists interested in the organisation of sound systems (e.g. phones, syllables, stress and intonational patterns) and their acquisition had not enjoyed the same level of computational support prior to the inception of the PhonBank project within CHILDES. There was no developed platform for phonological analysis and no system for data sharing. This situation negatively affected the study of natural language phonology and phonological development. (Rose et al. 2013: 30)

Nuestros datos tomados del corpus Koiné y del corpus Providence están tratados en Phon, por lo que constatamos que el proyecto PhonBank y su software asociado proporcionan a los investigadores un espacio de trabajo y unas herramientas de enorme valor y con amplísimas posibilidades para la explotación y el análisis de corpus de datos fónicos.

3. DETALLES DE LA MUESTRA

Las muestras seleccionadas descansan en pautas que le dan valor y significación. Como adelantábamos en §2, se han manejado criterios de relevancia adecuados a los intereses de la investigación. La perspectiva corpus-oriented ha guiado los ingredientes de composición de los inventarios diseñados. Como si fuese un tamiz cada vez más fino, los criterios de diseño van cribando la muestra. Partiendo de los corpus completos se van aplicando criterios, y cada sucesivo parámetro filtra únicamente los niños que cumplen el criterio anterior, hasta que, aplicados los cinco criterios, obtenemos la muestra de datos definitiva. Los criterios que se han manejado son los siguientes:

(1) En primer lugar, indiscutiblemente, está el parámetro de la edad. Los niños comienzan sus emisiones idiomáticas alrededor del primer año de vida; sin embargo, no será aproximadamente hasta el hasta el año y medio cuando empiece a emerger la habilidad fonológica propiamente dicha. Las características fonológicas de esa primera etapa, a la que Ingram llama Phonology of the First 50 Words, son diferentes de los aspectos fonológicos de etapas posteriores:

> In several ways the phonological aspects of the words during this time, which we can call the Phonology of the First 50 Words, is [sic] different from the development which follow. It therefore constitutes a separate stage of acquisition that merits separate investigation (Ingram 1976: 12).

También Pamela Grunwell, quien establece una periodización de los procesos fonológicos (1981: 175), señala el hito de las 50 primeras palabras (aproximadamente cuando el niño cuenta con 2 años de edad, apunta esta autora) como punto de comienzo de la adquisición de un sistema fonológico propiamente dicho.

Es precisamente este desarrollo fonológico posterior el que nos interesa, y da comienzo a partir del año y medio de edad. Por tanto, un año y seis meses (1;6) será el límite inferior de nuestra franja de edad. Como la mayor parte de los procesos fónicos se dominan al acercarse a los 4 años de edad (Ingram 1986: 223), hemos decidido establecer el límite superior de la horquilla de edad en los tres años y medio (3;6). De esta manera, el criterio de edad que manejamos incluye a todos aquellos niños con edades comprendidas entre un año y medio y tres años y medio (1;6 – 3;6).

De la totalidad corpus Koiné estarían comprendidos dentro de esa horquilla de edad 25 informantes y del Providence los 6 que lo integran, si bien el seguimiento de todos ellos no comienza exactamente al año y medio de edad ni acaba a los tres y medio; el seguimiento de algunos empieza y acaba antes, el de otros empieza y acaba dentro del período, y el de otros empieza dentro de la horquilla y acaba meses más tarde. Lo realmente importante, como de hecho es el caso, es que en el periodo contemplado haya seguimiento.

(2) El segundo criterio de diseño empleado para obtener el corpus piloto es el seguimiento evolutivo de cada informante. En tanto dinámica lingüística evolutiva, el habla infantil pide, además de un estudio transversal entre iguales, un estudio longitudinal que muestre la evolución de cada individuo. Es en la intersección de estos dos tipos de estudio donde podremos detectar si el desarrollo lingüístico de un niño transcurre con normalidad o, por el contrario, si su desarrollo es atípico o disfuncional. Teniendo esto en cuenta, en la muestra piloto hemos decidido quedarnos con los niños que presenten un seguimiento longitudinal igual o superior a los 12 meses; seguimiento que consideramos apropiado para encontrar características evolutivas suficientes. A este parámetro se ajustan los mismos niños que el criterio anterior: 25 niños de Koiné y los 6 de Providence tienen un seguimiento de 12 meses o más.

(3) El tercer criterio de diseño es el de la cadencia, y está ligado a la permanencia evolutiva. No basta únicamente con que haya un seguimiento evolutivo mínimamente extenso, sino que también es preciso que a lo largo de esa extensión de tiempo el niño mantenga su participación en un número suficiente de grabaciones (de sesiones de grabación), que sean al menos 10 o 12. En palabras de Enríquez (2015: 182): "Consideramos que este criterio, conjuntamente con la decisión de que el seguimiento longitudinal superara los doce meses, nos asegura al menos una media de una sesión de grabación al mes, esto es, asegura cierta regularidad en el seguimiento". El parámetro de la cadencia sí recorta la muestra. De los 25 niños del corpus Koiné que cumplían el parámetro anterior, se adecúan 10. En el corpus Providence, otra vez, permanecen los seis participantes.

(4) Otro parámetro importante manejado es el de la lengua de los informantes. En el corpus Providence todos sus integrantes son hablantes monolingües de

inglés, mientras que en el corpus Koiné, más heterogéneo y amplio, se encuentran informantes que ocasionalmente emplean una segunda lengua: el gallego. No obstante, es el castellano la principal lengua empleada en este banco de datos. De los 10 niños de Koiné que cumplían el criterio de la cadencia, 9 pasan por el filtro de este criterio y solamente descartamos a uno, ya que emplea mayoritariamente el gallego en sus emisiones.

(5) Finalmente, el último criterio que se ha aplicado para diseñar la muestra es de tipo físico-social, pero con incidencia en el plano adquisitivo. Se trata del sexo. Por ello, se ha buscado un número similar de niños y de niñas en cada subcorpus. Una vez más, en el Providence no hubo que hacer ajuste alguno, puesto que está conformado por tres niños y tres niñas, pero no sucede lo mismo en Koiné. Los nueve informantes que nos devuelve el criterio anterior son 6 niños y 3 niñas, por lo que nos vimos en la obligación de prescindir de tres niños, concretamente de los que participaban en menos sesiones para obtener una muestra equilibrada por sexos.

En la tabla 1 presentamos de forma resumida la muestra de informantes que hemos diseñado, que describimos con más detalle en §3:

	Niño	Sexo	Lengua	Periodo de seguimiento	Muestras
KOINÉ	CEC	f	español	1;10 – 3;1	12
	DIA*	f	español	1;10 – 3;3	10
	ANP	f	español	2;1 – 3;9	19
	RIC	m	español	1;10 – 3;11	23
	JOR	m	español	1;11 – 3;7	22
	IAG	m	español	1;11 – 4;1	28
PROVIDENCE	Lily**	f	inglés	1;1 – 4	80
	Naima**	f	inglés	0;11 – 3;10	88
	Violet	f	inglés	1;2 – 3;11	51
	Alex	m	inglés	1;4 – 3;5	51
	Ethan	m	inglés	0;11 – 2;11	50
	William	m	inglés	1;4 – 3;4	44

Tabla 1: Subcorpus Koiné y subcorpus Providence

**No incluida en PhonBank.*
***Seguimiento denso en ciertos periodos.*

Como se observa en la tabla 1, nuestra muestra de datos con valor está compuesta por dos subcorpus, un repertorio con niños que adquieren el español y otro inventario con niños que desarrollan el inglés. Ambos subcorpus contienen seis in-

formantes, con un equilibrio en el parámetro del sexo, tres niños y tres niñas en cada uno. Los niños de la muestra para el español –subcorpus Koiné– proceden de diferentes escuelas: CEC y DIA de la escuela infantil Breogán , ANP y JOR de la escuela infantil de Vite y RIC y IAG de la escuela Elfos. En cuanto a las lenguas empleadas, todos presentan un uso mayoritario de castellano con emisiones ocasionales en gallego. El seguimiento de estos niños comienza, de media, a los 22 meses (1;10) y termina a los 3;7 años, y además su participación en las sesiones de grabación es constante, lo que nos garantiza que efectivamente hay regularidad en el seguimiento. Las grabaciones siguieron el patrón más habitual en la elaboración de corpus, sesiones de 15-20 minutos cada dos semanas.

En la muestra para el inglés –subcorpus Providence– la información de los participantes es más limitada que para los participantes de Koiné. Como adelantábamos en §2.2, todos los niños son hablantes monolingües de inglés procedentes del sur de Nueva Inglaterra (EEUU). El seguimiento de los informantes duró dos años y las sesiones de grabación tuvieron lugar cada dos semanas durante una hora. A dos de las tres niñas, sin embargo, se les hizo un seguimiento denso en los períodos 1;3-2;10 (Naima) y 2;0-3;0 (Lily) con grabaciones semanales. La edad media en la que comienza el seguimiento es ligeramente más temprana que la de Koiné (13 meses y medio) y lo mismo sucede con la edad de finalización del seguimiento (aproximadamente 3;6 años). Dado que las diferencias son mínimas y ambas horquillas de edad son, en general, bastante parecidas, no consideramos que vaya a suponer ningún problema a la hora de rastrear procesos fónicos.

4. DISCUSIÓN Y CONCLUSIONES

Siguiendo una perspectiva corpus-driven, se ha diseñado una muestra de datos significativa empleando cinco criterios psicolingüísticos de relevancia adecuados a los intereses de una investigación particular, a saber, la periodización comparada de procesos fónicos en castellano e inglés. Esta muestra piloto, por ocuparse de un sistema lingüístico peculiar y emergente, como es el habla de los niños, no busca ser representativa (cualidad de los corpus de habla adulta, ya conformada), sino ser significativa, es decir, tener garantías en cuanto a la calidad de los datos, a las características de los informantes y a los entornos de grabación. Garantías que se cumplen, precisamente, gracias a los mencionados criterios de diseño.

Repasaremos brevemente los cinco criterios empleados, señalando por qué son importantes y para qué nos han servido. Además, cada uno de ellos podrá extender, complementar o incrementar el número de informantes de los subcorpus si fuese

necesario. El primer parámetro empleado, fundamental en cualquier estudio de habla infantil, es la edad cronológica. Siguiendo el patrón de lo que es esperable en la emergencia del lenguaje, y dado que los informantes de los dos repertorios de datos presentan un desarrollo típico, con el criterio de la edad se seleccionan aquellos niños de entre 1;6 y 3;6 años de edad, momento en el que operan la mayor parte de procesos fónicos. El segundo y el tercer parámetro, muy relacionados entre sí, son el seguimiento y la cadencia evolutivos. Cuestiones ambas relacionadas ya no con los informantes en sí, sino con el modo de proceder de las sesiones de grabación. Para poder descubrir características lingüísticas evolutivas interesantes, se necesita un seguimiento más o menos largo (para el repertorio Koiné-Providence se estableció en al menos 12 meses) y una presencia regular en sesiones de grabación (10 o 12 sesiones como mínimo). De esta forma nos aseguramos de que cada informante aparece en aproximadamente una sesión mensual. El cuarto criterio divide la muestra en dos grupos, según sea la lengua que adquieren los informantes: un subcorpus con hablantes de español y otro con hablantes de inglés. Por último, el parámetro del sexo. Se tomó la decisión de contar con un número igual de niños que de niñas en cada subcorpus, puesto que el sexo es una variable que repercute directamente tanto en el *input* como en el output lingüístico que reciben o emiten los niños.

Si la muestra que se ha diseñado no es suficiente, o si no se le puede extraer todo el partido que se esperaba, también se puede contemplar una ampliación, siguiendo los mismos criterios establecidos para dar relevancia a los datos. La horquilla de edad podría ser extendida, sea por el límite superior, sea por el inferior; si bien sería más interesante el inferior, para poder disponer de datos de adquisición en etapas más tempranas. O por el contrario, la franja de edad podría restringirse, y la investigación se concentraría únicamente en una etapa concreta. El seguimiento, que se estableció en un año, puede extenderse a un periodo de tiempo más largo, para que la muestra proporcione mayor cantidad de datos evolutivos individuales. Le acompaña la cadencia, parámetro que también es susceptible de ser ampliado, y en especial en el subcorpus Koiné (ya que sus participantes tienen una presencia mucho menor

[10] Situada en una zona periférica de Santiago de Compostela, adscrita al personal de la Universidade de Santiago de Compostela.

[11] También en una zona periférica de Santiago de Compostela, en este caso en un barrio obrero. Escuela infantil de titularidad privada en la zona centro de A Estrada (Pontevedra).

[12] Las DDB (o Dense Databases) se utilizan principalmente en investigaciones que se interesan por fenómenos de aparición poco frecuente. La recogida de datos es más exhaustiva que en las bases de datos tradicionales, con cinco sesiones semanales de una hora. Los corpus densos que se generan a partir de esta recogida contienen alrededor del 10% de las producciones diarias del niño, frente al 1-1.5% de las bases de datos tradicionales (cf. Tomasello & Stahl 2004).

que los del subcorpus Providence), incluyendo a niños con presencia en un mayor número de sesiones. Por último, la manera de ampliar el corpus Koiné-Providence considerando el criterio de la lengua, sería posible añadir un tercer o un cuarto grupo de datos de niños que adquieren otro idioma. Siguiendo en la línea de las investigaciones clásicas de comparación interlingüística (cf. Slobin 1985a, 1985b, 1992, 1997a, 1997b), lo ideal sería contar con lenguas pertenecientes a distintas familias lingüísticas, con el objeto de alcanzar generalizaciones interlingüísticas.

Como se ha ido viendo, la muestra de datos Koiné-Providence, además de tener ciertas garantías sobre criterios precisos que la respaldan, no es extremadamente rígida ni está totalmente cerrada, sino que permite grados de ajuste si se encontrasen dificultades a la hora de trabajar con ella. No parecen agotadas todas las opciones ni tampoco tomadas todas las decisiones; son estos dos asuntos, la significatividad de los datos que contiene y su flexibilidad, los que hacen del corpus Koiné-Providence una muestra de datos con valor.

5. REFERENCIAS BIBLIOGRÁFICAS

BIBER, Douglas (1993): "Representativeness in Corpus Design", en *Literary and Linguistic Computing*, Vol. 8, No. 4, pp. 243-257. Oxford: Oxford University Press.

BIBER, Douglas, Susan CONRAD y Randi RIEPEN (1998): *Corpus Linguistics. Investigating Language Structure and Use. Cambridge*: Cambridge University Press.

BIBER, Douglas y James K. JONES (2003): "Quantitative methods in corpus linguistics", en LÜDELING, A. y M. KYTÖ (eds.): *Corpus linguistics: An international handbook*, Vol. 2, pp. 1286-1304. Berlin: Walter de Gruyter.

DEMUTH, Katherine, Jennifer CULBERTSON y Jennifer ALTER (2006): "Word-minimality, epenthesis, and coda licensing in the acquisition of English, en *Language & Speech,* 49, pp. 137-174.

ENRÍQUEZ, Iván (2015): *La Adquisición de Construcciones Complejas: de la Interacción a la Gramática*. [Tesis doctoral] Santiago de Compostela: Universidade de Santiago de Compostela.

FERNÁNDEZ, Milagros (2003): "Dinamismo construccional en el lenguaje infantil y teoría lingüística", en *Estudios de Lingüística Universidad de Alicante (ELUA),* 17 (vol. especial), pp. 273-287. Alicante: Universidad de Alicante.

– (2005): "El lenguaje infantil. Algunos lugares comunes revisitados", en *Interlingüística,* 16 (1), pp. 21-42.

– (2006): "Usos verbales y adquisición de la gramática. Construcciones y procesos en el habla infantil, *en Revista Española de Lingüística (RSEL),* 36, pp. 319-347.
– (2007): "La actualidad de los estudios sobre lenguaje infantil", *en Lynx: Panorámica de estudios lingüísticos,* 6, pp. 3-40.
– (coord.) (2011): *Lingüística de corpus y adquisición de la lengua*. Madrid: Arco/Libros.
– (2015): "Lenguaje infantil y medidas de desarrollo verbal", en *ENSAYOS, Revista de la Facultad de Educación de Albacete,* 30 (2), pp. 53-69.
– (2017): [en revisión]: "Corpus lingüísticos y "representatividad". El valor de los datos en fuentes de habla infantil".
GRUNWELL, Pamela (1981): "The development of phonology", en *First Language*, 2, pp. 161-191.
HANSON, Norwood Russell (1958*): Patterns of Discovery. An Inquiry into the Conceptual Foundations of Science*. Cambridge: Cambridge University Press.
HUNSTON, Susan (2002*): Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.
INGRAM, David (1976): *Phonological disability in children.* London: Edward Arnold.
– (1986): "Phonological development: production" en FLETCHER, P. y M. GARMAN (eds.): *Language Acquisition. Studies in first language development,* pp. 223-239. Cambridge: Cambridge University Press.
LAKOFF, George y Mark JOHNSON (1980): *Metaphors We Live By*. Chicago and London: The University of Chicago Press.
LEECH, Geoffrey (2007): "New resources, or just better old ones?" en M. HUNDT, N. NESSELHAUF y C. BIEWER (eds*.): Corpus Linguistics and the Web*, pp. 134–49. Amsterdam: Rodopi.
MACWHINNEY, Brian y Catherine SNOW (1985): "The Child Language Data Exchange System", *en Journal of Child Language*, 12, pp. 271-472.
MCENERY, Tony y Andrew HARDIE (2012): *Corpus Linguistics. Method, Theory and Practice.* Cambridge: Cambridge University Press.
PETERS, Ann M. (1980): "The units of language acquisition", *University of Hawai'i Working Papers in Linguistics* 12 (1), pp.1-72.
– (1983): *The Units of Language Acquisition, Monographs in Applied Psycholinguistics*, Cambridge University Press.
ROSE, Yvan (2012): "Multilingual Phonological Corpus Analysis: The Tools behind the PhonBank Project", en SCHMIDT, T. y K. WÖRNER (eds*.): Multilingual Corpora and Multilingual Corpus Analysis,* pp. 365–381. Amsterdam:

John Benjamins Publishing Company.
– (2014): "Corpus-based Investigations of Child Phonological Development: Formal and Practical Considerations", en DURAND, J., U. GUT y G. KRISTOFFERSEN (eds*.), The Oxford Handbook of Corpus Phonology*, pp. 265-285. Oxford: Oxford University Press.

ROSE, Yvan y Brian MACWHINNEY (2014): "The PhonBank Project: Data and Software- Assisted Methods for the Study of Phonology and Phonological Development", en DURAND, J., U. GUT y G. KRISTOFFERSEN (eds.), The Oxford Handbook of Corpus Phonology, pp. 380–401. Oxford: Oxford University Press.

ROSE, Yvan, Brian MACWHINNEY, Rod BYRNE, Gregory HEDLUND, Keith MADDOCKS, Philip O'BRIEN y Todd WAREHAM (2006): "Introducing Phon: A Software Solution for the Study of Phonological Acquisition", en BAMMAN, D., T. MAGNITSKAIA y C. ZALLER (eds.): *Proceedings of the 30th Annual Boston University Conference on Language Development,* pp. 489-500. Somerville, MA: Cascadilla Press.

ROSE, Yvan, Gregory HEDLUND, Todd WAREHAM, Rod BYRNE y Brian MACWHINNEY (2013): "Phon: A Computational Basis for Phonological Database Building and Model Testing", en VILLAVICENCIO, A. et al. (eds.): *Cognitive Aspects of Computational Language Acquisition*, pp. 29-49. Berlin/Heidelberg: Springer.

SARDINHA, Tony Berber (2004): *Lingüística de Corpus*. Barueri, SP: Manole.

SLOBIN, Dan Isaac (ed.) (1985a*): The Crosslinguistic Study of Language Acquisition, Vol. 1: The data.* Hillsdale, NJ: Lawrence Erlbaum Associates.
– (ed.) (1985b): *The Crosslinguistic Study of Language Acquisition, Vol. 2: Theoretical issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.
– (ed.) (1992): *The Crosslinguistic Study of Language Acquisition, Vol. 3*. Hillsdale, NJ: Lawrence Erlbaum Associates.
– (ed.) (1997a*): The Crosslinguistic Study of Language Acquisition, Vol. 4.* Hillsdale, NJ: Lawrence Erlbaum Associates.
– (ed.) (1997b*): The Crosslinguistic Study of Language Acquisition, Vol. 5: Expanding the contexts*. Hillsdale, NJ: Lawrence Erlbaum Associates.

STAMPE, David. (1969): "The Acquisition of Phonetic Representation", en BINNICK, R. et al. (eds*.): Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*, pp. 443-454. Chicago: Chicago Linguistic Society.

TOGNINI-BONELLI, Elena (2001): *Corpus Linguistics at Work.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

TOMASELLO, Michael (2003): *Constructing a Language: A Usage-Based Theory*

of Language Acquisition. Cambridge, MA: Harvard University Press.
TOMASELLO, Michael & Daniel STAHL (2004): "Sampling children's spontaneous speech: how much is enough?" en *Journal of Child Language*, 31 (1), pp 101–121.

Fecha de recepción:19 de enero de 2018.
Fecha de aceptación: 28 de febrero de 2018.