Universidad Autónoma de Madrid
Departamento de Lingüística General, Lenguas modernas, Lógica y Filosofía de la Ciencia
Laboratorio de Lingüística Informática

CHIEDE
Corpus de Habla Infantil Espontánea del Español

Marta Garrote Salazar

Tesis doctoral dirigida por el Dr. Antonio Moreno Sandoval

2008

Agradecimientos

El presente trabajo es fruto, en gran medida, del apoyo y ayuda que he recibido de todos aquellos que han estado a mi lado durante su realización.

En primer lugar, debo dar las gracias a mi director de tesis, el Dr. Antonio Moreno Sandoval, por confiar en mí desde que inicié mis estudios en Lingüística, por darme las oportunidades necesarias para llegar hasta aquí y, por supuesto, por guiarme durante el proceso de realización de la tesis con su experiencia y consejos.

Debo agradecer también la gran ayuda que he recibido del Dr. José María Guirao, que ha compartido conmigo sus conocimientos en el campo de la informática y ha estado siempre disponible para ayudarme en la resolución de cualquier tipo de inconveniente que haya podido surgir.

Los miembros del Laboratorio de Lingüística Informática de la UAM también son parte de este trabajo. Todos ellos me han ofrecido y prestado su ayuda, en especial Ana González-Ledesma, Raúl de la Torre y Ana Valverde, quienes han participado de forma activa en el proyecto.

Un agradecimiento especial a mis familiares y amigos, por su paciencia y apoyo durante estos últimos años, especialmente a mis padres y a mi tía Mª Ángeles. Gracias Guillermo.

Finalmente, este trabajo ha sido posible por mi participación en el proyecto europeo C-ORAL-ROM, primero, y después en el proyecto RILARIM, gracias a los cuales he obtenido la formación y financiación necesarias para poder finalizar esta tesis.

Índice general

Índice de tablas, figuras y gráficos ____ **6**

0. Introducción ____ **8**
0.1 Motivos y objetivos del trabajo ____ 8
0.2 Estructura de la tesis ____ 10

PARTE PRIMERA Bases teóricas ____ ***12***

1. La Lingüística de Corpus ____ **13**
1.1 Antecedentes ____ 13
1.2 La Lingüística de Corpus en la actualidad ____ 15
1.2.1 Principales corpus actuales ____ 17
1.3 Conceptos básicos ____ 18
1.3.1 Definición de corpus ____ 18
1.3.2 Criterios y dificultades de recopilación ____ 21
1.3.3 Tipología de corpus ____ 22
1.3.4 Ventajas ____ 24
1.4 Perspectivas de futuro ____ 24

2. La ontogénesis del lenguaje ____ **28**
2.1 Reseña histórica sobre el estudio del lenguaje infantil ____ 28
2.2 Principales teorías ____ 30
2.3 Evolución del lenguaje infantil ____ 34

3. Los corpus y la ontogénesis del lenguaje ____ **37**
3.1 Estado de la cuestión ____ 37
3.2 Los corpus infantiles en la actualidad ____ 39
3.3 CHILDES ____ 40
3.3.1 Presentación ____ 40
3.3.2 Contenidos de la página ____ 41
3.3.3 El corpus español ____ 45
3.4 Análisis del método CHILDES ____ 46

4. Experiencia en C-ORAL-ROM ____ **54**
4.1 El proyecto C-ORAL-ROM ____ 54
4.2 El LLI-UAM ____ 55
4.3 Adaptación de la metodología ____ 58

PARTE SEGUNDA La aplicación _____ 63

1. Diseño del corpus _____ 64
1.1 Aspectos generales _____ 64
1.2 Dificultades previas y elección de los participantes _____ 67
1.3 Diseño final de CHIEDE _____ 71

2. Metodología _____ 74
2.1 Grabaciones y entorno _____ 74
2.2 Digitalización _____ 76
2.3 Transcripción _____ 78
2.3.1 Cabeceras (Metadatos) _____ 80
2.3.2 Convenciones _____ 81
2.3.2.1 Turnos _____ 81
2.3.2.2 Comentarios _____ 81
2.3.2.3 Etiquetas para marcar la información prosódica _____ 83
2.3.2.4 Apoyos vocálicos _____ 84
2.3.2.5 Interjecciones _____ 85
2.3.2.6 Los signos paralingüísticos _____ 85
2.3.2.7 Sonidos no reconocidos como signos lingüísticos _____ 86
2.3.2.8 Los reinicios o reformulaciones _____ 86
2.3.2.9 Interrupciones _____ 87
2.3.2.10 Pausa _____ 87
2.3.2.11 Solapamiento _____ 87
2.4 Alineamiento _____ 88
2.5 Conversión a XML _____ 90

3. La anotación _____ 93
3.1 Anotación morfosintáctica del corpus _____ 93
3.1.1 Etiquetario _____ 94
3.1.1.1 Problemas para el establecimiento de un etiquetario _____ 95
3.1.1.2 Concepción teórica del etiquetario en C-ORAL-ROM _____ 96
3.1.2 GRAMPAL _____ 97
3.1.3 Ejemplo en CHIEDE _____ 100
3.2 Transcripción fonológica _____ 101
3.3 Revisión y desambiguación _____ 104

4. Resultados _____ 106
4.1 Evaluación del etiquetado morfosintáctico _____ 106
4.2 Evaluación del transcriptor fonológico _____ 110
4.3 Listado de frecuencias de las unidades _____ 111
4.3.1 Datos extraídos del etiquetador morfosintáctico _____ 113
4.3.2 Datos extraídos del transcriptor fonológico _____ 120
Gráfico 4.22 Incremento de la LME _____ 127
4.4 Comparación de CHIEDE con C-ORAL-ROM _____ 127

5. Explotación del recurso _____ 130
5.1 Estudios cuantitativos vs. cualitativos _____ 130

5.2 Un ejemplo de investigación cualitativa: los marcadores discursivos en el lenguaje infantil _____ 131

6. La aplicación en Internet _____ 135

7. Conclusiones y trabajo futuro _____ 137

Bibliografía _____ 139

Apéndice A. Ejemplo de CHIEDE _____ **146**

Apéndice B. Listados completos de frecuencias de formas y categorías léxicas _____ **196**

Apéndice C. Listados completos de frecuencias de sílabas _____ **288**

Índice de tablas, figuras y gráficos

Parte primera

Tabla 2.1 Evolución lingüística en el niño_____ 36
Tabla 3.1 El corpus español en CHILDES _____ 46
Tabla 3.2 Comparación CHILDES y CHIEDE _____ 47
Tabla 3.3 Fases del diseño en CHILDES y CHIEDE _____ 48
Tabla 4.1 Nuevas convenciones de transcripción en CHIEDE ___ 62

Parte segunda

Figura 1.1 Diseño de CHIEDE _____ 69
Tabla 1.2 Asambleas en CHIEDE _____ 73
Tabla 1.3 Entrevistas en CHIEDE _____ 73
Tabla 1.4 Información general de CHIEDE _____ 73
Figura 2.1 Ejemplo de autorización de grabación _____ 75
Figura 2.2 Proceso de traslado de la DAT al ordenador _____ 77
Figura 2.3 Edición del sonido con WaveLab _____ 77
Figura 2.4 Programa de transcripción Transana _____ 89
Figura 2.5 Alineamiento en Transana _____ 90
Tabla 3.1 Etiquetario _____ 97
Tabla 3.2 Entradas en GRAMPAL (Moreno y Guirao, 2006) ____ 98
Figura 3.3 Revisión del etiquetado morfosintáctico _____ 104
Tabla 4.1 Evaluación del etiquetado del corpus _____ 106
Tabla 4.2 Porcentaje de error en CHIEDE _____ 107
Tabla 4.3 Principales errores de categoría _____ 109
Tabla 4.4 Errores de categoría por ambigüedad _____ 110
Tabla 4.5 Lemas y formas por archivo _____ 113
Tabla 4.6 Lemas y formas por grupos de edad _____ 114
Tabla 4.7 Frecuencia de formas por grupos de edad _____ 115

Tabla 4.8 Categorías léxicas por grupos de edad _____ 116
Gráfico 4.9 Formas por grupos de edad _____ 117
Gráfico 4.10 Lemas por grupos de edad _____ 117
Figura 4.11 Porcentajes de frecuencia en grupos de 500 palabras _ 118
Gráfico 4.12 Evolución de errores por analogía _____ 119
Tabla 4.13 Errores por analogía _____ 120
Tabla 4.14 Frecuencia de fonemas por grupos de edad _____ 122
Gráfico 4.15 Vocales por grupos de edad _____ 123
Gráfico 4.16 Oclusivas por grupos de edad _____ 123
Gráfico 4.17 Nasales por grupos de edad _____ 124
Gráfico 4.18 Líquidas por grupos de edad _____ 124
Gráfico 4.19 Fricativas por grupos de edad _____ 125
Tabla 4.20 Frecuencia de sílabas por grupos de edad _____ 126
Tabla 4.21 Longitud Media de Enunciado _____ 127
Tabla 4.22 Incremento de la LME _____ 127
Tabla 4.23 Frecuencias CHIEDE vs. C-ORAL-ROM _____ 128
Figura 5.1 Marcadores relacionales (González-Ledesma y Garrote, 2006) _____ 132
Figura 5.2 Marcadores de subjetividad (González-Ledesma y Garrote, 2006) _____ 132
Figura 5.3 Frecuencia de marcadores discursivos en CHIEDE (González-Ledesma y Garrote, 2006) _____ 133

0. Introducción

0.1 Motivos y objetivos del trabajo

La investigación en el campo de la Lingüística Computacional es actualmente una tarea no sólo justificada, sino también necesaria. Los avances tecnológicos e informáticos que se producen casi a diario facilitan dicha tarea y le conceden una nueva perspectiva. Ciertos trabajos que hace unos años parecían imposibles de realizar son hoy factibles si se saben aprovechar adecuadamente los recursos disponibles. Así, las nuevas herramientas informáticas, junto con la capacidad de los ordenadores actuales, hacen más fácil, y sobre todo posible, el trabajo con grandes cantidades de datos. De esta forma, hoy en día es posible construir y manejar enormes corpus lingüísticos con cierta comodidad.

El desarrollo de corpus realizados a partir de muestras de lengua oral se ha convertido en una actividad tan importante como habitual entre diferentes profesionales dentro del campo de las tecnologías lingüísticas y de la comunicación. Sin embargo, el tratamiento de la lengua oral espontánea, a diferencia de lo que ocurre con textos escritos, resulta una tarea difícil, debido a la naturaleza compleja del comportamiento lingüístico humano. La actuación espontánea en situaciones reales incluye además factores extralingüísticos como el ruido, interacciones no verbales, etc. El desarrollo de herramientas de análisis o generación de lenguaje natural, pues, tiene como principal obstáculo este carácter dinámico de la lengua oral, que en muchas ocasiones difiere en gran medida de la norma escrita establecida. Por ello, es necesario trabajar con grandes cantidades de datos de los que poder extraer patrones de comportamiento lingüístico o, si esto no es posible, al menos contar con muestras suficientemente representativas.

Los corpus orales pueden ser de diferentes tipos según los intereses sociolingüísticos de los investigadores; de esta forma, podemos encontrar corpus generales que recogen muestras de la variante estándar de una lengua usada en diferentes ámbitos; o corpus dialectales con muestras de una variante lingüística diatópica; o también se pueden recoger ejemplos de las variantes diastráticas de una lengua.

En nuestro caso, se decidió construir un corpus de habla infantil por diferentes motivos. En primer lugar, es llamativa la escasez de este tipo de recurso lingüístico para el español. Actualmente, contamos con corpus de habla infantil llevados a cabo por diversos organismos y universidades españolas; no obstante, la mayoría se basan en estudios longitudinales —lo que supone contar con un número de participantes muy reducido— y su tamaño suele ser bastante limitado. Por otra parte, algunos de ellos centran su estudio en niños de edades superiores a lo que se considera el periodo de

adquisición[1], trabajando con niños en edad escolar o incluso adolescentes. Nuestra intención por lo tanto era reunir muestras de lenguaje infantil de niños con edades entre los tres y los seis años en un corpus con un tamaño que se pudiese considerar representativo.

En la actualidad, el corpus de habla infantil que se presenta como referencia internacional es CHILDES, un corpus multilingüe que incluye muestras de español. Esta es otra de las razones por las que consideramos necesario el desarrollo de un corpus de esta variante lingüística: los resultados obtenidos tras una comparación de la metodología empleada en el proyecto CHILDES y la desarrollada en el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid (LLI-UAM) para el trabajo con corpus orales (Pérez Milans y Moreno, 2004) nos demuestran que CHILDES es una herramienta mejorable.

Para ello, partimos de una metodología asentada y reconocida internacionalmente, la del proyecto C-ORAL-ROM[2], realizado, en lo que a su apartado de lengua española se refiere, en el LLI-UAM. Además, se cuenta con herramientas de anotación fonológica y morfosintáctica probadas y eficientes. Por tanto, la aportación más significativa de este trabajo será la aplicación de la experiencia en corpus orales y Lingüística Computacional del LLI-UAM a la construcción de un corpus de habla infantil. La experiencia previa del equipo investigador en creación de corpus orales y herramientas informáticas para el análisis de textos le permite estar a la vanguardia de este tipo de investigaciones.

Los objetivos de nuestro trabajo han sido los siguientes:

- Recogida de muestras de habla infantil para la creación de un corpus con un tamaño de unas 60.000 palabras. Unido a los otros corpus del LLI-UAM se obtendría un corpus unificado de más de un millón de palabras con una muestra muy representativa de la variedad infantil.
- Transcripción ortográfica de las muestras de audio seleccionadas.
- Alineamiento de la transcripción y el sonido de todo el corpus.
- Anotación automática morfosintáctica de la transcripción con posterior revisión manual.
- Anotación automática fonológica y revisión de la misma.
- Extracción de los datos de los textos anotados mediante métodos automáticos y estadísticos.
- Análisis de los datos referentes al léxico, morfosintaxis y fonología.
- Creación de una página web mediante la cual se pueda acceder al corpus y a los datos extraídos del mismo de forma rápida y sencilla.

[1] Más adelante veremos que la mayoría de los autores sitúan el periodo crítico de aprendizaje lingüístico entre los tres y los seis años, considerando que después de esta edad los niños ya han adquirido un lenguaje similar al de los adultos.

[2] http://lablita.dit.unifi.it/**coralrom**

El resultado será una variada base de datos con una rica información en un área en la que escasean los recursos básicos. Los nuevos conocimientos adquiridos tendrán una clara aplicación para la comunidad científica.

0.2 Estructura de la tesis

El presente trabajo se divide en dos partes: la primera de ellas, dedicada a las bases teóricas sobre las que se fundamenta; y la segunda, en la que se presenta todo el trabajo práctico realizado.

La primera parte consta de cuatro capítulos. En el primero de ellos, se hace un breve repaso a la historia de la Lingüística de Corpus, desde sus orígenes a su estado en la actualidad. El siguiente, trata sobre la ontogénesis del lenguaje y las principales teorías sobre adquisición lingüística. Se describen las metodologías empleadas a lo largo de la historia para estudiar el lenguaje infantil y su estrecha relación con la Lingüística de Corpus. Finalmente, se exponen las etapas o periodos básicos de adquisición, desde que el niño empieza a dar las primeras muestras de comunicación hasta el momento en el que se considera que el proceso de adquisición está completado.

En el capítulo tercero, se ponen en relación la Lingüística de Corpus y la ontogénesis del lenguaje mediante una descripción del estado de la cuestión y un análisis de los principales corpus infantiles que existen en la actualidad, prestando especial interés a CHILDES, el corpus de habla infantil que actualmente se presenta como referencia internacional.

Por último, se establecen las bases metodológicas que guiarán nuestro trabajo, heredadas éstas de la tradición existente en el LLI-UAM y su experiencia en creación de corpus, en especial, en la creación del proyecto multilingüe C-ORAL-ROM, cuyas características han servido de ejemplo para el desarrollo de nuestro corpus.

La segunda parte del trabajo está dedicada a la aplicación, es decir, el corpus, y todo el trabajo práctico que ha supuesto su realización, desde el diseño del mismo hasta su concretización mediante una página web. Esta segunda parte consta de siete capítulos, de los cuales el primero presenta las cuestiones previas a la realización de un corpus, es decir, su diseño y las dificultades que esta primera etapa puede entrañar.

En el segundo capítulo se presenta la metodología, detallando todos los pasos que se han dado para grabar las interacciones, digitalizarlas, transcribirlas, alinear sonido y texto y finalmente generar ficheros en formato XML a partir de los de texto. Esta es la fase que más tiempo consume en la creación de cualquier corpus de lengua oral.

El capítulo número tres detalla el proceso de enriquecimiento de los textos del corpus mediante la anotación. Se describen los dos tipos de anotación que se han utilizado en CHIEDE, es decir, la morfosintáctica y la fonológica, y cómo estos procesos se llevan a cabo de forma automática a pesar de necesitar una posterior revisión manual.

Dentro del cuarto capítulo, se exponen los resultados obtenidos después de la evaluación del funcionamiento del *tagger* o etiquetador morfosintáctico automático. Además, se analizan las listas de frecuencias, extraídas mediante métodos estadísticos, y se comparan con los datos obtenidos del corpus español que se incluye en C-ORAL-ROM.

En contraposición al apartado anterior, en el que se realiza un estudio cuantitativo, en el quinto capítulo se presenta un ejemplo de estudio cualitativo realizado a partir de la cuantificación de datos en CHIEDE.

La aplicación en Internet se describe en el sexto capítulo, mostrando el diseño de la misma y sus funciones. El usuario podrá así consultar el corpus y los datos extraídos del mismo a través de la red.

Finalmente, presentamos las conclusiones y propuestas de investigación futura para continuar con el trabajo.

Todos los ejemplos de transcripción usados a lo largo de este trabajo proceden del corpus CHIEDE —exceptuando aquellos tomados de CHILDES para ilustrar su funcionamiento— y a continuación de cada uno de ellos se especifica, entre corchetes, el nombre del archivo de transcripción al que pertenece. En cuanto a las citas bibliográficas, para aquellas obtenidas de páginas de Internet se facilitará la URL o localizador de la página, no pudiendo especificar número de página o fecha debido a su carácter dinámico y cambiante.

PARTE PRIMERA

Bases teóricas

1. La Lingüística de Corpus

1.1 Antecedentes

La Lingüística de Corpus (en adelante LC) es una disciplina ya consolidada dentro de los estudios lingüísticos. En la actualidad, la LC se describe como "the study of language on the basis of text corpora" (Aijmer & Altenberg, 1991:1), entendiendo corpus como una gran colección de textos disponible en formato electrónico. Sin embargo, es necesario hacer una diferenciación entre la LC antes y después de la aparición del ordenador. Con anterioridad a la LC actual, muchos investigadores basaban sus teorías en la observación directa de la realidad y en la anotación de ejemplos de lenguaje natural. Los estudios sobre adquisición lingüística son un buen ejemplo de ello, en los que la metodología empleada para el análisis del lenguaje infantil era la recolección de muestras de habla en los llamados "diarios de bebés".

Nelson Francis (1992) utiliza el término "corpora B.C.", es decir "before the use of computers", para referirse a la LC anterior a la existencia del ordenador, y hace una descripción de diferentes corpus que se han recopilado de forma manual a lo largo de la historia. Francis habla de tres tipos de corpus B.C.:

- Corpus lexicográfico, recopilado en el proceso de realización de diccionarios.
- Corpus dialectológico, elaborado con el fin de producir atlas dialectales.
- Corpus gramaticales, mediante los cuales se establecían las reglas de una gramática.

Fillmore (1991) también diferencia dos tipos de actividades de la LC a finales de la década de los 50 del siglo pasado: la recolección de corpus textuales en lenguas poco documentadas con fines descriptivos y etnográficos; y el estudio de las propiedades estadísticas de las lenguas, para lo cual había entonces poca cantidad de datos. Con anterioridad a 1950, los estudios basados en corpus se restringían al campo de la psicolingüística (en concreto la adquisición de la primera lengua), la fonología, la pedagogía o la lingüística comparativa. Fue en ese momento cuando los estructuralistas americanos (con Zelig Harris como máximo representante) e ingleses (con J. Firth) comprendieron que los datos lingüísticos eran la fuente esencial de información para construir teorías acerca de las lenguas.

A pesar de ello, durante las siguientes dos décadas y media, los estudios basados en datos se vieron interrumpidos o, al menos, eclipsados por influencia del racionalismo promulgado por Chomsky, quien señaló el principal problema de los trabajos basados en corpus: ningún corpus es

capaz de recoger todos los ejemplos posibles de una lengua. Por el contrario, la introspección o búsqueda de datos en la competencia del lingüista nativo es una fuente disponible que proporciona la información necesaria acerca de la gramaticalidad de una construcción determinada.

Con el racionalismo de Chomsky y su método introspectivo, la LC sufrió un gran abandono. No obstante, el método chomskyano no tardó en recibir críticas. La principal basaba sus argumentos en el hecho de que el uso natural espontáneo de la lengua tiene mayor validez que los datos creados artificialmente. Čermak censura "[…] the ad hoc character of examples used for analysis, which were often made up by linguists themselves without any attempt at recourse to representative sources and any reference to context" (Čermak, 2002: 267). Además, el mismo autor recalca la posibilidad de que, en algunas, lenguas ciertas construcciones propias de la variante oral nunca se reproduzcan por escrito, es decir, que sólo sean accesibles mediante la observación.

De la misma manera, el método introspectivo no es válido para el psicolingüista que estudia la adquisición del lenguaje:

> Introspective judgements are only available to us when our meta-linguistic awareness has developed, and there is no evidence that a child at the one-word stage has meta-linguistic awareness. Even Chomsky (1964) cautioned the rejection of performance data as a source of evidence for language acquisition studies. (McEnery & Wilson, http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus1/1fra1.htm).

O para el lingüista interesado en el cálculo de frecuencias lingüísticas, ya que mediante la introspección no es posible hallar el porcentaje de frecuencia de uso de una estructura o una palabra determinadas.

Alrededor de 1960 se sitúan dos grandes eventos en la historia de la LC: la realización del corpus *Survey of English Usage (SEU)* por Randolph Quirk, que recogía muestras del inglés contemporáneo oral y escrito; y la aparición del ordenador como herramienta que permitía almacenar y tratar grandes cantidades de datos. Esta nueva herramienta supuso el resurgimiento de la LC, permitiendo que en 1961 Nelson Francis y Henry Kučera realizaran el Brown Corpus como muestra del inglés contemporáneo, almacenado por primera vez de forma digital para ser usado mediante un ordenador. El Brown Corpus sirvió como modelo a seguir por numerosos países en todo el mundo, y los estudios basados en corpus aumentaron de forma notable en un periodo de treinta años. De la misma manera, fue también revolucionaria la realización de los diccionarios ingleses COBUILD o LDOCE, en los que se utilizaron datos extraídos de un corpus para crear las entradas.

A la par que el trabajo con corpus se iba consolidando, se ha ido desarrollando una metodología sobre cómo compilarlos. Más adelante, presentaremos una síntesis de las propuestas más relevantes de manos de autores como Sinclair, Leech o McEnery.

1.2 La Lingüística de Corpus en la actualidad

El aumento de estudios lingüísticos basados en corpus ha sido espectacular a lo largo de las últimas dos décadas; no obstante, aún existen discrepancias entre aquellos que consideran a la LC como una rama de la lingüística, equiparable a la sociolingüística o a la psicolingüística, con sus propias bases teóricas, objeto de estudio y método; y aquellos que conciben la LC como una herramienta o base metodológica dentro de la lingüística, pero no equiparable a una ciencia teórica. Como herramienta era considerada en 1991 en el simposio celebrado en Estocolmo, *Directions in Corpus Linguistics*, bajo la dirección de Jan Svartvik. Los participantes, entre los que se encontraban figuras tan importantes dentro de la disciplina como Nelson Francis, J. Fillmore, M.A.K. Halliday o G. Leech, consideraban a la LC como un método para la investigación lingüística:

> The only other branch of linguistics which, like corpus linguistics, refers to a tool or methodology rather than a subject-matter is computational linguistics, defined as the investigation of language by means of computers. (Leech, 1991:106).

Sin embargo, en aquel momento la LC era una disciplina extremadamente joven, cuyas bases aún debían asentarse. Desde entonces han pasado casi dos décadas. Durante este tiempo, el desarrollo tecnológico ha sido impresionante y este hecho ha permitido conseguir resultados y alcanzar metas dentro de la LC que hace veinte años eran impensables. Además, el hecho de que la LC sea claramente interdisciplinaria, con múltiples campos de aplicación, hace que aún hoy parezca que sus límites están borrosos. Algunos lingüistas como Pérez Hernández (2002) entienden que la LC debe considerarse una ciencia si tiene como objetivo el estudio lingüístico; por el contrario, la realización de un corpus como mera herramienta sin fines teóricos no se considera ciencia, sino método. No obstante, para diseñar, recopilar y crear un corpus es necesario tener una base lingüística teórica, al mismo tiempo que seguir unos principios teóricos hoy en día ya establecidos. Por el contrario, A. Briz, en los preliminares de la revista *Oralia* (Vol. 8, 2005) afirma que un corpus que proporcione simplemente una descripción de datos puede suponer el origen de una futura teoría. Según este autor, la compilación de un corpus puede llevarse a cabo sin una hipótesis de partida.

De cualquier manera, la LC en la actualidad está plenamente integrada dentro de la lingüística, ya sea como rama de la misma o como método.

Algunas de las especialidades lingüísticas donde se ha consolidado el empleo de corpus son la lexicografía, la terminología, la traducción o los estudios sociolingüísticos y multiculturales.

Concretamente en el área que nos afecta de forma más directa, la Ingeniería Lingüística, los corpus representan el recurso lingüístico más empleado (junto con los lexicones y las ontologías). El éxito de la aplicación de métodos estadísticos a datos de grabaciones de habla para entrenar sistemas de reconocimiento de voz fue uno de los causantes del cambio de paradigma en la LC. A partir de ahí, los métodos estadísticos de inferencia de conocimiento se han aplicado a cualquier nivel lingüístico, desde la señal sonora hasta el análisis pragmático y discursivo.

Es necesario hacer aquí una distinción entre Lingüística Computacional y Lingüística de Corpus. Ambas especialidades han estado siempre unidas, sobre todo en las últimas décadas, ya que tienen un denominador común: el uso del ordenador como herramienta básica. Sin embargo, sus orígenes, motivaciones y objetivos son distintos: La LC tiene como finalidad la creación de grandes archivos de textos y el tratamiento de los mismos mediante herramientas informáticas. Sus principales campos de aplicación son la lexicografía, la terminología y la enseñanza, y sus herramientas son la estadística y el ordenador para el procesamiento del lenguaje natural. Por su parte, la Lingüística Computacional surge de la Traducción Automática como primer objetivo. Su principal herramienta para el estudio del lenguaje es también el ordenador, pero su objetivo es la mejora de la comunicación hombre-máquina y la creación de aplicaciones informáticas que empleen el lenguaje natural.

En la actualidad la colaboración mutua de la lingüística y la ingeniería (nuevas tecnologías) es un hecho. La necesidad de establecer teorías sobre el funcionamiento del lenguaje natural ha dejado atrás a la lingüística teórica tradicional y la investigación ahora se centra en el lenguaje oral espontáneo como elemento multimodal (contexto, lenguaje corporal, aspectos suprasegmentales, elementos pragmáticos, etc.). Algunos campos de estudio dentro del Procesamiento del Lenguaje Natural son Síntesis de Voz, Gestión de Diálogo o Reconocimiento de Voz. Rubio Ayuso y Hernáez Rioja (2005) hablan de cuatro grandes bloques de investigación dentro del Procesamiento del Lenguaje Natural:

- Tratamiento de la palabra: análisis léxico-morfológico (*PoS tagging*), anotación semántica (tratamiento de la ambigüedad) y modelos estadísticos del lenguaje.
- Tratamiento sintáctico: gramáticas, análisis sintáctico robusto, análisis sintáctico superficial (*chunkers*).
- Tratamiento semántico: análisis semánticos robustos, categorización semántica y clasificación de textos y ontologías.

- Tratamiento pragmático: es el campo más amplio y en él se incluyen la traducción automática, la búsqueda y recuperación de información textual, extracción de información, sistemas de pregunta-respuesta, elaboración automática de resúmenes, etc.

Para todo ello, no cabe duda de que los corpus son una herramienta de gran utilidad, una fuente de datos necesarios para cualquiera de las actividades mencionadas.

1.2.1 Principales corpus actuales

La mayor parte del trabajo realizado en LC se ha centrado en la lengua inglesa. El primer corpus moderno (esto es, en formato digital) fue el Brown Corpus en lengua inglesa, y a partir de ese momento los recursos lingüísticos proliferaron en dicho idioma. No obstante, muchos países europeos y asiáticos siguieron el ejemplo y crearon sus corpus de referencia para sus propias lenguas. A continuación presentamos un listado de los corpus pioneros y más importantes en lengua inglesa.

- *The Brown Corpus*. Realizado en 1961 en la Brown University, con un millón de palabras en inglés americano.
- *The Birmingham Collection of English Texts*. Recopilado por un grupo de investigación bajo la supervisión de J. Sinclair, contiene alrededor de veinte millones de palabras de textos en inglés británico contemporáneo.
- *The Helsinki Corpus of English Texts: Diachronic and Dialectal*. Recopilado en la Universidad de Helsinki bajo la dirección de M. Rissanen. El corpus diacrónico contiene 1,6 millones de palabras de textos británicos pertenecientes al periodo comprendido entre 850 y 1720. El material dialectal reúne unas cuatrocientas mil palabras en diversos dialectos británicos contemporáneos.
- *The Lancaster/IBM Spoken English Corpus (SEC)*. Contiene 52.000 palabras del inglés británico oral.
- *The Lancaster-Oslo/Bergen Corpus (LOB)*. Recopilado por grupos de investigación en Lancaster, Oslo y Bergen en 1961, contiene un millón de palabras de textos en inglés británico.
- *The London-Lund Corpus of Spoken English (LLC)*. Está formado por la parte oral del Survey of Spoken English, dirigido por J. Svartvik y contiene unas 500.000 palabras en inglés británico oral.
- *The Longman/Lancaster English Language Corpus*. Bajo la supervisión de R. Quirk y G. Leech, está formado por 30 millones de palabras en inglés británico y americano.
- *The British National Corpus (BNC)*. Colección de 100 millones de palabras con muestras de lengua oral y escrita de una amplia variedad de fuentes.

En España los primeros trabajos en LC empezaron a llevarse a cabo en la década de los noventa. Así, podemos hablar de CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea) como proyecto pionero. Fue el primer corpus de habla espontánea del español, bajo la dirección de F. Marcos Marín, recogido entre 1991 y 1992 y financiado por IBM.

Entre los proyectos de recopilación de corpus generales cabe mencionar los patrocinados por la Real Academia Española: CREA (Corpus de Referencia del Español Actual) y CORDE (Corpus Diacrónico del Español); este último contiene textos de tres épocas fundamentales: la Edad Media, el Siglo de Oro y la época Contemporánea.

En cuanto a corpus orales, destaca C-ORAL-ROM, recopilado, entre otros equipos europeos, por el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid. Se trata de un corpus oral multilingüe y general del habla de cuatro lenguas romances (francés, español, italiano y portugués) en sus diferentes registros. En el capítulo 5, explicamos con mayor detenimiento el proyecto C-ORAL-ROM, cuya metodología ha servido de modelo para el presente proyecto. Además, en el apartado dedicado a los corpus y la ontogénesis, se hará un breve análisis de los diferentes corpus de habla infantil que existen en la actualidad.

Otros trabajos que se han llevado a cabo en este terreno y que se recogen en la Oficina de Español en la Sociedad de la Información (OESI), del Instituto Cervantes, son:

- ANGLE, Archivo Gramatical de la Lengua Española, Centro Virtual Cervantes.
- BDS, Base de Datos Sintácticos del Español Actual, Departamento de Lengua Española, Universidad de Santiago de Compostela.
- LEXESP, Laboratori de Recerca en Lingüística Computacional, Secció de Lingüística Computacional, Departament de Filologia Romànica, Universitat de Barcelona.
- Corpus textual plurilingüe especializado, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
- Corpus Oral y Sonoro del Español Rural (COSER).

1.3 Conceptos básicos

Bajo este epígrafe, vamos a definir el concepto actual de corpus y a resumir algunos de los conceptos básicos dentro de la LC.

1.3.1 Definición de corpus

Los corpus son grandes muestras de texto con ejemplos reales. Según McEnery y Wilson (1996), un corpus es cualquier colección de textos. Sin

embargo, como término utilizado en el contexto de la lingüística moderna, tiene unas connotaciones más específicas. La definición de Sinclair (1991) destaca el carácter natural —como opuesto a artificialmente creado— de un corpus y la finalidad de su recolección: "A corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language" (Sinclair, 1991:171). A esto, podemos agregar las características que destacan McEnery y Wilson (1996) como propias de un corpus moderno:

- Muestreo y representatividad
- Tamaño finito
- En formato electrónico
- Referencia estándar

Los rasgos básicos de un corpus moderno se pueden resumir en lo siguiente:

- Estar codificado **en formato electrónico**: esta característica es esencial en cualquier corpus actual ya que permite su fácil almacenamiento (en la recogida) y manipulación (en la anotación y búsqueda de información). La tercera virtud del formato electrónico es que permite su transmisión y reutilización de manera que puede ser modificado o transformado en otro tipo de corpus.
- Ser **representativo** de una lengua o variedad: el concepto de representatividad de la muestra recogida se ha convertido en uno de los temas más discutidos. El punto de partida es que todo corpus es finito. Por tanto, no se puede aspirar a contener todos los datos de una variedad (salvo en casos especiales como el corpus del latín clásico escrito, las obras completas de un autor, etc.). La finitud de un corpus obliga a que el diseño del mismo se realice pensando más en la proporción y en la diversidad que en su tamaño. Efectivamente, imaginemos que queremos construir un corpus del habla infantil. Podríamos empezar a recoger una gran cantidad de datos siguiendo la creencia de que cuantos más mejor. Sin embargo, para que los estudios derivados tengan validez más allá de la observación de la muestra, deben estar proporcionados en función de parámetros como edad, sexo, situación comunicativa, etc.
- Estar **disponible para el público**: en muchos casos es difícil difundir públicamente un corpus debido a posibles problemas legales con respecto a los derechos de propiedad intelectual y de privacidad. Efectivamente, los participantes son los propietarios legítimos de los derechos sobre los textos empleados, y son ellos quienes deben autorizar su uso. De igual manera, las conversaciones privadas no pueden ser recogidas en un corpus sin el consentimiento de sus participantes.
- Tener una **estructura interna** que clasifique los datos: es especialmente útil que los corpus se organicen en clases y subclases, siguiendo la proporción definida en su diseño. Esto permite hacer

comparaciones internas y, en general, proporcionar una mejor explotación de los datos. Por ejemplo, en un corpus donde estén clasificadas las muestras por las edades de los participantes se puede hacer un estudio cronológico de los datos.

Otras características interesantes para un corpus, a juicio de Sinclair (1991) son:

- **Cantidad**: si nuestro corpus tiene un diseño proporcionado y representativo, será mejor cuantos más datos tenga. La cantidad es un factor muy importante para la utilización de un corpus como recurso de ingeniería lingüística. Sin embargo, en ocasiones es muy difícil de conseguir: los corpus orales son intrínsecamente mucho más pequeños que los corpus escritos, ya que la recogida de los datos y su transcripción implica un mayor esfuerzo y más tiempo que compilar un corpus de textos.
- **Calidad**: por este concepto podemos entender varias cosas. En primer lugar, que la muestra recogida sea un ejemplo fidedigno de la variedad. Por ejemplo, si se trata de un corpus de habla espontánea, que la grabación haya sido realizada sin planificación previa. Más importante aún es que no se censuren o corrijan partes del original. La transcripción y la anotación deben ser verificadas de forma independiente. Si se trata de un texto escrito, es necesario verificar los errores tipográficos o de edición. En general, la calidad es más importante cuanto más específico es el corpus.
- **Simplicidad**: este criterio tiene que ver con la forma de presentar los datos. Como veremos, la anotación es el valor añadido de un corpus, pero sin olvidar que los datos primarios son la base. Por tanto, en todo corpus deben estar claramente diferenciados el texto base y la anotación añadida por el lingüista. En este aspecto se han elaborado diferentes recomendaciones y guías para poder separar las marcas del texto original.
- **Documentación**: este requisito incide en la necesidad de especificar todos los aspectos de la compilación del corpus, desde su diseño y recogida hasta las normas seguidas en su anotación.

A modo de resumen, y siguiendo a F. Marcos Marín (1994), las características de los archivos digitales que componen un corpus son: el almacenamiento en soporte electrónico, la posibilidad de recuperación por el usuario, el establecimiento de reglas para poder acceder a la información, la codificación estándar, la clasificación tipológica de los textos y el almacenamiento y mantenimiento a cargo de una institución responsable que facilite su consulta y uso.

1.3.2 Criterios y dificultades de recopilación

Antes de nada, tenemos que aclarar que un corpus no es cualquier colección de datos, sino que se trata de una selección basada en criterios que corresponden a unos objetivos claros:

> […] corpus data may not be and, many people would say, should not be just texts taken from, for example, newspapers in any straightforward way, where to gather a mass of newspapers is an easy matter". (Čemak, 2002:270).

Existen unos criterios previos a la recogida de los datos que están sometidos a cambios dependiendo de los objetivos del estudio y de la tipología del corpus. Si se trata de un corpus general, entonces la selección de datos debería de representar los diferentes registros del lenguaje. En este aspecto, existen algunos estudios sobre las proporciones recomendables que ha de tener un corpus general. Por otro lado, si se trata de un corpus especializado, como en el caso de un corpus de habla infantil, entonces la recopilación será de textos especializados que representen el área en cuestión. En resumen, en cualquier tipo de corpus es imprescindible dejar claros los objetivos de la investigación y limitar el área de alcance del corpus. En nuestro trabajo, dicha tarea se explicará en el capítulo dedicado al diseño del corpus.

En cuanto a las dificultades de recopilación se pueden dividir en dos tipos: las relacionadas con la lengua y las dificultades técnicas.

En primer lugar, a pesar del incremento de proyectos de recopilación de corpus, éstos sólo cubren un número limitado de lenguas, especialmente el inglés y algunas lenguas europeas. Por tanto, los trabajos desarrollados sólo se pueden probar en un pequeño grupo de lenguas: aquellas que disponen de corpus.

En segundo lugar, si consideramos Internet como un recurso fundamental de textos para la construcción de corpus nos enfrentamos con el mismo problema, ya que la mayoría de la información contenida en la red está en unas pocas lenguas, como el inglés, el chino o el alemán, y la presencia de otras lenguas es mucho menor. Esta situación hace más difícil la tarea de recopilación de corpus de lenguas con poca presencia en Internet.

Las dificultades técnicas están relacionadas con las distintas lenguas y sus sistemas de escritura, de los que derivan ciertos problemas. Los sistemas de codificación están pensados desde y para el inglés, y no para aquellas lenguas que contienen caracteres que no están incluidos en el sistema de escritura inglés, o lenguas con un alfabeto distinto al latino. Es necesario conocer los distintos tipos de codificación y saber cuál es el indicado para cada lengua.

1.3.3 Tipología de corpus

Sinclair (1991) describe dos tipos básicos de corpus: el *sample corpus*, una colección de textos finita en la que las muestras se seleccionan de forma exhaustiva y se analizan con detenimiento; y el *monitor corpus*, que reutiliza textos digitalizados para otros propósitos, como periódicos o libros.

La abundante experiencia acumulada a lo largo de los años permite realizar una tipología bien fundamentada. Actualmente, contamos con distintas clasificaciones, aunque la mayoría de los autores consultados coinciden en diferenciar tipos básicos de corpus. Marcos Marín (1994) propone la siguiente tipología:

- Textos completos / muestras
- Sincrónico / diacrónico
- General / terminológico
- Monolingüe / bilingüe / plurilingüe
- Simple / parcelas dobles / parcelas triples, etc.
- Central / exterior

Además, el mismo autor establece la tipología textual de un corpus dependiendo de las características específicas de los textos que lo componen: lengua escrita u oral, autor del texto, género, situación, temas o tópicos, etc.

Si echamos un vistazo a la tipología establecida por la Text Encoding Initiative (TEI)[3], vemos que apenas existen diferencias con la anterior tipología expuesta:

- Corpus de referencia
- Corpus monitor
- Corpus oral
- Corpus de ejemplos
- Corpus especiales, especializados y diseñados con fines específicos
- Corpus bilingües o multilingües (paralelos / comparables).

A modo de resumen, podemos establecer cinco tipos de corpus:

- **Corpus escritos y corpus orales**: los corpus escritos son colecciones de textos que originariamente tuvieron una fuente escrita. Los orales están formados por textos transcritos de grabaciones. Normalmente, las transcripciones suelen ser ortográficas. Una característica importante de los corpus orales es que deben ir

[3] En 1987 surge la Text Encoding Initiative (TEI) con la intención de crear un formato de codificación estándar que facilite el intercambio de textos en formato electrónico (http://www.tei-c.org).

acompañados de la grabación original. De sus características hablaremos en el capítulo dedicado al diseño del corpus.

- **Corpus monolingües y multilingües**: la mayoría de los corpus actuales recogen muestras de una única lengua. Sin embargo, por necesidades de la ingeniería lingüística, se están desarrollando corpus en más de una lengua. Los corpus multilingües pueden ser comparables o paralelos. Los corpus comparables contienen textos en distintas lenguas, pero no son traducciones unos de otros, sino que comparten temática, origen, extensión, etc. Se emplean para realizar estudios contrastivos. Los corpus paralelos están formados por versiones diferentes (original y traducciones) del mismo texto en diferentes lenguas. Suelen estar alineados por párrafos, por oraciones o incluso por palabras, de manera que se hacen explícitas las relaciones de equivalencia entre lenguas. Estos corpus son utilizados especialmente en traducción y en terminología.
- **Corpus generales y corpus de especialidad**: los primeros recogen muestras representativas de una lengua y tienen diferentes clasificaciones internas. Los segundos se han compilado pensando en el estudio de uno o varios aspectos. Es el caso de los corpus terminológicos, que se concentran en un dominio concreto. Es importante señalar que los corpus pueden reutilizarse con una finalidad diferente para la que fueron diseñados. Así, de un corpus general se pueden extraer fragmentos para crear uno de especialidad y, de igual forma, se pueden añadir corpus de especialidad para aumentar uno general.
- **Corpus anotados**: la versión más simplificada de un corpus es la que proporciona el texto plano, sin apenas información adicional. Si marcamos información en el texto (por ejemplo, con etiquetas morfosintácticas), entonces estamos añadiendo valor a ese corpus. De esta manera, podremos buscar información a partir de los lemas o de la categoría sintáctica, algo que es imposible de hacer con un texto plano. Para la anotación se emplean distintos programas informáticos. En la actualidad, la mayoría de los corpus se ofrecen con anotaciones lingüísticas, incluso en varios niveles.
- **Corpus multimodales**: son corpus que incluyen varios formatos multimedia, es decir, que mezclan texto, sonido y a veces imagen. De esta manera, se puede ver u oír una grabación y leer al mismo tiempo la trascripción. O se puede buscar una palabra en la trascripción y visualizar o escucharla en su realización concreta. Estos corpus son recientes, ya que exigen una tecnología de anotación simultánea de imagen, sonido y texto no disponible hasta hace poco. Su aplicación para la enseñanza de lenguas, análisis de la conversación o desarrollo de sistemas informáticos interactivos hace que los corpus multimodales se vean como una de las aplicaciones con más porvenir. Sin embargo, su complejidad en cuanto a diseño, tratamiento y anotación también es considerablemente mayor.

1.3.4 Ventajas

Según las definiciones y los rasgos comentados, podemos resumir las ventajas del uso de un corpus en lo siguiente:

- La principal ventaja de los corpus modernos es su carácter digital y su tratamiento y análisis mediante el uso del ordenador. El hardware y software disponibles en la actualidad facilitan enormemente el trabajo y ahorran mucho tiempo. Los programas de concordancia permiten realizar tareas como la recuperación de texto, el cálculo de frecuencias, la limitación del contexto en la búsqueda, etc. en un tiempo récord.
- En segundo lugar, no cabe duda de la gran utilidad de los corpus orales. Su creación es una tarea muy trabajosa, pero es la única forma de poseer grandes cantidades de muestras orales espontáneas y naturales que no se podrían obtener de otra manera. Los corpus orales son el único sistema para estudiar el habla espontánea.
- Los corpus ofrecen también una aproximación empírica y objetiva al estudio lingüístico, ya que se trata de estudiar la lengua en su uso real, sin partir de teorías ni de especulaciones subjetivas. Este rasgo representa el núcleo del conflicto entre la escuela chomskyana y la LC mencionado anteriormente. Sin embargo, a partir de los años ochenta del pasado siglo, el enfoque empírico respaldado por los avances tecnológicos y el estudio de la lengua en uso volvió a cobrar más importancia, con lo cual, la LC empezó a ganar terreno en el panorama lingüístico general y, en particular, en los estudios de Lingüística Computacional y en el Procesamiento del Lenguaje Natural.
- El uso de corpus ofrece la posibilidad de validar las hipótesis lingüísticas, demostrando así su validez o falsedad.
- Los corpus son un recurso indispensable para los estudios cuantitativos. La anotación de los textos facilita este tipo de trabajo. Este aspecto es fundamental, sobre todo, para las aplicaciones de procesamiento estadístico del lenguaje natural que requieren una fase de entrenamiento en un gran corpus de datos.
- El uso de corpus no solamente es la base para los estudios de Lingüística Computacional y de Procesamiento del Lenguaje Natural, sino que también es imprescindible para el estudio y el desarrollo de varias aplicaciones en áreas como la lexicología y lexicografía, la terminología, la gramática, la sociolingüística, la semántica, la pragmática, la enseñanza de una lengua extranjera o de la lingüística, la lingüística histórica, la dialectología o la psicolingüística.

1.4 Perspectivas de futuro

La LC, como disciplina (para algunos método) joven que es, debe avanzar aún mucho en algunos de sus aspectos. Su evolución corre de forma paralela al desarrollo tecnológico en general y a los avances en *hardware* y

software, que, por otra parte, son espectaculares de un año para otro. En cuanto a su utilidad dentro del campo de la lingüística no caben dudas. A este respecto son relevantes las diez razones expuestas por Svartvik en 1991, por las cuales el uso de un corpus podía ser de gran ayuda para la lingüística en general:

- Enunciación de teorías generalizables más objetivas que las basadas en la observación introspectiva.
- Exigencia de verificación real en la investigación científica.
- Necesidad de la utilización de un corpus en estudios dialectológicos, de registro o de estilo, al igual que en los estudios diacrónicos.
- Establecimiento de frecuencias de uso (relaciones frecuencia-registro).
- El corpus no es una mera muestra, sino un recurso teórico para lingüistas que trabajan con sistemas de reglas con probabilidad intrínseca.
- Es una fuente de información esencial para algunos campos mencionados anteriormente (Procesamiento del Lenguaje Natural, enseñanza, traducción automática, etc.)
- Posibilidad de explicación total de diferentes características del lenguaje.
- Posibilidad de compartir una base común de análisis sobre un mismo fenómeno con otros investigadores.
- Los corpus disponibles de forma gratuita dan la posibilidad de acceder a material lingüístico de todo tipo desde cualquier lugar del mundo, especialmente para estudios de lenguaje oral.
- El método introspectivo no es válido para lingüistas no nativos.

Leech (1991) destaca tres grandes áreas donde la LC debe centrar sus esfuerzos:

- Desarrollo de corpus básicos
- Desarrollo de herramientas para el tratamiento de corpus
- Desarrollo de sistemas de anotación de corpus

A pesar de que han pasado casi dos décadas desde que Leech apuntó dichos propósitos, éstos siguen vigentes en la actualidad. No sólo se debe mejorar la metodología en cuanto a la elaboración de corpus y aumentar el número y tipos de los mismos, sino que aún estamos lejos de conseguir herramientas de tratamiento de texto que funcionen de forma absolutamente automática sin necesidad de una posterior revisión manual por parte del lingüista. Por otra parte, las tres actividades mencionadas por Leech deben perseguir un mismo fin: la estandarización, mediante la cual sería posible compartir y usar de forma útil cualquier corpus.

Čemak (2002) propone diez preceptos tanto para la Lingüística Computacional como para la LC:

- *Garbage in, garbage out*: dicho aforismo, muy conocido en el campo de la informática, hace referencia al hecho de que un ordenador es capaz de procesar y producir cualquier tipo de información por muy absurda que sea. Por este motivo, es necesaria la supervisión humana y, al menos hasta el día de hoy, imposible la realización de tareas lingüísticas de una forma totalmente automática.
- Cuantos más datos poseamos mejor, pero hay que tener en cuenta que nunca habrá suficientes datos para explicar todos los fenómenos.
- Los datos directos proporcionan la mejor información. A veces el tipo de anotación que se utilice para etiquetar un texto puede adulterar el producto inicial[4].
- No existe ningún algoritmo universal que abarque todas las posibilidades de tratamiento y que sea transferible de una lengua a otra.
- Los lematizadores inventan nuevos términos, creando entidades inexistentes y sugiriendo otras tantas falsas[5].
- Los estudios estadísticos no siempre son productivos. En ocasiones se limitan a facilitar cantidades numéricas que el lingüista no es capaz de interpretar.
- La lengua es regular e irregular, y en relación a esta última característica, existen fenómenos no procesables de forma automática mediante un algoritmo.
- El objetivo principal de la lengua es codificar y descodificar el significado. Y puesto que el significado no se restringe a las palabras únicamente, es un error concentrarse sólo en ellas.
- Actualmente la tarea de alineamiento no puede realizarse de forma automática, sino que sigue siendo un trabajo puramente manual.
- También hay que preguntarse cuántas de las teorías y programas realizados en el campo de la LC no funcionan y jamás se menciona.

Según este último autor, "Much less has been achieved, so far, than has been claimed, particularly by computational linguists" (Čemak, 2002:266). Y más adelante continua: "What is really needed is a steady increase and perpetual growth of even, by present standards, very large corpora of billion of words, which should be as much representative as possible". Está claro que aún queda un gran camino por recorrer para la LC, que sus herramientas deben mejorarse y que la recolección de corpus es una tarea casi tan infinita como la lengua. A pesar de ello, los corpus son una fuente de datos cada vez

[4] La anotación de un texto enriquece al mismo; no obstante, las etiquetas creadas para este propósito deben ser claras y mantener un criterio sistemático. Además, se debe proporcionar al usuario una guía que facilite la comprensión del etiquetario.

[5] Esto ocurre sólo si son *stemmers*, es decir, programas que eliminan los sufijos y crean una raíz. Los lematizadores creados para el español y otras lenguas flexivas se basan en diccionarios de lemas, luego no inventan nuevos términos.

más imprescindible, y ello queda bien reflejado en la siguiente afirmación de Fillmore:

> "I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way". (Fillmore, 1991:35)

2. La ontogénesis del lenguaje[6]

2.1 Reseña histórica sobre el estudio del lenguaje infantil

El interés por la adquisición y el desarrollo del lenguaje durante los primeros años de vida del ser humano ha existido desde siempre. Es relevante cómo San Agustín en sus *Confesiones* ya se planteaba esta duda, que trató de resolver haciendo un ejercicio de memoria con técnicas retrospectivas sobre su propio proceso de adquisición. Sin embargo, este tipo de disertaciones no pueden considerarse desde un punto de vista científico, ya que las limitaciones de dicha técnica son evidentes y este proceso de retrospección resulta poco sistemático y exhaustivo.

Sin embargo, en la Antigüedad el interés por el origen del lenguaje no era tanto de carácter científico, sino que estaba influido por cuestiones religiosas y políticas. Son conocidos los experimentos llevados a cabo por gobernantes con la intención de averiguar cuál fue la primera lengua y por tanto la de mayor prestigio y poder. Algunos de estos experimentos de los que hoy tenemos conocimiento son el realizado por el faraón Psamético I en el siglo VII a. C., quien encargó a un pastor que se hiciera cargo de dos niños y los mantuviese aislados de todo contacto lingüístico; o Jaime IV de Escocia que realizó un experimento similar en 1493.

Tras el Renacimiento, se produce una renovación en el estudio de la adquisición del lenguaje con el empirismo inglés y la Ilustración francesa. La nueva visión antropocentrista predominante en la época lleva a los científicos a interesarse por el funcionamiento del pensamiento y la comunicación humana.

En el siglo XIX se produce un cambio en el ámbito naturalista, filológico y pedagógico. Se inician los estudios observacionales, caracterizados por una mayor sistematicidad y centrados en el estudio de los cambios que se producen en el proceso hacia la madurez.

Quizás, los primeros estudios llevados a cabo con una mínima metodología científica sean los de Tiedermann en 1787, basados en la evolución lingüística de sus propios hijos mediante diarios de observación evolutiva; o el de Itard, científico que estudió el caso del famoso "niño salvaje" de Aveyron durante los primeros años del siglo XIX. No obstante, a pesar de que estudios casos seguían una metodología científica, no dejan de ser meros casos aislados.

[6] El presente capítulo es un breve resumen recopilado de las obras de A. Anula (1998), H. Boada (1986), C. Boston (1976), J. Brown (1973), R. Brown (1981), J. Bruner (1984), D. Crystal (1981), P. Dale (1992), S. Gili Gaya (1972), F. Giuseppe (1971), F. Hernández Pina (1984), A. Kuria (1956), J. Piaget (1965), M. Serra *et al.* (2000) y M. Siguán (1983/1984).

Las primeras muestras de preocupación por el estudio del desarrollo del lenguaje desde una perspectiva exclusivamente científica arrancan a principios del siglo XX. Aparecieron las técnicas de observación naturalista. Éstas, introducidas por Darwin a mediados del siglo XIX para el estudio de las especies animales, fueron aplicadas al estudio de cualquier aspecto del desarrollo humano durante la primera mitad del siglo XX. Así, aparecieron las notas biográficas o diarios de bebés. En ellas, los investigadores observaban y analizaban el desarrollo lingüístico de un bebé mientras tomaban nota de ello. Fue de esta forma como surgieron los primeros estudios longitudinales, en donde el propio hijo del investigador era el objeto de un estudio, que abarcaba todo su desarrollo inicial. Este es el caso del matrimonio Stern, que en 1907 publicaron su obra *Die Kinderspreche,* que incluía una descripción lingüística detallada de la evolución de sus hijos durante los primeros seis años de edad. Estudios de este tipo aumentaron en Europa y Estados Unidos.

Sin embargo, y a pesar del beneficio que suponía un estudio detallado en tiempo real, seguían existiendo importantes carencias metodológicas, ya que el uso de las notas biográficas era insuficiente para responder a la velocidad de progreso del niño y a la multitud de información que subyace a toda interacción, y el resultado final solía ser un cuaderno con anotaciones sobre léxico principalmente.

Uno de los primeros teóricos que escribió sobre la ontogénesis del lenguaje fue el biólogo suizo Jean Piaget, famoso por sus teorías en psicología evolutiva. En su obra *El lenguaje y el pensamiento en el niño* afirma el origen intelectivo del lenguaje. En contraposición, poco después el psicólogo ruso Lev Vygotsky expuso sus teorías sobre la adquisición del lenguaje como un proceso de origen social y cultural, y no intelectual como proponía Piaget.

A partir de 1930, aparece un nuevo método de estudio basado en la aplicación de tests, conocido como método psicométrico y diferencial o correlacional. Estos tests relacionaban lenguaje e inteligencia, y permitían recoger un amplio número de datos de una gran variedad de sujetos. Así, las muestras obtenidas eran amplias y se basaban principalmente en cuestiones gramaticales.

En los años 60 se produce una reacción contra el conductismo asociacionista que no resolvía gran parte de las dudas que se planteaban. Así, surge el enfoque racionalista por influencia de las teorías de Chomsky y la idea del carácter innato del lenguaje. En el siguiente apartado se hará una descripción más detallada de las principales teorías y posturas sobre la adquisición del lenguaje en los últimos años.

En la década de los 70 se produce un giro en las investigaciones dirigido hacia el carácter comunicativo del lenguaje, con una clara orientación funcional en oposición a la perspectiva formal imperante hasta ese momento. En ese momento, los estudios se centran en la interacción del niño con una clara visión pragmática. Así se empiezan a recoger muestras de habla espontánea, con la finalidad de obtener datos reales sobre cómo se comunica el niño. De hecho, el interés no sólo se centra en qué dice el niño, sino también en el papel del adulto en su comunicación con el niño, es decir, en el estudio de la interacción entre ambos.

Además, el desarrollo de las nuevas tecnologías en los últimos 50 años indujo un cambio cualitativo en las formas de actuación. En concreto, la aparición de las cintas magnetofónicas permitió la recolección de gran cantidad de grabaciones en diferentes contextos, así como la participación de numerosos sujetos en periodos de dos o tres años. Conseguido el almacenamiento de las interacciones en formatos de audio, ya era posible transcribir el contenido de estas, aunque por otra parte comenzó la polémica en torno al formato o código de la transcripción. Aunque surgieron numerosos grupos y proyectos de investigación con el objetivo de recoger muestras de lenguaje infantil, cada uno de ellos usaba un código de transcripción propio y distinto del resto. Este hecho dificultaba la comparación de resultados y el intercambio de información. Fue la generalización del uso de ordenadores, a finales del siglo XX, lo que propulsó un cambio en la mentalidad de la comunidad científica en torno a este campo de interés, abandonando su carácter individualista y sectario para confluir en una nueva idea de contraste e intercomunicación. Así comenzaron a utilizarse formatos de transcripción compartidos y aceptados que posibilitaban el aprovechamiento de los datos por todos aquellos interesados en el estudio de la adquisición y el desarrollo del lenguaje.

La paulatina introducción de nuevas herramientas tecnológicas y la asunción de metodologías y procedimientos comunes hizo posible, finalmente, el diseño de los primeros corpus de lenguaje infantil. En ellos se recopilaban cientos de grabaciones, realizadas a niños de diferentes edades, que eran transcritas siguiendo un único código de transcripción. Este tipo de recursos suponía una fuente importante de datos para cualquier tipo de investigación psicolingüística.

2.2 Principales teorías

A continuación, haremos un breve repaso de las principales teorías sobre la adquisición y desarrollo del lenguaje infantil y algunas de sus características. Como hemos mencionado en el apartado anterior, a pesar de que el interés por la adquisición lingüística ha existido desde siempre, no es hasta el siglo XX cuando se le da un enfoque más científico al estudio de este fenómeno. Es en dicho siglo cuando empiezan a aparecer teorías sobre

la ontogénesis del lenguaje, con la intención de esclarecer sus orígenes y funcionamiento. Podemos hablar de tres enfoques básicos: teorías conductistas, innatistas y cognitivistas.

- El conductismo de Skinner

El conductismo nació en los primeros años del siglo XX con los experimentos de científicos como Watson en la Universidad de Chicago o Pavlov en Rusia. Ambos llevaron a cabo experimentos con animales para demostrar que mediante el condicionamiento era posible modificar el comportamiento. El conductismo basa sus estudios en hechos observables y objetivos, eliminando el carácter mentalista de su ideología. Según esta teoría, el comportamiento humano puede ser explicado mediante la fórmula estímulo-respuesta.

Uno de los autores más destacados dentro del conductismo por sus estudios sobre la adquisición del lenguaje fue Skinner. Para él, el aprendizaje del lenguaje se produciría por simples mecanismos de condicionamiento. En un principio, los niños simplemente imitarían, para después asociar determinadas palabras a situaciones, objetos o acciones. El aprendizaje del vocabulario y de la gramática se haría por condicionamiento operante. Los adultos que se encuentran alrededor del niño recompensarán la vocalización de enunciados correctos gramaticalmente, la presencia de nuevas palabras en el vocabulario, la formulación de preguntas y respuestas, etc. y castigará con la desaprobación todas las formas del lenguaje incorrecto, como enunciados agramaticales.

La principal crítica a esta teoría es que, a pesar de aceptar que la imitación puede jugar un papel importante en el proceso de adquisición de una lengua, no explica hechos como los errores por analogía tan comunes en el lenguaje de los niños. El lenguaje infantil no es un fiel reflejo del lenguaje adulto, sino que posee características propias que no son atribuibles a la imitación.

- El innatismo de Chomsky

En contraposición al conductismo, y debido a la insatisfacción que producían muchas de sus explicaciones, surge un enfoque totalmente opuesto conocido como racionalismo o mentalismo. Su máximo representante es Noam Chomsky, quien defiende que la mente es la única fuente de conocimiento y sólo mediante la introspección se puede alcanzar una explicación a los orígenes y funcionamiento del lenguaje. Chomsky propone la teoría del innatismo entendiendo que los seres humanos nacemos con una serie de conocimientos lingüísticos genéticamente determinados que se concretan en la relación con el medio. Propone la existencia de un "dispositivo para la adquisición del lenguaje" o LAD (Language Acquisition

Device), capaz de recibir el *input* lingüístico y, a partir de él, derivar las reglas gramaticales universales. El niño es capaz de generar de dicho *input* una gramática que produce oraciones bien estructuradas y que determina cuál es la forma en que deben usarse y comprenderse éstas. Este LAD sólo se pone en marcha cuando el niño se expone a los datos externos; si no hay exposición, el LAD no se activa. Así se explican casos como el de Genie, la niña de California que fue privada de contacto lingüístico hasta los doce años y que, a pesar de los intentos de sus educadores, nunca consiguió adquirir una competencia lingüística completa.

Aunque muchos psicólogos, biólogos y lingüistas no dudan del hecho de que el hombre está genéticamente predeterminado para la adquisición y uso del lenguaje, la teoría de Chomsky sigue dejando muchas dudas. Por ejemplo, no explica el porqué de las diferencias en producción de un niño a otro durante el proceso de aprendizaje; ni por qué unas estructuras se aprenden antes que otras. Dicha teoría se centra en los universales lingüísticos, dejando a un lado las diferencias. Además, concibe al lenguaje como un sistema modular independiente de otras capacidades cognitivas. Esto dejaría fuera de estudio a la inteligencia o la memoria, entre otras capacidades humanas que pueden tener una gran influencia en el proceso de aprendizaje de una lengua.

- Teorías cognitivas defendidas por Piaget, Vygotsky, Luria y Bruner

Todos los autores que se comentan a continuación tienen como idea de partida establecer hasta qué punto el lenguaje determina a la cognición. Todos ven el lenguaje como una parte más del desarrollo cognitivo general, y las diferencias entre ellos residen en al grado de importancia que cada uno otorga al papel del lenguaje en dicho desarrollo cognitivo.

Piaget, biólogo y más tarde psicólogo, se interesó por la evolución del conocimiento en los niños y en el papel que jugaba el lenguaje en dicha evolución. Se centró en el estudio de las actividades de los niños como manifestación directa de la inteligencia y llegó a la conclusión de que el lenguaje es una consecuencia del pensamiento y no al revés. La cognición no depende del lenguaje para su desarrollo, sino que éste es una forma de la función simbólica que forma representaciones y diferencia entre significante y significado. Para Piaget, el lenguaje es una actividad cognitiva auxiliar y no central como creen otros autores. Además, considera al contexto relativamente poco importante y escasamente influyente en los cambios cualitativos de la cognición. El niño es visto como constructor activo de su conocimiento y, por lo tanto, del lenguaje. Se proponen dos mecanismos constructores de las estructuras cognitivas para tratar con entornos cada vez más complejos: la organización y la acomodación. Estos principios son aplicables al estudio del desarrollo del lenguaje; éste se centraría en una expresión cada vez más clara y lógica del pensamiento y en una progresiva

socialización, basada en la capacidad progresiva del niño para comprender puntos de vistas ajenos (de lenguaje egocéntrico a social).

Vygotsky es un teórico dialéctico que enfatiza tanto los aspectos culturales del desarrollo como las influencias históricas. Para él, la reciprocidad entre el individuo y la sociedad, siendo definida ésta tanto histórica como culturalmente, es muy importante. El contexto de cambio y desarrollo es el principal foco de atención, dado que ahí es donde podemos buscar las influencias sociales que promueven el progreso cognitivo y lingüístico. Para Vygotsky, el habla es, fundamentalmente, un producto social y juega un papel decisivo en la formación de los procesos mentales. El lenguaje precederá al pensamiento, al que va estructurando: los niveles de funcionamiento intelectual dependerían de un lenguaje más abstracto. Además, habla y acción están íntimamente unidas: cuanto más compleja es la conducta y más indirecta la meta, más importante es el rol de la lengua.

A. R. Luria, investigador soviético y seguidor de Vygotsky, entiende que el lenguaje es un factor de gran importancia en el desarrollo cognitivo. Concibe el lenguaje como un elemento de regularización de la actuación y el comportamiento. A raíz de una serie de experimentos que llevó a cabo con niños, llegó a la conclusión de que, a diferencia de lo que afirmaba Piaget, el lenguaje no era un mero acompañamiento de la acción, sino que su función era mucho más primaria e influía directamente en la formación de la actividad humana.

Por último, y dentro de las teorías cognitivas, J. S. Bruner intenta conciliar las dos posturas anteriores. Para él, ni el desarrollo del pensamiento se consigue mediante el lenguaje, ni el lenguaje es una función periférica al desarrollo cognitivo, sino que el lenguaje sería el agente de dicho desarrollo. Bruner sugirió que el niño aprende a usar el lenguaje para comunicarse en el contexto de la solución de problemas; se enfatiza el aspecto comunicativo del desarrollo lingüístico más que su naturaleza estructural o gramatical. De acuerdo con Bruner, el niño necesita dos fuerzas para lograr el aprendizaje del uso del lenguaje. Una de ellas es equivalente al LAD de Chomsky; la otra fuerza sería la presencia de un ambiente de apoyo que facilite el aprendizaje del lenguaje. Bruner denominó a éste Sistema de Apoyo para la Adquisición de un Lenguaje o LASS (Language Acquisition Support System). Dentro de este LASS sería relevante la presencia del *maternés*, forma de comunicación que tienen los padres con sus hijos pequeños, que se caracteriza por su lentitud, brevedad, repetitividad, concentración en el "aquí y ahora" y en su simplicidad. Esta manera de comunicarse le permite al niño extraer la estructura del lenguaje y formular principios generales. Este *maternés* aparecerá generalmente en un contexto de acción conjunta, en el que el tutor y el niño concentran su acción en un solo objeto y uno de ellos "vocaliza" sobre él.

2.3 Evolución del lenguaje infantil

Los estudios sobre la adquisición y el desarrollo del lenguaje infantil realizados hasta hoy han establecido ciertos periodos o etapas en dicho proceso. A pesar de que existen grandes diferencias entre las teorías comentadas en el apartado anterior, todos los investigadores coinciden en establecer las siguientes etapas de desarrollo lingüístico en el niño:

- Comunicación prelingüística

Desde muy temprano, el niño y sus padres empiezan a comunicarse; además, el niño presenta una especial atención al habla, dirigiendo su mirada hacia el adulto como respuesta a las voces.

La primera herramienta comunicativa del niño es su llanto, que producirá una acción paterna con relación al contexto, atendiendo a las necesidades del bebé. Ya a las seis semanas, el niño comienza a sonreír, estableciendo nuevos lazos comunicativos. A las doce semanas el niño sonríe cuando le hablan y produce sonidos de carácter vocal modulado, que son mantenidos durante quince o veinte segundos. A la edad de veinte semanas no sólo emite sonidos de carácter vocal, sino que éstos se entremezclan con consonantes.

Alcanzados los seis meses aparece un balbuceo semejante a emisiones monosilábicas. Las más comunes se semejan a sílabas del tipo /ma/, /mu/, /da/, /di/. A los ocho meses se hacen más frecuentes las repeticiones, se distinguen estructuras de entonación en las emisiones y estas pueden indicar emociones. Finalmente, sobre los diez meses las emisiones se mezclan con sonidos de juego como los gorgoritos y el niño realiza imitaciones de sonidos.

- Prelenguaje

Las funciones del lenguaje en esta etapa serían, primero, de orden pragmático (obtención de objetos y servicios, control de la conducta de las personas) y expresivo (manifestaciones de agrado o rechazo). Más adelante, el lenguaje tendrá una función heurística (obtención de información sobre el medio inmediato) e imaginativa (creación de realidades propias).

El niño emite sus primeras palabras al final del primer año. En un principio, se referirán a objetos y personas, pero aún no expresarán peticiones o descripciones del ambiente. Estas primeras palabras se caracterizan por usar un número limitado de elementos fonéticos y por referirse a categorías más amplias que las aceptadas por la lengua adulta, tanto en lo que se refiere a objetos como a acciones. Las palabras, en esta

época, expresan ideas que un adulto comunicaría mediante oraciones completas; a esto se le denomina habla polisintética u holofrástica.

Entre los dieciocho y los veinte meses se empiezan a formar frases que constan generalmente de dos palabras, pudiendo llegar a siete[7]. El niño reproduce las frases que escucha acortándolas; este acortamiento se conoce como "lenguaje telegráfico" (Brown, 1981) y se caracteriza por la retención de los elementos con alta carga semántica (sustantivos, verbos, adjetivos) y la omisión de aquellos elementos que cumplen una función gramatical. En el aspecto formal, estas frases se caracterizan por mantener el orden original. La elección de las palabras por parte del niño se haría en razón de la función referencial de los significantes (los sustantivos, los verbos y los adjetivos hacen referencia a objetos, acciones o cualidades presentes y actuales), la brevedad de la palabra, la imposibilidad de inferencia de la palabra por el contexto y la acentuación más fuerte que se le otorga a las palabras con mayor carga semántica dentro de la oración (Brown, 1981).

A partir de los dos años de edad, el niño posee ya un vocabulario que excede las cincuenta palabras. Comienza a utilizar la morfología flexiva del lenguaje, es decir, aquellos elementos que modifican el significado de las palabras, dándoles una connotación más específica (pluralidad, tiempo) o estableciendo relaciones entre elementos. El orden en que éstas se aprenden, al menos en inglés, es bastante consistente y depende, en parte, de la complejidad de los morfemas (Dale, 1992). Un hecho interesante de este aprendizaje de flexiones es la hiperregularización, es decir, el uso de los mismos morfemas de los verbos regulares en los irregulares, lo que muestra que el niño busca patrones en el lenguaje.

- Lenguaje

A esta etapa se llega cuando el niño alcanza los tres o cuatro años. Se produce "un abandono progresivo de las estructuras elementales del lenguaje infantil y de su vocabulario específico, sustituidos por construcciones cada vez más acordes con el lenguaje adulto" (Bouton, 1976). En el caso del inglés, se aprenden las principales estructuras gramaticales, que le permitirán formular oraciones declarativas, negativas, interrogativas e imperativas.

Ya en esta etapa se da la producción de verdadero diálogo, adquiriendo así el lenguaje una función informativa, es decir, la de producir lenguaje sin que éste tenga relación con algún elemento del contexto inmediato del niño (Boada, 1986). Se produce una interacción entre la actividad verbal libre, con la gramática autónoma de la etapa anterior, y una actividad mimética,

[7] La mayor parte de los estudios se han realizado teniendo como referencia la lengua inglesa. Por tanto estos datos pueden variar notablemente de una lengua a otra.

tendiente a adaptarse al mundo lingüístico del adulto. Según Luria (1956), a estas características se le sumaría la función autorreguladora de la conducta del lenguaje, que determina que el niño se plantee metas y busque los medios para lograrlas de acuerdo a las categorías que el mundo lingüístico le ofrece, pudiendo abstraerse de lo concreto e inmediato.

La siguiente tabla presenta un resumen de la evolución lingüística del niño. Para su realización, hemos seguido las etapas expuestas por Anula (1998):

Etapa	Edad	Características
Sonidos no lingüísticos	0-2 meses	Llanto, suspiros, etc.
Fase prelingüística	2-7 meses	Arrullo
Balbuceo	7-12 meses	Emisiones silábicas aisladas
Periodo holofrástico	12-18 meses	Palabras aisladas
Habla telegráfica	18-36 meses	Aparece la sintaxis
Estructura adulta	3-4 años	Mecanismos básicos
Estructura adquirida	4-6 años	Aprendizaje completo

Tabla 2.1 Evolución lingüística del niño

Las edades son siempre orientativas, nunca exactas, ya que cada individuo muestra patrones individuales durante todo el proceso de aprendizaje. No obstante, el proceso parece ser similar para cualquier niño que no presente ninguna anomalía o patología lingüística, y los periodos aquí establecidos son bastante representativos.

3. Los corpus y la ontogénesis del lenguaje

3.1 Estado de la cuestión

Los corpus siempre han sido el método indispensable para el estudio del lenguaje infantil, tarea para la cual era necesario recoger datos reales sobre la lengua del niño, anotando sus características. Ya se ha comentado en el capítulo anterior que el procedimiento más habitual era la anotación de todos los fenómenos y cambios que presentaba la lengua de un niño, desde las primeras palabras hasta que éste alcanzaba una competencia lingüística similar a la del adulto. En el siglo XIX, aparecieron los llamados "diarios de bebés" en los que el investigador iba anotando todos los datos de una forma más sistemática. No obstante, el concepto actual de corpus y la nueva tecnología disponible para la recolección del habla espontánea ha supuesto una revolución en los estudios sobre adquisición lingüística. Ahora es posible recopilar una gran cantidad de datos mediante los diferentes sistemas de grabación, transcribir el contenido de dichas grabaciones y almacenar ambos tipos de datos (audio y texto) en soporte electrónico. Al igual que para la LC, para la psicolingüística el ordenador ha supuesto una herramienta muy potente que ha dado una nueva perspectiva a la investigación, facilitando el trabajo y ampliando las posibilidades.

Actualmente, existen tres líneas de investigación básicas para el estudio de la adquisición del lenguaje:

- Una línea lingüística generativa, que mediante el método de la deducción parte de principios formales y busca su verificación en los datos.
- Una perspectiva de investigación más heterogénea, de base psicológica y metodología inductiva, que se basa en el trabajo con corpus y lleva a cabo diferentes experimentos.
- Una tercera línea de corte pragmático que centra su interés en los diferentes usos del lenguaje en la interacción.

Además, la tendencia actual se caracteriza por los trabajos sobre fenómenos lingüísticos muy específicos y la amplitud de los estudios a otras lenguas aparte del inglés. Hasta hace poco, la mayoría de las investigaciones llevadas a cabo sobre adquisición y desarrollo lingüístico, al igual que los corpus recogidos, eran sobre la lengua inglesa. Es necesaria la creación de corpus en una gran variedad de lenguas, ya que "[…] estamos tratando un ámbito de conocimiento en el cual las hipótesis y predicciones han de ser contrastadas sistemáticamente con lo datos procedentes del comportamiento lingüístico de los sujetos […]" (Serra *et al*., 2000:75).

Los corpus de lenguaje infantil siguen dos tipos de diseño: transversal o longitudinal. El primero está formado por interacciones de dos o más grupos de varios sujetos, normalmente divididos por edad; mientras que el corpus longitudinal suele estar formado por material obtenido de un solo sujeto grabado durante un largo periodo de tiempo. Entre las ventajas de un corpus transversal destacan la adquisición de un gran número de datos en un breve periodo de tiempo y la posibilidad de establecer patrones generales de comportamiento lingüístico; por el contrario, este tipo de corpus no muestra el patrón evolutivo de un sujeto, como ocurre con los corpus longitudinales, sino que solamente es posible establecer las diferencias entre los distintos grupos de edad.

Dependiendo de la finalidad del estudio que se vaya a llevar a cabo, será más adecuado un tipo de diseño u otro. Existe también un tercer tipo de diseño, el llamado secuencial, que consiste en combinar los dos tipos anteriores, siguiendo la evolución de los mismos grupos a lo largo de un periodo de tiempo.

En cuanto a la metodología, podemos hablar del método observacional y del método experimental. El primero de ellos procede de la tradición antropológica —también llamado método etnográfico— y su objeto de estudio son situaciones naturales en contexto familiar. Para ello, el investigador intenta acercarse al sujeto para que éste se familiarice con él y evitar así un intrusismo llamativo. Aunque la intención es la recogida de datos espontáneos, existe siempre un cierto control sobre la situación, guiándola por ejemplo con entrevistas semi-estructuradas. Los trabajos tradicionales han seguido en su mayoría el método observacional, pero en muchos casos los resultados han sido básicamente descriptivos, dejando muchas incógnitas sin aclarar. La principal ventaja de este método es que el sujeto produce el lenguaje de forma absolutamente libre. Por el contrario, el investigador debe conformarse con lo producido, pudiendo no obtener ningún tipo de dato sobre un determinado fenómeno.

Precisamente cuando lo que se pretende es estudiar un fenómeno concreto, lo ideal es recurrir al método experimental que, al contrario que el observacional, crea situaciones artificiales para provocar la producción del fenómeno que se quiere estudiar y verificar así hipótesis muy específicas. La principal desventaja del método experimental es que en muchas ocasiones se cuestiona su validez debido a la manipulación de la realidad que implica. Además, requiere una colaboración por parte de los sujetos difícil de conseguir en edades muy tempranas.

3.2 Los corpus infantiles en la actualidad

En España, la experiencia acumulada en el desarrollo de corpus orales de lenguaje infantil se ha limitado tradicionalmente a los estudios longitudinales de varios grupos de menores. En este tipo de estudios, autores particulares siguen a un sujeto o grupo de sujetos reducido durante un periodo de tiempo determinado, durante el cual se llevan a cabo grabaciones sistemáticas para comprovar su evolución. Por esta razón no se han dado las bases en nuestro país para la elaboración de un corpus representativo de población infantil accesible para todo investigador interesado en este campo de estudio. No obstante, en la actualidad comienzan a aparecer corpus más amplios, amparados por universidades o instituciones concretas, que pretenden estudiar algún aspecto del lenguaje infantil a partir de la recolección de muestras con un mayor número de participantes. Entre estos encontramos los siguientes:

- *Corpus de vocabulario del niño de 6 a 14 años* (Diccionario de frecuencias). Universidad de Granada.
- *Representación de categorías semánticas en niños ciegos de nacimiento de edad escolar (E.G.B).* Universidad Nacional de Educación a Distancia.
- *Adquisición, desarrollo y representación de categorías semánticas en niños de edad escolar*, UNED.
- *Diferencias individuales en la adquisición del lenguaje*, de la Universidad de Barcelona.
- *Corpus de habla infantil*, CSIC-UNED.
- *Disponibilidad léxica de los adolescentes*, Universidad de Salamanca
- *Corpus ASOPRACIPEIA (Análisis Socio-Pragmático de la Comunicación Intercultural en las Prácticas Educativas: hacia la Integración en el Aula)*, Universidad Autónoma de Madrid.

Los dos primeros corpus, el de la Universidad de Granada y el de la UNED, son corpus textuales, es decir, basados en muestras de lenguaje escrito. Respecto a los cuatro siguientes, sí son corpus de lengua oral. El corpus de la UNED, *Adquisición, desarrollo y representación de catagorías semánticas en niños de edad escolar*, aparece como no finalizado en el *Informe sobre recursos lingüísticos para el español (II)* elaborado por el Instituto Cervantes en el año 1996. Según este informe, cuenta con unas 20.000 palabras, lo que cuantitativamente no resulta muy significativo. *Diferencias individuales en la adquisición del lenguaje*, corpus desarrollado por la Universidad de Barcelona, se llevó a cabo con 10 sujetos grabados en diez sesiones por año entre los 3 y los 4 años. Algo parecido ocurre con el corpus elaborado por el CSIC y la UNED, *Corpus de habla infantil*, que se centra en 6 hablantes en 15 o 20 sesiones anuales desde 1991.

Esto nos conduce a la afirmación hecha anteriormente sobre los tipos de estudios tradicionales en el campo del desarrollo del lenguaje infantil: son estudios longitudinales basados en un número muy reducido de sujetos, hecho que puede afectar de forma negativa a la representatividad que se persigue en este tipo de investigaciones.

Por último, los corpus elaborados en la Universidad de Salamanca (*Disponibilidad léxica de los adolescentes*) y en la Universidad Autónoma de Madrid (*ASOPRACIPEIA*) se centran en sujetos adolescentes que ya han superado el periodo básico de adquisición lingüística.

3.3 CHILDES

Childes Language Data Exchange System (CHILDES) es el corpus de lenguaje infantil que actualmente se presenta como referencia internacional. Fue concebido en 1984 como un sistema de intercambio de datos, hoy accesible en Internet en la dirección <http://childes.psy.cmu.edu/>. En dicha página encontramos, además del corpus, manuales de transcripción, aplicaciones informáticas y herramientas de apoyo. Está caracterizado por ser un corpus multilingüe que cuenta con un subcorpus de lenguas romances, con muestras de catalán y castellano. CHILDES se nutre de aportaciones por parte de investigadores particulares que, siguiendo las instrucciones de los manuales de transcripción y codificación de los textos, colocan sus propios trabajos en la página web. Por su gran importancia en el campo de la investigación psicolingüística y por ser en gran medida el referente para la creación de nuestro corpus, creemos necesario hacer una descripción detallada de su contenido y posibilidades.

3.3.1 Presentación

CHILDES se fundó en 1984 en la Universidad Carnegie Mellon, dentro del departamento de Psicología y bajo la dirección de Brian MacWhinney. Su principal meta es la recopilación de grandes cantidades de datos con el fin de estudiar dos elementos principales en el lenguaje infantil: universales y diferencias, tanto dentro de una única lengua como comparando lenguas diferentes. El estudio de los dos factores mencionados requiere de una enorme cantidad de datos para poder formular teorías generales.

Las herramientas que se facilitan a través de la página de CHILDES son:

- La base de datos de las transcripciones
- Programas para el análisis de las transcripciones
- Métodos para la codificación lingüística
- Sistemas para enlazar las grabaciones (audio y video) con sus respectivas transcripciones.

El software necesario para trabajar con los datos facilitados en CHILDES es el siguiente: *Quicktime player*, para acceder a los archivos media; *Adobe reader*, para leer los manuales; *Winzip*, para descomprimir el corpus; y *Unicode fonts: Arial FixedSys*, para visualizar los caracteres. En la página se facilita la descarga de las fuentes Unicode.

El siguiente paso es descargar e instalar el programa CLAN, y posteriormente leer los manuales relacionados: *CHAT Transcript System*, que nos indica cómo transcribir una conversación en la forma estándar establecida para CHILDES, y *CLAN Program Manual*, que nos muestra cómo usar el programa en cuestión.

Si estamos interesados únicamente en estudiar la base de datos existente sin la intención de aportar material propio a la misma, habría que seguir los mismos pasos e instalar tanto CLAN como el sistema Unicode. Una vez tenemos los programas instalados, accederíamos a la base de datos y descargaríamos o bien todo el corpus o bien sólo uno de los subcorpus (si sólo queremos información sobre lenguas romances, accederíamos únicamente al subcorpus de las mismas). Tenemos la opción de acceder solamente al texto transcrito, o bien acompañarlo de su correspondiente sonido o vídeo.

Con CLAN tenemos la posibilidad de extraer información lingüística sobre palabras o estructuras concretas; para ello necesitamos leer el manual sobre el funcionamiento del programa. Dicho manual nos proporciona los diferentes comandos para la extracción de información, como *mean length utterance* (ratio de morfemas sobre enunciados), *frequency* (contar el número de palabras por fichero y calcular la ratio *type/token* o medida de diversidad léxica), *keyword and line searching* (busca información sobre palabras especificadas por el usuario y presenta dichas palabras en su contexto) y *combination search* (busca información sobre combinaciones específicas de palabras o grupos de caracteres).

3.3.2 Contenidos de la página

Cuando accedemos a la página <http://childes.psy.cmu.edu/> encontramos los siguientes apartados[8]:

- *System* (Sistema).
- *Programs and Database* (Programas y Base de datos).
- *Links* (Enlaces).
- *Manuals* (Manuales):
- *Phonology and Fonts* (Fonología y Fuentes).

[8] Hemos mantenido los contenidos de CHILDES en la lengua original, el inglés, añadiendo entre paréntesis la traducción.

- *Teaching with CHILDES* (Enseñanza con CHILDES).
- *Special Procedures* (Procedimientos especiales).
- *Morphology and Lexicon* (Morfología y Lexicón).

A continuación, haremos una breve descripción de lo que el usuario puede encontrar dentro de cada apartado.

Sistema

- *Ground rules* (directrices): en este primer aparatado se hace referencia a cuestiones legales y éticas sobre el uso del corpus, el tratamiento que se haga de la información extraída del mismo o el procedimiento correcto que deben seguir las contribuciones de los usuarios, en concreto en referencia a la necesidad de la existencia de permisos por parte de los participantes en la interacción.
- *Overviews and introductions* (perspectivas generales e introducciones): aquí podemos encontrar diferentes descripciones en distintas lenguas del proyecto CHILDES y su funcionamiento. La mayor parte de la literatura está escrita en inglés, aunque también podemos encontrar resúmenes en otras lenguas.
- *Membership list* (lista de miembros): CHILDES ofrece la posibilidad de pasar a formar parte de su lista de miembros y así poder mantener contacto con el resto de usuarios.

Programas y Base de datos

Este apartado, junto con el dedicado a los manuales, es básico para poder acceder a la información y comprender el funcionamiento de CHILDES. Dentro de *Programas y Base de datos* encontramos el programa CLAN de descarga gratuita, la base de datos con los distintos subcorpus divididos en lenguas o grupos de lenguas, los procedimientos para contribuir al crecimiento del corpus con aportaciones personales y el software relacionado que puede ser de utilidad. En lo que se refiere al programa CLAN, se indica al usuario cómo descargarlo e instalarlo en su ordenador personal, y posteriormente se le redirige al manual para la comprensión de su uso. En cuanto a la base de datos, presenta el siguiente aspecto:

- Inglés-EEUU: 41 archivos.
- Inglés-Reino Unido: 10 archivos.
- Celta: 3 archivos.
- Este de Asia: que a su vez se divide en seis lenguas, cantonés, chino, indonesio, japonés, coreano y tai, con un total de 16 archivos entre las diferentes lenguas.
- Germánico: dividido en afrikáans, danés, holandés, alemán y sueco, con un total de 18 archivos.

- Romance: que incluye seis lenguas, catalán, francés, italiano, portugués, rumano y español, con un total de 40 archivos.
- Eslavo: incluye croata, polaco y ruso en un total de 5 archivos.
- Otras lenguas: aquí se recogen subcorpus en vasco, estonio, farsi, griego, hebreo, húngaro, sesotho, tamil y turco, reuniendo 16 archivos.
- Bilingües: 21 archivos bilingües en diferentes lenguas.
- Clínicos: 23 archivos en varias lenguas.
- Cuentos de ranas: 13 archivos en diferentes lenguas.
- Otras narraciones: 5 archivos en diferentes lenguas.

La organización del corpus no nos permite conocer de forma sencilla datos como el número total de palabras de todo el corpus, el número de palabras de un subcorpus concreto o únicamente el de una transcripción determinada. Para conocer toda esa información sería necesario descargar el corpus completo en nuestro ordenador (los 211 archivos, algunos de los cuales no son un solo documento, sino una carpeta dentro de la cual encontramos varias transcripciones), visualizar todas y cada una de las transcripciones y aplicar el programa CLAN para tareas del tipo recuento de palabras. Sería de gran ayuda la inclusión de un esquema que nos facilite toda esa información y mediante el cual nos podamos hacer una idea, de forma gráfica, del diseño del corpus con sus diferentes subcorpus y descripción del contenido de cada uno.

Dentro de la sección *The Database* (La Base de datos), por tanto, encontramos todos los archivos de texto por una parte y los de audio y vídeo por otra. Para poder visualizar conjuntamente la transcripción y su correspondiente sonido o vídeo debemos descargar ambos elementos, guardarlos en una misma carpeta y abrirlos con el programa CLAN. El principal problema que puede encontrar el usuario en este aspecto es que no todas las transcripciones que se contienen en CHILDES incluyen la grabación original, es decir, que muchas de ellas sólo están disponibles en formato de texto y no es posible acceder al sonido. La información sobre el número de transcripciones que aportan la grabación en sonido o vídeo a modo de esquema sería igualmente de gran utilidad para el usuario.

Links

Por medio de esta sección podemos acceder a la *Talkbank Database* (proyecto interdisciplinario creado en la Carnegie Mellon University y la Universidad de Pennsylvania para el almacenamiento de grandes cantidades de transcripciones y archivos audiovisuales), a la IASCL, *Internacional Association for the Study of Child Language* y a otras tantas páginas relacionadas con el estudio del lenguaje infantil, entre las que encontramos una española: EQUIAL, Equipo de Investigación sobre la Adquisición del Lenguaje, dirigido por la Dra. López Ornat desde la Facultad de Psicología de la Universidad Complutense de Madrid. Dicho grupo de investigación

centra sus estudios en los primeros tres años de adquisición de la lengua y han elaborado un corpus longitudinal en el que ha participado una niña entre los 1;7 y los 4;0 años de edad.

Manuales

La sección está dedicada a los dos grandes manuales necesarios para trabajar con la información contenida en CHILDES y para la aportación de nuevos datos. Dichos manuales son *CHAT Transcription* y *CLAN Programs*. En el primero se dan las pautas necesarias para realizar una transcripción de forma adecuada, incluyendo una guía de las convenciones establecidas para el formato CHAT. El segundo, indica cómo descargar y utilizar el programa CLAN para el análisis de los textos transcritos.

Además de los manuales mencionados, se incluye un tercero sobre la transcripción del lenguaje de signos, el *BTS sign transcription system*. Y por último, encontramos una guía sobre el contenido de CHILDES en la que se hace una breve descripción de la información que se incluye en cada uno de los subcorpus.

Fonología y Fuentes

Este apartado está dedicado a la investigación en fonología y en él se presenta el proyecto *Phon*, *A Shared Database for the Study of Phonological Development*, dedicado al estudio fonológico basado en corpus; además se facilita al usuario descargar Unicode e IPA para Mac y Windows.

Enseñanza con CHILDES

CHILDES se propone aquí como herramienta académica para ilustrar el proceso de adquisición de la lengua. Se sugieren tópicos, metodologías (cómo utilizar muestras del corpus para que los estudiantes las analicen en los distintos niveles lingüísticos, o como ejemplo de transcripción y utilización de programas informáticos para su tratamiento). Además, se facilitan enlaces a diferentes páginas que contienen material de interés didáctico. No se propone la idea de utilizar CHILDES como herramienta para la Enseñanza de Lengua Extranjera.

Procedimientos especiales

Los procedimientos especiales incluyen el uso de CLAN y CHAT para el análisis conversacional, o cómo trabajar con vídeo o audio digitalizado.

Morfología y Lexicón

CHILDES proporciona el programa MOR para etiquetar de forma automática los corpus en formato CHAT, después de haber construido una gramática para la lengua específica que se vaya a tratar. Tras el etiquetado morfológico, existe la posibilidad de usar otro programa, POST, para desambiguar, aunque sólo está disponible para la lengua inglesa; para el resto de lenguas, el usuario debe realizar el entrenamiento de una base de datos POST. Además, se facilitan los enlaces a la MRC *Psycholonguistic Database* y al programa MEGRASP de análisis sintáctico.

3.3.3 El corpus español

La tabla 3.1 muestra la base de datos en español recopilada en CHILDES hasta el año 2007. En ella se proporcionan los datos relativos al nombre del subcorpus, el rango de edad de los participantes, el número de los mismos y breves comentarios sobre el contenido del subcorpus. En ninguno de los casos se facilita ningún dato relevante sobre la longitud del corpus, ya sea por horas de grabación o por número de palabras.

De los 17 subcorpus, sólo dos, *Irene* y *López Ornat*, cuentan con su versión en audio y vídeo, es decir, que son los únicos que dan la posibilidad de escuchar y visualizar la grabación original. Ambos son corpus longitudinales, es decir, con un único participante, y las grabaciones comprenden un periodo de tiempo que finaliza antes de que el participante alcance los cuatro años de edad.

Teniendo en cuenta que los subcorpus son independientes, es decir, que no han seguido un mismo patrón de diseño (éste se habrá adaptado a las necesidades y finalidad del investigador o grupo de investigadores), sería de gran utilidad que la descripción que se incluye en *Comments* (Comentarios) fuese más amplia y detallada. En especial en los casos de corpus transversales, en los que participa más de un sujeto; por ejemplo, en el caso del segundo subcorpus, *Becacesn*, sabemos que han participado 81 sujetos de entre 4 y 12 años en sesiones de una hora. Pero no sabemos cómo están distribuidos dichos participantes, si la distribución se ha hecho intentando mantener un equilibrio de edades (mismo número de sujetos por cada edad) o si se ha tenido en cuenta el sexo de los participantes; tampoco tenemos datos sobre la participación de adultos o sobre si la interacción se produjo en una situación familiar (en casa) o formal (en el colegio).

Por otra parte, aunque la principal característica del subcorpus es la lengua española, en el mismo se incluyen situaciones de bilingüismo (*Koiné*) o variantes hispanoamericanas (*Hess* o *Shiro*). El diseño del corpus, en este tipo de casos, carece de sistematicidad. Lo ideal sería incrementar

los datos diferenciando por países, y dentro de cada país, por zonas geográficas.

Corpus	Edad	N	Comentarios
Aguirre	1;11-2;11	1	Sin documentación
Becacesn	4-12	81	Sesiones de una hora
ColMex	6-8	90	Niños en edad escolar
Diez-Itza	3;0-3;11	20	Primeros 10 niños y 10 niñas de un corpus mayor en construcción
FernAguado	3;0, 3;6, 4;0	50	Longitudinal a corto plazo con tres situaciones de juego
Hess	6-9	24	Niños de colegios públicos y privados en México, en cinco tareas narrativas
Irene	0;11-3;2	1	Niño español de Asturias
JacksonThal	0;10-3;0	202	Corpus transversal de Queretaro, San Diego, y Santa Barbara
Koiné	1;6-4;5	71	Conversaciones de niños en preescolar en Galicia, la mayoría en castellano y algunas en gallego
Linaza	2-4	1	Estudio longitudinal de un niño aprendiendo español
Lopez Ornat	1;7-3;11	1	Estudio longitudinal de un niño aprendiendo español
Marrero/ Albala	1;6-8;0	6	Estudio longitudinal de niños españoles de Canarias, con 12 archivos
Montes	1;7-2;11	1	Estudio longitudinal de interacciones padres-hijo
Romero	2;0	1	Estudio transversal de un niño de dos años
Serra / Sole	1;4-3;10	1	Estudio longitudinal de un niño monolingüe español en Cataluña
Shiro	Primer y cuarto grado	111	Narraciones de niños venezolanos en cuatro tareas
Vila	0;11-4;8	1	Corpus longitudinal con 35 grabaciones

Tabla 3.1 El corpus español de CHILDES

3.4 Análisis del método CHILDES

El presente proyecto pretende ser una continuación de otro anterior, "Diseño de corpus orales de lenguaje infantil: análisis comparado de CHILDES y las nuevas herramientas del Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid", comunicación presentada por Antonio Moreno Sandoval y Miguel Pérez Milans en el XXXIV Simposio de la Sociedad Española de Lingüística. Dicho proyecto partía de una comparación exhaustiva de CHILDES con el trabajo en corpus orales desarrollado hasta el momento en el LLI-UAM, con la finalidad de elaborar un corpus oral de la variedad lingüística infantil, el Corpus de Habla Infantil Espontánea del Español (CHIEDE). La comparación entre CHILDES y la metodología empleada en el LLI-UAM produjo las conclusiones expuestas en la tabla 3.2.

		CHILDES	CHIEDE
Metodología	Distribución equilibrada y representativa del corpus	No (para facilitar el intercambio)	Sí (para facilitar la representatividad)
	Fases de diseño	Variabilidad y flexibilidad (pluralidad)	Secuencia fija
Formato	Cabecera de transcripción	Características de participantes Heterogeneidad	Participantes y contexto de uso Sistematicidad
	Formato de transcripción	CHAT: distintas variantes conviviendo	CHAT: un único criterio (fidelidad a la lengua oral)

Tabla 3.2 Comparación CHILDES y CHIEDE

La distribución del corpus es uno de los aspectos más importantes que debe ser estudiado previamente a la realización del mismo. Sin embargo, este aspecto no se tiene en cuenta en CHILDES, en el que no existe ningún criterio organizador, mientras que en el LLI-UAM este es un elemento fundamental. Para el buen equilibrio de un corpus oral que pretenda ser de referencia ha de haber una organización basada en dos factores de variación del habla: las características de los hablantes y el contexto de uso. Esta organización permite el control de variables, necesario para garantizar la representatividad del corpus.

Antes de iniciar cualquier grabación, es obligatorio contar con los permisos firmados por todos los participantes, y es especialmente importante en los casos de grabaciones a menores. El LLI-UAM cuenta con las autorizaciones pertinentes para cada uno de los corpus desarrollados. Además, todas las grabaciones se realizan con grabadora digital que proporciona una calidad de sonido excelente. En CHILDES, las grabaciones digitales conviven con las analógicas.

La validación de la transcripción es otro aspecto importante en un corpus que pretenda ser fiable. La transcripción debe ir seguida de una validación posterior a cargo de un investigador diferente, quien a su vez remita el resultado al primero. Esta metodología consigue unificar formalmente todos los textos, aumentando la fiabilidad de manera espectacular. CHILDES está formado por las aportaciones de más de cien profesionales que contribuyen a la creación del corpus con sus grabaciones y transcripciones. Esto supone una gran ventaja para el incremento de la base de datos; pero, al mismo tiempo, obstaculiza el control que se pueda tener sobre los datos: sería muy difícil revisar todas las transcripciones para evitar errores tanto humanos

como los generados por los programas de tratamiento de texto. Además, se producen continuos cambios en los programas y en la base de datos: "Because of the continually changing shape of the programs and the database, keeping this manual up to date has been an ongoing activity. (MacWhinney, 2007:11). La cantidad de datos se antepone a la calidad de los mismos.

El alineamiento sistemático sonido-transcripción es un requisito obligatorio en el LLI-UAM, ya que permite al usuario verificar la fiabilidad de la versión transcrita. En CHILDES este alineamiento es opcional y raramente se da —al menos en las muestras en lengua española, que son las que a nosotros nos interesan.

En el LLI el etiquetado automático o anotación se realiza mediante un programa informático específico con posterior revisión manual por parte de un equipo de lingüistas. La anotación, no obstante, es opcional en CHILDES y la desambigüación corre a cargo de particulares.

En resumen, algunas de las mejoras de CHIEDE con respecto a CHILDES son la autorización por escrito de los padres de los participantes y su formato unificado y fijo. La siguiente tabla muestra dichas mejoras:

Fases diseño	CHILDES	CHIEDE
Autorizaciones	¿?	Homogéneas y detalladas
Grabaciones	Audio analógico / digital / vídeo /ninguna	Todo grabaciones en audio digital
Transcripción	Sí	Sí
Revisión	No	Sí
Alineamiento	Opcional	Siempre
Anotación	Opcional	Siempre
Desambiguación	A cargo de particulares	Equipo de lingüistas

Tabla 3.3 Fases del diseño en CHILDES y CHIEDE

Como se ha mencionado con anterioridad, los corpus que podemos encontrar en CHILDES provienen de colaboraciones desinteresadas de diferentes profesionales que trabajan de forma unilateral, lo que no garantiza una sistematicidad en todos los corpus. A pesar de que la página de CHILDES facilita manuales que explican de qué manera hay que proceder para realizar un corpus, cada investigador o grupo de investigadores poseen una metodología propia. Este sistema de construcción de corpus supone anteponer la cantidad[9] a la calidad: garantiza un gran corpus, pero la

[9] La lengua mayoritaria en CHILDES es el inglés; las transcripciones en otras lenguas son minoritarias y por lo tanto tampoco encontramos un gran corpus en una lengua como el castellano.

fiabilidad y representatividad se pueden poner en duda. A continuación, vamos a exponer algunas de las desventajas del sistema CHILDES. Al consultar la página en busca de los corpus de lenguaje español, nos hemos encontrado con los siguientes inconvenientes:

- Codificación del texto: las transcripciones en CHILDES quedan codificadas en formato UTF-8. Este hecho se especifica en el manual disponible en la página web del corpus. Dicho manual puede resultar útil para aquellos que quieran colgar en la página sus propios corpus, ya que deben seguir las indicaciones en él facilitadas. No obstante, si un investigador sólo desea tener acceso directo a las transcripciones para extraer información, puede no reparar en el manual o simplemente no quiere perder el tiempo en ello. De esta forma, si el investigador no sabe que el formato adecuado para ver las transcripciones es UTF-8 puede encontrarse con lo siguiente:

```
*CHI: oy [!] oy, Ã©ste es chique [: chicle] [*] [= la
bola] .
%mor: co|oy co|oy pro:dem|Ã©ste=the_latter vpres|se-
3S&PRES=be n|chicle&MASC=chewing_gum.
*MJO: quieres que juguemos con la granja ¿
%mor: vpres|quere-2S&PRES=want rel|que=that vsub|juga-
1P&SUB&PRES=play prep|con=with det:art|el-
FEM=then|granja&FEM=grange ?
*CHI: <yo fubo [: subo] [*] bo [: por] [*] aquÃ-> [x 5]
.
%mor: pro:per|yo=I vpres|subi-1S&PRES=go_up
prep|por=for adv|aquÃ-=here .
%act: subiÃ©ndose al respaldo del sofÃ¡ .
*MJO: subes por ahÃ- ?
%mor: vpres|subi-2S&PRES=go_up prep|por=for
adv|ahÃ-=there ?

                              [Marrero / Albalá, alf9105.cha]
```

Al abrir cualquiera de los archivos que contienen las transcripciones con Microsoft Word u otro editor de texto, nos encontraríamos con el problema, ya que estos editores tienen predeterminado un tipo concreto de codificación. En la mayoría de los editores de texto tenemos una opción que nos permite elegir el tipo de codificación, pero debemos saber que la elección correcta es UTF-8.

- Características de los corpus: dentro de la clasificación en diferentes lenguas que encontramos en la página de CHILDES, encontramos la categoría Romance, y dentro de esta el subtipo Spanish. Una vez entramos en el apartado de español, encontramos un listado de títulos, cada uno correspondiente a un investigador o investigadores distintos. Sin embargo, son pocos lo datos que se nos facilitan a priori sobre el contenido de cada archivo; solamente podemos encontrar la tabla 3.1 que

hemos reproducido en el apartado anterior, dentro de la sección dedicada a los manuales. En ella, tenemos información sobre las edades de los participantes, el número de grabaciones o transcripciones, y una breve reseña sobre el contenido de las mismas. No obstante, no se aportan datos referentes al número de palabras —forma más habitual de "medir" los corpus—, ni a las horas de grabación; tendremos que descargar todos los archivos e ir abriendo uno por uno para conocer la información que contienen. Por otra parte, en ninguno de los corpus están alineados transcripción y sonido, y solamente dos de ellos aportan el video de la grabación, donde podemos escuchar de primera mano la secuencia original.

• Análisis morfosintáctico: CHILDES facilita un *PoS tagging* o programa de análisis morfosintáctico, MOR, que anota automáticamente el texto transcrito. Para ello, "it is necessary to construct a separate MOR grammar for each language" (http://childes.psy.cmu.edu/morgrams/). Así, en la página se facilita una versión para el castellano, basada en un diccionario o listado de palabras en español con su correspondiente traducción inglesa. De esta forma, el análisis morfosintáctico se realiza sobre las plantillas de una gramática diseñada para el inglés, buscando la palabra española correspondiente. El resultado final del análisis morfosintáctico es el siguiente:

```
*MJO: venga, Alfon, vamos a jugar un poco con la
granja, vale ¿
%mor: vimp|venga-2S&IMP=avenge n:prop|Alfon vpres|i-
1P&PRES=go prep|a=to vinf|juga-INF=play
det:art|un&MASC=one adv|poco=few prep|con=with
det:art|el-FEM=the n|granja&FEM=grange
co|okay ¿
%act: sacando la granja para jugar .
*CHI: hay &an [//] hay uno [*] gola [: bola] [*] ¡
%mor: vpres|habe-3S&PRES&SPEC=have vpres|habe-
3S&PRES&SPEC=have det:art|un-MASC=one
[*] n|bola&FEM=ball ¡
%exp: refiriéndose a una bola de chicle que había en la
caja de la
granja .
*MJO: jugamos con la granja ¿
%mor: vpres|juga-1P&PRES=play prep|con=with det:art|el-
FEM=the n|granja&FEM=grange
¿
                          [Marrero / Albalá, alf9105.cha]
```

Como vemos, después de cada intervención de un participante, se añade una línea de texto identificada con la secuencia **%mor**, donde se realiza el análisis morfosintáctico. Al utilizar el sistema explicado anteriormente, ocurren cosas como esta,

```
*MJO: venga, Alfon, vamos a jugar un poco con la
granja, vale ¿

%mor: vimp|venga-2S&IMP=avenge n:□res|Alfon vpres|i-
1P&PRES=go □rest|a=to vinf|juga-INF=play
det:art|un&MASC=one adv|poco=few □rest|con=with
det:art|el-FEM=the n|granja&FEM=grange co|okay ¿
```

donde la palabra "venga" se ha analizado como verbo en modo imperativo, segunda persona de singular del verbo "vengar"; o la locución adverbial "un poco" que se analiza por separado y el artículo indefinido "un" se considera como traducción del numeral inglés "one". O, por último, la traducción de "granja" como "grange" en lugar de "farm".

- Revisiones: esta es una de las partes más importantes para que un corpus pueda resultar fiable. Puesto que actualmente ningún etiquetador morfosintáctico cuenta con un nivel de acierto del cien por cien, es necesaria una revisión manual tras la anotación automática. CHILDES facilita "a POST disambiguation database for English, but for other languages, users will need to do the work of training a POST database for themselves" (http://childes.psy.cmu.edu/morgrams/). Esto significa que para el corpus español, los investigadores deben realizar la desambiguación porque el programa POST no es válido para nuestra lengua. Si dicha revisión no se produce, nos encontramos con los problemas expuestos arriba. Por otra parte, la necesidad de desambiguación manual se observa también en las transcripciones en lengua inglesa, a pesar de que el programa POST esté diseñado para este idioma:

```
*MOT: well you see that's dual and it runs and he won't
buy it he says
 <no I'll buy coal> ["] .
%mor: co|well pro|you v|see pro:dem|that~v|be&3S
adj|dual conj:coo|and pro|it
 n|run-PL conj:coo|and pro|he v:aux|will~neg|not v|buy
pro|it pro|he v|say-3S
 n|quote .
*MOT: I haven't a pick of oil out there so I haven't .
%mor: pro|I v|have~neg|not det|a n|pick prep|of n|oil
prep|out adv:loc|there
 co|so pro|I v|have~neg|not .
*INV: well I wouldn't worry .
%mor: co|well pro|I v:aux|would~neg|not v|worry .
*MOT: he says +…
%mor: pro|he v|say-3S +…
*INV: does that run your central heating as well ?
%mor: v:aux|do&3S det|that n|run pro:poss:det|your
adj|central n|heating prep|as
 n|well ?
```

[Belfast, barb08.cha]

> En este pequeño fragmento de una de las transcripciones que podemos encontrar en lengua inglesa observamos un error en el análisis morfosintáctico debido a la falta de una posterior revisión manual. El verbo "run" se ha entendido en los dos casos señalados en negrita como un sustantivo. En el primer caso, puede haberse debido al hecho de que esté morfológicamente flexionado con el sufijo de tercera persona (-s) y se ha entendido como plural. En el segundo caso, la existencia de un determinante ante una palabra que puede tener las categorías de sustantivo y verbo ha hecho que se opte por categorizarlo como sustantivo.

Según lo expuesto, podemos afirmar que la fiabilidad y representatividad de la base de datos de CHILDES es relativa. Todos estos problemas son comunes a cualquier corpus, pero por ello precisamente es necesaria la constante revisión de cada resultado que se obtenga para corregir este tipo de errores. Además, una buena planificación del corpus con anterioridad a la realización del mismo evita problemas futuros como los derivados de la codificación del texto. De la misma forma, aportar una ficha con una descripción detallada del contenido de las transcripciones —número de palabras, características de los participantes, contexto de uso, entre otras— facilitaría enormemente el trabajo. Todos estos factores se han tenido en cuenta durante el diseño y la realización de nuestro corpus de lenguaje infantil, CHIEDE.

El sistema CHILDES puede ser muy útil como herramienta para realizar estudios propios al completo, es decir, para el investigador que quiera crear su propio corpus y extraer datos sobre fenómenos concretos del mismo. No obstante, dicha tarea requiere de un nivel de manejo informático bastante elevado, además de tiempo y paciencia para leer y entender el funcionamiento de los programas (sin mencionar las convenciones de transcripción). En muchos casos, resultaría más rentable y práctico que los datos se facilitasen directamente, es decir, que el usuario tenga acceso al texto de la transcripción, al igual que a su correspondiente sonido alineado, pero que ciertos datos sobre número de palabras, listas de frecuencia o cualquier otro dato estadístico se incluyan de manera sistemática en cada archivo. De esta forma, el investigador tiene acceso a los datos y puede dedicarse a realizar su estudio, probar su hipótesis, o realizar su propia interpretación de los datos sin necesidad de realizar todo el trabajo inicial de extracción de los mismos.

El aprendizaje de CLAN requiere bastante dedicación, y el tiempo empleado para dicha tarea dependerá de las habilidades personales del investigador, al igual que de sus conocimientos previos sobre transcripción y posterior tratamiento informático de corpus. Los manuales facilitados tanto para la transcripción de material audiovisual como para su posterior tratamiento informático son muy completos, pero al mismo tiempo muy

extensos; esto significa que el uso de la base de datos CHILDES puede resultar demasiado costoso en cuanto a tiempo y dificultad, sobre todo si lo que se quiere es conocer datos muy concretos sobre las frecuencias de diferentes elementos, como pueden ser morfemas, categorías gramaticales o fonemas. En este sentido, puede resultar más práctico para el investigador que se le facilite ese tipo de información en lugar de que tenga que ser él mismo el que la extraiga del corpus. Si una parte de la investigación va a estar conformada por el proceso de extracción de datos de las transcripciones, entonces es necesario aprender a utilizar los programas pertinentes. Sin embargo, si dicho proceso no es relevante para el estudio, sería mucho más útil poder disponer de los datos de forma directa.

4. Experiencia en C-ORAL-ROM

4.1 El proyecto C-ORAL-ROM

C-ORAL-ROM es un corpus de habla espontánea recopilado en cuatro lenguas romances: francés, español, italiano y portugués. El proyecto fue financiado por la Comisión Europea y se desarrolló desde enero de 2001 hasta marzo de 2004. El trabajo se llevó a cabo desde cuatro universidades, la de Florencia, Provenza, Lisboa y la Autónoma de Madrid, trabajando cada una de ellas en la parte del corpus correspondiente a su lengua.

El corpus consta de un total de 772 textos orales, con más de 120 horas de grabación de habla espontánea, distribuidas equivalentemente en las cuatro lenguas con una extensión aproximada de unas 300.000 palabras en cada idioma. Cada texto incluye la siguiente información (Moreno Sandoval y Urresti, 2005):

- Metadatos de la grabación: características sociolingüísticas, contexto situacional, calidad acústica y contenido del texto.
- Transcripción ortográfica: siguiendo el formato CHAT.
- Alineamiento: sincronización del sonido y su correspondiente transcripción.
- Anotación: información morfosintáctica de cada palabra.

El principal obstáculo para el proyecto fue la necesidad de encontrar soluciones para tres problemas básicos: metodológicos (conseguir un diseño y formato común para los cuatro grupos de investigación), legales (obtener la autorización por escrito de todos los participantes para la publicación del corpus) y tecnológicos (aprovechamiento de todos los recursos que las nuevas tecnologías aportan). Una vez superadas estas limitaciones iniciales, los aspectos más destacados del corpus son los siguientes:

- Distribución del corpus: el elemento indispensable de un corpus es la representatividad de los datos. Las grabaciones en cada una de las lenguas de C-ORAL-ROM se realizaron partiendo de una primera diferenciación entre registro formal e informal. Con respecto al primero de ellos, se llevó a cabo una subclasificación en contextos naturales, medios de comunicación y conversaciones telefónicas. El registro informal se divide a su vez en dos grandes grupos: familiar o privado y público, que a su vez distinguen monólogos, diálogos y conversaciones. Además, se tuvo en cuenta que el número de participantes de ambos sexos fuera lo más equilibrado posible.

- Formato común: la transcripción de los datos se basa fundamentalmente en el modelo CHAT (Codes for the Human Analysis of Transcripts) y como lenguaje de marcación se ha empleado XML por su versatilidad.
- Validación de los datos: la fiabilidad del producto como recurso lingüístico es indispensable. C-ORAL-ROM se ha sometido a una validación interna (revisión en cuanto a contenido y forma, tanto de manera automática como manual, a cargo de un equipo de lingüistas) y a otra externa (por parte de una equipo independiente a la investigación de la empresa LOQUENDO).
- La anotación morfosintáctica: cada palabra del corpus está anotada con su categoría y lema, además de la información morfológica (género, número, tiempo, etc.). Dicha anotación se ha realizado automáticamente con programas informáticos específicos, y con una posterior revisión manual de gran parte del corpus.

En cuanto a su potencial y aplicabilidad, son muchos los campos científicos para los que una herramienta de estas características puede resultar de gran utilidad: la investigación lingüística, las investigaciones en pragmática y análisis del discurso, las tecnologías lingüísticas, la traducción, la enseñanza de lenguas extranjeras (ELE), o el campo de la edición y la publicación.

4.2 El LLI-UAM

El Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid (LLI-UAM) es pionero en la construcción de corpus orales de español. Desde su creación, se han recogido dos grandes corpus: Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC), a principios de los 90, y recientemente, Corpus Oral de las Lenguas Romances (C-ORAL-ROM). Ambos corpus suman unas 1.400.000 palabras transcritas y comparten múltiples aspectos. El principal objetivo de ambos es el registro fiel y preciso del lenguaje oral contemporáneo. Sin embargo, C-ORAL-ROM ha podido aprovechar las innovaciones técnicas y metodológicas, lo que lo convierte en un corpus de segunda generación.

En ambos casos se trata de corpus de grabaciones de habla espontánea que han sido transcritas ortográficamente. La característica más sobresaliente es la espontaneidad, es decir, no se trata de la lectura en voz alta de textos escritos, sino de manifestaciones directas de la actuación lingüística, en diferentes contextos y situaciones comunicativas. En ambos casos, cada texto transcrito va acompañado de una cabecera (o sección de metadatos) donde se especifica todo tipo de información relevante, desde la clasificación tipológica del texto hasta los rasgos sociolingüísticos de los hablantes (sexo, edad, educación, profesión, origen geográfico). También se incluyen datos sobre el número de palabras, la duración y los responsables

de la trascripción y de la revisión. Toda esta información es esencial para situar el texto y establecer correlaciones entre las variables lingüísticas y socio-contextuales.

Sin embargo, CORLEC y C-ORAL-ROM tienen importantes diferencias entre sí, debidas a la distancia de diez años entre la realización de uno y otro. CORLEC fue el primer corpus oral del español y uno de los primeros del mundo. C-ORAL-ROM se beneficia de una experiencia de más de diez años recopilando textos orales y sobre todo de los avances tecnológicos y de una financiación mucho mayor. En esencia, las diferencias son:

- **Monolingüe / Multilingüe**: C-ORAL-ROM es un corpus comparable de grabaciones espontáneas en español, francés, italiano y portugués. Parte de la experiencia individual en realización de corpus monolingües de cada uno de los equipos (UAM, Aix-en-Provence, Florencia y Lisboa) y para conseguir el objetivo de la comparabilidad, los diferentes equipos tuvieron que acordar un diseño (es decir, la proporción de los textos) y un formato de transcripción comunes. Además, se emplearon las mismas herramientas informáticas para el alineamiento del texto y del sonido y se adoptó el mismo conjunto de metadatos. Finalmente, se anotaron morfosintácticamente los cuatro corpus, de manera que el resultado es un corpus multimedia (multimodal) anotado en diferentes niveles y en cuatro lenguas.
- **Tecnología analógica / tecnología digital**: CORLEC se realizó con grabadoras analógicas portátiles con el micrófono incorporado. La calidad para su época era suficiente, ya que lo que se exigía era que se pudiera reproducir en un magnetófono analógico para realizar su transcripción. Nunca se pensó en la utilización directa de las cintas, hasta que un grupo de investigación en reconocimiento de señal solicitó los originales y se digitalizaron. Sin embargo, la calidad de sonido era tan mala que sólo una grabación de CORLEC se reutilizó en C-ORAL-ROM. En el segundo proyecto sólo se ha empleado tecnología digital: desde la grabación con una DAT y micrófonos profesionales omnidireccionales, pasando por la transferencia al ordenador y la manipulación con programas profesionales que no deterioran el original. La calidad es muy buena y en algunos casos excelente. Todos los pasos posteriores (transcripción, alineamiento y anotación) se han realizado con programas especiales. Finalmente, el corpus se ha presentado en una versión multimedia en formato DVD.
- **Sin autorización / con autorización**: cuando se compiló CORLEC no existía legislación sobre la recogida de grabaciones y su posterior publicación. Sin embargo, a mediados de los 90 se dictaron dos leyes que afectan directamente a este tipo de adquisición de datos. Por una parte, está el derecho a la intimidad, que protege a los hablantes de ser grabados sin su consentimiento y de reproducir sus conversaciones privadas. Por otra, están los derechos de propiedad intelectual en aquellas emisiones

públicas que implican una elaboración de contenidos que tienen una autoría. Nos estamos refiriendo a las grabaciones de los medios de comunicación y a las conferencias, clases magistrales en la universidad, etc. En ambos casos, se necesita un permiso por escrito de los participantes o de los propietarios de los derechos de autor para poder grabar, transcribir, publicar y comercializar las emisiones orales. Todos los textos de C-ORAL-ROM, por exigencias de la Unión Europea, cuentan con las autorizaciones para ser publicados.

El principal eje de la metodología es la distribución del corpus; se considera que existen dos elementos básicos que influyen en la variación: las características de los participantes y el contexto de uso. Estos dos factores guían el diseño del corpus.

Una vez que el corpus está organizado, el procedimiento metodológico es el siguiente: grabación, transcripción, revisión, alineamiento y etiquetado. Previamente a la fase de grabación, es un requisito imprescindible contar con el permiso de los participantes. La grabación se lleva a cabo con una grabadora digital (DAT), que garantiza una alta calidad de sonido. La transcripción sigue el formato CHAT y como lenguaje de marcación se usa XML debido a su versatilidad y facilidad de conversión a otros formatos de texto. Cada transcripción es revisada por un lingüista diferente, quien de nuevo reporta el resultado al anterior. Todo este proceso conduce a un trabajo final unificado y fiable.

El siguiente paso es el alineamiento de sonido y texto, que permite al usuario comprobar y revisar la transcripción. La fase de etiquetado tiene en cuenta todos los niveles del lenguaje, desde la prosodia a la morfosintaxis. El formato final de todos los archivos es la cabecera —con información sociolingüística sobre los participantes y la situación comunicativa— seguida de la transcripción.

La anotación lingüística es el aspecto más importante para el análisis y recuperación de información contenida en un corpus. La anotación manual es inviable para un corpus de un cierto tamaño si el resultado perseguido debe ser exhaustivo y completo. Sólo la anotación automática, seguida de revisión humana, garantiza exhaustividad y precisión. Para este propósito, contamos con programas informáticos (GRAMPAL y un transcriptor fonológico) que han sido desarrollados para anotar corpus. El lenguaje oral espontáneo tiene sus propios rasgos: es más innovador y no está sujeto al canon escrito. Estos rasgos son más evidentes en el vocabulario y las estructuras sintácticas. Por lo tanto, las herramientas deben incluir unidades de morfología derivativa y análisis sintáctico superficial. Además, un sistema semiautomático de transcripción fonológica se implementará a partir de la transliteración ortográfica.

4.3 Adaptación de la metodología

La construcción del corpus oral infantil CHIEDE se ha basado en la metodología empleada para la elaboración de C-ORAL-ROM y ha aprovechado la experiencia y los recursos del LLI-UAM. No obstante, durante el proceso de transcripción de las grabaciones, nos dimos cuenta de que era necesario realizar algunos cambios en el formato, para garantizar una mejor comprensión de los textos y mantener una mayor fidelidad al discurso oral. Es cierto que C-ORAL.ROM se concibió como un corpus multimedia en el que se puede acceder al sonido y al texto al mismo tiempo por medio del alineamiento de ambos. Sin embargo, y aunque los textos en CHIEDE también están alineados con su sonido, consideramos que en ciertos momentos puede ser más útil trabajar sólo con el texto, y para ello creemos que pueden ser necesarios una serie de cambios en las convenciones de transcripción que faciliten la lectura.

En primer lugar, tenemos que partir del hecho de que con la transcripción ortográfica no ocurre lo mismo que con la fonética, para la que existe un sistema de codificación universal, el AFI. Para la transcripción de textos orales existen distintos formatos que atienden a elementos diversos, dependiendo de la finalidad del corpus. Los hay que siguen las reglas de puntuación convencionales de la lengua escrita; también podemos encontrar sistemas de anotación que se centran en los rasgos suprasegmentales como INTSINT o ToBI. Pero a la hora de transcribir la lengua oral, lo más común es establecer convenciones siguiendo un criterio pragmático y prosódico y no tanto sintáctico. De esta forma, y dependiendo del grado de precisión o fidelidad a la lengua oral, se establecen distintos formatos de transcripción. Payrató (1994) establece tres niveles de transcripción dependiendo de la precisión con que se quieran reflejar las características propias de la lengua oral. Por lo tanto, la carencia de un sistema estándar y la finalidad de nuestro corpus nos llevarán a seleccionar los elementos que queremos reflejar.

En nuestro caso, uno de los intereses principales es la anotación morfosintáctica de los textos recopilados. Dicha anotación se realiza de automáticamente mediante un programa informático cuyo resultado es validado posteriormente de forma manual. De esta manera, se pueden extraer datos estadísticos de los niveles morfológico y sintáctico. También, mediante un proceso de lematización podremos realizar estudios en el campo de la semántica léxica. Pero donde hemos encontrado mayores problemas es en el nivel pragmático y suprasegmental, ya que consideramos que el formato de transcripción empleado para C-ORAL-ROM descuida estos niveles en algunos casos. Es cierto que, como ya hemos comentado, el corpus multilingüe se concibió de forma que el usuario tuviera acceso al sonido original de la grabación y de esta forma poder observar directamente elementos prosódicos o paralingüísticos. Aún así, consideramos que ciertos cambios en el formato de transcripción serían necesarios.

La primera de las cuestiones afecta al sistema de alineamiento en C-ORAL-ROM y los criterios empleados para dicho proceso. El alineamiento de un corpus oral suele hacerse o bien por turnos dialógicos o bien por enunciados. Esta última opción fue la elegida en C-ORAL-ROM. En el punto 1.2 de la obra *C-ORAL-ROM. Integrated Referente Corpora for Spoken Romance Languages* se plantea la problemática en la segmentación del lenguaje oral, para la que los criterios sintácticos propios de la lengua escrita no son los más adecuados. Así, la forma de segmentación atenderá a las rupturas prosódicas[10]. Pero el problema aparece cuando hay que decidir dónde se encuentran los límites de un enunciado y, por tanto, dónde establecer los puntos de alineamiento. Las pausas prosódicas de dividen en dos tipos, dependiendo de si son terminales o no terminales. Se considera que una pausa prosódica es terminal cuando el acto de habla de la emisión se ha llevado a cabo. De esta manera, "each string ending with a terminal break is considered an utterance" (Cresti y Moneglia, 2005:16). Además, en C-ORAL-ROM se entiende que un enunciado es "the minimal linguistic unit such that it allows a pragmatic interpretation in the world" (Cresti y Moneglia, 2005:16). Según esto, podemos establecer que un enunciado es la unidad mínima del discurso con significado pragmático, esto es, que constituye un acto de habla ilocutivo y que además está delimitada por pausas, siendo la pausa prosódica final de carácter terminal.

A esto, debemos añadir que las pausas prosódicas terminales se representan en el texto con cuatro símbolos, dependiendo de la entonación y de la función pragmática del enunciado. Así, tenemos doble barra (//) para la pausa prosódica con valor declarativo; signo de interrogación (?) para la pausa prosódica con valor interrogativo; tres puntos (…) para la entonación suspendida de forma intencionada; y el signo de suma (+) para los enunciados interrumpidos. Ninguno de los tres últimos símbolos representa una dificultad especial; no obstante, el uso de la doble barra (//) también se relaciona con criterios de tiempo de cada pausa. Generalmente, se utilizan dos símbolos para marcar las pausas y segmentar así el discurso: la barra simple (/) para pausas breves, a veces casi imperceptibles; y la doble barra para pausas más marcadas. Y se suele establecer que las pausas breves coinciden con pausas prosódicas no terminales mientras que las pausas prosódicas terminales son más perceptibles. De esta forma, todo final de enunciado afirmativo estaría marcado con el símbolo //. De hecho, para el alineamiento del texto y el audio en C-ORAL-ROM las marcas de tiempo se situaron en relación a los cuatro símbolos de pausa prosódica terminal comentados con anterioridad. Es decir, que en el lugar de cada uno de estos símbolos se colocaba una marca de tiempo que delimitaba el enunciado y lo alineaba con su sonido original. Sin embargo, no siempre se puede apreciar

[10] El término original en inglés es *prosodic break*, que nosotros hemos decidido traducir por ruptura prosódica.

una pausa marcada a final de enunciado, y por el contrario sí pueden apreciarse pausas prolongadas dentro de un mismo enunciado.

En resumen, lo que proponemos es establecer símbolos de transcripción diferentes para marcar las pausas y el final de enunciado. La barra simple se emplearía para pausas breves y la doble para pausas más marcadas en las que el hablante no ha concluido su proyecto informativo, mientras que para establecer el fin de un enunciado proponemos el símbolo ///. La triple barra marcará los límites entre enunciados afirmativos, independientemente de que dentro de uno de ellos podamos apreciar una ruptura en el flujo del discurso. El siguiente ejemplo extraído de CHIEDE muestra de forma clara el uso de la doble barra (//).

*TEA: ¡ups! todavía te queda mucho // para los seis ///

[INE5]

Después de pronunciar la palabra "mucho", el hablante hace una pausa descendente con la cual su enunciado se completa. Sin embargo, decide continuar y explicar o detallar su declaración. Si considerásemos la pausa con entonación descendente como fin de enunciado, la secuencia "para los seis" quedaría aislada de la primera parte y formaría un enunciado por sí sola. En casos como este, en el que el hablante prolonga su proyecto informativo con una aclaración, utilizaremos la doble barra y no la triple.

Cresti y Moneglia establecen que "what follows a terminal break necessarily belongs to a different utterance, and therefore to a different pragmatic and linguistic domain" (Cresti y Moneglia, 2005:17). Teniendo en cuenta que en el discurso oral espontáneo la falta de programación del mismo hace que en ocasiones un enunciado se vea interrumpido por una pausa marcada y terminal, consideramos que sería necesario establecer un símbolo específico (///) para marcar el final de un enunciado. Este símbolo sustituiría a la doble barra en las oraciones afirmativas para asegurar una segmentación por enunciados más fiable.

Otro cambio que hemos decidido realizar en la transcripción de nuestro corpus es la introducción del signo interrogativo de apertura (¿). En C-ORAL- ROM solamente se incluía el signo de cierre para indicar el final de la enunciación interrogativa. Es cierto que los inicios de pregunta van marcados por una pausa (ya sea breve o larga) y que aunque en ocasiones sea casi imperceptible siempre podemos advertir una variación en la entonación suficiente para incluir el símbolo /. Pero nos podemos encontrar con una pregunta que, ya sea por su extensión o por cualquier otro tipo de causa, en su interior contenga pausas. En este caso, esa pausa se marcaría con una barra simple, lo que daría pie a pensar que esa pausa interna marca el inicio del enunciado interrogativo. Para evitar esto, hemos decidido

incluir también el signo interrogativo de apertura que evite errores a la hora de marcar los límites de una pregunta.

*TEA: ¡eh! ¿ cómo se escribe / veintiocho ?

[A5-02]

En el ejemplo, la entonación interrogativa no se limita a la palabra "veintiocho", como podría inducir a pensar la inexistencia del signo de interrogación de apertura.

Por otra parte, el uso del signo de exclamación quedaba limitado en C-ORAL-ROM a las interjecciones. No obstante, nos encontramos con una entonación exclamativa en muchas otras ocasiones y creemos que es necesario incorporarlo. En el caso de las interjecciones, los signos de exclamación se transcriben sin dejar espacio entre ellos y la palabra, ya que es de esta forma como se incluyen en el diccionario. En el caso del resto de enunciados exclamativos, los signos de exclamación se transcribirán dejando un espacio entre estos y las palabras adyacentes, al igual que en el caso de los signos de interrogación. Y por el mismo motivo que consideramos necesario marcar los enunciados exclamativos, creemos que debe hacerse de la misma manera con las interrogaciones exclamativas, que se marcarán simultáneamente con los signos ¿¡ ¡?.Los siguientes ejemplos lo ilustran.

*TEA: ¡ah! ¡ pues cuéntame !
*JOR: pero primero lo de comprar / ¿ &eh ? vale /// me compraba mi papa y mi mama un &e [/] un cuento de [/] de Bugs Bunny ///

[JOR4]

*CAR: sí / la vamos a hacer / pero es que van a / venir / veinticinco niños ///
*TEA: ¿¡ tantos !?

[CAR5]

Por último, existen otros dos elementos que incluimos en CHIEDE y que no encontramos en C-ORAL-ROM:

- Alargamientos: en múltiples ocasiones se producen alargamientos de una vocal o una consonante en final de palabra. Esto es algo que podemos apreciar escuchando el sonido correspondiente a cada texto; aún así, consideramos que es necesario representarlo en la transcripción y para ello utilizaremos el símbolo →. En el ejemplo, el hablante duda ante la pregunta y alarga el sonido vocálico de la palabra "que" para pensar en su respuesta.

*TEA: ¿ y cuando faltan dos niños qué os dice Candi ?
*CAL: pues que -> / saltemos / hhh / cinco saltos / tres saltos ///

[CAL5]

• Articulación lenta de palabra: en algunas ocasiones el hablante articula una palabra haciendo una pausa después de cada sílaba, ya sea por una cuestión de énfasis o porque la esté leyendo. En este caso, incluiremos una línea dependiente después del turno del hablante introducida por *%alt* en la que escribiremos la palabra con barra simple indicando los lugares intersilábicos donde se ha producido la pausa. En el ejemplo, el participante identificado con VIC está leyendo, y la dificultad le hace pronunciar la palabra "Faustino" sílaba a sílaba.

*TEA: a ver Vicente // ayúdale un poquito a Arturo ///
*VIC: Faustino ///
%alt: (1) Fa/us/ti/no
*SEV: Faustino +
*TEA: Faustino ///

[A5-02]

En resumen, los cambios en las convenciones de transcripción en CHIEDE con respecto a C-ORAL-ROM son:

///	Delimitación de fin de enunciado; terminal
//	Pausa marcada; no terminal
¿?	Especificando el momento de apertura del enunciado interrogativo
¡!	Para enunciados exclamativos
¿¡!?	Para interrogativas exclamativas
→	Alargamientos de vocal o consonante final de palabra
x/x/x	Articulación lenta de palabra; se especifica en línea dependiente con %alt

Tabla 4.1 Nuevas convenciones de transcripción en CHIEDE

PARTE SEGUNDA

La aplicación

1. Diseño del corpus

1.1 Aspectos generales

MacWhinney (2007), en la introducción del manual de transcripción para CHILDES, hace la siguiente aclaración:

> Language acquisition research thrives on data collected from spontaneous interactions in naturally occurring situations. You can turn on a tape recorder or videotape, and, before you know it, you will have accumulated a library of dozens or even hundreds of hours of naturalistic interactions. But simply collecting data is only the beginning of a much larger task, because the process of transcribing and analyzing naturalistic samples is extremely time-consuming and often unreliable. (MacWhinney, 2007:5)

Crear un corpus no consiste en la simple tarea de poner la grabadora en funcionamiento y recoger una mayor o menor muestra de lengua. Como indica MacWhinney, eso es sólo el principio, y el trabajo duro y laborioso viene después, cuando hay que transcribir, anotar y analizar las muestras. Pero además del trabajo posterior a la recolección de muestras, también existe un trabajo previo de gran importancia: el diseño del corpus.

El diseño del corpus es la base sobre la que se asienta todo el trabajo; es un pilar básico para que el desarrollo posterior se pueda realizar sin problemas y para que los resultados sean óptimos. Por lo tanto, antes de recoger las muestras de un corpus, hay que establecer los criterios de diseño del mismo atendiendo a la finalidad de nuestra investigación. Es necesario realizar un plan de trabajo completo, en el que se tengan en cuenta los posibles inconvenientes que puedan surgir y las metas que se esperan alcanzar.

En el apartado 1.3 de la primera parte, dedicado a los conceptos básicos en LC, hicimos un breve repaso sobre las principales características de un corpus y los distintos tipos del mismo. Ya comentamos que los requisitos indispensables de un corpus moderno son su formato electrónico, la representatividad de su contenido, su disponibilidad para el público y su estructuración interna. Estos cuatro elementos se tuvieron en cuenta en la fase previa a la recolección de las muestras. Así, las grabaciones se realizarían digitalmente y, al igual que las transcripciones, se almacenarían en formato electrónico, lo que permitiría su consulta y tratamiento mediante un ordenador. En cuanto a la representatividad, teniendo en cuenta el carácter finito del corpus, era necesario realizar un diseño proporcionado y que priorizase la diversidad por encima de la cantidad. Su estructura interna se fundamenta en las variables que se han tenido en cuenta para garantizar la representatividad. El corpus estará disponible al público de forma gratuita a través de una página web.

Respecto a la tipología del corpus, es un corpus de lengua oral formado por las transcripciones realizadas a partir de grabaciones de habla espontánea. En este punto, habría que diferenciar entre dos tipos de corpus orales (Moreno Sandoval y Urresti, 2006):

- Corpus de lengua oral (spoken language corpora): transcripciones de producciones lingüísticas espontáneas.
- Corpus orales (speech corpora): el núcleo de la investigación es la señal sonora, y se caracterizan por su carácter controlado.

Así pues, nuestro corpus pertenecería al primer tipo, en el que las muestras recogidas son absolutamente naturales y por ello es necesario tener en cuenta ciertas características de un evento de habla espontáneo:

- Interacciones cara a cara multimodales: importancia del contexto.
- Referencia intersubjetiva a un espacio deíctico.
- Programación simultánea de la producción.
- Comportamiento lingüístico impredecible sujeto al contexto.
- Discontinuidad de la cadena hablada.
- Lenguaje corporal.
- Aspectos suprasegmentales.
- Aspectos pragmáticos.
- Características sociales y contextuales que influyen en el evento de habla.

Las características típicas de la lengua oral espontánea siempre han sido un problema para el transcriptor. Los rasgos prosódicos no pueden expresarse mediante la puntuación normativa establecida para la lengua escrita. Además, la comprensión del evento de habla no es completa si no se incluyen datos relativos al contexto social y situacional. Es debido a esta dificultad de codificar el lenguaje oral que no existe un sistema de transcripción estándar, ya que las convenciones de codificación que se establezcan dependerán de la finalidad del corpus. Más adelante, en la sección dedicada a la metodología seguida durante el proceso de transcripción, explicaremos con detalle los criterios fijados en CHIEDE.

También se mencionaron con anterioridad algunas de las utilidades de un corpus; en concreto, algunas de las ventajas de trabajar con corpus de lengua oral son:

- Poseer una muestra de habla con variables como las características de los participantes o el tipo de texto, y poder realizar estudios sobre ellas.
- Realizar teorías generales si el corpus es representativo.
- Naturalidad y espontaneidad del texto en contraposición a la artificialidad de una muestra creada.

- La anotación permite realizar análisis cuantitativos y relacionar distintos niveles de anotación.
- También es posible desarrollar estudios sobre la relación del léxico con las características de los participantes.

Otro de los puntos a tener en cuenta con anterioridad a la recogida de muestras fue el estado de la cuestión, es decir, la situación actual en España en lo que a corpus infantiles se refiere. Tradicionalmente, las muestras de habla infantil se han basado en estudios de corte longitudinal, para lo que se recogían muestras de un solo individuo a lo largo de sus primeros años de vida. En general, dichos estudios no se prolongaban más allá de los tres años, a pesar de que los expertos coinciden en aceptar que la adquisición lingüística no se puede considerar completa hasta los cinco o seis años de edad, y que el proceso de aprendizaje se puede extender hasta la pubertad. Por ello, consideramos que podría ser de gran utilidad la realización de un corpus transversal —con varios grupos de individuos clasificados por edad— en el que las edades de los sujetos abarcasen de los tres a los seis años de edad.

Quirk (1992) destaca los puntos clave en la planificación del British Nacional Corpus:

- Tamaño: medido en número de palabras.
- Disponibilidad: para otros investigadores.
- Equilibrio: representatividad.
- Software: herramientas de procesamiento de texto disponibles junto con el corpus.
- Estrategia: mejorar los recursos existentes.
- Estandarización: recomendaciones de la TEI; guías para la comprensión y manejo del corpus.

Durante la fase del diseño, se tuvieron en cuenta todos estos aspectos destacados por Quirk. Nos propusimos recoger un corpus de unas 60.000 palabras, cuya estructura garantizase la representatividad y fuese innovadora con respecto a los recursos existentes. El corpus estaría disponible mediante una página web para los investigadores interesados, con las guías necesarias para su comprensión y utilización.

Otro aspecto a tener en cuenta durante el proceso de diseño de un corpus es su utilidad posterior. En el caso de CHIEDE, sus posibles usos van desde el estudio en lexicografía (datos estadísticos sobre frecuencias de palabras o categorías), la investigación lingüística (ejemplos reales para probar teorías) o la tecnología lingüística (reconocimiento y síntesis de voz). La anotación morfosintáctica de los textos facilita todas estas tareas.

Así pues, la perspectiva general del programa de trabajo incluiría cinco operaciones principales:

- Diseño: tipo de textos que se van a grabar, tamaño y especificación de la codificación de las características de los textos.
- Permisos: es necesaria la autorización de todos los individuos que participen en el las interacciones.
- Recolección de datos y codificación de los mismos: establecer la metodología que se va a seguir.
- Almacenaje y distribución: en formato electrónico y de consulta gratuita a través de la red.
- Procesamiento del corpus: anotación morfosintáctica y transcripción fonética.

1.2 Dificultades previas y elección de los participantes

Como se ha mencionado ya, el primer paso en la elaboración de cualquier corpus, y siguiendo la metodología empleada en C-ORAL-ROM, es el diseño. Los dos principales factores que son relevantes en esta tarea son las características de los participantes y el contexto de uso, que deben ser prefijados con anterioridad al inicio de las grabaciones. Estos dos elementos variarán dependiendo del tipo de corpus que se vaya a realizar y la finalidad del mismo. En nuestro caso, el desarrollo de un corpus de habla infantil, era necesario establecer un criterio por edades y por el tipo de situación comunicativa.

En lo que respecta al primero de los elementos, las características de los participantes, el aspecto más importante es la edad de los mismos. Había que establecer de forma clara los límites del periodo del desarrollo lingüístico que podía ser más interesante para nuestro estudio. Teniendo en cuenta las diferentes teorías sobre adquisición y desarrollo del lenguaje comentadas con anterioridad, y la tabla 2.1 de la primera parte sobre la evolución lingüística del niño, se puede decir que la mayoría de los autores están de acuerdo en localizar el periodo más representativo del aprendizaje lingüístico entre los tres y los seis años —algunos autores extienden este periodo hasta la edad de siete años; sin embargo, otros consideran que el proceso se completa a la edad de cinco. Con anterioridad a los tres años de edad, el lenguaje del niño se encuentra en un periodo muy prematuro de desarrollo. Además, dentro de esa primera etapa es relevante el tipo de lenguaje que el adulto usa para comunicarse con el niño, es decir, el llamado *maternés*. Esto implicaría la realización de un corpus no sólo centrado en el lenguaje infantil, sino también a la variedad adulta que se usa para la comunicación con el niño. Por ello, nosotros hemos decidido establecer los límites de nuestro corpus desde el momento en que el niño comienza a dar muestras de un lenguaje suficientemente autónomo, con cierta semejanza ya al del adulto.

En cuanto a la edad máxima de los sujetos, se estableció el límite de los seis años, porque a esta edad el niño muestra una adquisición de la estructura lingüística prácticamente completa, cuyo perfeccionamiento se extenderá algunos años más, pero ya se puede considerar que el periodo básico de aprendizaje se ha completado.

El periodo de desarrollo lingüístico elegido para elaborar nuestro corpus, desde los tres a los seis años de edad, coincide con la primera etapa de escolarización de los niños españoles, Educación Infantil, que abarca el periodo desde los tres a los cinco años, aunque la mayoría de los estudiantes alcanzan los seis a lo largo del último curso. Para nuestro proyecto, contamos con la colaboración de un colegio público español que aprobó nuestra intención de grabar el habla de los niños de estos grupos de edades en Educación Infantil. Se concertó una reunión con los padres de los niños escolarizados en la que se les explicó en qué consistía el proyecto y cuál era su finalidad. En esta reunión se les pedía su colaboración y, por supuesto, una autorización por escrito de su consentimiento para grabar las voces de sus hijos.

Por cuestiones de privacidad y anonimato, hemos preferido no mencionar el nombre del colegio ni el municipio en el que se sitúa. Simplemente, aportaremos algunos datos de interés sociolingüístico: la grabación del corpus se llevó a cabo en un pueblo de Castilla la Mancha, en la provincia de Ciudad Real. Su población ronda los 7.000 habitantes, con algo más de 300 inmigrantes, provenientes en su mayoría del este de Europa —según el censo del año 2005. Su economía se basa en la agricultura y en la ganadería y, en menor escala, en la industria y el turismo. La variante lingüística del castellano, propia de la zona central de España.

En cuanto a la distribución del corpus, existió una primera idea sobre la misma que tuvo que sufrir algunos cambios, ya que pudimos confirmar que era difícil de llevar a la práctica. Es una tarea habitual dentro de la LC desarrollar un primer *corpus piloto* a modo de prueba, susceptible de sufrir modificaciones. De esta forma, si el diseño inicial no es factible —ya sea por problemas técnicos o por impedimentos impuestos por los participantes— es posible rectificar su estructura. En nuestro caso, el primer diseño consideraba la posibilidad de realizar las grabaciones de interacciones entre dos niños, sin la presencia de adultos, para obtener una muestra de espontaneidad y naturalidad. Sin embargo, después de varios intentos nos dimos cuenta de que los resultados no eran productivos debido a diferentes factores: en primer lugar, los niños de menor edad (especialmente aquellos con tres y cuatro años) no interaccionan como los adultos y tienden a presentar un comportamiento individualista que no ayuda al establecimiento de una conversación; en segundo lugar, la escena era artificial, ya que se les sacaba de su entorno natural (el aula) para evitar la grabación de ruidos o interrupciones por terceras partes, y esta situación

no ayudaba tampoco al establecimiento de una interacción. Este hecho supuso un gran inconveniente, ya que significaba reconsiderar el diseño inicial del corpus, además de perder la oportunidad de obtener muestras espontáneas de lenguaje entre niños. Sin embargo, encontramos otra posibilidad de obtener muestras similares grabando las *asambleas*, actividad realizada diariamente en clase con el objetivo de motivar la expresión oral de los niños. Es una actividad colectiva donde las producciones lingüísticas de los niños son espontáneas. Contando con esta posibilidad, decidimos eliminar las interacciones entre niños del diseño inicial y sustituirla por grabaciones de asambleas.

De esta forma, nuestro corpus presenta un diseño final formado por dos tipos de interacciones: **conversaciones colectivas espontáneas**, grabadas en las asambleas diarias y **entrevistas personales** hechas por un adulto a un único niño, donde la conversación pierde espontaneidad ya que está guiada por preguntas.

Otro de los objetivos del proyecto era obtener un corpus simétrico o equilibrado entre los participantes y las situaciones comunicativas. Para ello, se grabó al mismo número de niños de cada edad (tres, cuatro y cinco años), obteniendo tres subcorpus; y dentro de cada subgrupo, el mismo número de niños y niñas, haciendo así diferenciación entre sexos. El número de horas de grabación y palabras son similares para las intervenciones colectivas (asambleas) e individuales (entrevistas). El número total de participantes para las entrevistas fue de veinticuatro, en tres grupos de ocho sujetos —cada grupo se corresponde con una de las tres edades—, que a su vez se dividían en cuatro niños y cuatro niñas. Para las asambleas, el número de participantes no está tan equilibrado, ya que dependía del número de alumnos que hubiese en cada clase. Así, la clase de primero de infantil (3 años) tenía veintiún alumnos; la clase de segundo, también veintiuno; y la de tercero sólo diecisiete alumnos. Entre las tres suman un total de cincuenta y nueve participantes. El diseño final presenta el siguiente aspecto:

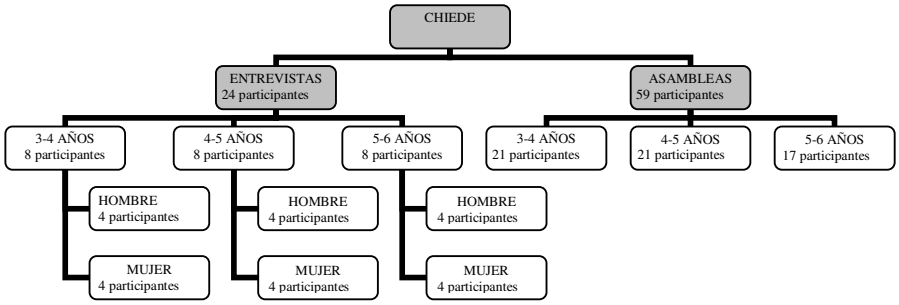


Figura 1.1 Diseño de CHIEDE

Según este diseño, consideramos que la representatividad queda resuelta en cuanto a que todos los grupos de edad están equilibrados y se proporcionan muestras de lenguaje oral en dos situaciones diferentes: en un contexto colectivo, donde el niño se siente cómodo y relajado entre iguales y conocidos; y un contexto artificial, donde es interrogado por un adulto desconocido para él. Además, este diseño permite al investigador acceder directamente a uno de los grupos en caso de que su interés recaiga solamente en niños de una sola edad, situaciones colectivas, etc.

Un punto importante, una vez recogidos los datos y antes de empezar a transcribir, fue decidir si era necesario hacer cambios en la metodología de transcripción para adaptarla a la variedad lingüística infantil. El problema con el que nos encontramos fue el tratamiento de la información fonética, ya que el lenguaje de algunos niños se situaba en una etapa aún primitiva de desarrollo y su sistema de fonemas estaba incompleto con respecto al sistema estándar adulto. Uno de nuestros propósitos era la aplicación de un transcriptor fonológico automático al texto, y realizar la transcripción fonológica desde la ortográfica. Pero no hay que olvidar que el etiquetador morfosintáctico también opera sobre la transcripción ortográfica y podría no reconocer una oración en la que las palabras se han escrito de la forma real en que han sido pronunciadas. La solución ha sido crear dos entradas diferentes para los turnos conversacionales de los niños con un sistema fonético prematuro: una primera línea, marcada *ABC (donde las mayúsculas representan las primeras tres letras del nombre del participante), en la que el enunciado es reconstruido siguiendo la forma normativa; y una segunda entrada marcada como *%pho*, donde la transcripción ortográfica es literal, basada en la pronunciación real. Así, el etiquetador morfosintáctico operará sobre la línea marcada con las iniciales del participante y el transcriptor fonético sobre la marcada como *%pho*. La siguiente muestra ejemplifica cómo sería el formato final de la transcripción:

*MAR: a ver /// siéntate conmigo /// en la alfombra / que te da frío /// aquí /// a ver /// ¿ qué me quieres contar ? ¿ qué habéis hecho en el gimnasio ?
*NAT: &eh /// &eh / corriendo / a [/] a [/] a [/] al &s / a pillarnos / el lobo ///
%pho: &eh /// &eh / codedo / a [/] a [/] a [/] al &s / a pillano / el lobo //
*MAR: ¡ah! / el lobo /// que habéis jugado al lobo //
%alt: (6) jugao
*NAT: &eh sí /// porque / me ha pillado el lobo ///
%pho: &eh sí /// poque / ve ha pillao el lobo ///
*MAR: ¿ y quién era el lobo ?
*NAT: Aitana / y yo ///
%pho: Intana / y yo ///

[NAT3]

En este caso, el menor se identifica con las siglas NAT, y después de cada una de sus intervenciones se incluye la línea *%pho,* con el mismo contenido pero transcrito de la forma más fiel posible al sonido. Esta línea adicional tiene como única finalidad servir de muestra para el transcriptor fonético automático; por ello sólo se incluirá en aquellas transcripciones donde sea necesaria, es decir, en las que el menor muestre un sistema fonético prematuro —solamente en las intervenciones de niños del primer grupo de edad, tres años. Por lo tanto, la marca *%pho* no se incluirá en las transcripciones en las que los participantes muestren un desarrollo completo del sistema fonético, ya que supondría escribir dos veces el mismo enunciado.

1.3 Diseño final de CHIEDE

CHIEDE es un corpus transversal, formado por tres grupos de sujetos divididos por edades. En la figura 1.1 se pueden ver los subcorpus o niveles de división: interacciones colectivas y entrevistas, y dentro de estas, tres ramificaciones correspondientes a los tres grupos de edad. El método seguido durante las grabaciones ha sido el obsevacional, evitando al máximo el intrusismo. Para ello, antes de llevar a cabo las grabaciones, el investigador dedicó un periodo de tiempo a la estancia en el centro sin llevar a cabo ninguna actividad excepto la observación y familiarización con los niños. Las entrevistas están dirigidas por el adulto, y por ello los temas de conversación son similares; no obstante, no existía un guión previo, con la intención de dejar absoluta libertad al niño para expresarse.

CHIEDE consta de 58.163 palabras, distribuidas en 30 textos, con un total de 7 horas y 53 minutos de grabación y 59 participantes menores. Cada grabación está alineada con su correspondiente transcripción ortográfica, en la que se incluye una cabecera con los metadatos o información sociolingüística. Además de los archivos de audio y texto, se incluyen otros dos tipos de archivo más: aquellos en los que se ha realizado la transcripción fonológica automática y aquellos en los que el texto aparece en formato XML con la anotación morfosintáctica. Los archivos se identifican con un nombre en el que se recoge la edad del participante o participantes menores de edad. Las interacciones colectivas se denominan con una A mayúscula y a continuación un número que hace referencia a la edad; en el caso de una asamblea del grupo de tres años el nombre sería A3. Pero puesto que existen dos asambleas por cada grupo de edad, éstas se han numerado para poder diferenciarlas. Así, tenemos seis archivos de asambleas: A3-01 y A3-02 del grupo de tres años; A4-01 y A4-02 del grupo de cuatro; y A5-01 y A5-02 del grupo de cinco años. Las entrevistas se denominan con las tres letras iniciales del participante menor seguidas por un número que lo sitúa dentro de un grupo de edad.

De esta forma, tenemos 30 carpetas —una por transcripción— dentro de las cuales se incluyen cuatro archivos:

- ADI4.txt : transcripción ortográfica.
- ADI4.wav : sonido de la grabación.
- ADI4.fon : transcripción fonológica.
- ADI4.eti : texto etiquetado morfosintácticamente.

El Apéndice A incluye un ejemplo de los tres archivos de texto, es decir, la transcripción ortográfica, la fonológica y la anotación morfosintáctica de una de las grabaciones del corpus.

CHIEDE no es un estudio dialectológico. En un corpus de dimensiones restringidas no se pueden representar todas las variantes geográficas. CHIEDE se definiría como corpus especial: "El termino *corpus especial* (special corpus) suele usarse para describir córpora (normalmente pequeños) que han sido diseñados con algún propósito específico, como por ejemplo, los que contienen lenguaje infantil o de hablantes no nativos" (Pérez Hernández, 2002:16). Así pues, dentro de las limitaciones obvias de un corpus de las dimensiones de CHIEDE, hemos intentado que su valor recaiga, como ya se ha comentado, en su carácter representativo, basado en la variedad y en su estructura.

En cuanto al contenido de los textos, se tuvieron en cuenta los siguientes elementos:

- Estructura dialógica: diálogos / conversaciones colectivas.
- Dominio de uso: entrevistas / asambleas.
- Variante de registro: formal o menos espontáneo.
- Características de hablantes: cualidades sociolingüísticas.

A continuación, presentamos las tablas con los datos relativos al número de palabras, duración de los archivos de audio y número de turnos y enunciados por fichero. La tabla 1.2 contiene los datos de las interacciones colectivas o asambleas, mientras que en la tabla 1.3 se muestran los pertenecientes a las entrevistas. Por último, en la tabla 1.4 se presentan los datos totales del corpus.

FICHERO	MINUTOS	TURNOS	ENUNCIADOS	PALABRAS
A3-01	36´11´´	751	1.154	3.877
A3-02	23´38´´	459	761	3.024
A4-01	31´37´´	537	942	4.284
A4-02	26´44´´	530	879	3.700
A5-01	37´10´´	1.108	1.502	5.657
A5-02	47´40´´	1.352	1.853	6.838
TOTAL	202´20´´	4.737	7.091	27.380

Tabla 1.2 Asambleas en CHIEDE

FICHERO	MINUTOS	TURNOS	ENUNCIADOS	PALABRAS
ADI4	10´21´´	297	398	857
ADR3	06´26´´	146	233	1.057
AIT4	15´03´´	251	397	1.527
ANG4	09´58´´	188	291	1.210
BRU3	19´32´´	404	662	2.147
CAL5	11´05´´	199	330	1.046
CAR5	07´56´´	151	254	911
CLA3	13´16´´	238	381	1.470
DAI5	08´09´´	217	345	1.154
ELE3	10´14´´	171	273	876
INE5	08´36´´	153	232	1.093
JAV3	13´47´´	204	321	1.180
JOM4	13´16´´	231	405	1.607
JOR4	13´39´´	351	481	2.061
MAI4	10´36´´	194	331	1.236
MAR4	12´17´´	223	329	940
MJO4	13´36´´	215	318	1.362
NAT3	14´34´´	234	410	1.410
RAU5	12´15´´	216	407	1.630
ROD3	11´30´´	274	392	1.288
SEL3	11´43´´	217	322	1.366
SOL5	10´00´´	157	250	1.247
TAL4	08´59´´	206	321	964
VIC5	08´03´´	168	270	1.144
TOTAL	**270´51´´**	**5.305**	**8.353**	**30.783**

Tabla 1.3 Entrevistas en CHIEDE

CHIEDE	MINUTOS	TURNOS	ENUNCIADOS	PALABRAS
	473´11´´	**10.042**	**15.444**	**58.163**

Tabla 1.4 Información general de CHIEDE

2. Metodología

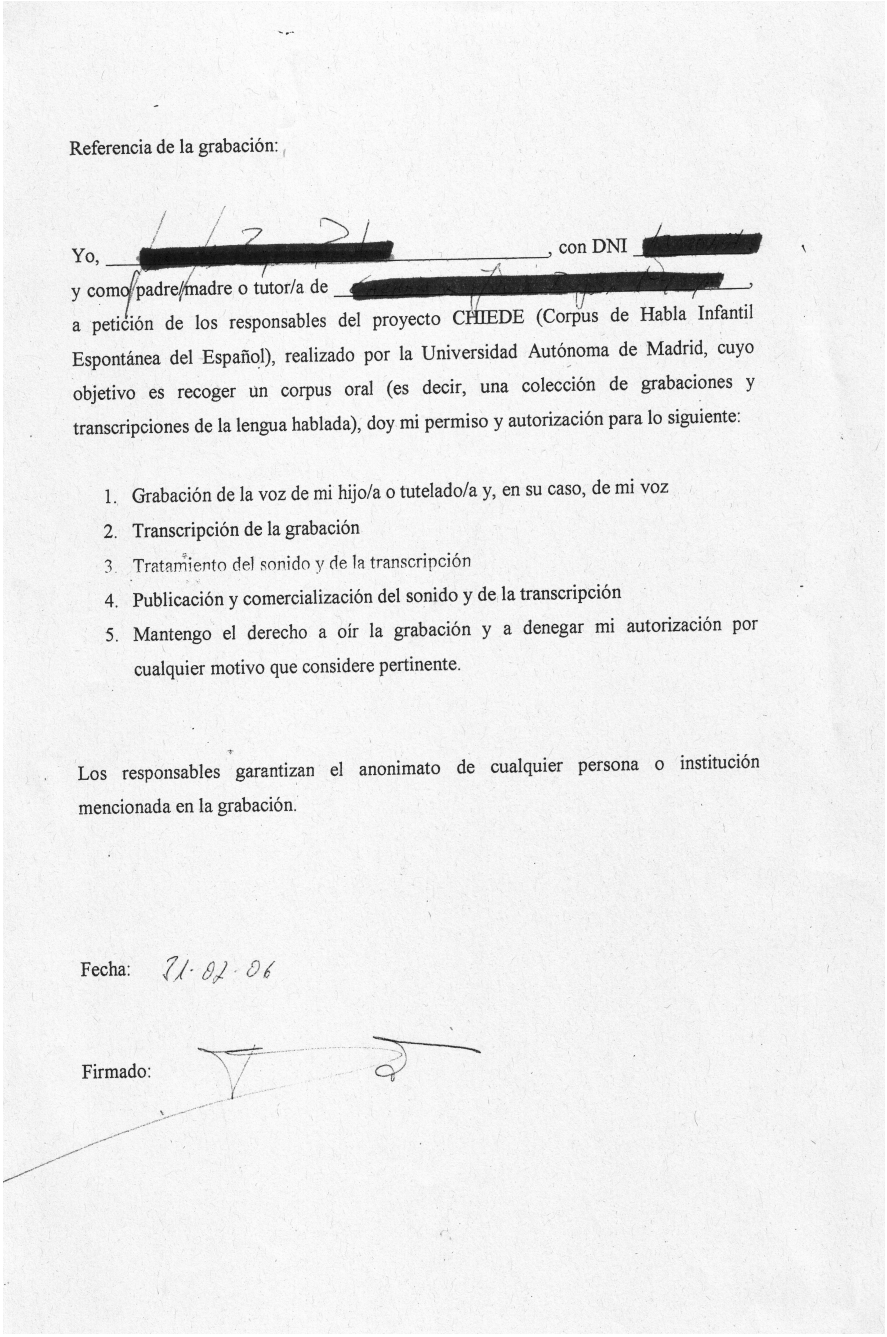
2.1 Grabaciones y entorno

La etnografía, y después la experiencia derivada de la lingüística de corpus, nos han aportando diferentes recomendaciones sobre cómo debe hacerse una grabación. Existen muchas posibilidades al respecto de cómo debería ser la situación ideal para captar el habla pura y espontánea de los hablantes. Algunas de ellas se refieren a la propia conducta del investigador; se habla, por ejemplo, de si influye o no el hecho de que este esté presente, oculto o participando en la conversación; también se tiene en cuenta si la grabadora está a la vista de los hablantes o si por el contrario está escondida; y en relación con esto último, se discute sobre si grabar sin que los hablantes lo sepan y después comunicárselo, o bien advertirles de todo con antelación.

Se podría discutir si todo este protocolo influye o no de forma negativa en la conducta de las personas que estaban siendo grabadas y por tanto en las muestras de oralidad y espontaneidad de su discurso. Es posible. Sin embargo, es un hecho bastante conocido que los hablantes comienzan a relajarse y algunos casi se olvidan de que se les está grabando después de los primeros minutos de grabación.

En nuestro caso estuvimos presentes en todas las conversaciones —siendo parte integrante en muchas de las entrevistas— porque este hecho aumentaba en gran medida el grado de comprensión de cara a la fase de transcripción, tanto de la situación comunicativa, como de lo que decían los hablantes, como de las acciones que tuvieron lugar, por ejemplo si un hablante abandonaba o no la conversación.

El hecho de que nuestro corpus pueda ser publicado para su uso por parte de investigadores de diversos campos científicos nos llevó a ser sumamente respetuosos y no violar el marco legal establecido para este tipo de situaciones. Es por ello que antes de la grabación se advirtió de la misma a los padres o tutores de los participantes y se les informó de que, bien antes, bien después de que esta se produjera, deberían firmar un permiso en el que autorizaban a la grabación de la voz de sus hijos. A continuación, presentamos el modelo de autorización que se empleó.



Referencia de la grabación:

Yo, [redacted], con DNI [redacted]
y como padre/madre o tutor/a de [redacted],
a petición de los responsables del proyecto CHIEDE (Corpus de Habla Infantil Espontánea del Español), realizado por la Universidad Autónoma de Madrid, cuyo objetivo es recoger un corpus oral (es decir, una colección de grabaciones y transcripciones de la lengua hablada), doy mi permiso y autorización para lo siguiente:

1. Grabación de la voz de mi hijo/a o tutelado/a y, en su caso, de mi voz
2. Transcripción de la grabación
3. Tratamiento del sonido y de la transcripción
4. Publicación y comercialización del sonido y de la transcripción
5. Mantengo el derecho a oír la grabación y a denegar mi autorización por cualquier motivo que considere pertinente.

Los responsables garantizan el anonimato de cualquier persona o institución mencionada en la grabación.

Fecha: 21-02-06

Firmado:

Figura 2.1 Ejemplo de autorización de grabación

Para esta primera etapa del corpus, es importante en la medida de lo posible controlar todos los aspectos que conciernen al investigador a la hora de grabar y que estos se conviertan casi en un hábito para evitar después contratiempos irreversibles. Así por ejemplo, es necesario comprobar más de una vez si se lleva todo el material necesario. Antes de grabar, hay que

comprobar que los micrófonos están encendidos, que no están en contacto directo con una superficie lisa y dura —porque entonces suelen hacer ruido y disminuye la calidad de la grabación— y por último, hacer una prueba de grabación.

Al final, cuando todo haya terminado, no hay que olvidarse de que los participantes lean, rellenen y firmen el permiso como está estipulado.

2.2 Digitalización

En este apartado se explican los pasos para digitalizar el sonido, esto es, pasar la grabación al ordenador y convertirlo en un archivo de sonido para poder transcribir desde el ordenador. Este proceso depende de diferentes factores, como por ejemplo el tipo de aparato que se haya utilizado para la grabación. En nuestro caso, disponíamos de una DAT (Digital Audio Tape), que permite la grabación digital de sonido en una cinta con un nivel de calidad profesional. Con un cable RCA se conecta la grabadora al ordenador y se transfiere la información a tiempo real.

Un punto importante a tener en cuenta es que la grabación de sonido ambiente nunca va a presentar la misma calidad que se obtiene en un estudio. El carácter real de las situaciones comunicativas y su naturalidad supone la posibilidad de que se produzcan interrupciones por parte de personas no implicadas en la interacción, o que se capte en la grabación cualquier tipo de ruido ajeno a la situación comunicativa. A pesar de que se tomaron todas las precauciones necesarias para evitar estos posibles contratiempos, la calidad de las grabaciones puede variar ligeramente de una a otra. De hecho, en algunos casos, se ha utilizado un programa de edición de sonido con el que se ha mejorado la calidad de algunas de las grabaciones.

De la misma forma que actualmente existen diversos aparatos para la grabación (grabadoras MP3, minidisc, etc.), también contamos con diferentes programas para el tratamiento del sonido, cuyo funcionamiento es sencillo (apto para inexpertos en el campo del sonido) y que están al alcance de cualquiera. En nuestro caso se utilizó el programa WaveLab, con el que trasladamos el sonido de las cintas DAT al ordenador.

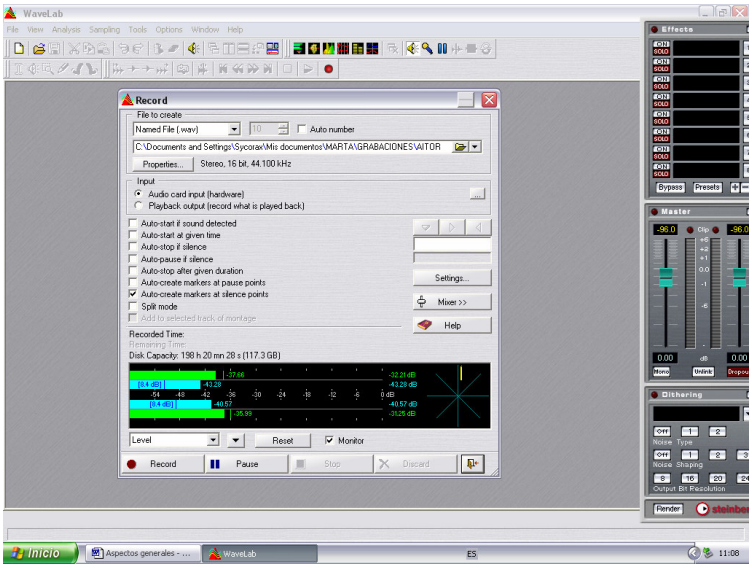


Figura 2.2 Proceso de traslado de la DAT al ordenador

Una vez capturado el sonido, WaveLab nos presenta una pantalla en la que aparece la onda sonora de la grabación. Este programa nos permite manipular el sonido mejorando su calidad, eliminando ruidos o cortando partes que no son relevantes o que preferimos no utilizar por cuestiones de anonimato e intimidad.

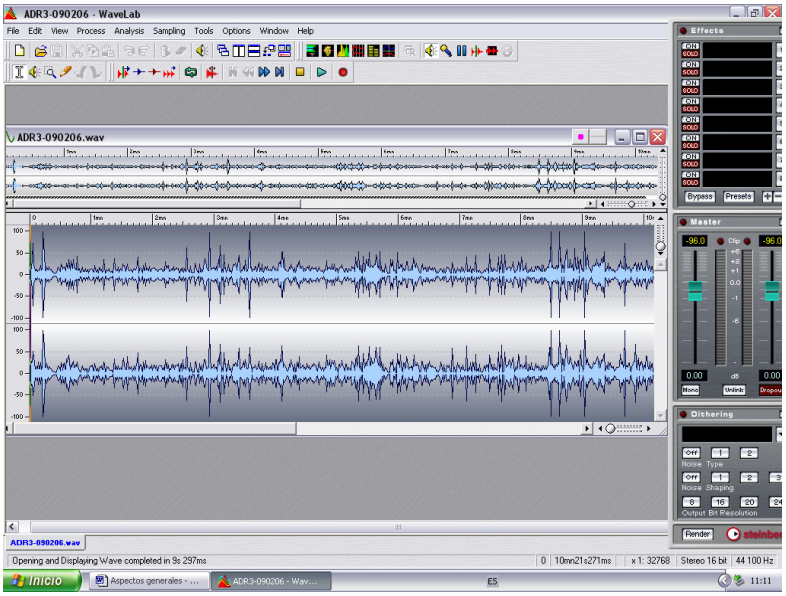


Figura 2.3 Edición del sonido con WaveLab

2.3 Transcripción

La transcripción es la tarea más costosa en la elaboración de un corpus de lengua oral por dos razones: la primera tiene que ver con el tiempo y la segunda con la metodología. Transcribir media hora de grabación puede llevar una jornada completa, si no más. Es un trabajo que consume una gran cantidad de tiempo, además de resultar agotador. Es la única fase en la creación de un corpus que, a día de hoy, debe hacerse de forma totalmente manual. Por esta razón, los corpus de lengua oral son menos abundantes que los de la variante escrita y sus dimensiones son significativamente menores.

El segundo de los problemas que se plantea a la hora de transcribir tiene que ver con la metodología y el establecimiento de los criterios y las convenciones que se van a seguir. Ya se ha comentado que no existe un sistema de transcripción ortográfica estándar como ocurre, por ejemplo, con la transcripción fonética. Existen varias propuestas, como el sistema CHAT (MacWhinney, 2007), o el propuesto por Payrató (1994); sin embargo, este último afirma:

> No existe ningún recetario ni existen fórmulas magistrales o mágicas que resuelvan de manera automática los conflictos derivados de la necesidad de transcribir; no existe, en definitiva, una transcripción ideal a la que acogerse de manera irreflexiva. (Payrató, 1994: 45).

La tendencia más habitual es que cada investigador fije unas convenciones concretas que se ajusten a la finalidad de su estudio. Hay que tener en cuenta que ni una transcripción refleja fielmente todos los fenómenos sonoros que se recogen en una grabación, ni una grabación recoge todos los elementos no verbales o extralingüísticos que pueden expresarse en una transcripción. Por esta razón, es necesario fijar unas convenciones de transcripción que reflejen de la forma más completa posible todos los elementos del evento comunicativo; además, si junto con la transcripción se facilita la grabación inicial, alineada con el texto, la utilidad del corpus se incrementa en gran medida.

Otro de los puntos importantes a la hora de establecer las convenciones de transcripción es tener en cuenta que estamos tratando la lengua oral. MacWhinney (2007) advierte que uno de los peligros a los que se enfrenta el transcriptor es la tendencia a tratar la lengua oral de la misma manera que la escrita:

> Transcribers have a tendency to write out spoken language with the punctuation conventions of written language. Written language is organized into clauses and sentences delimited by commas, periods, and other marks of punctuation. Spoken language, on the other hand, is organized into tone units clustered about a tonal nucleus and delineated by pauses and tonal contours […] (MacWhinney, 2007:16)

También Cresti y Moneglia advierten que "Usual syntactic schemes were unseccessful because of the high percentage of non-canonical patterns, and discourse units or speech acts were too elusive" (Cresti y Moneglia, 2005:XV). Para establecer los criterios de transcripción, es necesario tener en cuenta las diferencias entre la lengua oral y la escrita. Todos los autores destacan la incorrección de transcribir la lengua oral siguiendo los parámetros de la norma dictada para la lengua escrita y en especial cuando se trata de transcribir lenguaje infantil: las formas propias de esta variante lingüística no pueden ser reinterpretadas y representadas como formas propias del lenguaje adulto. Deben ser transcritas de la forma más fiel posible, aunque se reconstruyan en una línea dependiente. En este tipo de casos, el alineamiento facilita enormemente la tarea ya que permite al usuario comprobar por sí mismo cuál es el sonido exacto de la emisión.

Para la transcripción de las grabaciones de CHIEDE se ha utilizado la versión de CHAT usada en el proyecto C-ORAL-ROM, con algunas variaciones mencionadas en el punto 4.3 de la primera parte. El procedimiento de transcripción se divide en tres fases:

1. Transcripción ortográfica con inclusión simultánea del etiquetado de marcas prosódicas.
2. Revisión del etiquetado por un segundo etiquetador.
3. Segunda revisión por parte del primer etiquetador para garantizar al máximo la precisión.

El texto de una transcripción se va a estructurar en dos partes: cabecera[11] y la propia transcripción del sonido. Así pues, el formato textual comprende dos niveles:

1. Metadatos: información de la sesión de grabación (contexto, tamaño, participantes, calidad acústica, fuente, transcriptores y revisores).
2. Representación dialógica: información lingüística y no lingüística presentada en forma textual y con dos niveles de información:

    - Líneas de texto: transcripción ortográfica.
        - Vertical: turnos dialógicos.
        - Horizontal: anotación prosódica y límite de enunciados.
    - Líneas dependientes: información contextual.

A continuación, detallaremos las convenciones establecidas para la realización de las transcripciones.

[11] Para la información contenida en la cabecera se utiliza el inglés por ser la convención seguida en el proyecto C-ORAL-ROM.

2.3.1 Cabeceras (Metadatos)

En esta primera parte, se va a reflejar toda la información relacionada con la situación comunicativa y con la propia transcripción. Se aportan datos relativos al lugar dónde se hizo la grabación, quiénes participaron en ella, la longitud temporal de la interacción, etc.

@Title: título que ayude a reconocer el texto.
@File: nombre del archivo.
@Participants: datos de cada participante, donde se incluye:

- Nombre: tres letras mayúsculas, generalmente las tres primeras letras del nombre del participante.

- Diferenciación entre niños y adultos (child/adult).
- Sexo (man/woman).
- Edad: en el caso de los menores se indica la edad exacta; para los adultos, se utiliza el código A: 18-25; B 25-40; C: 40-60; D: >60.
- Educación: 1: primaria o menos; 2: secundaria; 3: estudios superiores.
- Origen geográfico y/o influencia lingüística (zona).

@Date: separado por barras, con las cuatro cifras del año.
@Place: lugar geográfico de la grabación.
@Situation: breve descripción del espacio e información relevante del lugar en el que se ha hecho la grabación.
@Topic: temas y subtemas tratados en el texto, con un máximo de 50 caracteres.
@Source: en este caso indicamos el nombre del proyecto.
@Class: tipo de interacción, es decir, formal, informal, privada, pública, etc.
@Length: minutos (') y segundos (").
@Words: número de palabras.
@Acoustic_quality: A (muy buena), B (buena), C (suficientemente buena).
@Transcriber: nombre del transcriptor.
@Revisor: nombre del revisor.
@Comments: comentarios del transcriptor sobre el texto; aquí hemos decidido incluir la información sobre la clase social del menor y su orden de nacimiento.

A continuación, se incluye un ejemplo de una cabecera donde se pueden observar todos los elementos expuestos arriba:

@Title: Adrián y Marta
@File: ADI4
@Participants: ADI, child, (man, 5;0, 1, Ciudad Real)
TEA, adult (woman, B, 3, Madrid)
@Date: 22/02/2006
@Place: Ciudad Real
@Situation: conversation in an empty classroom at school.
@Topic: daily matters
@Source: CHIEDE
@Class: informal, family/private, dialogue
@Lenght: 06´26"
@Words: 851
@Acoustic_quality: A
@Transcriber: Marta
@Revisor: Ana and Marta
@Comments: ADI (middle class; birth order: 1st)

2.3.2 Convenciones

2.3.2.1 Turnos

*ABC: los turnos se indican con asterisco, tres mayúsculas (pueden ser las tres primeras letras del nombre del participante), dos puntos y un espacio.

*TEA: ¿ sí ? ¿ fue tu cumple ? ¿ y qué hiciste ?
*ADI: pues jugar con mis amigos al fútbol ///

[ADI4]

En caso de imposibilidad de identificar al interlocutor se utiliza la variable *XYZ.

2.3.2.2 Comentarios

En los comentarios se introduce toda la información paralela a la conversación que se considera relevante para comprender lo que está sucediendo. Los comentarios se escriben inmediatamente después del turno. Y, normalmente, si se refieren a un hecho puntual, se suele especificar en qué momento se produce. En este caso, se especificará la posición de la palabra donde ocurre el hecho referido.

*INE: le vi un chupete hhh ///
%act: (5) INE laughs

[INE5]

Este comentario quiere decir que en el quinto segmento del enunciado la participante se ríe. Si por ejemplo la participante se riera varias veces a lo largo de la intervención, el comentario sería de esta forma:

%act: (9) and (12) INE laughs

Esto indica que entre las palabras novena y duodécima la participante se ríe. Mientras que si se riera en puntos independientes de su conversación, se pondría de esta manera:

%act: (6), (16) and (29) INE laughs

Como se podrá observar, los comentarios se escriben en inglés. En función de la clase de información que se quiera expresar en el texto, existen distintos tipos de comentarios:

- **%act**: son los comentarios más usuales y sirven fundamentalmente para expresar el paralenguaje. Pongamos un ejemplo:

  *TEA: hhh /// se le olvidó a mamá /// hhh // bueno / el año que viene / ¿no?
  %act: (1) laugh (7) clic

  [ADI4]

- **%com**: sirve para hacer referencia a algún cambio ocurrido en la situación, como por ejemplo, que de pronto un participante se levante, se cambie de lugar, o se marche de la escena, o también la llegada de otro participante. Todo lo que pueda influir en la forma de enfocar la situación comunicativa y en los temas de conversación. Por ejemplo, encender la televisión y comentar los anuncios publicitarios.

  *TEA: a ver // Adrián /// ¿ cuántos años tienes ?
  %com: ADR shows four fingers

  [ADI4]

- **%alt**: este tipo de comentario tiene dos funciones: en primer lugar, se utiliza para reflejar los errores de producción. Por ejemplo, imaginemos que un hablante en un momento determinado de la conversación se equivoca y dice "pacharrón" en lugar de "chaparrón". En nuestro caso, al tratarse de lenguaje infantil, bajo esta etiqueta se sitúan todos los errores debidos a la hiperregularización o errores por analogía. En la transcripción se presentaría así:

  *ADI: [<] <menos la tarta> // porque no traje ///
  %alt: (6) trayí

  [ADI4]

Así pues, en la transcripción se refleja la forma normativa, y lo que el hablante dijo realmente se pone en el comentario. En segundo lugar, los comentarios %alt sirven para recoger todo tipo de expresiones o fenómenos que son típicos del habla oral pero que no están aceptados por la norma. Por ejemplo, imaginemos que un hablante dice "fúbol" en lugar de "fútbol" o "m'acuerdo" en lugar de "me acuerdo". En todos estos casos, la forma que, desde un punto de vista normativo, se considera correcta se pone en el texto de la transcripción, mientras que la variante lingüística se pone en el comentario.

*DAI: no lo sé / ya no me acuerdo ///
%alt: (6-7) m'acuerdo

*DAI: &eh al fútbol /// hoy no se han traído balón / pues vamos a jugar ///
%alt: (2) fúbol

[DAI5]

2.3.2.3 Etiquetas para marcar la información prosódica

• **Pausas entonativas**: / sirve para indicar las unidades tonales. En términos generales, se pondrá una barra / cuando intuitivamente se perciba una pequeña pausa tonal, a pesar de que el hablante no haya terminado su unidad entonativa o enunciado.

*INE: xxx casita del bosque / vivía Caperucita Roja /// y le dijo su mamá // ven Caperucita Roja /// llévale la comida a la abuela // que no puede venir // está coja / muy malita /// entonces dijo &pa [/] Caperucita / vale ///

[INE5]

Como podemos observar, cuando el hablante llega a *bosque* hace una pausa pero no ha terminado de hablar. Normalmente, la / se suele corresponder con una breve pausa con entonación ascendente, que indica que el hablante no ha finalizado su discurso (independientemente de si desde un punto de vista sintáctico la oración esté completa o no), mientras que con las dos barras // el hablante hace una pausa más marcada, a veces con entonación descendente, pero a continuación prosigue su unidad informativa.

• **Final de enunciado**: estos cinco signos, **///**, **?**, **!**, **…**, **+** (según sea asertivo, interrogativo, exclamativo, suspensión por conocimiento compartido o interrupción por parte de otro hablante) sirven para indicar el final del enunciado.. Los tres puntos indican el final de un enunciado con entonación suspendida. Es un enunciado que no se termina porque el

hablante considera que su interlocutor ya ha inferido lo que quiere decir. Pongamos un ejemplo:

> *TEA: muy bien /// claro /// poquito a poco /// ya el año que viene sabrás más / y luego más / y luego más ...
>
> [INE5]

En este caso, el hablante considera que no es necesario terminar su enunciado porque se deduce lo que quiere decir. Así pues, los puntos suspensivos se ponen cuando la entonación suspendida guarda en sí misma una carga de significado que no se explicita pero que se da por supuesto. No hay que confundir los puntos suspensivos con los alargamientos vocálicos: en este caso, el hablante no considera que su interlocutor pueda inferir el resto de la información, sino que el alargamiento, ya sea de un sonido vocálico o uno consonántico, en una estrategia de planificación del discurso, pero el hablante culminará su enunciado. Los alargamientos se señalan con el signo → a continuación del fonema alargado.

La triple barra (///) se indica para señalar el final de todos los demás tipos de enunciados, de la clase que sean, afirmativos, dubitativos, imperativos, etc.

- **Continuación de turno**: se pondrá este símbolo ¬ a principio de turno para indicar que se continúa la unidad informativa del turno anterior, tras una interrupción o solapamiento.

> *MAT: Candi /// < pues yo >
> *TEA: [<] < ¿ qué ? >
> *MAT: ¬ por la noche ceno / un vaso de leche con cola-cao y cereales ///
>
> [A5-02]

2.3.2.4 Apoyos vocálicos

Son signos de transcripción que se utilizan para expresar aquellas vocales que pronunciamos como por ejemplo “eh” y “ah”, con el fin de alargar el tiempo de planificación de habla. En nuestro corpus se van a transcribir de esta forma: &ah o bien &eh.

> *INE: no [/] no me gusta lo de princesas // sólo lo de &eh animales ///
>
> [INE5]

2.3.2.5 Interjecciones

Las interjecciones se transcriben entre signos de exclamación ¡!, sin espacio entre la palabra y los signos, y se incluyen en un glosario de interjecciones ampliable. El DRAE (diccionario de la Real Academia) tiene su propia lista de interjecciones. Cuando el transcriptor se encuentre con una palabra que él considera que se comporta como una interjección (y que por tanto debe transcribirse entre signos de exclamación) pero que sin embargo no está recogida por el DRAE debe añadirla a la lista de interjecciones de C-ORAL-ROM. Recordemos que para que una palabra se considere interjección no tiene que poseer ningún tipo de significado referencial. Las interjecciones no sirven para designar realidades sino únicamente para expresar el estado de ánimo del hablante.

2.3.2.6 Los signos paralingüísticos

Los sonidos que no son lingüísticos pero que tienen significado, por ejemplo, de duda, de interrogación, etc. se reflejan como *hhh*. Algunos de estos signos paralingüísticos son:

laugh (risa)
assent (asentimiento)
click (chasqueo)
cough (tos)
negation (negación)
doubt (duda)
interrogation (interrogación)
blow (soplido)
clearing of throat (carraspeo)
yawn (bostezo)

A la hora de transcribir, en el texto de la transcripción se pondrá **hhh** mientras que el significado se pondrá en un comentario de tipo %act, tal y como ya hemos explicado más arriba.

*TEA: [<] < ¡ah! > ¡ah! / un año / hhh /// entonces es muy pequeñita / claro /// sólo tienes esa hermana ///
%act: (5) blow
*DAI: sí ///
%act: whisper

[DAI5]

2.3.2.7 Sonidos no reconocidos como signos lingüísticos

- **Las tres equis**: se recurre a este signo de transcripción (xxx) cuando el transcriptor no entiende lo que ha dicho el hablante. Indica sonido ininteligible.

> *DAN: sí /// una hermana / que xxx / que tiene la varicela ///
> [DAI5]

- **El ampersan**: es el signo (&) que tiene que preceder a palabras inconclusas o fragmentos.

> *DAN: yo &siem [///] unas veces soy el portero yo ///
> [DAI5]

2.3.2.8 Los reinicios o reformulaciones

Los signos para expresar reinicio o reformulación son dos: [/] y [///]. Se utilizan para aquellos casos de reprogramación del habla en los que el hablante se traba o se confunde al decir una palabra y necesita pronunciarla de nuevo. También sirve para aquellos casos en los que el hablante no se muestra satisfecho con lo dicho y prefiere comenzar otra vez. En el reinicio siempre hay dos partes:

> *JOM: yo no lo vi /// fui [/] fue mi padre ///
> [A4-02]

A la izquierda del signo de reinicio siempre se encuentra el fragmento desechado; mientras que a la derecha del reinicio se encuentra la opción que el hablante ha dado por válida. Dependiendo del grado de semejanza o variación que haya a un lado y a otro se elegirá entre [/] y [///]. Normalmente, se suele utilizar [/] para aquellos casos en los que a un lado y a otro existe o bien la misma cadena de habla, o bien dos palabras que a pesar de ser distintas tienen la misma categoría gramatical. Veamos un ejemplo de cada tipo:

> *DAN: no /// con Miguel que me di una voltereta /// que [/] que [///] encima de él /// me &eh / y [/] y [///] sí ///
> [DAI5]

Como vemos, en estos ejemplos marcados con [/] hay repetición pero no se produce una ruptura de la sintaxis de la oración. Mientras que cuando se utiliza el signo con tres barras [///], la reprogramación del habla es completa.

2.3.2.9 Interrupciones

Existen dos tipos de interrupciones.

- **La autointerrupción** (=)**:** esto es, cuando el hablante corta su discurso de una forma brusca para hablar de otro tema completamente distinto al que estaba tratando anteriormente. La diferencia entre = y [///] es temática, es decir, que mientras = implica un cambio de tema dentro del discurso, [///] no.

  *TEA: ¿ y qué estabas haciendo ?
  *SOL: pues es que estaba buscando piedras // porque [///] para = y me estaba divirtiendo mucho // para el -> / para el patito /// es que ha traído patitos ///

  [SOL5]

- **La interrupción** de un participante a otro cuando el primero no ha concluido su enunciado. Esto se marca con el signo + en el punto de la interrupción, para indicar que el enunciado está inconcluso.

  *TEA: por dónde &teni +
  *ADI: por el cuello ///

  [A4-01]

2.3.2.10 Pausa

El signo # se utiliza para marcar una pausa prolongada. Es un silencio, un espacio temporalmente muy perceptible y que no se pueda considerar una pausa prosódica.

2.3.2.11 Solapamiento

El solapamiento se produce cuando dos hablantes hablan a la vez. El texto solapado aparece entre los signos < >. En el siguiente turno, [<] indica que se solapa con el anterior.

*TEA: dame que te ponga la -> /// < la pinza >
*SOL: [<] < la horquilla > /// más / que ya he perdido muchas ///

[SOL5]

Como se ha comentado en varias ocasiones a lo largo de este documento, todas las transcripciones son revisadas, y este proceso implica a investigadores diferentes. Una vez realizada la transcripción inicial, una persona diferente la revisa, remitiendo los resultados al primer transcriptor. Este, a su vez, vuelve a repasar el texto para asegurarse de que no existe ningún error, y que la transcripción es lo más fiel posible al sonido.

2.4 Alineamiento

El alineamiento consiste en la sincronización del texto transcrito con el sonido original, de forma segmentada, es decir, o bien por turnos conversacionales o bien por enunciados. Nuestro corpus está alineado por enunciados, siguiendo la metodología del LLI-UAM. El sistema de alineamiento permite al investigador o usuario del corpus verificar la transcripción escuchando el fragmento de sonido.

En cuanto a la forma de hacerlo, se ha producido un cambio metodológico con respecto a C-ORAL-ROM. Para éste, se había utilizado el programa WinPitch Corpus, desarrollado específicamente para tareas de transcripción y análisis de datos lingüísticos orales. El proceso consistía en transcribir el texto y, una vez completo, se activa el sonido mientras el usuario selecciona la parte del texto correspondiente a la unidad de sonido percibida. De esta forma, se construye automáticamente una base de datos con los segmentos definidos. No obstante, para nuestro proyecto decidimos utilizar otro programa, Transana, por diversas razones.

Transana es un programa de transcripción y análisis de datos creado en la Universidad de Wisconsin y disponible de forma gratuita a través de su página web[12]. En nuestra opinión, este programa presenta tres ventajas básicas:

- **Simultaneidad de operaciones**: Transana se presenta en un formato que permite mantener activas simultáneamente varias ventanas, entre las que se incluyen la ventana de texto y la de sonido (también está preparado para vídeo, en caso de que la grabación incluya imágenes). De esta forma, se puede operar en la ventana de texto al mismo tiempo que se escucha el sonido. Esta característica facilita enormemente el trabajo del transcriptor, ya que agiliza el proceso. En la imagen que se incluye a continuación podemos observar el aspecto de Transana:

[12] www.transana.org

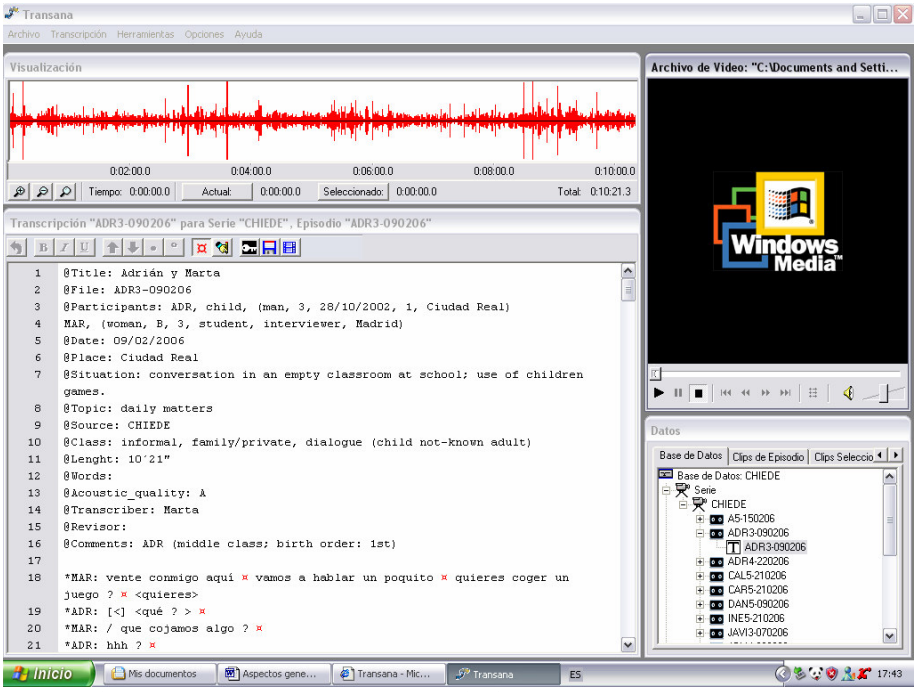


Figura 2.4 Programa de transcripción Transana

Como se puede apreciar, el programa se compone de cuatro ventanas: arriba, en la parte izquierda, aparece la ventana de sonido con la imagen de la onda sonora e información sobre el tiempo total de duración de la grabación, o el tiempo actual de la transcripción; debajo de esta ventana encontramos la de texto, con una pequeña barra de herramientas para la edición; el la parte superior derecha tenemos la ventana de vídeo; y debajo de ella está la ventana de datos, donde almacenamos los distintos archivos de sonido y texto.

- **Ergonomía**: Transana dispone de un número de atajos en el teclado con los que se facilita la labor del transcriptor. Con ellos, se pude detener el sonido sin necesidad de recurrir al uso del ratón. Además, permiten restituir el sonido en diferentes puntos, ya sea antes, después o en el momento justo en el que se paró la grabación. Los atajos y sus funciones son los siguientes:

Ctr+S	Parada/marcha; retrocede automáticamente dos segundos
Ctr+D	Parada/marcha; sin retroceso
Ctr+A	Reinicio diez segundos atrás
Ctr+F	Reinicio diez segundos adelante
Ctr+P	Comienza en el principio de la grabación

- **Alineamiento simultáneo**: Transana permite alinear el sonido durante el mismo proceso de transcripción. Utilizando el atajo Ctr+D se para el sonido en el momento preciso de la finalización de cada enunciado, y se marca con el símbolo ▫ destinado a esta tarea, sincronizando o alineando el segmento de sonido con su correspondiente fragmento textual. Una vez alineado todo el texto, se puede escuchar la grabación con un sistema similar al de un karaoke, en el que el texto correspondiente al sonido se va marcando.

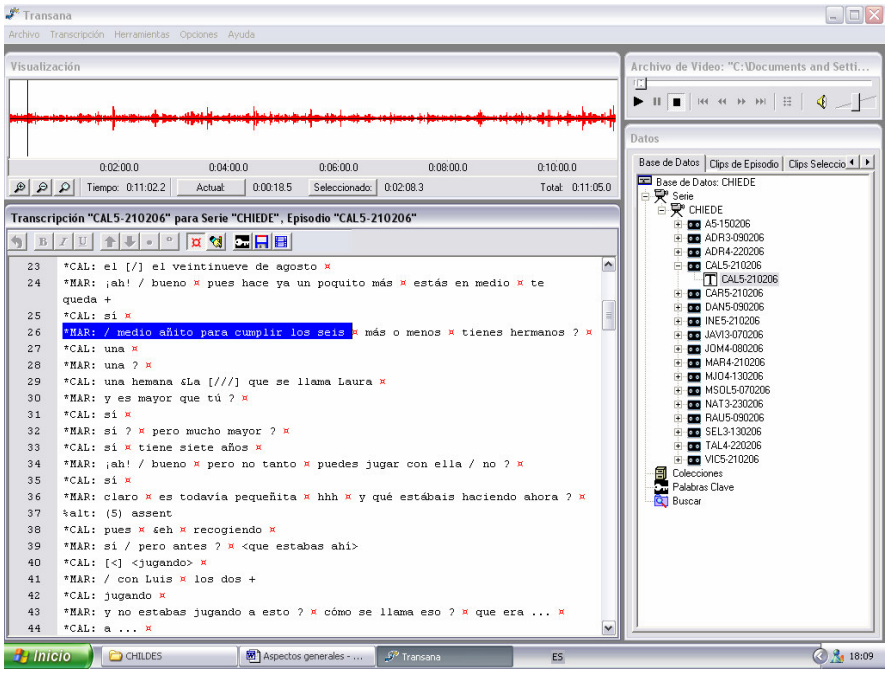


Figura 2.5 Alineamiento en Transana

2.5 Conversión a XML

El texto de salida en Transana se puede exportar a cualquier otro editor de texto, como por ejemplo Microsoft Word o Edit Plus, simplemente seleccionando el texto, copiando y pegándolo en el documento elegido. En nuestro caso, nos interesaba convertir el formato de texto de Transana en un documento en XML. Sin embargo, las marcas de tiempo establecidas en Transana para alinear se convierten en cifras numéricas al exportar el texto a otro tipo de documento. Esas cifras numéricas expresan el tiempo exacto del alineamiento de texto y sonido en milisegundos, quedando de la siguiente manera:

*TEA: a ver Daniel /// ¿ cuántos años tienes tú ? ¤<1559>
*DAN: cinco ¤<2260>
*TEA: ¿ cinco ? ¤<2829> ¿ y cuándo los cumples ? ¤<3653>
*DAN: el verano ¤<4946>
*TEA: <o sea>
*DAN: [<] <marzo> ¤<6215>
*TEA: ¡ah! / en marzo ¤<7079> o sea / que te queda / bueno / te queda un -> [/] un mes ¤<9917>
*DAN: sí ¤<10400>
*TEA: ¿ en marzo ? ¤<11234> ¿ o en mayo ? ¤<12443>

[DAI5]

Transana facilita la labor de alineamiento ya que permite realizarlo de forma simultánea el proceso de transcripción; pero en el resultado final encontramos cifras de tiempo en lugar del símbolo ///. Por lo tanto, era necesario convertir el formato de salida de Transana al formato convencional de C-ORAL-ROM, es decir, desglosando el texto en enunciados limitados por una marca de tiempo al inicio y otra al final de los mismos. Así pues, se hizo un programa que realizara el cambio de forma automática para que el documento final mostrase el mismo aspecto y formato que las transcripciones alineadas de C-ORAL-ROM, quedando así:

<UNIT speaker="TEA" startTime="0" endTime="0.509"> a ver Daniel </UNIT>
<UNIT speaker="TEA" startTime="0.509" endTime="1.559"> ¿ cuántos años tienes tú ? </UNIT>
<UNIT speaker="DAI" startTime="1.559" endTime="2.26"> cinco </UNIT>
<UNIT speaker="TEA" startTime="2.26" endTime="2.829"> ¿ cinco ? </UNIT>
<UNIT speaker="TEA" startTime="2.829" endTime="3.653"> ¿ y cuándo los cumples ? </UNIT>
<UNIT speaker="DAI" startTime="3.653" endTime="4.946"> el verano </UNIT>
<UNIT speaker="TEA" startTime="4.946" endTime="6.215"> < o sea > </UNIT>
<UNIT speaker="DAI" startTime="4.946" endTime="6.215"> [<] < marzo > </UNIT>
<UNIT speaker="TEA" startTime="6.215" endTime="7.079"> ¡ah! / en marzo </UNIT>
<UNIT speaker="TEA" startTime="7.079" endTime="9.917"> o sea / que te queda / bueno / te queda un -> [/] un mes </UNIT>
<UNIT speaker="DAI" startTime="9.917" endTime="10.4"> sí </UNIT>
<UNIT speaker="TEA" startTime="10.4" endTime="11.234"> ¿ en marzo ? </UNIT>

<UNIT speaker="TEA" startTime="11.234" endTime="12.443"> ¿ o en mayo ? </UNIT>

La etiqueta <UNIT> incluye un solo enunciado, especificando el tiempo de inicio y el de finalización en formato XML. En Transana, la primera marca de tiempo no puede corresponderse con 0, ya que ha tenido que pasar al menos un segundo para ponerla. Así, la primera marca en el ejemplo es <1559>, que indica el final del primer enunciado en milisegundos. El programa de cambio a formato C-ORAL-ROM establece el inicio del primer enunciado en 0 y el final en "1.559"; esta cifra será a su vez el tiempo de inicio del siguiente enunciado, cuyo tiempo de finalización será de nuevo el tiempo de inicio de la siguiente y así sucesivamente.

Para los ficheros de texto en los que se presenta la transcripción ortográfica sin ningún tipo de anotación —a excepción del etiquetado prosódico— las cifras numéricas de Transana se sustituyeron por la marca de fin de enunciado ///.

3. La anotación

3.1 Anotación morfosintáctica del corpus

Una vez transcrito y revisado todo el corpus, el siguiente paso es su anotación morfosintáctica. Ésta se hace de forma automática mediante un *etiquetado morfosintáctico* o programa informático que analiza un *input* dado, asignando una etiqueta con sus rasgos morfológicos y gramaticales. Aunque el *etiquetado morfosintáctico* facilita enormemente el trabajo, ya que hacerlo de forma manual sería una tarea que implicaría mucho tiempo y esfuerzo, sus resultados no son satisfactorios al cien por cien. Así pues, es necesaria una posterior revisión manual para corregir posibles errores de anotación.

La anotación de un corpus supone el enriquecimiento de los textos. La transcripción consiste únicamente en la transliteración del discurso oral a los caracteres ortográficos convencionales; mientras que la codificación de un texto incluye el reconocimiento, análisis y anotación de los fenómenos del discurso transcrito.

La información lingüística que se anota es variada, de nuevo dependiendo de los intereses de la investigación. Esto se debe, como ocurre con los sistemas de transcripción, a que no existe un formato de anotación estándar reconocido por la comunidad científica. No obstante, proyectos como EAGLES (Expert Advisory Group on Language Engineering Standards)[13] o TEI (Text Encoding Iniciative), en un intento por aunar criterios para la codificación e intercambio de textos en formato electrónico, establecieron unos requisitos fundamentales entre los que se encuentran la utilización del lenguaje SGML, además de la división del texto en dos partes: cabecera y transcripción.

También son conocidas las máximas de Leech (1993) sobre anotación. Dicho autor establece siete normas fundamentales a tener en cuenta:

- Posibilidad de eliminar la anotación para volver al texto plano.
- Posibilidad de extraer las anotaciones del texto.
- El esquema de anotación debe estar basado en pautas disponibles para el usuario en forma de guía.
- Especificación de cómo y por quién se hizo la anotación (si se hizo de forma automática y si fue corregida de forma manual posteriormente).
- Aviso al usuario de que la anotación no es infalible, sino sólo una herramienta útil.

[13]http:// www.ilc.cnr.it/EAGLES/home.html

- La anotación debe estar basada, en lo posible, en principios teóricos extendidos y neutrales.
- Ningún esquema de anotación se puede considerar un estándar a priori, sino que será la comunidad científica quien, por consenso práctico, le otorgue esa cualidad.

En cuanto a los tipos de anotación, McEnery y Wilson (http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fra1.htm) establecen la siguiente tipología:

- *PoS (Part of Speech) tagging*: la anotación de categorías lingüísticas es el procedimiento más común. Facilita la recuperación de información y la posterior anotación sintáctica y semántica. Además, ayuda en la distinción de homógrafos y en la resolución de la ambigüedad categorial.
- Lematización: reducción de las palabras a lexemas, es decir, a la forma básica de la palabra sin flexiones. Esto permite contabilizar el número de lemas y también el número de formas o palabras diferentes que aparecen en el corpus.
- *Parsing*: anotación mediante árboles sintácticos. Es menos habitual que el *PoS tagging*, debido a las complicaciones que entraña, en concreto en lenguas en las que el orden oracional es libre.
- Anotación semántica: se codifica la relación entre palabras o la información semántica de cada palabra.
- Anotación lingüística discursiva y textual: mediante etiquetas discursivas, anotación anafórica, etc.
- Transcripción fonética: siguiendo las convenciones del AFI. En la actualidad, es más habitual dentro de las tecnologías lingüísticas la utilización de SAMPA (Speech Assessment Methods Phonetic Alphabet), un sistema de transcripción fonética en formato electrónico que traduce los símbolos del AFI a códigos ASCII, lo que le confiere robustez, sobre todo en aquellos casos en los que Unicode no es apropiado o no está disponible.
- Prosodia: es el tipo de anotación básica para los corpus de lengua oral, ya que, como hemos comentado, las convenciones de puntuación normativas para la lengua escrita no son válidas para la transcripción de la lengua oral. Payrató (1994) establece tres niveles de anotación prosódica, dependiendo de lo rigurosos que queramos ser en dicha tarea. Así, el Nivel I recogería únicamente las pausas prosódicas, el Nivel II sería el utilizado en nuestro corpus, en el que se detallan además fenómenos paralingüísticos, y por último, en el Nivel III se realizaría una transcripción fonética estrecha, detallando elementos suprasegmentales.

3.1.1 Etiquetario

El proceso de etiquetado categorial de un corpus exige la creación de un sistema de etiquetado o etiquetario bien definido. Cada una de las etiquetas

empleadas debe fijar sus límites según criterios morfológicos, semánticos, sintácticos y pragmáticos. Nuestro corpus ha contado con las herramientas y metodología empleadas en el LLI-UAM para la realización del corpus C-ORAL-ROM, entre las que se encuentran el sistema de etiquetado; a continuación, pasaremos a describir dicho sistema.

Por etiqueta se entiende un símbolo descriptivo que se asigna de forma manual o automática a una palabra simple o compleja en un texto. Según esta definición, es necesario establecer el concepto de palabra que se va a tener en cuenta para el diseño del sistema de etiquetado. En C-ORAL-ROM se considera que una palabra es la unión de una cadena de sonidos que remite a un solo significado. Este criterio semántico obliga a distinguir dos tipos de palabras: simples y complejas. Las primeras se representan desde un punto de vista gráfico entre dos espacios en blanco. Las complejas, por el contrario, son grupos de dos o más unidades gráficas, unidas por guiones bajos, y reciben el nombre de *multipalabra*[14].

Según nuestros criterios de etiquetado, para que una palabra se considere multipalabra debe cumplir dos requisitos:

- Ausencia de significado composicional: el grupo de palabras debe designar un nuevo significado composicional independiente del significado propio de cada una de las palabras de forma aislada.
- Imposibilidad de insertar otras palabras dentro del grupo: esto le otorga el carácter de unidad al grupo. Por ejemplo, la multipalabra *fin_de_semana* no aceptaría la inclusión de otra palabra que complementase a ninguno de los elementos que la forman. Así, **fin de larga semana* sería una construcción incorrecta, o cuando menos anómala.

3.1.1.1 Problemas para el establecimiento de un etiquetario

Al igual que ocurre con las convenciones de transcripción, la inexistencia de un sistema de etiquetado categorial estándar implica la necesidad de diseñar uno propio basado en criterios claros y coherentes. Las tres últimas máximas de Leech (1993) sobre anotación, mencionadas en el apartado anterior, hacen referencia a este hecho. Los problemas de clasificación categorial siguen provocando muchas controversias. Como indica Bosque (1999), las distintas clasificaciones categoriales suelen estar basadas en una mezcla de criterios, principalmente semánticos y sintácticos. Esto se debe a la dificultad que entraña tal clasificación.

Es debido a esta problemática que creemos oportuno hacer una breve aclaración sobre el concepto de categoría gramatical que ha guiado la

[14] Del término inglés *multiword.*

clasificación categorial en nuestro sistema de etiquetado. En primer lugar, los criterios que se han seguido para la definición de las categorías han tenido en cuenta tanto aspectos sintácticos como morfológicos y semánticos. Esta mezcla de puntos de vista se debe precisamente al hecho que menciona Bosque sobre las diferencias existentes entre categorías y la dificultad de tratarlas a todas desde una única perspectiva. Por otra parte, el enfoque a la hora de realizar la clasificación ha sido de corte funcionalista y no generativista. Para esta última corriente, las palabras poseen por sí mismas una categoría gramatical y como consecuencia de ello tienen un comportamiento determinado dentro de la sintaxis de una lengua. Por el contrario, el concepto funcionalista de categoría es inverso, considerando que es la función determinada que desempeña una palabra dentro de la oración la que define la categoría a la que puede pertenecer dicha palabra. Es decir, que según el funcionalismo, una palabra por sí misma no pertenece a ninguna categoría gramatical de antemano, sino que según su comportamiento sintáctico adquirirá una categorización u otra. Esta última es la visión que se ha tenido en cuenta a la hora de realizar el sistema de etiquetado. Un claro ejemplo de categorización funcionalista es el del adjetivo en función de sustantivo cuando va precedido de artículo. En este caso, se considerará la capacidad sustantivadora del artículo, y una palabra que habitualmente está categorizada como adjetivo, se recategorizaría como sustantivo.

No obstante, la posición sintáctica que ocupa una palabra dentro de la oración puede considerarse un argumento insuficiente para la recategorización de una unidad léxica. Es por ello por lo que nuestro sistema de etiquetado se basa también en criterios morfosintácticos y, especialmente, en criterios semánticos.

3.1.1.2 Concepción teórica del etiquetario en C-ORAL-ROM

Al contrario de lo que plantea el enfoque funcionalista, que privilegia el criterio sintáctico en su concepción categorial, el sistema de etiquetado desarrollado en el LLI-UAM parte de unas bases semánticas sobre las cuales se asentarán el resto de parámetros. Por lo tanto, sería la semántica de una palabra la que determinaría su comportamiento morfológico y sintáctico en un determinado enunciado. Así, volviendo al ejemplo del adjetivo, éste se define como un elemento que designa cualidades o propiedades de las entidades representadas por los sustantivos. En castellano, las palabras que designan cualidades poseen flexión de género y número heredada de los nombres a los que complementan y desempeñan ciertas funciones a nivel oracional como complemento del nombre, atributo y complemento predicativo. No obstante, un adjetivo, por sus características semánticas, puede ser recategorizado mediante diferentes recursos morfológicos o sintácticos que proporciona cada lengua, dependiendo de lo que queramos expresar. En concreto, ciertos adjetivos carecen de un sustantivo equivalente

que cumpla la misma función, como señala Bosque (1990) que ocurre con los adjetivos relativos a características físicas (*la rubia*), características morales o anímicas (*pecador*) y rasgos sociales (*los ricos).* Así, dentro de los campos semánticos mencionados, podemos encontrar palabras que en unas ocasiones designen cualidades y en otras grupos de individuos.

De la misma manera, serían las restricciones semánticas de una palabra, y no las morfológicas o sintácticas, las que impedirían que un determinado adjetivo pueda funcionar como sustantivo. A continuación, presentamos una tabla a modo de resumen de las dieciséis categorías que conforman el etiquetario, junto con la etiqueta que las designa y un ejemplo de cada una de ellas.

Categoría	**Etiqueta**	**Ejemplo**
Sustantivo	N	*mesa*
Nombre propio	NPR	*Natalia*
Adjetivo	ADJ	*rojo*
Artículo	ART	*el, la, lo*
Posesivo	POSS	*mi, tu, su*
Demostrativo	DEM	*este, ese*
Cuantificador	Q	*uno, primero, muchos*
Pronombre	P	*yo, tú, lo_que*
Relativo	REL	*que*
Verbo	V	*comer*
Auxiliar	AUX	***he** comido*
Preposición	PREP	*ante, bajo, con*
Adverbio	ADV	*aquí, así, allí*
Conjunción	C	*y, pero, ni*
Marcador discursivo	MD	*oye, o_sea, es_decir*
Interjección	INTJ	*madre_mía, uy*

Tabla 3.1 Etiquetario

3.1.2 GRAMPAL

Con anterioridad a la creación etiquetador morfosintáctico es necesario el diseño del etiquetario, descrito arriba, y el desarrollo de una base de datos léxica o diccionario.

GRAMPAL fue desarrollado en origen como un procesador morfológico del español para la lengua escrita (Moreno 1991, Moreno y Goñi 1995). Más tarde, con la necesidad de anotar el corpus de lengua oral C-ORAL-ROM, fueron creados nuevos módulos: un tokenizador[15], módulos de

[15] Un tokenizador delimita automáticamente los elementos léxicos de un enunciado, segmentando la cadena en tokens o elementos individuales —en este caso, palabras.

desambiguación y un reconocedor de palabras desconocidas (Moreno y Guirao 2003). Dichos módulos se crearon para el español, con la intención de adaptar el procesador morfológico a las características propias de la lengua oral.

La base de datos léxica de GRAMPAL cuenta con unas 50.000 entradas entre raíces, sufijos y multipalabras. Todas las entradas de diccionario, esto es, exceptuando a los morfemas dependientes y a las multipalabras, poseen información morfológica y categorial, incluyendo sus respectivos lemas. Además, GRAMPAL cuenta con 239 prefijos que forman nuevas palabras sin que se produzca un cambio de categoría. La tabla 3.1 muestra los datos exactos de los tres tipos fundamentales de entradas en GRAMPAL.

Sufijos		**Raíces**		**Multi-words**	
Morfemas nominales	4	N	25.426	ADV	507
Morfemas verbales	179	ADJ	11.290	PREP	349
		V	10.568	C	91
		ADV	189	INTJ	70
		P	109	Palabras extranjeras	30
		MD	74	Q	24
		PREP	40	ADJ	15
		C	26		
		POSS	26		
		REL	16		
		ART	5		
Total	**183**		**47.769**		**1086**

Tabla 3.2 Entradas en GRAMPAL (Moreno y Guirao, 2006)

A continuación, explicaremos cómo funciona GRAMPAL y cómo se realiza la anotación morfológica automática. El procedimiento del *etiquetador morfosintáctico* sería el siguiente (Moreno y Guirao, 2006):

- Detección de palabra desconocida.
- Procesamiento léxico: separación de clíticos y *portmanteau*.
- Reconocimiento de multipalabras: mediante un lexicón.
- Reconocimiento de palabras individuales.
- Reconocimiento de palabra desconocida.
- Fase 1 de desambiguación: gramática de restricciones basada en rasgos[16].
- Fase 2: etiquetador estadístico TNT.

[16] Del inglés feature-based Constraint Grammar

La primera fase de desambiguación categorial se hace con un modelo de reglas basadas en el contexto. Las reglas presentan la siguiente forma:

"word" -> <cat="X"> / _ <cat="Y">
"word" -> <cat="Z"> / <cat="W"> _

Y su lectura sería: a una palabra ambigua se le adjudica la categoría X en un contexto en el que se sitúe ante otra palabra con categoría Y. Así, la distribución oracional de una palabra ambigua determinaría su categoría. Estas reglas permiten añadir información no sólo categorial, sino también sobre lemas o posición inicial o final de enunciado:

"word" -> <cat="X"> / _ <lem="Y">
"word" -> <cat="X"> / _ #[17]

Además, existen reglas genéricas en las que no se hace referencia a una única palabra, sino que son válidas para cualquier palabra ambigua en español:

<cat="N">, <cat="V"> -> <cat="N"> / <cat="V"> _

Esta regla indica que ante la posibilidad de que una palabra pueda adquirir tanto la categoría N (nombre) como la categoría V (verbo), se le adjudicará de forma automática la categoría N cuando se sitúe a continuación de otra palabra con categoría V.

Las palabras que no sufren desambiguación se tratan con TNT, un etiquetador categorial que se basa en procedimientos estadísticos. En cuanto a las palabras desconocidas, pueden ser de diferente índole. Moreno y Guirao (2003), tras la evaluación del etiquetado de C-ORAL-ROM, obtuvieron un 8% de palabras desconocidas por ser extranjerismos, palabras que no se encontraban en el lexicón por diferentes causas: ser propias de la variante oral, errores de transcripción y neologismos. GRAMPAL permite corregir o añadir nuevas entradas de forma manual con facilidad.

Después de que todo el proceso automático haya concluido, es necesaria una revisión manual para comprobar posibles errores de etiquetado. La desambiguación manual se realiza, pues, sobre el *output* o texto de salida de GRAMPAL. A cada palabra ambigua se le añaden de forma automática todos los análisis posibles, dando prioridad a uno de ellos según el contexto (probabilidad). Cuando ninguna de las etiquetas sea correcta, se añade manualmente y se agrega esa nueva información a GRAMPAL. En el último apartado del presente epígrafe se explica con más detalle el proceso de revisión manual.

[17] El símbolo # representa los límites de un enunciado, es decir, tanto el inicio como el fin.

Como último paso, una vez que todo el corpus está revisado y tenemos la seguridad de que no existen ya errores de ningún tipo, se procede al entrenamiento del procesador estadístico con nuevos datos, lo que incrementará su fiabilidad. Un *etiquetador morfosintáctico* por probabilidad contextual suele producir un mayor número de acierto en lenguas con orden fijo que en aquellas cuyo orden es libre y poseen un sistema flexivo más amplio (por ejemplo, en inglés la eficacia suele ser mayor que en español). Por ello, el entrenamiento con un gran número de datos es necesario e incrementa su probabilidad de acierto.

3.1.3 Ejemplo en CHIEDE

El resultado final después de la revisión del *etiquetado morfosintáctico* es un fichero en formato XML donde encontramos el texto transcrito analizado morfosintácticamente y en donde cada una de las palabras que lo forman aparece con la información morfológica y gramatical correspondiente:

*TEA: a ver Daniel /// ¿ cuántos años tienes tú ?
*DAI: cinco ///

```
<p>
<f h="TEA" st="0.0" et="0.509" id="1">
<sf t="enu" id="1-1">
<w cat="MD" lem="a ver" id="1-1-1"> a ver </w>
<w cat="NPR" lem="Daniel" id="1-1-2"> Daniel </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="0.509" et="1.559" id="2">
<sf t="int" id="2-1">
<w cat="PUNCT" lem="¿" id="2-1-1"> ¿ </w>
<w cat="P" lem="cuántos" gen="masc" id="2-1-2"> cuántos </w>
<w cat="N" lem="año" gen="masc" num="plu" id="2-1-3"> años </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="2-1-4"> tienes
</w>
<w cat="P" lem="tú" num="sing" per="2" id="2-1-5"> tú </w>
<w cat="PUNCT" lem="?" id="2-1-6"> ? </w>
</sf>
</f>
</p>
<p>
<f h="DAI" st="1.559" et="2.26" id="3">
<sf t="enu" id="3-1">
<w cat="Q" lem="cinco" id="3-1-1"> cinco </w>
</sf>
</f>
</p>
```

[DAI5]

Esta anotación y su formato en XML nos permitirán con posterioridad extraer datos relevantes acerca de cualquier aspecto morfológico o categorial, ya sea de un individuo concreto o de uno de los grupos de edad.

3.2 Transcripción fonológica

No cabe duda de que la transcripción fonética de un texto es una tarea en la que se necesita invertir muchas horas de trabajo. Si la transliteración ortográfica ya consume gran parte del tiempo dedicado a la creación de un corpus, la transcripción fonética supondría cuando menos duplicar ese tiempo:

> "Not many examples of publicly available phonetically transcribed corpora exists at the time of writing. This is possible because phonetic transcription is a form of annotation which needs to be carried out by humans rather than computers. […]. Phonetic transcription is therefore a very time consuming task" (McEnery y Wilson, http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fra1.htm).

Como McEnery y Wilson afirman, es difícil encontrar corpus que incluyan una transcripción fonética por ser una tarea a la que muchos no se querrían enfrentar. Para evitar esto, en el LLI-UAM se ha creado un transcriptor fonético automático que transcribe textos completos en pocos segundos.

El proceso de transcripción fonética funciona de la misma forma que el etiquetador morfosintáctico. El transcriptor fonético se aplica al texto de forma automática, posibilitando la transcripción del mismo en poco tiempo. Con posterioridad, será necesaria una revisión manual que garantice su fiabilidad y corrija posibles errores. La posibilidad de contar con un transcriptor fonético automático hace factible la tarea de la transcripción fonética en una gran cantidad de textos; la realización de dicha tarea de forma manual sería muy costosa sobre un corpus de 60.000 palabras. De esta forma, contaríamos con dos archivos paralelos: uno de ellos contendría la transcripción ortográfica y el otro la transcripción fonética. A continuación presentamos un ejemplo de ambas:

Transcripción ortográfica

*JOR: aquí ///
*TEA: a ver si puedes /// ¿ cuántos años tienes Jorge ?
*JOR: &eh tengo -> / cuatro ///
*TEA: cuatro /// que fue tu cumple el otro día /// ¿ a que sí ?
*JOR: cinco sí ///
*TEA: ¡ah! ¿ cinco ? ¿ o cuatro ?
*JOR: bueno / hoy &cum [/] mañana cumplí cinco // pero ahora / tengo cuatro ///

*TEA: ¡ah! ¿ y qué has hecho en tu cumple ?
*JOR: ¿ qué he hecho en mi cumple ? ¡ abrir los regalos ! / ¿ cómo que qué he hecho en el cumple ?
*TEA: no / pero que [/] ¿ has hecho una fiesta en tu casa / o qué ?
*JOR: sí [/] no / mi mama sí / pero yo no ///

[JOR4]

Transcripción fonológica

*JOR: . a . k i+ . ///
*TEA: . a+ . . b e+ ɾ. . s i+ . . p u e+ . d e s . /// ¿ . k u a+ n . t o s . . a+ . ɲ o s . . t i e+ . n e s . . x o+ ɾ. x e . ?
*JOR: &eh . t e+ n . g o . -> / . k u a+ . t ɾ o . ///
*TEA: . k u a+ . t ɾ o . /// . k e+ . . f u e+ . . t u+ . . k u+ m . p l e . . e+ l . . o+ . t ɾ o . . d i+ . a . /// ¿ . a+ . . k e+ . . s i+ . ?
*JOR: . θ i+ n . k o . . s i+ . ///
*TEA: ¡ah! ¿ . θ i+ n . k o . ? ¿ . o+ . . k u a+ . t ɾ o . ?
*JOR: . b u e+ . n o . / . o+ i . &cum [/] . m a . ɲ a+ . n a . . k u m . p l i+ . . θ i+ n . k o . // . p e+ . ɾ o . . a . o+ . ɾ a . / . t e+ n . g o . . k u a+ . t ɾ o . ///
*TEA: ¡ah! ¿ . i+ . . k e+ . . a+ s . . e+ . t ʃ o . . e+ n . . t u+ . . k u+ m . p l e . ?
*JOR: ¿ . k e+ . . e+ . . e+ . t ʃ o . . e+ n . . m i+ . . k u+ m . p l e . ? ¡ . a . b ɾ i+ ɾ. . l o+ s . . r e . g a+ . l o s . ! / ¿ . k o+ . m o . . k e+ . . k e+ . . e+ . . e+ . t ʃ o . . e+ n . . e+ l . . k u+ m . p l e . ?
*TEA: . n o+ . / . p e+ . ɾ o . . k e+ . [/] ¿ . a+ s . . e+ . t ʃ o . . u+ . n a . . f i e+ s . t a . . e+ n . . t u+ . . k a+ . s a . / . o+ . . k e+ . ?
*JOR: . s i+ . [/] . n o+ . / . m i+ . . m a+ . m a . . s i+ . / . p e+ . ɾ o . . y o+ . . n o+ . ///

Pero además, como se comentó en el apartado 1.2 de esta segunda parte, existía un problema con el lenguaje de alguno de los niños —aquellos que no superaban los tres años de edad—, que se encontraba en una etapa de evolución muy primitiva y por lo tanto no presentaba un sistema fonético completamente desarrollado. Teniendo esto en cuenta, y para que los datos finales fueran lo más realistas posible, se decidió hacer dos entradas para los turnos de estos niños: una de ellas introducida por las tres primeras letras de su nombre (según la convención de transcripción) y en la que se ha transcrito el texto de la forma normativa; y una segunda entrada marcada como *%pho*, en la que la transcripción se ha hecho de la forma más fiel posible a la pronunciación real. En estos casos, el transcriptor fonológico se aplica sobre la línea introducida por *%pho*. Todo ello se aprecia de forma clara en los ejemplos que se exponen a continuación:

Transcripción ortográfica

*BRU: la grúa ///
%pho: la gúa ///
*TEA: ¿ la grúa ?
*BRU: sí ///
%pho: ti ///
*TEA: ¿ y qué haces con ella ? ¿ tiene ruedas ?
*BRU: hhh ///
%act: negation
*TEA: ¿ no tiene ruedas ?
*BRU: tiene un mando ///
%pho: tene un mando ///
*TEA: ¡ah! / un mando +
*BRU: sube y baja ///
%pho: zube y baja ///
*TEA: ¡ah! ¿ y puedes coger cosas ? ¿ o no coge ?
*BRU: sí ///
%pho: ti ///
*TEA: ¿ las pones encima / o cómo ?
*BRU: las pongo encima ///
%pho: la pono encima ///

[BRU3]

Transcripción fonológica

*BRU: . l a+ . . g u+ . a . ///
*TEA: ¿ . l a+ . . g ɾ u+ . a . ?
*BRU: . t i+ . ///
*TEA: ¿ . i+ . . k e+ . . a+ . θ e s . . k o+ n . . e+ . ʌ a . ? ¿ . t i e+ . n e . . r u e+ . d a s . ?
*BRU: hhh ///
%act: negation
*TEA: ¿ . n o+ . . t i e+ . n e . . r u e+ . d a s . ?
*BRU: . t e+ . n e . . u+ n . . m a+ n . d o . ///
*TEA: ¡ah! / . u+ n . . m a+ n . d o . +
*BRU: . θ u+ . b e . . i+ . . b a+ . x a . ///
*TEA: ¡ah! ¿ . i+ . . p u e+ . d e s . . k o . x e+ ɾ. . k o+ . s a s . ? ¿ . o+ . . n o+ . . k o+ . x e . ?
*BRU: . t i+ . ///
*TEA: ¿ . l a+ s . . p o+ . n e s . . e n . θ i+ . m a . / . o+ . . k o+ . m o . ?
*BRU: . l a+ . . p o+ n . o . . e n . θ i+ . m a . ///

3.3 Revisión y desambiguación

El siguiente paso, una vez que tanto el etiquetado morfosintáctico como el transcriptor fonológico han hecho su trabajo, es revisar manualmente los resultados. Aunque el grado de fiabilidad de ambos es bastante elevado, es inevitable que se cometan ciertos errores. Y aunque la revisión debe ser manual, archivo por archivo y línea por línea, la mayor parte del trabajo ya se ha realizado de forma automática. Las causas de los errores responden a diversos factores; no obstante, en el caso de la anotación morfosintáctica uno de los mayores problemas con los que nos enfrentamos es la ambigüedad categorial. Por ello, los resultados de la anotación morfológica y gramatical se presentan en un formato determinado donde el revisor puede seleccionar la categoría adecuada de aquellas palabras que presentan más de una opción. La siguiente figura muestra dicho proceso (Moreno Sandoval y Guirao, 2006):

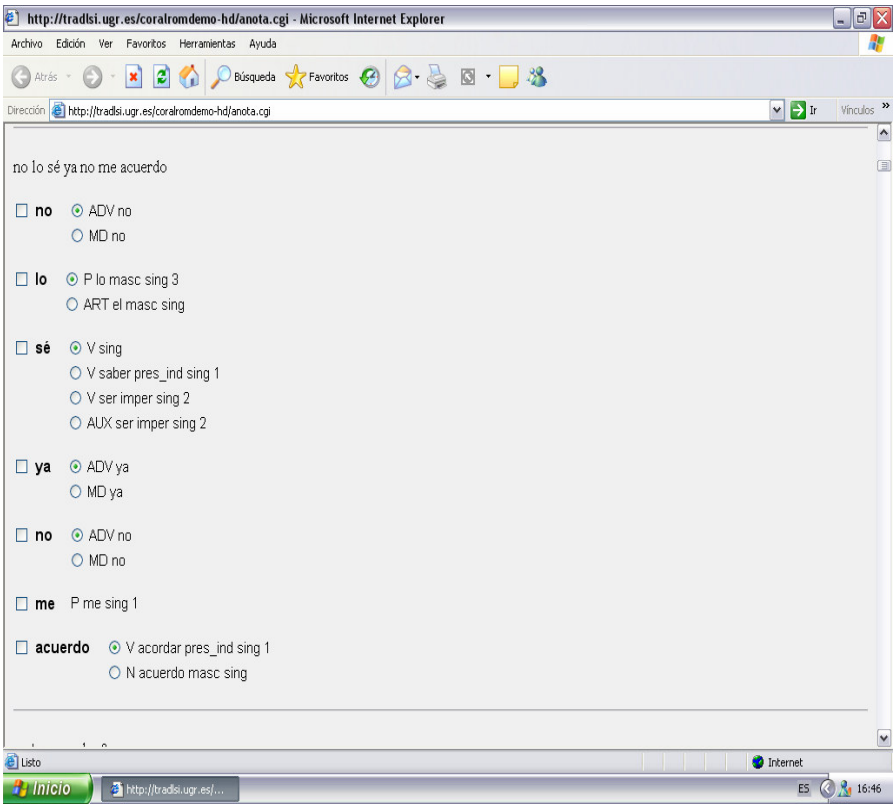


Figura 3.3 Revisión del etiquetado morfosintáctico

La figura muestra el texto desglosado en palabras y al lado de cada una de ellas su correspondiente información categorial y morfológica. Como podemos observar, una palabra como el pronombre personal *me* no presenta ninguna complicación. Sin embargo, el pronombre *lo* también puede

funcionar como un artículo. El programa sitúa las distintas opciones por orden de probabilidad; no obstante, si la primera de las opciones fuera errónea, se marcaría la casilla de la opción correcta.

En el caso del transcriptor fonológico, la mayoría de los errores son causados por palabras extrajeras, en muchos casos nombres propios. Todos estos casos se detallan más adelante en la sección dedicada a la evaluación del programa (capítulo 4). La revisión se hizo a partir de un listado de todas las palabras que aparecen en el corpus, al que se le aplicó el transcriptor fonológico. De esta forma, garantizamos la correcta transcripción de todas y cada una de las palabras sin necesidad de revisar los textos uno por uno, salvo en el caso de los niños con fonología prematura.

4. Resultados

En este capítulo presentamos dos tipos de resultados: por una parte, los relativos a la evaluación de los procesos automáticos, es decir, el etiquetado morfosintáctico y el transcriptor fonológico; por otra, los datos lingüísticos extraídos del corpus, facilitados por la anotación que se ha realizado previamente.

4.1 Evaluación del etiquetado morfosintáctico

La evaluación del etiquetador morfológico automático o *tagger* nos sirve para comprobar su grado de fiabilidad y, atendiendo a los errores cometidos, mejorar la herramienta. Como se explicó en el capítulo anterior, las correcciones se hacen de forma manual, pero todas ellas quedan grabadas en el programa de corrección de forma que posteriormente los datos se extraen automáticamente. En la tabla 4.1, se presenta el porcentaje de acierto y de error del etiquetado, facilitando también los datos en términos cuantitativos[18]:

	Nº de palabras	**Porcentaje**
Etiquetado correcto	55911	96,30%
Fallos en palabras simples	1337	2,30%
Fallos en multipalabras	560	0,96%
Palabras desconocidas	261	0,44%
Fallos de transcripción	51	0.08%

Tabla 4.1 Evaluación del etiquetado de corpus

Siguiendo los valores presentados en la tabla, el porcentaje total de error en la anotación del corpus sería del 3,78%. No obstante, hay que hacer una distinción entre tres tipos de errores:

- Los fallos cometidos por el etiquetador morfosintáctico se deben a la ambigüedad categorial que presentan algunas palabras y que el programa no es capaz de diferenciar por su falta de manejo de un contexto amplio. En estos casos es necesaria la elección del lingüista y la corrección manual. Este tipo de fallo agrupa la mayoría de los errores, con un porcentaje del 3,26%.
- Las carencias en GRAMPAL (palabras desconocidas) no pueden considerarse errores del *tagger,* ya que son entradas que no se contemplan en el diccionario y que, por tanto, son fallos de cobertura, que no superan el 0,44% del total de errores.

[18] Hemos seguido el método de evaluación expuesto en Moreno y Guirao (2006).

- Los errores debidos a descuidos durante el proceso de transcripción, es decir, a fallos ortográficos cometidos por el lingüista al teclear, tampoco pueden incluirse en la evaluación del *tagger,* porque son descuidos humanos.

Las conclusiones que se pueden extraer, por tanto, en cuanto al porcentaje de fallos en el corpus, son las siguientes:

	Porcentaje de error		
	Máquina		**Humano**
	Tagger	**Cobertura GRAMPAL**	**Transcripción**
	3,26%	0,44%	0,08%
Total	**3,78%**		

Tabla 4.2 Porcentaje de error en CHIEDE

Si eliminamos los errores de cobertura y los humanos, obtenemos un porcentaje de acierto del *tagger* de un 96,74%, una cifra bastante satisfactoria teniendo en cuenta que un programa similar para la lengua inglesa como CLAWS (Constituent Likelihood Automatic Word-tagging System)[19] ha conseguido un grado de precisión del 96-97%, del que un 3.3% de los fallos se deben a la ambigüedad categorial. La gran mayoría de los programas de etiquetado morfosintáctico están en torno al 95% de índice de acierto.

Además, debemos decir que GRAMPAL incluye dos elementos que la gran mayoría de *taggers* no contemplan y que incrementan el riesgo de fallo: las multipalabras y la categoría gramatical marcador discursivo (MD). Si consideramos que la secuencia "es que" siempre se va a categorizar como V C (verbo y conjunción), el etiquetador no va a producir ningún error; sin embargo, dicha secuencia de palabras no tiene el mismo significado en los dos enunciados siguientes:

> *EST: yo le invité un día a Laura a mi casa /// lo que pasa **es que** no vino ///
> *TEA: que a lo mejor no pudo /// **es que** a veces la mamá ... ¿ tú se lo dijiste a su mamá ?
>
> [A5-01]

Mientras que en el primero de los enunciados, "es" tiene función de verbo y "que" de conjunción, en el segundo, "es que" funciona como una

[19] CLAWS ha sido desarrollado en e l University Centre for Computer Corpus Research on Language (UCREL) en Lancaster y se ha aplicado al British Nacional Corpus (http://ucrel.lancs.ac.uk/claws/).

única unidad de significado y su función es la de marcador discursivo relacional de justificación[20]. Lo mismo ocurre con la palabra "bien". Podemos seguir el criterio tradicional y categorizarla como adverbio (ADV); no obstante, si observamos los dos ejemplos que se exponen a continuación, vemos que su valor semántico, sintáctico y pragmático no es el mismo en ambos casos:

> *TEA: **bien** /// por aquí veo manos alzadas // a ver ///
> *TEA: de baloncesto /// ¿ bota **bien** ?
>
> [A4-02]

En el segundo de los ejemplos, "bien" funciona como ADV, modificando al verbo al que acompaña, pero esto no ocurre en el primero, donde "bien" tiene una clara función de marcador discursivo.

En definitiva, introducir en el análisis categorial elementos como las multipalabras o la categoría MD supone incrementar la ambigüedad y por tanto el riesgo de error. Moreno Sandoval y Guirao (2006) describen exhaustivamente este problema y estiman que el tratamiento de dichos elementos puede hacer variar un 8-10% el resultado de la precisión. Pero al mismo tiempo, se garantiza un análisis más exhaustivo y por ello consideramos que es necesario optar por la inclusión de esos elementos. Como se puede apreciar en la tabla 4.1, el incremento del índice de error por causa de las multipalabras es de un 0,96%.

En cuanto a los errores de análisis categorial, la tabla 4.3 muestra los más problemáticos. Los puestos primero y cuarto de la lista los ocupa la palabra "que", a causa de su ambigüedad categorial, junto con su alta frecuencia de uso en nuestra lengua. Moreno y Guirao (2003) presentaron una evaluación del *tagger* aplicado al subcorpus de español incluido en C-ORAL-ROM, y entre los principales errores de categorización aparecían la palabra "que" y los marcadores discursivos. Son dos de los elementos más problemáticos en nuestra lengua y que por ahora requieren de una revisión manual por parte de los lingüistas.

[20] Siguiendo la tipología de marcadores discursivos en González-Ledesma y Garrote (2007), expuesta en el capítulo 5 del presente trabajo.

Errores por cambio de categoría		
Palabra	**Cambio de categoría**	**Frecuencia**
que	REL -> C	117
sí	MD -> ADV	79
este	DEM -> P	63
que	C -> REL	57
ese	DEM -> P	55
mira	MD -> V	33
pelota	ADJ -> N	23
esa	DEM -> P	21
ya	MD -> ADV	21
no	MD -> ADV	18
esta	DEM -> P	17
cumple	V -> N	11
nada	MD -> Q	10
estos	DEM -> P	6
esos	DEM -> P	6
estas	DEM -> P	6
sabes	MD -> V	6
bueno	MD -> ADJ	5
bien	MD -> ADV	4

Tabla 4.3 Principales errores de categoría

Volviendo a la lista de los principales errores, observamos que la gran mayoría de ellos se reducen a dos elementos: los deícticos por una parte y los marcadores discursivos por otra. En el caso de los deícticos, el *tagger* ha seleccionado como categoría más probable el demostrativo (DEM). Pero tras la corrección manual hemos visto que en muchos casos palabras como "este" o "ese" han tenido que recategorizarse como pronombre (P). Respecto a los marcadores discursivos, se ha tenido que hacer la recategorización manual de MD a la categoría tradicional. Si tenemos en cuenta que GRAMPAL está entrenado en un corpus de lenguaje adulto, estos errores podrían deberse a dos hechos:

- El uso de los determinantes demostrativos es más común en la lengua adulta que en la infantil. Los niños tienden a hacer más uso de los pronombres que sustituyen a otras palabras a causa de sus carencias léxicas.
- Los marcadores discursivos pertenecen al nivel pragmático de la lengua, y éste requiere una mayor destreza lingüística. Como veremos en el capítulo quinto, ya a los tres años de edad aparecen los marcadores discursivos en la lengua. Sin embargo, su uso es mucho más moderado que en el lenguaje adulto.

Con esto último también está relacionado el 0,44% de fallo de cobertura de GRAMPAL. Todas las palabras que no se encontraban en el diccionario son propias del lenguaje infantil —"chuches", "cole", "mami", "chuli"— y han tenido que ser agregadas con posterioridad a la revisión manual.

A continuación, presentamos un resumen de la tabla 4.3, agrupando los tipos de errores de categoría por ambigüedad. De esta forma podemos observar que el cambio de relativo a conjunción pasa de la primera a la tercera posición, pasando a ser los errores más frecuentes el cambio de demostrativo a pronombre y, en segundo lugar, el de marcador discursivo a adverbio.

Errores de categoría por ambigüedad	
Cambio de categoría	**Frecuencia**
DEM -> P	174
MD -> ADV	122
REL -> C	117
C -> REL	57
MD -> V	39
ADJ -> N	23
V -> N	11
MD -> Q	10
MD -> ADJ	5

Tabla 4.4 Errores de categoría por ambigüedad

Como el *tagger* está entrenado sobre la lengua adulta, nos ofrece una pista inversa de la recategorización. Es decir, los niños usan antes el pronombre que el demostrativo, el adverbio antes que el marcador discursivo, o la recategorización de nombre a adjetivo o verbo. Estos datos confirman hechos ya conocidos, pero aportando datos cuantificados.

4.2 Evaluación del transcriptor fonológico[21]

El transcriptor fonológico automático se aplica a la transcripción ortográfica, obteniendo en unos minutos una transcripción fonológica completa del corpus, además de la silabificación de las palabras que lo componen. Las ventajas del transcriptor fonológico son principalmente dos:

- La obtención de los textos en formato fonológico de una forma rápida y eficaz. La realización de la transcripción fonológica de forma

[21] Moreno Sandoval, *et al.* (2006).

manual consumiría una gran cantidad de tiempo cuando trabajamos con corpus de un tamaño considerable.

- La posibilidad de extraer todo tipo de información fonológica: frecuencia de fonemas y sílabas, longitud media de enunciado con respecto al número de fonemas y al número de sílabas, etc.

Para evaluar la fiabilidad del transcriptor, se ha revisado de forma manual la transcripción fonológica automática de la lista de formas diferentes que aparecen en todo el corpus. El número total de fallos en términos cuantitativos es de 35, es decir, un 0.06% de error. Sin embargo, todos los casos en los que el transcriptor ha fallado se corresponden con palabras extranjeras, la mayoría de ellas nombres propios. A continuación presentamos las 34 formas en las que el transcriptor fonológico ha fallado:

Spiderman	Mickey Mouse	American Dragon
hamster	Bugs Bunny	Donald
Wendy	Phoskitos	Barbie
block	Rompehuesos	Goofy
Giuly	Messi	King Kong
hello	Hansel	Minnie
Gretel	Kinder	Daisy
Larsson	Belletti	Ronaldinho
Shin-Chan	Halloween	Zidane
Playmobil	Cassano	Eto'o
Playmobiles	Gameboy	
Play	Jorquera	

Nuestro transcriptor está pensado y diseñado para la ortografía del castellano, por lo tanto es normal que cometa errores con palabras que pertenecen a otra lengua. Según esto, los fallos cometidos no se pueden considerar como tal, al no haberse producido en palabras del castellano. Así pues, se puede concluir que el margen de error del transcriptor fonológico, al menos en su aplicación a nuestro corpus, es del 0%. Éste es un resultado esperable, ya que el sistema fonológico de los niños de 3 a 5 años es más simple que el del adulto.

4.3 Listado de frecuencias de las unidades

El proceso de diseño, recogida de muestras, transcripción y revisión es la etapa que más tiempo consume en de la elaboración de un corpus de lengua oral. Una vez que todo ese trabajo está hecho, podemos empezar a extraer diferentes tipos de información del corpus para analizar fenómenos lingüísticos variados. En nuestro caso, al contar con un corpus anotado categorial y morfológicamente por un lado, y transcrito fonológicamente por

otro, podemos obtener fácilmente la información relativa a estos dos niveles lingüísticos.

Una de las posibilidades que nos permite la tecnología actual es la de extraer datos de un texto de forma automática. Sin la posibilidad de utilizar un ordenador, la tarea de recuento de frecuencias tendría que hacerse de forma manual con dos grandes inconvenientes: por una parte, trabajar con grandes cantidades de texto eternizaría la tarea; por otra, en este tipo de tareas automáticas el índice de acierto de una máquina es superior al del humano, ya que éste último puede equivocarse con facilidad mientras que la máquina realiza las órdenes dadas de forma estricta. El humano es más creativo para resolver problemas difíciles, pero el ordenador es más sistemático y rápido, y, sobre todo, cuenta mejor.

El tipo de anotación que contienen los textos nos permite extraer información de distinta índole. Por una parte, las transcripciones contienen información sobre las características sociolingüísticas de los participantes y sobre el contexto situacional. Por otra, las interacciones transcritas están estructuradas según turnos dialógicos, enunciados e interrupciones prosódicas. Todo ello, junto con la anotación morfosintáctica y fonética, nos facilita extraer datos generales sobre las interacciones, al igual que de forma individual sobre un solo participante o sobre uno de los subcorpus o una única transcripción.

La información extraída del texto nos permite contestar preguntas como cuántas palabras utilizan los hablantes de nuestro corpus, qué palabras son esas y con cuánta frecuencia aparecen, qué tipo de uso se hace de una palabra o una categoría concreta, etc. Los estudios cuantitativos nos dan la información empírica necesaria para establecer teorías generales sobre el lenguaje y el uso que de éste hacen los hablantes.

Respecto al tipo de información que se puede extraer de un corpus anotado, dependerá siempre de la clase de corpus que estemos manejando y de la finalidad de la investigación. Ya se comentó que tanto la tipología de corpus como las clases de anotación de los mismos responden a los intereses concretos de un estudio determinado, al igual que los datos que se extraigan del corpus. Éstos variarán de un proyecto a otro dependiendo del diseño del corpus y las anotaciones que contenga.

En el caso de nuestro corpus, CHIEDE, algunos de los datos que se pueden proporcionar de forma automática son los siguientes:

- Longitud media de enunciado: por sílabas y fonemas.
- Frecuencia de uso de lemas y formas.
- Ratio type/token: media de diversidad léxica.
- Palabras más frecuentes.

- Categorías más frecuentes.

Todos estos datos nos permitirán realizar descripciones reales sobre el uso de la lengua y establecer patrones de actuación lingüística.

4.3.1 Datos extraídos del etiquetador morfosintáctico

La información que se puede extraer automáticamente del etiquetador morfosintáctico atañe a tres niveles lingüísticos: el morfológico, el sintáctico y el léxico. En nuestro caso, hemos utilizado la información morfológica que se incluye en las entradas del lexicón para obtener los diferentes lemas que aparecen en el corpus; y la información categorial que se añade a cada palabra o multipalabra.

Formas y Lemas por archivos			
Archivo	**Palabras**	**Formas**	**Lemas**
A3-01	3877	607	436
A3-02	3024	532	366
A4-01	4284	756	510
A4-02	3700	742	532
A5-01	5657	820	576
A5-02	6838	1022	711
ADI4	857	255	201
ADR3	1057	212	171
AIT4	1527	418	347
ANG4	1210	329	269
BRU3	2147	421	319
CAL5	1046	319	250
CAR5	911	284	231
CLA3	1470	361	286
DAI5	1154	314	250
ELE3	876	234	184
INE5	1093	360	282
JAV3	1180	327	237
JOM4	1607	403	314
JOR4	2061	374	288
MAI4	1236	324	244
MAR4	940	237	199
MJO4	1362	362	262
NAT3	1410	307	230
RAU5	1630	418	329
ROD3	1288	299	232
SEL3	1366	313	236
SOL5	1247	381	282
TAL4	964	270	225
VIC5	1144	315	248

Tabla 4.5 Lemas y formas por archivo

En primer lugar, presentamos en la tabla 4.5 el número de palabras totales, lemas diferentes y formas distintas que se incluyen en cada una de las transcripciones de nuestro corpus. Las cifras son bastante parecidas, siendo más elevadas en los seis primeros archivos, ya que éstos son los pertenecientes a las interacciones colectivas.

En la siguiente tabla encontramos la misma información que en la anterior, pero en este caso los datos hacen referencia al corpus en su totalidad —incluyendo el lenguaje adulto— y también por grupos de edad:

Formas y Lemas por edades				
	Palabras totales	**Formas diferentes**	**Lemas diferentes**	**Ratio forma/lema**
Adultos	36.905	2.910	1.804	1,61
Grupo 3	5.713	985	718	1,37
Grupo 4	6.374	1.155	839	1,38
Grupo 5	8.993	1.450	1.056	1,37

Tabla 4.6 Lemas y formas por grupos de edad

En esta última tabla 4.6 se incluye en la primera columna el número total de palabras que forman cada grupo, en la segunda columna, el número de formas distintas que aparecen y en la tercera, el número de lemas diferentes. El número total de palabras del subcorpus infantil es 21.080.La última columna indica la media de diversidad léxica, es decir, la proporción de formas distintas existentes por cada lema. Esta última cifra no varía apenas en lo que se refiere a los tres grupos de menores; sí lo hace en el recuento general y en el grupo de adultos, y esto es debido a que en el lenguaje adulto aumenta la flexión de las palabras, es decir, el número de formas por lema.

Sin embargo, sí pueden apreciarse las diferencias entre los tres grupos de edad en cuanto al incremento tanto de formas como de lemas. Se observa una progresión desde los tres a los cinco años, fruto del paulatino aumento de léxico. Durante el periodo de los tres a los cuatro años se observa un incremento de 170 formas y 121 lemas; de los cuatro a los cinco años, la cifra aumenta a 295 formas y 217 lemas. Esto nos indica que durante el primero de estos periodos —de los tres a los cuatro años— el aprendizaje es un poco más lento que en el segundo de ellos.

También podemos observar que a pesar de que el número de horas de grabación es muy similar para los tres grupos, hay una diferencia de 3219 palabras entre el grupo de tres años de edad y el de cinco. Los niños de cinco años ya son capaces de mantener una conversación pseudoadulta, mientras que los de tres años tienen más dificultades para ello. Esto se aprecia de forma más clara calculando la longitud media de enunciado (LME) por fonemas y sílabas. En el siguiente apartado, dedicado a los datos

extraídos del transcriptor fonológico, veremos como la LME por fonemas es de 10,29 en el grupo de tres años, mientras que en el grupo de cinco esta cifra incrementa 3,82 puntos.

Frecuencia de formas (35 primeras formas)					
Grupo de edad: 3		**Grupo de edad: 4**		**Grupo de edad: 5**	
Forma	**Frecuencia**	**Forma**	**Frecuencia**	**Forma**	**Frecuencia**
y	322	y	341	y	502
no	271	que	213	a	289
sí	204	el	206	no	274
el	162	sí	195	que	259
yo	157	no	179	el	251
un	148	a	165	sí	248
la	141	la	153	la	217
a	128	de	148	me	189
me	119	en	136	de	172
se	106	mi	131	se	171
en	92	un	128	mi	156
una	83	me	115	un	155
de	82	yo	113	yo	147
lo	76	se	107	en	136
que	74	una	102	le	116
los	66	es	82	los	106
es	57	pues	72	una	106
con	48	los	71	lo	99
mi	47	pero	66	con	94
aquí	42	lo	58	pero	74
ya	42	para	56	porque	68
pero	41	tengo	50	pues	68
he	40	porque	47	es	67
tengo	40	con	47	Candi	66
ha	40	le	44	dos	59
qué	38	qué	38	ya	56
Guadalupe	37	uno	37	también	54
está	36	también	36	las	53
porque	35	las	34	sé	48
mira	32	es que	33	tiene	44
esto	32	cuando	32	está	43
por	32	ya	30	es que	42
le	32	madre	30	más	41
así	31	tiene	29	tengo	40
este	30	casa	29	para	40

Tabla 4.7 Frecuencia de formas por grupos de edad

En cuanto a la frecuencia de formas (tabla 4.7) destaca el abundante uso de la conjunción copulativa "y". Esto es debido a dos hechos: por una parte, es la primera forma de coordinación que aprenden los niños; por otra, "y" tiene un uso diferente al de nexo, y se utiliza con frecuencia como un marcador discursivo, no sólo relacional, sino también de subjetividad: sirve como estrategia para tomar el turno de palabra. El siguiente ejemplo muestra de forma clara esta última función de la conjunción copulativa:

*DAI: y [/] y yo sé +
*TEA: ¡ qué bonito ! y a ver María ///
*DAI: y mi &pa +
*TEA: espera / espera a María /// a ver María ///
*MRI: mi papá no le regala / nada a mi mamá ///
%alt: (6) na
*TEA: pero bueno +
*MRI: ¬ y mi mamá sí ///

[A5-02]

Además de esto, destaca la alta frecuencia de los adverbios negativo y afirmativo, y el uso de pronombres y determinantes. En definitiva, en cualquier corpus de lengua oral los primeros puestos de la lista de frecuencias van a estar ocupados por categorías gramaticales; no es hasta las últimas filas de esta selección de las primeras treinta y cinco formas cuando empiezan a aparecer categorías léxicas. Lo más llamativo es quizá la carencia de marcadores discursivos como "bueno", "¿sabes?" o "¿no?", tan frecuentes en la lengua oral adulta.

Si obviamos las categorías gramaticales, la adquisición de categorías léxicas como verbos, nombres o adjetivos es mucho más rápida de los cuatro a los cinco años de edad que de los tres a los cuatro (tabla 4.8). En el segundo período (4 a 5 años), la aparición de nuevos lemas aumenta más del doble que durante el primero de ellos. Por esta razón, y según lo descrito hasta ahora, podemos afirmar que durante entre el cuarto y quinto año de vida, los niños pasan por una etapa de explosión de vocabulario, es decir, que durante esta época la adquisición de categorías léxicas sufre un aumento considerable en comparación con los años anteriores.

		Verbos	Nombres	Adjetivos
Grupo de edad: 3	**Formas**	906	836	157
	Lemas	138	313	50
Grupo de edad: 4	**Formas**	937	998	141
	Lemas	143	357	57
Grupo de edad: 5	**Formas**	1324	1340	161
	Lemas	190	439	72

Tabla 4.8 Categorías léxicas por grupos de edad

En los dos gráficos siguientes, se puede apreciar de forma más clara el aumento de formas (gráfico 4.9) y de lemas (gráfico 4.10) en los tres grupos de edad.

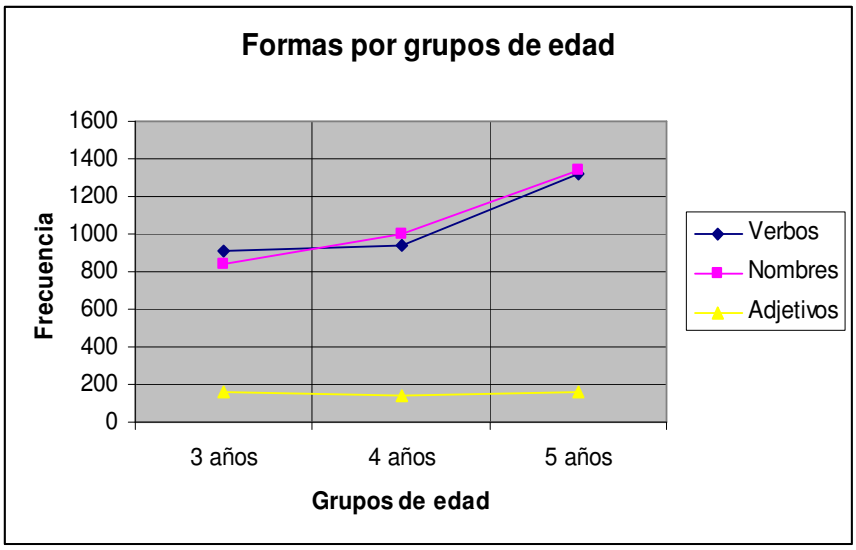


Gráfico 4.9 Formas por grupos de edad

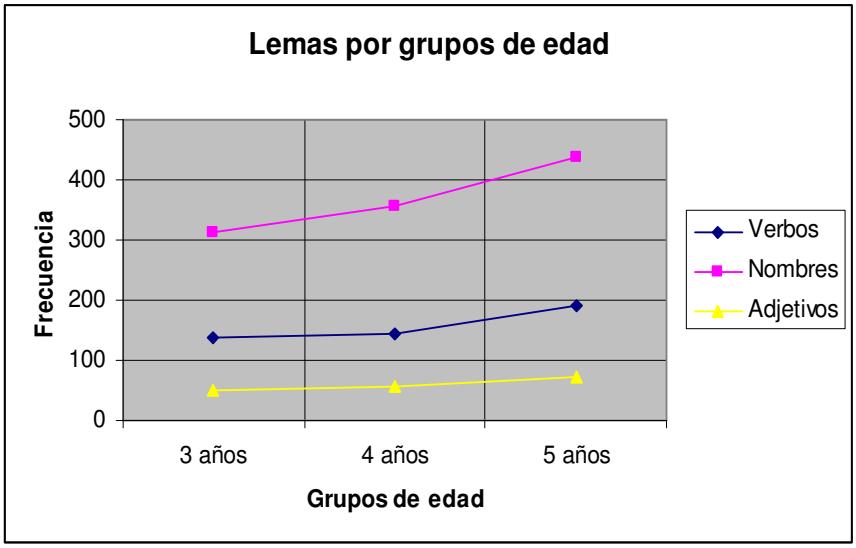


Gráfico 4.10 Lemas por grupos de edad

Otro hecho significativo de nuestro estudio, y que ya han apuntado con anterioridad otros autores (O'Keeffe, McCarthy & Carter, 2007) es que más de un 80% de la comunicación oral humana se compone de un número fijo y bastante reducido de palabras. El siguiente grafico muestra cómo con apenas

500 palabras se cubre un 80,15% de las interacciones en CHIEDE. Las segundas 500 palabras más frecuentes del corpus —de 500 a 1000— sólo abarcan un 7,25% de las interacciones; las siguientes 500 —de 1000 a 1500— suponen un 3,63% del total de palabras del corpus; y así sucesivamente. En el Apéndice B se incluye el listado completo de formas y lemas de todo el corpus por grupos de edad. En dichos listados, podemos comprobar cuáles son esas primeras 500 palabras y su frecuencia de uso.

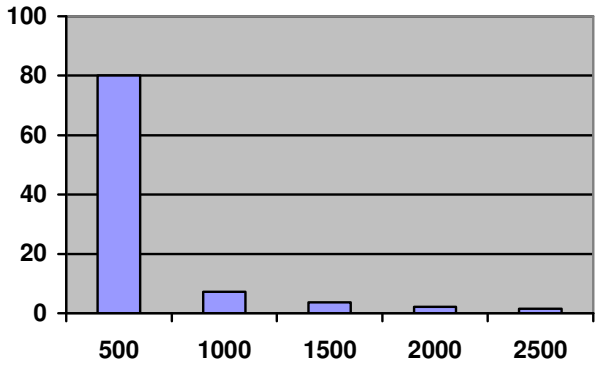


Figura 4.11 Porcentajes de frecuencia en grupos de 500 palabras

El último punto que vamos a tratar en esta sección es el de la hiperregulación propia del lenguaje infantil. Este fenómeno, también conocido como *errores por analogía*, se produce en los primeros años de adquisición, cuando el niño ya ha adquirido los conocimientos suficientes sobre morfología y flexión en nuestra lengua. Una vez aprendidas las reglas morfológicas generales de, por ejemplo, la conjugación de los verbos españoles, el niño las aplica por igual a todos los verbos, sin discriminar entre regulares e irregulares. La hiperregularización ha sido siempre uno de los principales argumentos de aquellos que critican las teorías conductistas o del aprendizaje por imitación. Argumentan que si el niño se limitase a reproducir lo que escucha, este tipo de errores no se producirían; por el contrario, es fácil encontrar una explicación a la hiperregularización si consideramos que el niño es capaz de aprender las reglas morfológicas de su lengua y aplicarlas de forma creativa, innovando y no imitando. De hecho, es curioso cómo ante la corrección del adulto, generalmente el niño no rectifica y sigue manteniendo su versión.

A continuación, presentamos la lista de errores por analogía que hemos extraído de nuestro corpus. En ella sólo hemos incluido los verbos, ya que son éstos los que mayor confusión producen por su naturaleza flexiva. En un corpus de 21.080 palabras se producen un total de 31 errores, esto es, un 0,15% del total. Si reducimos la lista de formas a lemas, vemos que son un total de 17 los que resultan más problemáticos: acordar, cerrar, conducir,

decir, escribir, hacer, ir, leer, mentir, morder, poder, poner, querer, ser, soltar, tener, traer.

Errores por analogía		
	Error	**Palabra**
Grupo 3 **(9 formas)**	ponío	puesto
	acorda	acuerda
	fió	fue
	muerdía	mordía
	ponió	puso
	dijió	dijo
	dició	dijo
	hació	hizo
	hací	hice
Grupo 4 **(17 formas)**	quedrán	querrán
	cierrar	cerrar
	dijí	dije
	hacimos	hicimos
	podimos	pudimos
	podí	pude
	fío	fue
	fió	fue
	trayí	traje
	leendo	leyendo
	mentió	mintió
	escribío	escrito
	ponío	puesto
	hació	hizo
	fi	fui
	podieron	pudieron
	hacieron	hicieron
Grupo 5 **(5 formas)**	tení	tuve
	hació	hizo
	sueltó	soltó
	conduzo	conduzco
	traí	traje

Tabla 4.12 Errores por analogía

Según nuestros datos, este fenómeno es más abundante entre los niños de cuatro años, reduciéndose en gran medida entre los niños de cinco. Parece ser que es a la edad de cinco años cuando los niños empiezan a comprender

la flexión irregular de los verbos del español. En nuestro corpus, la forma irregular más problemática y persistente en las tres edades es "hació/hizo".

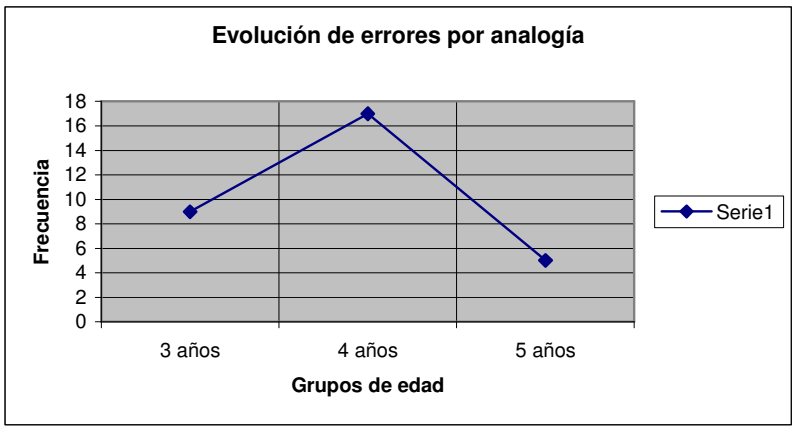


Gráfico 4.13 Evolución de errores por analogía

4.3.2 Datos extraídos del transcriptor fonológico

La psicolingüística, en su intento de dar una explicación al proceso de adquisición de una primera lengua, ha recurrido de forma habitual a establecer patrones de medición de distinta índole. De esta forma, se calculan las distintas etapas de evolución lingüística del niño, ya sea midiendo el número de palabras que forman su léxico a una determinada edad, el número de fonemas que se manejan o el tipo de sílabas más frecuentes.

La mayoría de los expertos coinciden en establecer el periodo de adquisición del sistema fonológico de una lengua desde los nueve meses hasta los cuatro años de edad. No obstante, y como ya hemos mencionado en alguna ocasión a lo largo de este trabajo, la mayoría de los estudios sobre lenguaje infantil no utilizan sujetos que superen los 36 meses de edad. Casi todos los trabajos consultados facilitan una descripción del lenguaje infantil que se detiene cuando el niño alcanza los tres años, a pesar de que insisten en que el proceso de adquisición continúa hasta el umbral de la pubertad. Por ello, nosotros consideramos que nuestro corpus puede ser de gran utilidad para la comunidad científica, ya que facilitamos los datos obtenidos de sujetos cuyas edades van desde los tres a los seis años de edad.

En la tabla 4.10 presentamos los datos fonológicos referentes a los tres grupos de edad que conforman nuestro corpus. El número total de fonemas que se han contabilizado es de 75.535, y la longitud media de enunciado, en el recuento general, es de 12,72 fonemas por turno.

Lo más llamativo de esta tabla es que a la edad de tres años el niño ya ha adquirido el sistema fonológico del español al completo —a excepción de tres sujetos del grupo de tres años para los que hubo que adaptar la transcripción. Nuestra tabla coincide en gran medida con los datos expuestos por Anula Rebollo (1998), según los cuales, los niños adquieren primero las vocales abiertas y las oclusivas bilabiales (sorda y sonora) entre los 9 y los 12 meses, después las vocales cerradas y el resto de oclusivas (sobre los 20 meses) y más tarde las nasales, fricativas y líquidas (en este orden) concluyendo el aprendizaje alrededor de los 30 meses.

En nuestro caso, la tabla que aportamos no representa el orden de aparición de los fonemas, ya que éstos ya están adquiridos por los sujetos que participan en el corpus, sino su frecuencia de uso. Es sabido que los niños tienden a usar más frecuentemente aquellos sonidos que dominan, mientras que evitan aquellos que les resultan más dificultosos. En la tabla de frecuencias, podemos observar cómo los fonemas que ocupan las últimas posiciones son los menos frecuentes en nuestra lengua y por tanto es lógico que su frecuencia de uso disminuya en relación a otros sonidos más habituales; no obstante, se puede apreciar cómo incrementan las cifras con el aumento de edad de los niños. Esto nos demuestra que desde los tres a los cinco años de edad el proceso de adquisición lingüística aún es muy visible y por ello los estudios sobre adquisición de la lengua materna no deben detenerse a los 36 meses.

Además de los datos fonológicos extraídos de CHIEDE, hemos agregado a la tabla 4.14 una columna más que incluye las frecuencias de fonemas en C-ORAL-ROM (Moreno *et al.*, 2006). A pesar de que los datos son similares para el lenguaje infantil y para el adulto, se observa, sobre todo en los primeros puestos de la tabla, una mayor similitud entre el grupo de cinco años y el grupo de adultos, que entre éste último y los grupos de tres y cuatro años.

Grupo de edad: 3			Grupo de edad: 4			Grupo de edad: 5			C-ORAL-ROM		
Fonema	**Frecuencia**		**Fonema**	**Frecuencia**		**Fonema**	**Frecuencia**		**Fonema**	**Frecuencia**	
	Absoluta	**Relativa**		**Absoluta**	**Relativa**		**Absoluta**	**Relativa**		**Absoluta**	**Relativa**
a	2616	13,29	e	3190	13,44	e	4360	13,58	e	188196	15,12
e	2539	12,90	a	3108	13,09	a	4215	13,13	a	152664	12,27
o	2300	11,68	o	2484	10,46	o	3472	10,81	o	129208	10,38
i	1656	8,41	i	1922	8,10	s	2551	7,94	s	100881	8,11
n	1447	7,35	s	1792	7,55	i	2550	7,94	i	89799	7,22
s	1444	7,34	n	1709	7,20	n	2159	6,72	n	87775	7,05
l	884	4,49	l	1139	4,80	ɾ	1553	4,84	ɾ	63702	5,12
ɾ	869	4,41	ɾ	1130	4,76	l	1523	4,74	t	56287	4,52
t	793	4,03	k	1041	4,39	k	1472	4,58	l	56107	4,51
u	765	3,89	u	927	3,91	m	1258	3,92	k	55863	4,49
k	732	3,72	m	914	3,85	t	1214	3,78	d	54284	4,36
m	681	3,46	t	885	3,73	d	1134	3,53	m	39278	3,15
p	625	3,18	d	739	3,11	u	1134	3,53	u	39146	3,14
d	524	2,66	p	709	2,99	p	871	2,71	p	34135	2,74
b	477	2,42	b	644	2,71	b	777	2,42	b	31126	2,50
ʎ	344	1,75	g	308	1,30	ʎ	361	1,12	θ	18940	1,52
g	296	1,50	ʎ	274	1,15	g	342	1,06	g	11359	0,91
x	166	0,84	x	221	0,93	θ	330	1,03	ʎ	10356	0,83
θ	165	0,84	θ	210	0,88	x	214	0,67	x	7681	0,62
tʃ	125	0,64	tʃ	106	0,45	tʃ	200	0,62	f	6217	0,50
r	111	0,56	r	101	0,43	r	199	0,62	r	5236	0,42
f	68	0,35	f	95	0,40	f	126	0,39	tʃ	3744	0,30
ɲ	58	0,29	ɲ	89	0,37	ɲ	98	0,31	ɲ	2427	0,19
Total	19685	100,00		23737	100,00		32113	100,00		1244411	100

Tabla 4.14 Frecuencia de fonemas por grupos de edad

Para poder apreciar de manera más clara el funcionamiento del sistema fonológico infantil —al menos el presentado por los sujetos que han participado en nuestro corpus—, hemos realizado cinco gráficos independientes. El primero de ellos incluye la frecuencia de las vocales y su evolución de un grupo de edad a otro. Los siguientes recogen la frecuencia de consonantes, divididas a su vez en cuatro grupos, según su modo de articulación: oclusivas, nasales, líquidas y fricativas.

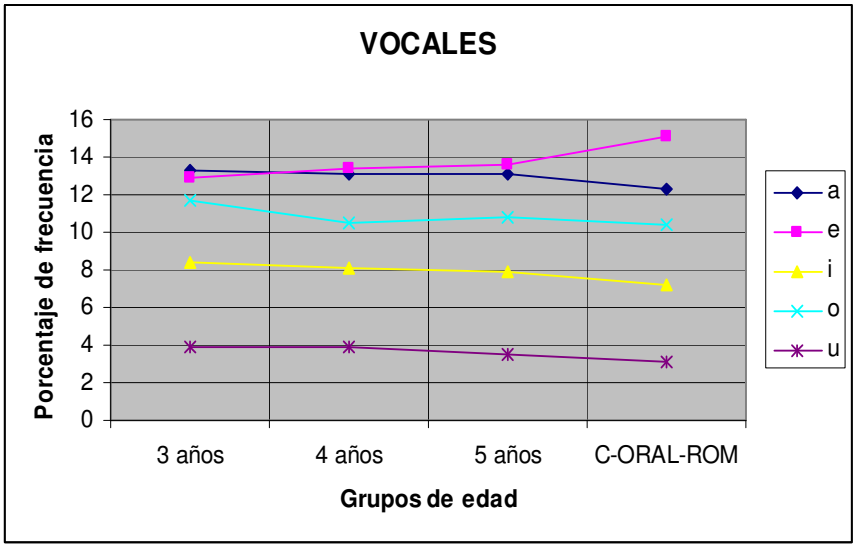


Gráfico 4.15 Vocales por grupos de edad

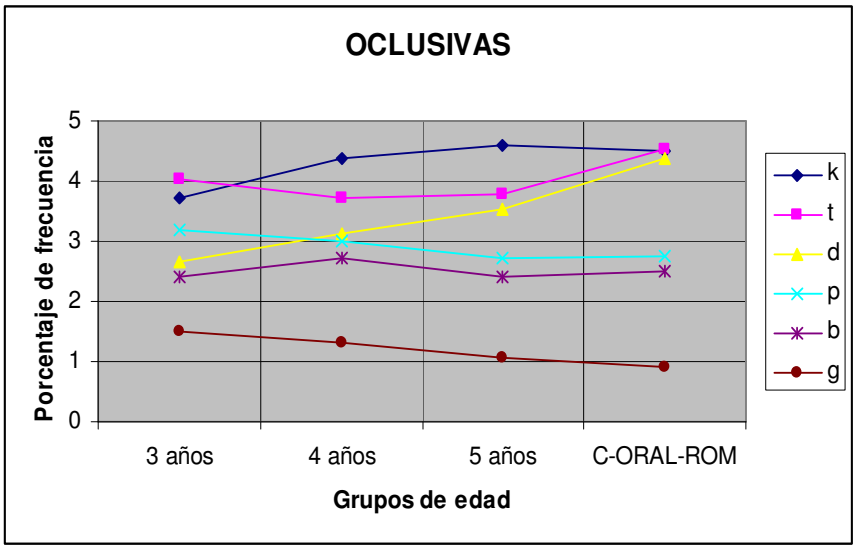


Gráfico 4.16 Oclusivas por grupos de edad

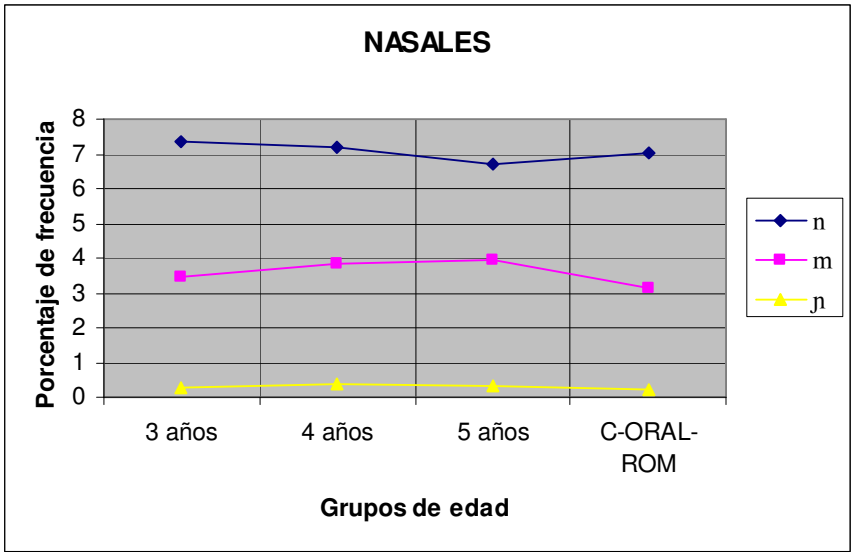


Gráfico 4.17 Nasales por grupos de edad

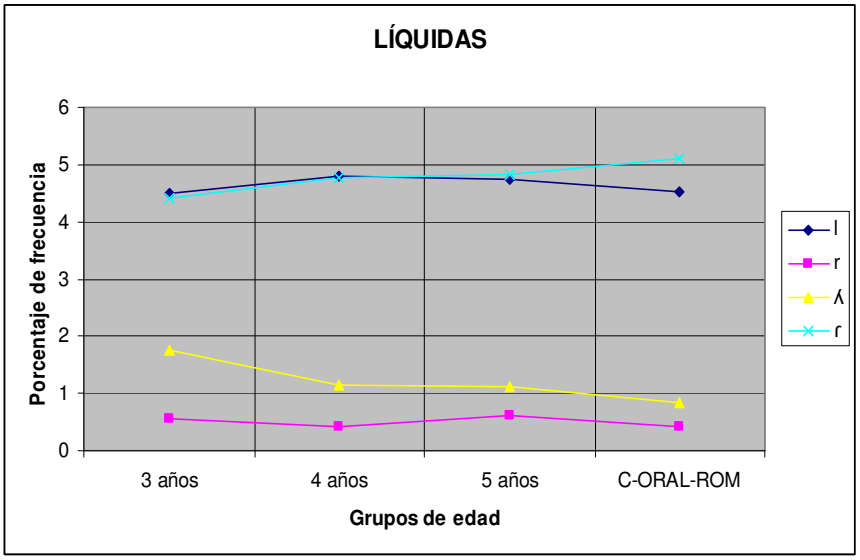


Gráfico 4.18 Líquidas por grupos de edad

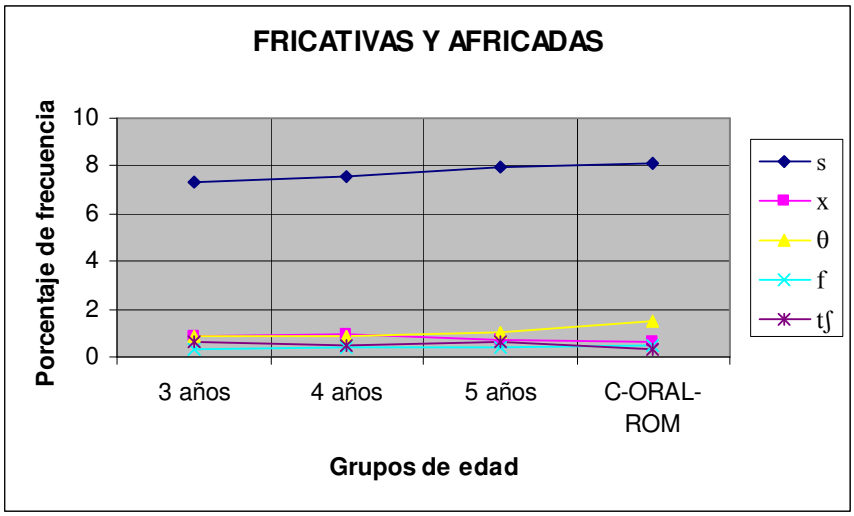


Gráfico 4.19 Fricativas por grupos de edad

Otra de las posibilidades que nos ofrece el transcriptor fonológico automático es la segmentación de las palabras en sílabas. De esta forma, podemos saber con rapidez y fiabilidad el número total de sílabas que forman nuestro corpus, cuáles son esas sílabas y cuál es su frecuencia de uso. El número total es de 35.086, y la longitud media de enunciado es de 5,91 sílabas.

A modo de ejemplo, hemos seleccionado las 30 primeras sílabas del listado completo de cada grupo de edad. En los tres casos, observamos cómo las sílabas no llegan a los tres fonemas, y en su mayoría siguen el patrón CV. Las sílabas trabadas (CVC) o los grupos consonánticos del tipo CCV suponen un grado de dificultad articulatoria mayor y por tanto su frecuencia de uso desciende en comparación con las sílabas abiertas compuestas por no más de dos fonemas.

Grupo de edad: 3		Grupo de edad: 4		Grupo de edad: 5	
Sílaba	**Frecuencia**	**Sílaba**	**Frecuencia**	**Sílaba**	**Frecuencia**
.a.	490	.a.	504	.a.	707
.i.	379	.ke.	395	.i.	594
.no.	300	.i.	391	.ke.	463
.si.	279	.no.	280	.no.	381
.ke.	208	.si.	250	.si.	318
.es.	202	.mi.	231	.es.	288
.la.	193	.la.	223	.se.	283
.se.	189	.es.	214	.la.	273
.yo.	185	.de.	204	.de.	269
.ta.	174	.na.	188	.ta.	243
.e.	160	.se.	187	.me.	242
.na.	156	.pa.	180	.do.	239
.me.	156	.u.	176	.ko.	233
.un.	151	.ma.	169	.na.	227
.el.	149	.el.	169	.mi.	221
.do.	148	.ta.	156	.ka.	208
.de.	147	.ra.	155	.el.	205
.ko.	139	.me.	154	.le.	202
.pe.	138	.ko.	154	.ma.	201
.to.	131	.un.	154	.en.	191
.lo.	130	.en.	152	.pa.	189
.te.	124	.do.	152	.lo.	189
.ma.	123	.lo.	144	.ba.	186
.ba.	121	.ka.	142	.di.	182
.mi.	115	.yo.	142	.yo.	177
.en.	114	.pe.	136	.to.	166
.ra.	114	.o.	125	.un.	162
.ka.	113	.e.	121	.u.	162
.pa.	110	.to.	117	.pe.	158
.o.	104	.ba.	116	.ro.	156

Tabla 4.20 Frecuencia de sílabas por grupos de edad

Con estos datos, podemos calcular fácilmente y con exactitud la Longitud Media de Enunciado (LME) por fonemas y sílabas en cada grupo de edad. En la siguiente tabla podemos ver las cifras exactas y cuál es el porcentaje de incremento desde los tres a los cinco años. En el Apéndice C se incluye el listado completo de sílabas extraídas del corpus con sus respectivas frecuencias de uso, por grupos de edad.

	Longitud Media de Enunciado	
	Fonemas	**Sílabas**
Grupo de edad: 3	10,29	4,88
Grupo de edad: 4	13,57	6,26
Grupo de edad: 5	14,11	6,49

Tabla 4.21 Longitud Media de Enunciado

El gráfico 4.22 muestra de forma más clara el incremento de la LMD desde los tres a los cinco años de edad.

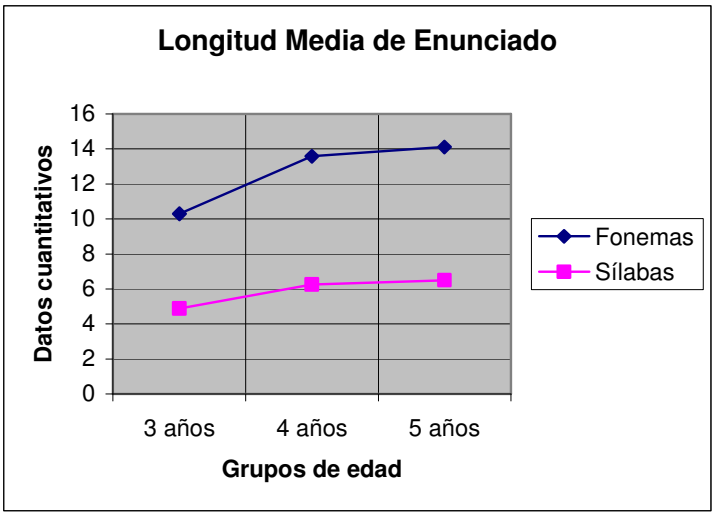


Gráfico 4.22 Incremento de la LME

4.4 Comparación de CHIEDE con C-ORAL-ROM

Por último, dentro del presente apartado dedicado a los datos obtenidos del corpus, presentamos una pequeña comparación entre un corpus de lengua infantil, CHIEDE, y otro de lenguaje adulto, C-ORAL-ROM. Los datos relativos a éste último proceden del estudio presentado por Moreno *et al.* (2005), en el que se proporcionan las listas de frecuencias de las diez formas más habituales de todo el corpus y las diez categorías léxicas más frecuentes.

Nosotros hemos hecho lo mismo, para comprobar si con un ranking de sólo diez elementos se puede apreciar alguna diferencia entre el lenguaje adulto y el infantil.

	CHIEDE	C-ORAL-ROM
Verbos	ser	ser
	tener	decir
	estar	estar
	ir	tener
	saber	hacer
	hacer	haber
	ver	ir
	haber	ver
	jugar	dar
	decir	saber
Nombres	casa	día
	niño	año
	madre	cosa
	padre	tío
	mamá	vez
	cosa	casa
	vez	gente
	día	persona
	abuelo	tiempo
	hermano	momento
Adjetivos	grande	mismo
	pequeño	bueno
	rojo	grande
	malo	mejor
	presente	pequeño
	azul	importante
	verde	último
	bueno	solo
	solo	nuevo
	chiquitín	normal
Palabras más frecuentes	y	de
	no	el
	sí	y
	el	a
	a	que
	que	la
	la	no
	un	en
	me	es
	yo	se

Tabla 4.23 Frecuencias CHIEDE vs. C-ORAL-ROM

Los resultados se incluyen en la tabla 4.21, en la que las principales diferencias residen en la lista de nombres y adjetivos. Las palabras que forman ambas categorías hacen referencia a conceptos cotidianos y a cualidades que los describen. Sin embargo, los nombres y adjetivos usados por los niños son más concretos, frente a la abstracción de los conceptos empleados en la lengua adulta, y abundan especialmente nombres de distintos miembros de la familia ("padre", "madre", "mamá", "abuelo", "hermano"). Es curioso que en el ranking de nombres usados por los adultos aparezca también un miembro de la familia: "tío". Sin embargo, en la lengua adulta oral lo más común es que "tío" no se refiera al hermano de tu padre o de tu madre, sino que funciona, en la lengua coloquial, como un apelativo amistoso o para designar a alguien cuyo nombre no se conoce. En el caso del lenguaje infantil —al menos cuando hablamos de niños que no superan los seis años de edad— no es habitual el uso de la palabra "tío" para referirse a otra persona que no sea el hermano de alguno de tus padres.

En cuanto a los diez adjetivos más frecuentes, en la lengua infantil éstos hacen referencia en su mayoría al campo semántico de los colores y los tamaños, exceptuando el adjetivo "presente", cuya frecuencia de uso aumenta debido al contexto situacional académico —es la respuesta de los niños cuando el maestro los nombra para comprobar su asistencia a clase. En el caso de los adultos, la variedad semántica de los diez adjetivos más habituales aumenta.

En el caso de los verbos más frecuentes, las diferencias no son tan notables. En ambos casos todos los verbos coinciden excepto por "jugar", ya que es una acción básica en la cotidianidad de los niños.

Finalmente, el listado de las diez palabras más frecuentes apenas aporta información en cuanto a las diferencias entre las dos variedades lingüísticas. Está formado únicamente por categorías gramaticales, por lo que habría que ampliar el número de elementos del listado para observar alguna diferencia significativa.

5. Explotación del recurso

5.1 Estudios cuantitativos vs. cualitativos

Una de las principales críticas que ha recibido la LC es el carácter cuantitativo de sus estudios. Muchos investigadores, sobre todo aquellos que defienden una perspectiva generativista de la lingüística, han destacado la falta de profundidad de los estudios cuantitativos, argumentando que en muchas ocasiones terminan siendo enormes listados de datos que carecen de una interpretación lingüística y que no alcanzan ninguna conclusión concreta. Es cierto que los análisis cualitativos se caracterizan por su exhaustividad, ya que se basan en un fenómeno concreto y éste se estudia con detalle. No obstante, los datos en los que se basan dichas investigaciones suelen ser pocos y creados de forma artificial. Por ello, sus conclusiones son difícilmente generalizables y no es posible saber si se deben a la casualidad o si realmente son relevantes.

> The main disadvantage of qualitative approaches to corpus analysis is that their findings can not be extended to wider populations with the same degree of certainty that quantitative analysis can. This is because the findings of the research are not tested to discover whether they are statistically significant or due to chance. (McEnery and Wilson, http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus3/3fra1.htm).

Por el contrario, los defensores de los análisis cuantitativos destacan su fiabilidad estadística y la posibilidad de establecer teorías generales por medio de sus resultados. Los estudios cuantitativos se basan en la clasificación de fenómenos, en su recuento y en la construcción de modelos estadísticos complejos que sirvan para dar una explicación a los datos. Los resultados obtenidos de un análisis cuantitativo permiten no sólo generalizar, sino observar hasta qué punto éstos se deben a una mera casualidad, o si por el contrario son fenómenos representativos de la lengua o variedad estudiada. Además, el hecho de tener una gran cantidad de datos cuantificados y medidos en cuanto a su representatividad nos permite realizar análisis que de otra forma no serían factibles. Por ejemplo, es posible comparar datos obtenidos de dos corpus distintos y establecer diferencias entre la lengua oral y la escrita, entre dos lenguas diferentes o entre dos variedades de una misma lengua.

En nuestra opinión, ambos tipos de análisis pueden ser complementarios: "Un corpus que proporcione simplemente una descripción de datos puede suponer el origen de una futura teoría." (Briz, 2005:7). La obtención y presentación de los datos puede no resultar relevante por sí misma; sin embargo, una vez ese trabajo está realizado, puede dar pie a futuras investigaciones. Por el contrario, una teoría basada en un breve listado de ejemplos creados de forma artificial puede considerarse irrelevante por su falta de representatividad. Cualquier tipo de teoría sobre el comportamiento

y la producción lingüísticos debe tomar como punto de partida ejemplos reales que le proporcionen unos cimientos sólidos.

A continuación, vamos a presentar un ejemplo de investigación cualitativa realizada a partir de un análisis cuantitativo de los datos.

5.2 Un ejemplo de investigación cualitativa: los marcadores discursivos en el lenguaje infantil

González-Ledesma y Garrote (2007) presentaron en el XXII Encuentro Internacional de la Asociación de Jóvenes Lingüistas un estudio realizado a partir del corpus CHIEDE sobre un fenómeno pragmático: los marcadores discursivos. La finalidad del estudio era observar hasta qué punto son utilizados los marcadores discursivos en el lenguaje infantil y qué repercusiones pueden tener dentro de una teoría discursiva y cognitiva.

La motivación principal de la investigación fue la carencia de estudios pragmáticos sobre la adquisición del lenguaje infantil en general, y en especial entre los grupos de edad que conforman nuestro corpus (3-6 años) a juzgar por la escasa bibliografía encontrada. La mayoría de los estudios se centran en la primera etapa de adquisición, esto es, desde que el niño empieza a emitir sus primeras palabras hasta que alcanza una competencia lingüística básica, es decir, hasta los 36 meses. Crystal afirma:

> Existe una gran porción de gramática que el niño *no* ha aprendido aún antes de ingresar en la escuela, y el proceso de ir completando dicho conocimiento de su lengua se mantendrá hasta casi el umbral de la pubertad. (Crystal, 1981:52)

Pero a pesar de ello, los estudios sobre lenguaje infantil no suelen extenderse más allá de los tres años de edad. A esto se suma el hecho de que la metodología empleada en dichos estudios (Gili Gaya, 1972; Crystal, 1981; Siguán, 1983; Anula, 1998) se basa en los planos fonológico, morfosintáctico y léxico, pero no hace ninguna mención al plano pragmático, es decir, que el nivel discursivo e interaccional de la lengua no se contempla.

Quizá por haber descuidado el nivel pragmático de la lengua es por lo que se suele pensar que hasta al menos los siete años el niño carece de habilidades pragmático-comunicativas (Piaget, 1923; Vigotsky, 1964; Bates, 1976).

Por todo ello, se consideró de gran interés realizar un estudio del fenómeno mencionado. Como punto de partida se realizó una selección de marcadores discursivos según su relevancia en términos cognitivos para el objeto final de nuestro estudio. A su vez, el inventario de marcadores se divide en dos clases: marcadores relacionales y marcadores de subjetividad.

Los primeros hacen referencia a las relaciones entre enunciados; su función es la de organizar el discurso y desde el punto de vista cognitivo implican ciertas operaciones formales de gran importancia en la relación entre lengua y pensamiento. La tabla 5.1 muestra los marcadores relacionales seleccionados para nuestro estudio:

MARCADORES RELACIONALES	
Relación semántica	**Codificación lingüística en español**
ADICIÓN	*y, luego*
OPOSICIÓN	*pero*
CONSECUCIÓN	*pues, así es que, porque, así que, entonces*
CONCESIÓN	*aunque*
JUSTIFICACIÓN	*es que, lo que pasa es que*
FINALIDAD	*para que*
HIPÓTESIS	*si*

Figura 5.1 Marcadores relacionales (González-Ledesma y Garrote 2007)

En segundo lugar, los marcadores discursivos de subjetividad muestran la percepción que tiene el hablante de la transmisión del conocimiento que está comunicando durante la interacción verbal. Éstos a su vez se subdividen en *conativos*, que mantienen abierto el canal comunicativo, y *moralizadores epistémicos*, que hacen referencia al grado de compromiso que muestra el hablante hacia la verdad del enunciado que está comunicando.

MARCADORES DE SUBJETIVIDAD		
CONATIVOS		*fíjate, sabes qué, ¿eh?, ¿vale?, ¿a que sí?, ¿sabes_X?, ¿a_que_no_sabes X?, ¿o no?*
MODALIZADORES EPISTÉMICOS	**Atenuantes**	*me parece que, a lo mejor, bueno*
	Intensificadores	*seguro que, vale, claro, eso sí, sí que, claro_que_sí*

Figura 5.2 Marcadores de subjetividad (González-Ledesma y Garrote 2007)

Una vez que se han definido los elementos de estudio, es necesario un tratamiento computacional del corpus para posteriormente extraer de forma automática la información que nos interesa. Un punto importante en dicha tarea es la resolución de la ambigüedad que caracteriza generalmente a los marcadores discursivos. Para ello, es necesaria una revisión manual posterior al etiquetado automático que garantice la fiabilidad de los datos. Después de todo este proceso, en cuyos detalles no nos detendremos ahora, los datos obtenidos fueron los siguientes:

MARCADOR DISCURSIVO			EDAD		
Tipo		**Subtipo**	**3-4**	**4-5**	**5-6**
MARCADORES DE RELACIÓN		Adición	206	359	409
		Consecución	22	46	95
		Oposición	20	55	46
		Justificación	14	17	36
		Hipótesis	5	12	9
		Finalidad	0	4	3
		Concesión	0	0	4
COORDINACIÓN VS. SUBORDINACIÓN	CATEGORÍA GRAMATICAL	Conjunción	44	105	135
		Relativo	12	49	65
	POSICIÓN DISCURSIVA	Ppio. turno	2	7	18
		Ppio. enunciado	17	34	38
		Interior enunciado	25	64	79
MARCADORES DE SUBJETIVIDAD		Conativos	3	7	15
		Intensificadores	2	3	5
		Atenuantes	2	11	15

Figura 5.3 Frecuencia de marcadores discursivos en CHIEDE
(González-Ledesma y Garrote 2007)

Entre las conclusiones que se extrajeron tras la observación de los datos, destacan hechos como que las estrategias de adición o consecución son las más usadas, mientras que en el punto opuesto se encontrarían la hipótesis y la concesividad. O que la coordinación como táctica sintáctica para unir enunciados se adquiere antes, mientras que la adquisición de la subordinación es más tardía. De la mima manera, los datos ponen de manifiesto que el uso de marcadores relacionales es más habitual que el de marcadores de subjetividad, lo que nos indica que aprendemos antes a estructurar nuestro discurso que a ser conscientes de sus efectos en el interlocutor. En definitiva, en dicho estudio se aclararon ciertas dudas sobre las habilidades comunicativas de los niños y el uso que éstos hacen de ciertas estrategias pragmáticas en sus interacciones.

La posibilidad de obtener estos datos reales, procedentes de un corpus de lengua oral espontánea, nos permite realizar una interpretación de los mismos y elaborar teorías generales sobre la variante lingüística estudiada. Y para ello, es necesario contar con un número de datos suficientemente

representativo, contabilizar dichos datos y extraer la información estadística relevante. De esta forma, un estudio cuantitativo podrá ser el origen de futuros análisis cualitativos.

6. La aplicación en Internet

Con el fin de facilitar la consulta de CHIEDE, hemos diseñado una página web por medio de la cual los profesionales interesados en el lenguaje infantil puedan utilizar el corpus en sus investigaciones. De esta forma, se puede acceder tanto a las transcripciones (alineadas con su sonido) como a cualquier tipo de información extraída de las mismas (número de palabras, listados de frecuencias). El único requisito es el registro del usuario, que nos permita garantizar el correcto uso de la información contenida en la página.

Según su diseño, la página web de CHIEDE consta de los siguientes apartados:

- **Inicio**: <http://www.lllf.uam.es/chiede>.
- **Introducción**: en la que se explican los motivos que nos llevaron a realizar el presente proyecto y se plantea el estado de la cuestión sobre la lingüística de corpus y la ontogénesis del lenguaje.
- **Objetivos**: donde se detallan los objetivos de nuestro trabajo.
- **Corpus**: en este apartado se describen el diseño del corpus y sus características.
- **Transcripciones**: el usuario puede consultar las transcripciones en formato de texto, realizadas mediante la transliteración ortográfica de las grabaciones realizadas.
- **Alineamiento**: en esta sección facilitamos el texto de las transcripciones alineado con su correspondiente sonido. Esto permite que en caso de duda respecto de las convenciones de transcripción, el usuario pueda escuchar la grabación original y juzgar por sí mismo.
- **Etiquetado**: aquí se encuentran los textos en formato XML con la anotación morfosintáctica. La anotación automática y su posterior revisión manual garantizan la fiabilidad de los datos.
- **Transcripción fonológica**: en este caso los textos han sido tratados con un transcriptor fonológico automático y en este apartado podemos consultarlos en dicho formato.
- **Medidas**: todos los datos obtenidos de forma automática con métodos estadísticos pueden ser consultados aquí. Se facilitan pues listados de frecuencias por lemas, categorías, longitud media de enunciado (LME), etc.
- **Guías**: se facilitan al usuario dos guías básicas para la comprensión y utilización del corpus. En la primera de ellas se incluyen todas las convenciones de transcripción para la comprensión de los textos transcritos; la segunda contiene el *tagset* o sistema de etiquetado categorial con especificaciones sobre los criterios seguidos para el establecimiento del mismo.
- **Sugerencias**: aquí el usuario podrá hacer cualquier tipo de comentario o crítica; por ejemplo, en caso de que encuentre algún error

en cualquiera de las partes del corpus, podría notificárnoslo mediante un correo electrónico y así podríamos subsanarlo.

7. Conclusiones y trabajo futuro

El trabajo realizado ha dado como resultado un Corpus de Habla Infantil Espontánea del Español, CHIEDE, que cuenta con cerca de 60.000 palabras. Aproximadamente un tercio del corpus está formado por habla infantil y los dos tercios restantes por habla adulta. El corpus pasará a formar parte de otro mayor, el compilado en el LLI-UAM de lengua oral adulta. De esta forma, incrementamos las muestras ya existentes, aportando una nueva variante lingüística: el lenguaje infantil.

CHIEDE cumple con todas las características que debe poseer un corpus de lengua oral actual. Su formato es electrónico, permitiendo el almacenamiento y la manipulación de los datos y su posible intercambio con otros investigadores interesados. Por su diseño proporcionado y su diversidad —variables de sexo, edad y situación comunicativa— garantiza una representatividad de la variedad lingüística en cuestión. Su presentación en una página web facilita su disponibilidad para todo aquel que esté interesado en su consulta. Finalmente, posee una estructura interna de clasificación de datos que posibilita una óptima explotación de los mismos.

Además, la posibilidad de contar con el respaldo de una metodología y unas herramientas de probada eficacia, desarrolladas en el LLI-UAM, otorga a nuestro corpus una mayor calidad.

La principal contribución a la comunidad científica con la realización de CHIEDE es la aportación de un recurso lingüístico que aún hoy escasea. El estudio de la adquisición y desarrollo del lenguaje debe basarse en la observación directa de la realidad. Su objeto de estudio es el lenguaje infantil en su manifestación más natural y espontánea. Con nuestro corpus, facilitamos una amplia muestra de la variedad lingüística infantil, tanto en formato de audio como de texto. Además, dichos textos están enriquecidos con anotación fonológica y morfosintáctica, de la cual se puede extraer automáticamente todo tipo de información referente a esos dos niveles lingüísticos.

Aparte del corpus, los listados de frecuencias que se facilitan en los Apéndices B y C pueden resultar de gran utilidad para ejemplificar teorías sobre adquisición y desarrollo lingüístico, comprobar hipótesis o realizar comparaciones con otras variedades lingüísticas.

En cuanto al trabajo futuro, las líneas de investigación serían principalmente dos:

- La ampliación del corpus no sólo en cuanto al número de palabras, sino también incrementando el rango de edad de los participantes e incluyendo una mayor variedad de situaciones comunicativas.

- La realización de estudios cualitativos a partir de los datos cuantitativos, como en el caso expuesto en el capítulo quinto del presente trabajo. Los diferentes fenómenos fonológicos y morfosintácticos anotados en los textos son susceptibles de ser analizados y pueden ser el objeto de estudio de futuras investigaciones.

Bibliografía

ÁLVAREZ DE MON Y REGO, I. *Lenguaje y Comunicación Científica y Técnica. Los nuevos caminos de la lingüística aplicada en el siglo XXI*. Madrid: Acta, 2006.

ANULA REBOLLO, A. *El abecé de la psicolingüística*. Madrid: Arco-Libros, D. L., 1998.

ATKINS, S.; CLEAR, J.; & OSLER, N. "Corpus design criteria". En: *Literary and Linguistic Computing*. 1992, vol 7, núm. 1, p. 1-16.

BATES, E. *Language and context: the acquisition of pragmatics*. New York: Academic Press, 1976.

BOADA, H. *El desarrollo de la comunicación en el niño*. Barcelona: Anthropos, 1986.

BOSQUE, I. *Las Categorías Gramaticales*. Madrid: Síntesis, 1990.

BOSQUE, I. y DEMONTE, V. (ed.). *Gramática Descriptiva de la Lengua Española*. Madrid: Espasa, 1999.

BOUTON, C. *El desarrollo del lenguaje*. Buenos Aires: Huemul, 1976.

BRANTS, T. "TnT: A statistical part-of-speech tagger". En: *Proceedings of the 6th Conference on Applied Natural Language Processing.* Seattle, USA, 2000. p. 224-231.

BRILL, E. "A simple rule-based part of speech tagger". En: *Proceedings of the DARPA Speech and Natural Language Workshop*. San Mateo: Morgan Kauffman, 1993. p. 112-116.

BRIZ, A. "Los corpus de español hablado. Presentación". En: *Oralia. Análisis del discurso oral*. Madrid: Arco Libros, 2005. p. 7-12.

BROWN, J. *A first language. The early stages*. Cambridge: Harvard University Press, 1973.

BROWN, R. *Psicolingüística: algunos aspectos acerca de la adquisición del lenguaje*. México D. F.: Trillas, 1981.

BRUNER, J. *Acción, pensamiento y lenguaje*. Madrid: Alianza, 1984.

CARTER, R. & MNCARTHY, M. (1995). "Grammar and the spoken language". En: *Applied linguistics*. 1995, vol 16, núm. 2, p. 141-158.

ČERMÁK, F. “Today’s corpus linguistics: Some open questions”. En: *International Journal of Corpus Linguistics*. 2002, vol 7, núm. 2, p. 265-282.

CHAFE, W. (1992) "The importance of corpus linguistics to understanding the nature of language". En: SVARTVIK, J. (ed.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, *Stockholm*. Berlin - New York: Mouton de Gruyter, 1992. p. 79-97.

CHAT [en línea]: <http://childes.psy.cmu.edu/manuals/CHAT.pdf> [Consulta: marzo de 2007]

CHURCH, K., *et al*. “Using statistics in lexical analysis”. En: ZERNICK, U. (ed.). *Lexical Acquisition*. Englewood Cliffs, NJ: Lawrence Erlbaum, 1991. p. 115-165.

CLAWS [en línea]: <http://ucrel.lancs.ac.uk/claws/> [Consulta: octubre de 2007]

CLEAR, J. “Corpus sampling”. En: LEITNER, G. (ed.). *New Directions in English Language Corpora. Methodology, Results, Software Development*. Berlin: Mouton de Gruyter, 1992. p. 21-32.

COUPER-KUHLEN, E. & SELTING, M. (ed.). *Prosody in Conversation: Interactional Studies.* Cambridge: Cambridge University Press, 1996.

CRESTI, E., *et al*. “The C-ORAL-ROM project: New methods for spoken language archives in a multilingual romance corpus”. En: M.C. RODRIGUEZ & C. SUAREZ ARAUJO (ed.), *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Paris: ELRA, 2002. p. 2-10.

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

CRYSTAL, D. *Lenguaje infantil. Aprendizaje y lingüística*. Barcelona: Editorial Médica y Técnica, 1981.

DALE, P. S. *Desarrollo del lenguaje: un enfoque psicolingüístico*. México D. F.: Trillas, 1992.

DUTCH CORPUS [en línea]: <http://lands.let.kun.nl/cgn/doc_English/topics/project/pro_info.htm> [Consulta: febrero de 2007]

EAGLES [en línea] : <http://www.ilc.cnr.it/EAGLES/home.html> [Consulta: febrero de 2007]

FILLMORE, C. J. "Corpus Linguistics or Computer-aided armchair linguistics". En: SVARTVIK, J. (ed.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, *Stockholm*. Berlin - New York: Mouton de Gruyter, 1992. p. 35-60.

GARSIDE, R., LEECH, G., & MCENERY, T. *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London - New York: Longman, 1997.

GILI GAYA, S. *Estudios de lenguaje infantil*. Barcelona: Biblograf, 1972.

GIUSEPPE, F. *El lenguaje infantil. Estructuración y* aprendizaje. Barcelona: Península, 1971.

GONZÁLEZ-LEDESMA, A. y GARROTE, M. "Los marcadores discursivos en CHIEDE, un corpus de habla infantil espontánea". En: *XXII Encuentro Internacional de la Asociación de Jóvenes Lingüistas*, (*Sevilla 9-11 de mayo de 2007).*

GONZÁLEZ, A., *et al*. "Orality and difficulties in the transcription of a spoken corpus". En: *Proceedings of the Worshop on Compiling and Processing Spoken Corpora*. LREC 2004, International Conference on Language Resources and Evaluation. Lisboa, 2004. p. 12-19.

GUIRAO, J. M. & MORENO-SANDOVAL, A. (2004). "A 'toolbox' for tagging the Spanish C-ORAL-ROM corpus". En: LINO, M. T. *et al*. (ed.). *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004) – Workshop "Compiling and Processing Spoken Language Corpora"*. Paris: ELRA, 2004. p. 28-32.

GUIRAO, J. M., *et al*. "Relating linguistic units to socio-contextual information in a spontaneous speech corpus of Spanish". En: *Corpus Linguistics Across the World*. Amsterdam: Rodopi, (en prensa).

MORENO SANDOVAL, A. Y GUIRAO, J. M.. "Morpho-syntactic Tagging of the Spanish C-ORAL-ROM corpus: methodology, tools and evaluation". En: *Spoken Language Corpus and Linguistic Informatics*. Amsterdam: John Benjamins, 2006.

HALLIDAY, M.A.K. (1992) "Language as system and language as instance: The corpus as a theoretical construct". En: SVARTVIK, J. (ed.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm*. Berlin - New York: Mouton de Gruyter, 1992. p. 61-77.

HERNÁNDEZ PINA, F. *Teorías psico-sociolingüísticas y su aplicación a la adquisición del español como lengua maternal*. Madrid: Siglo XXI de España Editores, 1984.

HIDALGO NAVARRO, A. Y SANMARTÍN SÁEZ, J. "Los sistemas de transcripción de la lengua hablada". En: *Oralia. Análisis del discurso oral*. Madrid: Arco Libros, 2005. p.13-36.

IDE, N. & VÉRONIS, J. (ed.). *The Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer Academic, 1995.

JOHANSSON, S. "Times change, and so do corpora". En: AIJMER, K. & ALTENBERG, B. (ed.). *English Corpus Linguistics*. London: Longman, 1991. p. 305-314.

KENNEDY, G. *An Introduction to Corpus Linguistics*. London and New York: Longman, 1998.

LEECH, G. (1991). "The state of the art in corpus linguistics". En: AIJMER, K. & ALTENBERG, B. (ed.). *English Corpus Linguistics*. London: Longman, 1991. p. 8-29.

LEECH, G. "Corpora and theories of linguistic performance". En: SVARTVIK, J. (ed.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm*. Berlin - New York: Mouton de Gruyter, 1992. p. 105-122.

LEECH, G. & WILSON, A. *EAGLES Recommendatios for the Morphosyntactic Annotation of Corpora* [en línea]: <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html> [Consulta: febrero de 2007]

LEECH, G. & FLIGELSTONE, S. "Computers and Corpus Analysis". En: BUTLER, C. S. (ed.). *Computers and Written Texts*. Oxford: Basil Blackwell, 1992. p. 115-140.

LLISTERRI, J. Informe sobre recursos lingüísticos para el español (II). Corpus escritos y orales disponibles y en desarrollo en España. Alcalá de Henares: Instituto Cervantes, 1996.

LLISTERRI, J. “Transcripción, etiquetado y codificación de corpus orales” [en línea]: <http://liceu.uab.es/~joaquim/publicacions/FDS97.html> [Consulta: marzo de 2007]

LLISTERRI, J., *et al*. “Corpus orales para el desarrollo de las tecnologías del habla en español”. En: *Oralia. Análisis del discurso oral*. Madrid Arco Libros, 2005. p. 289-328.

LURIA, A. R. *Lenguaje y desarrollo intelectual en el niño*. Madrid: Pablo del Río, 1956.

MACWHINNEY, B. *The CHILDES Project: Tools for Analyzing Talk* [en línea]: <http://childes.psy.cmu.edu/manuals/chat.pdf> [Consulta: enero de 2007]

MARCOS MARÍN, F. “El Corpus Oral de Referencia de la Lengua Española Contemporánea” [en línea] <ftp://ftp.lllf.uam.es/pub/corpus/oral> [Consulta: enero de 2007]

MARCOS MARÍN, F. *Informática y Humanidades*. Madrid: Gredos, 1994.

MCENERY, T. & WILSON, A. *Corpus Linguistics* [en línea]: <http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm> [Consulta: febrero de 2007]

MORENO SANDOVAL, A. *Un Modelo Computacional Basado en la Unificación para el Análisis y Generación de la Morfología del Español.* PhD thesis: Universidad Autónoma de Madrid, 1991.

MORENO SANDOVAL, A. & GOÑI, J. M. “A morphological model and processor for Spanish implemented in Prolog”. En: ALPUENTE, M. & SESSA, M.I. (ed.). *Proceedings of Joint Conference on Declarative Programming.* Salerno: Palladio Editrice, 1995. p. 321-331.

MORENO SANDOVAL, A. & GOÑI, J. M. “Spanish inflectional morphology in DATR”. *Journal of Logic, Language and Information*. 2002, vol 11, p. 79-105.

MORENO, A. & GUIRAO, J. M. “Tagging a spontaneous speech corpus of Spanish”. En: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.). Borovets, Bulgaria, 2003. p. 292-296.

MORENO SANDOVAL, A., *et al*. “Developing a syntactic annotation scheme and tools for a Spanish Treebank”. En: ABEILLÉ, A. (ed.),

Building and using syntactically annotated corpora. Dordrecht: Kluwerp, 2003. p.149-163.

MORENO SANDOVAL, A. Y URRESTI, J. "El Proyecto C-ORAL-ROM y su aplicación a la enseñanza del español". *Oralia. Análisis del discurso oral*. Madrid: Arco Libros, 2005. p. 81-104.

MORENO SANDOVAL, A. *et al.* ""Inventario de frecuencias fonémicas silábicas del castellano espontáneo y escrito". En: *Actas de las IV Jornadas de Tecnologías del Habla*. Zaragoza, 2006.

MORENO SANDOVAL, A. "Los corpus orales del LLI-UAM: Primera generación y segunda generación" [en línea]: <http://www.uclm.es/lamusa/ver_articulo.asp?articulo=44&lengua=es> [Consulta: enero de 2007]

NELSON FRANCIS, W. (1992) "Language corpora B.C.". En: SVARTVIK, J. (ed.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, *Stockholm*. Berlin - New York: Mouton de Gruyter, 1992. p. 17-32.

OFICINA DE ESPAÑOL EN LA SOCIEDAD DE LA INFORMACIÓN (OESI) [en línea]:
<http://oesi.cervantes.es/TLTODOS/recursos_linguisticos_2.htm>
[Consulta: marzo de 2007]

PÉREZ HERNÁNDEZ, M. C. *Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento* [en línea]: <http://elies.rediris.es/elies18/index.html> [Consulta: marzo de 2007]

PÉREZ MILANS, M. y MORENO, A "Diseño de corpus orales de lenguaje infantil: análisis comparado de CHILDES y CHIEDE". *Simposio anual de la Sociedad Española de Lingüística*. Madrid, 2004.

PIAGET, J. *El lenguaje y el pensamiento en el niño*, Buenos Aires: Paidós, 1965.

QUIRK, R. (1992) "On corpus principles and design". En: SVARTVIK, J. (ed.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, *Stockholm*. Berlin - New York: Mouton de Gruyter, 1992. p. 457-469.

RUBIO AYUSO, A. Y HERNÁEZ RIOJA, I. *Libro blanco de Tecnologías del Habla*. Granada: Universidad de Granada – Red Temática en Tecnologías del Habla, 2005.

SAMPA [en línea]: <http://www.phon.ucl.ac.uk/home/sampa/index.htm> [Consulta: junio de 2007]

SÁNCHEZ SÁNCHEZ, M. "El corpus de referencia del español actual (CREA). El CREA oral". En: *Oralia. Análisis del discurso oral*. Madrid: Arco Libros, 2005. p. 37-56.

SERRA, M., *et al*. *La adquisición del lenguaje*. Barcelona: Ariel, 2000.

SIGUÁN. M. *Metodología para el estudio del lenguaje en la infancia.* Universidad de Barcelona, 1983.

SIGUÁN, M. *Estudios sobre psicología del lenguaje infantil*. Madrid: Pirámide, 1984.

SINCLAIR, J. M. *Corpus, concordance, collocation.* Oxford: University Press, 1991.

SINCLAIR, J. M. "Corpus linguistics at the millennium". En: KOHN, J.; RÜSCHOFF, B. & WOLFF, D. (ed.). *New Horizons in CALL. Proceedings of European Association for Computer Assisted Language* Learning. Szombathely: Bersenyi Daniel College, 1997. p. 1-10.

STEIN, G. & QUIRK, R. "On having a look in a corpus". En: AIJMER, K. & ALTENBERG, B. (ed.). *English Corpus Linguistics*. London: Longman, 1991. p. 197-203.

SVARTVIK, J. (1992) "Corpus Linguistics comes of age". En: SVARTVIK, J. (ed.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, *Stockholm*. Berlin - New York: Mouton de Gruyter, 1992. p. 7-13.

SVARTVIK, J. "English corpus studies: Past, present and future". *English Corpus Studies*. 1999, vol 6, p. 1-16.

TEI (Text Encoding Initiative) [en línea]: <http://www.tei-c.org> [Consulta : enero de 2007]

TRANSANA [en línea]: <http://www.transana.org> [Consulta: diciembre de 2006]

Apéndice A

Ejemplo de CHIEDE

Transcripción ortográfica

@Title: Inés y Marta
@File: INE5
@Participants: INE, child, (woman, 5:2, 1, Ciudad Real)
MAR, adult, (woman, B, 3, Madrid)
@Date: 21/02/2006
@Place: Ciudad Real
@Situation: conversation in an empty classroom at school.
@Topic: daily matters
@Source: CHIEDE
@Class: informal, family/private, dialogue
@Lengh: 8´36"
@Words: 1093
@Acoustic_quality: A
@Transcriber: Marta
@Revisor: Ana and Marta
@Comments: INE (middle class; birth order: 1st)

*MAR: ¿ cuántos años tienes Inés ? ¿ cinco ?
%com: chair noise
*INE: sí ///
*MAR: ¿ y cuándo los cumples ? ¿ lo sabes ?
*INE: no ///
*MAR: ¿ no ?
*INE: en diciembre ///
*MAR: ¡ups! todavía te queda mucho // para los seis ///
*INE: yo los cumplí antes // pero / los &c [/] los [/] los seis me queda mucho ///
*MAR: claro /// hhh /// ¿ y tu hermanito cuántos tiene ?
%act: (2) assent
*INE: me parece que tiene tres /// tres añitos ///
*MAR: ¿ y él cuándo los cumple ?
*INE: los &c [/] ¿ los cuatro ? no sé ///
*MAR: ¿ no sabes ? yo tampoco lo sé /// bueno / y a ver /// ¿ qué estabais [///] qué habéis hecho hoy con Candi ? en clase ///
*INE: hemos hecho fichas /// ha venido Jesús y hemos hecho las caretas ///
%alt: (5) venío
*MAR: ¿ que Jesús qué os da ?
*INE: nos da unas cosas para pintar ///
*MAR: ¿ en qué idioma ?
*INE: algunas muy grandes / y nos cansamos un poco ///
*MAR: ¿ sí ? ¿ pintar ? < ¿ mucho pintar ? >
*INE: [<] < sí > ///
*MAR: pero Jesús os da [///] ¿ qué os da Jesús ?
*INE: Jesús +
*MAR: la asignatura ///
*INE: ¿ Arturo ?
*MAR: la &asignatu [/] asignatura ///
*INE: no lo sé ///
*MAR: ¿ no lo sabes ? ¿ qué idioma ?
*INE: idioma / inglés ///
*MAR: ¡ah! / pues eso /// inglés /// ¿ y sabes hablar en inglés ?
*INE: hhh ///
%act: (1) negation

*MAR: ¿ no ? pero sí sabes decir hola y adiós / ¿ no ? cuando viene /// y esas cosas ///
*INE: hello ///
*MAR: hhh ///
%act: (1) assent
*INE: bye bye ///
*MAR: muy bien /// claro /// poquito a poco /// ya el año que viene sabrás más / y luego más / y luego más ...
*INE: sí ///
%act: (1) whisper
*MAR: hhh ///
%act: (1) assent
*INE: sí ///
*MAR: que te gusta contar cuentos / ¿ a que sí ?
%com: INE nods
*MAR: xxx en las asambleas en cuanto puedes ... un cuento ///
*INE: sí ///
%act: (1) whisper
*MAR: ¿ y te los inventas tú ?
*INE: sí ///
*MAR: ¿ todo ?
*INE: sí ///
*MAR: tienes imaginación / ¿ eh ? ¿ me cuentas uno a mí ?
*INE: vale ///
*MAR: a ver / dime ///
*INE: había una vez un [/] un [/] un = a ver cuál es /// Caperucita Roja ///
*MAR: vale ///
*INE: había una vez una casita en el bosque / vivía Caperucita Roja /// y le dijo su mamá // ven Caperucita Roja /// llévale la comida a la abuela // que no puede venir // está coja / muy malita /// entonces dijo &pa [/] Caperucita / vale /// cogió la cestita // se puso la caperuza // y hhh voy a mi casita / hhh /// entonces / oyó una cosa detrás de los arbustos ///
%act: (42 and 47) singing
*MAR: ¿ dónde estaba ?
*INE: hhh ///
%act: (1) howl
*MAR: estaba [///] ¿ dónde estaba Caperucita ?
*INE: en el bosque ///
*MAR: en el bosque ///
*INE: y oía / hhh /// y salió un lobo muy / feo muy malo /// y le dijo a Caperucita // ¿ dónde vas Caperucita ? voy a casa de mi abuelita // a llevarle la comidita /// entonces / tú ve por el camino ese más largo // y yo por el más corto / así tú llegarás antes /// el lobo / la engañó /// sabía / que si llegaba por el camino más corto / llegaba antes /// entonces / el lobo corría más /// y Caperucita se entretuvo / viendo unas mariposas / cogiendo flores ... entonces / cuando llamó a la puerta afinando el lobo la voz /// &ton &ton / &ton &ton /// pasa / pasa /// la puerta está abierta /// y en cuanto vio el lobo // se [/] se la escondió en el armario /// cuando vino Caperucita / vino / y le dijo la [///] ¡ qué ojos más grandes tienes ! ¡ son para verte mejor ! ¡ qué nariz más grande tienes ! ¡ son para olerte mejor ! y qué orejas más grandes tienes /// ¡ son para oírte mejor ! y qué boca tan grande tienes /// ¡ son para comerte mejor ! se abalanzó /// y un xxx / el lobo ¡pum! y no se dio cuenta / que había fuego en el culo suyo ///
%act: (3) gol
%alt: (8 and 10) mu
*MAR: ¿ se estaba quemando la colita ?
*INE: sí / y salió corriendo /// y metió la cola en el charco // y le salía humo humo /// y colorín colorado / este cuento se ha acabado ///
*MAR: ¿ y quién había [///] por qué tenía fuego ?
*INE: porque el cazador le había disparado en el culo ///
*MAR: ¡ah! que lo del cazador no me lo has dicho ///
*INE: hhh ///
%act: (1) laugh
*MAR: hhh /// ¿ y esos cuentos quién te los cuenta a ti ?
%act: (1) assent
*INE: hhh yo me los invento ///

%act: (1) doubt
*MAR: ¿ pero el de Caperucita ?
*INE: nadie me los cuenta /// estamos haciendo un cuento de Caperucita /// pero no tiene tantas hojas ///
*MAR: ¿ no ?
*INE: no tiene [///] cuando se [///] la escondió en el armario y esas cosas ///
*MAR: hhh ///
%act: (1) assent
*INE: nos falta colorear el dibujo ///
*MAR: pero lo estáis haciendo en clase ///
*INE: sí ///
%act: (1) whisper
*MAR: ¿ y qué es otra cosa que vais a hacer / que sé yo ?
*INE: hhh / sumar ///
%act: (1) doubt
*MAR: no /// una cosita que vais a escribir también entre todos ///
*INE: ¿ entre todos juntos ?
*MAR: que lo dijo Candi el otro día /// dijo / tenéis que traer ... < una >
*INE: [<] < una -> > /// noticia ///
*MAR: eso /// ¿ para qué ?
*INE: para hacer un periódico ///
*MAR: un periódico ///
*INE: y mi perra [///] mira / fíjate /// te voy a contar qué le pasó a mi perra /// íbamos mi abuelo y yo paseando por el campo /// y cuando Aila ya se fue más lejos que nada // nos [///] se desapareció / fíjate /// y Furi nuestra perra // que era un poco chica // le [/] la llamó xxx / porque [/] porque Furi no &ha [/] sí hace caso ///
%alt: (34) na
*MAR: Furi es la pequeña ///
*INE: sí ///
*MAR: ¿ y la grande cómo se llama ?
*INE: pero [///] la grande se llama Aila ///
*MAR: ¿ Aila ?
*INE: pero / es un poco más alta / Furi /// Aila es la más alta /// y Furi es un poquito más ///
*MAR: más pequeña ///
#
*MAR: ¿ y qué pasó ? ¿ se fue ?
*INE: ya vino /// &mi [/] mira cuando / Furi era así de chiquitina / se salió a la calle /// al campo ///
*MAR: hhh ///
%act: (1) assent
*INE: y [/] y la madre de Furi estaba [///] se llamaba Aila / que ya te lo he dicho ///
*MAR: hhh ///
%act: (1) assent
*INE: y vino una perra /// y quería comerse a la chiquituja ///
*MAR: ¿ sí ?
*INE: pero salió la perra / Aila / y le pegó unos mordiscos que +
*MAR: dijo / hhh / cuidado /// ¿ no ?
%act: (2) click
*INE: sí ///
*MAR: ¿ y las tienes en casa a las perras ?
*INE: sí /// en mi casa no /// en la casa de los abuelos ///
*MAR: ¡ah!
*INE: porque yo tengo sólo nave /// no tengo patio ///
*MAR: ¡ah!
*INE: mi abuelo es que tiene nave / patio / casa ...
*MAR: ¡jolín! ahí puede tener un montón de perros ///
*INE: sólo tiene dos ///
*MAR: dos /// y son mamá e hija / entonces ///
%alt: (5) y
*INE: sí ///
*MAR: ¡ah! / o sea +

*INE: el padre de Furi no está // pero sí lo conozco / no sé el nombre /// aunque se fue con otro amo ///
*MAR: ¿ se fue ?
*INE: sí ///
*MAR: ¿ pero él se escapó ?
*INE: no se escapó /// < se fue con otro amo >
*MAR: [<] <¡ah! / se lo diste > /// ¡ah! / claro /// hhh /// vale /// ¿ tienes más animales ?
%act: (7) assent
*INE: sí /// muchos /// < en mi casa >
*MAR: [<] < ¿ más ? > ¿ sí ?
*INE: no [/] no me gusta lo de princesas // sólo &l [/] lo de -> animales +
*MAR: lo de los animales ///
*INE: ¿ sabes qué animal le gusta más a mi hermano ?
*MAR: a ver ///
*INE: el elefante ///
*MAR: tenía uno el otro día de plástico / que lo vi yo ///
*INE: ¿ sí ?
*MAR: venía muy contento con el elefante ///
*INE: ¿ ese gris / chiquitín ?
*MAR: sí /// uno gris chiquitín ///
*INE: ¡ah!
*MAR: se lo guardaba en el bolsillo del babi ///
*INE: es que no quiere que Abel se lo quite // porque ayer le quitó / la moto y la tiró fuera ///
*MAR: ¿ quién ? < ¿ otro niño ? >
*INE: [<] < Abel > ///
*MAR: ¿ sí ? ¡ahí va! sí / pues lo llevaba y le dijo la profesora // guárdatelo en el babi ///
*INE: ¿ y sabes que le encontré el [/] en el bolsillo ?
*MAR: ¿ el qué ?
*INE: le encontré un chupete ///
*MAR: ¿ un chupete ?
*INE: ¡ sí !
*MAR: bueno / pero todavía es chiquitín / ¿ no ?
*INE: ya / pero / eso es cuando sea bebé ///
*MAR: cuando es bebé sólo ///
*INE: le vi un chupete hhh ///
%act: (5) laugh
*MAR: bueno / hay niños que -> tienen el chupete -> / mucho tiempo ///
*INE: sí ///
*MAR: ya cuando son mayores +
*INE: pero / Abel no lo lleva puesto ///
*MAR: ¿ no ? sólo lo lleva en el bolsillo / por si acaso /// ¿ no ?
*INE: si nunca lo necesita ///
*MAR: no /// hhh /// pues nada / muy bien Inés /// nos vamos a ir para clase / ¿ vale ?
%act: (2) laugh
*INE: vale ///
*MAR: ¡ale! vámonos ///

Transcripción fonológica[1]

@Title: Inés y Marta
@File: INE5
@Participants: INE, child, (woman, 5:2, 1, Ciudad Real)
TEA, adult, (woman, B, 3, Madrid)
@Birth_Date: 4/12/2000
@Date: 21/02/2006
@Place: Ciudad Real
@Situation: conversation in an empty classroom at school.
@Topic: daily matters
@Source: CHIEDE
@Class: informal, family/private, dialogue
@Lengh: 8´36"
@Words: 1067
@Acoustic_quality: A
@Transcriber: Marta
@Revisor: Ana and Marta
@Comments: INE (middle class; birth order: 1st)

```
*TEA: ¿ . k u a+ n . t o s .  . a+ . N o s .  . t i e+ . n e s .  . i
. n e+ s .  ? ¿ . T i+ n . k o .  ?
%com: chair noise
*INE: . s i+ .  ///
*TEA: ¿ . i+ .  . k u a+ n . d o .  . l o+ s .  . k u+ m . p l e s .
? ¿ . l o+ .  . s a+ . b e s .  ?
*INE: . n o+ .  ///
*TEA: ¿ . n o+ .  ?
*INE: . e+ n .  . d i . T i e+ m . b r e .  ///
*TEA: ¡ups! . t o . d a . b i+ . a .  . t e+ .  . k e+ . d a .  . m u+
. C o .  // . p a+ . r a .  . l o+ s .  . s e+ i s .  ///
*INE: . y o+ .  . l o+ s .  . k u m . p l i+ .  . a+ n . t e s .  // .
p e+ . r o .  / . l o+ s .  &c [/] . l o+ s .  [/] . l o+ s .  . s e+
i s .  . m e+ .  . k e+ . d a .  . m u+ . C o .  ///
*TEA: . k l a+ . r o .  /// hhh /// ¿ . i+ .  . t u+ .  . e r . m a .
n i+ . t o .  . k u a+ n . t o s .  . t i e+ . n e .  ?
%act: (2) assent
*INE: . m e+ .  . p a . r e+ . T e .  . k e+ .  . t i e+ . n e .  . t
r e+ s .  /// . t r e+ s .  . a . N i+ . t o s .  ///
*TEA: ¿ . i+ .  . e+ l .  . k u a+ n . d o .  . l o+ s .  . k u+ m . p
l e .  ?
*INE: . l o+ s .  &c [/] ¿ . l o+ s .  . k u a+ . t r o .  ? . n o+ .
. s e+ .  ///
*TEA: ¿ . n o+ .  . s a+ . b e s .  ? . y o+ .  . t a m . p o+ . k o .
. l o+ .  . s e+ .  /// . b u e+ . n o .  / . i+ .  . a+ .  . b e+ r .
/// ¿ . k e+ .  . e s . t a+ . b a i s .  [///] . k e+ .  . a . b e+ i
s .  . e+ . C o .  . o+ i .  . k o+ n .  . k a+ n . d i .  ? . e+ n .
. k l a+ . s e .  ///
*INE: . e+ . m o s .  . e+ . C o .  . f i+ . C a s .  /// . a+ .  . b
e . n i+ . d o .  . x e . s u+ s .  . i+ .  . e+ . m o s .  . e+ . C o
.  . l a+ s .  . k a . r e+ . t a s .  ///
%alt: (5) venío
*TEA: ¿ . k e+ .  . x e . s u+ s .  . k e+ .  . o+ s .  . d a+ .  ?
*INE: . n o+ s .  . d a+ .  . u+ . n a s .  . k o+ . s a s .  . p a+ .
r a .  . p i n . t a+ r .  ///
```

[1] La transcripción fonológica está en formato SAMPA, ya que esta es la forma que emplea el transcriptor fonológico automático. Las diferencias con el formato IPA afectan únicamente a cuatro fonemas: θ (=T), r (=R), ʧ(=C) y ɲ (=N).

```
*TEA: ¿ . e+ n .  . k e+ .  . i . d i o+ . m a .  ?
*INE: . a l . g u+ . n a s .  . m u i+ .  . g r a+ n . d e s .  / . i+
.  . n o+ s .  . k a n . s a+ . m o s .  . u+ n .  . p o+ . k o .  ///
*TEA: ¿ . s i+ .  ? ¿ . p i n . t a+ r .  ? <  ¿ . m u+ . C o .  . p i
n . t a+ r .  ?  >
*INE: [< ] < .  s i+ .  > ///
*TEA: . p e+ . r o .  . x e . s u+ s .  . o+ s .  . d a+ .  [///] ¿ .
k e+ .  . o+ s .  . d a+ .  . x e . s u+ s .  ?
*INE: . x e . s u+ s .  +
*TEA: . l a+ .  . a . s i g . n a . t u+ . r a .  ///

*INE: ¿ . a r . t u+ . r o .  ?
*TEA: . l a+ .  &asignatu [/] . a . s i g . n a . t u+ . r a .  ///
*INE: . n o+ .  . l o+ .  . s e+ .  ///
*TEA: ¿ . n o+ .  . l o+ .  . s a+ . b e s .  ? ¿ . k e+ .  . i . d i
o+ . m a .  ?
*INE: . i . d i o+ . m a .  / . i n . g l e+ s .  ///
*TEA: ¡ah! / . p u e+ s .  . e+ . s o .  /// . i n . g l e+ s .  /// ¿
. i+ .  . s a+ . b e s .  . a . b l a+ r .  . e+ n .  . i n . g l e+ s
.  ?
*INE: hhh ///
%act: (1) negation
*TEA: ¿ . n o+ .  ? . p e+ . r o .  . s i+ .  . s a+ . b e s .  . d e
. T i+ r .  . o+ . l a .  . i+ .  . a . d i o+ s .  / ¿ . n o+ .  ? .
k u a+ n . d o .  . b i e+ . n e .  /// . i+ .  . e+ . s a s .  . k o+
. s a s .  ///
*INE: . x e+ . l o u .  ///
*TEA: hhh ///
%act: (1) assent
*INE: . b a+ i .  . b a+ i .  ///
*TEA: . m u i+ .  . b i e+ n .  /// . k l a+ . r o .  /// . p o . k i+
. t o .  . a+ .  . p o+ . k o .  /// . y a+ .  . e+ l .  . a+ . N o .
. k e+ .  . b i e+ . n e .  . s a . b r a+ s .  . m a+ s .  / . i+ .
. l u e+ . g o .  . m a+ s .  / . i+ .  . l u e+ . g o .  . m a+ s .
...
*INE: . s i+ .  ///
%act: (1) whisper
*TEA: hhh ///
%act: (1) assent
*INE: . s i+ .  ///
*TEA: . k e+ .  . t e+ .  . g u+ s . t a .  . k o n . t a+ r .  . k u
e+ n . t o s .  / ¿ . a+ .  . k e+ .  . s i+ .  ?
%com: INE nods
*TEA: xxx . e+ n .  . l a+ s .  . a . s a m . b l e+ . a s .  . e+ n .
. k u a+ n . t o .  . p u e+ . d e s .  ... . u+ n .  . k u e+ n . t o
.  ///
*INE: . s i+ .  ///
%act: (1) whisper
*TEA: ¿ . i+ .  . t e+ .  . l o+ s .  . i m . b e+ n . t a s .  . t u+
.  ?
*INE: . s i+ .  ///
*TEA: ¿ . t o+ . d o .  ?
*INE: . s i+ .  ///
*TEA: . t i e+ . n e s .  . i . m a . x i . n a . T i o+ n .  / ¿ &eh
? ¿ . m e+ .  . k u e+ n . t a s .  . u+ . n o .  . a+ .  . m i+ .  ?
*INE: . b a+ . l e .  ///
*TEA: . a+ .  . b e+ r .  / . d i+ . m e .  ///
*INE: . a . b i+ . a .  . u+ . n a .  . b e+ T .  . u+ n .  [/] . u+ n
.  [/] . u+ n .  = . a+ .  . b e+ r .  . k u a+ l .  . e+ s .  /// . k
a . p e . r u . T i+ . t a .  . R o+ . x a .  ///
*TEA: . b a+ . l e .  ///
```

```
*INE: . a . b i+ . a . . u+ . n a . . b e+ T . . u+ . n a . . k a
. s i+ . t a . . e+ n . . e+ l . . b o+ s . k e . / . b i . b i+ .
a . . k a . p e . r u . T i+ . t a . . R o+ . x a . /// . i+ . . l
e+ . . d i+ . x o . . s u+ . . m a . m a+ . // . b e+ n . . k a .
p e . r u . T i+ . t a . . R o+ . x a . /// . y e+ . b a . l e . .
l a+ . . k o . m i+ . d a . . a+ . . l a+ . . a . b u e+ . l a .
// . k e+ . . n o+ . . p u e+ . d e . . b e . n i+ r . // . e s .
t a+ . . k o+ . x a . / . m u i+ . . m a . l i+ . t a . /// . e n
. t o+ n . T e s . . d i+ . x o . &pa [/] . k a . p e . r u . T i+ .
t a . / . b a+ . l e . /// . k o . x i o+ . . l a+ . . T e s . t
i+ . t a . // . s e+ . . p u+ . s o . . l a+ . . k a . p e . r u+
. T a . // . i+ . hhh . b o+ i . . a+ . . m i+ . . k a . s i+ .
t a . / hhh /// . e n . t o+ n . T e s . / . o . y o+ . . u+ . n a
. . k o+ . s a . . d e . t r a+ s . . d e+ . . l o+ s . . a r . b
u+ s . t o s . ///
%act: (42 and 47) singing
*TEA: ¿ . d o+ n . d e . . e s . t a+ . b a . ?
*INE: hhh ///
%act: (1) howl
*TEA: . e s . t a+ . b a . [///] ¿ . d o+ n . d e . . e s . t a+ . b
a . . k a . p e . r u . T i+ . t a . ?
*INE: . e+ n . . e+ l . . b o+ s . k e . ///
*TEA: . e+ n . . e+ l . . b o+ s . k e . ///
*INE: . i+ . . o . i+ . a . / hhh /// . i+ . . s a . l i o+ . .
u+ n . . l o+ . b o . . m u i+ . / . f e+ . o . . m u i+ . . m a+
. l o . /// . i+ . . l e+ . . d i+ . x o . . a+ . . k a . p e . r
u . T i+ . t a . // ¿ . d o+ n . d e . . b a+ s . . k a . p e . r u
. T i+ . t a . ? . b o+ i . . a+ . . k a+ . s a . . d e+ . . m
i+ . . a . b u e . l i+ . t a . // . a+ . . y e . b a+ r . l e . .
l a+ . . k o . m i . d i+ . t a . /// . e n . t o+ n . T e s . / .
t u+ . . b e+ . . p o+ r . . e+ l . . k a . m i+ . n o . . e+ . s
e . . m a+ s . . l a+ r . g o . // . i+ . . y o+ . . p o+ r . .
e+ l . . m a+ s . . k o+ r . t o . / . a . s i+ . . t u+ . . y e
. g a . r a+ s . . a+ n . t e s . /// . e+ l . . l o+ . b o . / .
l a+ . . e n . g a . N o+ . /// . s a . b i+ . a . / . k e+ . . s
i+ . . y e . g a+ . b a . . p o+ r . . e+ l . . k a . m i+ . n o .
. m a+ s . . k o+ r . t o . / . y e . g a+ . b a . . a+ n . t e s .
/// . e n . t o+ n . T e s . / . e+ l . . l o+ . b o . . k o . R i+
. a . . m a+ s . /// . i+ . . k a . p e . r u . T i+ . t a . . s
e+ . . e n . t r e . t u+ . b o . / . b i e+ n . d o . . u+ . n a s
. . m a . r i . p o+ . s a s . / . k o . x i e+ n . d o . . f l o+
. r e s . ... . e n . t o+ n . T e s . / . k u a+ n . d o . . y a .
m o+ . . a+ . . l a+ . . p u e+ r . t a . . a . f i . n a+ n . d o
. . e+ l . . l o+ . b o . . l a+ . . b o+ T . /// &ton &ton /
&ton &ton /// . p a+ . s a . / . p a+ . s a . /// . l a+ . . p u e+
r . t a . . e s . t a+ . . a . b i e+ r . t a . /// . i+ . . e+ n
. . k u a+ n . t o . . b i o+ . . e+ l . . l o+ . b o . // . s e+
. [/] . s e+ . . l a+ . . e s . k o n . d i o+ . . e+ n . . e+ l
. . a r . m a+ . r i o . /// . k u a+ n . d o . . b i+ . n o . . k
a . p e . r u . T i+ . t a . / . b i+ . n o . / . i+ . . l e+ . .
d i+ . x o . . l a+ . [///] ¡ . k e+ . . o+ . x o s . . m a+ s .
. g r a+ n . d e s . . t i e+ . n e s . ! ¡ . s o+ n . . p a+ . r a
. . b e+ r . t e . . m e . x o+ r . ! ¡ . k e+ . . n a . r i+ T .
. m a+ s . . g r a+ n . d e . . t i e+ . n e s . ! ¡ . s o+ n . .
p a+ . r a . . o . l e+ r . t e . . m e . x o+ r . ! . i+ . . k e+
. . o . r e+ . x a s . . m a+ s . . g r a+ n . d e s . . t i e+ .
n e s . /// ¡ . s o+ n . . p a+ . r a . . o . i+ r . t e . . m e .
x o+ r . ! . i+ . . k e+ . . b o+ . k a . . t a+ n . . g r a+ n .
d e . . t i e+ . n e s . /// ¡ . s o+ n . . p a+ . r a . . k o . m
e+ r . t e . . m e . x o+ r . ! . s e+ . . a . b a . l a n . T o+ .
/// . i+ . . u+ n . xxx / . e+ l . . l o+ . b o . ¡pum! . i+ . .
```

n o+ . . s e+ . . d i o+ . . k u e+ n . t a . / . k e+ . . a . b i+ . a . . f u e+ . g o . . e+ n . . e+ l . . k u+ . l o . . s u+ . y o . ///
%act: (3) gol
%alt: (8 and 10) mu
*TEA: ¿ . s e+ . . e s . t a+ . b a . . k e . m a+ n . d o . . l a+ . . k o . l i+ . t a . ?
*INE: . s i+ . / . i+ . . s a . l i o+ . . k o . R i e+ n . d o . /// . i+ . . m e . t i o+ . . l a+ . . k o+ . l a . . e+ n . . e+ l . . C a+ r . k o . // . i+ . . l e+ . . s a . l i+ . a . . u+ . m o . . u+ . m o . /// . i+ . . k o . l o . r i+ n . . k o . l o . r a+ . d o . / . e+ s . t e . . k u e+ n . t o . . s e+ . . a+ . . a . k a . b a+ . d o . ///
*TEA: ¿ . i+ . . k i e+ n . . a . b i+ . a . [///] . p o+ r . . k e+ . . t e . n i+ . a . . f u e+ . g o . ?
*INE: . p o+ r . k e . . e+ l . . k a . T a . d o+ r . . l e+ . . a . b i+ . a . . d i s . p a . r a+ . d o . . e+ n . . e+ l . . k u+ . l o . ///
*TEA: ¡ah! . k e+ . . l o+ . . d e+ l . . k a . T a . d o+ r . . n o+ . . m e+ . . l o+ . . a+ s . . d i+ . C o . ///
*INE: hhh ///
%act: (1) laugh
*TEA: hhh /// ¿ . i+ . . e+ . s o s . . k u e+ n . t o s . . k i e+ n . . t e+ . . l o+ s . . k u e+ n . t a . . a+ . . t i+ . ?
%act: (1) assent
*INE: hhh . y o+ . . m e+ . . l o+ s . . i m . b e+ n . t o . ///
%act: (1) doubt
*TEA: ¿ . p e+ . r o . . e+ l . . d e+ . . k a . p e . r u . T i+ . t a . ?
*INE: . n a+ . d i e . . m e+ . . l o+ s . . k u e+ n . t a . /// . e s . t a+ . m o s . . a . T i e+ n . d o . . u+ n . . k u e+ n . t o . . d e+ . . k a . p e . r u . T i+ . t a . /// . p e+ . r o . . n o+ . . t i e+ . n e . . t a+ n . t a s . . o+ . x a s . ///
*TEA: ¿ . n o+ . ?
*INE: . n o+ . . t i e+ . n e . [///] . k u a+ n . d o . . s e+ . [///] . l a+ . . e s . k o n . d i o+ . . e+ n . . e+ l . . a r . m a+ . r i o . . i+ . . e+ . s a s . . k o+ . s a s . ///
*TEA: hhh ///
%act: (1) assent
*INE: . n o+ s . . f a+ l . t a . . k o . l o . r e . a+ r . . e+ l . . d i . b u+ . x o . ///
*TEA: . p e+ . r o . . l o+ . . e s . t a+ i s . . a . T i e+ n . d o . . e+ n . . k l a+ . s e . ///
*INE: . s i+ . ///
%act: (1) whisper
*TEA: ¿ . i+ . . k e+ . . e+ s . . o+ . t r a . . k o+ . s a . . k e+ . . b a+ i s . . a+ . . a . T e+ r . / . k e+ . . s e+ . . y o+ . ?
*INE: hhh / . s u . m a+ r . ///
%act: (1) doubt
*TEA: . n o+ . /// . u+ . n a . . k o . s i+ . t a . . k e+ . . b a+ i s . . a+ . . e s . k r i . b i+ r . . t a m . b i e+ n . . e+ n . t r e . . t o+ . d o s . ///
*INE: ¿ . e+ n . t r e . . t o+ . d o s . . x u+ n . t o s . ?
*TEA: . k e+ . . l o+ . . d i+ . x o . . k a+ n . d i . . e+ l . . o+ . t r o . . d i+ . a . /// . d i+ . x o . / . t e . n e+ i s . . k e+ . . t r a . e+ r . ... < . u+ . n a . >
*INE: [<] < . u+ . n a . - > > /// . n o . t i+ . T i a . ///
*TEA: . e+ . s o . /// ¿ . p a+ . r a . . k e+ . ?
*INE: . p a+ . r a . . a . T e+ r . . u+ n . . p e . r i o+ . d i . k o . ///

```
*TEA: . u+ n .  . p e . r i o+ . d i . k o .  ///
*INE: . i+ .  . m i+ .  . p e+ . R a .  [///] . m i+ . r a .  / . f i+
. x a . t e .  /// . t e+ .  . b o+ i .  . a+ .  . k o n . t a+ r .
. k e+ .  . l e+ .  . p a . s o+ .  . a+ .  . m i+ .  . p e+ . R a .
/// . i+ . b a . m o s .  . m i+ .  . a . b u e+ . l o .  . i+ .  . y
o+ .  . p a . s e . a+ n . d o .  . p o+ r .  . e+ l .  . k a+ m . p o
.  /// . i+ .  . k u a+ n . d o .  . a+ i . l a .  . y a+ .  . s e+ .
. f u e+ .  . m a+ s .  . l e+ . x o s .  . k e+ .  . n a+ . d a .  //
. n o+ s .  [///] . s e+ .  . d e . s a . p a . r e . T i o+ .  / . f
i+ . x a . t e .  /// . i+ .  . f u+ . r i .  . n u e+ s . t r a .  .
p e+ . R a .  // . k e+ .  . e+ . r a .  . u+ n .  . p o+ . k o .  . C
i+ . k a .  // . l e+ .  [/] . l a+ .  . y a . m o+ .  xxx / . p o+ r
. k e .  [/] . p o+ r . k e .  . f u+ . r i .  . n o+ .  &ha [/] . s
i+ .  . a+ . T e .  . k a+ . s o .  ///
%alt: (34) na
*TEA: . f u+ . r i .  . e+ s .  . l a+ .  . p e . k e+ . N a .  ///

*INE: . s i+ .  ///
*TEA: ¿ . i+ .  . l a+ .  . g r a+ n . d e .  . k o+ . m o .  . s e+ .
. y a+ . m a .  ?
*INE: . p e+ . r o .  [///] . l a+ .  . g r a+ n . d e .  . s e+ .  .
y a+ . m a .  . a+ i . l a .  ///
*TEA: ¿ . a+ i . l a .  ?
*INE: . p e+ . r o .  / . e+ s .  . u+ n .  . p o+ . k o .  . m a+ s .
. a+ l . t a .  / . f u+ . r i .  /// . a+ i . l a .  . e+ s .  . l a+
.  . m a+ s .  . a+ l . t a .  /// . i+ .  . f u+ . r i .  . e+ s .  .
u+ n .  . p o . k i+ . t o .  . m a+ s .  ///
*TEA: . m a+ s .  . p e . k e+ . N a .  /// #

*TEA: ¿ . i+ .  . k e+ .  . p a . s o+ .  ? ¿ . s e+ .  . f u e+ .  ?
*INE: . y a+ .  . b i+ . n o .  /// &mi [/] . m i+ . r a .  . k u a+ n
. d o .  / . f u+ . r i .  . e+ . r a .  . a . s i+ .  . d e+ .  . C i
. k i . t i+ . n a .  / . s e+ .  . s a . l i o+ .  . a+ .  . l a+ .
. k a+ . y e .  /// . a+ l .  . k a+ m . p o .  ///
*TEA: hhh ///
%act: (1) assent
*INE: . i+ .  [/] . i+ .  . l a+ .  . m a+ . d r e .  . d e+ .  . f u+
. r i .  . e s . t a+ . b a .  [///] . s e+ .  . y a . m a+ . b a .  .
a+ i . l a .  / . k e+ .  . y a+ .  . t e+ .  . l o+ .  . e+ .  . d i+
. C o .  ///
*TEA: hhh ///
%act: (1) assent
*INE: . i+ .  . b i+ . n o .  . u+ . n a .  . p e+ . R a .  /// . i+ .
. k e . r i+ . a .  . k o . m e+ r . s e .  . a+ .  . l a+ .  . C i .
k i . t u+ . x a .  ///
*TEA: ¿ . s i+ .  ?
*INE: . p e+ . r o .  . s a . l i o+ .  . l a+ .  . p e+ . R a .  / .
a+ i . l a .  / . i+ .  . l e+ .  . p e . g o+ .  . u+ . n o s .  . m
o r . d i+ s . k o s .  . k e+ .  +
*TEA: . d i+ . x o .  / hhh / . k u i . d a+ . d o .  /// ¿ . n o+ .
?
%act: (2) click
*INE: . s i+ .  ///
*TEA: ¿ . i+ .  . l a+ s .  . t i e+ . n e s .  . e+ n .  . k a+ . s a
.  . a+ .  . l a+ s .  . p e+ . R a s .  ?
*INE: . s i+ .  /// . e+ n .  . m i+ .  . k a+ . s a .  . n o+ .  ///
. e+ n .  . l a+ .  . k a+ . s a .  . d e+ .  . l o+ s .  . a . b u e+
. l o s .  ///
*TEA: ¡ah!
*INE: . p o+ r . k e .  . y o+ .  . t e+ n . g o .  . s o+ . l o .  .
n a+ . b e .  /// . n o+ .  . t e+ n . g o .  . p a+ . t i o .  ///
```

```
*TEA: ¡ah!
*INE: . m i+ . . a . b u e+ . l o . . e+ s . . k e+ . . t i e+ . n
e . . n a+ . b e . / . p a+ . t i o . / . k a+ . s a . ...
*TEA: ¡jolín! . a . i+ . . p u e+ . d e . . t e . n e+ r . . u+ n .
. m o n . t o+ n . . d e+ . . p e+ . R o s . ///
*INE: . s o+ . l o . . t i e+ . n e . . d o+ s . ///
*TEA: . d o+ s . /// . i+ . . s o+ n . . m a . m a+ . . e+ . . i+
. x a . / . e n . t o+ n . T e s . ///
%alt: (5) y
*INE: . s i+ . ///
*TEA: ¡ah! / . o+ . . s e+ . a . +
*INE: . e+ l . . p a+ . d r e . . d e+ . . f u+ . r i . . n o+ .
. e s . t a+ . // . p e+ . r o . . s i+ . . l o+ . . k o . n o+ T
. k o . / . n o+ . . s e+ . . e+ l . . n o+ m . b r e . /// . a+
u n . k e . . s e+ . . f u e+ . . k o+ n . . o+ . t r o . . a+ .
m o . ///
*TEA: ¿ . s e+ . . f u e+ . ?
*INE: . s i+ . ///
*TEA: ¿ . p e+ . r o . . e+ l . . s e+ . . e s . k a . p o+ . ?
*INE: . n o+ . . s e+ . . e s . k a . p o+ . /// < . s e+ . . f
u e+ . . k o+ n . . o+ . t r o . . a+ . m o . >
*TEA: [< ] < ¡ah! / . s e+ . . l o+ . . d i+ s . t e . > /// ¡ah!
/ . k l a+ . r o . /// hhh /// . b a+ . l e . /// ¿ . t i e+ . n e s
. . m a+ s . . a . n i . m a+ . l e s . ?
%act: (7) assent
*INE: . s i+ . /// . m u+ . C o s . /// < . e+ n . . m i+ . . k
a+ . s a . >
*TEA: [< ] < ¿ . m a+ s . ? > ¿ . s i+ . ?
*INE: . n o+ . [/] . n o+ . . m e+ . . g u+ s . t a . . l o+ . .
d e+ . . p r i n . T e+ . s a s . // . s o+ . l o . &l [/] . l o+ .
. d e+ . - > . a . n i . m a+ . l e s . +
*TEA: . l o+ . . d e+ . . l o+ s . . a . n i . m a+ . l e s . ///
*INE: ¿ . s a+ . b e s . . k e+ . . a . n i . m a+ l . . l e+ . .
g u+ s . t a . . m a+ s . . a+ . . m i+ . . e r . m a+ . n o . ?
*TEA: . a+ . . b e+ r . ///
*INE: . e+ l . . e . l e . f a+ n . t e . ///
*TEA: . t e . n i+ . a . . u+ . n o . . e+ l . . o+ . t r o . . d
i+ . a . . d e+ . . p l a+ s . t i . k o . / . k e+ . . l o+ . .
b i+ . . y o+ . ///
*INE: ¿ . s i+ . ?
*TEA: . b e . n i+ . a . . m u i+ . . k o n . t e+ n . t o . . k o+
n . . e+ l . . e . l e . f a+ n . t e . ///
*INE: ¿ . e+ . s e . . g r i+ s . / . C i . k i . t i+ n . ?
*TEA: . s i+ . /// . u+ . n o . . g r i+ s . . C i . k i . t i+ n .
///
*INE: ¡ah!
*TEA: . s e+ . . l o+ . . g u a r . d a+ . b a . . e+ n . . e+ l .
. b o l . s i+ . y o . . d e+ l . . b a+ . b i . ///
*INE: . e+ s . . k e+ . . n o+ . . k i e+ . r e . . k e+ . . a .
b e+ l . . s e+ . . l o+ . . k i+ . t e . // . p o+ r . k e . . a
. y e+ r . . l e+ . . k i . t o+ . / . l a+ . . m o+ . t o . . i+
. . l a+ . . t i . r o+ . . f u e+ . r a . ///
*TEA: ¿ . k i e+ n . ? < ¿ . o+ . t r o . . n i+ . N o . ? >
*INE: [< ] < . a . b e+ l . > ///
*TEA: ¿ . s i+ . ? ¡ahí va! . s i+ . / . p u e+ s . . l o+ . . y e
. b a+ . b a . . i+ . . l e+ . . d i+ . x o . . l a+ . . p r o .
f e . s o+ . r a . // . g u a+ r . d a . t e . l o . . e+ n . . e+
l . . b a+ . b i . ///
*INE: ¿ . i+ . . s a+ . b e s . . k e+ . . l e+ . . e n . k o n .
t r e+ . . e+ l . [/] . e+ n . . e+ l . . b o l . s i+ . y o . ?
*TEA: ¿ . e+ l . . k e+ . ?
```

*INE: . l e+ . . e n . k o n . t r e+ . . u+ n . . C u . p e+ . t e . ///
*TEA: ¿ . u+ n . . C u . p e+ . t e . ?
*INE: ¡ . s i+ . !
*TEA: . b u e+ . n o . / . p e+ . r o . . t o . d a . b i+ . a . . e+ s . . C i . k i . t i+ n . / ¿ . n o+ . ?
*INE: . y a+ . / . p e+ . r o . / . e+ . s o . . e+ s . . k u a+ n . d o . . s e+ . a . . b e . b e+ . ///
*TEA: . k u a+ n . d o . . e+ s . . b e . b e+ . . s o+ . l o . ///
*INE: . l e+ . . b i+ . . u+ n . . C u . p e+ . t e . hhh ///
%act: (5) laugh
*TEA: . b u e+ . n o . / . a+ i . . n i+ . N o s . . k e+ . - > . t i e+ . n e n . . e+ l . . C u . p e+ . t e . - > / . m u+ . C o . . t i e+ m . p o . ///
*INE: . s i+ . ///
*TEA: . y a+ . . k u a+ n . d o . . s o+ n . . m a . y o+ . r e s . +
*INE: . p e+ . r o . / . a . b e+ l . . n o+ . . l o+ . . y e+ . b a . . p u e+ s . t o . ///
*TEA: ¿ . n o+ . ? . s o+ . l o . . l o+ . . y e+ . b a . . e+ n . . e+ l . . b o l . s i+ . y o . / . p o+ r . . s i+ . . a . k a+ . s o . /// ¿ . n o+ . ?
*INE: . s i+ . . n u+ n . k a . . l o+ . . n e . T e . s i+ . t a . ///
*TEA: . n o+ . /// hhh /// . p u e+ s . . n a+ . d a . / . m u i+ . . b i e+ n . . i . n e+ s . /// . n o+ s . . b a+ . m o s . . a+ . . i+ r . . p a+ . r a . . k l a+ . s e . / ¿ . b a+ . l e . ?
%act: (2) laugh
*INE: . b a+ . l e . ///
*TEA: ¡ale! . b a+ . m o . n o s . ///

Anotación morfosintáctica

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- pre-etiquetado tradlsi: mar may 15 11:38:11 CEST 2007 -->
<!-- Etiquetado por marta: mar jul  3 12:07:46 CEST 2007 -->
<fml>
<Header>
<Title> Inés y Marta </Title>
<File> INE5 </File>
<Participants>
<Speaker>
<ShortName> INE </ShortName>
<Sex Type="woman"/>
<Age Type="5:2"/>
<Education Type="1"/>
<Role> child </Role>
<Origin> Ciudad Real </Origin>
</Speaker>
<Speaker>
<ShortName> TEA </ShortName>
<Sex Type="woman"/>
<Age Type="B"/>
<Education Type="3"/>
<Role> adult </Role>
<Origin> Madrid </Origin>
</Speaker>
</Participants>
<Date> 21/02/2006 </Date>
<Place> Ciudad Real </Place>
<Situation> conversation in an empty classroom at school. </Situation>
<Topic> daily matters </Topic>
<Source> CHIEDE </Source>
<Class Type1="informal" Type2="family/private" Type3="dialogue"/>
<Length> 8´36" </Length>
<Words> 1067 </Words>
<Acoustic_quality Type="A"/>
<Transcriber> Marta </Transcriber>
<Revisor> Ana and Marta </Revisor>
<Comments> INE (middle class; birth order: 1st) </Comments>
</Header>
<Text>
<p>
<f h="TEA" st="0.0" et="7.848" id="1">
<sf t="int" id="1-1">
<w cat="PUNCT" lem="¿" id="1-1-1"> ¿ </w>
<w cat="P" lem="cuántos" gen="masc" id="1-1-2"> cuántos </w>
<w cat="N" lem="año" gen="masc" num="plu" id="1-1-3"> años </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="1-1-4"> tienes </w>
<w cat="NPR" lem="Inés" id="1-1-5"> Inés </w>
<w cat="PUNCT" lem="?" id="1-1-6"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="7.848" et="8.996" id="2">
<sf t="int" id="2-1">
<w cat="PUNCT" lem="¿" id="2-1-1"> ¿ </w>
<w cat="Q" lem="cinco" id="2-1-2"> cinco </w>
<w cat="PUNCT" lem="?" id="2-1-3"> ? </w>
</sf>
```

```
</f>
</p>
<p>
<f h="INE" st="8.996" et="9.352" id="3">
<sf t="enu" id="3-1">
<w cat="ADV" lem="sí" id="3-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="9.352" et="9.917" id="4">
<sf t="int" id="4-1">
<w cat="PUNCT" lem="¿" id="4-1-1"> ¿ </w>
<w cat="C" lem="y" id="4-1-2"> y </w>
<w cat="P" lem="cuándo" id="4-1-3"> cuándo </w>
<w cat="P" lem="lo" gen="masc" num="plu" per="3" id="4-1-4"> los </w>
<w cat="V" lem="cumplir" tie="pres_ind" num="sing" per="2" id="4-1-5"> cumples </w>
<w cat="PUNCT" lem="?" id="4-1-6"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="9.917" et="10.556" id="5">
<sf t="int" id="5-1">
<w cat="PUNCT" lem="¿" id="5-1-1"> ¿ </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="5-1-2"> lo </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="2"id="5-1-3"> sabes </w>
<w cat="PUNCT" lem="?"id="5-1-4"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="10.556" et="11.358" id="6">
<sf t="enu" id="6-1">
<w cat="ADV" lem="no" id="6-1-1"> no </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="11.358" et="11.787" id="7">
<sf t="int" id="7-1">
<w cat="PUNCT" lem="¿" id="7-1-1"> ¿ </w>
<w cat="ADV" lem="no"id="7-1-2"> no </w>
<w cat="PUNCT" lem="?"id="7-1-3"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="11.787" et="13.135" id="8">
<sf t="enu" id="8-1">
<w cat="PREP" lem="en" id="8-1-1"> en </w>
<w cat="N" lem="diciembre" gen="masc" num="sing" id="8-1-2"> diciembre </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="13.135" et="14.51" id="9">
<sf t="exc" id="9-1">
<w cat="PUNCT" lem="¡" id="9-1-1"> ¡ </w>
<w cat="INTJ" lem="ups" id="9-1-2"> ups </w>
```

```
<w cat="PUNCT" lem="!" id="9-1-3"> ! </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="14.51" et="16.526" id="10">
<sf t="enu" id="10-1">
<w cat="ADV" lem="todavía" id="10-1-1"> todavía </w>
<w cat="P" lem="te" num="sing" per="2" id="10-1-2"> te </w>
<w cat="V" num="sing" lem="quedar" id="10-1-3"> queda </w>
<w cat="Q" lem="mucho" gen="masc" id="10-1-4"> mucho </w>
<ps t="complex"/>
<w cat="PREP" lem="para" id="10-1-5"> para </w>
<w cat="ART" lem="el" gen="masc" num="plu" id="10-1-6"> los </w>
<w cat="N" lem="seis" id="10-1-7"> seis </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="16.526" et="22.932" id="11">
<sf t="enu" id="11-1">
<w cat="P" lem="yo" num="sing" per="1" id="11-1-1"> yo </w>
<w cat="P" lem="lo" gen="masc" num="plu" per="3" id="11-1-2"> los </w>
<w cat="V" lem="cumplir" tie="indf_ind" num="sing" per="1" id="11-1-3"> cumplí </w>
<w cat="P" lem="antes" id="11-1-4"> antes </w>
<ps t="complex"/>
<w cat="C" lem="pero" id="11-1-5"> pero </w>
<ps/>
<w cat="ART" lem="el" gen="masc" num="plu" id="11-1-6"> los </w>
<frag>
</frag>
<ps t="pr"/>
<w cat="ART" lem="el" gen="masc" num="plu" id="11-1-7"> los </w>
<ps t="pr"/>
<w cat="ART" lem="el" gen="masc" num="plu" id="11-1-8"> los </w>
<w cat="N" lem="seis" id="11-1-9"> seis </w>
<w cat="P" lem="me" num="sing" per="1" id="11-1-10"> me </w>
<w cat="V" num="sing" lem="quedar" id="11-1-11"> queda </w>
<w cat="Q" lem="mucho" gen="masc" id="11-1-12"> mucho </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="22.932" et="23.723" id="12">
<sf t="enu" id="12-1">
<w cat="MD" lem="claro" id="12-1-1"> claro </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="23.723" et="24.723" id="13">
<sf t="enu" id="13-1">
<nl/>
</sf>
</f>
</p>
<p>
<f h="TEA" st="24.723" et="26.346" id="14">
<sf t="int" id="14-1">
<w cat="PUNCT" lem="¿" id="14-1-1"> ¿ </w>
```

<w cat="C" lem="y" id="14-1-2"> y </w>
<w cat="POSS" lem="tu" num="sing" id="14-1-3"> tu </w>
<w cat="N" lem="hermanito" gen="masc" num="sing" id="14-1-4"> hermanito </w>
<w cat="P" lem="cuántos" gen="masc" id="14-1-5"> cuántos </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="3" id="14-1-6"> tiene </w>
<w cat="PUNCT" lem="?" id="14-1-7"> ? </w>
</sf>
</f>

<f h="INE" st="26.346" et="29.647" id="15">
<sf t="enu" id="15-1">
<w cat="P" lem="me" num="sing" per="1" id="15-1-1"> me </w>
<w cat="V" lem="parecer" num="sing" id="15-1-2"> parece </w>
<w cat="C" lem="que" id="15-1-3"> que </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="3" id="15-1-4"> tiene </w>
<w cat="N" lem="tres" id="15-1-5"> tres </w>
</sf>
</f>

<f h="INE" st="29.647" et="31.21" id="16">
<sf t="enu" id="16-1">
<w cat="Q" lem="tres" id="16-1-1"> tres </w>
<w cat="N" lem="añito" gen="masc" num="plu" id="16-1-2"> añitos </w>
</sf>
</f>

<f h="TEA" st="31.21" et="32.377" id="17">
<sf t="int" id="17-1">
<w cat="PUNCT" lem="¿" id="17-1-1"> ¿ </w>
<w cat="C" lem="y" id="17-1-2"> y </w>
<w cat="P" lem="él" gen="masc" num="sing" per="3" id="17-1-3"> él </w>
<w cat="P" lem="cuándo" id="17-1-4"> cuándo </w>
<w cat="P" lem="lo" gen="masc" num="plu" per="3" id="17-1-5"> los </w>
<w cat="V" lem="cumplir" num="sing" id="17-1-6"> cumple </w>
<w cat="PUNCT" lem="?" id="17-1-7"> ? </w>
</sf>
</f>

<f h="INE" st="32.377" et="35.155" id="18">
<sf t="enu" id="18-1">
<w cat="ART" lem="el" gen="masc" num="plu" id="18-1-1"> los </w>
<frag>
</frag>
<ps t="pr"/>
</sf>
<sf t="int" id="18-2">
<w cat="PUNCT" lem="¿" id="18-2-1"> ¿ </w>
<w cat="ART" lem="el" gen="masc" num="plu" id="18-2-2"> los </w>
<w cat="N" lem="cuatro" id="18-2-3"> cuatro </w>
<w cat="PUNCT" lem="?" id="18-2-4"> ? </w>
</sf>
</f>

<f h="INE" st="35.155" et="36.068" id="19">
<sf t="enu" id="19-1">

<w cat="ADV" lem="no" id="19-1-1"> no </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="1" id="19-1-2"> sé </w>
</sf>
</f>

<f h="TEA" st="36.068" et="37.268" id="20">
<sf t="int" id="20-1">
<w cat="PUNCT" lem="¿" id="20-1-1"> ¿ </w>
<w cat="MD" lem="no" id="20-1-2"> no </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="2"id="20-1-3"> sabes </w>
<w cat="PUNCT" lem="?"id="20-1-4"> ? </w>
</sf>
</f>

<f h="TEA" st="37.268" et="38.301" id="21">
<sf t="enu" id="21-1">
<w cat="P" lem="yo" num="sing" per="1" id="21-1-1"> yo </w>
<w cat="ADV" lem="tampoco" id="21-1-2"> tampoco </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="21-1-3"> lo </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="1" id="21-1-4"> sé </w>
</sf>
</f>

<f h="TEA" st="38.301" et="41.509" id="22">
<sf t="enu" id="22-1">
<w cat="MD" lem="bueno" id="22-1-1"> bueno </w>
<ps/>
<w cat="C" lem="y" id="22-1-2"> y </w>
<w cat="MD" lem="a ver" id="22-1-3"> a ver </w>
</sf>
</f>

<f h="TEA" st="41.509" et="43.308" id="23">
<sf t="int" id="23-1">
<w cat="PUNCT" lem="¿" id="23-1-1"> ¿ </w>
<w cat="P" lem="qué" id="23-1-2"> qué </w>
<w cat="V" lem="estar" tie="impf_ind" num="plu" per="2"id="23-1-3"> estabais </w>
<ps t="tr"/>
<w cat="P" lem="qué" id="23-1-4"> qué </w>
<w cat="AUX" lem="haber" tie="pres_ind" num="plu" per="2" id="23-1-5"> habéis </w>
<w cat="V" lem="hacer" tie="part" id="23-1-6"> hecho </w>
<w cat="P" lem="hoy" id="23-1-7"> hoy </w>
<w cat="PREP" lem="con" id="23-1-8"> con </w>
<w cat="NPR" lem="Candi" id="23-1-9"> Candi </w>
<w cat="PUNCT" lem="?" id="23-1-10"> ? </w>
</sf>
</f>

<f h="TEA" st="43.308" et="44.332" id="24">
<sf t="enu" id="24-1">
<w cat="PREP" lem="en" id="24-1-1"> en </w>
<w cat="N" lem="clase" gen="fem" num="sing" id="24-1-2"> clase </w>
</sf>
</f>

```
<p>
<f h="INE" st="44.332" et="46.915" id="25">
<sf t="enu" id="25-1">
<w cat="AUX" lem="haber" tie="pres_ind" num="plu" per="1" id="25-1-1"> hemos </w>
<w cat="V" lem="hacer" tie="part" id="25-1-2"> hecho </w>
<w cat="N" lem="ficha" gen="fem" num="plu" id="25-1-3"> fichas </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="46.915" et="50.45" id="26">
<sf t="enu" id="26-1">
<w cat="AUX" lem="haber" tie="pres_ind" num="sing" per="3" id="26-1-1"> ha </w>
<w cat="V" lem="venir" tie="part" id="26-1-2"> venido </w>
<w cat="NPR" lem="Jesús" id="26-1-3"> Jesús </w>
<w cat="C" lem="y" id="26-1-4"> y </w>
<w cat="AUX" lem="haber" tie="pres_ind" num="plu" per="1" id="26-1-5"> hemos </w>
<w cat="V" lem="hacer" tie="part" id="26-1-6"> hecho </w>
<w cat="ART" lem="el" gen="fem" num="plu" id="26-1-7"> las </w>
<w cat="N" lem="careta" gen="fem" num="plu" id="26-1-8"> caretas </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="50.45" et="51.844" id="27">
<sf t="int" id="27-1">
<w cat="PUNCT" lem="¿" id="27-1-1"> ¿ </w>
<w cat="C" lem="que" id="27-1-2"> que </w>
<w cat="NPR" lem="Jesús" id="27-1-3"> Jesús </w>
<w cat="P" lem="qué" id="27-1-4"> qué </w>
<w cat="P" lem="os" num="plu" per="2" id="27-1-5"> os </w>
<w cat="V" lem="dar" num="sing" id="27-1-6"> da </w>
<w cat="PUNCT" lem="?" id="27-1-7"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="51.844" et="55.905" id="28">
<sf t="enu" id="28-1">
<w cat="P" lem="nos" num="plu" per="1" id="28-1-1"> nos </w>
<w cat="V" lem="dar" num="sing" id="28-1-2"> da </w>
<w cat="Q" lem="un" gen="fem" id="28-1-3"> unas </w>
<w cat="N" lem="cosa" gen="fem" num="plu" id="28-1-4"> cosas </w>
<w cat="PREP" lem="para" id="28-1-5"> para </w>
<w cat="V" lem="pintar" tie="inf" id="28-1-6"> pintar </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="55.905" et="56.305" id="29">
<sf t="int" id="29-1">
<w cat="PUNCT" lem="¿" id="29-1-1"> ¿ </w>
<w cat="PREP" lem="en" id="29-1-2"> en </w>
<w cat="P" lem="qué" id="29-1-3"> qué </w>
<w cat="N" lem="idioma" gen="masc" num="sing" id="29-1-4"> idioma </w>
<w cat="PUNCT" lem="?" id="29-1-5"> ? </w>
</sf>
</f>
</p>
<p>
```

<f h="INE" st="56.305" et="60.189" id="30">
<sf t="enu" id="30-1">
<w cat="Q" lem="algún" gen="fem" id="30-1-1"> algunas </w>
<w cat="Q" lem="muy" id="30-1-2"> muy </w>
<w cat="ADJ" lem="grande" num="plu" id="30-1-3"> grandes </w>
<ps/>
<w cat="C" lem="y" id="30-1-4"> y </w>
<w cat="P" lem="nos" num="plu" per="1" id="30-1-5"> nos </w>
<w cat="V" lem="cansar" num="plu" per="1" id="30-1-6"> cansamos </w>
<w cat="Q" lem="un poco" id="30-1-7"> un poco </w>
</sf>
</f>

<f h="TEA" st="60.189" et="61.036" id="31">
<sf t="int" id="31-1">
<w cat="PUNCT" lem="¿" id="31-1-1"> ¿ </w>
<w cat="ADV" lem="sí" id="31-1-2"> sí </w>
<w cat="PUNCT" lem="?" id="31-1-3"> ? </w>
</sf>
</f>

<f h="TEA" st="61.036" et="61.906" id="32">
<sf t="int" id="32-1">
<w cat="PUNCT" lem="¿" id="32-1-1"> ¿ </w>
<w cat="V" lem="pintar" tie="inf" id="32-1-2"> pintar </w>
<w cat="PUNCT" lem="?" id="32-1-3"> ? </w>
</sf>
</f>

<f h="TEA" st="61.906" et="62.814" id="33">
<sf t="int" id="33-1">
<w cat="PUNCT" lem="¿" id="33-1-1"> ¿ </w>
<w cat="Q" lem="mucho" gen="masc" id="33-1-2"> mucho </w>
<w cat="V" lem="pintar" tie="inf" id="33-1-3"> pintar </w>
<w cat="PUNCT" lem="?" id="33-1-4"> ? </w>
</sf>
</f>

<f h="INE" st="61.906" et="62.814" id="34">
<sf t="enu" id="34-1">
<bw/>
<w cat="ADV" lem="sí" id="34-1-1"> sí </w>
</sf>
</f>

<f h="TEA" st="62.814" et="65.682" id="35">
<sf t="enu" id="35-1">
<w cat="C" lem="pero" id="35-1-1"> pero </w>
<w cat="NPR" lem="Jesús" id="35-1-2"> Jesús </w>
<w cat="P" lem="os" num="plu" per="2" id="35-1-3"> os </w>
<w cat="V" lem="dar" num="sing" id="35-1-4"> da </w>
<ps t="tr"/>
</sf>
<sf t="int" id="35-2">
<w cat="PUNCT" lem="¿" id="35-2-1"> ¿ </w>

```
<w cat="P" lem="qué" id="35-2-2"> qué </w>
<w cat="P" lem="os" num="plu" per="2" id="35-2-3"> os </w>
<w cat="V" lem="dar" num="sing" id="35-2-4"> da </w>
<w cat="NPR" lem="Jesús" id="35-2-5"> Jesús </w>
<w cat="PUNCT" lem="?" id="35-2-6"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="65.682" et="67.234">
<sf t="enu">
<w cat="NPR" lem="Jesús"> Jesús </w>
<ps t="interrup"/>
</sf>
</f>
</p>
<p>
<f h="TEA" st="67.234" et="68.205" id="36">
<sf t="enu" id="36-1">
<w cat="ART" lem="el" gen="fem" num="sing" id="36-1-1"> la </w>
<w cat="N" lem="asignatura" gen="fem" num="sing" id="36-1-2"> asignatura </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="68.205" et="69.97" id="37">
<sf t="int" id="37-1">
<w cat="PUNCT" lem="¿" id="37-1-1"> ¿ </w>
<w cat="NPR" lem="Arturo" id="37-1-2"> Arturo </w>
<w cat="PUNCT" lem="?" id="37-1-3"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="69.97" et="71.546" id="38">
<sf t="enu" id="38-1">
<w cat="ART" lem="el" gen="fem" num="sing" id="38-1-1"> la </w>
<frag>
</frag>
<ps t="pr"/>
<w cat="N" lem="asignatura" gen="fem" num="sing" id="38-1-2"> asignatura </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="71.546" et="72.777" id="39">
<sf t="enu" id="39-1">
<w cat="ADV" lem="no" id="39-1-1"> no </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="39-1-2"> lo </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="1" id="39-1-3"> sé </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="72.777" et="73.315" id="40">
<sf t="int" id="40-1">
<w cat="PUNCT" lem="¿" id="40-1-1"> ¿ </w>
<w cat="ADV" lem="no" id="40-1-2"> no </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="40-1-3"> lo </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="2"id="40-1-4"> sabes </w>
```

```
<w cat="PUNCT" lem="?"id="40-1-5"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="73.315" et="74.097" id="41">
<sf t="int" id="41-1">
<w cat="PUNCT" lem="¿" id="41-1-1"> ¿ </w>
<w cat="P" lem="qué" id="41-1-2"> qué </w>
<w cat="N" lem="idioma" gen="masc" num="sing" id="41-1-3"> idioma </w>
<w cat="PUNCT" lem="?" id="41-1-4"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="74.097" et="76.455" id="42">
<sf t="enu" id="42-1">
<w cat="N" lem="idioma" gen="masc" num="sing" id="42-1-1"> idioma </w>
<ps/>
<w cat="N" lem="inglés" gen="masc" num="sing" id="42-1-2"> inglés </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="76.455" et="77.516" id="43">
<sf t="exc" id="43-1">
<w cat="PUNCT" lem="¡" id="43-1-1"> ¡ </w>
<w cat="INTJ" lem="ah" id="43-1-2"> ah </w>
<w cat="PUNCT" lem="!" id="43-1-3"> ! </w>
<ps/>
<w cat="MD" lem="pues"> pues </w>
<w cat="P" lem="ése" num="sing" per="3"> eso </w>
</sf>
</p>
<p>
<f h="TEA" st="77.516" et="78.393" id="44">
<sf t="enu" id="44-1">
<w cat="N" lem="inglés" gen="masc" num="sing" id="44-1-1"> inglés </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="78.393" et="80.085" id="45">
<sf t="int" id="45-1">
<w cat="PUNCT" lem="¿" id="45-1-1"> ¿ </w>
<w cat="C" lem="y" id="45-1-2"> y </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="2" id="45-1-3"> sabes </w>
<w cat="V" lem="hablar" tie="inf" id="45-1-4"> hablar </w>
<w cat="PREP" lem="en" id="45-1-5"> en </w>
<w cat="N" lem="inglés" gen="masc" num="sing" id="45-1-6"> inglés </w>
<w cat="PUNCT" lem="?" id="45-1-7"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="80.085" et="80.796" id="46">
<sf t="enu" id="46-1">
<nl/>
</sf>
</f>
```

```
</p>
<p>
<f h="TEA" st="80.796" et="81.54" id="47">
<sf t="int" id="47-1">
<w cat="PUNCT" lem="¿" id="47-1-1"> ¿ </w>
<w cat="MD" lem="no ?" id="47-1-2"> no ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="81.54" et="84.434" id="48">
<sf t="enu" id="48-1">
<w cat="C" lem="pero" id="48-1-1"> pero </w>
<w cat="ADV" lem="sí" id="48-1-2"> sí </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="2" id="48-1-3"> sabes </w>
<w cat="V" lem="decir" tie="inf" id="48-1-4"> decir </w>
<w cat="INTJ" lem="hola" id="48-1-5"> hola </w>
<w cat="C" lem="y" id="48-1-6"> y </w>
<w cat="INTJ" lem="adiós" id="48-1-7"> adiós </w>
<ps/>
</sf>
<sf t="int" id="48-2">
<w cat="PUNCT" lem="¿" id="48-2-1"> ¿ </w>
<w cat="MD" lem="no ?" id="48-2-2"> no ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="84.434" et="84.989" id="49">
<sf t="enu" id="49-1">
<w cat="C" lem="cuando" id="49-1-1"> cuando </w>
<w cat="V" lem="venir" tie="pres_ind" num="sing" per="3" id="49-1-2"> viene </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="84.989" et="85.826" id="50">
<sf t="enu" id="50-1">
<w cat="C" lem="y" id="50-1-1"> y </w>
<w cat="DEM" lem="ese" gen="fem" num="plu" id="50-1-2"> esas </w>
<w cat="N" lem="cosa" gen="fem" num="plu" id="50-1-3"> cosas </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="85.826" et="87.443" id="51">
<sf t="enu" id="51-1">
<w id="51-1-1"> hello </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="87.443" et="88.347" id="52">
<sf t="enu" id="52-1">
<nl/>
</sf>
</f>
</p>
<p>
<f h="INE" st="88.347" et="90.014" id="53">
```

```
<sf t="enu" id="53-1">
<w id="53-1-1"> bye </w>
<w id="53-1-2"> bye </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="90.014" et="91.103" id="54">
<sf t="enu" id="54-1">
<w cat="Q" lem="muy" id="54-1-1"> muy </w>
<w cat="ADV" lem="bien" id="54-1-2"> bien </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="91.103" et="92.055" id="55">
<sf t="enu" id="55-1">
<w cat="MD" lem="claro" id="55-1-1"> claro </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="92.055" et="93.416" id="56">
<sf t="enu" id="56-1">
<w cat="Q" lem="poquito" gen="masc" id="56-1-1"> poquito </w>
<w cat="PREP" lem="a" id="56-1-2"> a </w>
<w cat="Q" lem="poco" gen="masc" id="56-1-3"> poco </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="93.416" et="96.839" id="57">
<sf t="sup" id="57-1">
<w cat="ADV" lem="ya" id="57-1-1"> ya </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="57-1-2"> el </w>
<w cat="N" lem="año" gen="masc" num="sing" id="57-1-3"> año </w>
<w cat="REL" lem="que" id="57-1-4"> que </w>
<w cat="V" lem="venir" tie="pres_ind" num="sing" per="3" id="57-1-5"> viene </w>
<w cat="V" lem="saber" tie="fut_ind" num="sing" per="2" id="57-1-6"> sabrás </w>
<w cat="Q" lem="más" id="57-1-7"> más </w>
<ps/>
<w cat="C" lem="y" id="57-1-8"> y </w>
<w cat="P" lem="luego" id="57-1-9"> luego </w>
<w cat="Q" lem="más" id="57-1-10"> más </w>
<ps/>
<w cat="C" lem="y" id="57-1-11"> y </w>
<w cat="P" lem="luego" id="57-1-12"> luego </w>
<w cat="Q" lem="más" id="57-1-13"> más </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="96.839" et="97.776" id="58">
<sf t="enu" id="58-1">
<w cat="ADV" lem="sí" id="58-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="97.776" et="98.113" id="59">
```

```
<sf t="enu" id="59-1">
<nl/>
</sf>
</f>
</p>
<p>
<f h="INE" st="98.113" et="98.839" id="60">
<sf t="enu" id="60-1">
<w cat="ADV" lem="sí" id="60-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="98.839" et="100.805" id="61">
<sf t="enu" id="61-1">
<w cat="C" lem="que" id="61-1-1"> que </w>
<w cat="P" lem="te" num="sing" per="2" id="61-1-2"> te </w>
<w cat="V" lem="gustar" num="sing" id="61-1-3"> gusta </w>
<w cat="V" lem="contar" tie="inf" id="61-1-4"> contar </w>
<w cat="N" lem="cuento" gen="masc" num="plu" id="61-1-5"> cuentos </w>
<ps/>
</sf>
<sf t="int" id="61-2">
<w cat="PUNCT" lem="¿" id="61-2-1"> ¿ </w>
<w cat="PREP" lem="a" id="61-2-2"> a </w>
<w cat="C" lem="que" id="61-2-3"> que </w>
<w cat="ADV" lem="sí" id="61-2-4"> sí </w>
<w cat="PUNCT" lem="?" id="61-2-5"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="100.805" et="105.723" id="62">
<sf t="sup" id="62-1">
<unint/>
<w cat="PREP" lem="en" id="62-1-1"> en </w>
<w cat="ART" lem="el" gen="fem" num="plu" id="62-1-2"> las </w>
<w cat="N" lem="asamblea" gen="fem" num="plu" id="62-1-3"> asambleas </w>
<w cat="C" lem="en cuanto" id="62-1-4"> en cuanto </w>
<w cat="V" lem="poder" tie="pres_ind" num="sing" per="2" id="62-1-5"> puedes </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="105.723" et="106.832" id="63">
<sf t="enu" id="63-1">
<w cat="Q" lem="un" gen="masc" id="63-1-1"> un </w>
<w cat="N" lem="cuento" gen="masc" num="sing" id="63-1-2"> cuento </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="106.832" et="107.604" id="64">
<sf t="enu" id="64-1">
<w cat="ADV" lem="sí" id="64-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="107.604" et="108.642" id="65">
```

<sf t="int" id="65-1">
<w cat="PUNCT" lem="¿" id="65-1-1"> ¿ </w>
<w cat="C" lem="y" id="65-1-2"> y </w>
<w cat="P" lem="te" num="sing" per="2" id="65-1-3"> te </w>
<w cat="P" lem="lo" gen="masc" num="plu" per="3" id="65-1-4"> los </w>
<w cat="V" lem="inventar" tie="pres_ind" num="sing" per="2" id="65-1-5"> inventas </w>
<w cat="P" lem="tú" num="sing" per="2" id="65-1-6"> tú </w>
<w cat="PUNCT" lem="?" id="65-1-7"> ? </w>
</sf>
</f>

<f h="INE" st="108.642" et="109.3" id="66">
<sf t="enu" id="66-1">
<w cat="ADV" lem="sí" id="66-1-1"> sí </w>
</sf>
</f>

<f h="TEA" st="109.3" et="110.123" id="67">
<sf t="int" id="67-1">
<w cat="PUNCT" lem="¿" id="67-1-1"> ¿ </w>
<w cat="Q" lem="todo" gen="masc" id="67-1-2"> todo </w>
<w cat="PUNCT" lem="?" id="67-1-3"> ? </w>
</sf>
</f>

<f h="INE" st="110.123" et="110.853" id="68">
<sf t="enu" id="68-1">
<w cat="ADV" lem="sí" id="68-1-1"> sí </w>
</sf>
</f>

<f h="TEA" st="110.853" et="113.413" id="69">
<sf t="enu" id="69-1">
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="69-1-1"> tienes </w>
<w cat="N" lem="imaginación" gen="fem" num="sing" id="69-1-2"> imaginación </w>
<ps/>
</sf>
<sf t="int" id="69-2">
<w cat="PUNCT" lem="¿" id="69-2-1"> ¿ </w>
<dm>
</dm>
<w cat="PUNCT" lem="?" id="69-2-2"> ? </w>
</sf>
</f>

<f h="TEA" st="113.413" et="115.58" id="70">
<sf t="int" id="70-1">
<w cat="PUNCT" lem="¿" id="70-1-1"> ¿ </w>
<w cat="P" lem="me" num="sing" per="1" id="70-1-2"> me </w>
<w cat="V" lem="contar" tie="pres_ind" num="sing" per="2" id="70-1-3"> cuentas </w>
<w cat="Q" lem="uno" id="70-1-4"> uno </w>
<w cat="PREP" lem="a" id="70-1-5"> a </w>
<w cat="P" lem="mí" num="sing" per="1" id="70-1-6"> mí </w>
<w cat="PUNCT" lem="?" id="70-1-7"> ? </w>
</sf>

```
</f>
</p>
<p>
<f h="INE" st="115.58" et="116.359" id="71">
<sf t="enu" id="71-1">
<w cat="MD" lem="vale" id="71-1-1"> vale </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="116.359" et="117.343" id="72">
<sf t="enu" id="72-1">
<w cat="MD" lem="a ver" id="72-1-1"> a ver </w>
<ps/>
<w cat="V" lem="decir" tie="imper" num="sing" per="2" clitic="yes" id="72-1-2"> di </w>
<w cat="P" lem="me" num="sing" per="1" id="72-1-3"> me </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="117.343" et="121.458" id="73">
<sf t="enu" id="73-1">
<w cat="V" lem="haber" num="sing" tie="impf_ind" id="73-1-1"> había </w>
<w cat="Q" lem="un" gen="fem" id="73-1-2"> una </w>
<w cat="N" lem="vez" gen="fem" num="sing" id="73-1-3"> vez </w>
<w cat="Q" lem="un" gen="masc" id="73-1-4"> un </w>
<ps t="pr"/>
<w cat="Q" lem="un" gen="masc" id="73-1-5"> un </w>
<ps t="pr"/>
<w cat="Q" lem="un" gen="masc" id="73-1-6"> un </w>
<ps t="self_interrup"/>
<w cat="MD" lem="a ver" id="73-1-7"> a ver </w>
<w cat="P" lem="cuál" gen="sing" id="73-1-8"> cuál </w>
<w cat="V" lem="ser" tie="pres_ind" num="sing" per="3" id="73-1-9"> es </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="121.458" et="123.421" id="74">
<sf t="enu" id="74-1">
<w cat="NPR" lem="Caperucita Roja" id="74-1-1"> Caperucita Roja </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="123.421" et="123.88" id="75">
<sf t="enu" id="75-1">
<w cat="MD" lem="vale" id="75-1-1"> vale </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="123.88" et="127.852" id="76">
<sf t="enu" id="76-1">
<w cat="V" lem="haber" num="sing" tie="impf_ind" id="76-1-1"> había </w>
<w cat="Q" lem="un" gen="fem" id="76-1-2"> una </w>
<w cat="N" lem="vez" gen="fem" num="sing" id="76-1-3"> vez </w>
<w cat="Q" lem="un" gen="fem" id="76-1-4"> una </w>
<w cat="N" lem="casita" gen="fem" num="sing" id="76-1-5"> casita </w>
<w cat="PREP" lem="en" id="76-1-6"> en </w>
```

<w cat="ART" lem="el" gen="masc" num="sing" id="76-1-7"> el </w>
<w cat="N" lem="bosque" gen="masc" num="sing" id="76-1-8"> bosque </w>
<ps/>
<w cat="V" lem="vivir" num="sing" tie="impf_ind" id="76-1-9"> vivía </w>
<w cat="NPR" lem="Caperucita Roja" id="76-1-10"> Caperucita Roja </w>
</sf>
</f>

<f h="INE" st="127.852" et="131.604" id="77">
<sf t="enu" id="77-1">
<w cat="C" lem="y" id="77-1-1"> y </w>
<w cat="P" lem="lo" num="sing" per="3" id="77-1-2"> le </w>
<w cat="V" lem="decir" tie="indf_ind" num="sing" per="3" id="77-1-3"> dijo </w>
<w cat="POSS" lem="su" num="sing" id="77-1-4"> su </w>
<w cat="N" lem="mamá" gen="fem" num="sing" id="77-1-5"> mamá </w>
<ps t="complex"/>
<w cat="V" lem="venir" tie="imper" num="sing" per="2"id="77-1-6"> ven </w>
<w cat="NPR" lem="Caperucita Roja" id="77-1-7"> Caperucita Roja </w>
</sf>
</f>

<f h="INE" st="131.604" et="138.027" id="78">
<sf t="enu" id="78-1">
<w cat="V" lem="llevar" tie="imper" num="sing" per="2" clitic="yes" id="78-1-1"> lléva </w>
<w cat="P" lem="lo" num="sing" per="3" id="78-1-2"> le </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="78-1-3"> la </w>
<w cat="N" lem="comida" gen="fem" num="sing" id="78-1-4"> comida </w>
<w cat="PREP" lem="a" id="78-1-5"> a </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="78-1-6"> la </w>
<w cat="N" lem="abuelo" gen="fem" num="sing" id="78-1-7"> abuela </w>
<ps t="complex"/>
<w cat="REL" lem="que" id="78-1-8"> que </w>
<w cat="ADV" lem="no" id="78-1-9"> no </w>
<w cat="AUX" lem="poder" num="sing" id="78-1-10"> puede </w>
<w cat="V" lem="venir" tie="inf" id="78-1-11"> venir </w>
<ps t="complex"/>
<w cat="V" lem="estar" num="sing" id="78-1-12"> está </w>
<w cat="ADJ" lem="cojo" gen="fem" num="sing" id="78-1-13"> coja </w>
<ps/>
<w cat="Q" lem="muy" id="78-1-14"> muy </w>
<w cat="ADJ" lem="malito" gen="fem" num="sing" id="78-1-15"> malita </w>
</sf>
</f>

<f h="INE" st="138.027" et="141.079" id="79">
<sf t="enu" id="79-1">
<w cat="MD" lem="entonces" id="79-1-1"> entonces </w>
<w cat="V" lem="decir" tie="indf_ind" num="sing" per="3" id="79-1-2"> dijo </w>
<frag>
</frag>
<ps t="pr"/>
<w cat="NPR" lem="Caperucita" id="79-1-3"> Caperucita </w>
<ps/>
<w cat="MD" lem="vale" id="79-1-4"> vale </w>
</sf>
</f>

```
<p>
<f h="INE" st="141.079" et="151.902" id="80">
<sf t="enu" id="80-1">
<w cat="V" lem="coger" tie="indf_ind" num="sing" per="3" id="80-1-1"> cogió </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="80-1-2"> la </w>
<w cat="N" lem="cesta" gen="fem" num=" sing"id="80-1-3"> cestita </w>
<ps t="complex"/>
<w cat="P" lem="se" id="80-1-4"> se </w>
<w cat="V" lem="poner" tie="indf_ind" num="sing" per="3" id="80-1-5"> puso </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="80-1-6"> la </w>
<w cat="N" lem="caperuza" gen="fem" num="sing" id="80-1-7"> caperuza </w>
<ps t="complex"/>
<w cat="C" lem="y" id="80-1-8"> y </w>
<nl/>
<w cat="V" lem="ir" tie="pres_ind" num="sing" per="1" id="80-1-9"> voy </w>
<w cat="PREP" lem="a" id="80-1-10"> a </w>
<w cat="POSS" lem="mi" num="sing" id="80-1-11"> mi </w>
<w cat="N" lem="casita" gen="fem" num="sing" id="80-1-12"> casita </w>
<ps/>
<nl/>
</sf>
</f>
</p>
<p>
<f h="INE" st="151.902" et="156.139" id="81">
<sf t="enu" id="81-1">
<w cat="MD" lem="entonces" id="81-1-1"> entonces </w>
<ps/>
<w cat="V" lem="oír" tie="indf_ind" num="sing" per="3" id="81-1-2"> oyó </w>
<w cat="Q" lem="un" gen="fem" id="81-1-3"> una </w>
<w cat="N" lem="cosa" gen="fem" num="sing" id="81-1-4"> cosa </w>
<w cat="P" lem="detrás" id="81-1-5"> detrás </w>
<w cat="PREP" lem="de" id="81-1-6"> de </w>
<w cat="ART" lem="el" gen="masc" num="plu" id="81-1-7"> los </w>
<w cat="N" lem="arbusto" gen="masc" num="plu" id="81-1-8"> arbustos </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="156.139" et="156.752" id="82">
<sf t="int" id="82-1">
<w cat="PUNCT" lem="¿" id="82-1-1"> ¿ </w>
<w cat="P" lem="dónde" id="82-1-2"> dónde </w>
<w cat="V" lem="estar" num="sing" tie="impf_ind" id="82-1-3"> estaba </w>
<w cat="PUNCT" lem="?" id="82-1-4"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="156.752" et="159.161" id="83">
<sf t="enu" id="83-1">
<nl/>
</sf>
</f>
</p>
<p>
<f h="TEA" st="159.161" et="160.705" id="84">
<sf t="enu" id="84-1">
<w cat="V" lem="estar" num="sing" tie="impf_ind" id="84-1-1"> estaba </w>
<ps t="tr"/>
```

```
</sf>
<sf t="int" id="84-2">
<w cat="PUNCT" lem="¿" id="84-2-1"> ¿ </w>
<w cat="P" lem="dónde" id="84-2-2"> dónde </w>
<w cat="V" lem="estar" num="sing" tie="impf_ind" id="84-2-3"> estaba </w>
<w cat="NPR" lem="Caperucita" id="84-2-4"> Caperucita </w>
<w cat="PUNCT" lem="?" id="84-2-5"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="160.705" et="161.534" id="85">
<sf t="enu" id="85-1">
<w cat="PREP" lem="en" id="85-1-1"> en </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="85-1-2"> el </w>
<w cat="N" lem="bosque" gen="masc" num="sing" id="85-1-3"> bosque </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="161.534" et="162.511" id="86">
<sf t="enu" id="86-1">
<w cat="PREP" lem="en" id="86-1-1"> en </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="86-1-2"> el </w>
<w cat="N" lem="bosque" gen="masc" num="sing" id="86-1-3"> bosque </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="162.511" et="165.026" id="87">
<sf t="enu" id="87-1">
<w cat="C" lem="y" id="87-1-1"> y </w>
<w cat="V" lem="oír" num="sing" tie="impf_ind" id="87-1-2"> oía </w>
<ps/>
<nl/>
</sf>
</f>
</p>
<p>
<f h="INE" st="165.026" et="168.738" id="88">
<sf t="enu" id="88-1">
<w cat="C" lem="y" id="88-1-1"> y </w>
<w cat="V" lem="salir" tie="indf_ind" num="sing" per="3" id="88-1-2"> salió </w>
<w cat="Q" lem="un" gen="masc" id="88-1-3"> un </w>
<w cat="N" lem="lobo" gen="masc" num="sing" id="88-1-4"> lobo </w>
<w cat="Q" lem="muy" id="88-1-5"> muy </w>
<ps/>
<w cat="ADJ" lem="feo" gen="masc" num="sing" id="88-1-6"> feo </w>
<w cat="Q" lem="muy" id="88-1-7"> muy </w>
<w cat="ADJ" lem="malo" gen="masc" num="sing" id="88-1-8"> malo </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="168.738" et="173.546" id="89">
<sf t="enu" id="89-1">
<w cat="C" lem="y" id="89-1-1"> y </w>
<w cat="P" lem="lo" num="sing" per="3" id="89-1-2"> le </w>
<w cat="V" lem="decir" tie="indf_ind" num="sing" per="3" id="89-1-3"> dijo </w>
<w cat="PREP" lem="a" id="89-1-4"> a </w>
```

<w cat="NPR" lem="Caperucita" id="89-1-5"> Caperucita </w>
<ps t="complex"/>
</sf>
<sf t="int" id="89-2">
<w cat="PUNCT" lem="¿" id="89-2-1"> ¿ </w>
<w cat="P" lem="dónde" id="89-2-2"> dónde </w>
<w cat="V" lem="ir" tie="pres_ind" num="sing" per="2" id="89-2-3"> vas </w>
<w cat="NPR" lem="Caperucita" id="89-2-4"> Caperucita </w>
<w cat="PUNCT" lem="?" id="89-2-5"> ? </w>
</sf>
</f>

<f h="INE" st="173.546" et="177.978" id="90">
<sf t="enu" id="90-1">
<w cat="V" lem="ir" tie="pres_ind" num="sing" per="1" id="90-1-1"> voy </w>
<w cat="PREP" lem="a" id="90-1-2"> a </w>
<w cat="N" lem="casa" gen="fem" num="sing" id="90-1-3"> casa </w>
<w cat="PREP" lem="de" id="90-1-4"> de </w>
<w cat="POSS" lem="mi" num="sing" id="90-1-5"> mi </w>
<w cat="N" lem="abuelito" gen="fem" num="sing" id="90-1-6"> abuelita </w>
<ps t="complex"/>
<w cat="PREP" lem="a" id="90-1-7"> a </w>
<w cat="V" lem="llevar" tie="inf" clitic="yes" id="90-1-8"> llevar </w>
<w cat="P" lem="lo" num="sing" per="3" id="90-1-9"> le </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="90-1-10"> la </w>
<w cat="N" lem="comida" gen="fem" num="sing"id="90-1-11"> comidita </w>
</sf>
</f>

<f h="INE" st="177.978" et="185.496" id="91">
<sf t="enu" id="91-1">
<w cat="MD" lem="entonces" id="91-1-1"> entonces </w>
<ps/>
<w cat="P" lem="tú" num="sing" per="2" id="91-1-2"> tú </w>
<w cat="V" lem="ir" tie="imper" num="sing" per="2"id="91-1-3"> ve </w>
<w cat="PREP" lem="por" id="91-1-4"> por </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="91-1-5"> el </w>
<w cat="N" lem="camino" gen="masc" num="sing" id="91-1-6"> camino </w>
<w cat="DEM" lem="ese" gen="masc" num="sing" id="91-1-7"> ese </w>
<w cat="Q" lem="más" id="91-1-8"> más </w>
<w cat="ADJ" lem="largo" gen="masc" num="sing" id="91-1-9"> largo </w>
<ps t="complex"/>
<w cat="C" lem="y" id="91-1-10"> y </w>
<w cat="P" lem="yo" num="sing" per="1" id="91-1-11"> yo </w>
<w cat="PREP" lem="por" id="91-1-12"> por </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="91-1-13"> el </w>
<w cat="Q" lem="más" id="91-1-14"> más </w>
<w cat="ADJ" lem="corto" gen="masc" num="sing" id="91-1-15"> corto </w>
<ps/>
<w cat="ADV" lem="así" id="91-1-16"> así </w>
<w cat="P" lem="tú" num="sing" per="2" id="91-1-17"> tú </w>
<w cat="V" lem="llegar" tie="fut_ind" num="sing" per="2" id="91-1-18"> llegarás </w>
<w cat="P" lem="antes" id="91-1-19"> antes </w>
</sf>
</f>

<f h="INE" st="185.496" et="188.247" id="92">

```
<sf t="enu" id="92-1">
<w cat="ART" lem="el" gen="masc" num="sing" id="92-1-1"> el </w>
<w cat="N" lem="lobo" gen="masc" num="sing" id="92-1-2"> lobo </w>
<ps/>
<w cat="P" lem="lo" gen="fem" num="sing" per="3" id="92-1-3"> la </w>
<w cat="V" lem="engañar" tie="indf_ind" num="sing" per="3" id="92-1-4"> engañó </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="188.247" et="194.259" id="93">
<sf t="enu" id="93-1">
<w cat="V" lem="saber" num="sing" tie="impf_ind" id="93-1-1"> sabía </w>
<ps/>
<w cat="C" lem="que" id="93-1-2"> que </w>
<w cat="C" lem="si" id="93-1-3"> si </w>
<w cat="V" lem="llegar" num="sing" tie="impf_ind" id="93-1-4"> llegaba </w>
<w cat="PREP" lem="por" id="93-1-5"> por </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="93-1-6"> el </w>
<w cat="N" lem="camino" gen="masc" num="sing" id="93-1-7"> camino </w>
<w cat="Q" lem="más" id="93-1-8"> más </w>
<w cat="ADJ" lem="corto" gen="masc" num="sing" id="93-1-9"> corto </w>
<ps/>
<w cat="V" lem="llegar" num="sing" tie="impf_ind" id="93-1-10"> llegaba </w>
<w cat="P" lem="antes" id="93-1-11"> antes </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="194.259" et="198.215" id="94">
<sf t="enu" id="94-1">
<w cat="MD" lem="entonces" id="94-1-1"> entonces </w>
<ps/>
<w cat="ART" lem="el" gen="masc" num="sing" id="94-1-2"> el </w>
<w cat="N" lem="lobo" gen="masc" num="sing" id="94-1-3"> lobo </w>
<w cat="V" lem="correr" num="sing" tie="impf_ind" id="94-1-4"> corría </w>
<w cat="Q" lem="más" id="94-1-5"> más </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="198.215" et="204.255" id="95">
<sf t="sup" id="95-1">
<w cat="C" lem="y" id="95-1-1"> y </w>
<w cat="NPR" lem="Caperucita" id="95-1-2"> Caperucita </w>
<w cat="P" lem="se" id="95-1-3"> se </w>
<w cat="V" lem="entretener" tie="indf_ind" num="sing" per="3" id="95-1-4"> entretuvo </w>
<ps/>
<w cat="V" lem="ver" tie="ger" id="95-1-5"> viendo </w>
<w cat="Q" lem="un" gen="fem" id="95-1-6"> unas </w>
<w cat="N" lem="mariposa" gen="fem" num="plu" id="95-1-7"> mariposas </w>
<ps/>
<w cat="V" lem="coger" tie="ger" id="95-1-8"> cogiendo </w>
<w cat="N" lem="flor" gen="fem" num="plu" id="95-1-9"> flores </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="204.255" et="210.92" id="96">
<sf t="enu" id="96-1">
```

<w cat="MD" lem="entonces" id="96-1-1"> entonces </w>
<ps/>
<w cat="C" lem="cuando" id="96-1-2"> cuando </w>
<w cat="V" lem="llamar" tie="indf_ind" num="sing" per="3" id="96-1-3"> llamó </w>
<w cat="PREP" lem="a" id="96-1-4"> a </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="96-1-5"> la </w>
<w cat="N" lem="puerta" gen="fem" num="sing" id="96-1-6"> puerta </w>
<w cat="V" lem="afinar" tie="ger" id="96-1-7"> afinando </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="96-1-8"> el </w>
<w cat="N" lem="lobo" gen="masc" num="sing" id="96-1-9"> lobo </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="96-1-10"> la </w>
<w cat="N" lem="voz" gen="fem" num="sing" id="96-1-11"> voz </w>
</sf>
</f>

<f h="INE" st="210.92" et="212.957" id="97">
<sf t="enu" id="97-1">
<w id="97-1-1"> tón </w>
<w id="97-1-2"> tón </w>
<ps/>
<w id="97-1-3"> tón </w>
<w id="97-1-4"> tón </w>
</sf>
</f>

<f h="INE" st="212.957" et="215.226" id="98">
<sf t="enu" id="98-1">
<w cat="V" num="sing" lem="pasar" id="98-1-1"> pasa </w>
<ps/>
<w cat="V" num="sing" lem="pasar" id="98-1-2"> pasa </w>
</sf>
</f>

<f h="INE" st="215.226" et="216.836" id="99">
<sf t="enu" id="99-1">
<w cat="ART" lem="el" gen="fem" num="sing" id="99-1-1"> la </w>
<w cat="N" lem="puerta" gen="fem" num="sing" id="99-1-2"> puerta </w>
<w cat="AUX" lem="estar" num="sing" id="99-1-3"> está </w>
<w cat="V" lem="abrir" tie="part" id="99-1-4"> abierta </w>
</sf>
</f>

<f h="INE" st="216.836" et="222.402" id="100">
<sf t="enu" id="100-1">
<w cat="C" lem="y" id="100-1-1"> y </w>
<w cat="C" lem="en cuanto" id="100-1-2"> en cuanto </w>
<w cat="V" lem="ver" tie="indf_ind" num="sing" per="3" id="100-1-3"> vio </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="100-1-4"> el </w>
<w cat="N" lem="lobo" gen="masc" num="sing" id="100-1-5"> lobo </w>
<ps t="complex"/>
<w cat="P" lem="se" id="100-1-6"> se </w>
<ps t="pr"/>
<w cat="P" lem="se" id="100-1-7"> se </w>
<w cat="P" lem="lo" gen="fem" num="sing" per="3" id="100-1-8"> la </w>
<w cat="V" lem="esconder" tie="indf_ind" num="sing" per="3" id="100-1-9"> escondió </w>
<w cat="PREP" lem="en" id="100-1-10"> en </w>

```
<w cat="ART" lem="el" gen="masc" num="sing" id="100-1-11"> el </w>
<w cat="N" lem="armario" gen="masc" num="sing" id="100-1-12"> armario </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="222.402" et="228.833" id="101">
<sf t="enu" id="101-1">
<w cat="C" lem="cuando" id="101-1-1"> cuando </w>
<w cat="V" lem="venir" tie="indf_ind" num="sing" per="3" id="101-1-2"> vino </w>
<w cat="NPR" lem="Caperucita" id="101-1-3"> Caperucita </w>
<ps/>
<w cat="V" lem="venir" tie="indf_ind" num="sing" per="3" id="101-1-4"> vino </w>
<ps/>
<w cat="C" lem="y" id="101-1-5"> y </w>
<w cat="P" lem="lo" num="sing" per="3" id="101-1-6"> le </w>
<w cat="V" lem="decir" tie="indf_ind" num="sing" per="3" id="101-1-7"> dijo </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="101-1-8"> la </w>
<ps t="tr"/>
</sf>
<sf t="exc" id="101-2">
<w cat="PUNCT" lem="¡" id="101-2-1"> ¡ </w>
<w cat="P" lem="qué" id="101-2-2"> qué </w>
<w cat="N" lem="ojo" gen="masc" num="plu" id="101-2-3"> ojos </w>
<w cat="Q" lem="más" id="101-2-4"> más </w>
<w cat="ADJ" lem="grande" num="plu" id="101-2-5"> grandes </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="101-2-6"> tienes </w>
<w cat="PUNCT" lem="!" id="101-2-7"> ! </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="228.833" et="232.14" id="102">
<sf t="exc" id="102-1">
<w cat="PUNCT" lem="¡" id="102-1-1"> ¡ </w>
<w cat="V" lem="ser" tie="pres_ind" num="plu" per="3" id="102-1-2"> son </w>
<w cat="PREP" lem="para" id="102-1-3"> para </w>
<w cat="V" lem="ver" tie="inf" clitic="yes" id="102-1-4"> ver </w>
<w cat="P" lem="te" num="sing" per="2" id="102-1-5"> te </w>
<w cat="ADV" lem="mejor" id="102-1-6"> mejor </w>
<w cat="PUNCT" lem="!" id="102-1-7"> ! </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="232.14" et="235.75" id="103">
<sf t="exc" id="103-1">
<w cat="PUNCT" lem="¡" id="103-1-1"> ¡ </w>
<w cat="P" lem="qué" id="103-1-2"> qué </w>
<w cat="N" lem="nariz" gen="fem" num="sing" id="103-1-3"> nariz </w>
<w cat="Q" lem="más" id="103-1-4"> más </w>
<w cat="ADJ" lem="grande" num="sing" id="103-1-5"> grande </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="103-1-6"> tienes </w>
<w cat="PUNCT" lem="!" id="103-1-7"> ! </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="235.75" et="239.57" id="104">
<sf t="exc" id="104-1">
```

```
<w cat="PUNCT" lem="¡" id="104-1-1"> ¡ </w>
<w cat="V" lem="ser" tie="pres_ind" num="plu" per="3" id="104-1-2"> son </w>
<w cat="PREP" lem="para" id="104-1-3"> para </w>
<w cat="V" lem="oler" tie="inf" clitic="yes" id="104-1-4"> oler </w>
<w cat="P" lem="te" num="sing" per="2" id="104-1-5"> te </w>
<w cat="ADV" lem="mejor" id="104-1-6"> mejor </w>
<w cat="PUNCT" lem="!" id="104-1-7"> ! </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="239.57" et="243.111" id="105">
<sf t="enu" id="105-1">
<w cat="C" lem="y" id="105-1-1"> y </w>
<w cat="P" lem="qué" id="105-1-2"> qué </w>
<w cat="N" lem="oreja" gen="fem" num="plu" id="105-1-3"> orejas </w>
<w cat="Q" lem="más" id="105-1-4"> más </w>
<w cat="ADJ" lem="grande" num="plu" id="105-1-5"> grandes </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="105-1-6"> tienes </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="243.111" et="245.809" id="106">
<sf t="exc" id="106-1">
<w cat="PUNCT" lem="¡" id="106-1-1"> ¡ </w>
<w cat="V" lem="ser" tie="pres_ind" num="plu" per="3" id="106-1-2"> son </w>
<w cat="PREP" lem="para" id="106-1-3"> para </w>
<w cat="V" lem="oír" tie="inf" clitic="yes" id="106-1-4"> oír </w>
<w cat="P" lem="te" num="sing" per="2" id="106-1-5"> te </w>
<w cat="ADV" lem="mejor" id="106-1-6"> mejor </w>
<w cat="PUNCT" lem="!" id="106-1-7"> ! </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="245.809" et="248.459" id="107">
<sf t="enu" id="107-1">
<w cat="C" lem="y" id="107-1-1"> y </w>
<w cat="P" lem="qué" id="107-1-2"> qué </w>
<w cat="N" lem="boca" gen="fem" num="sing" id="107-1-3"> boca </w>
<w cat="Q" lem="tan" id="107-1-4"> tan </w>
<w cat="ADJ" lem="grande" num="sing" id="107-1-5"> grande </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="107-1-6"> tienes </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="248.459" et="251.158" id="108">
<sf t="exc" id="108-1">
<w cat="PUNCT" lem="¡" id="108-1-1"> ¡ </w>
<w cat="V" lem="ser" tie="pres_ind" num="plu" per="3" id="108-1-2"> son </w>
<w cat="PREP" lem="para" id="108-1-3"> para </w>
<w cat="V" lem="comer" tie="inf" clitic="yes" id="108-1-4"> comer </w>
<w cat="P" lem="te" num="sing" per="2" id="108-1-5"> te </w>
<w cat="ADV" lem="mejor" id="108-1-6"> mejor </w>
<w cat="PUNCT" lem="!" id="108-1-7"> ! </w>
</sf>
</f>
</p>
```

```
<p>
<f h="INE" st="251.158" et="252.479" id="109">
<sf t="enu" id="109-1">
<w cat="P" lem="se" id="109-1-1"> se </w>
<w cat="V" lem="abalanzar" tie="indf_ind" num="sing" per="3" id="109-1-2"> abalanzó </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="252.479" et="256.031" id="110">
<sf t="enu" id="110-1">
<w cat="C" lem="y" id="110-1-1"> y </w>
<w cat="Q" lem="un" gen="masc" id="110-1-2"> un </w>
<unint/>
<ps/>
</sf>
<sf t="exc" id="110-2">
<w cat="PUNCT" lem="¡" id="110-2-1"> ¡ </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="110-2-2"> el </w>
<w cat="N" lem="lobo" gen="masc" num="sing" id="110-2-3"> lobo </w>
<w cat="INTJ" lem="pum" id="110-2-4"> pum </w>
<w cat="PUNCT" lem="!" id="110-2-5"> ! </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="256.031" et="261.555" id="111">
<sf t="enu" id="111-1">
<w cat="C" lem="y" id="111-1-1"> y </w>
<w cat="ADV" lem="no" id="111-1-2"> no </w>
<w cat="P" lem="se" id="111-1-3"> se </w>
<w cat="V" lem="dar" tie="indf_ind" num="sing" per="3" id="111-1-4"> dio </w>
<w cat="N" lem="cuenta" gen="fem" num="sing" id="111-1-5"> cuenta </w>
<ps/>
<w cat="C" lem="que" id="111-1-6"> que </w>
<w cat="V" lem="haber" num="sing" tie="impf_ind" id="111-1-7"> había </w>
<w cat="N" lem="fuego" gen="masc" num="sing" id="111-1-8"> fuego </w>
<w cat="PREP" lem="en" id="111-1-9"> en </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="111-1-10"> el </w>
<w cat="N" lem="culo" gen="masc" num="sing" id="111-1-11"> culo </w>
<w cat="POSS" lem="suyo" gen="masc" num="sing" id="111-1-12"> suyo </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="261.555" et="262.895" id="112">
<sf t="int" id="112-1">
<w cat="PUNCT" lem="¿" id="112-1-1"> ¿ </w>
<w cat="P" lem="se" id="112-1-2"> se </w>
<w cat="AUX" lem="estar" num="sing" tie="impf_ind" id="112-1-3"> estaba </w>
<w cat="V" lem="quemar" tie="ger" id="112-1-4"> quemando </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="112-1-5"> la </w>
<w cat="N" lem="colita" gen="fem" num="sing" id="112-1-6"> colita </w>
<w cat="PUNCT" lem="?" id="112-1-7"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="262.895" et="265.249" id="113">
<sf t="enu" id="113-1">
```

```
<w cat="ADV" lem="sí" id="113-1-1"> sí </w>
<ps/>
<w cat="C" lem="y" id="113-1-2"> y </w>
<w cat="AUX" lem="salir" tie="indf_ind" num="sing" per="3" id="113-1-3"> salió </w>
<w cat="V" lem="correr" tie="ger" id="113-1-4"> corriendo </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="265.249" et="269.283" id="114">
<sf t="enu" id="114-1">
<w cat="C" lem="y" id="114-1-1"> y </w>
<w cat="V" lem="meter" tie="indf_ind" num="sing" per="3" id="114-1-2"> metió </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="114-1-3"> la </w>
<w cat="N" lem="cola" gen="fem" num="sing" id="114-1-4"> cola </w>
<w cat="PREP" lem="en" id="114-1-5"> en </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="114-1-6"> el </w>
<w cat="N" lem="charco" gen="masc" num="sing" id="114-1-7"> charco </w>
<ps t="complex"/>
<w cat="C" lem="y" id="114-1-8"> y </w>
<w cat="P" lem="lo" num="sing" per="3" id="114-1-9"> le </w>
<w cat="V" lem="salir" num="sing" tie="impf_ind" id="114-1-10"> salía </w>
<w cat="N" lem="humo" gen="masc" num="sing" id="114-1-11"> humo </w>
<w cat="N" lem="humo" gen="masc" num="sing" id="114-1-12"> humo </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="269.283" et="273.625" id="115">
<sf t="enu" id="115-1">
<w cat="C" lem="y" id="115-1-1"> y </w>
<w cat="N" lem="colorín" gen="masc" num="sing" id="115-1-2"> colorín </w>
<w cat="ADJ" lem="colorado" gen="masc" num="sing" id="115-1-3"> colorado </w>
<ps/>
<w cat="DEM" lem="este" gen="masc" num="sing" id="115-1-4"> este </w>
<w cat="N" lem="cuento" gen="masc" num="sing" id="115-1-5"> cuento </w>
<w cat="P" lem="se" id="115-1-6"> se </w>
<w cat="AUX" lem="haber" tie="pres_ind" num="sing" per="3" id="115-1-7"> ha </w>
<w cat="V" lem="acabar" tie="part" id="115-1-8"> acabado </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="273.625" et="275.92" id="116">
<sf t="int" id="116-1">
<w cat="PUNCT" lem="¿" id="116-1-1"> ¿ </w>
<w cat="C" lem="y" id="116-1-2"> y </w>
<w cat="P" lem="quién" id="116-1-3"> quién </w>
<w cat="V" lem="haber" num="sing" tie="impf_ind" id="116-1-4"> había </w>
<ps t="tr"/>
<w cat="P" lem="por qué" id="116-1-5"> por qué </w>
<w cat="V" lem="tener" num="sing" tie="impf_ind" id="116-1-6"> tenía </w>
<w cat="N" lem="fuego" gen="masc" num="sing" id="116-1-7"> fuego </w>
<w cat="PUNCT" lem="?" id="116-1-8"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="275.92" et="279.565" id="117">
<sf t="enu" id="117-1">
```

<w cat="C" lem="porque" id="117-1-1"> porque </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="117-1-2"> el </w>
<w cat="ADJ" lem="cazador" gen="masc" num="sing" id="117-1-3"> cazador </w>
<w cat="P" lem="lo" num="sing" per="3" id="117-1-4"> le </w>
<w cat="AUX" lem="haber" num="sing" tie="impf_ind" id="117-1-5"> había </w>
<w cat="V" lem="disparar" tie="part" id="117-1-6"> disparado </w>
<w cat="PREP" lem="en" id="117-1-7"> en </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="117-1-8"> el </w>
<w cat="N" lem="culo" gen="masc" num="sing" id="117-1-9"> culo </w>
</sf>
</f>

<f h="TEA" st="279.565" et="279.91" id="118">
<sf t="exc" id="118-1">
<w cat="PUNCT" lem="¡" id="118-1-1"> ¡ </w>
<w cat="INTJ" lem="ah" id="118-1-2"> ah </w>
<w cat="PUNCT" lem="!" id="118-1-3"> ! </w>
</sf>
</f>

<f h="TEA" st="279.91" et="281.286" id="119">
<sf t="enu" id="119-1">
<w cat="C" lem="que" id="119-1-1"> que </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="119-1-2"> lo </w>
<w cat="PREP" lem="de" id="119-1-3"> de </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="119-1-4"> el </w>
<w cat="ADJ" lem="cazador" gen="masc" num="sing" id="119-1-5"> cazador </w>
<w cat="ADV" lem="no" id="119-1-6"> no </w>
<w cat="P" lem="me" num="sing" per="1" id="119-1-7"> me </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="119-1-8"> lo </w>
<w cat="AUX" lem="haber" tie="pres_ind" num="sing" per="2" id="119-1-9"> has </w>
<w cat="V" lem="decir" tie="part" id="119-1-10"> dicho </w>
</sf>
</f>

<f h="INE" st="281.286" et="282.206" id="120">
<sf t="enu" id="120-1">
<nl/>
</sf>
</f>

<f h="TEA" st="282.206" et="283.583" id="121">
<sf t="enu" id="121-1">
<nl/>
</sf>
</f>

<f h="TEA" st="283.583" et="285.221" id="122">
<sf t="int" id="122-1">
<w cat="PUNCT" lem="¿" id="122-1-1"> ¿ </w>
<w cat="C" lem="y" id="122-1-2"> y </w>
<w cat="DEM" lem="ese" gen="masc" num="plu" id="122-1-3"> esos </w>
<w cat="N" lem="cuento" gen="masc" num="plu" id="122-1-4"> cuentos </w>
<w cat="P" lem="quién" id="122-1-5"> quién </w>
<w cat="P" lem="te" num="sing" per="2" id="122-1-6"> te </w>

<w cat="P" lem="lo" gen="masc" num="plu" per="3" id="122-1-7"> los </w>
<w cat="V" lem="contar" tie="pres_ind" num="sing" per="3" id="122-1-8"> cuenta </w>
<w cat="PREP" lem="a" id="122-1-9"> a </w>
<w cat="P" lem="ti" num="sing" per="2" id="122-1-10"> ti </w>
<w cat="PUNCT" lem="?" id="122-1-11"> ? </w>
</sf>
</f>

<f h="INE" st="285.221" et="288.059" id="123">
<sf t="enu" id="123-1">
<nl/>
<w cat="P" lem="yo" num="sing" per="1" id="123-1-1"> yo </w>
<w cat="P" lem="me" num="sing" per="1" id="123-1-2"> me </w>
<w cat="P" lem="lo" gen="masc" num="plu" per="3" id="123-1-3"> los </w>
<w cat="V" lem="inventar" tie="pres_ind" num="sing" per="1" id="123-1-4"> invento </w>
</sf>
</f>

<f h="TEA" st="288.059" et="289.984" id="124">
<sf t="int" id="124-1">
<w cat="PUNCT" lem="¿" id="124-1-1"> ¿ </w>
<w cat="C" lem="pero" id="124-1-2"> pero </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="124-1-3"> el </w>
<w cat="PREP" lem="de" id="124-1-4"> de </w>
<w cat="NPR" lem="Caperucita" id="124-1-5"> Caperucita </w>
<w cat="PUNCT" lem="?" id="124-1-6"> ? </w>
</sf>
</f>

<f h="INE" st="289.984" et="291.357" id="125">
<sf t="enu" id="125-1">
<w cat="Q" lem="nadie" id="125-1-1"> nadie </w>
<w cat="P" lem="me" num="sing" per="1" id="125-1-2"> me </w>
<w cat="P" lem="lo" gen="masc" num="plu" per="3" id="125-1-3"> los </w>
<w cat="V" lem="contar" tie="pres_ind" num="sing" per="3" id="125-1-4"> cuenta </w>
</sf>
</f>

<f h="INE" st="291.357" et="293.607" id="126">
<sf t="enu" id="126-1">
<w cat="AUX" lem="estar" tie="pres_ind" num="plu" per="1" id="126-1-1"> estamos </w>
<w cat="V" lem="hacer" tie="ger" id="126-1-2"> haciendo </w>
<w cat="Q" lem="un" gen="masc" id="126-1-3"> un </w>
<w cat="N" lem="cuento" gen="masc" num="sing" id="126-1-4"> cuento </w>
<w cat="PREP" lem="de" id="126-1-5"> de </w>
<w cat="NPR" lem="Caperucita" id="126-1-6"> Caperucita </w>
</sf>
</f>

<f h="INE" st="293.607" et="295.92" id="127">
<sf t="enu" id="127-1">
<w cat="C" lem="pero" id="127-1-1"> pero </w>
<w cat="ADV" lem="no" id="127-1-2"> no </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="3" id="127-1-3"> tiene </w>
<w cat="Q" lem="tanto" gen="fem" id="127-1-4"> tantas </w>

<w cat="N" lem="hoja" gen="fem" num="plu" id="127-1-5"> hojas </w>
</sf>
</f>

<f h="TEA" st="295.92" et="296.558" id="128">
<sf t="int" id="128-1">
<w cat="PUNCT" lem="¿" id="128-1-1"> ¿ </w>
<w cat="MD" lem="no ?" id="128-1-2"> no ? </w>
</sf>
</f>

<f h="INE" st="296.558" et="300.934" id="129">
<sf t="enu" id="129-1">
<w cat="ADV" lem="no" id="129-1-1"> no </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="3" id="129-1-2"> tiene </w>
<ps t="tr"/>
<w cat="C" lem="cuando" id="129-1-3"> cuando </w>
<w cat="P" lem="se" id="129-1-4"> se </w>
<ps t="tr"/>
<w cat="P" lem="lo" gen="fem" num="sing" per="3" id="129-1-5"> la </w>
<w cat="V" lem="esconder" tie="indf_ind" num="sing" per="3" id="129-1-6"> escondió </w>
<w cat="PREP" lem="en" id="129-1-7"> en </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="129-1-8"> el </w>
<w cat="N" lem="armario" gen="masc" num="sing" id="129-1-9"> armario </w>
<w cat="C" lem="y" id="129-1-10"> y </w>
<w cat="DEM" lem="ese" gen="fem" num="plu" id="129-1-11"> esas </w>
<w cat="N" lem="cosa" gen="fem" num="plu" id="129-1-12"> cosas </w>
</sf>
</f>

<f h="TEA" st="300.934" et="302.875" id="130">
<sf t="enu" id="130-1">
<nl/>
</sf>
</f>

<f h="INE" st="302.875" et="306.626" id="131">
<sf t="enu" id="131-1">
<w cat="P" lem="nos" num="plu" per="1" id="131-1-1"> nos </w>
<w cat="V" lem="faltar" tie="pres_ind" num="sing" per="3" id="131-1-2"> falta </w>
<w cat="V" lem="colorear" tie="inf" id="131-1-3"> colorear </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="131-1-4"> el </w>
<w cat="N" lem="dibujo" gen="masc" num="sing" id="131-1-5"> dibujo </w>
</sf>
</f>

<f h="TEA" st="306.626" et="308.213" id="132">
<sf t="enu" id="132-1">
<w cat="C" lem="pero" id="132-1-1"> pero </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="132-1-2"> lo </w>
<w cat="AUX" lem="estar" tie="pres_ind" num="plu" per="2" id="132-1-3"> estáis </w>
<w cat="V" lem="hacer" tie="ger" id="132-1-4"> haciendo </w>
<w cat="PREP" lem="en" id="132-1-5"> en </w>
<w cat="N" lem="clase" gen="fem" num="sing" id="132-1-6"> clase </w>
</sf>

```
</f>
</p>
<p>
<f h="INE" st="308.213" et="309.164" id="133">
<sf t="enu" id="133-1">
<w cat="ADV" lem="sí" id="133-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="309.164" et="311.168" id="134">
<sf t="int" id="134-1">
<w cat="PUNCT" lem="¿" id="134-1-1"> ¿ </w>
<w cat="C" lem="y" id="134-1-2"> y </w>
<w cat="P" lem="qué" id="134-1-3"> qué </w>
<w cat="V" lem="ser" tie="pres_ind" num="sing" per="3" id="134-1-4"> es </w>
<w cat="Q" lem="otro" gen="fem" id="134-1-5"> otra </w>
<w cat="N" lem="cosa" gen="fem" num="sing" id="134-1-6"> cosa </w>
<w cat="REL" lem="que" id="134-1-7"> que </w>
<w cat="AUX" lem="ir" tie="pres_ind" num="plu" per="2" id="134-1-8"> vais </w>
<w cat="PREP" lem="a" id="134-1-9"> a </w>
<w cat="V" lem="hacer" tie="inf" id="134-1-10"> hacer </w>
<ps/>
<w cat="REL" lem="que" id="134-1-11"> que </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="1"id="134-1-12"> sé </w>
<w cat="P" lem="yo" num="sing" per="1" id="134-1-13"> yo </w>
<w cat="PUNCT" lem="?" id="134-1-14"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="311.168" et="313.823" id="135">
<sf t="enu" id="135-1">
<nl/>
<ps/>
<w cat="V" lem="sumar" tie="inf" id="135-1-1"> sumar </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="313.823" et="314.352" id="136">
<sf t="enu" id="136-1">
<w cat="ADV" lem="no" id="136-1-1"> no </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="314.352" et="316.752" id="137">
<sf t="enu" id="137-1">
<w cat="Q" lem="un" gen="fem" id="137-1-1"> una </w>
<w cat="N" lem="cosa" gen="fem" num="sing"id="137-1-2"> cosita </w>
<w cat="REL" lem="que" id="137-1-3"> que </w>
<w cat="AUX" lem="ir" tie="pres_ind" num="plu" per="2" id="137-1-4"> vais </w>
<w cat="PREP" lem="a" id="137-1-5"> a </w>
<w cat="V" lem="escribir" tie="inf" id="137-1-6"> escribir </w>
<w cat="ADV" lem="también" id="137-1-7"> también </w>
<w cat="PREP" lem="entre" id="137-1-8"> entre </w>
<w cat="Q" lem="todo" gen="masc" id="137-1-9"> todos </w>
</sf>
</f>
```

```
</p>
<p>
<f h="INE" st="316.752" et="319.426" id="138">
<sf t="int" id="138-1">
<w cat="PUNCT" lem="¿" id="138-1-1"> ¿ </w>
<w cat="PREP" lem="entre" id="138-1-2"> entre </w>
<w cat="Q" lem="todo" gen="masc" id="138-1-3"> todos </w>
<w cat="ADJ" lem="junto" gen="masc" num="plu" id="138-1-4"> juntos </w>
<w cat="PUNCT" lem="?" id="138-1-5"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="319.426" et="320.94" id="139">
<sf t="enu" id="139-1">
<w cat="C" lem="que" id="139-1-1"> que </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="139-1-2"> lo </w>
<w cat="V" lem="decir" tie="indf_ind" num="sing" per="3" id="139-1-3"> dijo </w>
<w cat="NPR" lem="Candi" id="139-1-4"> Candi </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="139-1-5"> el </w>
<w cat="Q" lem="otro" gen="masc" id="139-1-6"> otro </w>
<w cat="N" lem="día" gen="masc" num="sing" id="139-1-7"> día </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="320.94" et="323.123" id="140">
<sf t="sup" id="140-1">
<w cat="V" lem="decir" tie="indf_ind" num="sing" per="3" id="140-1-1"> dijo </w>
<ps/>
<w cat="AUX" lem="tener" tie="pres_ind" num="plu" per="2" id="140-1-2"> tenéis </w>
<w cat="C" lem="que" id="140-1-3"> que </w>
<w cat="V" lem="traer" tie="inf" id="140-1-4"> traer </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="323.123" et="324.913" id="141">
<sf t="enu" id="141-1">
<w cat="Q" lem="un" gen="fem" id="141-1-1"> una </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="323.123" et="324.913" id="142">
<sf t="enu" id="142-1">
<bw/>
<w cat="Q" lem="un" gen="fem" id="142-1-1"> una </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="324.913" et="326.138" id="143">
<sf t="enu" id="143-1">
<w cat="N" lem="noticia" gen="fem" num="sing" id="143-1-1"> noticia </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="326.138" et="326.659" id="144">
```

```
<sf t="enu" id="144-1">
<w cat="P" lem="ése" num="sing" per="3" id="144-1-1"> eso </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="326.659" et="327.354" id="145">
<sf t="int" id="145-1">
<w cat="PUNCT" lem="¿" id="145-1-1"> ¿ </w>
<w cat="PREP" lem="para" id="145-1-2"> para </w>
<w cat="P" lem="qué" id="145-1-3"> qué </w>
<w cat="PUNCT" lem="?" id="145-1-4"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="327.354" et="329.565" id="146">
<sf t="enu" id="146-1">
<w cat="PREP" lem="para" id="146-1-1"> para </w>
<w cat="V" lem="hacer" tie="inf" id="146-1-2"> hacer </w>
<w cat="Q" lem="un" gen="masc" id="146-1-3"> un </w>
<w cat="N" lem="periódico" gen="masc" num="sing" id="146-1-4"> periódico </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="329.565" et="330.464" id="147">
<sf t="enu" id="147-1">
<w cat="Q" lem="un" gen="masc" id="147-1-1"> un </w>
<w cat="N" lem="periódico" gen="masc" num="sing" id="147-1-2"> periódico </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="330.464" et="332.938" id="148">
<sf t="enu" id="148-1">
<w cat="C" lem="y" id="148-1-1"> y </w>
<w cat="POSS" lem="mi" num="sing" id="148-1-2"> mi </w>
<w cat="N" lem="perra" gen="fem" num="sing" id="148-1-3"> perra </w>
<ps t="tr"/>
<w cat="MD" lem="mira" id="148-1-4"> mira </w>
<ps/>
<w cat="V" lem="fijar" tie="imper" num="sing" per="2" clitic="yes" id="148-1-5"> fíja </w>
<w cat="P" lem="te" num="sing" per="2" id="148-1-6"> te </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="332.938" et="335.266" id="149">
<sf t="enu" id="149-1">
<w cat="P" lem="te" num="sing" per="2" id="149-1-1"> te </w>
<w cat="AUX" lem="ir" tie="pres_ind" num="sing" per="1" id="149-1-2"> voy </w>
<w cat="PREP" lem="a" id="149-1-3"> a </w>
<w cat="V" lem="contar" tie="inf" id="149-1-4"> contar </w>
<w cat="P" lem="qué" id="149-1-5"> qué </w>
<w cat="P" lem="lo" num="sing" per="3" id="149-1-6"> le </w>
<w cat="V" lem="pasar" tie="indf_ind" num="sing" per="3" id="149-1-7"> pasó </w>
<w cat="PREP" lem="a" id="149-1-8"> a </w>
<w cat="POSS" lem="mi" num="sing" id="149-1-9"> mi </w>
<w cat="N" lem="perra" gen="fem" num="sing" id="149-1-10"> perra </w>
```

```
    </sf>
    </f>
    </p>
    <p>
    <f h="INE" st="335.266" et="338.574" id="150">
    <sf t="enu" id="150-1">
    <w cat="V" lem="ir" tie="impf_ind" num="plu" per="1" id="150-1-1"> íbamos </w>
    <w cat="POSS" lem="mi" num="sing" id="150-1-2"> mi </w>
    <w cat="N" lem="abuelo" gen="masc" num="sing" id="150-1-3"> abuelo </w>
    <w cat="C" lem="y" id="150-1-4"> y </w>
    <w cat="P" lem="yo" num="sing" per="1" id="150-1-5"> yo </w>
    <w cat="V" lem="pasear" tie="ger" id="150-1-6"> paseando </w>
    <w cat="PREP" lem="por" id="150-1-7"> por </w>
    <w cat="ART" lem="el" gen="masc" num="sing" id="150-1-8"> el </w>
    <w cat="N" lem="campo" gen="masc" num="sing" id="150-1-9"> campo </w>
    </sf>
    </f>
    </p>
    <p>
    <f h="INE" st="338.574" et="344.586" id="151">
    <sf t="enu" id="151-1">
    <w cat="C" lem="y" id="151-1-1"> y </w>
    <w cat="C" lem="cuando" id="151-1-2"> cuando </w>
    <w cat="NPR" lem="Aila" id="151-1-3"> Aila </w>
    <w cat="ADV" lem="ya" id="151-1-4"> ya </w>
    <w cat="P" lem="se" id="151-1-5"> se </w>
    <w cat="V" lem="ir" tie="indf_ind" num="sing" per="3"id="151-1-6"> fue </w>
    <w cat="Q" lem="más" id="151-1-7"> más </w>
    <w cat="ADV" lem="lejos" id="151-1-8"> lejos </w>
    <w cat="C" lem="que" id="151-1-9"> que </w>
    <w cat="Q" lem="nada" num="sing" id="151-1-10"> nada </w>
    <ps t="complex"/>
    <w cat="P" lem="nos" num="plu" per="1" id="151-1-11"> nos </w>
    <ps t="tr"/>
    <w cat="P" lem="se" id="151-1-12"> se </w>
    <w cat="V" lem="desaparecer" tie="indf_ind" num="sing" per="3" id="151-1-13"> desapareció
</w>
    <ps/>
    <w cat="V" lem="fijar" tie="imper" num="sing" per="2" clitic="yes" id="151-1-14"> fíja </w>
    <w cat="P" lem="te" num="sing" per="2" id="151-1-15"> te </w>
    </sf>
    </f>
    </p>
    <p>
    <f h="INE" st="344.586" et="353.529" id="152">
    <sf t="enu" id="152-1">
    <w cat="C" lem="y" id="152-1-1"> y </w>
    <w cat="NPR" lem="Furi" id="152-1-2"> Furi </w>
    <w cat="POSS" lem="nuestro" gen="fem" num="sing" id="152-1-3"> nuestra </w>
    <w cat="N" lem="perra" gen="fem" num="sing" id="152-1-4"> perra </w>
    <ps t="complex"/>
    <w cat="REL" lem="que" id="152-1-5"> que </w>
    <w cat="V" num="sing" lem="ser" tie="impf_ind" id="152-1-6"> era </w>
    <w cat="Q" lem="un poco" id="152-1-7"> un poco </w>
    <w cat="N" lem="chico" gen="fem" num="sing" id="152-1-8"> chica </w>
    <ps t="complex"/>
    <w cat="P" lem="lo" num="sing" per="3" id="152-1-9"> le </w>
    <ps t="pr"/>
    <w cat="P" lem="lo" gen="fem" num="sing" per="3" id="152-1-10"> la </w>
    <w cat="V" lem="llamar" tie="indf_ind" num="sing" per="3" id="152-1-11"> llamó </w>
```

```
<unint/>
<ps/>
<w cat="C" lem="porque" id="152-1-12"> porque </w>
<ps t="pr"/>
<w cat="C" lem="porque" id="152-1-13"> porque </w>
<w cat="NPR" lem="Furi" id="152-1-14"> Furi </w>
<w cat="ADV" lem="no" id="152-1-15"> no </w>
<frag>
</frag>
<ps t="pr"/>
<w cat="ADV" lem="sí" id="152-1-16"> sí </w>
<w cat="V" lem="hacer" tie="pres_ind" num="sing" per="3" id="152-1-17"> hace </w>
<w cat="N" lem="caso" gen="masc" num="sing" id="152-1-18"> caso </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="353.529" et="354.565" id="153">
<sf t="enu" id="153-1">
<w cat="NPR" lem="Furi" id="153-1-1"> Furi </w>
<w cat="V" lem="ser" tie="pres_ind" num="sing" per="3" id="153-1-2"> es </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="153-1-3"> la </w>
<w cat="ADJ" lem="pequeño" gen="fem" num="sing" id="153-1-4"> pequeña </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="354.565" et="354.919" id="154">
<sf t="enu" id="154-1">
<w cat="ADV" lem="sí" id="154-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="354.919" et="355.658" id="155">
<sf t="int" id="155-1">
<w cat="PUNCT" lem="¿" id="155-1-1"> ¿ </w>
<w cat="C" lem="y" id="155-1-2"> y </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="155-1-3"> la </w>
<w cat="ADJ" lem="grande" num="sing" id="155-1-4"> grande </w>
<w cat="P" lem="cómo" id="155-1-5"> cómo </w>
<w cat="P" lem="se" id="155-1-6"> se </w>
<w cat="V" num="sing" lem="llamar" id="155-1-7"> llama </w>
<w cat="PUNCT" lem="?" id="155-1-8"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="355.658" et="357.783" id="156">
<sf t="enu" id="156-1">
<w cat="C" lem="pero" id="156-1-1"> pero </w>
<ps t="tr"/>
<w cat="ART" lem="el" gen="fem" num="sing" id="156-1-2"> la </w>
<w cat="ADJ" lem="grande" num="sing" id="156-1-3"> grande </w>
<w cat="P" lem="se" id="156-1-4"> se </w>
<w cat="V" num="sing" lem="llamar" id="156-1-5"> llama </w>
<w cat="NPR" lem="Aila" id="156-1-6"> Aila </w>
</sf>
</f>
</p>
```

```
<p>
<f h="TEA" st="357.783" et="358.677" id="157">
<sf t="int" id="157-1">
<w cat="PUNCT" lem="¿" id="157-1-1"> ¿ </w>
<w cat="NPR" lem="Aila" id="157-1-2"> Aila </w>
<w cat="PUNCT" lem="?" id="157-1-3"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="358.677" et="362.08" id="158">
<sf t="enu" id="158-1">
<w cat="C" lem="pero" id="158-1-1"> pero </w>
<ps/>
<w cat="V" lem="ser" tie="pres_ind" num="sing" per="3" id="158-1-2"> es </w>
<w cat="Q" lem="un poco" id="158-1-3"> un poco </w>
<w cat="Q" lem="más" id="158-1-4"> más </w>
<w cat="ADJ" lem="alto" gen="fem" num="sing" id="158-1-5"> alta </w>
<ps/>
<w cat="NPR" lem="Furi" id="158-1-6"> Furi </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="362.08" et="363.934" id="159">
<sf t="enu" id="159-1">
<w cat="NPR" lem="Aila" id="159-1-1"> Aila </w>
<w cat="V" lem="ser" tie="pres_ind" num="sing" per="3" id="159-1-2"> es </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="159-1-3"> la </w>
<w cat="Q" lem="más" id="159-1-4"> más </w>
<w cat="ADJ" lem="alto" gen="fem" num="sing" id="159-1-5"> alta </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="363.934" et="365.175" id="160">
<sf t="enu" id="160-1">
<w cat="C" lem="y" id="160-1-1"> y </w>
<w cat="NPR" lem="Furi" id="160-1-2"> Furi </w>
<w cat="V" lem="ser" tie="pres_ind" num="sing" per="3" id="160-1-3"> es </w>
<w cat="Q" lem="un" gen="masc" id="160-1-4"> un </w>
<w cat="Q" lem="poquito" gen="masc" id="160-1-5"> poquito </w>
<w cat="Q" lem="más" id="160-1-6"> más </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="365.175" et="365.829" id="161">
<sf t="enu" id="161-1">
<w cat="Q" lem="más" id="161-1-1"> más </w>
<w cat="ADJ" lem="pequeño" gen="fem" num="sing" id="161-1-2"> pequeña </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="365.829" et="369.413" id="162">
<sf t="int" id="162-1">
<w cat="PUNCT" lem="¿" id="162-1-1"> ¿ </w>
<w cat="C" lem="y" id="162-1-2"> y </w>
<w cat="P" lem="qué" id="162-1-3"> qué </w>
```

```
<w cat="V" lem="pasar" tie="indf_ind" num="sing" per="3" id="162-1-4"> pasó </w>
<w cat="PUNCT" lem="?" id="162-1-5"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="369.413" et="370.42" id="163">
<sf t="int" id="163-1">
<w cat="PUNCT" lem="¿" id="163-1-1"> ¿ </w>
<w cat="P" lem="se" id="163-1-2"> se </w>
<w cat="V" lem="ir" tie="indf_ind" num="sing" per="3"id="163-1-3"> fue </w>
<w cat="PUNCT" lem="?" id="163-1-4"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="370.42" et="372.13" id="164">
<sf t="enu" id="164-1">
<w cat="MD" lem="ya" id="164-1-1"> ya </w>
<w cat="V" lem="venir" tie="indf_ind" num="sing" per="3" id="164-1-2"> vino </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="372.13" et="376.928" id="165">
<sf t="enu" id="165-1">
<frag>
</frag>
<ps t="pr"/>
<w cat="MD" lem="mira" id="165-1-1"> mira </w>
<w cat="C" lem="cuando" id="165-1-2"> cuando </w>
<ps/>
<w cat="NPR" lem="Furi" id="165-1-3"> Furi </w>
<w cat="V" num="sing" lem="ser" tie="impf_ind" id="165-1-4"> era </w>
<w cat="ADV" lem="así" id="165-1-5"> así </w>
<w cat="PREP" lem="de" id="165-1-6"> de </w>
<w cat="ADJ" lem="chiquitín" gen="fem" num="sing" id="165-1-7"> chiquitina </w>
<ps/>
<w cat="P" lem="se" id="165-1-8"> se </w>
<w cat="V" lem="salir" tie="indf_ind" num="sing" per="3" id="165-1-9"> salió </w>
<w cat="PREP" lem="a" id="165-1-10"> a </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="165-1-11"> la </w>
<w cat="N" lem="calle" gen="fem" num="sing" id="165-1-12"> calle </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="376.928" et="378.263" id="166">
<sf t="enu" id="166-1">
<w cat="PREP" lem="a" id="166-1-1"> a </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="166-1-2"> el </w>
<w cat="N" lem="campo" gen="masc" num="sing" id="166-1-3"> campo </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="378.263" et="378.626" id="167">
<sf t="enu" id="167-1">
<nl/>
</sf>
```

```
</f>
</p>
<p>
<f h="INE" st="378.626" et="383.539" id="168">
<sf t="enu" id="168-1">
<w cat="C" lem="y" id="168-1-1"> y </w>
<ps t="pr"/>
<w cat="C" lem="y" id="168-1-2"> y </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="168-1-3"> la </w>
<w cat="N" lem="madre" gen="fem" num="sing" id="168-1-4"> madre </w>
<w cat="PREP" lem="de" id="168-1-5"> de </w>
<w cat="NPR" lem="Furi" id="168-1-6"> Furi </w>
<w cat="V" lem="estar" num="sing" tie="impf_ind" id="168-1-7"> estaba </w>
<ps t="tr"/>
<w cat="P" lem="se" id="168-1-8"> se </w>
<w cat="V" lem="llamar" num="sing" tie="impf_ind" id="168-1-9"> llamaba </w>
<w cat="NPR" lem="Aila" id="168-1-10"> Aila </w>
<ps/>
<w cat="C" lem="que" id="168-1-11"> que </w>
<w cat="ADV" lem="ya" id="168-1-12"> ya </w>
<w cat="P" lem="te" num="sing" per="2" id="168-1-13"> te </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="168-1-14"> lo </w>
<w cat="AUX" lem="haber" num="sing" id="168-1-15"> he </w>
<w cat="V" lem="decir" tie="part" id="168-1-16"> dicho </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="383.539" et="383.874" id="169">
<sf t="enu" id="169-1">
<nl/>
</sf>
</f>
</p>
<p>
<f h="INE" st="383.874" et="385.131" id="170">
<sf t="enu" id="170-1">
<w cat="C" lem="y" id="170-1-1"> y </w>
<w cat="V" lem="venir" tie="indf_ind" num="sing" per="3" id="170-1-2"> vino </w>
<w cat="Q" lem="un" gen="fem" id="170-1-3"> una </w>
<w cat="N" lem="perra" gen="fem" num="sing" id="170-1-4"> perra </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="385.131" et="388.188" id="171">
<sf t="enu" id="171-1">
<w cat="C" lem="y" id="171-1-1"> y </w>
<w cat="AUX" lem="querer" num="sing" tie="impf_ind" id="171-1-2"> quería </w>
<w cat="V" lem="comer" tie="inf" clitic="yes" id="171-1-3"> comer </w>
<w cat="P" lem="se" id="171-1-4"> se </w>
<w cat="PREP" lem="a" id="171-1-5"> a </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="171-1-6"> la </w>
<w cat="ADJ" lem="chico" gen="fem" num="sing"id="171-1-7"> chiquituja </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="388.188" et="388.827" id="172">
<sf t="int" id="172-1">
```

```
<w cat="PUNCT" lem="¿" id="172-1-1"> ¿ </w>
<w cat="ADV" lem="sí" id="172-1-2"> sí </w>
<w cat="PUNCT" lem="?" id="172-1-3"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="388.827" et="393.428">
<sf t="int">
<w cat="C" lem="pero"> pero </w>
<w cat="V" lem="salir" tie="indf_ind" num="sing" per="3"> salió </w>
<w cat="ART" lem="el" gen="fem" num="sing"> la </w>
<w cat="N" lem="perra" gen="fem" num="sing"> perra </w>
<ps/>
<w cat="NPR" lem="Aila"> Aila </w>
<ps/>
<w cat="C" lem="y"> y </w>
<w cat="P" lem="lo" num="sing" per="3"> le </w>
<w cat="V" lem="pegar" tie="indf_ind" num="sing" per="3"> pegó </w>
<w cat="Q" lem="un" gen="masc"> unos </w>
<w cat="N" lem="mordisco" gen="masc" num="plu"> mordiscos </w>
<w cat="REL" lem="que"> que </w>
<ps t="interrup"/>
</sf>
</f>
</p>
<p>
<f h="TEA" st="393.428" et="394.548" id="173">
<sf t="enu" id="173-1">
<w cat="V" lem="decir" tie="indf_ind" num="sing" per="3" id="173-1-1"> dijo </w>
<ps/>
<nl/>
<ps/>
<w cat="N" lem="cuidado" gen="masc" num="sing" id="173-1-2"> cuidado </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="394.548" et="395.184" id="174">
<sf t="int" id="174-1">
<w cat="PUNCT" lem="¿" id="174-1-1"> ¿ </w>
<w cat="MD" lem="no ?" id="174-1-2"> no ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="395.184" et="396.227" id="175">
<sf t="enu" id="175-1">
<w cat="ADV" lem="sí" id="175-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="396.227" et="401.006" id="176">
<sf t="int" id="176-1">
<w cat="PUNCT" lem="¿" id="176-1-1"> ¿ </w>
<w cat="C" lem="y" id="176-1-2"> y </w>
<w cat="P" lem="lo" gen="fem" num="plu" per="3" id="176-1-3"> las </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="176-1-4"> tienes </w>
<w cat="PREP" lem="en" id="176-1-5"> en </w>
```

```
<w cat="N" lem="casa" gen="fem" num="sing" id="176-1-6"> casa </w>
<w cat="PREP" lem="a" id="176-1-7"> a </w>
<w cat="ART" lem="el" gen="fem" num="plu" id="176-1-8"> las </w>
<w cat="N" lem="perra" gen="fem" num="plu" id="176-1-9"> perras </w>
<w cat="PUNCT" lem="?" id="176-1-10"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="401.006" et="401.406" id="177">
<sf t="enu" id="177-1">
<w cat="ADV" lem="sí" id="177-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="401.406" et="402.691" id="178">
<sf t="enu" id="178-1">
<w cat="PREP" lem="en" id="178-1-1"> en </w>
<w cat="POSS" lem="mi" num="sing" id="178-1-2"> mi </w>
<w cat="N" lem="casa" gen="fem" num="sing" id="178-1-3"> casa </w>
<w cat="ADV" lem="no" id="178-1-4"> no </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="402.691" et="404.061" id="179">
<sf t="enu" id="179-1">
<w cat="PREP" lem="en" id="179-1-1"> en </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="179-1-2"> la </w>
<w cat="N" lem="casa" gen="fem" num="sing" id="179-1-3"> casa </w>
<w cat="PREP" lem="de" id="179-1-4"> de </w>
<w cat="ART" lem="el" gen="masc" num="plu" id="179-1-5"> los </w>
<w cat="N" lem="abuelo" gen="masc" num="plu" id="179-1-6"> abuelos </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="404.061" et="404.969" id="180">
<sf t="exc" id="180-1">
<w cat="PUNCT" lem="¡" id="180-1-1"> ¡ </w>
<w cat="INTJ" lem="ah" id="180-1-2"> ah </w>
<w cat="PUNCT" lem="!" id="180-1-3"> ! </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="404.969" et="407.028" id="181">
<sf t="enu" id="181-1">
<w cat="C" lem="porque" id="181-1-1"> porque </w>
<w cat="P" lem="yo" num="sing" per="1" id="181-1-2"> yo </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="1" id="181-1-3"> tengo </w>
<w cat="ADV" lem="sólo" id="181-1-4"> sólo </w>
<w cat="N" lem="nave" gen="fem" num="sing" id="181-1-5"> nave </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="407.028" et="408.209" id="182">
<sf t="enu" id="182-1">
```

<w cat="ADV" lem="no" id="182-1-1"> no </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="1" id="182-1-2"> tengo </w>
<w cat="N" lem="patio" gen="masc" num="sing" id="182-1-3"> patio </w>
</sf>
</f>

<f h="TEA" st="408.209" et="409.366" id="183">
<sf t="exc" id="183-1">
<w cat="PUNCT" lem="¡" id="183-1-1"> ¡ </w>
<w cat="INTJ" lem="ah" id="183-1-2"> ah </w>
<w cat="PUNCT" lem="!" id="183-1-3"> ! </w>
</sf>
</f>

<f h="INE" st="409.366" et="413.182" id="184">
<sf t="sup" id="184-1">
<w cat="POSS" lem="mi" num="sing" id="184-1-1"> mi </w>
<w cat="N" lem="abuelo" gen="masc" num="sing" id="184-1-2"> abuelo </w>
<w cat="V" lem="ser" tie="pres_ind" num="sing" per="3" id="184-1-3"> es </w>
<w cat="C" lem="que" id="184-1-4"> que </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="3" id="184-1-5"> tiene </w>
<w cat="N" lem="nave" gen="fem" num="sing" id="184-1-6"> nave </w>
<ps/>
<w cat="N" lem="patio" gen="masc" num="sing" id="184-1-7"> patio </w>
<ps/>
<w cat="N" lem="casa" gen="fem" num="sing" id="184-1-8"> casa </w>
</sf>
</f>

<f h="TEA" st="413.182" et="413.415" id="185">
<sf t="exc" id="185-1">
<w cat="PUNCT" lem="¡" id="185-1-1"> ¡ </w>
<w cat="INTJ" lem="jolín" id="185-1-2"> jolín </w>
<w cat="PUNCT" lem="!" id="185-1-3"> ! </w>
</sf>
</f>

<f h="TEA" st="413.415" et="415.008" id="186">
<sf t="enu" id="186-1">
<w cat="P" lem="ahí" id="186-1-1"> ahí </w>
<w cat="AUX" lem="poder" num="sing" id="186-1-2"> puede </w>
<w cat="V" lem="tener" tie="inf" id="186-1-3"> tener </w>
<w cat="Q" lem="un" gen="masc" id="186-1-4"> un </w>
<w cat="N" lem="montón" gen="masc" num="sing" id="186-1-5"> montón </w>
<w cat="PREP" lem="de"id="186-1-6"> de </w>
<w cat="N" lem="perro" gen="masc" num="plu"id="186-1-7"> perros </w>
</sf>
</f>

<f h="INE" st="415.008" et="416.363" id="187">
<sf t="enu" id="187-1">
<w cat="ADV" lem="sólo" id="187-1-1"> sólo </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="3" id="187-1-2"> tiene </w>
<w cat="Q" lem="dos" id="187-1-3"> dos </w>
</sf>

```
</f>
</p>
<p>
<f h="TEA" st="416.363" et="417.027" id="188">
<sf t="enu" id="188-1">
<w cat="Q" lem="dos" id="188-1-1"> dos </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="417.027" et="419.595" id="189">
<sf t="enu" id="189-1">
<w cat="C" lem="y" id="189-1-1"> y </w>
<w cat="V" lem="ser" tie="pres_ind" num="plu" per="3" id="189-1-2"> son </w>
<w cat="N" lem="mamá" gen="fem" num="sing" id="189-1-3"> mamá </w>
<w cat="C" lem="e" id="189-1-4"> e </w>
<w cat="N" lem="hijo" gen="fem" num="sing" id="189-1-5"> hija </w>
<ps/>
<w cat="MD" lem="entonces" id="189-1-6"> entonces </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="419.595" et="420.113" id="190">
<sf t="enu" id="190-1">
<w cat="ADV" lem="sí" id="190-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="420.113" et="424.902" id="191">
<sf t="enu" id="191-1">
<w cat="ART" lem="el" gen="masc" num="sing" id="191-1-1"> el </w>
<w cat="N" lem="padre" gen="masc" num="sing" id="191-1-2"> padre </w>
<w cat="PREP" lem="de" id="191-1-3"> de </w>
<w cat="NPR" lem="Furi" id="191-1-4"> Furi </w>
<w cat="ADV" lem="no" id="191-1-5"> no </w>
<w cat="V" lem="estar" num="sing" id="191-1-6"> está </w>
<ps t="complex"/>
<w cat="C" lem="pero" id="191-1-7"> pero </w>
<w cat="ADV" lem="sí" id="191-1-8"> sí </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="191-1-9"> lo </w>
<w cat="V" lem="conocer" tie="pres_ind" num="sing" per="1" id="191-1-10"> conozco </w>
<ps/>
<w cat="ADV" lem="no" id="191-1-11"> no </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="1" id="191-1-12"> sé </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="191-1-13"> el </w>
<w cat="N" lem="nombre" gen="masc" num="sing" id="191-1-14"> nombre </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="424.902" et="427.639" id="192">
<sf t="enu" id="192-1">
<w cat="C" lem="aunque" id="192-1-1"> aunque </w>
<w cat="P" lem="se" id="192-1-2"> se </w>
<w cat="V" lem="ir" tie="indf_ind" num="sing" per="3"id="192-1-3"> fue </w>
<w cat="PREP" lem="con" id="192-1-4"> con </w>
<w cat="Q" lem="otro" gen="masc" id="192-1-5"> otro </w>
<w cat="N" lem="amo" gen="masc" num="sing" id="192-1-6"> amo </w>
```

```
</sf>
</f>
</p>
<p>
<f h="TEA" st="427.639" et="428.915" id="193">
<sf t="int" id="193-1">
<w cat="PUNCT" lem="¿" id="193-1-1"> ¿ </w>
<w cat="P" lem="se" id="193-1-2"> se </w>
<w cat="V" lem="ir" tie="indf_ind" num="sing" per="3"id="193-1-3"> fue </w>
<w cat="PUNCT" lem="?" id="193-1-4"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="428.915" et="429.483" id="194">
<sf t="enu" id="194-1">
<w cat="ADV" lem="sí" id="194-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="429.483" et="430.625" id="195">
<sf t="int" id="195-1">
<w cat="PUNCT" lem="¿" id="195-1-1"> ¿ </w>
<w cat="C" lem="pero" id="195-1-2"> pero </w>
<w cat="P" lem="él" gen="masc" num="sing" per="3" id="195-1-3"> él </w>
<w cat="P" lem="se" id="195-1-4"> se </w>
<w cat="V" lem="escapar" tie="indf_ind" num="sing" per="3" id="195-1-5"> escapó </w>
<w cat="PUNCT" lem="?" id="195-1-6"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="430.625" et="431.882" id="196">
<sf t="enu" id="196-1">
<w cat="ADV" lem="no" id="196-1-1"> no </w>
<w cat="P" lem="se" id="196-1-2"> se </w>
<w cat="V" lem="escapar" tie="indf_ind" num="sing" per="3" id="196-1-3"> escapó </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="431.882" et="433.175" id="197">
<sf t="enu" id="197-1">
<w cat="P" lem="se" id="197-1-1"> se </w>
<w cat="V" lem="ir" tie="indf_ind" num="sing" per="3"id="197-1-2"> fue </w>
<w cat="PREP" lem="con" id="197-1-3"> con </w>
<w cat="Q" lem="otro" gen="masc" id="197-1-4"> otro </w>
<w cat="N" lem="amo" gen="masc" num="sing" id="197-1-5"> amo </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="431.882" et="433.175" id="198">
<sf t="enu" id="198-1">
<bw/>
</sf>
<sf t="exc" id="198-2">
<w cat="PUNCT" lem="¡" id="198-2-1"> ¡ </w>
<w cat="INTJ" lem="ah" id="198-2-2"> ah </w>
```

```
<w cat="PUNCT" lem="!" id="198-2-3"> ! </w>
<ps/>
<w cat="P" lem="se"> se </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3"> lo </w>
<w cat="V" lem="dar" tie="indf_ind" num="sing" per="2"> diste </w>
</sf>
</p>
<p>
<f h="TEA" st="433.175" et="434.706" id="199">
<sf t="exc" id="199-1">
<w cat="PUNCT" lem="¡" id="199-1-1"> ¡ </w>
<w cat="INTJ" lem="ah" id="199-1-2"> ah </w>
<w cat="PUNCT" lem="!" id="199-1-3"> ! </w>
<ps/>
<w cat="MD" lem="claro"> claro </w>
</sf>
</p>
<p>
<f h="TEA" st="434.706" et="435.236" id="200">
<sf t="enu" id="200-1">
<nl/>
</sf>
</f>
</p>
<p>
<f h="TEA" st="435.236" et="436.084" id="201">
<sf t="enu" id="201-1">
<w cat="MD" lem="vale" id="201-1-1"> vale </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="436.084" et="437.203" id="202">
<sf t="int" id="202-1">
<w cat="PUNCT" lem="¿" id="202-1-1"> ¿ </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="202-1-2"> tienes </w>
<w cat="Q" lem="más" id="202-1-3"> más </w>
<w cat="N" lem="animal" gen="masc" num="plu" id="202-1-4"> animales </w>
<w cat="PUNCT" lem="?" id="202-1-5"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="437.203" et="437.883" id="203">
<sf t="enu" id="203-1">
<w cat="ADV" lem="sí" id="203-1-1"> sí </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="437.883" et="438.45" id="204">
<sf t="enu" id="204-1">
<w cat="Q" lem="mucho" gen="masc" id="204-1-1"> muchos </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="438.45" et="439.531" id="205">
<sf t="enu" id="205-1">
<w cat="PREP" lem="en" id="205-1-1"> en </w>
```

<w cat="POSS" lem="mi" num="sing" id="205-1-2"> mi </w>
<w cat="N" lem="casa" gen="fem" num="sing" id="205-1-3"> casa </w>
</sf>
</f>

<f h="TEA" st="438.45" et="439.531" id="206">
<sf t="enu" id="206-1">
<bw/>
</sf>
<sf t="int" id="206-2">
<w cat="PUNCT" lem="¿" id="206-2-1"> ¿ </w>
<w cat="Q" lem="más" id="206-2-2"> más </w>
<w cat="PUNCT" lem="?" id="206-2-3"> ? </w>
</sf>
</f>

<f h="TEA" st="439.531" et="439.974" id="207">
<sf t="int" id="207-1">
<w cat="PUNCT" lem="¿" id="207-1-1"> ¿ </w>
<w cat="ADV" lem="sí" id="207-1-2"> sí </w>
<w cat="PUNCT" lem="?" id="207-1-3"> ? </w>
</sf>
</f>

<f h="INE" st="439.974" et="444.469" id="208">
<sf t="enu">
<w cat="ADV" lem="no"> no </w>
<ps t="pr"/>
<w cat="ADV" lem="no"> no </w>
<w cat="P" lem="me" num="sing" per="1"> me </w>
<w cat="V" lem="gustar" num="sing" id="61-1-3"> gusta </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3"> lo </w>
<w cat="PREP" lem="de"> de </w>
<w cat="N" lem="princesa" gen="fem" num="plu"> princesas </w>
<ps t="complex"/>
<w cat="ADV" lem="sólo"> sólo </w>
<frag>
</frag>
<ps t="pr"/>
<w cat="P" lem="lo" gen="masc" num="sing" per="3"> lo </w>
<w cat="PREP" lem="de"> de </w>
<lengthening/>
<w cat="N" lem="animal" gen="masc" num="plu"> animales </w>
<ps t="interrup"/>
</sf>
</f>

<f h="TEA" st="444.469" et="445.409" id="208">
<sf t="enu" id="208-1">
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="208-1-1"> lo </w>
<w cat="PREP" lem="de" id="208-1-2"> de </w>
<w cat="ART" lem="el" gen="masc" num="plu" id="208-1-3"> los </w>
<w cat="N" lem="animal" gen="masc" num="plu" id="208-1-4"> animales </w>
</sf>
</f>

```
<p>
<f h="INE" st="445.409" et="449.322" id="209">
<sf t="int" id="209-1">
<w cat="PUNCT" lem="¿" id="209-1-1"> ¿ </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="2" id="209-1-2"> sabes </w>
<w cat="P" lem="qué" id="209-1-3"> qué </w>
<w cat="N" lem="animal" gen="masc" num="sing" id="209-1-4"> animal </w>
<w cat="P" lem="lo" num="sing" per="3" id="209-1-5"> le </w>
<w cat="V" lem="gustar" num="sing" id="209-1-6"> gusta </w>
<w cat="Q" lem="más" id="209-1-7"> más </w>
<w cat="PREP" lem="a" id="209-1-8"> a </w>
<w cat="POSS" lem="mi" num="sing" id="209-1-9"> mi </w>
<w cat="N" lem="hermano" gen="masc" num="sing" id="209-1-10"> hermano </w>
<w cat="PUNCT" lem="?" id="209-1-11"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="449.322" et="449.548" id="210">
<sf t="enu" id="210-1">
<w cat="MD" lem="a ver" id="210-1-1"> a ver </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="449.548" et="450.661" id="211">
<sf t="enu" id="211-1">
<w cat="ART" lem="el" gen="masc" num="sing" id="211-1-1"> el </w>
<w cat="N" lem="elefante" gen="masc" num="sing" id="211-1-2"> elefante </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="450.661" et="453.006" id="212">
<sf t="enu" id="212-1">
<w cat="V" lem="tener" num="sing" tie="impf_ind" id="212-1-1"> tenía </w>
<w cat="Q" lem="uno" id="212-1-2"> uno </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="212-1-3"> el </w>
<w cat="Q" lem="otro" gen="masc" id="212-1-4"> otro </w>
<w cat="N" lem="día" gen="masc" num="sing" id="212-1-5"> día </w>
<w cat="PREP" lem="de" id="212-1-6"> de </w>
<w cat="N" lem="plástico" gen="masc" num="sing" id="212-1-7"> plástico </w>
<ps/>
<w cat="REL" lem="que" id="212-1-8"> que </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="212-1-9"> lo </w>
<w cat="V" lem="ver" tie="indf_ind" num="sing" per="1" id="212-1-10"> vi </w>
<w cat="P" lem="yo" num="sing" per="1" id="212-1-11"> yo </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="453.006" et="454.135" id="213">
<sf t="int" id="213-1">
<w cat="PUNCT" lem="¿" id="213-1-1"> ¿ </w>
<w cat="ADV" lem="sí" id="213-1-2"> sí </w>
<w cat="PUNCT" lem="?" id="213-1-3"> ? </w>
</sf>
</f>
</p>
<p>
```

<f h="TEA" st="454.135" et="455.661" id="214">
<sf t="enu" id="214-1">
<w cat="V" lem="venir" num="sing" tie="impf_ind" id="214-1-1"> venía </w>
<w cat="Q" lem="muy" id="214-1-2"> muy </w>
<w cat="ADJ" lem="contento" gen="masc" num="sing" id="214-1-3"> contento </w>
<w cat="PREP" lem="con" id="214-1-4"> con </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="214-1-5"> el </w>
<w cat="N" lem="elefante" gen="masc" num="sing" id="214-1-6"> elefante </w>
</sf>
</f>

<f h="INE" st="455.661" et="457.423" id="215">
<sf t="int" id="215-1">
<w cat="PUNCT" lem="¿" id="215-1-1"> ¿ </w>
<w cat="P" lem="ése" gen="masc" num="sing" per="3" id="215-1-2"> ese </w>
<w cat="ADJ" lem="gris" num="sing" id="215-1-3"> gris </w>
<ps/>
<w cat="ADJ" lem="chiquitín" gen="masc" num="sing" id="215-1-4"> chiquitín </w>
<w cat="PUNCT" lem="?" id="215-1-5"> ? </w>
</sf>
</f>

<f h="TEA" st="457.423" et="457.715" id="216">
<sf t="enu" id="216-1">
<w cat="ADV" lem="sí" id="216-1-1"> sí </w>
</sf>
</f>

<f h="TEA" st="457.715" et="458.816" id="217">
<sf t="enu" id="217-1">
<w cat="Q" lem="uno" id="217-1-1"> uno </w>
<w cat="ADJ" lem="gris" num="sing" id="217-1-2"> gris </w>
<w cat="ADJ" lem="chiquitín" gen="masc" num="sing" id="217-1-3"> chiquitín </w>
</sf>
</f>

<f h="INE" st="458.816" et="459.456" id="218">
<sf t="exc" id="218-1">
<w cat="PUNCT" lem="¡" id="218-1-1"> ¡ </w>
<w cat="INTJ" lem="ah" id="218-1-2"> ah </w>
<w cat="PUNCT" lem="!" id="218-1-3"> ! </w>
</sf>
</f>

<f h="TEA" st="459.456" et="461.222" id="219">
<sf t="enu" id="219-1">
<w cat="P" lem="se" id="219-1-1"> se </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="219-1-2"> lo </w>
<w cat="V" lem="guardar" num="sing" tie="impf_ind" id="219-1-3"> guardaba </w>
<w cat="PREP" lem="en" id="219-1-4"> en </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="219-1-5"> el </w>
<w cat="N" lem="bolsillo" gen="masc" num="sing" id="219-1-6"> bolsillo </w>
<w cat="PREP" lem="de" id="219-1-7"> de </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="219-1-8"> el </w>
<w cat="N" lem="babi" gen="masc" num="sing" id="219-1-9"> babi </w>

```
</sf>
</f>
</p>
<p>
<f h="INE" st="461.222" et="467.754" id="220">
<sf t="enu" id="220-1">
<w cat="MD" lem="es que" id="220-1-1"> es que </w>
<w cat="ADV" lem="no" id="220-1-2"> no </w>
<w cat="V" lem="querer" num="sing" id="220-1-3"> quiere </w>
<w cat="C" lem="que" id="220-1-4"> que </w>
<w cat="NPR" lem="Abel" id="220-1-5"> Abel </w>
<w cat="P" lem="se" id="220-1-6"> se </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="220-1-7"> lo </w>
<w cat="V" num="sing" lem="quitar" id="220-1-8"> quite </w>
<ps t="complex"/>
<w cat="C" lem="porque" id="220-1-9"> porque </w>
<w cat="P" lem="ayer" id="220-1-10"> ayer </w>
<w cat="P" lem="lo" num="sing" per="3" id="220-1-11"> le </w>
<w cat="V" lem="quitar" tie="indf_ind" num="sing" per="3" id="220-1-12"> quitó </w>
<ps/>
<w cat="ART" lem="el" gen="fem" num="sing" id="220-1-13"> la </w>
<w cat="N" lem="moto" gen="fem" num="sing" id="220-1-14"> moto </w>
<w cat="C" lem="y" id="220-1-15"> y </w>
<w cat="P" lem="lo" gen="fem" num="sing" per="3" id="220-1-16"> la </w>
<w cat="V" lem="tirar" tie="indf_ind" num="sing" per="3" id="220-1-17"> tiró </w>
<w cat="ADV" lem="fuera" id="220-1-18"> fuera </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="467.754" et="468.457" id="221">
<sf t="int" id="221-1">
<w cat="PUNCT" lem="¿" id="221-1-1"> ¿ </w>
<w cat="P" lem="quién" id="221-1-2"> quién </w>
<w cat="PUNCT" lem="?" id="221-1-3"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="468.457" et="469.509" id="222">
<sf t="int" id="222-1">
<w cat="PUNCT" lem="¿" id="222-1-1"> ¿ </w>
<w cat="Q" lem="otro" gen="masc" id="222-1-2"> otro </w>
<w cat="N" lem="niño" gen="masc" num="sing" id="222-1-3"> niño </w>
<w cat="PUNCT" lem="?" id="222-1-4"> ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="468.457" et="469.509" id="223">
<sf t="enu" id="223-1">
<bw/>
<w cat="NPR" lem="Abel" id="223-1-1"> Abel </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="469.509" et="470.437" id="224">
<sf t="int" id="224-1">
<w cat="PUNCT" lem="¿" id="224-1-1"> ¿ </w>
```

```
<w cat="ADV" lem="sí" id="224-1-2"> sí </w>
<w cat="PUNCT" lem="?" id="224-1-3"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="470.437" et="471.707" id="225">
<sf t="exc" id="225-1">
<w cat="PUNCT" lem="¡" id="225-1-1"> ¡ </w>
<w cat="INTJ" lem="ahí va" id="225-1-2"> ahí va </w>
<w cat="PUNCT" lem="!" id="225-1-3"> ! </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="471.707" et="475.419" id="226">
<sf t="enu" id="226-1">
<w cat="MD" lem="sí" id="226-1-1"> sí </w>
<ps/>
<w cat="MD" lem="pues" id="226-1-2"> pues </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="226-1-3"> lo </w>
<w cat="V" lem="llevar" num="sing" tie="impf_ind" id="226-1-4"> llevaba </w>
<w cat="C" lem="y" id="226-1-5"> y </w>
<w cat="P" lem="lo" num="sing" per="3" id="226-1-6"> le </w>
<w cat="V" lem="decir" tie="indf_ind" num="sing" per="3" id="226-1-7"> dijo </w>
<w cat="ART" lem="el" gen="fem" num="sing" id="226-1-8"> la </w>
<w cat="N" lem="profesor" gen="fem" num="sing" id="226-1-9"> profesora </w>
<ps t="complex"/>
<w cat="V" lem="guardar" tie="imper" num="sing" per="2" clitic="yes" id="226-1-10"> guárda
</w>
<w cat="P" lem="te" num="sing" per="2" id="226-1-11"> te </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="226-1-12"> lo </w>
<w cat="PREP" lem="en" id="226-1-13"> en </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="226-1-14"> el </w>
<w cat="N" lem="babi" gen="masc" num="sing" id="226-1-15"> babi </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="475.419" et="478.762" id="227">
<sf t="int" id="227-1">
<w cat="PUNCT" lem="¿" id="227-1-1"> ¿ </w>
<w cat="C" lem="y" id="227-1-2"> y </w>
<w cat="V" lem="saber" tie="pres_ind" num="sing" per="2" id="227-1-3"> sabes </w>
<w cat="C" lem="que" id="227-1-4"> que </w>
<w cat="P" lem="lo" num="sing" per="3" id="227-1-5"> le </w>
<w cat="V" lem="encontrar" tie="indf_ind" num="sing" per="1" id="227-1-6"> encontré </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="227-1-7"> el </w>
<ps t="pr"/>
<w cat="PREP" lem="en" id="227-1-8"> en </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="227-1-9"> el </w>
<w cat="N" lem="bolsillo" gen="masc" num="sing" id="227-1-10"> bolsillo </w>
<w cat="PUNCT" lem="?" id="227-1-11"> ? </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="478.762" et="479.196" id="228">
<sf t="int" id="228-1">
<w cat="PUNCT" lem="¿" id="228-1-1"> ¿ </w>
```

<w cat="ART" lem="el" gen="masc" num="sing" id="228-1-2"> el </w>
<w cat="P" lem="qué" id="228-1-3"> qué </w>
<w cat="PUNCT" lem="?" id="228-1-4"> ? </w>
</sf>
</f>

<f h="INE" st="479.196" et="481.375" id="229">
<sf t="enu" id="229-1">
<w cat="P" lem="lo" num="sing" per="3" id="229-1-1"> le </w>
<w cat="V" lem="encontrar" tie="indf_ind" num="sing" per="1" id="229-1-2"> encontré </w>
<w cat="Q" lem="un" gen="masc" id="229-1-3"> un </w>
<w cat="N" lem="chupete" gen="masc" num="sing" id="229-1-4"> chupete </w>
</sf>
</f>

<f h="TEA" st="481.375" et="482.297" id="230">
<sf t="int" id="230-1">
<w cat="PUNCT" lem="¿" id="230-1-1"> ¿ </w>
<w cat="Q" lem="un" gen="masc" id="230-1-2"> un </w>
<w cat="N" lem="chupete" gen="masc" num="sing" id="230-1-3"> chupete </w>
<w cat="PUNCT" lem="?" id="230-1-4"> ? </w>
</sf>
</f>

<f h="INE" st="482.297" et="483.355" id="231">
<sf t="exc" id="231-1">
<w cat="PUNCT" lem="¡" id="231-1-1"> ¡ </w>
<w cat="ADV" lem="sí" id="231-1-2"> sí </w>
<w cat="PUNCT" lem="!" id="231-1-3"> ! </w>
</sf>
</f>

<f h="TEA" st="483.355" et="485.011" id="232">
<sf t="enu" id="232-1">
<w cat="MD" lem="bueno" id="232-1-1"> bueno </w>
<ps/>
<w cat="C" lem="pero" id="232-1-2"> pero </w>
<w cat="ADV" lem="todavía" id="232-1-3"> todavía </w>
<w cat="V" lem="ser" tie="pres_ind" num="sing" per="3" id="232-1-4"> es </w>
<w cat="ADJ" lem="chiquitín" gen="masc" num="sing" id="232-1-5"> chiquitín </w>
<ps/>
</sf>
<sf t="int" id="232-2">
<w cat="PUNCT" lem="¿" id="232-2-1"> ¿ </w>
<w cat="MD" lem="no ?" id="232-2-2"> no ? </w>
</sf>
</f>

<f h="INE" st="485.011" et="488.604" id="233">
<sf t="enu" id="233-1">
<w cat="MD" lem="ya" id="233-1-1"> ya </w>
<ps/>
<w cat="C" lem="pero" id="233-1-2"> pero </w>
<ps/>
<w cat="P" lem="ése" num="sing" per="3" id="233-1-3"> eso </w>

<w cat="V" lem="ser" tie="pres_ind" num="sing" per="3" id="233-1-4"> es </w>
<w cat="C" lem="cuando" id="233-1-5"> cuando </w>
<w cat="V" lem="ser" num="sing" id="233-1-6"> sea </w>
<w cat="N" lem="bebé" num="sing" id="233-1-7"> bebé </w>
</sf>
</f>

<f h="TEA" st="488.604" et="490.369" id="234">
<sf t="enu" id="234-1">
<w cat="C" lem="cuando" id="234-1-1"> cuando </w>
<w cat="V" lem="ser" tie="pres_ind" num="sing" per="3" id="234-1-2"> es </w>
<w cat="N" lem="bebé" num="sing" id="234-1-3"> bebé </w>
<w cat="ADV" lem="sólo" id="234-1-4"> sólo </w>
</sf>
</f>

<f h="INE" st="490.369" et="492.268" id="235">
<sf t="enu" id="235-1">
<w cat="P" lem="lo" num="sing" per="3" id="235-1-1"> le </w>
<w cat="V" lem="ver" tie="indf_ind" num="sing" per="1" id="235-1-2"> vi </w>
<w cat="Q" lem="un" gen="masc" id="235-1-3"> un </w>
<w cat="N" lem="chupete" gen="masc" num="sing" id="235-1-4"> chupete </w>
<nl/>
</sf>
</f>

<f h="TEA" st="492.268" et="496.511" id="236">
<sf t="enu" id="236-1">
<w cat="MD" lem="bueno" id="236-1-1"> bueno </w>
<ps/>
<w cat="V" lem="haber" tie="pres_ind" num="sing" per="3" id="236-1-2"> hay </w>
<w cat="N" lem="niño" gen="masc" num="plu" id="236-1-3"> niños </w>
<w cat="REL" lem="que" id="236-1-4"> que </w>
<lengthening/>
<w cat="V" lem="tener" tie="pres_ind" num="plu" per="3" id="236-1-5"> tienen </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="236-1-6"> el </w>
<w cat="N" lem="chupete" gen="masc" num="sing" id="236-1-7"> chupete </w>
<ps/>
<w cat="Q" lem="mucho" gen="masc" id="236-1-8"> mucho </w>
<w cat="N" lem="tiempo" gen="masc" num="sing" id="236-1-9"> tiempo </w>
</sf>
</f>

<f h="INE" st="496.511" et="496.93" id="237">
<sf t="enu" id="237-1">
<w cat="ADV" lem="sí" id="237-1-1"> sí </w>
</sf>
</f>

<f h="INE" st="496.93" et="499.906" id="238">
<sf t="enu" id="238-1">
<w cat="C" lem="pero" id="238-1-1"> pero </w>
<ps/>
<w cat="NPR" lem="Abel" id="238-1-2"> Abel </w>
<w cat="ADV" lem="no" id="238-1-3"> no </w>

<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="238-1-4"> lo </w>
<w cat="AUX" lem="llevar" num="sing" id="238-1-5"> lleva </w>
<w cat="V" lem="poner" tie="part" id="238-1-6"> puesto </w>
</sf>
</f>

<f h="TEA" st="499.906" et="500.254" id="239">
<sf t="int" id="239-1">
<w cat="PUNCT" lem="¿" id="239-1-1"> ¿ </w>
<w cat="MD" lem="no ?" id="239-1-2"> no ? </w>
</sf>
</f>

<f h="TEA" st="500.254" et="503.219" id="240">
<sf t="enu" id="240-1">
<w cat="ADV" lem="sólo" id="240-1-1"> sólo </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="240-1-2"> lo </w>
<w cat="V" lem="llevar" num="sing" id="240-1-3"> lleva </w>
<w cat="PREP" lem="en" id="240-1-4"> en </w>
<w cat="ART" lem="el" gen="masc" num="sing" id="240-1-5"> el </w>
<w cat="N" lem="bolsillo" gen="masc" num="sing" id="240-1-6"> bolsillo </w>
<ps/>
<w cat="ADV" lem="por si acaso" id="240-1-7"> por si acaso </w>
</sf>
</f>

<f h="TEA" st="503.219" et="504.495" id="241">
<sf t="int" id="241-1">
<w cat="PUNCT" lem="¿" id="241-1-1"> ¿ </w>
<w cat="MD" lem="no ?" id="241-1-2"> no ? </w>
</sf>
</f>

<f h="INE" st="504.495" et="506.404" id="242">
<sf t="enu" id="242-1">
<w cat="C" lem="si" id="242-1-1"> si </w>
<w cat="ADV" lem="nunca" id="242-1-2"> nunca </w>
<w cat="P" lem="lo" gen="masc" num="sing" per="3" id="242-1-3"> lo </w>
<w cat="V" lem="necesitar" num="sing" id="242-1-4"> necesita </w>
</sf>
</f>

<f h="TEA" st="506.404" et="506.918" id="243">
<sf t="enu" id="243-1">
<w cat="ADV" lem="no" id="243-1-1"> no </w>
</sf>
</f>

<f h="TEA" st="506.918" et="509.407" id="244">
<sf t="enu" id="244-1">
<nl/>
</sf>
</f>

```
<p>
<f h="TEA" st="509.407" et="510.586" id="245">
<sf t="enu" id="245-1">
<w cat="MD" lem="pues nada" id="245-1-1"> pues nada </w>
<ps/>
<w cat="Q" lem="muy" id="245-1-2"> muy </w>
<w cat="ADV" lem="bien" id="245-1-3"> bien </w>
<w cat="NPR" lem="Inés" id="245-1-4"> Inés </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="510.586" et="512.102" id="246">
<sf t="enu" id="246-1">
<w cat="P" lem="nos" num="plu" per="1" id="246-1-1"> nos </w>
<w cat="AUX" lem="ir" tie="pres_ind" num="plu" per="1" id="246-1-2"> vamos </w>
<w cat="PREP" lem="a" id="246-1-3"> a </w>
<w cat="V" lem="ir" tie="inf" id="246-1-4"> ir </w>
<w cat="PREP" lem="para" id="246-1-5"> para </w>
<w cat="N" lem="clase" gen="fem" num="sing" id="246-1-6"> clase </w>
<ps/>
</sf>
<sf t="int" id="246-2">
<w cat="PUNCT" lem="¿" id="246-2-1"> ¿ </w>
<w cat="MD" lem="vale ?" id="246-2-2"> vale ? </w>
</sf>
</f>
</p>
<p>
<f h="INE" st="512.102" et="512.647" id="247">
<sf t="enu" id="247-1">
<w cat="MD" lem="vale" id="247-1-1"> vale </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="512.647" et="513.469" id="248">
<sf t="exc" id="248-1">
<w cat="PUNCT" lem="¡" id="248-1-1"> ¡ </w>
<w id="248-1-2"> ale </w>
<w cat="PUNCT" lem="!" id="248-1-3"> ! </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="513.469" et="514.485" id="249">
<sf t="enu" id="249-1">
<w cat="V" lem="ir" tie="pres_ind" num="plu" per="1" clitic="yes" id="249-1-1"> vámo </w>
<w cat="P" lem="nos" num="plu" per="1"id="249-1-2"> nos </w>
</sf>
</f>
</p>
</Text>
</fml>
```

Apéndice B

Listados completos de frecuencias de formas y categorías léxicas

1. Grupo de edad: 3, 4 y 5 años

a. Frecuencia de formas

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
y	1139
no	708
sí	641
que	531
la	488
a	479
el	467
un	425
me	393
yo	386
se	363
de	348
en	348
mi	322
una	284
los	237
lo	215
es	200
con	191
le	177
pero	175
porque	150
pues	150
ya	128
tengo	128
también	116
para	113
qué	104
las	99
está	98
ha	97
sé	96
por	96
al	92
dos	89
tiene	87
uno	86
cuando	84
casa	84
más	82
es que	79
aquí	78

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
he	74
otro	72
muy	67
Candi	65
madre	61
padre	59
mira	59
si	59
hay	58
así	58
como	54
ahí	54
va	51
estaba	48
cinco	48
tres	46
eso	46
grande	45
su	45
del	44
voy	44
otra	44
este	43
o	41
esto	39
mamá	39
te	38
jugar	38
luego	36
niño	35
era	34
muchos	34
cuatro	34
ese	33
por qué	33
lobo	33
había	33
tenía	33
Guadalupe	33
mí	33
todos	33
mucho	32

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
nada	32
son	32
quién	31
Carmen	31
nos	31
llama	30
sólo	29
sabes	29
día	28
quiero	27
cómo	27
Balbi	27
fue	26
niña	26
tú	25
vamos	25
hermana	24
cosa	24
pone	23
hacer	23
hoy	23
coche	23
veces	23
ahora	23
dónde	23
a ver	23
ni	23
cama	22
Raúl	22
ella	22
después	21
hemos	21
abuela	21
perro	21
hasta	21
fui	21
mis	21
algunas	21
dijo	21
bueno	21
seis	20
estás	20
ocho	20
quería	20
hecho	20
vez	20
patio	20
parece	19
han	19
martes	19

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
sol	19
agua	19
cuento	19
nueve	19
fútbol	19
Daniel	18
presente	18
elefante	18
cosas	18
oye	18
vale	18
lo que	18
rojo	18
entonces	18
azul	17
buenos días	17
dicho	17
papá	17
niños	17
bien	17
ah	17
iba	17
pis	16
e	16
casi	16
juguetes	16
siete	16
abuelo	16
están	16
un poco	16
Esther	16
soy	16
ir	15
vino	15
claro	15
tienes	15
jueves	15
grandes	15
lunes	15
él	15
unos	15
unas	15
señorita	15
Mulan	14
euro	14
Abel	14
cole	14
vi	14
malo	14
Marta	14

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
corriendo	14
da	14
visto	14
Arturo	14
Marisol	14
ayer	14
cabeza	13
pasa	13
todo	13
esta	13
leche	13
muchas	13
Carlos	13
primero	13
sábado	13
diez	13
ser	13
arriba	13
gusta	13
pelota	13
tu	13
cayó	13
hace	13
allí	13
roto	13
tele	13
donde	13
primo	12
domingo	12
Javier	12
van	12
días	12
hermano	12
clase	12
vaso	12
castillo	12
María	12
prima	12
doce	12
veo	12
perros	11
años	11
payaso	11
viene	11
esa	11
siempre	11
Laura	11
gato	11
verde	11
Ana Pilar	11

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
galletas	11
Vicente	11
chica	11
Caperucita	11
amigo	11
cumple	11
hago	11
miércoles	11
mía	11
Jorge	10
tarta	10
Rodrigo	10
mío	10
acuerdo	10
ojos	10
tampoco	10
deja	10
estaban	10
mal	10
comprar	10
dice	10
jaula	10
algo	10
hablar	10
piscina	10
perra	10
venga	10
lleva	10
tía	10
Tomelloso	10
hizo	10
quince	10
puesto	10
ver	10
otros	10
tenemos	10
regalado	10
piedras	10
nunca	9
colacao	9
cumpleaños	9
trabaja	9
puede	9
ratón	9
juego	9
puedo	9
pasado	9
pelo	9
viernes	9
Furi	9

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
campo	9
caballo	9
jugando	9
algunos	9
ordenador	9
once	9
Inés	9
tan	9
cara	9
decir	8
casita	8
semáforo	8
pintar	8
ay	8
Mari	8
Cristina	8
chiquitín	8
ropa	8
rosa	8
abuelita	8
Ana	8
fuera	8
varicela	8
céntimos	8
bebé	8
de leche	8
febrero	8
vio	8
amigos	8
pequeño	8
guerra	8
ido	8
cogió	8
dibujos	8
animales	8
hice	8
pala	8
comer	8
Spiderman	8
has	8
pasó	8
abajo	8
año	7
menos	7
comida	7
tiro	7
As	7
león	7
toca	7
igual que	7

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
tortuga	7
a lo mejor	7
toda	7
pupa	7
salió	7
cerdito	7
antes	7
di	7
caído	7
el que	7
película	7
chico	7
José	7
Sofía	7
metió	7
jugamos	7
echa	7
puso	7
lila	7
cuello	7
comió	7
señorito	7
Sergio	7
pequeña	7
número	7
por la noche	7
comprado	7
Madrid	7
subir	7
deprisa	7
falta	7
mama	7
café	7
cien	7
pera	7
haciendo	7
todavía	7
meter	7
venir	7
piano	7
Marcos	7
Luis	7
traído	7
esas	6
vivo	6
mala	6
pica	6
picar	6
mañana	6
cuchara	6

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
pueden	6
mayor	6
podía	6
acosté	6
ala	6
poner	6
Manuel	6
vas	6
feo	6
rompe	6
rana	6
Antonio	6
Liana	6
calle	6
esos	6
montón	6
trabajar	6
murió	6
números	6
acabado	6
porta	6
cae	6
pantalones	6
gimnasio	6
fuego	6
queda	6
buena	6
dragón	6
jugado	6
cuentos	6
Iván	6
perdió	6
moto	6
chicos	6
compró	6
venido	6
vive	6
ce	6
pollo	6
quiere	6
Aila	6
ea	6
cojo	6
rinoceronte	6
serpiente	6
culo	6
poquito	6
negro	6
leer	6
barco	6

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
pille pille	6
río	6
lejos	6
habitación	5
tambor	5
Mónica	5
salen	5
cabritillos	5
catorce	5
trece	5
de verdad	5
tenedor	5
tirar	5
noche	5
doy	5
hacía	5
corazón	5
rapar	5
la que	5
pijama	5
pájaro	5
Selene	5
tobogán	5
hora	5
reyes	5
Saranya	5
estoy	5
Evelio José	5
chuches	5
pato	5
morro	5
sea	5
igual	5
oveja	5
será	5
alta	5
subo	5
llegó	5
compra	5
segundo	5
coger	5
boca	5
daba	5
zapatos	5
líquido	5
puedes	5
tío	5
cartera	5
entera	5
sola	5

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
mano	5
gallina	5
Octopus	5
parque	5
llorando	5
regaló	5
sitio	5
Rubén	5
silla	5
taza	5
de pie	5
sabes ?	5
cuál	5
cero	5
huevos	5
colegio	5
trabajo	5
sólido	5
gol	5
amarillo	5
camiseta	5
portero	5
ningún	5
puerta	5
echo	5
Aitor	5
Natalia	5
juguete	5
oso	5
sale	5
zapatillas	5
quita	5
Miguel Ángel	5
carnaval	5
estamos	5
Pepe	5
Papá Noel	5
sido	5
Ronaldinho	5
Antonio José	5
José Miguel	5
solo	5
gustan	5
dieciocho	5
naranja	5
Julia	5
míos	5
camiones	5
hucha	5
comido	5

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
mataba	5
Granada	5
piso	5
ponemos	5
mejor	5
Zaragoza	5
escondite	5
ellos	5
chicas	5
brazos	5
lee	5
pequeñito	4
tocó	4
aunque	4
colorado	4
coge	4
fantasma	4
invierno	4
tiempo	4
debajo de	4
china	4
regalos	4
bota	4
tigre	4
fueron	4
hablando	4
llamaba	4
digital	4
Rosa Gema	4
salta	4
abrió	4
cerditos	4
palo	4
escalera	4
creo	4
olvidó	4
seguro que	4
salgo	4
veranos	4
sábados	4
tacones	4
manos	4
Bruno	4
hacemos	4
allá	4
pueblo	4
gordo	4
manzana	4
por eso	4
domingos	4

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
muere	4
pico	4
soñado	4
partido	4
baño	4
a veces	4
otras	4
mató	4
animal	4
hijo	4
anda	4
llevo	4
salía	4
fruta	4
salido	4
Aitana	4
monstruo	4
gatos	4
Simpsons	4
habla	4
peine	4
detrás	4
céntimo	4
cantado	4
conejo	4
gafas	4
colorín	4
suelo	4
daño	4
rumanos	4
disco	4
llamó	4
estos	4
sangre	4
be	4
caí	4
enero	4
quitado	4
mordió	4
adiós	4
mariposa	4
señor	4
verano	4
venía	4
abuelito	4
come	4
ven	4
regalaron	4
ramo	4
cuna	4

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
Guadix	4
Emilio	4
trajes	4
les	4
dentro	4
collar	4
pie	4
juega	4
cortar	4
los que	4
pegan	4
bosque	4
camión	4
debajo	4
pasar	4
Adrián	4
viendo	4
convierte	4
Pepón	4
traje	4
puntitos	4
chupete	4
cerca	4
marrón	4
Andrés	4
Juanito	4
Elena	4
podemos	4
dormir	4
volando	4
bajé	4
hicimos	4
bolsillo	4
euros	4
creía	4
tira	4
casco	4
pelado	4
quitan	4
diciembre	4
lápiz	4
Play	4
piernas	4
mayores	4
montar	4
Eva	4
tiró	4
papás	4
Pachín	4
serpientes	4

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
comí	4
contar	4
balón	4
biberón	3
pájaros	3
pollito	3
señalar	3
pinto	3
calentita	3
sillón	3
pintamos	3
encanta	3
Manga	3
cuidado	3
árbol	3
cometa	3
caja	3
enseñó	3
Palma	3
Azorín	3
dientes	3
duerme	3
llevó	3
entre	3
plus	3
Talía	3
pones	3
veintiocho	3
periódico	3
Donald	3
rueda	3
gustaba	3
dinero	3
mucha	3
fuimos	3
dos mil seis	3
espada	3
cogen	3
monta	3
cerdo	3
en coche	3
garbanzos	3
bonito	3
luchar	3
sus	3
comía	3
frío	3
hojas	3
arena	3
corto	3

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
cereales	3
camino	3
echó	3
Héctor	3
conejito	3
tuvo	3
magia	3
flechas	3
veía	3
siéntate	3
nuestra	3
calla	3
Goofy	3
Mickey	3
Cora	3
suave	3
lavarme	3
celebramos	3
llegaba	3
guardo	3
cuánto	3
bebés	3
casen	3
pilas	3
tarde	3
tiritas	3
mesa	3
pego	3
ducha	3
eran	3
rompió	3
canción	3
pescado	3
mamás	3
Ángel	3
leemos	3
pisado	3
basura	3
estar	3
fecha	3
llora	3
nosotros	3
primos	3
careta	3
salimos	3
nadie	3
Santa Claus	3
Riki	3
llave	3
diecisiete	3

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
cuadra	3
ves	3
Minnie	3
ninguno	3
mueve	3
yayo	3
chimenea	3
echaba	3
escrito	3
acera	3
tigre	3
hacen	3
tocar	3
ochenta	3
soles	3
gané	3
quedan	3
traer	3
chocolate	3
guiña	3
metes	3
veinticuatro	3
monstruos	3
cogimos	3
tijeras	3
Caperucita Roja	3
plancha	3
sillita	3
temprano	3
tienen	3
caramelo	3
caía	3
blanca	3
paré	3
Jesús	3
despacio	3
alto	3
lobos	3
Manzanares	3
duende	3
cuándo	3
bigote	3
viento	3
granos	3
fichas	3
hache	3
uy	3
ve	3
porto	3
todas	3

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
películas	3
foca	3
base	3
pulsera	3
bombones	3
caer	3
cristal	3
pierna	3
Superman	3
patos	3
vueltas	3
enamorados	3
caso	3
cocina	3
haces	3
carretera	3
fuerte	3
metían	3
jugué	3
dejan	3
Miguel	3
abeja	3
escondió	3
cada	3
trae	3
flores	3
dio	3
brasero	3
parar	3
mojado	3
dieciséis	3
palomos	3
humo	3
Torres	3
alguna	3
diario	3
tortugas	3
pan	3
sin	3
subido	3
escribir	3
dormí	3
nave	3
desayunado	3
cogía	3
buenos	3
coches	3
rey	3
madera	3
caballito	3

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
hoja	3
contenedores	3
empieza	2
cabe	2
perdón	2
niñas	2
King Kong	2
recreo	2
cazador	2
Diana	2
llaman	2
ganas	2
papel	2
quedado	2
juntamos	2
Balbino	2
caiga	2
camas	2
cajón	2
sabe	2
vestirse	2
gomina	2
golpe	2
enciende	2
cola	2
agosto	2
Jose	2
pinchó	2
llevar	2
Adelardo	2
cumplo	2
nombre	2
cuenta	2
abrigo	2
policía	2
abril	2
canasta	2
Lanza	2
abeto	2
perdido	2
repasamos	2
juegos	2
saltos	2
bebido	2
llevaron	2
pavo	2
colchonetas	2
Shin-Chan	2
coges	2
mosca	2

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
pequeñitas	2
levanté	2
alguien	2
Mallorca	2
atacaron	2
negros	2
saco	2
encima	2
acuerda	2
muñeco	2
dibujo	2
cantaba	2
libro	2
dame	2
septiembre	2
ganó	2
discos	2
Pepito	2
bañar	2
bye	2
creían	2
Barbie	2
tenían	2
bajar	2
ayudar	2
cansado	2
saliendo	2
araña	2
sabían	2
Paco	2
copa	2
cachorros	2
adentro	2
malas	2
julio	2
pata	2
lengua	2
cartel	2
marrones	2
soplarlas	2
hijas	2
encender	2
suyo	2
truco	2
hazme	2
picado	2
merendé	2
albañiles	2
aire	2
arroz	2

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
oscuro	2
poco	2
huevo	2
oh	2
calcetines	2
cera	2
patas	2
acá	2
palas	2
gente	2
búho	2
cierra	2
beber	2
dije	2
plátano	2
contra	2
delante	2
Arturito	2
encontré	2
gusano	2
colonia	2
manda	2
gateando	2
marzo	2
pillado	2
encima de	2
chupa	2
bebo	2
ayuda	2
abiertos	2
sopa	2
caballeros	2
sofá	2
soñó	2
haberte	2
ballena	2
armario	2
enrolla	2
amiguitos	2
echar	2
y todo	2
vestir	2
cierran	2
Dulce	2
Rollo	2
ellas	2
grúa	2
quedo	2
os	2
vuela	2

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
botas	2
cosieron	2
literas	2
darle	2
profesor	2
cepas	2
toma	2
cuchillos	2
mechero	2
boli	2
nido	2
charco	2
entrar	2
cosió	2
señalarlo	2
calendario	2
hasta que	2
radio	2
andar	2
dan	2
maestros	2
ficha	2
subí	2
enterré	2
secreto	2
ponerle	2
aparcados	2
logopeda	2
perrito	2
Albacete	2
portado	2
conozco	2
al final	2
fíjate	2
zumo	2
tiré	2
noviembre	2
corría	2
mami	2
sueño	2
sientas	2
leído	2
demás	2
quietos	2
levantó	2
coloco	2
parada	2
maestro	2
guapa	2
empresa	2

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
compran	2
vemos	2
gorda	2
mete	2
frente	2
teléfono	2
pagar	2
Pedro	2
dejé	2
coleta	2
levantado	2
cazó	2
esperando	2
tripas	2
luna	2
pongo	2
de hierro	2
aros	2
atenta	2
salir	2
dar	2
columpio	2
muevo	2
Villamuriel	2
doctor	2
pepino	2
veintidós	2
bola	2
hola	2
quite	2
baranda	2
familia	2
zapato	2
sube	2
limpia	2
claro que sí	2
tigres	2
comerte	2
llamas	2
pantano	2
sentado	2
cajitas	2
móvil	2
carrito	2
metieron	2
pisa	2
por la mañana	2
quinto	2
reina	2
estuve	2

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
sujetado	2
manguitos	2
meten	2
de miedo	2
contenedor	2
Pokémon	2
quemé	2
veinticinco	2
rota	2
puntos	2
conmigo	2
duermen	2
fino	2
tuyo	2
pedido	2
pisos	2
espadas	2
Zipi	2
portería	2
vimos	2
escaleras	2
patitos	2
deportivas	2
gorro	2
podido	2
viéndolos	2
leerlo	2
Marca	2
escape	2
chinas	2
horario	2
muñecas	2
Jorgete	2
papa	2
colores	2
pegó	2
Inma	2
El País	2
encendía	2
haber	2
litera	2
Patricia	2
pies	2
despacito	2
pequeños	2
caían	2
estuches	2
guitarra	2
Carla	2
Socuéllamos	2

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
enferma	2
ponga	2
llevas	2
despertado	2
Discolandia	2
buscar	2
ge	2
galleta	2
gitanos	2
contaron	2
pino	2
octubre	2
al lado de	2
El Mundo	2
echado	2
dicen	2
dobermanes	2
i	2
muerde	2
oyó	2
colorear	2
babi	2
Cassano	2
mar	2
noches	2
pega	2
digan	2
goles	2
correr	2
equipo	2
convirtieron	2
letra	2
jope	2
ventana	2
digo	2
playa	2
peligrosas	2
señores	2
Jorge Torres	2
buscando	2
espalda	2
amo	2
enseguida	2
dura	2
tiraba	2
responsable	2
mayo	2
juntos	2
contado	2
somos	2

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
sabía	2
piedra	2
pantera	2
nueva	2
saqué	2
vea	2
sacar	2
déjalo	2
mujer	2
Eto'o	2
barba	2
pelotas	2
carro	2
obra	2
tonto	2
acabó	2
larguita	2
vale ?	2
picó	2
globos	2
tarjetas	2
íbamos	2
cansino	2
cena	2
pintalabios	2
quieren	2
llevarle	2
cincuenta	2
punta	2
orejas	2
mancha	2
corren	2
vestido	2
dejaba	2
tanto	2
sigue	2
maestra	2
Puyol	2
tercero	2
castaña	2
Vane	2
sabemos	2
tiran	2
chicles	2
pinchar	2
mas	2
mordía	2
abuelos	2
bajo	2
dando	2

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
macarrones	2
escuela	2
peinado	2
gorra	2
índice	2
hundían	2
hamster	2
patada	2
noventa	2
cumplí	2
tocando	2
Simpson	2
compraba	2
baloncesto	2
hueco	2
amigas	2
barriga	2
furgoneta	2
ganado	2
duermo	2
Mari Flor	2
Carlota	2
vaya	2
peluche	2
pirulo	2
levanto	2
voltereta	2
baja	2
estufa	2
rula	2
peluca	2
burro	2
por el suelo	2
trabajamos	2
saltan	2
Playmobiles	2
tacón	2
encontró	2
castaño	2
abrir	2
supermercado	2
vivía	2
ruiseñor	1
chino	1
Carmencita	1
Phoskitos	1
sucio	1
Zidane	1
en vez de	1
nadan	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
dormirme	1
manchegas	1
dejáis	1
dándole	1
lana	1
leí	1
vuelta	1
miel	1
DVD	1
botones	1
música	1
enseñado	1
i griega	1
cajones	1
jo	1
madre mía	1
empiezas	1
lías	1
leyendo	1
cambiarla	1
barbilla	1
carpeta	1
peina	1
pinté	1
inventamos	1
leones	1
pollitos	1
Messi	1
eché	1
coloreado	1
engañó	1
Hansel	1
ventanas	1
conducir	1
trozo	1
trampa	1
cuidadito	1
Belletti	1
cesta	1
voz	1
valga	1
enseña	1
permiso	1
levanta	1
rápido	1
mandos	1
marcar	1
Chuki	1
disparado	1
Tomasa	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
granitos	1
Antonio Delgado	1
ovejita	1
chupachuses	1
Halloween	1
torre	1
peguen	1
blancos	1
malita	1
veinte	1
subida	1
nana	1
buhardilla	1
sacó	1
verlo	1
pudimos	1
compre	1
rosco	1
Gameboy	1
canguro	1
muchacho	1
acerca	1
llevaba	1
afinando	1
verla	1
Jorquera	1
colocamos	1
doble	1
nubes	1
farmacia	1
tela	1
jugó	1
montaña	1
acuerdas	1
búhos	1
columna	1
pirata	1
largo	1
saliera	1
vocales	1
lavarte	1
escapar	1
barrita	1
recoger	1
no ?	1
bolso	1
sitios	1
saca	1
paso	1
Manolo Torres	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
abierta	1
regalaba	1
convertía	1
bruja	1
montando	1
cielo	1
compras	1
nueces	1
nombres	1
aprender	1
ardilla	1
patatas	1
ja	1
enfría	1
corta	1
metérmelo	1
vengo	1
lado	1
siendo	1
espera	1
caes	1
Betis	1
ciprés	1
comiste	1
abrochar	1
rojas	1
tiraban	1
Soco	1
Bugs Bunny	1
palos	1
toco	1
rascarme	1
puente	1
utilizaste	1
primer	1
estiras	1
Tintín	1
Loreya	1
baila	1
estas	1
damos	1
pintura	1
pierda	1
control	1
teles	1
último	1
vía	1
pedales	1
atropelló	1
al principio	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
mando	1
colorín colorado	1
volvían	1
idioma	1
intentó	1
enrollas	1
calladito	1
tazón	1
conduzco	1
jarabe	1
caíste	1
hundió	1
comprarse	1
grapar	1
ponerme	1
selva	1
pintaba	1
asfixió	1
echarme	1
meto	1
fin de semana	1
papi	1
mentiras	1
Iniesta	1
meterlo	1
poquitín	1
Ramón	1
Irene	1
tumba	1
jugaba	1
aburro	1
saltemos	1
sueltas	1
máquina	1
anillo	1
matarlo	1
polvos	1
acuesta	1
llamarle	1
rectangular	1
contigo	1
fría	1
cumplido	1
descosió	1
gatear	1
chocamos	1
tostada	1
fiesta	1
zancadilla	1
Oscar	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
pasearlo	1
comiera	1
escupir	1
romperte	1
soñando	1
traigo	1
cresta	1
caperuza	1
salto	1
bañé	1
Julio	1
portaba	1
quemar	1
comprarme	1
perece	1
vieja	1
veintitrés	1
anginas	1
barcos	1
grano	1
pingüino	1
Juan	1
pinchado	1
entrenadora	1
Alejandro	1
arrancado	1
mentirosa	1
Lili	1
pinta	1
jardín	1
Daisy	1
punto	1
fresa	1
debes	1
nadar	1
gustaban	1
tenga	1
huía	1
callarse	1
jugaste	1
roja	1
hurga	1
inventar	1
póster	1
trabajas	1
cuchillo	1
seas	1
bailamos	1
bata	1
despierte	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
bici	1
blancas	1
guardarlos	1
Popeye	1
colgante	1
llego	1
Batman	1
crema	1
Maxi López	1
salga	1
comprármelo	1
hicieron	1
venda	1
enanitos	1
María José	1
compañeros	1
misa	1
arco	1
Fernando Torres	1
comidita	1
prepara	1
limpitas	1
cartón	1
cantar	1
enganchada	1
agujero	1
sapo	1
Lourdes	1
invité	1
señora	1
segunda	1
dale	1
colegios	1
señoreada	1
pudiste	1
mismos	1
Cango	1
preferido	1
acercar	1
rama	1
quemó	1
san	1
queso	1
alcanzaba	1
sentarte	1
blanco	1
sombrero	1
ovejitas	1
cosita	1
habido	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
rompieron	1
Manchitas	1
viva	1
escondidas	1
rato	1
zampó	1
perdonas	1
nube	1
empates	1
erres	1
enganchaba	1
disfrazado	1
riendo	1
lanzas	1
canciones	1
asustan	1
trabajó	1
tantas	1
abierto	1
bicicleta	1
a gatas	1
cayeron	1
encontraron	1
traga	1
rectángulo	1
pilla besos	1
entendemos	1
echamos	1
mordiéndola	1
termine	1
trozos	1
sombra	1
funciona	1
fritas	1
regalar	1
matar	1
chalé	1
dejara	1
esté	1
picaba	1
tomado	1
quieres	1
tristes	1
horquilla	1
mochila	1
arroparme	1
perdieron	1
lavabo	1
sacan	1
vomitó	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
en cuanto	1
restaurante	1
tata	1
perdíamos	1
regalarle	1
enanito	1
demonio	1
saludo	1
leotardo	1
camioneros	1
ey	1
izquierda	1
Izán	1
echaré	1
disfrace	1
malos	1
coloretes	1
empezamos	1
Noel	1
televisión	1
pintando	1
llamar	1
encuentra	1
junto	1
magdalenas	1
perdiz	1
sofás	1
flojo	1
bolsas	1
estirada	1
animalitos	1
borra	1
helicóptero	1
octavo	1
fácil	1
despertó	1
dejo	1
rompí	1
chiquituja	1
González	1
sierras	1
escondí	1
estrecha	1
paramos	1
Rompehuesos	1
saques	1
Carmina	1
braguitas	1
pequeñas	1
carpa	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
pajarito	1
lavando	1
herida	1
pulgas	1
cogieron	1
espachurra	1
Rosa	1
colorines	1
cogiendo	1
rosas	1
bañamos	1
libros	1
dibujado	1
Piolín	1
purpurina	1
pide	1
chupachús	1
sal	1
pajita	1
jugaban	1
encontraban	1
suena	1
magos	1
sierra	1
profesores	1
enseño	1
echaron	1
junio	1
Blancanieves	1
deportes	1
fuente	1
granja	1
vacaciones	1
limpiar	1
contarlo	1
Tasmania	1
inventado	1
bocadillos	1
delfín	1
divirtiendo	1
extraterrestre	1
añitos	1
vendrá	1
fideos	1
guapo	1
a primeros	1
salva	1
chaqueta	1
arar	1
probarla	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
pum	1
felicitarla	1
banda	1
explotó	1
mintió	1
toallero	1
pagando	1
Playmobil	1
cosido	1
olvidado	1
buscado	1
etiqueta	1
baile	1
masajes	1
Papá	1
cochinillas	1
regalé	1
merendar	1
cuarto	1
recogiendo	1
lío	1
verdad ?	1
faltan	1
ruido	1
vámonos	1
Benidorm	1
columpiándose	1
Jorgito	1
olivo	1
además	1
polvorones	1
avispa	1
nieve	1
celebrar	1
paz	1
disparó	1
efe	1
pastel	1
coneja	1
paseo	1
estiraban	1
apellido	1
atrancó	1
pingüinos	1
lechuza	1
acaba	1
caramelos	1
miau	1
mata	1
cambian	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
terminar	1
guantes	1
hagas	1
quedar	1
pez	1
partidos	1
quedando	1
jefe	1
usar	1
andando	1
Wendy	1
Sierra Nevada	1
felicitar	1
cocodrilo	1
ponía	1
vamos a ver	1
block	1
antena	1
mayúscula	1
duermas	1
llévate	1
tonterías	1
gata	1
caca	1
quedé	1
cuántos	1
ten	1
bebía	1
dimos	1
quedarse	1
cuadro	1
pincho	1
primas	1
flor	1
quemada	1
terminado	1
resfriar	1
compro	1
tartas	1
chilla	1
iguales	1
americana	1
guardería	1
limón	1
recojo	1
habrás	1
círculos	1
noticia	1
regulín regulán	1
gallo	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
rosquillo	1
hermanitos	1
recoge	1
cerrado	1
estirando	1
pintado	1
susto	1
violonchelo	1
hello	1
San Valentín	1
subibaja	1
dioses	1
quemaron	1
lista	1
cañón	1
regala	1
pasillo	1
guardar	1
así que	1
africano	1
celebré	1
Martín	1
Zaragoza Madrid	1
volaba	1
soldados	1
sabéis	1
tendré	1
agarrado	1
caemos	1
corté	1
lavo	1
balones	1
cansamos	1
aprendiendo	1
tripita	1
empujado	1
ponérselo	1
gusanito	1
cuerda	1
chillando	1
patito	1
manchan	1
conejos	1
paseos	1
borrador	1
escucha	1
caza	1
ceno	1
podías	1
disfraz	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
adoptar	1
escapa	1
perdiéramos	1
raro	1
joven	1
añito	1
chiquitina	1
mismo	1
llévale	1
Felisa	1
echada	1
vacía	1
quitó	1
pude	1
morada	1
perdí	1
quieran	1
ruedas	1
tardó	1
sentada	1
morado	1
Kinder	1
cuerpo	1
miedo	1
podíamos	1
qué tal	1
de menos	1
veintiséis	1
empezó	1
siquiera	1
escondido	1
diciendo	1
diablo	1
pinche	1
castigó	1
Elisa	1
comérselos	1
ESO	1
vestirme	1
treinta y cinco	1
yegua	1
comemos	1
metemos	1
frutas	1
cacé	1
lloré	1
Campo de Criptana	1
verte	1
miro	1
beso	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
Manolo	1
tractores	1
separados	1
colaron	1
Shan	1
flaca	1
merienda	1
cogemos	1
pasta	1
acostarse	1
más que	1
vienen	1
pudieron	1
entiende	1
tranquila	1
labios	1
rotulador	1
por favor	1
corbata	1
telas	1
regañamos	1
salieron	1
encontrado	1
empujándole	1
oía	1
enfada	1
calzado	1
superhéroe	1
secado	1
echemos	1
nuestro	1
té	1
rascarle	1
después de	1
papeles	1
pinturines	1
veintinueve de agosto	1
mataban	1
tardaron	1
cantas	1
liebres	1
yaya	1
hermanita	1
llevan	1
elefantes	1
Jorge González	1
alfombra	1
ranas	1
ratones	1
lava	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
caretas	1
vinieron	1
estabas	1
borrará	1
panda	1
Gretel	1
abalanzó	1
así es que	1
Mickey Mouse	1
liebre	1
vídeos	1
nuevas	1
Barce	1
hija	1
empezar	1
navidad	1
uf	1
medio	1
dices	1
vete	1
apuntado	1
sopita	1
nació	1
José María Delgado	1
Rosi	1
pingoretas	1
letras	1
sustos	1
columpios	1
dedos	1
buh	1
lavado	1
comimos	1
inglés	1
Argamasilla	1
superhéroes	1
trepa	1
traen	1
tazas	1
amarilla	1
picaron	1
escuchado	1
llegáis	1
cáscara	1
cacahuete	1
castiga	1
poquitos	1
mojada	1
migas	1
profe	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
rastro	1
debido	1
Fernando Alonso	1
escribí	1
gustó	1
ratoncito	1
lavarse	1
pedí	1
pupas	1
pinchándole	1
fantasmas	1
cebra	1
salvarla	1
Larsson	1
dieciséis de agosto	1
hembra	1
zanahoria	1
casados	1
pisar	1
saludé	1
tomar	1
ahí va	1
pegarme	1
despertar	1
recortamos	1
mariposas	1
diste	1
teletexto	1
veas	1
soltó	1
Barcelona	1
entreno	1
mastica	1
entré	1
gris	1
negra	1
piruleta	1
mezclamos	1
cuenta cuentos	1
balonmano	1
durmiendo	1
repartes	1
llegando	1
subía	1
cestita	1
garaje	1
Aurelio	1
olerte	1
abrimos	1
porche	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
comerse	1
terraza	1
rodeo	1
veintiuno	1
sumar	1
ocurre	1
pasan	1
cinta	1
aparato	1
cosquillitas	1
hospital	1
otoño	1
Adriana	1
gatito	1
cerrar	1
tarda	1
encontramos	1
plato	1
American Dragon	1
necesita	1
paseando	1
llamo	1
bollo	1
enfado	1
goma	1
toqué	1
conocimos	1
chocaba	1
asusta	1
duele	1
expresión	1
corneja	1
nevó	1
en seguida	1
arbustos	1
gritaban	1
erizo	1
estudiar	1
enganchar	1
aparcan	1
de momento	1
subíamos	1
tuve	1
cortamos	1
Pilar	1
primera	1
bañaba	1
apaga	1
cambiado	1
jueguen	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
chiquitines	1
Balto	1
pinchan	1
edad	1
encendió	1
bañó	1
cenar	1
vistiendo	1
mirando	1
canarios	1
pasteles	1
Valdés	1
churrería	1
morir	1
Sancho Panza	1
pila	1
costado	1
invento	1
Pachán	1
subiéndose	1
cumplió	1
ele	1
dejar	1
manga	1
existe	1
tendrá	1
oímos	1
hermanos	1
piel	1
abraza	1
jota	1
rascacielos	1
Vanesa	1
estuvimos	1
treinta	1
foto	1
guardamos	1
bolsa	1
trabajando	1
llevando	1
quedaba	1
color	1
cumples	1
diecinueve	1
osito	1
oírte	1
veintitrés de febrero	1
alcanza	1
bajando	1
guardada	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
cerebro	1
Don Quijote	1
preguntado	1
apagaron	1
pillarnos	1
mover	1
desayuno	1
esponja	1
tenido	1
vieron	1
por la tarde	1
disfrazarme	1
bajaba	1
ardillas	1
Dulce María	1
guay	1
es más	1
llena	1
hipopótamo	1
difícil	1
alcachofa	1
rica	1
columpiar	1
estuche	1
siesta	1
violín	1
viaje	1
Juli	1
tirábamos	1
gitano	1
desayunar	1
perenne	1
iban	1
quitársela	1
hambre	1
cohete	1
entró	1
mil	1
tomate	1
grises	1
en medio	1
metí	1
puesta	1
hueso	1
Claus	1
tren	1
tragó	1
entrado	1
salón	1
Homer	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
querrán	1
pulgar	1
llaves	1
huelen	1
puzles	1
carretilla	1
parte	1
ti	1
cocacola	1
botón	1
pastelitos	1
a una	1
crecer	1
acabamos	1
mocos	1
correcaminos	1
quedó	1
tirachinas	1
u	1
médico	1
entretuvo	1
redondeles	1
alas	1
disfrazarse	1
chache	1
entiendo	1
Nayala	1
única	1
enseñé	1
Claudia	1
peiné	1
envenenada	1
corporal	1
manchas	1
notas	1
escondieron	1
aposta	1
apagaban	1
derecha	1
felicidades	1
adelante	1
un rato	1
pilla	1
desapareció	1
Sara	1
Faustino	1
lago	1
puzle	1
entro	1
Faustina	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
guarde	1
dedo	1
disfrazar	1
pirulí	1
levante	1
anular	1
levántate	1
montarlos	1
hermanas	1
despistado	1
Giuly	1
gritó	1
metiendo	1
subiendo	1
dejó	1
superfuerte	1
pues nada	1
campos	1
quema	1
loca	1
regalo	1
saltarina	1
paro	1
muñecos	1
planchar	1
descanse	1
montarlo	1
columpiarse	1
monedas	1
como si	1
soldado	1
traía	1
pares	1
sillitas	1
monto	1
sentamos	1
amiga	1
castigan	1
misma	1
chicha	1
Lucas	1
seguimos	1
corre	1
ganan	1
tripa	1
maletín	1
clavó	1
Félix	1
Fabián	1
nariz	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
pronto	1
antes de	1
cazar	1
vaca	1
patinete	1
niñitos	1
cortó	1
cumplir	1
reloj	1
nadaban	1
acaban	1
rastrillo	1
estantería	1
meñique	1
irme	1
haga	1
pelean	1
lloró	1

Grupo de edad: 3, 4 y 5	
Forma	**Frecuencia**
fregaron	1
vosotros	1
petardo	1
cortos	1
clases	1
espejo	1
últimos	1
abren	1
tiburón	1
raja	1
pasamos	1
pregunto	1
llegarás	1
falda	1
pagarte	1
coja	1
vampiro	1
escapó	1

b. Frecuencia de categorías léxicas: N, V y ADJ

Frecuencia de N	
Lema	**Frecuencia**
casa	83
niño	81
madre	61
padre	59
mamá	44
cosa	43
vez	43
día	40
abuelo	39
hermano	38
lobo	36
perro	32
primo	28
coche	26
cuento	24
cama	24
chico	24
amigo	22
sol	22
juguete	21
papá	21
patio	20
agua	19
elefante	19
fútbol	19
martes	19

Frecuencia de N	
Lema	**Frecuencia**
año	18
euro	18
sábado	17
pis	16
domingo	16
jueves	15
lunes	15
señorita	15
gato	15
pelota	14
cole	14
tío	14
tele	14
leche	13
galleta	13
clase	13
número	13
cabeza	13
piedra	12
e	12
animal	12
céntimo	12
vaso	12
castillo	12
abuelito	12
ratón	11

Frecuencia de N	
Lema	**Frecuencia**
bebé	11
nueve	11
miércoles	11
cerdito	11
payaso	11
a	11
tarta	11
película	10
perra	10
tortuga	10
pala	10
campo	10
serpiente	10
ojo	10
piscina	10
jaula	10
dos	10
dibujo	10
viernes	9
cumpleaños	9
silla	9
pelo	9
cara	9
colacao	9
camión	9
mano	9
caballo	9
ordenador	9
casita	8
semáforo	8
juego	8
comida	8
rey	8
varicela	8
pájaro	8
pupa	8
pato	8
guerra	8
febrero	8
león	8
ropa	8
verano	8
mama	7
noche	7
rana	7
hijo	7
ocho	7
huevo	7
cuello	7
piano	7

Frecuencia de N	
Lema	**Frecuencia**
oveja	7
señorito	7
barco	7
pierna	7
gol	7
zapato	7
monstruo	7
pera	7
café	7
pie	6
diez	6
señor	6
hoja	6
cumple	6
montón	6
colegio	6
fuego	6
ce	6
escalera	6
pantalón	6
tacón	6
disco	6
cuchara	6
sitio	6
piso	6
dragón	6
moto	6
pollo	6
taza	6
gimnasio	6
culo	6
pille pille	6
río	6
calle	6
rinoceronte	6
conejo	6
uno	6
traje	6
tobogán	5
tenedor	5
boca	5
mariposa	5
oso	5
hucha	5
trabajo	5
regalo	5
camiseta	5
cartera	5
palo	5
carnaval	5

Frecuencia de N	
Lema	**Frecuencia**
zapatilla	5
gallina	5
partido	5
fruta	5
hora	5
contenedor	5
sólido	5
tambor	5
fantasma	5
parque	5
ficha	5
chuche	5
brazo	5
corazón	5
escondite	5
portero	5
morro	5
espada	5
balón	5
habitación	5
colorín	5
cabrito	5
pijama	5
ese	5
líquido	5
puerta	5
tigre	5
lápiz	4
pata	4
cuna	4
bosque	4
caramelo	4
mayor	4
careta	4
manzana	4
tiempo	4
yayo	4
collar	4
pico	4
gafa	4
rumano	4
vuelta	4
peine	4
baño	4
flor	4
puntito	4
daño	4
chupete	4
pueblo	4
be	4

Frecuencia de N	
Lema	**Frecuencia**
bolsillo	4
pila	4
rosa	4
diciembre	4
doce	4
once	4
tripa	4
canción	4
invierno	4
casco	4
ramo	4
litera	4
llave	4
rueda	4
grano	4
enero	4
suelo	4
pollito	4
sangre	4
conejito	3
cajón	3
viento	3
fecha	3
ventana	3
plus	3
cometa	3
profesor	3
patito	3
camino	3
cuadra	3
mesa	3
salto	3
diario	3
cristal	3
cerdo	3
arena	3
gusano	3
duende	3
gitano	3
búho	3
dinero	3
enamorado	3
periódico	3
hache	3
punto	3
diente	3
cinco	3
chocolate	3
papel	3
tres	3

Frecuencia de N	
Lema	**Frecuencia**
abeja	3
base	3
madera	3
flecha	3
cuidado	3
malo	3
bombón	3
acera	3
lila	3
sillón	3
estuche	3
magia	3
muñeco	3
bota	3
cereal	3
libro	3
pastel	3
humo	3
biberón	3
foca	3
color	3
columpio	3
nombre	3
bigote	3
nave	3
cocina	3
pescado	3
sopa	3
árbol	3
carretera	3
caso	3
pan	3
garbanzo	3
sofá	3
cuchillo	3
brasero	3
chimenea	3
letra	3
pulsera	3
palomo	3
tijera	3
basura	3
caja	3
ducha	3
caballito	3
maestra	2
ge	2
soldado	2
índice	2
charco	2

Frecuencia de N	
Lema	**Frecuencia**
pintalabios	2
pingüino	2
pirulo	2
doctor	2
nido	2
espalda	2
paseo	2
sueño	2
mayo	2
guitarra	2
patada	2
castaño	2
marzo	2
logopeda	2
gente	2
tela	2
cesta	2
manguito	2
calendario	2
nube	2
amiguito	2
lengua	2
ballena	2
araña	2
playa	2
aro	2
botón	2
babi	2
macarrón	2
globo	2
perrito	2
sierra	2
coleta	2
cachorro	2
tarjeta	2
pantera	2
pino	2
boli	2
amo	2
canasta	2
hermanito	2
añito	2
armario	2
quince	2
susto	2
punta	2
barba	2
colchoneta	2
cola	2
cena	2

Frecuencia de N	
Lema	**Frecuencia**
supermercado	2
baloncesto	2
siete	2
frente	2
escuela	2
albañil	2
o	2
teléfono	2
cartel	2
ala	2
castaña	2
mando	2
carro	2
luna	2
plátano	2
cajita	2
colonia	2
aire	2
cera	2
hamster	2
pepino	2
liebre	2
octubre	2
gorro	2
obra	2
noviembre	2
radio	2
perdón	2
maestro	2
chicle	2
trozo	2
plancha	2
septiembre	2
peluca	2
estufa	2
carrito	2
bola	2
abeto	2
caballero	2
ardilla	2
mujer	2
mechero	2
chupachús	2
pantano	2
seis	2
calcetín	2
empresa	2
truco	2
escape	2
pavo	2

Frecuencia de N	
Lema	**Frecuencia**
policía	2
i	2
hueco	2
papa	2
arroz	2
portería	2
dedo	2
agosto	2
dos mil seis	2
horario	2
burro	2
grúa	2
golpe	2
cepa	2
TEA	2
gomina	2
abril	2
bolsa	2
móvil	2
oreja	2
familia	2
puzle	2
veintiocho	2
gorra	2
baranda	2
muñeca	2
peluche	2
julio	2
reina	2
mosca	2
zumo	2
cero	2
recreo	2
furgoneta	2
enanito	2
voltereta	2
barriga	2
equipo	2
copa	2
carpa	1
queso	1
abrigo	1
beso	1
vaca	1
tazón	1
hipopótamo	1
plato	1
cazador	1
selva	1
perdiz	1

Frecuencia de N	
Lema	**Frecuencia**
mochila	1
ele	1
miel	1
manchego	1
etiqueta	1
cuadro	1
i griega	1
pajarito	1
trece	1
labio	1
cresta	1
rato	1
gallo	1
fuente	1
masaje	1
nieve	1
fresa	1
pedal	1
corneja	1
canario	1
desayuno	1
anillo	1
bicicleta	1
bolso	1
sombra	1
lado	1
angina	1
buhardilla	1
gatito	1
restaurante	1
anular	1
bata	1
terraza	1
tirachinas	1
navidad	1
dieciocho	1
farmacia	1
extraterrestre	1
moneda	1
agujero	1
parte	1
cohete	1
avispa	1
sombrero	1
salón	1
herida	1
rascacielos	1
magdalena	1
noticia	1
médico	1

Frecuencia de N	
Lema	**Frecuencia**
superhéroe	1
bici	1
osito	1
antena	1
corbata	1
barbilla	1
tomate	1
vídeo	1
profe	1
panda	1
tren	1
piruleta	1
patata	1
cuarto	1
chaqueta	1
dóberman	1
vacación	1
tiburón	1
rotulador	1
cochinilla	1
paz	1
diablo	1
cosquillas	1
baile	1
cañón	1
idioma	1
jota	1
pincho	1
círculo	1
tractor	1
cocacola	1
purpurina	1
horquilla	1
manga	1
puente	1
pirulí	1
columna	1
guante	1
cáscara	1
rastro	1
helicóptero	1
hembra	1
cuerpo	1
junio	1
jardín	1
guardería	1
cuerda	1
pulgar	1
cinta	1
espejo	1

Frecuencia de N	
Lema	**Frecuencia**
bollo	1
quema	1
lavabo	1
erizo	1
torre	1
petardo	1
deporte	1
pasta	1
lechuza	1
subibaja	1
violín	1
zanahoria	1
fin de semana	1
entrenador	1
lista	1
cielo	1
patinete	1
polvorón	1
canguro	1
chalé	1
yegua	1
raja	1
siesta	1
block	1
carretilla	1
máquina	1
viaje	1
derecha	1
vía	1
china	1
arco	1
nariz	1
nuez	1
hambre	1
alcachofa	1
teletexto	1
disfraz	1
sapo	1
churrería	1
delfín	1
barra	1
miedo	1
trampa	1
caza	1
caperuza	1
reloj	1
compra	1
moco	1
ruido	1
carpeta	1

Frecuencia de N	
Lema	**Frecuencia**
pulga	1
té	1
violonchelo	1
fideo	1
u	1
misa	1
alfombra	1
correcaminos	1
balonmano	1
póster	1
lana	1
papi	1
tiro	1
pingoreta	1
chicha	1
falda	1
zancadilla	1
rectángulo	1
estantería	1
expresión	1
borrador	1
banda	1
tata	1
cuenta	1
control	1
pasillo	1
compañero	1
edad	1
pirata	1
rosco	1
pez	1
arbusto	1
muchacho	1
colorete	1
izquierda	1
pajita	1
rama	1
medio	1
lío	1
televisión	1
esponja	1
jefe	1
lago	1
piel	1
borra	1
rosquillo	1
nana	1
cerebro	1
miau	1
cacahuete	1

Frecuencia de N	
Lema	**Frecuencia**
leotardo	1
hospital	1
par	1
mancha	1
mago	1
maletín	1
fiesta	1
voz	1
colgante	1
sal	1
crema	1
veinticuatro	1
caca	1
mayúscula	1
brujo	1
bocadillo	1
cebra	1
olivo	1
toallero	1
hueso	1
vocal	1
subida	1
granja	1
debe	1
música	1
empate	1
foto	1
señora	1
gana	1
mentira	1
meñique	1
montaña	1
naranja	1
porche	1
cartón	1
braguita	1
otoño	1
apellido	1
pinturín	1
frío	1
erre	1
goma	1

Frecuencia de N	
Lema	**Frecuencia**
ciprés	1
tontería	1
cuatro	1
cocodrilo	1
responsable	1
limón	1
camionero	1
aparato	1
redondel	1
falta	1
cuenta cuentos	1
polvo	1
pilla besos	1
dios	1
efe	1
catorce	1
demonio	1
garaje	1
pintura	1
jarabe	1
permiso	1
ruiseñor	1
inglés	1
merienda	1
enganchado	1
vampiro	1
nota	1
veintiséis	1
granito	1
miga	1
gata	1
chache	1
rastrillo	1

Frecuencia de V	
Lema	**Frecuencia**
ser	325
tener	254
estar	161
ir	136
saber	133
hacer	116
ver	85
haber	81
jugar	75
decir	68
poner	64
dar	47
venir	46
llamar	45
caer	44
comer	43
comprar	37
salir	36
pasar	34
coger	33
meter	32
echar	29
llevar	28
tirar	27
dejar	24
regalar	24
gustar	23
subir	23
correr	21
leer	20
trabajar	20
quedar	19
querer	19
parecer	19
pintar	19
mirar	18
romper	18
traer	18
hablar	18
tocar	18
picar	18
dormir	17
quitar	17
matar	13
perder	13
acordar	13
cumplir	13
pegar	13
portar	12
llegar	12

Frecuencia de V	
Lema	**Frecuencia**
morir	11
montar	11
contar	11
levantar	11
llorar	10
vivir	10
acabar	10
bajar	10
creer	10
parar	10
sacar	10
ganar	9
encontrar	9
abrir	9
sentar	9
lavar	9
poder	9
morder	9
pinchar	8
vestir	8
acostar	8
cantar	8
guardar	8
convertir	7
saltar	7
esconder	7
pisar	7
beber	7
volar	7
enseñar	7
cortar	7
escribir	7
soñar	7
faltar	7
oír	7
encender	7
entrar	6
mover	6
quemar	6
bañar	6
cerrar	6
coser	5
despertar	5
señalar	5
olvidar	5
andar	5
celebrar	5
buscar	5
rapar	5
disfrazar	4

Frecuencia de V	
Lema	**Frecuencia**
pagar	4
recoger	4
desayunar	4
tomar	4
empezar	4
ayudar	4
peinar	4
cazar	4
pedir	4
casar	4
pillar	4
inventar	4
callar	4
enrollar	3
apagar	3
esperar	3
estirar	3
aparcar	3
botar	3
luchar	3
cambiar	3
conocer	3
columpiar	3
nadar	3
seguir	3
entender	3
gatear	3
manchar	3
hundir	3
juntar	3
escapar	3
encantar	3
colocar	3
limpiar	3
tardar	3
guiñar	3
castigar	3
tiritar	3
merendar	3
terminar	3
acercar	2
mandar	2
escuchar	2
cenar	2
rular	2
felicitar	2
pasear	2
valer	2
aprender	2
enganchar	2

Frecuencia de V	
Lema	**Frecuencia**
chillar	2
colorear	2
enterrar	2
atacar	2
gritar	2
rascar	2
preguntar	2
chupar	2
sujetar	2
chocar	2
asustar	2
caber	2
tragar	2
saludar	2
repasar	2
alcanzar	2
cansar	2
soplar	2
salvar	2
bailar	2
planchar	2
disparar	2
conducir	2
fijar	2
oler	2
empujar	2
enfadar	2
perdonar	1
enterar	1
reír	1
arar	1
explotar	1
necesitar	1
recortar	1
secar	1
entrenar	1
espachurrar	1
crecer	1
doler	1
enfriar	1
tostar	1
zampar	1
hurgar	1
resfriar	1
adoptar	1
ocurrir	1
divertir	1
engañar	1
afinar	1
preparar	1

Frecuencia de V	
Lema	**Frecuencia**
rodear	1
abrochar	1
mojar	1
funcionar	1
mentir	1
abrigar	1
repartir	1
masticar	1
mezclar	1
probar	1
fregar	1
abalanzar	1
nacer	1
atrancar	1
clavar	1
entretener	1
señorear	1
borrar	1
existir	1
liar	1
sonar	1
utilizar	1
grapar	1
descansar	1
sumar	1
regañar	1
marcar	1
costar	1
pelear	1
usar	1
estudiar	1
podar	1
tumbar	1
trepar	1

Lema	**Frecuencia**
asfixiar	1
invitar	1
arrancar	1
lanzar	1
volver	1
desaparecer	1
colar	1
agarrar	1
arropar	1
atropellar	1
abrazar	1
intentar	1
envenenar	1
dibujar	1
escupir	1
aburrir	1
apuntar	1
vender	1
perecer	1
calzar	1
nevar	1
huir	1
vomitar	1
soltar	1
descoser	1

Frecuencia de ADJ	
Lema	**Frecuencia**
grande	60
pequeño	24
rojo	20
malo	20
presente	18
azul	17
bueno	12
verde	11
solo	10
chiquitín	10
roto	9
negro	9
de leche	8
alto	8
amarillo	6
gordo	6
feo	6
chico	6
mayor	6
marrón	6
chino	6
blanco	6
rosa	5
vivo	5
cojo	5
corto	5
de verdad	5
entero	4
colorado	4
pelado	4
naranja	4
digital	4
lila	4
fuerte	3
mismo	3
nuevo	3
pasado	3
abierto	3
calentito	3
bonito	3
guapo	3
largo	3
frío	3
suave	3
mojado	3
de hierro	2
fino	2
peligroso	2
duro	2
parado	2

Frecuencia de ADJ	
Lema	**Frecuencia**
gris	2
junto	2
bajo	2
igual	2
morado	2
deportivo	2
de miedo	2
oscuro	2
último	2
maestro	2
cansino	2
atento	2
quieto	2
enfermo	2
tonto	2
secreto	2
frito	1
flaco	1
joven	1
superfuerte	1
responsable	1
loco	1
cazador	1
triste	1
sucio	1
guay	1
octavo	1
corporal	1
cansado	1
despistado	1
pelota	1
sentado	1
viejo	1
callado	1
coloreado	1
flojo	1
único	1
malito	1
limpio	1
rico	1
estirado	1
africano	1
saltarín	1
estrecho	1
perenne	1
rectangular	1
fácil	1
difícil	1
disfrazado	1
suelto	1

Frecuencia de ADJ	
Lema	**Frecuencia**
lleno	1
americano	1
raro	1
separado	1
doble	1
san	1

Lema	Frecuencia
casado	1
mentiroso	1
temprano	1
vacío	1
preferido	1
tranquilo	1
rápido	1

2. Grupo de edad: 3 años

a. Frecuencia de formas

Grupo de edad: 3	
Forma	**Frecuencia**
y	316
no	261
sí	203
un	146
el	141
yo	138
la	133
me	110
a	110
se	101
en	90
una	78
de	74
que	73
los	64
lo	64
es	54
con	50
mi	46
ya	43
aquí	41
pero	41
tengo	40
está	37
ha	37
qué	36
he	35
porque	35
por	33
Guadalupe	32
esto	31
así	31
le	31
sé	28
este	28
también	27

Grupo de edad: 3	
Forma	**Frecuencia**
mira	27
muy	27
por qué	26
otro	25
muchos	24
mamá	22
hay	22
ahí	21
eso	21
quiero	19
si	19
grande	19
presente	18
para	18
cuando	17
casa	17
más	16
otra	16
tiene	15
ese	15
voy	15
las	15
pis	14
elefante	14
como	14
pues	14
dónde	14
al	14
ni	14
uno	13
era	13
mucho	13
ahora	13
estaba	12
va	12
había	12

Grupo de edad: 3	
Forma	**Frecuencia**
perro	12
rojo	12
azul	11
pone	11
luego	11
hacer	11
oye	11
juguetes	11
ella	11
vamos	11
sólo	10
Javier	10
cosa	10
tú	10
tres	10
día	10
tenía	10
mis	10
jugar	10
lobo	9
coche	9
papá	9
hablar	9
gato	9
primero	9
visto	9
soy	9
niña	9
niño	9
Rodrigo	8
semáforo	8
han	8
dos	8
hemos	8
Abel	8
hasta	8
corriendo	8
están	8
nada	8
son	8
es que	8
pupa	7
cerdito	7
vale	7
tenemos	7
pelota	7
hago	7
deprisa	7
del	6
cuchara	6

Grupo de edad: 3	
Forma	**Frecuencia**
chiquitín	6
Ana	6
abuela	6
Manuel	6
rompe	6
números	6
te	6
él	6
hecho	6
caballo	6
Sergio	6
pollo	6
hace	6
iba	6
pala	6
cabeza	5
pasa	5
todo	5
tenedor	5
ay	5
picar	5
Selene	5
Saranya	5
Cristina	5
Evelio José	5
rosa	5
martes	5
varicela	5
juego	5
ala	5
caído	5
sola	5
parque	5
piscina	5
verde	5
ocho	5
cómo	5
Iván	5
niños	5
bien	5
ningún	5
Natalia	5
nos	5
pequeña	5
agua	5
unas	5
sido	5
Antonio José	5
serpiente	5
ordenador	5

Grupo de edad: 3	
Forma	**Frecuencia**
roto	5
nueve	5
pille pille	5
llama	5
donde	5
nunca	4
casita	4
Rosa Gema	4
mala	4
seis	4
e	4
corazón	4
reyes	4
van	4
tacones	4
Bruno	4
muchas	4
siempre	4
baño	4
fruta	4
Aitana	4
peine	4
cantado	4
gallina	4
venga	4
mariposa	4
de pie	4
quién	4
pelo	4
quería	4
Emilio	4
trajes	4
diez	4
venido	4
juguete	4
oso	4
mí	4
puntitos	4
amigo	4
gusta	4
rinoceronte	4
subir	4
Elena	4
regalado	4
bajé	4
mataba	4
leer	4
pelado	4
cuatro	4
todos	4

Grupo de edad: 3	
Forma	**Frecuencia**
veo	4
habitación	3
señalar	3
calentita	3
tarta	3
bota	3
salta	3
hacía	3
claro	3
pica	3
tiro	3
payaso	3
monta	3
cerdo	3
garbanzos	3
viene	3
mayor	3
tienes	3
tampoco	3
oveja	3
conejito	3
magia	3
siéntate	3
ratón	3
toda	3
subo	3
estaban	3
boca	3
comprar	3
salía	3
pilas	3
Antonio	3
tiritas	3
pego	3
puedo	3
pescado	3
cole	3
pisado	3
vi	3
llorando	3
cinco	3
bebé	3
siete	3
be	3
mueve	3
yayo	3
señor	3
metió	3
lunes	3
acera	3

Grupo de edad: 3	
Forma	**Frecuencia**
tía	3
abuelito	3
Carmen	3
gimnasio	3
tigre	3
huevos	3
jugado	3
puso	3
puesto	3
cuentos	3
ver	3
guiña	3
un poco	3
cuello	3
unos	3
compró	3
campo	3
comió	3
veces	3
lobos	3
Pepón	3
dibujos	3
castillo	3
comprado	3
ce	3
viento	3
algunas	3
uy	3
ah	3
Julia	3
vez	3
cristal	3
míos	3
tu	3
patos	3
animales	3
culo	3
hice	3
cocina	3
fuerte	3
tan	3
o	3
barco	3
parar	3
pan	3
a ver	3
serpientes	3
contar	3
bueno	3
fútbol	3

Grupo de edad: 3	
Forma	**Frecuencia**
coches	3
traído	3
cabe	2
coge	2
perdón	2
King Kong	2
pinchó	2
decir	2
vivo	2
repasamos	2
mío	2
llamaba	2
pintar	2
doy	2
esta	2
alguien	2
cerditos	2
saco	2
encima	2
cantaba	2
parece	2
mucha	2
seguro que	2
Barbie	2
bajar	2
pájaro	2
en coche	2
copa	2
lengua	2
comía	2
pueden	2
truco	2
hazme	2
pato	2
aire	2
arroz	2
hoy	2
oh	2
calcetines	2
esa	2
león	2
búho	2
puede	2
pueblo	2
cierra	2
abuelita	2
plátano	2
veía	2
será	2
pillado	2

Grupo de edad: 3	
Forma	**Frecuencia**
alta	2
deja	2
sopa	2
dicho	2
enrolla	2
cierran	2
coger	2
grúa	2
fue	2
daba	2
llevo	2
zapatos	2
cuchillos	2
puedes	2
bebés	2
salido	2
señalarlo	2
algo	2
ducha	2
portado	2
mami	2
conejo	2
leído	2
coloco	2
parada	2
montón	2
suelo	2
fui	2
malo	2
estar	2
llora	2
esperando	2
Marta	2
sol	2
atenta	2
diecisiete	2
Rubén	2
acabado	2
muevo	2
Villamuriel	2
ninguno	2
lleva	2
silla	2
taza	2
tigres	2
cajitas	2
cuál	2
Tomelloso	2
come	2
da	2

Grupo de edad: 3	
Forma	**Frecuencia**
amigos	2
queda	2
buena	2
echa	2
quince	2
pedido	2
espadas	2
quedan	2
les	2
leerlo	2
viéndolos	2
collar	2
escape	2
puerta	2
echo	2
bosque	2
sillita	2
enferma	2
quita	2
ido	2
galleta	2
pasar	2
jugando	2
caramelo	2
traje	2
pega	2
convirtieron	2
granos	2
cuento	2
arriba	2
gustan	2
pantera	2
déjalo	2
dieciocho	2
naranja	2
barba	2
podemos	2
Superman	2
acabó	2
larguita	2
vale ?	2
globos	2
cayó	2
comido	2
bolsillo	2
poquito	2
mama	2
tira	2
café	2
sigue	2

Grupo de edad: 3	
Forma	**Frecuencia**
piso	2
Vane	2
trae	2
pinchar	2
mordía	2
bajo	2
macarrones	2
índice	2
meter	2
ayer	2
mayores	2
mojado	2
dieciséis	2
venir	2
hueco	2
montar	2
dijo	2
cara	2
comí	2
río	2
sin	2
pirulo	2
baja	2
subido	2
rula	2
su	2
burro	2
Inma	2
mía	2
caballito	2
supermercado	2
pequeñito	1
empieza	1
niñas	1
pájaros	1
sucio	1
pollito	1
recreo	1
nadan	1
lana	1
Diana	1
papel	1
botones	1
música	1
enseñado	1
cama	1
ir	1
i griega	1
buenos días	1
jo	1

Grupo de edad: 3	
Forma	**Frecuencia**
madre mía	1
lías	1
cambiarla	1
barbilla	1
peina	1
pinté	1
sabe	1
pollitos	1
enciende	1
ventanas	1
conducir	1
trozo	1
enseña	1
cuidado	1
árbol	1
salen	1
Chuki	1
Tomasa	1
cometa	1
granitos	1
regalos	1
catorce	1
trece	1
nana	1
caja	1
verlo	1
llevaba	1
abrigo	1
policía	1
colocamos	1
hablando	1
juegos	1
montaña	1
bebido	1
acuerdas	1
pirata	1
padre	1
llevaron	1
saliera	1
recoger	1
mosca	1
bolso	1
convertía	1
cielo	1
después	1
nombres	1
ardilla	1
patatas	1
vino	1
enfría	1

Grupo de edad: 3	
Forma	**Frecuencia**
metérmelo	1
sabes	1
comida	1
espera	1
negros	1
palo	1
comiste	1
acuerda	1
muñeco	1
dibujo	1
rueda	1
tiraban	1
gustaba	1
creo	1
utilizaste	1
la que	1
Loreya	1
pintura	1
salgo	1
pierda	1
acuerdo	1
pijama	1
cansado	1
mando	1
colorín colorado	1
enrollas	1
sabían	1
calladito	1
tobogán	1
selva	1
meto	1
papi	1
mentiras	1
malas	1
jugaba	1
anillo	1
matarlo	1
estoy	1
contigo	1
fría	1
chocamos	1
arena	1
días	1
casi	1
leche	1
portaba	1
pingüino	1
pinchado	1
arrancado	1
picado	1

Grupo de edad: 3	
Forma	**Frecuencia**
pinta	1
jardín	1
punto	1
oscuro	1
fresa	1
nadar	1
huevo	1
manos	1
huía	1
cera	1
roja	1
igual	1
patas	1
cuchillo	1
palas	1
bata	1
bici	1
guardarlos	1
echó	1
llego	1
limpitas	1
cantar	1
agujero	1
flechas	1
delante	1
señoreada	1
Cango	1
toca	1
acercar	1
queso	1
blanco	1
ovejitas	1
tortuga	1
a lo mejor	1
escondidas	1
perdonas	1
ayuda	1
fuera	1
enganchaba	1
disfrazado	1
trabajó	1
traga	1
mató	1
echamos	1
mordiéndola	1
cosas	1
ballena	1
funciona	1
fritas	1
chalé	1

Grupo de edad: 3	
Forma	**Frecuencia**
amiguitos	1
echar	1
mal	1
y todo	1
compra	1
mochila	1
vestir	1
lavabo	1
ellas	1
quedo	1
anda	1
poner	1
toma	1
cuánto	1
jaula	1
charco	1
animalitos	1
feo	1
di	1
tío	1
radio	1
dejo	1
dan	1
estrecha	1
paramos	1
subí	1
braguitas	1
rana	1
entera	1
cogieron	1
colorines	1
eran	1
perrito	1
calle	1
sal	1
pajita	1
jugaban	1
encontraban	1
conozco	1
canción	1
magos	1
enseño	1
mamás	1
corría	1
vacaciones	1
contarlo	1
esos	1
leemos	1
sientas	1
delfín	1

Grupo de edad: 3	
Forma	**Frecuencia**
fideos	1
guapo	1
guapa	1
toallero	1
buscado	1
el que	1
vemos	1
frente	1
faltan	1
teléfono	1
polvorones	1
Pedro	1
fecha	1
película	1
paseo	1
atrancó	1
pingüinos	1
caramelos	1
mata	1
nosotros	1
quedar	1
pez	1
quedando	1
careta	1
andando	1
ponía	1
murió	1
antena	1
estos	1
caca	1
cuántos	1
bebía	1
nadie	1
pongo	1
sitio	1
compro	1
clase	1
aros	1
tartas	1
chilla	1
salir	1
dar	1
iguales	1
limón	1
recojo	1
círculos	1
gallo	1
recoge	1
quitado	1
mordió	1

Grupo de edad: 3	
Forma	**Frecuencia**
ves	1
guardar	1
pasillo	1
así que	1
adiós	1
africano	1
echaba	1
venía	1
volaba	1
agarrado	1
zapato	1
vaso	1
sube	1
empujado	1
gusanito	1
cuerda	1
chillando	1
limpia	1
manchan	1
borrador	1
vio	1
Felisa	1
cae	1
abuelo	1
carrito	1
ven	1
hizo	1
manguitos	1
morado	1
cuerpo	1
miedo	1
fuego	1
siquiera	1
escondido	1
jugamos	1
diciendo	1
diablo	1
castigó	1
pasado	1
puntos	1
comemos	1
frutas	1
lloré	1
conmigo	1
merienda	1
fino	1
cuna	1
por favor	1
secado	1
vimos	1

Grupo de edad: 3	
Forma	**Frecuencia**
echemos	1
traer	1
amarillo	1
deportivas	1
moto	1
metes	1
yaya	1
dentro	1
pie	1
sábado	1
pequeño	1
monstruos	1
muñecas	1
papa	1
juega	1
Esther	1
los que	1
navidad	1
sopita	1
tijeras	1
ser	1
pingoretas	1
letras	1
Furi	1
columpios	1
traen	1
sale	1
zapatillas	1
cacahuete	1
castiga	1
poquitos	1
mojada	1
migas	1
camión	1
caía	1
blanca	1
pupas	1
pinchándole	1
al lado de	1
Adrián	1
despacio	1
número	1
carnaval	1
alto	1
mastica	1
colorear	1
repartes	1
llegando	1
mar	1
cuándo	1

Grupo de edad: 3	
Forma	**Frecuencia**
bigote	1
encontramos	1
letra	1
ventana	1
marrón	1
fichas	1
digo	1
Papá Noel	1
conocimos	1
chocaba	1
otros	1
señores	1
quiere	1
espalda	1
erizo	1
enganchar	1
foca	1
películas	1
cortamos	1
apaga	1
Aila	1
chiquitines	1
tiraba	1
ea	1
pulsera	1
contado	1
sabía	1
vea	1
caer	1
color	1
Don Quijote	1
preguntado	1
pillarnos	1
mover	1
vieron	1
disfrazarme	1
bajaba	1
ardillas	1
guay	1
cansino	1
llena	1
hipopótamo	1
volando	1
iban	1
hambre	1
falta	1
once	1
patio	1
punta	1
mancha	1

Grupo de edad: 3	
Forma	**Frecuencia**
negro	1
tomate	1
hueso	1
allí	1
salón	1
metían	1
pulgar	1
llaves	1
jugué	1
puzles	1
doce	1
comer	1
Miguel	1
acabamos	1
cada	1
pera	1
chicles	1
médico	1
alas	1
Nayala	1
enseñé	1
Claudia	1
aposta	1
gorra	1
pilla	1
lápiz	1
pasó	1
tele	1
lago	1
entro	1
puzle	1
disfrazar	1
todavía	1
piernas	1
anular	1
montarlos	1
despistado	1
subiendo	1
muñecos	1
montarlo	1
tiró	1
columpiarse	1
papás	1
traía	1
sillitas	1
barriga	1
monto	1
vaya	1
chicha	1
seguimos	1

Grupo de edad: 3	
Forma	**Frecuencia**
maletín	1
abajo	1
madre	1
escribir	1
peluca	1
patinete	1
niñitos	1
vaca	1
dormí	1
reloj	1
nadaban	1
acaban	1

Forma	Frecuencia
meñique	1
cogía	1
lloró	1
piano	1
lee	1
espejo	1
señorita	1
abren	1
rey	1
tiburón	1
hoja	1
vivía	1

b. Frecuencia de categorías léxicas: N, V y ADJ

Frecuencia de N	
Lema	**Frecuencia**
niño	25
mamá	25
casa	17
juguete	15
elefante	14
pis	14
perro	12
lobo	12
coche	12
día	11
cosa	11
papá	10
cerdito	9
gato	9
semáforo	8
serpiente	8
pupa	8
número	7
pala	7
abuelo	7
a	7
pelota	6
cuchara	6
amigo	6
pollo	6
caballo	6
vez	6
traje	6
tenedor	5
agua	5
bebé	5
rey	5

Frecuencia de N	
Lema	**Frecuencia**
varicela	5
silla	5
cabeza	5
cuento	5
fruta	5
parque	5
pato	5
piscina	5
martes	5
pille pille	5
ordenador	5
abuelito	5
tigre	5
casita	4
mariposa	4
oso	4
señor	4
juego	4
huevo	4
yayo	4
oveja	4
gallina	4
pelo	4
peine	4
baño	4
puntito	4
tacón	4
tío	4
corazón	4
ese	4
rinoceronte	4
dibujo	4

Frecuencia de N	
Lema	**Frecuencia**
tarta	4
conejito	3
boca	3
viento	3
caramelo	3
e	3
ratón	3
cristal	3
cerdo	3
cuello	3
lunes	3
pájaro	3
animal	3
ce	3
campo	3
be	3
cole	3
acera	3
pila	3
payaso	3
magia	3
barco	3
fútbol	3
castillo	3
gimnasio	3
cocina	3
pescado	3
habitación	3
sopa	3
culo	3
pan	3
garbanzo	3
cuchillo	3
zapato	3
índice	2
bosque	2
pingüino	2
pirulo	2
película	2
ventana	2
mama	2
galleta	2
mayor	2
lengua	2
macarrón	2
collar	2
globo	2
búho	2
montón	2
pantera	2

Frecuencia de N	
Lema	**Frecuencia**
punto	2
barba	2
supermercado	2
pueblo	2
bolsillo	2
piso	2
rosa	2
muñeco	2
plátano	2
cajita	2
aire	2
perdón	2
espada	2
cara	2
taza	2
ardilla	2
calcetín	2
león	2
truco	2
escape	2
hueco	2
arroz	2
burro	2
grúa	2
río	2
letra	2
puzle	2
grano	2
sol	2
conejo	2
suelo	2
café	2
ducha	2
pollito	2
copa	2
puerta	2
caballito	2
tobogán	1
queso	1
lápiz	1
abrigo	1
pata	1
vaca	1
charco	1
leche	1
hipopótamo	1
cuna	1
pie	1
selva	1
mochila	1

Frecuencia de N	
Lema	**Frecuencia**
i griega	1
fecha	1
gallo	1
diez	1
fresa	1
espalda	1
anillo	1
bolso	1
rana	1
cometa	1
anular	1
bata	1
paseo	1
navidad	1
careta	1
agujero	1
manguito	1
salón	1
amiguito	1
médico	1
ballena	1
bici	1
arena	1
antena	1
gusano	1
aro	1
botón	1
barbilla	1
regalo	1
tomate	1
comida	1
hoja	1
piano	1
clase	1
perrito	1
patata	1
tortuga	1
vacación	1
tiburón	1
diablo	1
palo	1
círculo	1
carnaval	1
zapatilla	1
padre	1
señorita	1
fuego	1
punta	1
cuerpo	1
papel	1

Frecuencia de N	
Lema	**Frecuencia**
jardín	1
cuerda	1
pulgar	1
sábado	1
espejo	1
erizo	1
lavabo	1
siete	1
frente	1
flecha	1
cuidado	1
ficha	1
cielo	1
patinete	1
polvorón	1
chalé	1
teléfono	1
ala	1
hambre	1
sitio	1
delfín	1
mando	1
miedo	1
reloj	1
vaso	1
fideo	1
lana	1
papi	1
pingoreta	1
chicha	1
moto	1
cera	1
borrador	1
foca	1
radio	1
color	1
pasillo	1
pirata	1
columpio	1
chicle	1
pez	1
cama	1
trozo	1
jaula	1
pajita	1
nombre	1
peluca	1
bigote	1
carrito	1
lago	1

Frecuencia de N	
Lema	**Frecuencia**
nana	1
cacahuete	1
canción	1
mago	1
colorín	1
pierna	1
árbol	1
maletín	1
policía	1
patio	1
papa	1
sal	1
camión	1
caca	1
toallero	1
hueso	1
pijama	1
llave	1
rueda	1
mano	1
música	1
mentira	1

Lema	Frecuencia
meñique	1
montaña	1
braguita	1
calle	1
dos	1
limón	1
gorra	1
pulsera	1
muñeca	1
madre	1
monstruo	1
tijera	1
pera	1
mosca	1
pintura	1
recreo	1
caja	1
merienda	1
barriga	1
uno	1
granito	1
miga	1
tele	1

Frecuencia de V	
Lema	**Frecuencia**
ser	94
tener	64
estar	50
hacer	39
saber	32
haber	31
ver	30
ir	25
jugar	22
poner	20
comer	17
mirar	13
subir	12
venir	12
comprar	11
caer	10
romper	10
leer	10
hablar	10

Frecuencia de V	
Lema	**Frecuencia**
correr	9
pasar	9
salir	9
decir	9
bajar	9
echar	9
meter	9
picar	9
dar	8
tirar	8
traer	8
montar	8
matar	7
quedar	7
gustar	7
llorar	7
cantar	7
llamar	7
querer	7

Frecuencia de V	
Lema	**Frecuencia**
pinchar	6
mover	6
llevar	6
parar	6
coger	6
señalar	5
acabar	5
dejar	5
contar	5
pegar	5
sentar	4
cerrar	4
enseñar	4
pillar	4
regalar	4
pintar	4
morder	4
enrollar	3
recoger	3
esperar	3
convertir	3
saltar	3
botar	3
pisar	3
nadar	3
portar	3
seguir	3
acordar	3
colocar	3
guiñar	3
tiritar	3
quitar	3
disfrazar	2
rular	2
esconder	2
enganchar	2
conocer	2
vivir	2
chillar	2
chocar	2
caber	2
repasar	2
beber	2
pedir	2
volar	2
encontrar	2
guardar	2
manchar	2
llegar	2

Lema	**Frecuencia**
faltar	2
oír	2
parecer	2
poder	2
castigar	2
sacar	2
acercar	1
perdonar	1
entrar	1
apagar	1
secar	1
enfriar	1
valer	1
tomar	1
cambiar	1
mojar	1
funcionar	1
empezar	1
vestir	1
colorear	1
columpiar	1
repartir	1
preguntar	1
ayudar	1
andar	1
masticar	1
peinar	1
morir	1
creer	1
tragar	1
atrancar	1
perder	1
buscar	1
señorear	1
cansar	1
liar	1
utilizar	1
abrir	1
dormir	1
podar	1
tocar	1
arrancar	1
conducir	1
cortar	1
agarrar	1
escribir	1
limpiar	1
encender	1
empujar	1
trabajar	1
huir	1

Frecuencia de ADJ	
Lema	**Frecuencia**
grande	19
presente	18
rojo	13
azul	11
pequeño	7
malo	7
chiquitín	7
solo	5
verde	5
pelado	4
fuerte	3
rosa	3
mayor	3
calentito	3
bueno	3
alto	3
parado	2
bajo	2
naranja	2
guapo	2
negro	2
largo	2
atento	2
enfermo	2
blanco	2

Lema	Frecuencia
mojado	2
frito	1
entero	1
amarillo	1
roto	1
sucio	1
fino	1
guay	1
vivo	1
despistado	1
pelota	1
callado	1
feo	1
igual	1
morado	1
limpio	1
deportivo	1
africano	1
estrecho	1
marrón	1
oscuro	1
disfrazado	1
cansino	1
frío	1
lleno	1

3. Grupo de edad: 4 años

a. Frecuencia de formas

Grupo de edad: 4	
Forma	**Frecuencia**
y	328
que	203
sí	193
no	177
la	148
el	140
en	127
a	126
un	125
de	124
mi	123
me	103
yo	102
una	100
se	100
es	79

Grupo de edad: 4	
Forma	**Frecuencia**
los	70
pues	68
pero	61
lo	57
para	56
tengo	48
porque	47
con	46
le	38
uno	37
qué	36
también	35
las	34
al	34
cuando	30
tiene	29

Grupo de edad: 4	
Forma	**Frecuencia**
ya	29
madre	29
es que	29
casa	27
más	25
por	25
Balbi	25
mira	24
ha	23
niño	23
padre	22
aquí	22
dos	22
hay	22
del	21
otro	21
muy	21
sé	20
está	19
si	18
cómo	18
quién	17
llama	17
sabes	16
cinco	16
fui	15
a ver	15
su	15
lo que	14
va	13
tres	13
otra	13
o	13
todos	13
he	12
hasta	12
eso	12
son	12
bueno	12
lobo	11
unos	11
grande	11
agua	11
jugar	11
patio	11
niña	11
era	10
claro	10
hermano	10
ahí	10

Grupo de edad: 4	
Forma	**Frecuencia**
cumple	10
tu	10
cuatro	10
dijo	10
e	9
luego	9
ese	9
tú	9
había	9
como	9
grandes	9
tenía	9
te	9
voy	9
mí	9
prima	9
fútbol	9
estaba	8
coche	8
abuela	8
jaula	8
este	8
clase	8
mis	8
niños	8
nada	8
así	8
tele	8
Jorge	7
años	7
payaso	7
mucho	7
fue	7
vi	7
José	7
abuelo	7
lila	7
nos	7
veces	7
castillo	7
ahora	7
algunas	7
gusta	7
piedras	7
perros	6
azul	6
sólo	6
primo	6
hermana	6
ojos	6

Grupo de edad: 4	
Forma	**Frecuencia**
esto	6
tienes	6
trabaja	6
martes	6
hemos	6
tortuga	6
mal	6
dice	6
vale	6
Carmen	6
bien	6
hecho	6
número	6
chica	6
vez	6
algunos	6
donde	6
veo	6
ir	5
pasa	5
pintar	5
domingo	5
pone	5
Mari	5
hacer	5
mamá	5
abuelita	5
llegó	5
cole	5
piscina	5
el que	5
malo	5
trabajar	5
siete	5
lleva	5
lunes	5
pantalones	5
jugamos	5
amigos	5
ver	5
están	5
sábado	5
Aitor	5
ido	5
Miguel Ángel	5
soy	5
Caperucita	5
cuento	5
José Miguel	5
camiones	5

Grupo de edad: 4	
Forma	**Frecuencia**
mama	5
once	5
café	5
pera	5
haciendo	5
nueve	5
ayer	5
mía	5
Mónica	4
cama	4
salen	4
tigre	4
de verdad	4
todo	4
vivo	4
noche	4
después	4
quiero	4
mañana	4
por qué	4
tobogán	4
días	4
podía	4
hoy	4
sea	4
esa	4
allá	4
por eso	4
ratón	4
jueves	4
cosas	4
hijo	4
oye	4
feo	4
monstruo	4
algo	4
gatos	4
cartera	4
habla	4
rana	4
mano	4
rumanos	4
película	4
disco	4
Marta	4
sabes ?	4
da	4
echa	4
dragón	4
trabajo	4

Grupo de edad: 4	
Forma	**Frecuencia**
quince	4
viernes	4
quería	4
cuello	4
pequeño	4
pegan	4
ser	4
visto	4
dibujos	4
cerca	4
unas	4
Marisol	4
Andrés	4
María	4
dónde	4
iba	4
tan	4
has	4
Play	4
Eva	4
biberón	3
pinto	3
Manga	3
esas	3
tarta	3
cumpleaños	3
año	3
decir	3
Azorín	3
seis	3
Talía	3
pones	3
comida	3
parece	3
Donald	3
pica	3
Raúl	3
cogen	3
pájaro	3
van	3
estoy	3
cosa	3
casi	3
pato	3
Héctor	3
gordo	3
manzana	3
Abel	3
Goofy	3
Mickey	3

Grupo de edad: 4	
Forma	**Frecuencia**
animal	3
coger	3
casen	3
tío	3
Simpsons	3
perro	3
montón	3
chico	3
estos	3
sol	3
Riki	3
febrero	3
cuadra	3
porta	3
Minnie	3
vaso	3
Tomelloso	3
queda	3
pasado	3
jugado	3
ella	3
gané	3
amarillo	3
camiseta	3
un poco	3
cogimos	3
los que	3
plancha	3
sale	3
cogió	3
viendo	3
convierte	3
carnaval	3
Papá Noel	3
otros	3
Madrid	3
Ronaldinho	3
amigo	3
ah	3
base	3
pelota	3
ea	3
solo	3
Juanito	3
hago	3
pierna	3
volando	3
haces	3
allí	3
piso	3

Grupo de edad: 4	
Forma	**Frecuencia**
doce	3
comer	3
quitan	3
lápiz	3
todavía	3
piernas	3
palomos	3
tortugas	3
ni	3
lejos	3
contenedores	3
colorado	2
pájaros	2
pollito	2
llaman	2
cajón	2
Jose	2
cabeza	2
cometa	2
llevar	2
catorce	2
trece	2
menos	2
Daniel	2
tirar	2
canasta	2
llevó	2
colchonetas	2
mala	2
esta	2
Mallorca	2
atacaron	2
palo	2
libro	2
dame	2
discos	2
gustaba	2
olvidó	2
la que	2
bañar	2
fuimos	2
dos mil seis	2
veranos	2
Paco	2
hora	2
adentro	2
bonito	2
viene	2
Cristina	2
marrones	2

Grupo de edad: 4	
Forma	**Frecuencia**
frío	2
arena	2
leche	2
merendé	2
poco	2
morro	2
manos	2
acá	2
tampoco	2
gente	2
oveja	2
muere	2
gusano	2
manda	2
será	2
Ana	2
bebo	2
a lo mejor	2
deja	2
a veces	2
varicela	2
otras	2
dicho	2
Dulce	2
Rollo	2
juego	2
anda	2
botas	2
cosieron	2
llevo	2
cepas	2
boli	2
nido	2
calendario	2
caído	2
papá	2
maestros	2
ficha	2
Antonio	2
entera	2
ponerle	2
aparcados	2
logopeda	2
calle	2
Albacete	2
puedo	2
mamás	2
Ángel	2
gafas	2
gato	2

Grupo de edad: 4	
Forma	**Frecuencia**
colorín	2
levantó	2
suelo	2
día	2
empresa	2
compran	2
mete	2
verde	2
bebé	2
cazó	2
juguetes	2
murió	2
luna	2
sitio	2
llave	2
columpio	2
acabado	2
ves	2
veintidós	2
silla	2
taza	2
escrito	2
familia	2
claro que sí	2
él	2
vio	2
pantano	2
sentado	2
móvil	2
cero	2
come	2
hizo	2
de miedo	2
fuego	2
quemé	2
buena	2
huevos	2
regalaron	2
pelo	2
colegio	2
puesto	2
muchos	2
pisos	2
portería	2
podido	2
primero	2
les	2
dentro	2
horario	2
Jorgete	2

Grupo de edad: 4	
Forma	**Frecuencia**
Patricia	2
tijeras	2
estuches	2
comió	2
guitarra	2
ponga	2
tienen	2
Discolandia	2
gitanos	2
jugando	2
contaron	2
blanca	2
Jesús	2
Adrián	2
dobermanes	2
vive	2
Cassano	2
chupete	2
bigote	2
jope	2
playa	2
peligrosas	2
ve	2
Jorge Torres	2
quiere	2
películas	2
foca	2
dura	2
dieciocho	2
naranja	2
mujer	2
obra	2
tarjetas	2
vueltas	2
miércoles	2
animales	2
pintalabios	2
hicimos	2
quieren	2
falta	2
negro	2
corren	2
pala	2
metían	2
Puyol	2
roto	2
abeja	2
tiran	2
mas	2
dando	2

Grupo de edad: 4	
Forma	**Frecuencia**
escuela	2
barco	2
hundían	2
pasó	2
venir	2
escondite	2
furgoneta	2
ellos	2
cara	2
río	2
estufa	2
saltan	2
lee	2
piano	2
pequeñito	1
empieza	1
chino	1
coge	1
nunca	1
habitación	1
Zidane	1
recreo	1
dormirme	1
tambor	1
dejáis	1
cazador	1
papel	1
DVD	1
Balbino	1
leyendo	1
carpeta	1
pintamos	1
leones	1
eché	1
gomina	1
enciende	1
cola	1
cuidadito	1
cesta	1
permiso	1
cuidado	1
regalos	1
chupachuses	1
peguen	1
blancos	1
bota	1
veinte	1
subida	1
cumplo	1
pudimos	1

Grupo de edad: 4	
Forma	**Frecuencia**
enseñó	1
muchacho	1
nombre	1
abrigo	1
policía	1
abril	1
fueron	1
hablando	1
juegos	1
dientes	1
mío	1
llamaba	1
digital	1
llevaron	1
salta	1
plus	1
pis	1
mosca	1
sitios	1
paso	1
regalaba	1
bruja	1
abrió	1
pequeñitas	1
aprender	1
levanté	1
siendo	1
Betis	1
abrochar	1
acuerda	1
corazón	1
Soco	1
Bugs Bunny	1
rascarme	1
puente	1
mucha	1
creo	1
Tintín	1
estas	1
tiro	1
damos	1
espada	1
control	1
acuerdo	1
cansado	1
picar	1
saliendo	1
intentó	1
jarabe	1
pintaba	1

Grupo de edad: 4	
Forma	**Frecuencia**
asfixió	1
echarme	1
malas	1
pata	1
sueltas	1
polvos	1
sus	1
acuesta	1
chiquitín	1
comía	1
descosió	1
Oscar	1
pasearlo	1
escupir	1
han	1
ropa	1
rosa	1
salto	1
pueden	1
Julio	1
quemar	1
suyo	1
veintitrés	1
vieja	1
mayor	1
anginas	1
barcos	1
grano	1
Juan	1
picado	1
Lili	1
Daisy	1
oscuro	1
debes	1
gustaban	1
tenga	1
cera	1
camino	1
igual	1
patas	1
hurga	1
trabajas	1
Popeye	1
crema	1
salga	1
hicieron	1
enanitos	1
puede	1
María José	1
compañeros	1

Grupo de edad: 4	
Forma	**Frecuencia**
misa	1
beber	1
muchas	1
siempre	1
dije	1
sapo	1
Lourdes	1
señora	1
pico	1
toca	1
colonia	1
alcanzaba	1
sentarte	1
habido	1
chupa	1
alta	1
viva	1
rato	1
nube	1
empates	1
abiertos	1
fuera	1
calla	1
riendo	1
canciones	1
asustan	1
Laura	1
abierto	1
bicicleta	1
sofá	1
a gatas	1
cayeron	1
entendemos	1
mató	1
termine	1
trozos	1
estaban	1
ballena	1
amiguitos	1
salió	1
picaba	1
quieres	1
tristes	1
compra	1
arroparme	1
perdieron	1
vestir	1
tata	1
lavarme	1
enanito	1

Grupo de edad: 4	
Forma	**Frecuencia**
demonio	1
boca	1
saludo	1
ala	1
camioneros	1
vuela	1
izquierda	1
comprar	1
Izán	1
daba	1
echaré	1
disfrace	1
zapatos	1
coloretes	1
poner	1
darle	1
profesor	1
televisión	1
pintando	1
junto	1
puedes	1
vas	1
estirada	1
borra	1
salido	1
entrar	1
di	1
radio	1
hasta que	1
despertó	1
González	1
pequeñas	1
carpa	1
tarde	1
pajarito	1
lavando	1
herida	1
libros	1
Piolín	1
purpurina	1
chupachús	1
rompió	1
canción	1
profesores	1
zumo	1
hablar	1
tiré	1
Blancanieves	1
fuente	1
granja	1

Grupo de edad: 4	
Forma	**Frecuencia**
Tasmania	1
esos	1
sueño	1
bocadillos	1
conejo	1
sientas	1
extraterrestre	1
vendrá	1
salva	1
quietos	1
mintió	1
maestro	1
basura	1
cosido	1
daño	1
olvidado	1
etiqueta	1
vemos	1
llorando	1
cochinillas	1
merendar	1
gorda	1
vámonos	1
Benidorm	1
columpiándose	1
Jorgito	1
además	1
Pedro	1
avispa	1
paz	1
fecha	1
pastel	1
llora	1
dejé	1
apellido	1
estiraban	1
acaba	1
miau	1
primos	1
terminar	1
de leche	1
careta	1
Wendy	1
block	1
duermas	1
tonterías	1
gata	1
quedé	1
pongo	1
primas	1

Grupo de edad: 4	
Forma	**Frecuencia**
sangre	1
de hierro	1
terminado	1
aros	1
salir	1
dar	1
be	1
diecisiete	1
guardería	1
regulín regulán	1
rosquillo	1
corriendo	1
cerrado	1
estirando	1
pintado	1
susto	1
subibaja	1
ninguno	1
adiós	1
doctor	1
ocho	1
bola	1
quite	1
celebré	1
Martín	1
echaba	1
venía	1
de pie	1
sabéis	1
tendré	1
zapato	1
corté	1
sube	1
tripita	1
conejos	1
paseos	1
escucha	1
comerte	1
disfraz	1
joven	1
echada	1
pude	1
abuelito	1
carrito	1
metieron	1
estuve	1
morada	1
quieran	1
ven	1
ruedas	1

Grupo de edad: 4	
Forma	**Frecuencia**
manguitos	1
meten	1
tardó	1
contenedor	1
gimnasio	1
veinticinco	1
veintiséis	1
comérselos	1
tocar	1
metemos	1
cacé	1
Campo de Criptana	1
separados	1
puso	1
cuna	1
pudieron	1
entiende	1
cuentos	1
tuyo	1
labios	1
rotulador	1
gol	1
telas	1
Iván	1
empujándole	1
superhéroe	1
vimos	1
nuestro	1
quedan	1
té	1
rascarle	1
traer	1
patitos	1
papeles	1
deportivas	1
pinturines	1
tardaron	1
perdió	1
moto	1
metes	1
Guadalupe	1
liebres	1
cantas	1
hermanita	1
llevan	1
elefantes	1
Jorge González	1
alfombra	1
ranas	1
chinas	1

Grupo de edad: 4	
Forma	**Frecuencia**
lava	1
borrará	1
panda	1
veinticuatro	1
Mickey Mouse	1
liebre	1
monstruos	1
papa	1
cortar	1
juega	1
diez	1
colores	1
compró	1
Barce	1
pegó	1
empezar	1
dices	1
vete	1
echo	1
pies	1
Rosi	1
despacito	1
sustos	1
campo	1
dedos	1
buh	1
juguete	1
oso	1
amarilla	1
zapatillas	1
llevas	1
llegáis	1
quita	1
camión	1
ge	1
pasar	1
caía	1
pedí	1
fantasmas	1
pequeña	1
al lado de	1
echado	1
dicen	1
i	1
casados	1
saludé	1
ahí va	1
pegarme	1
veas	1
alto	1

Grupo de edad: 4	
Forma	**Frecuencia**
por la noche	1
oyó	1
entré	1
Pepón	1
balonmano	1
traje	1
babi	1
porche	1
rodeo	1
terraza	1
veintiuno	1
mar	1
noches	1
gatito	1
cerrar	1
digan	1
American Dragon	1
marrón	1
granos	1
llamo	1
fichas	1
digo	1
asusta	1
duele	1
señores	1
en seguida	1
buscando	1
espalda	1
gritaban	1
estudiar	1
de momento	1
aparcan	1
subíamos	1
bañaba	1
arriba	1
Balto	1
pinchan	1
cenar	1
mirando	1
responsable	1
mayo	1
juntos	1
Sancho Panza	1
pila	1
contado	1
cojo	1
subiéndose	1
cumplió	1
piedra	1
nueva	1

Grupo de edad: 4	
Forma	**Frecuencia**
tendrá	1
oímos	1
vea	1
abraza	1
subir	1
Vanesa	1
estuvimos	1
Eto'o	1
treinta	1
foto	1
pelotas	1
trabajando	1
carro	1
diecinueve	1
cumples	1
míos	1
osito	1
alcanza	1
guardada	1
serpiente	1
picó	1
íbamos	1
cayó	1
Dulce María	1
cansino	1
rojo	1
es más	1
enamorados	1
cena	1
ordenador	1
hace	1
columpiar	1
estuche	1
siesta	1
Juli	1
tirábamos	1
gitano	1
comido	1
bolsillo	1
llevarle	1
cincuenta	1
cohete	1
creía	1
orejas	1
mancha	1
mil	1
tira	1
carretera	1
puesta	1
vamos	1

Grupo de edad: 4	
Forma	**Frecuencia**
tragó	1
entrado	1
Homer	1
querrán	1
maestra	1
jugué	1
ti	1
Miguel	1
cocacola	1
pastelitos	1
crecer	1
mocos	1
correcaminos	1
cien	1
tirachinas	1
trae	1
u	1
mejor	1
redondeles	1
disfrazarse	1
chache	1
Candi	1
peiné	1
envenenada	1
escondieron	1
derecha	1
un rato	1
Sara	1
cumplí	1
dedo	1
meter	1
hermanas	1
mayores	1
mojado	1
gritó	1
Simpson	1
dieciséis	1
compraba	1
campos	1
amigas	1
montar	1
quema	1
saltarina	1
planchar	1
alguna	1
tiró	1
papás	1
duermo	1
castigan	1
misma	1

Grupo de edad: 4	
Forma	**Frecuencia**
serpientes	1
Lucas	1
comí	1
corre	1
ganan	1
Félix	1
levanto	1
subido	1
Fabián	1
peluca	1
nave	1

Forma	Frecuencia
cumplir	1
brazos	1
estantería	1
vosotros	1
cortos	1
clases	1
tacón	1
pasamos	1
encontró	1
caballito	1
balón	1
abrir	1

b. Frecuencia de categorías léxicas: N, V y ADJ

Frecuencia de N	
Lema	**Frecuencia**
niño	42
madre	29
casa	26
padre	22
primo	17
hermano	17
abuelo	15
vez	13
agua	11
lobo	11
patio	11
año	10
perro	9
e	9
clase	9
tortuga	9
fútbol	9
amigo	9
piedra	8
coche	8
jaula	8
tele	8
payaso	7
mamá	7
castillo	7
cosa	7
película	6
cumple	6
número	6
cuento	6
disco	6
ojo	6
día	6

Frecuencia de N	
Lema	**Frecuencia**
martes	6
pierna	6
camión	6
mano	6
abuelito	6
gato	6
mama	5
noche	5
rana	5
lunes	5
pájaro	5
animal	5
sábado	5
pantalón	5
cole	5
piscina	5
domingo	5
monstruo	5
pera	5
café	5
tobogán	4
viernes	4
trabajo	4
hijo	4
ratón	4
pelota	4
jueves	4
cuello	4
nueve	4
cartera	4
rumano	4
contenedor	4
piso	4

Frecuencia de N	
Lema	**Frecuencia**
dragón	4
cama	4
once	4
dibujo	4
chico	4
a	4
lápiz	3
cuadra	3
cumpleaños	3
manzana	3
comida	3
camiseta	3
gitano	3
montón	3
carnaval	3
base	3
ficha	3
lila	3
sitio	3
estuche	3
tío	3
pato	3
bota	3
vaso	3
juguete	3
barco	3
libro	3
biberón	3
febrero	3
sol	3
palomo	3
papá	3
tarta	3
pata	2
leche	2
cajón	2
pintalabios	2
nido	2
cometa	2
profesor	2
bebé	2
guitarra	2
juego	2
logopeda	2
gente	2
calendario	2
huevo	2
arena	2
playa	2
gusano	2

Frecuencia de N	
Lema	**Frecuencia**
piano	2
oveja	2
varicela	2
tarjeta	2
gafa	2
miércoles	2
palo	2
silla	2
boli	2
cabeza	2
canasta	2
colegio	2
pelo	2
fuego	2
vuelta	2
susto	2
quince	2
papel	2
hora	2
abeja	2
colchoneta	2
pala	2
chupete	2
campo	2
escuela	2
serpiente	2
luna	2
escondite	2
pastel	2
liebre	2
foca	2
obra	2
morro	2
maestro	2
columpio	2
plancha	2
estufa	2
bigote	2
doce	2
cara	2
taza	2
mujer	2
chupachús	2
pantano	2
empresa	2
canción	2
colorín	2
portería	2
dedo	2
llave	2

Frecuencia de N	
Lema	**Frecuencia**
horario	2
cepa	2
río	2
calle	2
móvil	2
zapato	2
familia	2
grano	2
conejo	2
tijera	2
suelo	2
furgoneta	2
verano	2
enanito	2
pollito	2
carpa	1
maestra	1
boca	1
ge	1
cuna	1
cazador	1
pie	1
etiqueta	1
fecha	1
pajarito	1
doctor	1
oso	1
trece	1
labio	1
rato	1
fuente	1
señor	1
plus	1
espalda	1
bicicleta	1
angina	1
gatito	1
paseo	1
mayo	1
sueño	1
tirachinas	1
terraza	1
dieciocho	1
patito	1
elefante	1
camino	1
tela	1
careta	1
extraterrestre	1
salto	1

Frecuencia de N	
Lema	**Frecuencia**
cohete	1
cesta	1
manguito	1
avispa	1
nube	1
herida	1
amiguito	1
ballena	1
osito	1
aro	1
regalo	1
babi	1
panda	1
dóberman	1
rotulador	1
pico	1
cochinilla	1
paz	1
cocacola	1
enamorado	1
purpurina	1
zapatilla	1
puente	1
diente	1
hermanito	1
pis	1
tres	1
guardería	1
daño	1
tambor	1
fantasma	1
cola	1
quema	1
cena	1
siete	1
tacón	1
subibaja	1
be	1
bolsillo	1
cuidado	1
siesta	1
block	1
derecha	1
o	1
china	1
ala	1
pila	1
brazo	1
disfraz	1
sapo	1

Frecuencia de N	
Lema	**Frecuencia**
corazón	1
carro	1
moco	1
carpeta	1
té	1
u	1
misa	1
correcaminos	1
balonmano	1
alfombra	1
colonia	1
tiro	1
moto	1
cera	1
estantería	1
tata	1
radio	1
control	1
color	1
compañero	1
trozo	1
muchacho	1
colorete	1
izquierda	1
nombre	1
peluca	1
espada	1
televisión	1
carrito	1
balón	1
bola	1
borra	1
rosquillo	1
gimnasio	1
miau	1
tripa	1
nave	1
león	1
habitación	1
policía	1
i	1
carretera	1

Lema	Frecuencia
papa	1
crema	1
gol	1
brujo	1
bocadillo	1
ese	1
dos mil seis	1
sofá	1
subida	1
granja	1
debe	1
rueda	1
empate	1
foto	1
señora	1
porche	1
apellido	1
pinturín	1
tontería	1
gomina	1
abril	1
dos	1
oreja	1
responsable	1
camionero	1
redondel	1
polvo	1
ropa	1
ordenador	1
catorce	1
demonio	1
mosca	1
zumo	1
jarabe	1
permiso	1
recreo	1
basura	1
uno	1
veintiséis	1
gata	1
chache	1
caballito	1
sangre	1

Frecuencia de V	
Lema	**Frecuencia**
ser	122
tener	88
ir	42
saber	36
ver	36
haber	28
hacer	28
decir	27
jugar	24
estar	24
llamar	21
poner	18
llevar	15
coger	14
pasar	13
comer	13
trabajar	13
dar	12
pintar	12
salir	11
echar	11
gustar	10
meter	9
cumplir	9
tirar	8
querer	7
pegar	7
picar	7
comprar	6
llegar	6
hablar	6
venir	6
quedar	5
caer	5
subir	5
quitar	5
correr	4
ganar	4
mirar	4
morir	4
dejar	4
casar	4
volar	4
sentar	4
levantar	4
poder	4
entrar	3
coser	3
convertir	3

Frecuencia de V	
Lema	**Frecuencia**
saltar	3
aparcar	3
traer	3
vivir	3
acabar	3
leer	3
olvidar	3
cazar	3
portar	3
quemar	3
beber	3
bañar	3
contar	3
dormir	3
regalar	3
lavar	3
parecer	3
merendar	3
terminar	3
disfrazar	2
mandar	2
estirar	2
llorar	2
empezar	2
atacar	2
columpiar	2
rascar	2
gritar	2
asustar	2
creer	2
entender	2
perder	2
saludar	2
acordar	2
alcanzar	2
abrir	2
hundir	2
planchar	2
cerrar	2
tocar	2
cortar	2
escribir	2
faltar	2
oír	2
tardar	2
matar	1
reír	1
escuchar	1
crecer	1

Frecuencia de V	
Lema	**Frecuencia**
doler	1
pinchar	1
despertar	1
cenar	1
hurgar	1
esconder	1
pasear	1
aprender	1
romper	1
rodear	1
abrochar	1
pisar	1
mentir	1
abrigar	1
vestir	1
chupar	1
andar	1
peinar	1
tragar	1
celebrar	1
acostar	1
buscar	1

Lema	**Frecuencia**
borrar	1
cantar	1
montar	1
pedir	1
salvar	1
encontrar	1
guardar	1
manchar	1
estudiar	1
juntar	1
asfixiar	1
enseñar	1
arropar	1
abrazar	1
encender	1
callar	1
intentar	1
envenenar	1
escupir	1
castigar	1
empujar	1
descoser	1

Frecuencia de ADJ	
Lema	**Frecuencia**
grande	20
pequeño	8
malo	8
azul	6
chico	5
amarillo	4
vivo	4
gordo	4
feo	4
lila	4
de verdad	4
bueno	4
solo	3
marrón	3
blanco	3
entero	2
roto	2
colorado	2
peligroso	2
duro	2
naranja	2
abierto	2

Frecuencia de ADJ	
Lema	**Frecuencia**
mayor	2
bonito	2
de miedo	2
negro	2
verde	2
frío	2
alto	2
de hierro	1
joven	1
triste	1
mismo	1
rosa	1
cansado	1
nuevo	1
viejo	1
junto	1
morado	1
rojo	1
deportivo	1
estirado	1
saltarín	1
oscuro	1

Frecuencia de ADJ	
Lema	**Frecuencia**
maestro	1
digital	1
cansino	1
chino	1
suelto	1
de leche	1

Lema	Frecuencia
separado	1
quieto	1
corto	1
casado	1
chiquitín	1
mojado	1

4. Grupo de edad: 5 años

a. Frecuencia de formas

Grupo de edad: 5	
Forma	**Frecuencia**
y	495
no	270
que	255
sí	245
a	243
la	207
el	186
me	181
se	162
un	154
mi	153
de	150
yo	146
en	131
le	108
una	106
los	103
con	95
lo	94
pero	73
porque	68
pues	68
es	67
Candi	64
dos	59
ya	56
también	54
las	50
sé	48
al	44
tiene	43
está	42
es que	42
más	41
tengo	40
casa	40

Grupo de edad: 5	
Forma	**Frecuencia**
para	39
por	38
cuando	37
ha	37
uno	36
padre	36
qué	32
como	31
madre	31
cinco	29
estaba	28
su	28
he	27
va	26
otro	26
o	25
tres	23
ahí	23
te	24
si	22
Carmen	22
estás	20
voy	20
mí	20
cuatro	20
Raúl	19
nos	19
muy	19
así	19
hermana	18
entonces	18
cama	17
del	17
hoy	17
fue	17
jugar	17

Grupo de edad: 5	
Forma	**Frecuencia**
buenos días	16
Daniel	16
después	16
luego	16
día	16
nada	16
todos	16
aquí	15
Esther	15
grande	15
otra	15
Mulan	14
vino	14
parece	14
euro	14
tenía	14
sol	14
ocho	14
hay	14
Arturo	14
señorita	14
sólo	13
seis	13
lobo	13
cosas	13
dicho	13
eso	13
Carlos	13
veces	13
vamos	13
sabes	12
mamá	12
casi	12
mucho	12
había	12
quería	12
cuento	12
son	12
era	11
cosa	11
jueves	11
Ana Pilar	11
galletas	11
Vicente	11
algunas	11
ah	11
vez	11
leche	10
han	10
Laura	10

Grupo de edad: 5	
Forma	**Frecuencia**
perra	10
quién	10
un poco	10
Marisol	10
arriba	10
cayó	10
ir	9
colacao	9
esta	9
ese	9
miércoles	9
allí	9
Inés	9
nueve	9
dijo	9
acuerdo	8
mira	8
muchas	8
martes	8
céntimos	8
Marta	8
siete	8
vaso	8
abuelo	8
da	8
ella	8
muchos	8
hecho	8
diez	8
ser	8
Furi	8
guerra	8
María	8
patio	8
doce	8
Spiderman	8
llama	8
mío	7
domingo	7
pone	7
hacer	7
As	7
días	7
ropa	7
igual que	7
hemos	7
abuela	7
antes	7
malo	7
de leche	7

Grupo de edad: 5	
Forma	**Frecuencia**
este	7
Sofía	7
lunes	7
él	7
tía	7
hizo	7
sábado	7
señorito	7
iba	7
ayer	7
abajo	7
fútbol	7
Marcos	7
Luis	7
primo	6
cabeza	6
cumpleaños	6
viene	6
tienes	6
acosté	6
coche	6
tú	6
puede	6
siempre	6
deja	6
fuera	6
estaban	6
salió	6
comprar	6
papá	6
Liana	6
perro	6
grandes	6
cole	6
venga	6
bien	6
chicos	6
por la noche	6
Caperucita	6
unas	6
otros	6
regalado	6
hace	6
niña	6
roto	6
cien	6
bueno	6
ni	6
perros	5
cabritillos	5

Grupo de edad: 5	
Forma	**Frecuencia**
menos	5
rapar	5
van	5
chuches	5
tampoco	5
esa	5
león	5
toca	5
segundo	5
líquido	5
vas	5
di	5
vale	5
Octopus	5
regaló	5
febrero	5
corriendo	5
vio	5
cae	5
Tomelloso	5
pasado	5
viernes	5
puesto	5
sólido	5
perdió	5
portero	5
campo	5
cogió	5
jugando	5
chica	5
estamos	5
Pepe	5
Aila	5
cojo	5
hucha	5
rojo	5
hice	5
dónde	5
Granada	5
ponemos	5
Zaragoza	5
pasó	5
a ver	5
cara	5
chicas	5
tocó	4
aunque	4
nunca	4
fantasma	4
tambor	4

Grupo de edad: 5	
Forma	**Frecuencia**
invierno	4
tiempo	4
debajo de	4
china	4
tarta	4
año	4
años	4
todo	4
casita	4
escalera	4
quiero	4
pijama	4
sábados	4
ojos	4
hacemos	4
elefante	4
domingos	4
a lo mejor	4
soñado	4
partido	4
toda	4
poner	4
dice	4
algo	4
detrás	4
céntimo	4
puedo	4
esos	4
vi	4
fui	4
verde	4
llamó	4
chico	4
caí	4
enero	4
lo que	4
verano	4
metió	4
ramo	4
quince	4
cómo	4
gol	4
Guadix	4
niños	4
moto	4
debajo	4
vive	4
comprado	4
Madrid	4
amigo	4

Grupo de edad: 5	
Forma	**Frecuencia**
dormir	4
poquito	4
falta	4
euros	4
casco	4
comer	4
mejor	4
diciembre	4
has	4
tele	4
meter	4
Pachín	4
brazos	4
piano	4
mía	4
traído	4
Jorge	3
sillón	3
encanta	3
esas	3
pasa	3
decir	3
Palma	3
fueron	3
tirar	3
duerme	3
digital	3
ay	3
entre	3
doy	3
Mari	3
veintiocho	3
abrió	3
comida	3
e	3
periódico	3
dinero	3
tiro	3
salgo	3
por qué	3
hora	3
luchar	3
hojas	3
corto	3
pueden	3
cereales	3
trabaja	3
morro	3
igual	3
tuvo	3

Grupo de edad: 5	
Forma	**Frecuencia**
pico	3
nuestra	3
Abel	3
mal	3
Cora	3
compra	3
suave	3
celebramos	3
llegaba	3
guardo	3
oye	3
mesa	3
calle	3
daño	3
bebé	3
juguetes	3
salimos	3
murió	3
Santa Claus	3
sangre	3
clase	3
Rubén	3
porta	3
quitado	3
mordió	3
lleva	3
chimenea	3
cuál	3
cero	3
mis	3
fuego	3
hacen	3
pelo	3
ochenta	3
colegio	3
puso	3
soles	3
están	3
chocolate	3
pie	3
puerta	3
pequeño	3
cortar	3
Caperucita Roja	3
zapatillas	3
caballo	3
temprano	3
paré	3
Manzanares	3
duende	3

Grupo de edad: 5	
Forma	**Frecuencia**
ce	3
agua	3
ahora	3
hache	3
porto	3
quiere	3
todas	3
tenemos	3
pelota	3
gustan	3
bombones	3
algunos	3
ordenador	3
animales	3
culo	3
caso	3
prima	3
once	3
creía	3
negro	3
dejan	3
escondió	3
niño	3
flores	3
dio	3
brasero	3
todavía	3
venir	3
humo	3
escondite	3
Torres	3
diario	3
ellos	3
desayunado	3
lejos	3
buenos	3
madera	3
piedras	3
balón	3
pequeñito	2
colorado	2
ganas	2
quedado	2
juntamos	2
caiga	2
camas	2
pintamos	2
vestirse	2
golpe	2
agosto	2

Grupo de edad: 5	
Forma	**Frecuencia**
árbol	2
regalos	2
catorce	2
trece	2
Rodrigo	2
Adelardo	2
caja	2
enseñó	2
cuenta	2
Lanza	2
abeto	2
hablando	2
perdido	2
saltos	2
dientes	2
pavo	2
Shin-Chan	2
coges	2
plus	2
hacía	2
Javier	2
cerditos	2
claro	2
rueda	2
septiembre	2
ganó	2
Pepito	2
creo	2
olvidó	2
la que	2
seguro que	2
bye	2
creían	2
espada	2
tenían	2
veranos	2
ayudar	2
mañana	2
araña	2
cachorros	2
julio	2
cartel	2
esto	2
soplarlas	2
sus	2
hijas	2
encender	2
rosa	2
mayor	2
podía	2

Grupo de edad: 5	
Forma	**Frecuencia**
albañiles	2
camino	2
echó	2
pueblo	2
flechas	2
contra	2
Arturito	2
muere	2
encontré	2
gateando	2
marzo	2
encima de	2
alta	2
ratón	2
caballeros	2
calla	2
a veces	2
subo	2
soñó	2
otras	2
haberte	2
mató	2
armario	2
juego	2
lavarme	2
os	2
daba	2
zapatos	2
literas	2
mechero	2
cuánto	2
puedes	2
cosió	2
andar	2
enterré	2
tarde	2
secreto	2
entera	2
eran	2
rompió	2
al final	2
noviembre	2
hermano	2
leemos	2
gafas	2
demás	2
colorín	2
basura	2
pagar	2
película	2

Grupo de edad: 5	
Forma	**Frecuencia**
coleta	2
levantado	2
nosotros	2
primos	2
tripas	2
nadie	2
sitio	2
acabado	2
adiós	2
fíjate	2
pepino	2
hola	2
baranda	2
venía	2
llamas	2
por la mañana	2
pisa	2
quinto	2
reina	2
sujetado	2
ven	2
gimnasio	2
Pokémon	2
rota	2
buena	2
dragón	2
tocar	2
regalaron	2
duermen	2
cuna	2
cuentos	2
ver	2
Zipi	2
escaleras	2
gorro	2
primero	2
camiseta	2
collar	2
Marca	2
veinticuatro	2
juega	2
compró	2
venido	2
El País	2
encendía	2
haber	2
litera	2
echo	2
pequeños	2
caían	2

Grupo de edad: 5	
Forma	**Frecuencia**
comió	2
Carla	2
Socuéllamos	2
bosque	2
quita	2
despertado	2
buscar	2
camión	2
pino	2
octubre	2
El Mundo	2
muerde	2
despacio	2
soy	2
castillo	2
chupete	2
cuándo	2
goles	2
correr	2
equipo	2
marrón	2
Ronaldinho	2
amo	2
enseguida	2
gusta	2
ea	2
solo	2
pulsera	2
somos	2
saqué	2
rinoceronte	2
sacar	2
caer	2
Balbi	2
subir	2
Julia	2
podemos	2
tonto	2
enamorados	2
hicimos	2
comido	2
carretera	2
leer	2
vestido	2
dejaba	2
tanto	2
tercero	2
castaña	2
sabemos	2
cada	2

Grupo de edad: 5	
Forma	**Frecuencia**
haciendo	2
tan	2
abuelos	2
peinado	2
hamster	2
patada	2
noventa	2
tocando	2
baloncesto	2
alguna	2
tiró	2
papás	2
ganado	2
Mari Flor	2
Carlota	2
peluche	2
río	2
voltereta	2
escribir	2
dormí	2
nave	2
por el suelo	2
trabajamos	2
cogía	2
donde	2
lee	2
Playmobiles	2
rey	2
castaño	2
hoja	2
veo	2
ruiseñor	1
coge	1
Carmencita	1
habitación	1
niñas	1
Phoskitos	1
en vez de	1
manchegas	1
cazador	1
dándole	1
Diana	1
leí	1
vuelta	1
Mónica	1
miel	1
Balbino	1
cajones	1
empiezas	1
inventamos	1

Grupo de edad: 5	
Forma	**Frecuencia**
sabe	1
Messi	1
coloreado	1
gomina	1
engañó	1
Hansel	1
cola	1
trampa	1
Belletti	1
voz	1
valga	1
levanta	1
rápido	1
mandos	1
cuidado	1
marcar	1
disparado	1
Antonio Delgado	1
ovejita	1
Halloween	1
torre	1
malita	1
buhardilla	1
cumplo	1
sacó	1
de verdad	1
compre	1
rosco	1
Gameboy	1
canguro	1
acerca	1
nombre	1
afinando	1
verla	1
abril	1
Jorquera	1
doble	1
nubes	1
farmacia	1
tela	1
jugó	1
bebido	1
búhos	1
columna	1
llamaba	1
largo	1
pintar	1
llevó	1
vocales	1
lavarte	1

Grupo de edad: 5	
Forma	**Frecuencia**
noche	1
escapar	1
barrita	1
pis	1
no ?	1
saca	1
Manolo Torres	1
abierta	1
montando	1
compras	1
nueces	1
pequeñitas	1
ja	1
levanté	1
corta	1
vengo	1
lado	1
caes	1
negros	1
palo	1
ciprés	1
muñeco	1
dibujo	1
rojas	1
palos	1
toco	1
primer	1
estiras	1
baila	1
fuimos	1
dos mil seis	1
teles	1
último	1
payaso	1
vía	1
pedales	1
atropelló	1
al principio	1
volvían	1
saliendo	1
idioma	1
en coche	1
sabían	1
tazón	1
conduzco	1
caíste	1
hundió	1
comprarse	1
grapar	1
ponerme	1

Grupo de edad: 5	
Forma	**Frecuencia**
bonito	1
fin de semana	1
Iniesta	1
meterlo	1
poquitín	1
Ramón	1
Irene	1
reyes	1
tumba	1
pata	1
aburro	1
saltemos	1
máquina	1
Cristina	1
estoy	1
chiquitín	1
llamarle	1
rectangular	1
cumplido	1
frío	1
gatear	1
tostada	1
fiesta	1
zancadilla	1
comiera	1
romperte	1
soñando	1
traigo	1
cresta	1
caperuza	1
bañé	1
comprarme	1
suyo	1
perece	1
entrenadora	1
Alejandro	1
mentirosa	1
huevo	1
manos	1
sea	1
callarse	1
jugaste	1
inventar	1
póster	1
seas	1
bailamos	1
palas	1
despierte	1
blancas	1
colgante	1

Grupo de edad: 5	
Forma	**Frecuencia**
Batman	1
Maxi López	1
comprármelo	1
venda	1
arco	1
Fernando Torres	1
comidita	1
prepara	1
beber	1
cartón	1
enganchada	1
gordo	1
abuelita	1
dije	1
manzana	1
invité	1
segunda	1
delante	1
dale	1
colegios	1
pudiste	1
mismos	1
veía	1
preferido	1
rama	1
quemó	1
colonia	1
san	1
sombrero	1
será	1
cosita	1
rompieron	1
Manchitas	1
chupa	1
zampó	1
ayuda	1
abiertos	1
erres	1
lanzas	1
tantas	1
sofá	1
encontraron	1
rectángulo	1
pilla besos	1
varicela	1
sombra	1
regalar	1
matar	1
dejara	1
esté	1

Grupo de edad: 5	
Forma	**Frecuencia**
tomado	1
echar	1
animal	1
y todo	1
horquilla	1
sacan	1
vomitó	1
en cuanto	1
restaurante	1
perdíamos	1
ellas	1
regalarle	1
boca	1
quedo	1
leotardo	1
vuela	1
ey	1
anda	1
malos	1
empezamos	1
darle	1
Noel	1
profesor	1
toma	1
salía	1
encuentra	1
llamar	1
magdalenas	1
bebés	1
perdiz	1
jaula	1
sofás	1
flojo	1
bolsas	1
charco	1
salido	1
entrar	1
feo	1
tío	1
helicóptero	1
octavo	1
hasta que	1
fácil	1
dan	1
rompí	1
chiquituja	1
sierras	1
Simpsons	1
escondí	1
Rompehuesos	1

Grupo de edad: 5	
Forma	**Frecuencia**
saques	1
cartera	1
subí	1
Carmina	1
rana	1
Antonio	1
pulgas	1
espachurra	1
Rosa	1
ducha	1
cogiendo	1
rosas	1
perrito	1
bañamos	1
dibujado	1
pide	1
conozco	1
canción	1
suena	1
sierra	1
zumo	1
echaron	1
junio	1
tiré	1
deportes	1
corría	1
mano	1
Ángel	1
limpiar	1
inventado	1
sueño	1
conejo	1
divirtiendo	1
añitos	1
gallina	1
a primeros	1
chaqueta	1
arar	1
probarla	1
quietos	1
pum	1
felicitarla	1
banda	1
explotó	1
montón	1
maestro	1
guapa	1
Playmobil	1
pagando	1
hasta	1

Grupo de edad: 5	
Forma	**Frecuencia**
baile	1
el que	1
llorando	1
masajes	1
Papá	1
regalé	1
cuarto	1
recogiendo	1
lío	1
verdad ?	1
gorda	1
frente	1
teléfono	1
ruido	1
estar	1
olivo	1
nieve	1
celebrar	1
disparó	1
efe	1
fecha	1
coneja	1
dejé	1
lechuza	1
cambian	1
guantes	1
hagas	1
partidos	1
jefe	1
careta	1
trabajar	1
usar	1
Sierra Nevada	1
felicitar	1
cocodrilo	1
vamos a ver	1
mayúscula	1
llévate	1
ten	1
dimos	1
quedarse	1
pincho	1
cuadro	1
flor	1
quemada	1
de hierro	1
resfriar	1
llave	1
americana	1
habrás	1

Grupo de edad: 5	
Forma	**Frecuencia**
noticia	1
hermanitos	1
violonchelo	1
San Valentín	1
hello	1
dioses	1
quemaron	1
lista	1
cañón	1
regala	1
doctor	1
bola	1
quite	1
señor	1
silla	1
Zaragoza Madrid	1
taza	1
echaba	1
soldados	1
escrito	1
caemos	1
lavo	1
sabes ?	1
balones	1
cansamos	1
aprendiendo	1
ponérselo	1
limpia	1
patito	1
comerte	1
caza	1
ceno	1
podías	1
escapa	1
adoptar	1
perdiéramos	1
raro	1
añito	1
chiquitina	1
mismo	1
llévale	1
vacía	1
quitó	1
pantalones	1
metieron	1
estuve	1
perdí	1
meten	1
contenedor	1
sentada	1

Grupo de edad: 5	
Forma	**Frecuencia**
Kinder	1
podíamos	1
qué tal	1
veinticinco	1
de menos	1
empezó	1
jugamos	1
amigos	1
queda	1
echa	1
pinche	1
Elisa	1
ESO	1
vestirme	1
treinta y cinco	1
yegua	1
puntos	1
verte	1
miro	1
conmigo	1
beso	1
Manolo	1
tractores	1
colaron	1
Shan	1
flaca	1
cogemos	1
trabajo	1
pasta	1
fino	1
acostarse	1
más que	1
vienen	1
tranquila	1
tuyo	1
corbata	1
regañamos	1
salieron	1
encontrado	1
oía	1
enfada	1
calzado	1
traer	1
después de	1
patitos	1
amarillo	1
veintinueve de agosto	1
mataban	1
metes	1
dentro	1

Grupo de edad: 5	
Forma	**Frecuencia**
chinas	1
ratones	1
caretas	1
vinieron	1
estabas	1
Gretel	1
así es que	1
abalanzó	1
unos	1
monstruos	1
muñecas	1
vídeos	1
colores	1
nuevas	1
hija	1
pegó	1
uf	1
medio	1
apuntado	1
nació	1
José María Delgado	1
pies	1
despacito	1
lavado	1
comimos	1
inglés	1
Argamasilla	1
superhéroes	1
trepa	1
visto	1
sale	1
tazas	1
sillita	1
picaron	1
escuchado	1
llevas	1
cáscara	1
tienen	1
ido	1
profe	1
rastro	1
Sergio	1
debido	1
Fernando Alonso	1
escribí	1
gustó	1
ge	1
pasar	1
caramelo	1
caía	1

Grupo de edad: 5	
Forma	**Frecuencia**
ratoncito	1
lavarse	1
cebra	1
pequeña	1
Jesús	1
salvarla	1
Larsson	1
echado	1
Adrián	1
dicen	1
dieciséis de agosto	1
i	1
hembra	1
zanahoria	1
pisar	1
viendo	1
tomar	1
convierte	1
despertar	1
recortamos	1
mariposas	1
carnaval	1
diste	1
teletexto	1
soltó	1
alto	1
Barcelona	1
entreno	1
oyó	1
gris	1
negra	1
piruleta	1
mezclamos	1
colorear	1
cuenta cuentos	1
durmiendo	1
traje	1
subía	1
garaje	1
cestita	1
babi	1
Aurelio	1
abrimos	1
olerte	1
comerse	1
dibujos	1
sumar	1
ocurre	1
pasan	1
cinta	1

Grupo de edad: 5	
Forma	**Frecuencia**
aparato	1
cosquillitas	1
noches	1
hospital	1
otoño	1
Adriana	1
digan	1
tarda	1
plato	1
letra	1
ventana	1
necesita	1
paseando	1
bollo	1
fichas	1
enfado	1
Papá Noel	1
goma	1
toqué	1
expresión	1
corneja	1
ve	1
nevó	1
arbustos	1
buscando	1
tuve	1
Pilar	1
primera	1
cambiado	1
jueguen	1
edad	1
tiraba	1
encendió	1
bañó	1
vistiendo	1
responsable	1
canarios	1
mayo	1
pasteles	1
Valdés	1
churrería	1
juntos	1
cumple	1
Juanito	1
morir	1
costado	1
sabía	1
invento	1
Pachán	1
ele	1

Grupo de edad: 5	
Forma	**Frecuencia**
piedra	1
nueva	1
dejar	1
manga	1
existe	1
hermanos	1
piel	1
dieciocho	1
jota	1
rascacielos	1
hago	1
naranja	1
Eto'o	1
guardamos	1
pelotas	1
bolsa	1
carro	1
llevando	1
quedaba	1
míos	1
oírte	1
veintitrés de febrero	1
Superman	1
bajando	1
cerebro	1
apagaron	1
desayuno	1
picó	1
esponja	1
tenido	1
íbamos	1
por la tarde	1
vueltas	1
difícil	1
alcachofa	1
rica	1
cena	1
violín	1
viaje	1
desayunar	1
bolsillo	1
perenne	1
quitársela	1
mataba	1
llevarle	1
cincuenta	1
punta	1
entró	1
orejas	1
grises	1

Grupo de edad: 5	
Forma	**Frecuencia**
en medio	1
tira	1
metí	1
Claus	1
tren	1
maestra	1
huelen	1
jugué	1
carretilla	1
parte	1
Miguel	1
botón	1
a una	1
abeja	1
pera	1
quedó	1
chicles	1
quitan	1
entretuvo	1
entiendo	1
única	1
corporal	1
manchas	1
notas	1
apagaban	1
felicidades	1
adelante	1
gorra	1
barco	1
desapareció	1
Faustino	1
cumplí	1
Faustina	1
guarde	1
pirulí	1
levante	1
levántate	1
mayores	1
Giuly	1
Simpson	1
metiendo	1
compraba	1
dejó	1
superfuerte	1
pues nada	1

Forma	**Frecuencia**
amigas	1
montar	1
loca	1
regalo	1
paro	1
descanse	1
monedas	1
como si	1
soldado	1
pares	1
barriga	1
sentamos	1
amiga	1
duermo	1
vaya	1
pille pille	1
comí	1
sin	1
tripa	1
clavó	1
levanto	1
pronto	1
nariz	1
antes de	1
cazar	1
cortó	1
contar	1
rastrillo	1
irme	1
haga	1
pelean	1
fregaron	1
petardo	1
últimos	1
raja	1
tacón	1
encontró	1
pregunto	1
pagarte	1
falda	1
llegarás	1
vampiro	1
coja	1
escapó	1
abrir	1
vivía	1

b. Frecuencia de categorías léxicas: N, V y ADJ

Frecuencia de N	
Lema	**Frecuencia**
casa	40
padre	36
madre	31
cosa	25
vez	24
día	23
hermano	21
chico	20
cama	19
euro	18
abuelo	17
sol	17
niño	14
señorita	14
cuento	13
lobo	13
céntimo	12
mamá	12
perro	11
galleta	11
jueves	11
primo	11
sábado	11
domingo	11
leche	10
perra	10
miércoles	9
colacao	9
año	8
vaso	8
guerra	8
martes	8
patio	8
dos	8
papá	8
ocho	7
nueve	7
lunes	7
señorito	7
tío	7
fútbol	7
amigo	7
ropa	7
cumpleaños	6
cabeza	6
escalera	6
cole	6
coche	6

Frecuencia de N	
Lema	**Frecuencia**
gol	6
verano	6
viernes	5
diez	5
hucha	5
hoja	5
partido	5
sólido	5
campo	5
chuche	5
portero	5
febrero	5
cara	5
león	5
cabrito	5
líquido	5
tele	5
casita	4
pie	4
piedra	4
ratón	4
bebé	4
elefante	4
pelota	4
tiempo	4
comida	4
piano	4
colegio	4
flor	4
animal	4
tambor	4
fantasma	4
brazo	4
ojo	4
diciembre	4
moto	4
balón	4
invierno	4
casco	4
pijama	4
ramo	4
litera	4
enero	4
uno	4
tarta	4
agua	3
hijo	3
mesa	3

Frecuencia de N	
Lema	**Frecuencia**
diario	3
regalo	3
clase	3
duende	3
rey	3
pico	3
dinero	3
zapatilla	3
periódico	3
hache	3
cinco	3
pelo	3
fuego	3
chocolate	3
hora	3
ce	3
daño	3
madera	3
malo	3
bombón	3
sillón	3
cereal	3
juguete	3
escondite	3
humo	3
morro	3
tripa	3
culo	3
caso	3
caballo	3
brasero	3
chimenea	3
calle	3
ordenador	3
puerta	3
sangre	3
soldado	2
cuna	2
bosque	2
película	2
noche	2
plus	2
mayor	2
patada	2
castaño	2
juego	2
patito	2
marzo	2
camino	2
careta	2

Frecuencia de N	
Lema	**Frecuencia**
salto	2
araña	2
collar	2
camiseta	2
sierra	2
coleta	2
cachorro	2
gafa	2
palo	2
pino	2
enamorado	2
silla	2
amo	2
diente	2
añito	2
vuelta	2
armario	2
tres	2
cerdito	2
baloncesto	2
chupete	2
pueblo	2
flecha	2
albañil	2
cartel	2
sitio	2
castaña	2
rosa	2
dragón	2
hamster	2
pepino	2
gorro	2
octubre	2
noviembre	2
septiembre	2
espada	2
doce	2
taza	2
castillo	2
abeto	2
caballero	2
gimnasio	2
mechero	2
nave	2
seis	2
colorín	2
árbol	2
pavo	2
carretera	2
camión	2

Frecuencia de N	
Lema	**Frecuencia**
agosto	2
sofá	2
rueda	2
mano	2
golpe	2
río	2
bolsa	2
zapato	2
veintiocho	2
pulsera	2
baranda	2
rinoceronte	2
peluche	2
conejo	2
julio	2
dibujo	2
reina	2
cero	2
basura	2
caja	2
voltereta	2
equipo	2
pata	1
maestra	1
boca	1
beso	1
ge	1
tazón	1
mariposa	1
charco	1
plato	1
cajón	1
perdiz	1
ele	1
miel	1
manchego	1
caramelo	1
cuadro	1
fecha	1
doctor	1
cresta	1
ventana	1
señor	1
masaje	1
trabajo	1
nieve	1
pedal	1
corneja	1
canario	1
desayuno	1

Frecuencia de N	
Lema	**Frecuencia**
sombra	1
lado	1
rana	1
buhardilla	1
restaurante	1
profesor	1
sueño	1
mayo	1
tela	1
farmacia	1
moneda	1
parte	1
cesta	1
sombrero	1
nube	1
huevo	1
rascacielos	1
magdalena	1
noticia	1
superhéroe	1
corbata	1
botón	1
manzana	1
vídeo	1
babi	1
profe	1
tren	1
piruleta	1
oveja	1
perrito	1
cuarto	1
chaqueta	1
varicela	1
búho	1
cosquillas	1
montón	1
baile	1
cañón	1
cartera	1
idioma	1
jota	1
pincho	1
carnaval	1
tractor	1
horquilla	1
gallina	1
manga	1
pirulí	1
columna	1
cáscara	1

Frecuencia de N	
Lema	**Frecuencia**
guante	1
punto	1
hermanito	1
rastro	1
pis	1
helicóptero	1
punta	1
hembra	1
junio	1
contenedor	1
abeja	1
cinta	1
bollo	1
cola	1
cena	1
pala	1
torre	1
pantalón	1
petardo	1
deporte	1
pasta	1
tacón	1
frente	1
lechuza	1
bolsillo	1
cuidado	1
violín	1
zanahoria	1
fin de semana	1
ficha	1
entrenador	1
lista	1
canguro	1
yegua	1
raja	1
carretilla	1
máquina	1
viaje	1
vía	1
o	1
arco	1
nariz	1
teléfono	1
nuez	1
alcachofa	1
teletexto	1
payaso	1
churrería	1
muñeco	1
barra	1

Frecuencia de N	
Lema	**Frecuencia**
mando	1
trampa	1
caza	1
carro	1
caperuza	1
compra	1
ruido	1
pulga	1
violonchelo	1
barco	1
póster	1
colonia	1
falda	1
pastel	1
zancadilla	1
rectángulo	1
expresión	1
banda	1
cuenta	1
color	1
edad	1
chicle	1
rosco	1
arbusto	1
jaula	1
nombre	1
rama	1
medio	1
lío	1
esponja	1
jefe	1
piel	1
bola	1
cerebro	1
leotardo	1
hospital	1
par	1
mancha	1
canción	1
habitación	1
i	1
fiesta	1
voz	1
colgante	1
veinticuatro	1
mayúscula	1
cebra	1
olivo	1
vocal	1
dos mil seis	1

Frecuencia de N	
Lema	**Frecuencia**
llave	1
pille pille	1
gana	1
naranja	1
cartón	1
otoño	1
frío	1
erre	1
goma	1
ciprés	1
gomina	1
abril	1
letra	1
cuatro	1
cocodrilo	1
oreja	1
gorra	1
aparato	1

Lema	Frecuencia
falta	1
cuenta cuentos	1
muñeca	1
pilla besos	1
dios	1
monstruo	1
efe	1
garaje	1
pera	1
zumo	1
ruiseñor	1
abuelito	1
inglés	1
enganchado	1
ducha	1
barriga	1
vampiro	1
nota	1
rastrillo	1

Frecuencia de V	
Lema	**Frecuencia**
ser	109
tener	102
estar	87
ir	69
saber	65
hacer	49
decir	32
caer	29
jugar	29
venir	28
dar	27
poner	26
haber	22
comprar	20
ver	19
llamar	17
regalar	17
salir	16
dejar	15
tocar	15
meter	14
parecer	14
comer	13
dormir	13
coger	13
pasar	12
tirar	11

Frecuencia de V	
Lema	**Frecuencia**
perder	10
echar	9
quitar	9
correr	8
acordar	8
sacar	8
quedar	7
romper	7
traer	7
leer	7
creer	7
acostar	7
soñar	7
llevar	7
levantar	7
gustar	6
vestir	6
morir	6
subir	6
portar	6
encontrar	6
abrir	6
lavar	6
trabajar	6
matar	5
ganar	5
vivir	5

Frecuencia de V	
Lema	**Frecuencia**
rapar	5
guardar	5
querer	5
encender	5
morder	5
pagar	4
desayunar	4
despertar	4
esconder	4
celebrar	4
llegar	4
cortar	4
cumplir	4
escribir	4
inventar	4
parar	4
luchar	3
tomar	3
pisar	3
ayudar	3
andar	3
quemar	3
buscar	3
gatear	3
bañar	3
contar	3
escapar	3
encantar	3
faltar	3
oír	3
callar	3
poder	3
pintar	3
entrar	2
apagar	2
coser	2
felicitar	2
cambiar	2
acabar	2
olvidar	2
enterrar	2
sujetar	2
peinar	2
beber	2
montar	2
soplar	2
bailar	2
hablar	2
juntar	2
disparar	2

Frecuencia de V	
Lema	**Frecuencia**
enseñar	2
fijar	2
limpiar	2
oler	2
enfadar	2
picar	2
acercar	1
enterar	1
arar	1
explotar	1
necesitar	1
recoger	1
escuchar	1
recortar	1
entrenar	1
espachurrar	1
convertir	1
pinchar	1
cenar	1
tostar	1
zampar	1
estirar	1
llorar	1
resfriar	1
saltar	1
adoptar	1
ocurrir	1
divertir	1
mirar	1
pasear	1
engañar	1
afinar	1
valer	1
aprender	1
preparar	1
conocer	1
empezar	1
colorear	1
preguntar	1
bajar	1
chupar	1
mezclar	1
probar	1
cazar	1
fregar	1
abalanzar	1
nacer	1
clavar	1
entender	1
entretener	1

Frecuencia de V	
Lema	**Frecuencia**
cansar	1
existir	1
pedir	1
sonar	1
salvar	1
volar	1
grapar	1
descansar	1
sumar	1
regañar	1
marcar	1
costar	1
hundir	1
pelear	1
usar	1
tumbar	1
sentar	1
trepar	1

Lema	Frecuencia
invitar	1
lanzar	1
volver	1
desaparecer	1
conducir	1
colar	1
atropellar	1
tardar	1
dibujar	1
aburrir	1
apuntar	1
vender	1
pegar	1
perecer	1
calzar	1
nevar	1
vomitar	1
soltar	1

Frecuencia de ADJ	
Lema	**Frecuencia**
grande	21
pequeño	9
de leche	7
roto	6
rojo	6
cojo	5
negro	5
malo	5
chino	5
bueno	5
verde	4
corto	4
pasado	3
digital	3
suave	3
alto	3
solo	2
mismo	2
colorado	2
nuevo	2
gris	2
gordo	2
marrón	2
último	2
chiquitín	2
tonto	2
secreto	2
de hierro	1

Frecuencia de ADJ	
Lema	**Frecuencia**
flaco	1
entero	1
superfuerte	1
responsable	1
loco	1
cazador	1
amarillo	1
fino	1
rosa	1
octavo	1
corporal	1
sentado	1
junto	1
feo	1
coloreado	1
flojo	1
chico	1
igual	1
único	1
abierto	1
mayor	1
malito	1
rico	1
bonito	1
guapo	1
perenne	1
rectangular	1
fácil	1

Frecuencia de ADJ	
Lema	**Frecuencia**
difícil	1
maestro	1
largo	1
americano	1
raro	1
doble	1
quieto	1
san	1

Lema	Frecuencia
blanco	1
mentiroso	1
de verdad	1
temprano	1
vacío	1
preferido	1
tranquilo	1
rápido	1

Apéndice C

Listados completos de frecuencias de sílabas

1. Grupo de edad: 3 años

Sílaba	Frecuencia
.a.	490
.i.	379
.no.	300
.si.	279
.ke.	208
.es.	202
.la.	193
.se.	189
.yo.	185
.ta.	174
.e.	160
.na.	156
.me.	156
.un.	151
.el.	149
.do.	148
.de.	147
.ko.	139
.pe.	138
.to.	131
.lo.	130
.te.	124
.ma.	123
.ba.	121
.mi.	115
.en.	114
.ra.	114
.ka.	113
.pa.	110
.o.	104
.u.	100
.ro.	98
.por.	97
.ya.	92
.go.	91
.le.	86
.sa.	85
.los.	73
.da.	69
.kon.	62
.so.	61
.ni.	59
.ne.	56
.be.	56
.ki.	53

Sílaba	Frecuencia
.mu.	50
.ti.	50
.Ro.	47
.gua.	46
.po.	45
.di.	45
.mos.	42
.ga.	42
.ten.	41
.Te.	41
.lu.	40
.bi.	40
.pi.	39
.xu.	37
.ai.	37
.bo.	37
.Co.	36
.xo.	35
.bien.	35
.ye.	35
.xa.	34
.tra.	33
.tro.	32
.tam.	30
.mo.	27
.Ter.	25
.Na.	25
.mui.	24
.al.	24
.Cos.	24
.tos.	22
.an.	22
.don.	22
.las.	22
.pre.	22
.tes.	22
.kie.	22
.ri.	21
.Ti.	21
.pis.	21
.kuan.	21
.ban.	20
.Ca.	20
.gran.	19
.su.	19

Sílaba	Frecuencia
.nes.	18
.No.	18
.tar.	18
.pues.	18
.bue.	18
.ge.	18
.tie.	18
.tas.	17
.Re.	17
.sen.	17
.tu.	16
.mas.	16
.li.	16
.ku.	16
.pri.	16
.Ci.	15
.ser.	15
.xe.	15
.fan.	15
.tan.	14
.kom.	14
.nos.	14
.gar.	14
.pin.	14
.yi.	13
.re.	13
.Ce.	13
.ran.	13
.dos.	13
.pu.	13
.ber.	13
.nas.	12
.Ri.	12
.boi.	12
.as.	12
.mon.	12
.Ta.	12
.kar.	11
.Tul.	11
.kan.	11
.pue.	11
.dor.	11
.bos.	11
.ron.	10
.bu.	10
.Ra.	10
.fo.	10
.bis.	10
.lue.	10
.tres.	10
.mis.	10

Sílaba	Frecuencia
.bier.	10
.blar.	10
.nio.	9
.bar.	8
.bel.	8
.bol.	8
.lar.	8
.lia.	8
.gu.	8
.ros.	8
.kris.	8
.soi.	8
.er.	8
.son.	8
.pien.	8
.gi.	8
.ben.	8
.mar.	8
.nu.	8
.gus.	7
.Rien.	7
.nin.	7
.ue.	7
.per.	7
.xi.	7
.lio.	7
.tin.	7
.xue.	7
.pun.	7
.die.	7
.del.	7
.bir.	7
.bes.	6
.xio.	6
.gun.	6
.kas.	6
.Cas.	6
.len.	6
.dri.	6
.par.	6
.Rom.	6
.gan.	6
.kin.	6
.man.	6
.seis.	6
.les.	6
.or.	5
.nue.	5
.sien.	5
.os.	5
.dar.	5

Sílaba	Frecuencia
.ar.	5
.Tio.	5
.fe.	5
.fru.	5
.pan.	5
.sal.	5
.Car.	5
.kuen.	5
.sie.	5
.mue.	5
.tio.	5
.Tie.	5
.on.	5
.ton.	5
.kuer.	5
.dan.	5
.fue.	5
.pei.	5
.kam.	4
.kua.	4
.des.	4
.in.	4
.ui.	4
.Ton.	4
.fri.	4
.yes.	4
.pie.	4
.bie.	4
.gos.	4
.kla.	4
.men.	4
.res.	4
.pro.	4
.pra.	4
.nuel.	4
.xos.	4
.nun.	4
.Ne.	4
.mor.	4
.Tion.	4
.xes.	4
.pon.	4
.Tir.	4
.Ru.	4
.kien.	4
.das.	4
.dieT.	4
.siem.	4
.Nos.	4
.kue.	3
.xer.	3

Sílaba	Frecuencia
.sio.	3
.Tes.	3
.Nor.	3
.dre.	3
.xim.	3
.bru.	3
.tor.	3
.ter.	3
.du.	3
.dia.	3
.tal.	3
.tre.	3
.fra.	3
.To.	3
.prar.	3
.dis.	3
.den.	3
.mer.	3
.Ces.	3
.Ran.	3
.blo.	3
.Tin.	3
.mio.	3
.gra.	3
.fi.	3
.yor.	3
.Tos.	3
.blan.	3
.beT.	3
.yar.	3
.xar.	3
.Ni.	3
.yas.	3
.rar.	3
.fut.	3
.gre.	3
.xia.	3
.glo.	3
.pes.	3
.guar.	2
.ras.	2
.Can.	2
.pla.	2
.fer.	2
.gui.	2
.lum.	2
.ru.	2
.tru.	2
.sar.	2
.kual.	2
.fuer.	2

Sílaba	Frecuencia
.gru.	2
.yos.	2
.sue.	2
.lan.	2
.fal.	2
.pas.	2
.puT.	2
.em.	2
.yer.	2
.tros.	2
.lox.	2
.Tar.	2
.RoT.	2
.dio.	2
.lim.	2
.aT.	2
.sin.	2
.fui.	2
.bas.	2
.bri.	2
.riel.	2
.gres.	2
.sol.	2
.doi.	2
.pul.	2
.gien.	2
.pier.	2
.Tien.	2
.kre.	2
.kal.	2
.tri.	1
.kles.	1
.xau.	1
.fu.	1
.fren.	1
.Rei.	1
.dro.	1
.guan.	1
.biar.	1
.kau.	1
.bres.	1
.noT.	1
.din.	1
.bal.	1
.gel.	1
.ner.	1
.puen.	1
.rien.	1
.Cu.	1
.pios.	1
.grie.	1

Sílaba	Frecuencia
.bio.	1
.sos.	1
.lir.	1
.gro.	1
.lon.	1
.toi.	1
.pol.	1
.oi.	1
.dios.	1
.rin.	1
.Nas.	1
.piar.	1
.klau.	1
.bre.	1
.gas.	1
.bor.	1
.ir.	1
.kos.	1
.kui.	1
.xie.	1
.fa.	1
.nom.	1
.Ron.	1
.gual.	1
.yu.	1
.tran.	1
.pel.	1
.pal.	1
.fre.	1
.gros.	1
.tir.	1
.mons.	1
.bren.	1
.due.	1
.bus.	1
.rio.	1
.lie.	1
.bra.	1
.piT.	1
.pia.	1
.kor.	1
.puer.	1
.drian.	1
.guai.	1
.fle.	1
.truos.	1
.nir.	1
.mie.	1
.lor.	1
.fin.	1

Sílaba	Frecuencia
.mal.	1
.peT.	1
.som.	1
.yan.	1
.kri.	1
.sel.	1
.Rue.	1
.fun.	1

Sílaba	Frecuencia
.dien.	1
.am.	1
.pos.	1
.dad.	1
.xor.	1
.Tu.	1
.sas.	1

2. Grupo de edad: 4 años

Sílaba	Frecuencia
.a.	504
.ke.	395
.i.	391
.no.	280
.si.	250
.mi.	231
.la.	223
.es.	214
.de.	204
.na.	188
.se.	187
.pa.	180
.u.	176
.ma.	169
.el.	169
.ta.	156
.ra.	155
.me.	154
.ko.	154
.un.	154
.en.	152
.do.	152
.lo.	144
.ka.	142
.yo.	142
.pe.	136
.o.	125
.e.	121
.to.	117
.ba.	116
.ro.	104
.sa.	101
.go.	99
.te.	96
.por.	93
.le.	91
.ni.	81
.bi.	79

Sílaba	Frecuencia
.los.	78
.kon.	76
.ya.	75
.pues.	75
.be.	70
.al.	65
.Te.	60
.ten.	57
.da.	57
.di.	57
.dre.	56
.ne.	56
.dos.	54
.so.	53
.ti.	52
.mo.	51
.bo.	50
.bien.	49
.ga.	44
.tro.	44
.No.	43
.tu.	42
.tam.	42
.xo.	41
.nos.	41
.mos.	40
.tra.	40
.nes.	40
.las.	40
.ki.	40
.tie.	38
.bue.	37
.ri.	37
.mas.	36
.li.	36
.po.	35
.ye.	32
.bes.	32

Sílaba	Frecuencia
.ber.	31
.kuan.	31
.bal.	30
.ai.	30
.xu.	27
.Ti.	27
.xa.	27
.Ro.	26
.Co.	26
.su.	25
.an.	25
.Ca.	25
.gran.	24
.xe.	24
.Ra.	24
.as.	24
.mui.	23
.del.	23
.nas.	23
.kla.	22
.gu.	22
.Ce.	22
.tos.	21
.fui.	20
.Ta.	20
.ron.	20
.kar.	20
.er.	19
.Tin.	19
.tor.	19
.re.	19
.pri.	19
.Nos.	19
.pi.	18
.kum.	18
.fe.	18
.mar.	18
.des.	18
.mu.	18
.tas.	18
.tar.	17
.Na.	17
.pin.	17
.tan.	17
.kien.	16
.tes.	16
.xor.	16
.ban.	16
.Re.	16
.xos.	16
.lu.	15

Sílaba	Frecuencia
.tres.	15
.gua.	14
.Tes.	14
.boi.	14
.tio.	14
.kan.	14
.Ter.	14
.kua.	14
.man.	14
.Ci.	14
.ple.	13
.son.	13
.fue.	12
.gan.	12
.gar.	12
.kom.	12
.bol.	11
.don.	11
.dra.	11
.dis.	11
.kie.	11
.ku.	11
.gel.	11
.pie.	11
.bier.	10
.seis.	10
.gus.	10
.kas.	10
.fut.	10
.ran.	9
.xi.	9
.sas.	9
.sol.	9
.xue.	9
.die.	9
.res.	9
.lue.	9
.mal.	9
.Tien.	9
.les.	9
.pan.	9
.mis.	9
.To.	8
.sie.	8
.bar.	8
.tre.	8
.on.	8
.pis.	8
.ton.	8
.Ri.	8
.sus.	8

Sílaba	Frecuencia
.kue.	8
.xau.	8
.yi.	8
.pu.	8
.bein.	8
.dras.	8
.men.	7
.xel.	7
.min.	7
.pier.	7
.kuen.	7
.tros.	7
.nu.	7
.nue.	7
.sal.	7
.Ros.	7
.dor.	7
.em.	7
.fi.	7
.gon.	7
.den.	7
.gre.	7
.par.	7
.mer.	6
.fa.	6
.gos.	6
.gi.	6
.xen.	6
.sen.	6
.lan.	6
.gas.	6
.beT.	6
.ru.	6
.len.	6
.bra.	6
.rio.	6
.Cu.	6
.bre.	6
.ser.	6
.bu.	6
.Tul.	6
.ner.	5
.bis.	5
.kre.	5
.ren.	5
.mon.	5
.oi.	5
.dan.	5
.rin.	5
.nal.	5
.xer.	5

Sílaba	Frecuencia
.blan.	5
.yer.	5
.xar.	5
.xio.	5
.kos.	5
.ros.	5
.pas.	5
.dar.	5
.ir.	5
.pue.	5
.pra.	4
.gro.	4
.ge.	4
.truo.	4
.Ru.	4
.plan.	4
.mons.	4
.os.	4
.dres.	4
.simp.	4
.ben.	4
.kin.	4
.plei.	4
.nio.	4
.pio.	4
.mil.	4
.lum.	4
.pon.	4
.gor.	4
.soi.	4
.ter.	4
.kam.	4
.tin.	4
.Nor.	4
.dad.	4
.per.	3
.bla.	3
.drid.	3
.toi.	3
.bel.	3
.bie.	3
.sons.	3
.bas.	3
.mio.	3
.yor.	3
.Tir.	3
.fo.	3
.sue.	3
.kor.	3
.sio.	3
.fri.	3

Sílaba	Frecuencia
.mier.	3
.ol.	3
.kuer.	3
.tri.	3
.dul.	3
.Rar.	3
.Tar.	3
.pre.	3
.sos.	3
.bos.	3
.xua.	3
.niel.	3
.rar.	3
.ul.	3
.pro.	3
.piT.	3
.yos.	3
.Car.	3
.Cas.	3
.bios.	3
.buel.	3
.Res.	3
.Ces.	3
.bro.	3
.Cos.	2
.Tion.	2
.Ras.	2
.duer.	2
.nia.	2
.xun.	2
.Tia.	2
.gra.	2
.Nar.	2
.bio.	2
.ses.	2
.pla.	2
.Ren.	2
.tras.	2
.fur.	2
.dios.	2
.nues.	2
.glan.	2
.Tos.	2
.pran.	2
.lon.	2
.dri.	2
.flo.	2
.fan.	2
.lia.	2
.xas.	2
.san.	2

Sílaba	Frecuencia
.lie.	2
.sar.	2
.bir.	2
.mue.	2
.nar.	2
.fas.	2
.ras.	2
.gri.	2
.dia.	2
.leT.	2
.lio.	2
.Ni.	2
.xan.	2
.sien.	2
.bil.	2
.nir.	2
.du.	2
.kri.	2
.fra.	2
.Rien.	2
.kui.	2
.fal.	2
.gual.	2
.kris.	2
.pien.	2
.kol.	2
.ur.	2
.guar.	2
.yol.	2
.xon.	2
.dio.	2
.yas.	2
.ar.	2
.lin.	2
.xai.	2
.nen.	2
.ue.	2
.mie.	2
.glo.	2
.or.	1
.blar.	1
.Nas.	1
.kues.	1
.tien.	1
.yen.	1
.pur.	1
.Ron.	1
.Tie.	1
.pei.	1
.Ran.	1
.gol.	1

Sílaba	Frecuencia
.tal.	1
.lir.	1
.fik.	1
.Tan.	1
.der.	1
.nan.	1
.paT.	1
.truos.	1
.biar.	1
.in.	1
.dai.	1
.brir.	1
.diar.	1
.Ne.	1
.xim.	1
.fraT.	1
.puen.	1
.bres.	1
.piar.	1
.plus.	1
.xuan.	1
.pies.	1
.gais.	1
.Ton.	1
.trol.	1
.brio.	1
.Ten.	1
.Cus.	1
.bru.	1
.lau.	1
.eks.	1
.miau.	1
.ples.	1
.pol.	1
.bug.	1
.trar.	1
.dok.	1
.tios.	1
.bri.	1
.fuen.	1
.sies.	1
.prar.	1
.Tio.	1
.pos.	1
.bor.	1
.kle.	1
.sor.	1
.pian.	1
.dien.	1
.nom.	1

Sílaba	Frecuencia
.Rom.	1
.pal.	1
.trein.	1
.bril.	1
.plir.	1
.s.	1
.lien.	1
.bus.	1
.pren.	1
.pel.	1
.krip.	1
.yus.	1
.plio.	1
.blok.	1
.tel.	1
.ie.	1
.Tu.	1
.mir.	1
.pir.	1
.fom.	1
.due.	1
.pli.	1
.iT.	1
.dorm.	1
.Rue.	1
.suel.	1
.nin.	1
.uen.	1
.xais.	1
.Can.	1
.dro.	1
.beis.	1
.kier.	1
.bros.	1
.tir.	1
.nie.	1
.gen.	1
.nun.	1
.nald.	1
.dieT.	1
.us.	1
.plo.	1
.lar.	1
.das.	1
.bian.	1
.siem.	1
.sion.	1
.tis.	1
.tris.	1

3. Grupo de edad: 5 años

Sílaba	Frecuencia
.a.	707
.i.	594
.ke.	463
.no.	381
.si.	318
.es.	288
.se.	283
.la.	273
.de.	269
.ta.	243
.me.	242
.do.	239
.ko.	233
.na.	227
.mi.	221
.ka.	208
.el.	205
.le.	202
.ma.	201
.en.	191
.pa.	189
.lo.	189
.ba.	186
.di.	182
.yo.	177
.to.	166
.un.	162
.u.	162
.pe.	158
.ro.	156
.te.	155
.o.	147
.ra.	146
.e.	144
.sa.	134
.por.	131
.los.	124
.kon.	113
.go.	100
.da.	99
.ya.	98
.mos.	95
.pues.	91
.ne.	89
.so.	89
.dos.	86
.be.	79
.mo.	78

Sílaba	Frecuencia
.Te.	77
.kan.	76
.ri.	73
.ga.	72
.bi.	71
.ti.	71
.al.	71
.dre.	67
.Co.	64
.tam.	63
.bien.	62
.po.	61
.Ra.	56
.pi.	55
.tro.	55
.las.	53
.kar.	53
.nos.	52
.tie.	51
.ni.	49
.mas.	48
.mu.	47
.ki.	47
.No.	46
.bue.	45
.Ro.	44
.tas.	43
.su.	43
.ten.	42
.xo.	42
.kuan.	42
.xa.	41
.re.	40
.bo.	40
.Ce.	40
.as.	38
.tra.	38
.an.	37
.Ta.	37
.Tes.	36
.des.	35
.ye.	34
.Tin.	34
.xu.	34
.nes.	34
.Ci.	33
.tu.	33
.li.	32

Sílaba	Frecuencia
.mar.	32
.Ten.	32
.Re.	31
.bes.	30
.fue.	29
.les.	28
.Ti.	28
.er.	27
.sol.	27
.ar.	26
.ton.	26
.men.	26
.ter.	25
.tres.	25
.tar.	25
.kua.	24
.tes.	24
.ul.	23
.ban.	22
.ai.	22
.kos.	22
.kom.	22
.gu.	21
.gran.	21
.Na.	21
.nas.	21
.kuen.	20
.mui.	20
.ben.	20
.ber.	20
.gar.	20
.dio.	20
.lan.	20
.man.	19
.boi.	19
.sas.	19
.Ri.	19
.tos.	19
.lar.	19
.ron.	18
.bre.	18
.del.	18
.eu.	18
.per.	17
.kas.	17
.pue.	17
.xe.	16
.fe.	16
.xue.	16
.lue.	16
.Nos.	16

Sílaba	Frecuencia
.niel.	16
.seis.	16
.pri.	16
.tan.	15
.bol.	15
.fi.	15
.tio.	14
.oi.	14
.son.	14
.bier.	14
.bra.	13
.par.	13
.nue.	13
.Ca.	13
.ge.	12
.beT.	12
.ru.	12
.sie.	12
.lu.	12
.Cas.	12
.yi.	12
.kor.	12
.tre.	11
.pre.	11
.ir.	11
.To.	11
.rio.	11
.dor.	11
.xor.	11
.min.	11
.Cu.	11
.pra.	11
.der.	11
.ku.	10
.Ter.	10
.kum.	10
.lau.	10
.san.	10
.gual.	10
.res.	10
.kre.	9
.Tien.	9
.gos.	9
.gua.	9
.bar.	9
.ser.	9
.kien.	9
.Res.	9
.mier.	9
.xio.	9
.lio.	9

Sílaba	Frecuencia
.gan.	9
.pie.	9
.im.	9
.prar.	9
.on.	9
.kuer.	8
.Cos.	8
.yos.	8
.xos.	8
.Ros.	8
.lia.	8
.fu.	8
.kie.	8
.fan.	8
.yu.	8
.tros.	8
.bas.	8
.bein.	8
.ok.	8
.dieT.	8
.bio.	7
.ros.	7
.kin.	7
.ran.	7
.xas.	7
.bie.	7
.mer.	7
.gun.	7
.fut.	7
.kam.	7
.kla.	7
.yer.	7
.sal.	7
.don.	7
.lon.	7
.luis.	7
.Ni.	7
.gra.	7
.ple.	7
.tiem.	6
.ras.	6
.sos.	6
.xan.	6
.pas.	6
.Rom.	6
.os.	6
.tras.	6
.xi.	6
.gol.	6
.dar.	6
.duer.	6

Sílaba	Frecuencia
.bri.	6
.mon.	6
.ner.	6
.Ces.	6
.bal.	6
.gus.	6
.siem.	6
.pin.	5
.drid.	5
.fa.	5
.fui.	5
.guar.	5
.fal.	5
.bu.	5
.pus.	5
.bai.	5
.die.	5
.bir.	5
.Rien.	5
.mis.	5
.gas.	5
.Car.	5
.pro.	5
.tal.	5
.den.	4
.dri.	4
.em.	4
.diks.	4
.sio.	4
.bus.	4
.Ras.	4
.Tio.	4
.gre.	4
.Tos.	4
.nal.	4
.mir.	4
.pu.	4
.nun.	4
.Cin.	4
.bor.	4
.mal.	4
.Rei.	4
.aun.	4
.tien.	4
.yas.	4
.or.	4
.das.	4
.tri.	4
.rar.	4
.dis.	4
.Tiem.	4

Sílaba	Frecuencia
.dra.	4
.kri.	4
.sue.	4
.klaus.	4
.flo.	4
.dien.	3
.sar.	3
.is.	3
.dad.	3
.sus.	3
.blan.	3
.tem.	3
.tir.	3
.xun.	3
.gon.	3
.nues.	3
.tin.	3
.tel.	3
.xes.	3
.nan.	3
.puer.	3
.Ne.	3
.Tas.	3
.tor.	3
.fas.	3
.doi.	3
.Tir.	3
.Ru.	3
.ue.	3
.pul.	3
.yar.	3
.flor.	3
.bom.	3
.plai.	3
.sua.	3
.bel.	3
.yon.	3
.mor.	3
.nio.	3
.pal.	3
.duen.	3
.dras.	3
.nir.	3
.gro.	3
.brio.	3
.yor.	3
.Cen.	3
.pun.	3
.pio.	3
.kual.	3
.kai.	2

Sílaba	Frecuencia
.ble.	2
.bia.	2
.xar.	2
.biem.	2
.mue.	2
.ol.	2
.Til.	2
.xien.	2
.kue.	2
.sep.	2
.eks.	2
.dan.	2
.Rui.	2
.blo.	2
.nen.	2
.soi.	2
.gor.	2
.Tia.	2
.Nor.	2
.muer.	2
.lim.	2
.Rer.	2
.xams.	2
.sen.	2
.pon.	2
.Tion.	2
.faus.	2
.bos.	2
.fle.	2
.dro.	2
.pli.	2
.xim.	2
.som.	2
.bis.	2
.Rue.	2
.mio.	2
.pei.	2
.mun.	2
.in.	2
.gi.	2
.Rek.	2
.rin.	2
.dios.	2
.plar.	2
.ses.	2
.plus.	2
.plo.	2
.nie.	2
.fer.	2
.Ron.	2
.buel.	2

Sílaba	Frecuencia
.len.	2
.simp.	2
.mien.	1
.guan.	1
.bras.	1
.sin.	1
.Can.	1
.pia.	1
.trak.	1
.pras.	1
.tram.	1
.kol.	1
.mak.	1
.ral.	1
.bril.	1
.buar.	1
.fre.	1
.yet.	1
.uf.	1
.friar.	1
.tran.	1
.riT.	1
.nies.	1
.rir.	1
.bam.	1
.kuar.	1
.kop.	1
.fo.	1
.yes.	1
.trai.	1
.nu.	1
.dia.	1
.xua.	1
.fos.	1
.Tan.	1
.bil.	1
.n.	1
.dria.	1
.trar.	1
.Tam.	1
.kres.	1
.rios.	1
.mil.	1
.lie.	1
.gles.	1
.fla.	1
.pres.	1
.lum.	1
.lin.	1
.ria.	1
.gel.	1

Sílaba	Frecuencia
.yiu.	1
.pan.	1
.tau.	1
.bian.	1
.Tu.	1
.bro.	1
.sion.	1
.au.	1
.nar.	1
.ler.	1
.pies.	1
.gri.	1
.Ton.	1
.ei.	1
.kles.	1
.kes.	1
.piel.	1
.noT.	1
.Nas.	1
.Tie.	1
.kal.	1
.xau.	1
.mons.	1
.fri.	1
.ren.	1
.dur.	1
.too.	1
.duT.	1
.Tar.	1
.lien.	1
.sons.	1
.truos.	1
.gen.	1
.pier.	1
.sor.	1
.nia.	1
.dok.	1
.geim.	1
.toi.	1
.meT.	1
.pis.	1
.brir.	1
.prin.	1
.pren.	1
.lis.	1
.ui.	1
.nom.	1
.boT.	1
.kui.	1
.miel.	1
.mag.	1

Sílaba	Frecuencia
.trein.	1
.pos.	1
.fren.	1
.far.	1
.lou.	1
.bat.	1
.tren.	1
.fies.	1
.yus.	1
.gros.	1
.pron.	1
.tum.	1
.fuer.	1
.pum.	1
.piar.	1
.mie.	1
.mion.	1

Sílaba	Frecuencia
.Non.	1
.pla.	1
.kris.	1
.xel.	1
.ek.	1
.dop.	1
.brar.	1
.sis.	1
.fin.	1
.diT.	1
.sel.	1
.teks.	1
.ie.	1
.gris.	1
.xios.	1
.Nan.	1