Anotación de disfluencias en un corpus de habla espontánea no específico

Luis Javier Rodríguez, Inés Torres, Amparo Varona

Departamento de Electricidad y Electrónica
Facultad de Ciencias
Universidad del País Vasco
Apartado 644. 48080. BILBAO
{luisja,manes,amparo}@we.lc.ehu.es

Resumen

En esta comunicación presentamos la anotación de fenómenos de habla espontánea (también conocidos como disfluencias) en una parte del Corpus de Referencia de la Lengua Española Contemporánea de la Universidad Autónoma de Madrid. El subcorpus completo consiste en 132 entrevistas y conversaciones –tomadas de radio y televisión a finales de 1991–, de las cuales se han anotado hasta la fecha 42 entrevistas, con una duración aproximada de seis horas y media. Las anotaciones se han generado en dos fases: en primer lugar se han filtrado las transcripciones originales –creadas para realizar estudios lingüísticos del habla espontánea–, sobre todo porque algunas de las informaciones no eran de interés desde el punto de vista del reconocimiento automático del habla, pero también para adaptar las convenciones ortográficas y el formato de las anotaciones a nuestras necesidades; en segundo lugar se han revisado y aumentado las anotaciones resultantes, añadiendo marcas acústicas y léxicas: ruidos, pausas de silencio, pausas habladas, alargamientos, palabras cortadas o mal pronunciadas, etc. Aunque no se ha llegado a completar la anotación del subcorpus, las estadísticas de aparición de los fenómenos anotados hasta la fecha dan una idea de la importancia que puede tener su modelado en el rendimiento de los sistemas de reconocimiento y comprensión del habla.

1. Introducción

Las nuevas aplicaciones del reconocimiento y la comprensión automáticos del habla continua (traducción, acceso a información, venta de billetes, etc.) requieren manejar lo que se conoce como habla espontánea, es decir, habla producida sin restricciones, en la que el hablante puede cometer incorrecciones gramaticales, retroceder y corregir fragmentos, dudar, repetir palabras, producir todo tipo de ruidos, etc. Se suele aludir a estos fenómenos con el nombre de disfluencias, si bien el término puede llevar a confusión, ya que algunos autores lo utilizan en sentido muy amplio y otros lo restringen a un cierto tipo de fenómenos. Para evitar polémicas, nosotros utilizaremos el término más general de *fenómeno de habla espontánea* –FHE en lo sucesivo–, refiriéndonos a cualquier característica que distinga el habla espontánea del habla leída, incluyendo tanto condiciones ambientales (ruidos), condiciones de la interacción (habla simultánea), como características específicas de la lengua hablada.

Para obtener modelos acústicos y todavía en mayor medida modelos de lenguaje capaces de operar con ese tipo de habla, las metodologías actuales requieren el uso de una cantidad ingente de muestras, es decir, adquirir y anotar bases de datos de habla espontánea de grandes dimensiones. En el contexto de varios proyectos de ámbito español y europeo, ya se están adquiriendo y anotando bases de datos de esas características en lengua española, normalmente orientadas a una tarea –que suele implicar una interacción hombre-máquina–, y por tanto con un vocabulario y una sintaxis por lo general bastante restringidas. Cabe decir que el hecho de modelar el habla sobre una tarea concreta viene determinado tanto por la condición de que el sistema sea factible como por la de que sea útil.

En particular, nuestro grupo participa en el desarrollo un prototipo de atención al cliente de una línea de ferrocarriles, quien desea planificar un viaje y solicita información sobre el mismo: horarios, destinos, precios, etc. [1] –proyecto al que en adelante nos referiremos como INFOTREN. En el contexto de este proyecto se han adquirido un conjunto de diálogos utilizando la técnica del mago de Oz, se han anotado todo tipo de FHEs (acústicos, léxicos, sintácticos y pragmáticos) [2], y se han llegado a modelar los fenómenos acústicos [3].

Sin embargo, la espontaneidad de los hablantes que participan en una tarea como la descrita puede diferir tanto en tipología como en intensidad de la que de forma natural se produce en situaciones cotidianas. Nuestro interés se traslada pues al estudio del habla espontánea desde un punto de vista general, con objeto de disponer de una referencia con la que comparar el habla obtenida en tareas específicas. Para llevar a cabo dicho estudio se requiere la adquisición y anotación de habla no orientada, desde monólogos hasta conversaciones entre varias personas, pasando por entrevistas y pequeños diálogos del día a día.

La tarea de adquisición y anotación de una base de datos de las características mencionadas supone un trabajo de preparación y postproceso enorme que puede llevar varios años, razón por la cual se ha preferido utilizar una parte del Corpus de Referencia de la Lengua Española Contemporánea de la Universidad Autónoma de Madrid [4] –CRLEC en lo sucesivo–, cuyas características encajan perfectamente con nuestras necesidades. El corpus contiene tanto las señales como las transcripciones, y éstas a su vez incluyen muchos de los FHEs que deseamos anotar, como palabras cortadas, palabras mal pronunciadas, pausas, etc. No obstante, CRLEC presenta un gran inconveniente: la calidad de las grabaciones es por lo general muy baja, ya que fueron realizadas con material muy básico (grabadora y cinta de audio) y en condiciones ambientales adversas. De hecho, para el grupo de investigación que llevó a cabo la adquisición, una vez realizadas las transcripciones las señales tenían apenas valor de referencia, ya que su interés se centraba en aspectos morfológicos, sintácticos y semántico-pragmáticos de la lengua hablada, no tanto en aspectos acústicos, fonéticos o fonológicos.

Así pues, además de contener un alto grado de espontaneidad, CRLEC también representa un caso extremo en cuanto a las condiciones adversas de la señal.

El resto del artículo se estructura como sigue. En el apartado 2 se explica el camino seguido para escoger las señales que forman el subcorpus. El apartado 3 describe el inventario de marcas utilizado para anotar los FHEs, y el apartado 4, el proceso de filtrado de las marcas originales. En el apartado 5 se explican los criterios de cortado y anotación, y en el apartado 6, las herramientas utilizadas en el proceso de anotación, así como la problemática asociada a dicho proceso. Finalmente, en el apartado 7 se muestra la distribución de fenómenos en el subcorpus, con datos absolutos y relativos, acompañados de una breve discusión, y en el apartado 8 se indican líneas de actuación futuras.

2. Definición del subcorpus

CRLEC está formado por diálogos entre dos o más personas, grabados en radio, en televisión, en la calle, en aulas de clase, en mesas redondas, en casa, etc. Se trata, por tanto, de un corpus de habla espontánea en castellano, que abarca diversos dominios semánticos y pragmáticos, ya que los diálogos se desarrollan en distintos contextos socio-culturales y en distintas situaciones, algunas más formales, otras más familiares. El objetivo final de este trabajo es disponer de una base de datos genérica de habla espontánea, lo suficientemente grande como para obtener modelos acústicos robustos, con un vocabulario y una sintaxis no restringidos a una tarea.

No se ha procesado todo el corpus, sino sólo una parte, cuyas dimensiones se ajustan –creemos– al objetivo enunciado más arriba. Las cuentas realizadas sobre las anotaciones originales de CRLEC arrojan un total de 941386 palabras, con un tamaño del vocabulario de 39785 palabras. La base de datos consta de 17 bloques o secciones, creados atendiendo al área temática o al tipo de registro del habla. En la tabla 1 se muestra el identificador de cada uno de los bloques, el tipo de conversaciones que contiene, así como el número de palabras total, el tamaño del vocabulario y el promedio de muestras por palabra. Este último es un dato importante para el modelo de lenguaje del reconocedor.

Al no tratarse de diálogos enfocados todos ellos a una misma tarea, sino cada uno sobre un tema distinto, tendremos como resultado una gran dispersión léxica, es decir, muy pocas muestras de cada palabra, lo cual repercutirá en un modelo de lenguaje infraentrenado. Como dato de referencia, el número promedio de muestras por palabra obtenido para la tarea INFOTREN era de aproximadamente 23. Observando la tabla 1, vemos que –sorprendentemente– son dos bloques no temáticos: las conversaciones (14.03) y las entrevistas (11.51), los que arrojan un mayor ratio de muestras por palabra. Si unimos los dos bloques en uno solo, se obtiene un total de 355216 palabras, y un tamaño del léxico de 21537 palabras, de modo que el ratio de muestras por palabra resulta aún mayor: 16.49.

Por otra parte, dado que nuestro interés principal es estudiar los FHEs, para cada bloque se han contado los fenómenos ya presentes en las anotaciones originales de CRLEC: palabras cortadas, fonemas borrrados, pausas habladas, silencios, sonidos guturales o interjecciones con función fática, ruidos, etc. Resulta un total de FHEs de 92412, y un promedio de FHEs por palabra, para toda la base de datos, de 0.098. Los datos desglosados por bloques se muestran en la tabla 2, donde por un lado vemos que el 41.63% de los FHEs se producen en las conversaciones y entrevistas, y por otro que ambos bloques muestran un promedio de FHEs por palabra alrededor de la media: las conversaciones claramente por encima y las entrevistas ligeramente por debajo. No obstante, en cuanto a la *densidad* de FHEs no hay grandes diferencias. Si bien entre el mínimo (0.058) y el máximo (0.180) hay un factor de 3, 13 de los 17 bloques muestran valores comprendidos entre 0.08 y 0.12.

Tabla 1: Estadísticas de los diferentes bloques que conforman la base de datos CRLEC: número de palabras, tamaño del léxico y promedio de muestras por palabra (M/P). El tamaño del vocabulario es de 39785 palabras, sobre un total de 941386.

Bloque	Contenido	Palabras	Léxico	M/P
adm	Administrativos	6322	1080	5.8537
cie	Científicos	35172	4857	7.24151
con	Conversacionales	207748	14808	14.0294
deb	Debates	81928	8557	9.57438
dep	Deportivos	47165	5597	8.42684
doc	Documentales	26779	4721	5.67232
edu	Educativos	59240	6429	9.2145
en	Entrevistas	147468	12813	11.5092
hum	Humanísticos	53432	7150	7.47301
ins	Instrucciones	7175	1321	5.43149
jur	Jurídicos	34386	4247	8.09654
lud	Lúdicos	50347	6356	7.92118
no	Noticiario	65373	8389	7.7927
pol	Políticos	48604	5864	8.28854
pub	Publicitarios	24896	3864	6.44306
rel	Religiosos	11162	2298	4.85727
tec	Técnicos	34687	4333	8.00531

Como consecuencia de los datos anteriores, se establecen como bloques candidatos para conformar el subcorpus los de conversaciones y entrevistas. El bloque de conversaciones consta de diálogos por lo general abiertos, que suelen implicar a más de dos personas e incluyen múltiples solapamientos ya que no hay un moderador que conceda turnos. Las entrevistas son conversaciones más formales, normalmente entre dos personas, una de ellas haciendo las veces de entrevistador/moderador. En general la calidad acústica de las entrevistas es aceptable ya que en su mayor parte han sido tomadas de la radio, mientras que entre las conversaciones hay muchas grabadas en la calle, en bares o en reuniones caseras, con muchos ruidos, eco, etc.

En este punto era necesario comprobar por un lado los niveles de ruido, con objeto de descartar señales demasiado ruidosas, y por otro el número de solapamientos, ya que tales segmentos no iban a servir para entrenar los modelos acústicos, así que diálogos con una pequeña fracción de señal útil también serían descartados. Tras escuchar y evaluar subjetivamente todas las señales candidatas, se consideraron acústicamente aceptables 67 de un total de 79 entrevistas y 65 de un total de 126 conversaciones (véase la tabla 3), que conforman un subcorpus de 132 diálogos, al cual nos referiremos en adelante como CRLEC-EHU.

3. Inventario de fenómenos

El trabajo realizado con INFOTREN permitió establecer un primer inventario de FHEs, adecuado para diálogos hombre-máquina [2]. Sin embargo, en diálogos entre humanos aparece una gama más amplia de fenómenos. En concreto, se han añadido afirmaciones y negaciones guturales, solapamientos, es decir, fragmentos de señal en los que dos o más voces se super-

Tabla 2: Estadísticas de FHEs en las anotaciones originales de CRLEC, desglosadas por bloques: número absoluto de FHEs (NFHE), porcentaje que este número representa sobre el total de FHEs (%) y ratio de FHEs por palabra (FHE/P).

Bloque	NFHE	%	FHE/P
adm	1136	1.22928	0.17969
cie	3634	3.93239	0.103321
con	25095	27.1556	0.120795
deb	6651	7.19712	0.081181
dep	4326	4.68121	0.0917206
doc	2139	2.31463	0.079876
edu	5778	6.25243	0.0975354
ent	13376	14.4743	0.0907044
hum	4723	5.11081	0.0883927
ins	809	0.87542	0.112753
jur	3624	3.92157	0.105392
lud	5066	5.48197	0.100622
not	3804	4.11635	0.0581892
pol	4072	4.40635	0.0837791
pub	2846	3.07969	0.114316
rel	769	0.83214	0.0688945
tec	4564	4.93875	0.131577

Tabla 3: Número de conversaciones y entrevistas que conforman el subcorpus CRLEC-EHU. Se dan asimismo el número de turnos y la duración de las señales.

Bloque	Señales	Turnos	Duración (seg)
con	65	9691	38383
ent	67	4502	38907

ponen, y una marca especial que señala la continuación de una frase cortada (por un ruido o por una voz solapada) en un turno posterior. También se han añadido marcas específicas para palabras que no forman parte de la lengua, como siglas (deletreadas total o parcialmente) y palabras extranjeras. Por otro lado, las condiciones de grabación han forzado la definición de marcas especiales para segmentos no transcritos o ininteligibles, especialmente frecuentes en ambientes ruidosos o cuando varias personas hablan a la vez, y para cortes en la grabación, producidos originalmente por el fin de la cinta de audio o en el postproceso por errores en la transferencia de la señal a formato digital.

4. Filtrado de las transcripciones originales

El inventario ampliado de FHEs –definido específicamente para CRLEC-EHU– se expresa en el mismo formato simplificado utilizado en INFOTREN –al cual nos referiremos en adelante como formato EHU simplificado–, adecuado para el proceso de anotación –pero que es traducido después automáticamente a un formato más portable, basado en XML. Las anotaciones de CRLEC-EHU en formato EHU simplificado se han de ubicar en ficheros de texto sin cabecera, siguiendo las mismas pautas, convenciones ortográficas, etc. utilizadas en INFOTREN. Sin embargo, las anotaciones originales de CRLEC se ubican en ficheros de texto con una pequeña cabecera seguida de las transcripciones. Tanto la cabecera como las transcripciones contienen marcas de tipo SGML, tomadas de un inventario de marcas distinto al nuestro, y aplicando también distintas convenciones ortográficas [5]. Así pues, ha sido necesario escribir y aplicar un pequeño *script* Perl para pasar las 132 transcripciones del subcorpus CRLEC-EHU a formato EHU simplificado, según un conjunto de reglas de las que extraemos las principales:

- La cabecera se elimina.
- La transcripción de cada turno irá precedida por el identificador del locutor y el índice del turno.
- Las dobles comillas (") y los puntos suspensivos aislados (...) se eliminan.
- Los caracteres dos_puntos (:) y punto_y_coma (;) son sustituidos por comas (,).
- Las palabras cortadas, así como las siglas, palabras extranjeras y silencios simplemente cambian de formato.
- Los borrados de fonemas se anotan como palabras mal pronunciadas.
- Los puntos suspensivos (...) cuando van inmediatamente tras una palabra se anotan como alargamiento de fonema más pausa de silencio, si la palabra termina en vocal o en 'n', 'l' o 's', o simplemente como pausa de silencio en caso contrario.
- Los sonidos fáticos de afirmación y negación se transcriben como afirmaciones y negaciones guturales, respectivamente.
- El resto de sonidos fáticos, así como las marcas de vacilación y la secuencia *"eh..."* se transcriben como pausas habladas.
- Las marcas *"<ininteligible>"* y *"<texto no transcrito>"* se transcriben como *"[NO TRANSCRITO]"*.
- Los borrados involuntarios se transcriben como cortes.
- Las marcas *"<simultáneo> X"* y *"X </simultáneo>"* se transcriben ambas como solapamientos: *"(o X)"*.
- Las marcas que indican el tipo de habla o el modo en que se produce el habla: *"<texto leído>"*, *"<cantando>"*, *"<onomatopéyico>"* y *"<argot>"*, se eliminan.
- Toses, carraspeos y demás marcas de ruidos (aplausos, risas, música, etc.) se transcriben todas como ruido genérico.

5. Criterios de cortado y anotación

El resultado las transformaciones anteriores fue tomado como punto de partida por el anotador, que debía llevar a cabo dos tareas: (1) cortar la señal correspondiente a cada diálogo en tantos ficheros como turnos resultaran para dicho diálogo, y (2) corregir y aumentar con FHEs las anotaciones ya existentes. El proceso de cortado y anotación se efectuó con dos criterios principales:

1. Cortar sólo señales acústicamente aceptables, es decir, útiles para entrenar modelos acústicos. Esto significa que si un segmento de señal presentaba ruido o música de fondo apreciables, o contenía voces solapadas, debía ser descartado.
2. Anotar todo: ruidos de fondo, música de fondo, cortes, solapamientos, respiraciones, tics de los labios, etc. Esto significa que incluso un segmento de señal descartado debía ser transcrito completamente.

Para compatibilizar ambos criterios, las anotaciones indican dónde empiezan y dónde terminan los tramos ruidosos, de manera que aún disponiendo de toda la transcripción, es posible quedarse únicamente con los fragmentos de transcripción para los que se tiene señal cortada. El final de un tramo ruidoso determina el final de un turno, de modo que el siguiente tramo acústicamente aceptable tiene lugar ya en el turno siguiente. Ello permite asociar a cada señal cortada un índice de turno distinto. Si, por ejemplo, se produce un solapamiento entre el final de la intervención de un hablante y el comienzo de la intervención de otro, el segmento solapado quedará fuera de los dos ficheros de señal que se generan. Tal como muestra la Figura 1, en un fichero estaría la señal hasta el inicio del solapamiento y en el otro la señal a partir del final del mismo. La definición de los puntos de inicio y final del solapamiento se ha hecho de forma que abarquen palabras completas. Es decir, si una palabra estuviera afectada por el solapamiento, aunque sólo fuera muy al principio o muy al final, dicha palabra sería incluida en el segmento solapado.

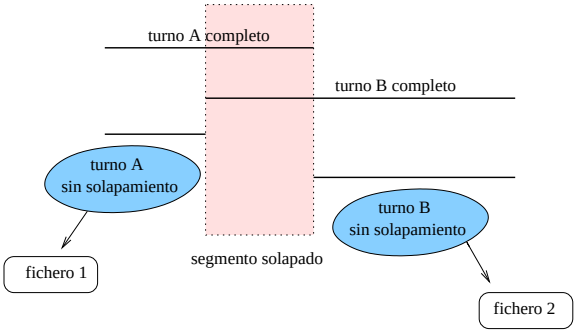

Figura 1: Esquema de cortado de dos intervenciones solapadas: se cortan los fragmentos anterior y posterior al solapamiento y se descarta el segmento de señal solapado.

6. Proceso de anotación

El cortado de las señales se realizó de forma manual, visualizando por un lado el texto de la transcripción original –en formato simplificado EHU– mediante un editor de texto básico, configurado para colorear las marcas y facilitar su localización e inserción, y por otro la forma de onda de la señal –en formato ESPS, a frecuencia de muestreo de 16 kHz– mediante la herramienta *XWAVES*, que permite escuchar, seleccionar y cortar los fragmentos correspondientes a las intervenciones de los hablantes. No se utilizaron las herramientas de transcripción y segmentación del *XWAVES* debido a que (1) no era necesario sincronizar las marcas con la señal, y (2) no eran sólo transcripciones lo que queríamos generar, sino marcas sobre las transcripciones, para lo cual dichas herramientas no servían.

El proceso de anotación fue dividido en dos fases. La primera fase, realizada a la vez que el cortado de las señales, debía cubrir los fenómenos acústicos y léxicos: ruidos externos, ruidos producidos por el locutor, pausas de silencio, pausas habladas, alargamientos de sonidos (principalmente vocales), palabras cortadas, palabras mal pronunciadas, sonidos guturales que pueden indicar afirmación o negación, siglas, palabras extranjeras y solapamientos. Asimismo, en la primera fase se revisaría la corrección de las transcripciones, y se efectuarían determinadas conversiones: por ejemplo, números, ordinales y fechas se transcriben ortográficamente tal como han sido pronunciados, etc. En la segunda fase las señales estarían ya cortadas, y las anotaciones acústicas y léxicas corregidas y aumentadas, de modo que el anotador podría concentrarse únicamente en la detección y caracterización de los fenómenos sintácticos (reformulaciones) y pragmáticos (marcadores de discurso). Esta segunda fase se efectuaría siguiendo estrictamente el mismo esquema: procedimiento, marcas, etc. aplicado en INFOTREN, al cual nos remitimos [6]. Hasta la fecha tan sólo ha llegado a completarse la primera fase de anotación sobre 47 de las 67 entrevistas, de las cuales 5 han sido descartadas y 42 han pasado a formar parte de la base de datos preliminar, a la que llamaremos CRLEC-EHU-1.

La tarea de anotación se desarrolló entre el 24 de abril y el 19 de julio de 2002. Se contrató a una persona a tiempo parcial, para que desempeñase la tarea de anotación durante 3 horas al día. Esta persona, que ya había desempeñado previamente tareas de anotación similares, cumplió 60 días de trabajo, lo que hace un total de 180 horas. Se había estimado que, una vez entrenada en el uso de las herramientas, podría anotar del orden de dos diálogos al día. Lamentablemente, el rendimiento obtenido ha sido muy inferior, del orden de 0.7 diálogos al día si sólo contamos los 42 diálogos anotados. Teniendo en cuenta que la duración de dichos diálogos es de 23106 segundos (6.41 horas), se deduce que cada hora de diálogo ha llevado cerca de 30 horas de trabajo de cortado y anotación. Obviamente, estos números podrían reducirse aumentando la ergonomía de las herramientas de anotación, pero baste el dato para poner de manifiesto una vez más los costes enormes asociados a la creación de bases de datos de habla, y la necesidad de unir esfuerzos para la creación de dichas bases de datos.

7. Distribución de fenómenos en el subcorpus

En lo que se refiere al proceso de cortado, las 42 entrevistas que conforman el subcorpus CRLEC-EHU-1 han producido 2090 turnos, que suman una duración de 20197 segundos (5.61 horas), es decir, un 87.41% de la duración total de los diálogos. La duración media de cada turno resulta de 9.66 segundos, pero con una desviación típica muy alta (14.40). De hecho, el histograma de duraciones (Figura 2) muestra un pico muy claro de turnos muy cortos (de 0 a 5 segundos) –con poblaciones de más de 100 turnos–, que va seguido de poblaciones paulatinamente menores, hasta llegar a los 60 segundos, donde la población se hace prácticamente nula. En números: por un lado, hay 1090 turnos de duración inferior a 5 segundos (un 52.26% de todos los turnos), que suman una duración de tan sólo 1994 segundos (un 9.87% de la duración total); por otro lado, sólo hay 33 turnos de duración superior a 60 segundos (un 1.58% de todos los turnos), pero suman una duración de 2845 segundos (un 14.09 de la duración total).

Los datos anteriores reflejan el esquema de interacción típico de una entrevista, con preguntas bastante concisas por parte del entrevistador, seguidas de largos monólogos a cargo del entrevistado, con eventuales gestos de asentimiento, acompañados a veces de sonidos guturales y otras veces de palabras como ”claro”, ”por supuesto”, ”sí”, ”ya”, etc. Es de prever que el cortado y anotación de las conversaciones –donde el esquema de interacción es más abierto– resulte en una distribución de duraciones distinta.

En lo que se refiere a los FHEs, la Tabla 4 muestra las cuentas obtenidas sobre las 42 entrevistas, así como el ratio de FHEs

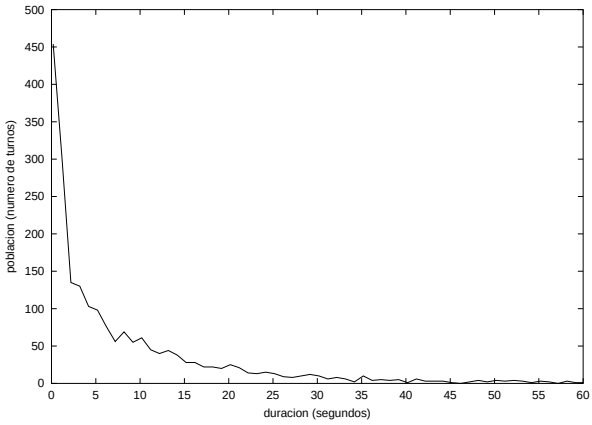

Figura 2: Histograma de los turnos de CRLEC-EHU-1 con respecto a su duración.

por palabra, teniendo en cuenta que en dichas entrevistas se han contabilizado 64905 palabras (sin contar los FHEs). A continuación discutiremos brevemente los números de dicha tabla.

En primer lugar, llama la atención el gran número de solapamientos: 1808 para un total de 2090 turnos. No obstante, sobre este dato hay que hacer algunas observaciones: (1) muchos de tales solapamientos corresponden a turnos completos que no han sido cortados y que por tanto no forman parte de los 2090 que *sí* han sido cortados; (2) en algunas ocasiones –no muchas– un mismo turno incluía dos solapamientos, uno al principio con el turno anterior y otro al final con el turno siguiente; y (3) no siempre pero sí la mayor parte de las veces, la marca *CONTINUA* se añadía tras un solapamiento breve que no interrumpía el discurso del locutor principal, de ahí que podamos estimar en alrededor de 400 el número de ese tipo de solapamientos.

Es notable también la cantidad de aspiraciones anotadas (3005, del orden de una cada 20 palabras, casi el 94% de los ruidos producidos por los hablantes). Esto responde a la necesidad de los hablantes de tomar aire periódicamente mientras desarrollan su discurso. Estos fenómenos operan, por tanto, como *pausas técnicas*, que generalmente no corresponden a una frontera entre unidades lingüísticas. Es importante anotar estas aspiraciones, distinguiéndolas de las pausas de silencio, ya que éstas sí suelen ir colocadas entre unidades lingüísticas, y a veces marcan la presencia de FHEs más complejos, como reformulaciones o frases abandonadas.

En cuanto a los ruidos externos, la cantidad anotada parece anormalmente pequeña (530) sobre todo teniendo en cuenta las condiciones ambientales de las grabaciones. Este número se explica por el hecho de que solamente estamos contabilizando ruidos externos *aislados* que suceden en segmentos por lo demás relativamente libres de ruido. Los segmentos de señal afectados por una secuencia ininterrumpida de clicks o por un ruido o música de fondo demasiado fuertes, al igual que los que contienen voces solapadas, no han sido cortados y por tanto los FHEs que puedan aparecer en su interior no han sido contabilizados.

En lo que respecta al resto de FHEs acústicos, destaca el gran número de alargamientos (3638, uno cada 18 palabras), que resulta en parte debido a la falta de entrenamiento del anotador, que en algunos casos confundía segmentos enfatizados (es decir, más energéticos) con alargamientos, y en parte a un exceso de celo, ya que muchos de los alargamientos son

Tabla 4: Cuentas de aparición de los FHEs (NFHE) y promedio de FHEs por cada 100 palabras (NFHE/100P).

FHE	NFHE	NFHE/100P
Aspiración	3005	4.62984
Click labios	161	0.248055
Tos	40	0.0616285
Ruido genérico	530	0.816578
Pausa silencio	1945	2.99669
Pausa hablada /a/	25	0.0385178
Pausa hablada /e/	800	1.23257
Pausa hablada /m/	323	0.49765
Pausa hablada /?/	616	0.949079
Alargamiento	3638	5.60512
Mala pronunciación	1013	1.56074
Palabra cortada	212	0.326631
Afirmación gutural	295	0.45451
Siglas	36	0.0554657
Palabra extranjera	187	0.288113
Solapamiento de voces	1808	2.78561
CONTINUA	440	0.677914
NO_TRANSCRITO	71	0.109391
CORTE	9	0.0138664

prácticamente inapreciables y podrían haber sido absorbidos como segmentos normales. En cualquier caso, aunque se redujera a la mitad, el número de alargamientos es muy significativo, sobre todo si lo comparamos con el número de pausas de silencio (1945, también bastante sobredimensionado, ya que muchas de estas pausas duran entre 100 y 200 milisegundos) y con el número total de pausas habladas (1764). Esto pone de manifiesto que los tres tipos de fenómenos, como recursos del habla espontánea, resultan igualmente útiles, bien para mantener la posesión del turno, bien para señalar a la audiencia que debe esperar mientras el hablante elabora el discurso. Los números de la tabla dan un reparto de 26.74% para pausas de silencio, 24.00% para pausas habladas y 49.52% para alargamientos. La distribución interna de las pausas habladas dice que la realización /e/ es con mucho la más frecuente en castellano (800 instancias, un 45.35% del total), seguida de /m/ (323,un 18.31% del total) y la realización /a/, cuya presencia es prácticamente anecdótica (25, 1.42% del total). La realización indeterminada /?/ (616 instancias, un 24.92% del total) corresponde por lo general a glotalizaciones o distorsiones de sonidos vocálicos, que suelen suceder al final de alargamientos o de pausas habladas de otro tipo, o también en determinados contextos que no permiten articular una de las realizaciones más ortodoxas.

Finalmente, dentro de los fenómenos léxicos destaca el gran número de *malas* pronunciaciones (1013, un 58.12% de todos los fenómenos léxicos). Esto es debido a las estrictas condiciones impuestas al anotador, que debía anotar cualquier desviación de la pronunciación *standard*, por ejemplo ”Madri”en lugar de ”Madrid”, ”pasao”en lugar de ”pasado”, ”desir”en lugar de ”decir”, etc. Estas condiciones buscan obtener modelos acústicos más ajustados, si bien ello implica una modelización explícita de las variantes de pronunciación al construir el vocabulario del reconocedor. Es destacable también el número de afirmaciones guturales (295 instancias), lo cual confirma la necesidad de modelar a nivel acústico tales fenómenos.

8. Trabajo futuro

Evidentemente, queda pendiente completar la primera fase de cortado y anotación, y después efectuar la segunda fase, correspondiente a los fenómenos sintácticos y pragmáticos, que en la literatura sobre disfluencias reciben una atención primordial, sobre todo desde los campos de la psicolingüística y el procesamiento del lenguaje natural, pero también desde el punto de vista de los sistemas de reconocimiento y comprensión del habla, ya que por un lado los marcadores de discurso facilitarían la tarea de segmentación de los turnos en unidades más pequeñas, como frases, sintagmas, etc. y por otro lado, la detección de reformulaciones resultaría vital para evitar errores en la interpretación de la secuencia de palabras reconocida.

Una vez realizado el cortado y anotación de CRLEC-EHU, el estudio brevemente esbozado en esta comunicación deberá ser realizado con mayor profundidad, especialmente con objeto de comparar tanto la tipología como la frecuencia de los fenómenos que aparecen en situaciones totalmente espontáneas como las de CRLEC-EHU, con respecto a los que aparecen en el contexto de aplicaciones de diálogo hombre-máquina como INFOTREN.

Finalmente, están en curso de realización varios experimentos de reconocimiento de habla espontánea sobre CRLEC-EHU-1. A pesar de que los resultados que se obtengan sin duda estarán afectados por las adversas condiciones de grabación de CRLEC, podrán servir como referencia para otros experimentos sobre bases de datos *menos espontáneas*.

9. Referencias

[1] A. Bonafonte, P. Aibar, N. Castell, E. Lleida, J.B. Mariño, E. Sanchís, and I. Torres, "Desarrollo de un sistema de diálogo oral en dominios restringidos," in *Actas de las I Jornadas en Tecnología del Habla*, University of Sevilla, Spain, November 6-10 2000, Project website: http://gps-tsc.upc.es/veu/basurde.

[2] L. J. Rodríguez, I. Torres, and A. Varona, "Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish," in *Proceedings of the Workshop on Disfluency in Spontaneous Speech*, University of Edinburgh, Scotland, August 29-31 2001, pp. 1–4.

[3] L. J. Rodríguez, I. Torres, and A. Varona, "Evaluation of sublexical and lexical models of acoustic disfluencies for spontaneous speech recognition in Spanish," in *Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH)*, Aalborg, Denmark, September 2-7 2001.

[4] CRLEC, "Corpus de referencia de la lengua española contemporánea," Corpus website: http://www.lllf.uam.es /corpus/corpus_oral.html, 1992.

[5] Almudena Ballester, Carmen Santamaría, and Francisco A. Marcos-Marín, "Transcription conventions used for the corpus of spoken contemporary spanish," *Literary and Linguistic Computing*, vol. 8, no. 4, pp. 283–292, 1993.

[6] L. J. Rodríguez, I. Torres, and A. Varona, "Manual para el etiquetado de disfluencias," Technical Report BS12BV30, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad del País Vasco, May 2000.