El análisis lingüístico en la transcripción automática de la lengua hablada, el Proyecto COAST

Manuel Alcántara Plá

Instituto Austriaco de Investigación de la Inteligencia Artificial (OFAI)
Freyung 6/6, 1010, Viena
manuel.alcantara@ofai.at

Resumen

El presente artículo expone los problemas principales encontrados en la transcripción automática de lengua hablada en la actualidad. Se repasan algunos de los proyectos más avanzados en las tareas que implica (transcripción, análisis morfosintáctico y semántico) y se describe con detalle el trabajo dentro del proyecto austriaco COAST de transcripción automática de informes médicos.

Palabras clave: habla, transcipción automática, corpus, semántica.

Laburpena

Artikulo honetan ahozko hizkuntzako transkripzio automatikoan orain aurkitutako arazo nagusiak azaltzen dira. Horren zereginak betzeteko (transkripzioa, analisi morfosintaktikoa eta semantikoa) berrienetakoak diren hainbat proiektu errepasatzen dira , eta zehatz adierazten da COAST izeneko austriar proiektu barnean medikuen txostenen transkripzio automatikoari buruz egindako lana.

Gako-hitzak: transkripzio automatikoari, korpus, analisi semantikoa.

Abstract

This paper shows the main problems faced nowadays when developing automatic speech recognition systems (ASR). Some of the most advanced projects are reported as well as a description of the Austrian COAST project, which deals with transcription of medical domain reports.

Key words: spoken language, ASR, corpus, Semantics.

Tabla de contenidos

1. Introducción

El análisis de la lengua hablada es uno de los mayores retos a los que nos enfrentamos en la actualidad dentro de las disciplinas de la lingüística computacional y de las tecnologías del lenguaje así como de la inteligencia artificial en general. El reto se debe a distintos factores de entre los que cabe destacar el carácter novedoso de los recursos

que nos permiten su análisis (especialmente los corpus) y, como consecuencia de ello, la falta de una tradición en el análisis sistemático de la oralidad basado en datos empíricos.

Como ocurre en otros ámbitos como el del análisis lingüístico de documentos multimedia, estos factores se suman al estado de la cuestión en otras disciplinas que determinan nuestro trabajo, en especial aquellas relacionadas con el reconocimiento automático de los sonidos. Aunque estos sistemas de reconocimiento han avanzado de forma importante en los últimos años, la calidad de sus resultados dificulta aún claramente el trabajo lingüístico posterior en una relación bidireccional: la lingüística mejora los sistemas de reconocimiento a la vez que estos nos ofrecen cada vez mejores datos para nuestra labor como lingüistas (o, dicho desde el lado pesimista, los resultados de los reconocedores no son mejores porque no cuentan con información lingüística para la que es necesario disponer de corpus transcritos automáticamente...).

En cualquier caso, se trata de un campo en evidente auge por unas comunicaciones que, tras una nueva época dorada del texto escrito por las limitaciones de Internet, parecen volver a ser fundamentalmente orales: los servicios telefónicos y los podcasts son dos claros ejemplos.

La transcripción automática se encuentra en un punto intermedio entre dos objetos de estudio que rara vez se mezclan: la lengua hablada (que es el producto de que se parte) y la lengua escrita (que es el producto que normalmente se desea obtener). Este artículo se centra en la primera parte, el análisis de lo que se dice, pero no se debe olvidar que la otra mitad es igualmente compleja. Esto se debe a que incluso un texto perfectamente transcrito es difícil de leer por dos motivos que requieren de soluciones distintas. Por un lado, la gramática de la lengua hablada es diferente de la escrita, con abundancia de elementos elididos que pueden impedir la comprensión una vez que se han eliminado la entonación y el contexto originales. Por otro, los documentos escritos responden a unas convenciones formales que guían su interpretación. Por poner un ejemplo, una carta o una enumeración con elementos dispuestos con diversas tabulaciones pueden perder su sentido si se presentan como un párrafo simple.

Los sistemas que se han utilizado hasta el momento (para reconocimiento con vocabulario extenso) se han basado en una estrategia que partía del entrenamiento del sistema y su adaptación al tipo de documentos que debía reconocer. El modelo acústico, el léxico y el modelo de habla son los tres módulos centrales. De este modo, el habla se procesa, se reconoce, se transcribe, se corrige y el resultado se utiliza para mejorar la adaptación. Este modelo tiene, sin embargo, importantes limitaciones que impiden un desarrollo óptimo de los sistemas. Por un lado, es difícil conseguir datos en cantidad (y calidad) suficientes para el entrenamiento. La creación manual de estos recursos es muy costosa en tiempo y esfuerzo. Por otro, el sistema debería ser capaz de ofrecer un producto final formateado y sin elementos típicos de la oralidad tales como los reinicios o las dudas, lo que no es posible si está entrenado sólo sobre habla.

La tarea implica diversos niveles aun olvidándonos de la parte de la generación y concentrándonos únicamente en la del reconocimiento. El trabajo lingüístico con transcripciones de habla espontánea engloba varios apartados de entre los que destacan el estándar de la transcripción fonética/prosódica (p.ej. TOBI, CHAT o MATE), la anotación morfosintáctica (p.ej. CLAWS4 o GRAMPAL) y el análisis semántico (p.ej. USAS, NeXT o SESCO). Las siguientes secciones 2., 3. y 4. presentan algunas de las propuestas más relevantes hasta el momento en cada uno de estos apartados y su relación con los estándares (para una descripción más detallada, véase Alcántara 2007). Después nos centraremos con más detalle en el caso práctico de la transcripción

automática dentro del proyecto austriaco COAST, que se encuentra actualmente en desarrollo (2007).

2. Transcripción y prosodia

El modo en que se realiza la transcripción determina los posteriores análisis. 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 Sandoval et al. 2005) y en el Corpus Taiwanés de Lengua Infantil (TAICORP, Tsay 2005), el sistema SAMPA[4] en el CGN y las sílabas Kana en el CSJ.

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.

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 sonido. 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

---

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/.

4 http://www.phon.ucl.ac.uk/home/sampa/home.htm.

5 http://www.cs.vassar.edu/XCES/.

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, datos especialmente relevantes a la hora de evaluar la transcripción automática.

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[6] 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 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, hay que tener presente que algunas anotaciones legales pueden ser obligatorias dependiendo de la legislación vigente sobre todo si se piensa publicar o comercializar el corpus.

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 El marco europeo Isle Meta Data Initiative[7], que nos facilitarán en el futuro tanto el diseño de nuevos corpus como la utilización de los ya existentes.

2.1. Prosodia

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 ortografía (p.ej. el CORLEC[8]).

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 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

---

6 http://childes.psy.cmu.edu/.

7 http://www.mpi.nl/IMDI/.

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

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 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 entre 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 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).

Los 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 et al. 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).

---

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

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

3. Anotación morfosintáctica

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[11]). 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. "fox-hole" 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, caracterizados, 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 et al. 1994) y GRAMPAL (Moreno Sandoval 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).

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

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

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. 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.

4. Análisis semántico

La anotación semántica se realiza habitualmente desde dos perspectivas en principio diferentes:

- La conceptual. 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[12] 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 dependen 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 (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.
- La estructural. 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 2007b), 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 requirieran. 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

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

expuestos en la página del proyecto[13]. 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.

5. El proyecto COAST

La Red Austriaca de Competencia para Tecnologías Avanzadas de Habla (COAST14 - Competence Network for Advanced Speech Technologies en su nombre original) y está dedicado al avance científico en el campo del reconocimiento e interpretación automáticos de habla con léxicos extensos y para aplicaciones profesionales. Es, por lo tanto, una apuesta del gobierno austriaco por avanzar en el reconocimiento del habla de modo que los resultados puedan ser comercializados o utilizados de forma industrial. Este objetivo implica diversas áreas de investigación entre las que destaca la mejora y utilización de algoritmos aplicados al reconocimiento (de los campos de estadística, la física-acústica y el procesamiento de señales), el uso de nuevas técnicas de interpretación semántica utilizando inteligencia artificial para mejorar la utilidad de los resultados y la adaptación del reconocimiento a aplicaciones concretas (así como el descubrimiento de nuevas posibles aplicaciones).

Desde un punto de vista científico y tecnológico, el proyecto se centra en la optimización automática o semiautomática de los parámetros de reconocimiento y de los modelos estadísticos, incluyendo información semántica. Los objetivos concretos se pueden resumir en tres:

- Mejora del sistema de reconocimiento de habla con un componente de transformación modal (habla –> transcripción literal –> transcripción refinada con formato).
- Adaptación automática a las peculiaridades específicas de los hablantes y de las distintas partes de un texto. Los métodos de aprendizaje tradicionales no son capaces siempre de captar las características definitorias de un fragmento o hablante ni de, consecuentemente, adaptar los modelos estadísticos. Algunos ejemplos de aspectos a los que se tiene que adaptar el sistema son la tendencia (o falta de ella) del hablante a utilizar marcadores del discurso o a dudar, su predisposición para pronunciar los signos de puntuación, su forma particular de pronunciar algunas palabras (p.ej. abreviaturas) y su ritmo de habla.
- Adaptación al nivel de señal acústica. Los sistemas actuales reducen drásticamente sus resultados cuando las condiciones no son las esperadas (por ejemplo, si hay mucho ruido ambiental o varios hablantes intervienen a la vez). Esta reducción debería evitarse a través de nuevos métodos de procesamiento de la señal y el uso de micrófonos específicos (incluyendo el uso de varios micrófonos en las grabaciones).

---

13 http://gemini.uab.es:9080/SFNsite.

14 http://www.coast.at

Para lograr estos objetivos, la red cuenta con grupos científicos tanto del mundo académico (universidades como TU Graz y centros de investigación como nuestro OFAI) como del industrial (especialmente SailLabs y Philips). Para demostrar la potencialidad del reconocimiento automático, se eligió una aplicación práctica real en la que la transcripción automática tuviera una utilidad evidente: el dictado de informes médicos (para lo que se han buscado también centros hospitalarios que quisieran colaborar en el desarrollo a cambio de poder utilizar el sistema final gratuitamente).

6. Recursos

El trabajo puramente lingüístico en el reconocimiento automático del habla requiere de la disponibilidad de unos recursos básicos que, afortunadamente, se acercan a lo ideal en nuestro caso gracias a que es uno de los puntos en que más esfuerzo se está aplicando.

6.1. Corpus paralelos

COAST es, en más de un sentido, heredero del proyecto austriaco SPARC (Semantic Phonetic Automatic ReConstruction of dictations). Aunque COAST es claramente más ambicioso y práctico que su predecesor, ambos comparten suficientes rasgos como para aprovechar los recursos existentes. Probablemente el más valioso de estos recursos es un corpus paralelo de transcripciones manuales y automáticas.

En la aplicación que se ha elegido para el proyecto, el/la doctor/a realiza su informe habitualmente después de cada intervención y lo graba con un dictófono. Esa grabación es transcrita adecuadamente por un asistente que realiza una versión limpia y ordenada del informe. Gracias a este proceso, nosotros contamos con dos versiones de cada grabación: la realizada por el transcriptor automático a partir de lo contenido en el dictófono y la elaborada manualmente por el asistente del hospital.

El corpus paralelo ofrece múltiples posibilidades de gran valor para el estudio de los errores cometidos por el reconocedor automático dentro del análisis lingüístico. La siguiente imagen muestra la pantalla de la aplicación utilizada para contrastar las dos versiones. La tercera columna es quizá la de más interés en un primer momento puesto que advierte del paso necesario para que la transcripción automática coincida con la manual: añadir un elemento, borrarlo o cambiarlo.

El alineamiento se realiza automáticamente siguiendo un algoritmo inspirado en Burns (1997), que se guía básicamente por la región contigua coincidente más larga para cada lugar. El resultado es bastante exacto salvo en aquellos casos en que un fragmento extenso ha sido añadido, movido o borrado. Esto es algo que ocurre desgraciadamente con cierta frecuencia y que se debe a decisiones tomadas por el asistente que ha transcrito el texto o por la misma persona que lo ha dictado, que puede realizar peticiones del tipo “escriba también en el informe de antes que...”.

	CorClass	CorText	ORecons Align Instr	OJust	ORecons	PrepDictation	RecText	RecAll	RecClass	Phonetic
96	Word	heart	COR	identical	heart	heart	heart	heart	Word	{h A r t}
97	Word	bypass	COR	identical	bypass	bypass	bypass	bypass	Word	{b 2 p & s}
98	Word	.	=	sim_bigram			.		AP	
99	Phr	He	=	sim_bigram	he	he	He	he	FRM	{h i}
100				@?	@?			<hes>	Word	{A m}
101	Phr	had	=	sim_unigram	had	had	had	had	Word	{h & d}
102				@?	@?			<hes>	Word	{A m}
103			>	@?	@?	been	been	been	Word	{b I n}
104	Word	begun	=	sim_unigram	began	began	began	began	Word	{b i g & n}
105	Word	with	COR	identical	with	with	with	with	Word	{w I T}
106	Word	toe	COR	identical	toe	toe	toe	toe	Word	{t O}
107	Word	amputation	COR	identical	amputation	amputation	amputation	amputation	Word	{& m p j @ t Y S n}
108	Word	in	<	@?	@?					
109	<YEA>	1998	=	sim_unigram	ninety-eight	ninety-eight	98	ninety-eight	<NUM_B>	{n 2 n i Y t}
110	Word	,	<	@?	@?					
111	Word	has	=	pho_trigram	has	that	that	that	Word	{D & t}
112			=	pho_trigram		that	that	that	Word	{D & t}
113	Word	worked	=	pho_trigram	worked	walked	walked	walked	Word	{w Q k t}
114	Word	up	COR	identical	up	up	up	up	Word	{V p}
115	Word	to	COR	identical	to	to	to	to	Word	{t u}
116			>	@?	@?	this	this	this	Word	{D @ s}
117			>	@?	@?	ones	ones	ones	Word	{w V n z}

6.2. Léxico

Otro de los recursos fundamentales para el reconocimiento automático es el léxico, que incluye información morfológica de cada palabra. En COAST, el léxico original para el análisis lingüístico ha sido ampliado con terminología específica del dominio médico. Esta labor ha sido más sencilla para el inglés dada la gran cantidad de recursos prexistentes (como, por ejemplo, el Unified Medical Language System, UMLS). En el caso del alemán, se ha utilizado la traducción de algunas de estas colecciones a las que se ha añadido las abreviaturas típicas en alemán (muy frecuentes en los textos médicos). La actual versión contiene más de 2.2 millones de entradas para 677.608 palabras distintas. Aproximadamente 200.000 entradas del total pertenecen a la terminología médica.

6.3. Información semántica

El aspecto más novedoso del análisis lingüístico en la mejora del reconocimiento de voz se encuentra en el uso de la información semántica. Su función fundamental es, como veremos más adelante, permitir saber cómo de plausible es que aparezca una palabra en un contexto determinado.

Para ello contamos con recursos que nos dan idea de lo que significa cada término y, lo que es más importante, de las relaciones que existen entre los distintos términos (o entre los conceptos que estos expresan). La información ha sido obtenida de dos fuentes principales. Por un lado, el ya mencionado Unified Medical Language System (UMLS[15]); por otro, los documentos disponibles en la página de Internet de la European Medicines Agency (EMEA).

El UMLS es un metatesauro y una red semántica de conceptos biomédicos. La red da información detallada con 135 tipos semánticos y 54 relaciones distintas establecidas entre ellos tales como “físicamente relacionado con”, “funcionalmente relacionado con”, etc.

Los documentos de la EMEA tienen como objetivo documentar los nuevos medicamentos que se aceptan para su comercialización en la Unión Europea, información que nosotros hemos utilizado para relacionar los nombres y características

[15] http://www.nlm.nih.gov/research/umls/.

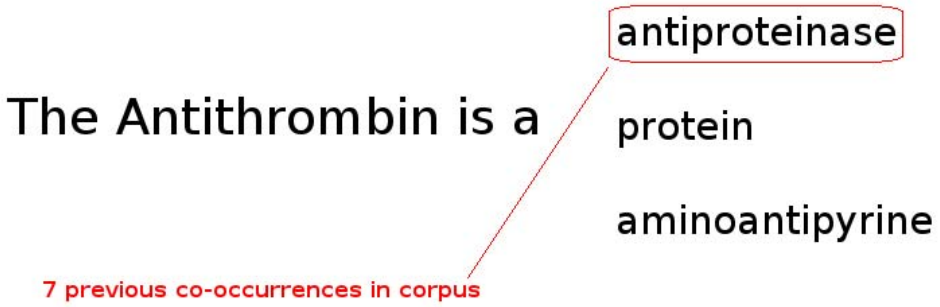
de los medicamentos con las enfermedades para las que se utilizan y los elementos químicos que los componen.

El problema fundamental al utilizar este tipo de recursos es la falta de una coherencia formal. Por ejemplo, el EMEA nos ofrece información valiosa tanto en sus páginas web como en documentos en PDF descargables. La información no aparece en ninguno de los dos casos etiquetada, por lo que nuestra labor como recopiladores consiste en recolectar todos los documentos que incluyen información, buscar ésta y traducirla a un formato estándar (en XML). Las incoherencias formales obligan necesariamente a una revisión exhaustiva de los datos recuperados automáticamente.

7. Medida de la plausibilidad a través de la semántica

La medida de la plausibilidad se realiza mediante un proceso con dos análisis paralelos que se corresponden con las relaciones semánticas entre los conceptos denotados por las palabras según los recursos anteriormente presentados y con la coaparición de los términos en corpus previamente analizados. De esta forma, a dos términos que coaparecen frecuentemente se les asume una relación semántica cuya naturaleza quizá se desconozca, pero que es igualmente relevante de cara a considerar su plausibilidad. En ambos casos, las relaciones se establecen entre el texto que ya ha sido reconocido y aceptado como válido y las candidatas a ser la siguiente palabra reconocida.

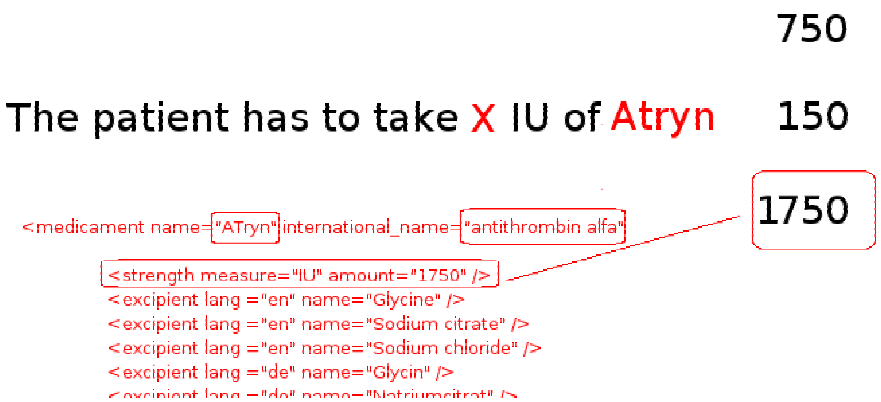
En el próximo ejemplo, el contexto reconocido está compuesto por la secuencia "The Antithrombin is a". Para la siguiente palabra, los sonidos reconocidos por el sistema se pueden corresponder con "aminoantipyrine", "protein" o "antiproteinase". Aunque ninguna de las tres candidatas aparece etiquetada en los recursos semánticos como relacionada con ninguno de los términos del contexto, "antiproteinase" es elegida como la más plausible porque ha parecido previamente junto a "Antithrombin" en 7 ocasiones.



En el siguiente ejemplo, el contexto ya reconocido es "The patient has to take X IU of Atryn", con la X en la posición en la que hay varias candidatas. En este caso, se trata de "750", "150" y "1750", algo común puesto que los sistemas de reconocimiento de habla suelen tener problemas distinguiendo números (donde un error puede ser fatal en medicina).

El sistema busca en la base de datos de relaciones semánticas donde descubrimos que "Atryn" es el nombre comercial del "Antithrombin" que veíamos en el

anterior ejemplo, y que la cantidad típica administrada es de 1750 IU (Unidad Internacional).



8. Conclusiones y trabajo futuro

Con esta breve introducción a algunos aspectos fundamentales de la transcripción automática de la lengua hablada se pretende mostrar la importancia de la aportación lingüística. La parte central de los sistemas actuales se basa en estudios estadísticos con rasgos de bajo nivel (principalmente acústicos). La mejora de estos estudios está consiguiendo resultados cada vez más exactos a la vez que está, siguiendo con una tendencia que parece general en las tecnologías del lenguaje, demostrando que existen unos límites que la estadística no puede superar si no se aporta mayor riqueza de contenidos lingüísticos.

Se ha mostrado que contar con buenos recursos es una condición necesaria para el análisis y que estos son, en la mayoría de los casos, costosos. Afortunadamente y gracias a los estándares, cada vez existen más recursos que se pueden reutilizar, pero suele requerirse un esfuerzo importante de adaptación. Es el caso, por ejemplo, del léxico utilizado en COAST.

La dificultad para conseguir generalizaciones sintácticas suficientes, en parte debida a la carencia de corpus extensos, nos lleva a aproximaciones de corte claramente semantista. En nuestro caso, el análisis semántico nos permite no depender completamente de los corpus aunque estos siguen utilizándose para enriquecer los datos que tenemos.

El trabajo futuro más evidente e inmediato en el punto en que nos encontramos es la realización de pruebas para comprobar en qué medida la aportación lingüística mejora los resultados originales del sistema de reconocimiento de habla. Se realizarán en distintas fases. Primero se medirán las diferencias absolutas para después calcular los pesos adecuados de influencia de la información lingüística en la toma de decisiones del sistema.

De forma paralela, seguiremos trabajando en la mejora de los recursos, limpiando errores contenidos en las bases de datos como consecuencia de la extracción automática de esa información y aumentando el léxico con los nuevos términos que aparezcan.

9. Agradecimientos

La parte de este artículo dedicada a la descripción del proyecto COAST se basa en los trabajos realizados en el grupo vienés de tecnologías del lenguaje del Centro Alemán para la Investigación de la Inteligencia Artificial (OFAI). El equipo de este proyecto lo dirige Harald Trost y lo conforman, además del autor, Johannes Matiasek, Alexandra Klein y Jeremy Jancsary.

También quiero mostrar mi agradecimiento a los otros equipos del proyecto y, en especial, al de Philips Speech Recognition Systems, en cuyo sistema trabajamos conjuntamente.

La asistencia y participación en el CLG8 se ha financiado con una ayuda del proyecto BRAVO-RL del MEC-CICYT (TIN2007-67407-C03-02) al que el autor pertenece como miembro externo al Laboratorio de Lingüística Informática de la UAM.

10. Referencias bibliográficas

Alcántara Plá, Manuel (2007). "Los retos en el análisis de los corpus de última generación". En *Actas del XXV Congreso AESLA*.

Alcántara Plá, Manuel (2007b). *Introducción al análisis de estructuras lingüísticas en corpus: aproximación semántica*. UAM Press.

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*. Tesis doctoral, Philadelphia.

Cresti, E., y M. Moneglia (eds.)(2005). *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Language.* Amsterdam: Benjamins.

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

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

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*. 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)*.

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*. Tesis doctoral, Universidad Autónoma de Madrid.

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

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

Tsay, Jane S (2005). "Taiwan Child Language Corpus: Data Collection and Annotation". En *Fifth Workshop on Asian Language Resources (ALR-05)*.