La anotación del habla en corpus de vídeo

Manuel Alcántara Plá
DFKI GmbH
Saarbrücken
manuel.alcantara@dfki.de

Resumen: La anotación lingüística del habla en corpus multimodales es una labor tan nueva como costosa, pero también es prometedora para tareas como la extracción y el resumen de contenido, así como para abrir nuevos caminos en el análisis del habla espontánea. El presente artículo repasa el estado de la cuestión en los distintos niveles de análisis con ejemplos de proyectos internacionales y nacionales, resaltando la importancia de encontrar una base común a pesar de la actual falta de estándares.
Palabras clave: anotación lingüística, corpus multimodal, estandarización

Abstract: The linguistic tagging of spoken language in multimodal corpora is a new and complex task. However, its possibilities for other tasks such as content extraction/summarization and for further linguistic analysis are promising. This article reports on the state-of-the-art in the different analysis levels including experiences from international projects and stressing the importance of a common ground in spite of the current lack of standards.
Keywords: linguistic tagging, multimodal corpora, standardization

1. Introducción

El análisis lingüístico de las transcripciones del habla extraídas de corpus de vídeo es un campo de investigación muy reciente dentro de la lingüística computacional. La cantidad de colecciones de habla es también muy limitada y aún más si sólo tenemos en consideración los corpus que incluyen algún tipo de anotación lingüística. La aplicación en los corpus orales de los etiquetados diseñados para la lengua escrita requiere de una adaptación costosa que empieza incluso en las bases teóricas gramaticales, sólo probadas hasta ahora -en el mejor de los casos- sobre textos escritos.

La necesidad de corpus anotados de estas características es cada vez más obvia y acuciante tanto en la lingüística como en las aplicaciones enmarcadas dentro de la inteligencia artificial. Por este motivo, el número de corpus de habla espontánea ha crecido de manera importante durante los últimos años y su desarrollo ha suscitado un buen número de cuestiones que se están multiplicando ahora al incluir las relaciones entre el habla y el resto de elementos presentes en un corpus multimodal.

Este artículo describe cuáles son los problemas más graves encontrados en este nuevo reto de la lingüística de corpus así como algunas de las medidas que han sido adoptadas hasta el momento para resolverlos. Dado que muchos de los proyectos mencionados están aún desarrollándose, he optado por citar a pie de página el respectivo sitio de internet en cada primera mención para facilitar el acceso a su estado actual. En las conclusiones finales, se resaltará la necesidad de una base de trabajo común para el etiquetado del habla.

2. La transcripción del habla

La anotación del habla depende en primera instancia de las características de la transcripción. La mayoría de las transcripciones se realizan o generan siguiendo las convenciones ortográficas de la lengua que se trate tal y como recomiendan, entre otros, el Corpus de Habla Holandés (CGN)[1], el Corpus Nacional Británico (BNC)[2] y el Corpus de Japonés Espontáneo (CSJ)[3]. Debido a que la transcripción fonética se considera aún demasiado compleja para el habla espontánea, los corpus que incluyen transcripciones de este tipo en lugar -o además- de ortográficas se basan en alfabetos fonémicos en lugar de fonéticos. Con este fin, se utiliza el AFI en la última versión del UAM-C-Oral-Rom (Moreno et al., 2005) y en el Corpus Taiwanés de Lengua Infantil (TAICORP) (Tsay, 2005), el sistema

[1] http://lands.let.kun.nl/cgn/ehome.htm
[2] http://www-dev.natcorp.ox.ac.uk/
[3] http://www2.kokken.go.jp/ csj/public/

ISSN: 1135-5948

SAMPA[4] en el CGN y las sílabas Kana en el CSJ. Precisamente este último es un buen ejemplo de intento de realizar transcripciones *fonéticas* con el objetivo de etiquetar fenómenos como la palatalización. Sus conclusiones no son, sin embargo, muy alentadoras puesto que no fueron capaces de etiquetar todos los rasgos fonéticos que pretendían originalmente por el bajo nivel de acuerdo que se encontraron entre los anotadores.

La transcripción, aun siendo ortográfica, implica un buen número de decisiones arbitrarias tales como el tratamiento de las mayúsculas, los acrónimos y los símbolos, la puntuación, las marcas diacríticas, los números, los préstamos lingüísticos y las palabras que no aparecen normalmente en fuentes escritas. Entre estas últimas, son especialmente importantes por su frecuencia las decisiones con respecto a los rasgos dialectales, las interjecciones y los marcadores discursivos. A este respecto, es importante señalar la existencia de guías como el Estándar de Codificación de Corpus (XCES) del grupo EAGLES[5], que desgraciadamente sólo cubren los aspectos más generales.

Las convenciones ortográficas han probado ser problemáticas por dos razones curiosamente opuestas. Por un lado, hay casos en los que son excesivamente ambiguas y necesitan ser restringidas. Un ejemplo es el CSJ, que hace un uso del Kanji (pictogramas chinos) y del Kana (silabario japonés) mucho más estricto que el propuesto por las normas ortográficas del japonés estándar de modo que a cada forma sólo le corresponda una cadena fónica.

Por otro lado, las convenciones pueden ser excesivamente restrictivas como para reflejar la creatividad del habla. El TAICORP es un ejemplo en el que se usa la ortografía china como base, pero se la acompaña del sistema de romanización Taiwan Southern Min para las palabras que no se pueden encontrar en los diccionarios tradicionales.

Otro aspecto importante a tener en cuenta a la hora de analizar un corpus de habla es el modo en que se ha realizado la transcripción: de forma manual o automática. El estado actual de los sistemas de reconocimiento automático de habla (ASR) no permite obtener aún unos resultados fiables para el análisis lingüístico (Alcántara y Declerck, 2007). Los sistemas más avanzados logran alrededor del 90 % de palabras correctas, pero sólo en las mejores condiciones (lo que significa habla con guión producida en un laboratorio). Si el corpus incluye diferentes hablantes y las grabaciones han sido realizadas en contextos *naturales*, el porcentaje baja a bastante menos de la mitad.

3. Elementos no lingüísticos

Las transcripciones de habla suelen incluir la anotación de rasgos no lingüísticos que ayudan a su posterior análisis. Estos datos, generalmente en la cabecera del documento o en un documento externo, están relacionados tanto con la transcripción como con la fuente original del vídeo. Con respecto a los documentos, datos típicos son su tamaño, su calidad acústica, los formatos, las fuentes, los hablantes que aparecen (generalmente con algunas características como su edad, nivel educativo y género), los responsables de las transcripciones y los enlaces a otros archivos o documentos relacionados. La información sobre la calidad acústica suele acompañarse de detalles de la grabación tales como el tipo de micrófonos, la frecuencia o si el tratamiento es digital o analógico. La información sobre la fuente es especialmente importante si los textos han sido tomados de corpus preexistentes. En cuanto a los enlaces a otros documentos, es recomendable realizarlos a través de un documento externo de modo que sea más sencilla su gestión y la posibilidad de compartir o reutilizar los contenidos del corpus. El marco europeo Isle Meta Data Initiative [6] está proponiendo un estándar para este tipo de gestión de corpus multimodales/multimedia.

En algunos casos, es fundamental la inclusión de información sobre el contexto y sobre los rasgos sociolingüísticos de la interacción contenida en el documento (como, por ejemplo, en CHILDES[7] o C-Oral-Rom). Etiquetas típicas sobre el contexto son las condiciones en las que se produjo la grabación (incluyendo el papel que tuvo el grabador y el nivel de espontaneidad), la fecha y el lugar en que se produjo. Las anotaciones sociolingüísticas informan sobre los participantes de la interacción (nombres, edades y lugares de nacimiento, géneros, papel en la conversación, nivel educativo, etc.) y son un criterio común para

[4] http://www.phon.ucl.ac.uk/home/sampa/home.htm
[5] http://www.cs.vassar.edu/XCES/
[6] http://www.mpi.nl/IMDI/
[7] http://childes.psy.cmu.edu/

el diseño de los corpus (p.ej. CGN, CHILDES o C-Oral-Rom). Si el discurso está dividido en turnos, un identificador único se relaciona con cada participante para permitir referencias en el diálogo a la información del hablante. Otros rasgos sociolingüísticos como el dialecto o el registro son, aunque también frecuentes, más dependientes del objetivo del corpus. El CSJ, por ejemplo, incluye datos específicos sobre el nivel de fluidez, de expresividad y de claridad articulatoria de los hablantes.

Por último, algunas anotaciones legales pueden ser obligatorias dependiendo de la legislación vigente. El consentimiento de los hablantes a ser grabados y los derechos de la propiedad intelectual tienen que aparecer explícitos en los corpus de la Unión Europea. Los consentimientos deben explicitar si el sonido puede ser transcrito, usado para la investigación y publicado. Aunque los derechos de propiedad intelectual son más típicos de los documentos escritos, también son relevantes en grabaciones literarias o con valor científico (por ejemplo, conferencias) así como en documentos tomados de medios de comunicación. Este aspecto puede repercutir en el valor del corpus de dos maneras diferentes. Por un lado, las ventajas de un corpus que cuenta con todos los permisos para su utilización y publicación son evidentes para una investigación exitosa. Por otro lado, estos requisitos legales pueden comprometer la espontaneidad de lo grabado puesto que es difícil lograr una interacción natural después de haber advertido a los interlocutores de que sus palabras no van a ser sólo grabadas, sino también minuciosamente analizadas y probablemente publicadas.

Como ocurre también con los demás niveles de anotación en el corpus, las etiquetas elegidas para los elementos no lingüísticos difieren completamente entre los distintos proyectos. Por este motivo, son de gran importancia iniciativas como la citada IMDI, que nos facilitarán en el futuro tanto el diseño de nuevos corpus como la utilización de los ya existentes.

4. Los límites prosódicos

La falta de una puntuación ortográfica en la lengua oral le da una especial relevancia a otros criterios más lingüísticos, en especial los límites prosódicos (p.ej. las proferencias) y pragmáticos (p.ej. los actos de habla). Debemos señalar, no obstante, que existen corpus, generalmente no entre los más recientes, que sí se guían por la puntuación (p.ej. el CORLEC[8]). El análisis de este último muestra que la puntuación normativa influye a veces en la fiabilidad de la transcripción. El transcriptor tiende a adaptar lo que escucha a las formas normativamente correctas ya que en muchas ocasiones no es posible de otro modo ponerle puntos y comas al habla espontánea.

Como consecuencia en parte de que los estudios se hayan centrado tradicionalmente en la lengua escrita, las unidades de análisis prosódicas son todavía controvertidas en cuanto a su definición y nomenclatura. La proferencia (*utterance*) es el término más común (Cresti y Moneglia, 2005; Miller y Weinert, 1998), pero no hay acuerdo en cuanto a su definición. Para algunos corpus como el CIAIR-Corpus de Diálogos en Coches (Kawaguchi et al., 2005) o el CSJ, los silencios son las pistas determinantes, pero la mayoría de corpus combinan criterios de otros niveles lingüísticos, sobre todo pragmáticos y sintácticos. Estos criterios son, no obstante, también discutidos con frecuencia. Mientras que los pragmáticos se critican por basarse en los actos de habla de Austin, considerados a menudo demasiado subjetivos para una anotación extensa y coherente, los sintácticos se critican por la dificultad de aplicar reglas fundamentadas en la lengua escrita sobre textos que tienen características diferentes como, por poner un ejemplo, un tercio de oraciones no verbales (Cresti y Moneglia, 2005).

Algunos proyectos proponen criterios mixtos para evitar estos problemas. El corpus TRAINS93, por ejemplo, se basa en dos claves para establecer los límites prosódicos: por un lado, se da una ruptura en el discurso del hablante y otro hablante interviene; por otro lado, se produce una ruptura en la entonación, en la sintaxis (coincidencia con un límite de categoría sintáctica) o hay una respiración (Heeman y Allen, 1995). En C-Oral-Rom, se distingue entre proferencias simples y complejas (con una o más de una unidad tonal) y se comparan las proferencias con los actos de habla de Austin (Austin, 1962) y las *unidades tonales* con las unidades informativas de Halliday (Halliday, 1976), pero siempre considerando los cambios entonativos la pista más

[8] ftp://ftp.lllf.uam.es/pub/corpus/oral/

determinante a la hora de anotar límites, con un fuerte protagonismo de los perfiles terminales (Crystal, 1975). Cabe señalar que este último ejemplo lo es de una experiencia exitosa puesto que el proyecto contó con un 95 % de acuerdo ent re los anotadores.

Otras unidades han sido utilizadas en otros proyectos dependiendo del objetivo de sus análisis. Por poner dos ejemplos distintos, el CGN tiene anotadas las sílabas prominentes, los límites prosódicos entre palabras y los alargamientos segmentales (Hoekstra et al., 2002) mientras que el sistema de Multilevel Annotation Tools Engineering (MATE[9]) etiqueta grupos de acentos, pies, sílabas y moras.

Entre las aproximaciones más acústicas, el sistema TOBI[10] (Tone and Break-Index) se ha utilizado como estándar para la transcripción de entonación y estructuras prosódicas al menos para el inglés, el alemán, el japonés, el coreano y el griego, con las adaptaciones pertinentes en cada caso. Junto con el contorno de la frecuencia fundamental y la transcripción ortográfica, el TOBI incluye un nivel para los tonos y otro para los índices de los distintos límites. Las etiquetas transcriben las variaciones de tono como secuencias de tonos altos (H) y bajos (L) e incluyen marcas diacríticas con su función (el inventario de eventos tonales está basado en análisis autosegmentales). Los límites marcan los grupos prosódicos en una proferencia etiquetando el final de cada palabra sobre una escala del 0 (la unión perceptible más fuerte con la siguiente palabra) al 4 (la mayor separación).

Un ejemplo de adaptación del sistema es el X-JTOBI, versión del TOBI de japonés leído para el habla espontánea[11]. Las etiquetas para los tonos y los límites fueron extendidas en el X-JTOBI para poder representar rasgos paralingüísticos propios de la entonación espontánea, incluyendo fenómenos de disfluencia tales como las pausas largas, las palabras fragmentadas y las pausas dentro de una palabra.

Los diferentes sistemas existentes no se diferencian sólo en el modo en que se definen los conceptos que manejan, sino también en cómo estos son anotados. Una convención muy extendida es la de Gross (Gross, Allen, y Traum, 1993) con las proferencias separadas en distintas líneas o incluso ficheros, numeradas según el número de turno y el número de proferencia dentro de ese turno (como describen Nakatani y Traum sobre su corpus (Nakatani y Traum, 1999)). Otra convención frecuentemente utilizada es la del asterisco (*) junto a un código que identifique al hablante para marcar el inicio de un turno y la de las dobles barras (//) para marcar los límites prosódicos (p.ej. en CHILDES y en C-Oral-Rom).

Además de los límites prosódicos, la lengua hablada incluye otros fenómenos que también suelen etiquetarse dentro de la anotación prosódica a pesar de que, dadas sus peculiaridades, afectan a prácticamente todos los niveles (González et al., 2004). El citado artículo los clasifica en dos grupos: rasgos de producción y rasgos de la interacción. Los primeros incluyen, entre otros, las palabras fragmentadas, los apoyos vocálicos y los reinicios. Los segundos son los cambios de turnos y los solapamientos.

5. Unidades morfosintácticas

La anotación morfosintáctica de la lengua hablada es diferente a la de la escrita y no puede llevarse a cabo con los sistemas de etiquetado preexistentes. La morfosintaxis de la lengua oral es aún controvertida incluso en los aspectos más fundamentales. Por poner un ejemplo básico, algunos corpus utilizan los blancos para delimitar palabras (lo hacen así, p.ej., el BNC y el CGN) mientras que otros prefieren considerar palabras aquellos grupos mínimos de sonidos que tienen un significado propio (p.ej. el UAM C-Oral-Rom o el USAS[12]). Esta última decisión, aunque arbitraria en muchos casos, evita circunstancias como la descrita en las especificaciones del BNC, con etiquetados diferentes para formas distintas de una misma palabra (p.ej. “foxhole” o “fox hole”).

En el habla se encuentran muchas partes difícilmente categorizables dentro de las tipologías morfológicas tradicionales. Un uso común es no transcribirlas como palabras, sino a través de símbolos (o simplemente no transcribirlas en absoluto, lo que merma considerablemente la riqueza del corpus). Esta última solución fue la adoptada por los primeros corpus tales como el CORLEC, carac-

[9] http://mate.nis.sdu.dk/

[10] http://www.ling.ohio-state.edu/ tobi/

[11] http://www.ling.ohio-state.edu/research/phonetics/J_ToBI/

[12] http://www.comp.lancs.ac.uk/ucrel/usas/

terizados, como hemos visto antes, por seguir una transcripción ortográfica normativa. Los corpus más modernos están intentando ampliar la tipología para dar cabida a estas palabras, con lo que están ganando prominencia categorías que antes eran marginales como es la de los marcadores discursivos.

Como era de esperar, las características de cada lengua influyen directamente en las decisiones tomadas con respecto al análisis morfológico de modo que la anotación de corpus como el CGN y el CSJ es claramente distinta. El último, por ejemplo, distingue entre palabras cortas (de uno o dos morfemas) y largas (compuestas de varias cortas y partículas), algo que no sería pertinente en un corpus de una lengua romance o germánica. Es importante señalar que esta influencia proviene frecuentemente más de la tradición lingüística que de la lengua en sí. Un ejemplo claro es la imposibilidad de acuerdo para las clases de palabras entre los cuatro grupos de C-Oral-Rom, cuyas respectivas lenguas (portugués, italiano, francés y español) eran en teoría muy parecidas.

Precisamente las clases de palabras son la información morfosintáctica más básica y frecuente en los corpus, casi siempre acompañada de los lemas de las palabras. Los sistemas de etiquetado automático basados en métodos estadísticos como el TnT (Brants, 2000) o el de E. Brill (Brill, 1993) han demostrado resultados satisfactorios (p.ej. en los sistemas CLAWS4 (Leech, Garside, y Bryant, 1994) y GRAMPAL (Moreno, 1991)), pero siempre después de su adaptación a la lengua hablada. Así la última versión de GRAMPAL incorpora marcadores discursivos y elementos enfáticos mientras que el BNC utiliza el mencionado sistema CLAWS4 adaptándolo a algunos fenómenos propios de la oralidad como son las repeticiones. La calidad de la anotación depende también de la adaptación de las categorías que son frecuentes en la escritura puesto que sus posiciones y frecuencias no suelen coincidir con las del habla. Los marcadores discursivos y las interjecciones, por ejemplo, son en general palabras utilizadas con otras funciones al escribir, lo que dificulta su desambiguación categorial hasta el punto de haber sido obviadas hasta ahora en la mayoría de los corpus (como los mencionados CGN, EAGLES, BNC y XCES). En los corpus en los que se ha optado por adaptar la anotación, la redefinición de las categorías se ha realizado desde criterios *funcionales* (p.ej. en el UAM C-Oral-Rom) o *formales* (p.ej. en el CGN).

Más allá de los problemas de definición, no podemos olvidar aquellos heredados de la transcripción, como son la pronunciación extraña de palabras, la alta frecuencia de préstamos lingüísticos y el uso de neologismos (casi siempre a través de morfemas derivativos), que añaden gran cantidad de ruido a los análisis morfosintácticos. Por regla general, las normas de etiquetado suelen incluir un protocolo describiendo las decisiones que se han tomado para anotar estos fenómenos orales.

En cuanto a la anotación puramente sintáctica, muy pocos corpus orales la incluyen por la dificultad de distinguir automáticamente unidades complejas (sintagmas y oraciones) en el habla. Algunos ejemplos de estas experiencias son el CGN y el CSJ. Un 10 % del primero fue etiquetado semi-automáticamente con el programa ANNOTATE siguiendo un análisis de dependencias diseñado con la máxima sencillez para minimizar los costes (Hoekstra et al., 2002). El mismo criterio llevó a elegir las proposiciones como unidad de anotación de un subcorpus del CSJ de 500.000 palabras tomadas de monólogos. Las proposiciones son más sencillas de segmentar que las oraciones porque los verbos conjugados y las conjunciones se colocan al final de ellas en japonés.

6. La semántica

La anotación semántica se realiza habitualmente desde dos perspectivas en principio diferentes: la *conceptual* y la *estructural*. Los sistemas conceptuales etiquetan documentos o palabras según el campo al que pertenecen y se distinguen entre sí por el número de categorías y los criterios involucrados en sus ontologías. Por ejemplo, cada noticia grabada de los telediarios en la Digital Video Library[13] se etiqueta automáticamente dentro de una de sus 3178 categorías temáticas gracias a un algoritmo de cercanía K. Un ejemplo de etiquetado de palabras para lengua escrita y hablada -en inglés- es el USAS utilizado en el software UCREL para análisis semánticos automáticos. Incluye 232 categorías divididas en 21 campos (como "educación" o "comida") y sus reglas de desambiguación depen-

[13] http://www.open-video.org/

den de la categoría morfológica de la palabra, de sus apariciones en el mismo texto, del contexto y del dominio en el que se encuadra el discurso.

Otro caso típico de etiquetado conceptual es el del reconocimiento de *entidades propias* (NE). En el Corpus Japonés de Diálogos para Análisis de Enfermería (itoh Ozaku et al., 2005), se utilizó la herramienta NExT para extraer nombres propios, medicamentos y enfermedades de modo que se pudieran inferir fácilmente las situaciones que aparecían en cada grabación. Gracias al carácter multimodal del corpus, la desambiguación se llevaba a cabo teniendo en cuenta datos extralingüísticos como la localización en la que se encontraba la enfermera cuando pronunciaba las palabras (las enfermeras llevaban unos sensores de posición, lo que también permitía saber quién participaba en cada interacción).

La anotación estructural difiere más de la lengua escrita que la conceptual y es, por lo tanto, uno de los grandes retos en los nuevos corpus. Su atractivo es grande debido a las ya mencionadas dificultades que plantea la estructuración sintáctica del habla espontánea y aún más si se utiliza conjuntamente con la información ontológica. Uno de los escasos ejemplos ya finalizados es SESCO (Alcántara, 2005), donde las estructuras eventivas fueron utilizadas en un etiquetado que buscaba, de nuevo, la mayor simplicidad para ser flexible en el análisis de un corpus de habla espontánea sin restricciones. La anotación se basó en la estructuración composicional de tres únicos tipos eventivos (estados, procesos y acciones) que podían ser subdivididos según los argumentos que requisieran. El resultado es un ejemplo claro de la potencialidad de este tipo de etiquetados puesto que sus estructuras se están utilizando en la actualidad como base para el análisis de otros niveles lingüísticos.

Otro ejemplo es el Spanish Framenet, actualmente en desarrollo. Aunque el corpus que se utiliza en este proyecto es básicamente de lengua escrita, incluye también un 12 % de habla espontánea (alrededor de 35 millones de palabras según los datos expuestos en la página del proyecto[14]). El etiquetado estructura la lengua en *marcos* relacionando los lexemas con situaciones prototípicas que incluyen diferentes tipos de participantes. Al contrario que en SESCO, aquí el proceso no comienza en el corpus, sino en la identificación de los marcos. Una vez que el marco está definido, se buscan oraciones en el corpus que ejemplifiquen su tipo, anotando las distintas partes con las etiquetas apropiadas. El primer lexicón derivado de este trabajo está anunciado para principios del 2008.

7. La pragmática

La codificación de elementos pragmáticos ha tenido un gran avance en las últimas décadas gracias al desarrollo de sistemas aplicados para tareas específicas. Un ejemplo conocido es el Corpus de Tareas con Mapas (MTC) de la Universidad de Edimburgo (Anderson et al., 1991), que cuenta con tres niveles de anotación discursiva. En la superior, el diálogo se divide en *transacciones* en las que se completan los pasos de la tareas. Esas tareas se subdividen a su vez en *juegos conversacionales* similares a lo que Grosz y Sidner denominan *segmentos discursivos* (Grosz y Sidner, 1986). Por último, estos juegos se componen de inicios y respuestas clasificados según tipos de *movimientos conversacionales*.

También relacionado con el modelo de Grosz y Sidner, el CSJ ha sido anotado con un sistema basado en el IAD de Nakatani (Nakatani et al., 1995). El anotador tiene que dividir manualmente el discurso en segmentos asignándoles su finalidad. El manual del proyecto aclara que ésta es una labor muy costosa que requiere trabajo en equipo y decisiones complejas. Sin embargo, han sido capaces de etiquetar un pequeño subcorpus de monólogos con *patrones de cohesión* (es decir, "oraciones que tienen una relación local entre ellas") y *subhistorias* (la finalidad de una parte completa del discurso).

Un ejemplo diferente, más conectado con los aspectos morfosintácticos, es el esquema propuesto por Marco de Rocha para el análisis de expresiones anafóricas en la lengua hablada (de Rocha, 1997). Cada discurso se etiqueta con un tema que está formado por segmentos, los cuales son anotados según sus *funciones discursivas* (p.ej. introducción de un tema). Por último, las expresiones anafóricas son etiquetadas junto a su tipo, el tipo morfosintáctico del antecedente, el estatus de topicalidad del antecedente y el tipo de conocimiento necesario para procesarla.

Nakatani y Traum ofrecen un ejemplo de etiquetado más centrado en los hablan-

[14] http://gemini.uab.es:9080/SFNsite

tes. Anotan *unidades de elementos comunes* (CGU) que marcan "el acuerdo entre los hablantes sobre su entendimiento de lo que se dice" (Nakatani y Traum, 1999). Cada CGU contiene las oraciones necesarias para fundamentar un contenido, mientras que varias de estas unidades son anotadas juntas como *unidades intencionales o informativas*.

Otro de los corpus mencionados anteriormente, el CIAR, también incluye la anotación de *actos de habla* con unas etiquetas denominadas *marcas de intención* (LIT), que indican la intención que tienen las oraciones para el hablante. Cada LIT está formado por cuatro niveles: acto discursivo, acción, objeto y argumento, y se asume que la *oración* -vinculada al LIT- es la unidad fundamental del diálogo. Varias oraciones forman una *parte del discurso* (PoD) que aparece etiquetada con la tarea principal que esté llevando a cabo el hablante.

8. El alineamiento del texto con el sonido y la imagen

La anotación prosódica está estrechamente relacionada con el alineamiento del sonido y el texto ya que se suelen tomar unidades de la prosodia para realizar el proceso. Las aplicaciones automáticas para el alineamiento se basan en rasgos acústicos (físicamente reconocibles) que generalmente se corresponden con perfiles terminales, pero sus resultados son aún muy limitados. Algunos proyectos han utilizado unidades de definición más compleja, pero realizando la tarea manualmente (C-ORAL-ROM), mientras que otros han sacrificado esta complejidad para facilitar su automatización, tomando unidades como las pausas mayores de tres segundos (p.ej. el CGN) o los fonemas (realizado con un sistema HMM para el CSJ y siendo revisado después manualmente).

El alineamiento del habla con las imágenes en corpus multimodales es un campo en el que apenas contamos con experiencias, pero los primeros intentos ya han evidenciado la dificultad de sus retos, centrados especialmente en la conciliación entre los rasgos lingüísticos y los puramente audiovisuales. La segmentación del documento en unidades que sean relevantes tanto desde un punto de vista visual como lingüístico es el primer problema a solucionar. Las divisiones para el análisis audiovisual se basan en rasgos acústicos y de la imagen detectados automáticamente, como pueden ser el cambio de cámara o el movimiento de la imagen. Estas unidades (denominadas *shots*) raramente coinciden con los límites lingüísticos. Aunque sería lo ideal para el análisis del contenido, parece que la relevancia de la segmentación visual para la anotación lingüística es escasa (Alcántara y Declerck, 2007).

9. Conclusiones para el futuro

La multimodalidad supone un paso más en la evolución que se ha venido produciendo en la lingüística de corpus durante las últimas dos décadas (Moreno, 2002). Esta nueva generación de corpus ofrece un gran potencial para el análisis lingüístico y el desarrollo de aplicaciones de inteligencia artificial dentro de un contexto en el que la dependencia de los corpus y de los avances tecnológicos está resultando ser claramente bidireccional. No obstante, las características de estas colecciones hacen que requieran de un esfuerzo importante en la anotación tanto si se parte de la reutilización de sistemas como si se crean otros nuevos.

El mayor problema que afrontamos al desarrollar corpus multimodales es, como se deduce de lo descrito en las secciones previas, la falta de una estandarización eficiente, un problema que en parte viene heredado de la brevísima tradición en el trabajo con corpus de habla (Llisterri, 1997). Como hemos descrito en este artículo, cada nivel de análisis cuenta en la actualidad con experiencias tan interesantes como dispares y la discrepancia no se da únicamente en el plano teórico, sino también en la forma en que se codifican las informaciones. El uso cada vez más extendido del XML (lo que incluye también la traducción de formatos antiguos a este formato) nos permite a este respecto compartir recursos con mayor facilidad ahora que en el pasado, pero compatibilizar las diferentes informaciones sigue resultando una tarea ardua.

El contar con sistemas compatibles entre sí nos ayudaría a reutilizar y mejorar recursos ya existentes. Además, es un requisito indispensable para poder realizar investigaciones que impliquen más de un nivel lingüístico. Este último paso facilitaría la resolución de muchos de los problemas aquí planteados. Un ejemplo claro es la mencionada segmentación del documento en unidades pertinentes lingüísticamente. Las experiencias con habla espontánea demuestran que no es una tarea

fácil en ningún nivel, pero el uso combinado de la información obtenida en varios de ellos nos está dando resultados prometedores (Alcántara, 2007).

Un problema relacionado es el de la excesiva especificidad de muchas anotaciones. Por poner un ejemplo, pocos proyectos de los mencionados en este artículo están diseñados para etiquetar más de una lengua. De hecho, en la mayoría de los casos ni tan siquiera se pretende cubrir una lengua completa, sino una pequeña parte estrictamente delimitada por rasgos como el dominio o el tipo de hablantes, dados por los objetivos inmediatos de cada proyecto. Esta especificidad dificulta también el intercambio de información y, lo que es aún más grave, impide la escalabilidad de los sistemas. Como es lógico, los estudios lingüísticos que estos corpus permiten tampoco se pueden extrapolar a las características generales de la lengua.

Nos encontramos en un momento positivo porque contamos, por primera vez, con corpus multimodales y las posibilidades técnicas necesarias para etiquetarlos incluyendo información lingüística. Sin embargo, la escasez de experiencias comunes y la necesidad de conseguir objetivos en cada caso diferentes nos sitúan en una posición en la que los avances no son tan importantes como cabría esperar por el interés y el trabajo dedicados. El progreso en la anotación del habla necesita que nos esforcemos en encontrar una base común tanto en lo que se etiqueta como en cómo se etiqueta. En otras ocasiones, como ocurrió con los corpus de lengua escrita, los *estándares* se han ido imponiendo de forma natural por sistemas de etiquetado que por diversos motivos han gozado de una aceptación mayoritaria, pero parece que la rapidez de los desarrollos actuales recomienda la puesta en marcha de propuestas como la de, por ejemplo, la red europea de excelencia K-Space[15], dirigidas a acelerar ese proceso de convergencia. Esfuerzos de estandarización como el ya mencionado de Eagles (y otros como el de la Text Encoding Initiative[16] o el de la Red de Corpus Europeos de Referencia-NERC (Teubert, 1993)) son una base de gran interés para este fin que debería tenerse en cuenta para los futuros desarrollos.

[15]http://kspace.qmul.net/
[16]http://www.tei-c.org/

10. Agradecimientos

El autor quiere mostrar aquí su agradecimiento a la citada red europea de excelencia K-Space (Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content, FP6-027026) de la que forma parte y, especialmente, a los otros miembros del equipo del DFKI que participan en dicha red, Thierry Declerck y Paul Buitelaar. El trabajo de este artículo ha sido financiado con una beca posdoctoral del Ministerio de Educación y Ciencia.

Bibliografía

Alcántara, Manuel. 2005. *Anotación y recuperación de información semántica eventiva en corpus*. Ph.D. tesis, Universidad Autónoma de Madrid.

Alcántara, Manuel. 2007. Merging semantics and prosody to structure spoken language. En *Proceedings of the IWCS-7*.

Alcántara, Manuel y Thierry Declerck. 2007. Shallow semantic analysis of asr transcripts associated with video shots. En *Proceedings of the IWCS-7*.

Anderson, A., M. Bader, E. Bard, E. Boyle, G.M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H.S. Thompson, y R. Weinert. 1991. The hcrc map task corpus. *Language and Speech*, 34.

Austin, J.L. 1962. *How to do Things With Words*. Harvard University Press.

Brants, Thorsten. 2000. Tnt - a statistical part-of-speech tagger. En *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*.

Brill, E. 1993. *A Corpus-Based Approach to Language Learning*. Ph.D. tesis, Philadelphia.

Cresti, Emanuela y Massimo Moneglia, editores. 2005. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Language*. Benjamins.

Crystal, David. 1975. *The English tone of voice: essays in intonation, prosody and paralanguage*. Edward Arnold.

de Rocha, Marco, 1997. *Corpus-Based and Computational Approaches to Discourse Anaphora*, capítulo Corpus-Based Study of Anaphora in English and Portuguese. UCL Press.

González, Ana, Guillermo de la Madrid, Manuel Alcántara, Raúl de la Torre, y Antonio Moreno. 2004. Orality and difficulties in the transcription of spoken corpora. En *IV International Conference on Language Resources and Evaluation (LREC2004)*.

Gross, Derek, James F. Allen, y David R. Traum. 1993. *The Trains 91 Dialogues*. University of Rochester.

Grosz, B.J. y C.L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3).

Halliday, M.A.K. 1976. *System and Function in Language*. Oxford University Press.

Heeman, Peter A. y James F. Allen. 1995. *The Trains spoken dialog corpus (CD-ROM)*. Linguistic Data Consortium.

Hoekstra, H., M. Moortgat, B. Renmans, M. Schouppe, I. Schuurman, y T. van der Wouden. 2002. Cgn syntactische annotatie. Informe técnico, Radboud University Nijmegen.

itoh Ozaku, Hiromi, Akinori Abe, Noriaki Kuwahara, Futoshi Naya, Kiyoshi Kogure, y Kaoru Sagara. 2005. Building dialogue corpora for nursing activity analysis. En *Proceedings of the LINC05*.

Kawaguchi, Nobuo, Shigeki Matsubara, Kazuya Takeda, y Fumitada Itakura. 2005. Ciair in-car speech corpus: Influence of driving status : Corpus-based speech technologies. *IEICE transactions on information and systems*.

Leech, G., R. Garside, y M. Bryant. 1994. Claws4: The tagging of the british national corpus. En *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*.

Llisterri, Joaquim. 1997. Transcripción, etiquetado y codificación de corpus orales. Seminario de Industrias de la Lengua - Fundación Duques de Soria.

Miller, J. y R. Weinert. 1998. *Spontaneous Spoken Language. Syntax and Discourse*. Oxford University Press.

Moreno, Antonio. 1991. *Un modelo computacional basado en la unificación para el análisis y la generación de la morfología del español*. Ph.D. tesis, Universidad Autónoma de Madrid.

Moreno, Antonio. 2002. La evolución de los corpus de habla espontánea: la experiencia del lli-uam. En *Actas de las Segundas Jornadas de Tecnologías del Habla*.

Moreno, Antonio, Guillermo De la Madrid, Ana González, Jose María Guirao, Raul De la Torre, y Manuel Alcántara, 2005. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, capítulo The Spanish corpus. Benjamins.

Nakatani, C. H. y David R. Traum. 1999. Coding discourse structure in dialogue (version 1.0). Informe técnico, University of Maryland.

Nakatani, Christine H., Barbara J. Grosz, David D. Ahn, y Julia Hirschberg. 1995. Instructions for annotating discourse. Informe técnico, Center for Research in Computing Technology.

Teubert, W. 1993. Phonetic/phonemic and prosodic annotation. final report. Informe técnico, IDS Mannheim.

Tsay, Jane S. 2005. Taiwan child language corpus: Data collection and annotation. En *Fifth Workshop on Asian Language Resources (ALR-05)*.