Extracción de unidades distintivas en adultos y niños de un corpus de lengua oral espontánea[1]

Marta Garrote, José M. Guirao, y Antonio Moreno

Universidad Autónoma de Madrid, Dpto. de Lingüística, Laboratorio de Lingüística Informática
Campus de Cantoblanco, 28049-Madrid
marta@maria.lllf.uam.es

Resumen

Presentamos en este artículo una extensión del método propuesto en Guirao *et al.* (2006). Nuestro corpus está compuesto por transcripciones de grabaciones de lengua oral, en las que se incluyen cabeceras con información sociolingüística sobre los participantes. Se utiliza un programa para asociar cada palabra del texto con el hablante y sus características sociolingüísticas. De esta forma, se generan subcorpus para la variable edad. En este experimento se contrasta la lengua entre adultos y niños.

Se aplica el test de ratio de verosimilitud (Dunning 1993) para identificar las distintas palabras y lemas en un corpus dado. Esta técnica estadística asume una distribución binomial, que resulta más apropiada para unidades distintivas y características. En otras palabras, detecta no las palabras más frecuentes usadas por los niños de una determinada edad, sino las palabras que ese grupo utiliza de forma más específica y que en el resto de grupos no aparecen.

Los resultados, aunque claramente insuficientes y probablemente parciales, proporcionan una nueva perspectiva a los estudios empíricos sobre variación léxica, combinando el trabajo en corpus orales con herramientas computacionales para el análisis de datos.

Palabras clave: corpus de lengua oral infantil, test de ratio de verosimilitud (Dunning 1993).

Summary

In this paper, we present an extension of the method proposed in Guirao *et al.* (2006). Our corpora consist of transcriptions of spoken recordings along with a header including metadata with socio-linguistic information about the speakers. A program is used to associate every word in the text with the speaker and his or her socio-linguistic features. This way, we generate sub-corpora for age variables. In this experiment, we consider adult versus child speech.

In order to identify the distinctive words and lemmas for a given sub-corpus, the log-likelihood ratio test (Dunning 1993) is applied. This statistical technique assumes a binomial distribution, which is more appropriate for rare but distinctive units. In other words, it detects not the most frequent words used by the children of a given age, but the words that this group uses more specifically and the other groups do not.

The results, although clearly insufficient and probably biased, provide a new perspective to the empirical studies on lexical variation, combining spoken corpora with computational tools for exploring data.

Keywords: corpus of spontaneous child speech, log-likelihood ratio test (Dunning 1993).

[1] Esta investigación ha sido parcialmente financiada por la Comunidad de Madrid en el marco del convenio MAVIR (S-0505/TIC/0267) y por el proyecto BRAVO-RL del MEC-CICYT (TIN2007-67407-C03-02).

Résumé

On montre dans cet article un approfondissement de la méthode proposée dans Guirao *et al.* (2006). Notre corpus est composé des différentes transcriptions d'enregistrements de la langue orale, qui comprennent en plus des entêtes avec l'information sociolinguistique des participants. On utilise un logiciel pour mettre en relation chaque mot du texte avec le locuteur et ses caractéristiques sociolinguistiques. Ainsi, on produit des subcorpus pour la variable « âge ». Dans cette expérience, on oppose la langue des adults à celle des enfants.
On applique le test de ratio de vraisemblance (Dunning 1993) pour identifier les différents mots et les lemmes d'un corpus donné. Cette technique statistique entraîne une distribution binomiale, qui devient la plus appropriée pour les unités distinctives et caractéristiques. En d'autres mots, elle ne montre pas les mots les plus fréquents utilisés par les enfants d'un âge détérminé, mais les mots que ce groupe utilise d'une façon plus spécifique et qui n'apparaissent pas dans le reste des groupes.
Les résultats, même s'ils sont nettement insuffisants et probablement partials, nous offrent une nouvelle perspective aux études empiriques sur la variation lexicale, en combinant le travail sur corpus oraux avec des outils informatiques pour l'analyse des données.

Mots clés : corpus de la langue orale des enfants, test de ratio de vraisemblance (Dunning 1993).

Tabla de Contenidos

1. CHIEDE, un Corpus de Habla Infantil Espontánea del Español

El Corpus de Habla Infantil Espontánea del Español, CHIEDE, 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. Su principal característica es la espontaneidad de las interacciones en él recogidas: los textos son grabaciones de situaciones comunicativas en su contexto natural. El recurso se presenta en diferentes formatos: una transcripción ortográfica, una transcripción fonológica automática, una versión etiquetada en XML y el alineamiento del sonido con el texto. Además, se facilitan los resultados obtenidos tras la extracción, mediante métodos estadísticos, de información de los textos anotados.

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 (http://www.lllf.uam.es/chiede) 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.

1.1 Diferencias entre una base de datos oral y un corpus de lengua oral

En primer lugar, es necesario hacer una distinción entre los diferentes tipos de recursos lingüísticos orales. La mayoría de los recursos lingüísticos disponibles en la actualidad son **bases de datos orales**: colecciones de grabaciones de alta calidad y transcripciones fonéticas detalladas de muestras de lengua oral producidas en espacios controlados (comúnmente servicios telefónicos). Estas bases de datos orales se emplean generalmente para entrenar y evaluar sistemas de voz y se desarrollan por y para la industria de la ingeniería lingüística. Su principal objetivo es servir como base para reconocer y producir habla en dominios restringidos y predecibles. En la mayoría de los casos, esas bases de datos contienen muchos ejemplos de la misma palabra (es decir, muchos *tokens* del mismo *type*). Normalmente, los enunciados se preparan y producen por locutores profesionales. La calidad acústica de la grabación es esencial. Las bases de datos orales normalmente facilitan descripciones fonéticas detalladas, incluyendo disfluencias, ruidos y otros sonidos. En general, estas bases de datos reflejan el registro estándar, y otras variantes (dialectos, jergas) se representan de forma pobre. Ejemplos de esto son SpeechDat (LRE-63314, Infrastructure for Spoken Language Resources), SpeechDat II (LRE2-4001, Speech Databases for the Creation of Voice Driven Teleservices), que han establecido un estándar para este tipo de recurso.

Por otra parte, los **corpus de lengua oral espontánea** son típicamente colecciones de una amplia variedad de registros orales. Estos corpus son recogidos principalmente para análisis y aplicaciones lingüísticos (enseñanza de la lengua, gramáticas y diccionarios). En dichos corpus la calidad acústica no es esencial. Lo que es importante es que los textos reflejen la mayor variación posible y que el hablante se comporte de forma espontánea. En algunos casos, estos corpus sólo se centran en un registro dado, por ejemplo, un dialecto o el habla infantil. Una diferencia importante con respecto a las bases de datos orales es la transcripción: los corpus de lengua oral espontánea normalmente son menos precisos en las partes acústicas y fonéticas. Por el contrario, incluyen información detallada sobre el contexto y los hablantes. Estos corpus se usan principalmente para

análisis sociolingüísticos, de tipología textual o psicolingüísticos. Dos ejemplos son los corpus CHILDES y London-Lund.

Nuestro corpus pertenecería al segundo tipo, en el que las muestras recogidas son absolutamente naturales y por ello es necesario tener en cuenta ciertas características. CHIEDE es un corpus de lengua oral espontánea, pero también posee algunos rasgos distintivos, como su calidad acústica o el alineamiento de la transcripción con el sonido original. Esto último es útil tanto para verificar la precisión de la transcripción como para la enseñanza y otros propósitos de investigación aplicada.

1.2 Diseño del corpus.

Para grabar el corpus CHIEDE, contamos con la colaboración de un colegio público español que aprobó nuestra intención de grabar el habla de los niños de los grupos de Educación Infantil (de 3 a 6 años de edad). CHIEDE es un corpus transversal, formado por tres grupos de sujetos divididos por edades (de 3 a 4 años, de 4 a 5 y de 5 a 6).

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

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 y contextual. Además de los archivos de audio y texto, se incluyen otros dos tipos de archivos adicionales: 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.

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

Figura 1: Medidas de CHIEDE

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:

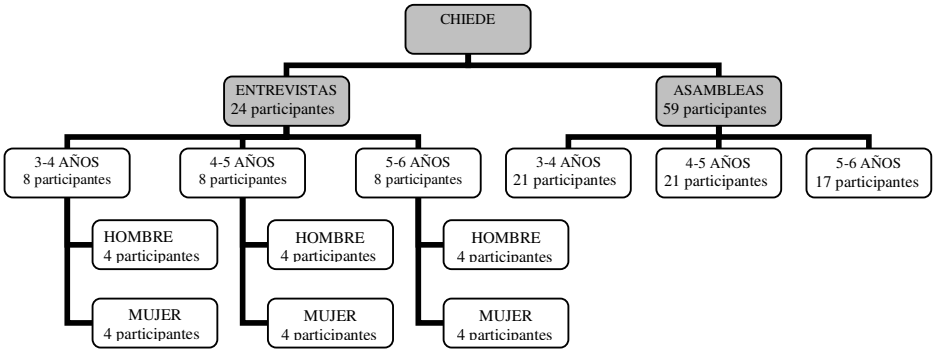

Figura 2: Diseño de CHIEDE

1.3 Formato.

Antes de transcribir las grabaciones, es necesario establecer un criterio de anotación consistente. Para ello, se siguió el sistema de anotación desarrollado para el proyecto C-ORAL-ROM, que a su vez está basado en el formato CHAT. Además, se proporciona una conversión a XML, que garantiza una fácil comprensión mediante su correspondiente DTD. El uso combinado de XML y la DTD asegura que cada texto cumple los mismos requisitos. De esta forma, se consigue una uniformidad textual a lo largo de todo el corpus.

El formato se divide en *cabecera* (con los metadatos) y *transcripción*. Los datos incluidos en la cabecera proporcionan una rica información sociolingüística al texto. La transcripción se estructura por turnos, introducidos por un código de tres letras mayúsculas que identifican al participante. La transcripción es ortográfica e incluye etiquetas que marcan las disfluencias, ruidos, solapamientos y unidades prosódicas. El etiquetado morfosintáctico se facilita en un nivel separado. La Figura 3 muestra un fragmento del texto.

@Title: Jorge y Marta
@File: JOR4
@Participants: JOR, child, (man, 5:0, 1, Ciudad Real)
TEA, adult, (woman, B, 3, Madrid)
@Birth_Date: 20/02/2001
@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 (child not-known adult)
@Lenght: 13´39"
@Words: 2097
@Acoustic_quality: A
@Transcriber: Marta
@Revisor: Ana and Marta
@Comments: JOR (middle class; birth order: 1st)

*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 ///

Figura 3: Fragmento de CHIEDE

1.4 Otros aspectos relevantes.

CHIEDE cumple con todas las características de un corpus moderno. Sus aspectos más significativos se resumen a continuación.

1.4.1 El aspecto legal

Durante la década de los noventa la legislación sobre los derechos de autor y la privacidad se sometió a cambios en muchos países europeos. En los corpus de lengua oral, la ley se aplica al grabar a individuos o al usar documentos sonoros de los medios de comunicación. En el primer caso, los hablantes tienen el derecho a preservar su intimidad, y deben dar su autorización expresa para que su discurso pueda ser transcrito y publicado. Para garantizar la espontaneidad, que es esencial para nuestros propósitos, el procedimiento consiste en pedir a cada participante que firme una autorización *después* de la grabación. Si un hablante rehúsa dar su consentimiento, entonces se desecha la grabación. El derecho a la intimidad se aplica a toda grabación realizada en un contexto privado, pero no a aquellas que se llevan a cabo en una situación pública (una conferencia, un discurso político, un sermón).

El hecho de que nuestro corpus vaya a ser publicado en una página web 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.

1.4.2 La calidad acústica

El corpus CHIEDE se ha recopilado partiendo de cero, es decir, que todas las grabaciones se han llevado a cabo al mismo tiempo al inicio del proyecto. De esta forma, nos hemos asegurado tanto de obtener todos los permisos de los participantes como de utilizar el equipo de grabación adecuado.

Todos las grabaciones se han realizado con una DAT Tascam (modelo DA-P1) y dos micrófonos unidireccionales. Con posterioridad, la fuente se convierte en un archivo WAV, mono, 16 bit, 22.050 Hz, mediante un puerto SPDIF en una Sound Blaster Live Platinum 5.1, usando el software WaveLab. La calidad acústica es esencial para la utilización del corpus en tecnologías del habla e ingeniería lingüística.

1.4.3 La anotación lingüística

Los corpus incrementan su valor según los niveles de anotación que aporten. Etiquetar un corpus de lengua oral espontánea es una tarea ligeramente diferente de la de etiquetar un corpus de lengua escrita (Uchimoto *el al*. 2002). La diferencia no recae en la información etiquetada sino en la menor eficiencia de los etiquetadores cuando se aplican a la lengua oral. Por ejemplo, algunos etiquetadores morfosintácticos están habitualmente entrenados en textos escritos, que muestran un orden de palabras más estable y determinado. Por el contrario, los corpus de lengua oral espontánea son altamente flexibles en lo que a su orden de palabras se refiere. Además, estos muestran repeticiones, reinicios, solapamientos y otras características de la sintaxis oral sobre las que hay que entrenar al programa de manera específica.

El léxico es también diferente. Se pueden encontrar muchas palabras que no están incluidas en los diccionarios (impresos), porque son neologismos o pertenecen a un registro informal, o simplemente porque son pronunciaciones incorrectas.

La anotación de CHIEDE incluye una completa lematización y un análisis morfosintáctico automáticos. Los programas que se han utilizado para ello se han desarrollado en el Laboratorio de Lingüística Informática de la UAM (Moreno y Guirao 2003).

1.4.4 La validación

Verificar la fiabilidad de los datos se ha convertido en un tema habitual en los últimos años. Los usuarios de recursos lingüísticos quieren saber cómo han sido recopilados los recursos y su grado de precisión y fidelidad.

CHIEDE ha sido sometido a una validación interna llevada a cabo por lingüistas. Cada texto se somete a cinco fases de evaluación: transcripción, primera revisión, etiquetado prosódico, segunda revisión y alineamiento del texto y el sonido. Al menos dos lingüistas revisan cada texto. Un programa verifica errores de formato, espacios en blanco, erratas, etiquetas mal formadas, etc. Por lo tanto, contenido y forma son validados exhaustivamente, garantizando que la transcripción es fiel a la fuente de sonido. Además, el alineamiento del sonido con el texto es la mejor forma de validar la transcripción: cualquier discrepancia entre el discurso real y su transcripción se detectaría fácilmente.

2. La herramienta computacional

Para el tratamiento y la extracción de datos de un texto no es suficiente con tener la transcripción ortográfica y trabajar sobre ella. En el LLI-UAM se han desarrollado herramientas informáticas para transformar el formato de plano del texto de una transcripción en un esquema de etiquetado más apropiado para relacionar los metadatos con los elementos léxicos y calcular las estadísticas correspondientes. En este apartado explicaremos dicho proceso, y posteriormente, veremos cómo se ha realizado el proceso de extracción de datos.

2.1 Uso de un corpus etiquetado en XML para relacionar metadatos y elementos lingüísticos

La anotación original de CHIEDE ha sido diseñada para registrar una amplia variedad de fenómenos, incluyendo los acústicos (marcas prosódicas, ruidos, etc.) que pueden ser usados por la comunidad de la tecnología del habla.

Nuestro objetivo en este experimento es buscar unidades léxicas significativas en dos subcorpus: uno de lenguaje adulto y otro de lenguaje infantil. El primer paso consiste en la separación de cada turno en enunciados para prevenir grupos de palabras mal formados. Esta tarea es similar a la tokenización en los corpus escritos. Esta división en enunciados también es necesaria para delimitar el contexto que el etiquetador morfosintáctico utiliza para la desambiguación.

Un programa genera un corpus etiquetado de nuevo con sólo una etiqueta: UNIT (enunciado), con atributos para *speaker* (hablante), *startTime* (tiempo de inicio) y *endTime* (tiempo de finalización).

En la **Figura 4** vemos este proceso de conversion a XML. Los valores numéricos corresponden a los tiempos de alineamiento con el sonido, expresados en milisegundos. De esta forma, se delimita cada enunciado, identificando a su correspondiente hablante.

El siguiente paso es la anotación morfosintáctica a partir de este segundo fichero generado. 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.
- Fase 2: etiquetador estadístico TNT.

```
<UNIT speaker="JOR" startTime="0" endTime="4.482">aquí </UNIT>
<UNIT speaker="TEA" startTime="4.482" endTime="7.655">a ver si puedes
</UNIT>
<UNIT speaker="TEA" startTime="7.655" endTime="9.246"> ¿ cuántos años
tienes Jorge ? </UNIT>
<UNIT speaker="JOR" startTime="9.246" endTime="12.459">&eh tengo -> /
cuatro </UNIT>
<UNIT speaker="TEA" startTime="12.459" endTime="13.131">cuatro
</UNIT>
<UNIT speaker="TEA" startTime="13.131" endTime="14.267"> que fue tu
cumple el otro día </UNIT>
<UNIT speaker="TEA" startTime="14.267" endTime="14.817"> ¿ a que sí ?
</UNIT>
<UNIT speaker="JOR" startTime="14.817" endTime="15.667">cinco sí
</UNIT>
<UNIT speaker="TEA" startTime="15.667" endTime="16.411">¡ah! </UNIT>
<UNIT speaker="TEA" startTime="16.411" endTime="17.09"> ¿ cinco ?
</UNIT>
<UNIT speaker="TEA" startTime="17.09" endTime="17.755"> ¿ o cuatro ?
</UNIT>
<UNIT speaker="JOR" startTime="17.755" endTime="23.601">bueno / hoy
&cum [/] mañana cumplí cinco // pero ahora / tengo cuatro </UNIT>
```

Figura 4: Conversión a XML

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

```
<Text>
<p>
<f h="JOR" st="0.0" et="4.482" id="1">
<sf t="enu" id="1-1">
<w cat="P" lem="aquí" id="1-1-1"> aquí </w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="4.482" et="7.655" id="2">
<sf t="enu" id="2-1">
<w cat="MD" lem="a ver" id="2-1-1"> a ver </w>
<w cat="C" lem="si" id="2-1-2"> si </w>
<w cat="V" lem="poder" tie="pres_ind" num="sing" per="2" id="2-1-3"> puedes
</w>
</sf>
</f>
</p>
<p>
<f h="TEA" st="7.655" et="9.246" id="3">
<sf t="int" id="3-1">
<w cat="PUNCT" lem="¿" id="3-1-1"> ¿ </w>
<w cat="P" lem="cuántos" gen="masc" id="3-1-2"> cuántos </w>
<w cat="N" lem="año" gen="masc" num="plu" id="3-1-3"> años </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="3-1-4"> tienes
</w>
<w cat="NPR" lem="Jorge" id="3-1-5"> Jorge </w>
<w cat="PUNCT" lem="?" id="3-1-6"> ? </w>
```

De esta forma, cada palabra del corpus puede ser relacionada con el hablante. El archivo mantiene en la cabecera toda la información socio-contextual, pudiendo crear tantos subcorpus como características diferentes aparezcan en la cabecera —un subcorpus de lenguaje adulto, un subcorpus infantil, un subcorpus por sexo, etc. Después de la partición en subcorpus, es posible calcular todas las ocurrencias (los *tokens*) para cada unidad léxica (los *types*). El procedimiento se puede aplicar a cualquier tipo de información lingüística anotada en el corpus.

2.2 Extracción de grupos de palabras

Si calculamos las estadísticas directamente sobre cada unidad, el resultado no sería correcto, ya que los elementos léxicos pluriverbales (es decir, las locuciones) no estarían incluidos en este recuento. Marcadores discursivos

tan frecuentes como "por ejemplo", "o sea" o "es decir" no aparecerían si trabajamos sobre unidades léxicas formadas por una sola palabra. Para solucionar esto, se ha creado una lista exhaustiva de locuciones por categorías, incluyendo compuestos nominales ("fin de semana"). Cada locución se considera una unidad léxica, equivalente a las palabras simples.

2.3 Aplicación de la estadística de la sorpresa

Para identificar las palabras, lemas o categorías distintivos de un subcorpus dado, hemos empleado el test de razón de verosimilitud (log-likelihood ratio test) propuesto por Dunning (1993). Este método no asume distribuciones estadísticas normales de las unidades de un corpus. Por el contrario, la ratio de probabilidad (logarítmica) λ asume una distribución binomial más apropiada para palabras poco comunes pero significativas. "Texts are composed largely of such rare events" (Dunning 1993). Además, este test no necesita subcorpus equilibrados para llevar a cabo la comparación.

Este método se ha aplicado con éxito para hallar colocaciones (Dunning 1993) y términos (Daille 1994). Para probar el método con la intención de encontrar unidades distintivas en dominios específicos, podemos trabajar con dos hipótesis:

i. Dos registros (o subcorpus) no muestran ninguna diferencia en unidades distintivas (*Hipótesis nula*).
ii. Para un subcorpus dado, podemos hallar unidades distintivas (*Hipótesis alternativa*).

Aplicamos el test a dos subcorpus bien definidos: lenguaje adulto e infantil. Los resultados se muestran en la **Tabla 5** y la **Tabla 6**.

FORMAS	ADULTOS (36.905)	NIÑOS (21.080)	DUNNING
qué	1.123	108	510.29
te	743	59	373.43
a ver	371	23	207.58
bien	304	14	189
ah	270	18	146.32
claro	231	15	126.53
tú	264	27	113.88
has	184	9	112.02
tu	197	14	103.64
cómo	249	27	103.26

Tabla 5: Formas distintivas en el lenguaje adulto

FORMAS	NIÑOS (21.080)	ADULTOS (36.905)	DUNNING
mi	334	24	524.66
yo	417	166	300.54
sí	647	428	255.77
me	423	248	198.53
tengo	130	28	141.16
Candi	67	9	88.55
porque	150	86	71.99
un	431	424	71.25
padre	60	13	64.86
he	79	27	64.09

Tabla 6: Formas distintivas en el lenguaje infantil

Los resultados confirman la hipótesis alternativa y la idoneidad del test de Dunning para esta tarea. La mayoría de las 10 formas más destacadas en ambos dominios tiene una baja ocurrencia, pero todos son términos típicos de ese dominio.

3. Resultados preliminares

Nuestro propósito es mostrar una gran variedad de posibilidades para la aplicación de este método a la extracción de información de un corpus. Por el momento, mostraremos un conjunto de datos bastante incompleto. Actualmente, hay una desproporción de características sociales y de registro con respecto a las características lingüísticas; nuestra intención es suplir esas carencias ampliando el corpus en un futuro.

En este trabajo, los fenómenos lingüísticos que se han tenido en cuenta son palabras y locuciones (como hemos visto en las **Tablas 5** y **6**), fonemas y categorías.

A continuación, presentamos los resultados del Test de Dunning para los dos subcorpus: lenguaje adulto e infantil. El primero de ellos está formado por 36.905 palabras y el segundo, por 21.080 palabras. La **Tabla 7** y la **Tabla 8** muestran las categorías distintivas de cada uno de los subcorpus.

CATEGORÍAS	ADULTOS (36.905)	NIÑOS (21.080)	DUNNING
MD	1.731	449	264.71
P	6.564	2.739	234.8
INTJ	524	81	162.52
V	6.450	3.167	59.07
AUX	1.278	522	44.88

Tabla 7: Categorías distintivas del lenguaje adulto

CATEGORÍAS	NIÑOS (21.080)	ADULTOS (36.905)	DUNNING
POSS	453	360	127.55
N	3.174	4.419	110.27
ADV	1.861	2.428	96.97
Q	1.739	2.497	42.94
NPR	910	1.242	33.33
C	2.184	3.403	19.83
PREP	1.773	2.786	13.63
ART	1.338	2.068	13.29

Tabla 8: Categorías distintivas del lenguaje infantil

Los resultados nos llevan a interpretar lo siguiente:

- En el lenguaje adulto, al contrario de lo que ocurre en el infantil, abundan elementos como los marcadores discursivos (MD) o las interjecciones (INTJ). Ambos elementos pertenecen al nivel pragmático de la lengua, y éste requiere una mayor destreza lingüística.
- Mientras que en la lengua adulta los verbos (V) son el elemento que guía el discurso, los niños de estas edades (3 a 6 años) aún apoyan el suyo sobre los nombres (N).
- Los pronombres posesivos (POSS) son el elemento más distintivo del lenguaje infantil. Según J. Piaget (1965), hasta los siete años de edad el lenguaje de los niños se caracteriza por ser egocéntrico, es decir, es un simple acompañamiento de la acción y el niño no posee otra perspectiva que no sea la propia. Si volvemos a la **Tabla 6**, observamos que algunas de las palabras más características del lenguaje infantil son "mi", "yo" o "me".
- En el lenguaje infantil son distintivas las categorías gramaticales como la conjunción (C), la preposición (P) o el artículo (ART). En concreto, el uso típico de la conjunción en este subcorpus se debe a la abundante utilización por parte de los niños 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.
- Por último, en la **Tabla 8** aparece como categoría distintiva del lenguaje infantil el nombre propio (NPR). Esto es una consecuencia derivada del contexto situacional escolar en el que se producen las interacciones: los niños se demandan constantemente la atención del profesor llamándole por su nombre.

Además de a las categorías, hemos aplicado el Test de Dunning al nivel fonológico. Las transliteraciones ortográficas obtenidas de las grabaciones se transcriben automáticamente en formato IPA. De esta forma, podemos

someter a los textos al mismo proceso explicado en el apartado 2.1 y extraer la información fonológica.

FONEMAS	ADULTOS (136.721)	NIÑOS (77.240)	DUNNING
t	6.408	2.949	90.86
b	4.197	1.914	63.6
e	19.938	10.342	58.27
k	6.851	3.352	49.62
s	11.382	5.924	28.72

Tabla 9: Fonemas distintivos del lenguaje adulto

FONEMAS	NIÑOS (77.240)	ADULTOS (136.721)	DUNNING
ʌ	1.024	1.108	128.15
i	6.277	9.635	82.59
p	2.265	3.176	72.57
m	2.928	4.348	55.2
o	8.490	14.247	16.88
u	2.872	4.667	13.39
r	407	571	12.71
g	957	1.461	12.66
x	618	940	8.54

Tabla 10: Fonemas distintivos del lenguaje infantil

Lo más llamativo de los resultados obtenidos es el carácter distintivo de los fonemas ʌ y r en el lenguaje infantil. El primero de ellos puede ser debido al abuso del pronombre personal "yo" en el lenguaje egocéntrico propio de los niños. Ambos fonemas son líquidos, curiosamente los últimos que se adquieren en el proceso de aprendizaje de la primera lengua, junto con los fricativos, por la dificultad que entraña su punto de articulación (Anula 1998).

4. Conclusiones y trabajo futuro

En este artículo, hemos probado la relevancia de este procedimiento, el Test de Dunning, como método empírico para la validación de hipótesis sociolingüísticas en lengua oral, al igual que para determinar una tipología de registros.

El método correlaciona datos lingüísticos con socio-contextuales aplicando las Estadísticas de Sorpresa de Dunning. Para conseguir esto, son esenciales un corpus enriquecido con anotación lingüística y el uso de XML. Los resultados preliminares son prometedores y no habían sido mostrados con anterioridad para el español. Sin embargo, es prematuro extraer

conclusiones e interpretaciones para estos datos, puesto que el corpus es claramente insuficiente.

La principal limitación de CHIEDE es su tamaño: 60.000 palabras no es un número suficiente para establecer clasificaciones y análisis estadísticamente significantes. No obstante, nuestra intención es incrementar el tamaño del corpus, el número de participantes y su variedad de situaciones comunicativas.

5. Referencias bibliográficas

Anula, Alberto (1998) *El abecé de la psicolingüística*. Madrid: Arco-Libros, D.L.

Biber, Douglas (1988). *Variation across speech and writing*. Cambridge: CUP.

Biber, Douglas (1995). *Dimensions of register variation*. Cambridge: CUP.

Biber Douglas, *et al.* (eds.) (1999), *The Longman grammar of spoken and written English*. London: Longman.

Cresti, Emmanuela *et al.* (2002). "The C-ORAL-ROM project. New methods for spoken language archives in a multilingual romance corpus". *Proceedings of LREC 2002*. Las Palmas de Gran Canaria.

Daille, Béatrice (1994). *Combined approach for terminology extraction: lexical statistics and linguistic filtering*. Ph.D. Thesis, Paris 7.

Dunning, Ted (1993). "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics*, 19 (1): 61-74.

Garrote, Marta (2008). *CHIEDE. Corpus de Habla Infantil Espontánea del Español*. Tesis Doctoral. Universidad Autónoma de Madrid.

Guirao, José M., *et al.* (2006). "Relating linguistics units to socio-contextual information in a spontaneous speech corpus of Spanish". En Andrew Wilson, Dawn Archer y Paul Rayson, eds., *Corpus linguistics around the world*. Amsterdam: Rodopi.

Labov, William (1966). *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.

Miller, Jim y Weinert, Regina (1999). *Spontaneous spoken language*. Oxford: Clarendon.

Moreno, Antonio (2002). "La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM". *Proceedings of II Jornadas de Tecnologías del Habla*. Granada.

Moreno, Antonio y Guirao, José M. (2003) "Tagging a spontaneous speech corpus of Spanish". *Proceedings of Recent Advances in NLP (RANLP-2003).* Borovets, Bulgaria.

Piaget, Jean (1965) *El lenguaje y el pensamiento en el niño*, Buenos Aires: Paidós.

Uchimoto, Kiyotaka (2002). "Morphological Analysis of the Spontaneous Speech corpus". *Proceedings of Conference of Computational Linguistics (COLING 2002).* Taipei, Taiwan.