LOS RETOS EN EL ANÁLISIS DE LOS CORPUS DE ÚLTIMA GENERACIÓN

MANUEL ALCÁNTARA PLÁ
DFKI GmbH

RESUMEN. *El presente artículo revisa los retos surgidos en la lingüística de corpus con la aparición de las colecciones más recientes, que incluyen el sonido y las imágenes relacionadas con las transcripciones. La revisión de los distintos niveles se realiza a través de proyectos que ejemplifican los acercamientos más vanguardistas en el panorama internacional. Los problemas se presentan divididos en fonética y alineamiento de sonido/imagen y texto, morfosintaxis, semántica y pragmática.*
PALABRAS CLAVE: *lengua hablada, anotación, corpus*

RESUMEN. *This paper reports on challenges arisen in the analysis of the most recent corpora, which relate transcriptions to images and sounds. The state-of-the art in every level is described through international projects that show cutting edge approaches. The challenges are divided into prosody and alignment of sound/images and text, morphology and syntax, semantics, and pragmatics.*
KEY WORDS: *spoken language, tagging, corpora*

1. INTRODUCCIÓN

La evolución reciente de los estudios lingüísticos y de las nuevas tecnologías ha convertido las colecciones de textos en una herramienta básica dentro de la lingüística y, en especial, de su vertiente más aplicada. Esta nueva situación ha significado una vuelta a los trabajos descriptivos después de un largo periodo en el que habían predominado los de sesgo teórico e introspectivo. Las conocidas críticas a los corpus por su parcialidad (partiendo de N. Chomsky) han dado paso a la necesidad de contar con muestras reales del lenguaje y la posibilidad cada vez mayor de satisfacer la exigencia de que "cuanta más cantidad de datos, mejores los datos" (Church y Mercer 1993). La mayoría de los corpus han estado históricamente acotados a la lengua escrita y a dominios relacionados con objetivos específicos. Los más generales requieren de una inversión mayor y tienen un interés más lingüístico y aparentemente menos inmediato que el de los corpus específicos, lo que los ha convertido en un bien lamentablemente escaso.

El protagonismo de la tecnología en la reciente transición desde la *lingüística temprana de corpus* (McEnery 1996) a la *moderna* se ha repetido en dos estadios. El primer salto consistió en la inclusión del sonido en los corpus y, consecuentemente, en la posibilidad del estudio de la lengua oral también en los niveles fonético y fonológico. La inclusión de la grabación de las imágenes (vídeo) en las que se encuadran las producciones lingüísticas ha significado un segundo cambio y una nueva generación de colecciones de textos.

Los retos planteados por estas nuevas generaciones no son exclusivos de la lingüística del español. En realidad, otras lenguas gozan de una experiencia mayor en su uso, especialmente el inglés, el japonés y el alemán. Por este motivo, el presente artículo ofrece una introducción a los problemas que afrontamos en su desarrollo y uso a través de las propuestas más vanguardistas del panorama internacional. La referencia a proyectos no pretende ser, por tanto, una clasificación exhaustiva, sino una ejemplificación de la variedad de acercamientos ya existente.

2. LAS TRANSCRIPCIONES

La anotación del habla comienza con su transcripción, de la que dependerán en gran medida los análisis subsiguientes. La mayoría de los corpus han optado por seguir las normas ortográficas de sus respectivas lenguas, como es el caso del Corpus de Holandés Hablado (CGN), el Corpus Nacional Británico (BNC) o el Corpus de Japonés Espontáneo (CSJ). Esta decisión conlleva el riesgo de que la transcripción no sea fiel al habla por el intento de adaptarla a las estructuras de la lengua escrita puesto que muchos de los elementos que nos encontramos en la lengua oral no tienen una representación gráfica normativa. Las transcripciones ortográficas implican un buen número de decisiones arbitrarias de las que destacan el uso de las mayúsculas, los acrónimos, los signos de puntuación, las marcas diacríticas, los números y las palabras extranjeras. Al tratar lengua oral, también se toman decisiones sobre cómo representar las palabras que no existen en la lengua escrita (p.ej. interjecciones y marcadores discursivos), los usos dialectales y otros fenómenos típicos de la oralidad como la asimilación, la epéntesis, la metátesis y las disfluencias. Para regular estas decisiones, se han establecido estándares de codificación como el muy extendido del grupo EAGLES (xCES).

Se debe tener en cuenta que las convenciones ortográficas son en algunos casos excesivamente ambiguas (así el CSJ es más estricto en el uso de caracteres kanji y kana que las normas ortográficas estándar del japonés para evitar su ambigüedad), mientras que en otros pueden ser excesivamente restrictivas para la creatividad del habla espontánea (así el TAICORP utiliza la ortografía china, pero opta por la romanización en los casos en que la palabra no se encuentra en los principales diccionarios).

Algunos de los corpus más recientes intentan ser fieles al habla añadiendo signos fuera de la norma ortográfica (p.ej. C-ORAL-ROM) o utilizando transcripciones fonológicas (p.ej. el AFI en el TAICORP de Taiwán, las sílabas de kana en el ya mencionado CSJ o el sistema SAMPA en el CGN). Los intentos de etiquetado fonético en lugar de fonológico han resultado hasta el momento demasiado complejos y con un escaso nivel de acuerdo entre los anotadores (CSJ).

3. La anotación prosódica y el alineamiento

Aunque existen muchos corpus segmentados con la puntuación ortográfica (p.ej. CORLEC), la tendencia actual más frecuente es dividir el discurso según otros criterios, normalmente prosódicos (p.ej. en *proferencias*) o pragmáticos (p.ej. en *actos de habla*),.

Las unidades de análisis prosódicas son todavía controvertidas en cuanto a su definición y nomenclatura. Las proferencias son las unidades más comunes (Cresti y Moneglia 2005; Miller y Weinert 1998), pero no siempre se definen igual. Para algunos corpus como el CIAIR o el CSJ, los silencios son determinantes en su definición, pero la mayoría combinan criterios de otros niveles lingüísticos, sobre todo pragmáticos y sintácticos. Los primeros, sin embargo, pueden ser criticados por basarse en los actos de habla de Austin, considerados a menudo demasiado subjetivos para una anotación extensa y coherente, mientras que los sintácticos lo son por la dificultad de aplicar reglas fundamentadas en la lengua escrita sobre textos que, por ejemplo, presentan un tercio de oraciones no verbales (Cresti y Moneglia 2005).

Algunos proyectos están proponiendo criterios mixtos para evitar estos problemas. En C-Oral-Rom se comparan las *proferencias* con los *actos de habla* de Austin y las *unidades tonales* con las *unidades informativas* de Halliday, pero se consideran al mismo tiempo los cambios prosódicos como la pista más determinante a la hora de anotar límites, con un fuerte

protagonismo de los *perfiles terminales* (Crystal 1975). La mezcla parece ser exitosa puesto que el proyecto contó con un 95% de acuerdo entre los anotadores.

Otros proyectos se han centrado en unidades diferentes dependiendo del objetivo del análisis. El CGN tiene anotadas las sílabas prominentes, los límites prosódicos entre palabras y los alargamientos segmentales (Hoekstra 2003) mientras que el MATE etiqueta grupos de acentos, pies, sílabas y moras. Entre las aproximaciones más acústicas, el sistema TOBI se ha utilizado como estándar al menos para el inglés, alemán, japonés, coreano y griego, con las adaptaciones pertinentes en cada caso.

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 son aún limitadas y se basan en límites acústicos (físicamente reconocibles) que generalmente se corresponden con perfiles terminales. 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 (CGN) o los fonemas (realizado con un sistema HMM para el CSJ, después convenientemente revisado).

El alineamiento del texto con las imágenes en corpus multimodales es un campo muy reciente, pero los primeros intentos ya han evidenciado la dificultad de sus retos, que se centran especialmente en la conciliación entre los rasgos lingüísticos y los puramente audiovisuales (Alcántara y Declerck 2007).

4. INFORMACIÓN MORFOSINTÁCTICA

Que la estructura morfosintáctica de la lengua hablada es diferente a la de la escrita es algo que ha quedado patente en los intentos de uso de las antiguas anotaciones para los nuevos corpus. La anotación morfosintáctica es controvertida incluso en los aspectos más básicos. Algunos corpus utilizan los blancos para distinguir entre palabras -transcritas- (p.ej. BNC y CGN) mientras que otros prefieren considerar palabras aquellos grupos mínimos de sonidos que tienen un significado propio (p.ej. UAM C-Oral-Rom, USAS). Esta última decisión, aunque arbitraria en muchos casos, evita circunstancias como la descrita en las especificaciones del BNC, con etiquetados diferentes para una misma palabra ("fox-hole" o "fox hole").

Las características de cada lengua influyen en estas decisiones. El CSJ 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. Cabe señalar que esta influencia proviene frecuentemente más de la tradición lingüística que de la lengua misma. Un ejemplo claro es la imposibilidad de acuerdo para las clases de palabras entre los cuatros grupos de C-Oral-Rom, cuyas respectivas lenguas (portugués, italiano, francés y español) eran en realidad muy parecidas.

Precisamente las clases de palabras son la información morfosintáctica más básica y frecuente en los corpus, casi siempre acompañada de los lemas. Los sistemas de etiquetado automático basados en métodos estadísticos como el TNT o el de Brill han arrojado resultados satisfactorios (p.ej. CLAWS4 o GRAMPAL). Algunas de las categorías utilizadas coinciden con las de los corpus escritos (nombres, adjetivos, verbos, etc.), pero otras son típicas de la lengua hablada (p.ej. los marcadores discursivos y los elementos enfáticos). La calidad de la anotación depende tanto de la inclusión de las últimas como de la adaptación de las primeras, puesto que sus posiciones y frecuencias no suelen coincidir con las de la

escritura. Los marcadores y las interjecciones, por ejemplo, son en general palabras utilizadas con otras funciones en la escritura, lo que dificulta su desambiguación categorial hasta el punto de haber sido obviadas hasta ahora en la mayoría de los corpus (CGN, EAGLES, BNC, XCES, etc.).

De este modo, aunque la mayoría de proyectos han optado por reutilizar herramientas preexistentes, su éxito depende del grado de adaptación de sus reglas, pudiéndose redefinir las categorías siguiendo criterios *funcionales* (p.ej. en el UAM C-Oral-Rom) o *formales* (p.ej. en el CGN).

Más allá de los problemas de definición, tampoco podemos olvidar aquellos relacionados con 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.

En cuanto a la anotación sintáctica, muy pocos corpus orales la incluyen por la dificultad de distinguir unidades (*sintagmas* y *oraciones*) automáticamente en el habla. Como ejemplo de una de estas experiencias, un 10% del CGN fue etiquetado semi-automáticamente (con el programa ANNOTATE) siguiendo un análisis de dependencias diseñado con la máxima sencillez para minimizar los costes (Hoekstra 2003). El mismo criterio llevó a elegir las *proposiciones* como unidad de anotación de un subcorpus del CSJ de 500.000 palabras tomadas de monólogos, puesto que estas fueron consideradas más sencillas de anotar que las oraciones por la colocación en japonés de los verbos conjugados y de las conjunciones al final de las proposiciones.

5. Anotación semántica

La anotación semántica se realiza habitualmente desde dos perspectivas: la *conceptual* y la *estructural*. Los sistemas conceptuales etiquetan documentos (p.ej. la Digital Video Library) 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. Un ejemplo para el análisis de lengua escrita y hablada –en inglés- es el USAS utilizado en el software UCREL para la desambiguación. Incluye 232 categorías divididas en 21 campos (como "educación" o "comida") y sus reglas dependen de la categoría morfológica de la palabra, de sus apariciones en el mismo texto, del contexto y del dominio en el que se encuadre el discurso.

Otro caso típico de etiquetado conceptual es el del reconocimiento de *entidades propias* o *named entities* (NE). En el Corpus Japonés de Diálogos para Análisis de Enfermería (Ozaku 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 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. A pesar de esto, su atractivo es grande debido a las ya mencionadas dificultades que plantea la estructuración sintáctica del habla espontánea. Uno de los escasos ejemplos finalizados es SESCO (Alcántara 2007), 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 50.000 palabras de habla espontánea. La anotación se basó en la estructuración composicional de tres únicos tipos eventivos que podían ser subdivididos según los argumentos que requirieran. Las estructuras resultantes fueron utilizadas como base para el análisis de otros niveles lingüísticos.

6. La anotación pragmática

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

Un ejemplo diferente, más conectado con los aspectos morfosintácticos, es el esquema propuesto por Marco de Rocha (1997) para el análisis de expresiones anafóricas en la lengua hablada. 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 de antecedente, el estatus de topicalidad del antecedente y el tipo de conocimiento necesario para procesarla.

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

La anotación pragmática es la que más varía según el objetivo final del etiquetado y, aunque algunas propuestas son generales (como los dos últimos ejemplos citados), la mayoría han sido diseñadas para sistemas muy específicos.

7. CONCLUSIONES

La nueva generación de corpus ofrece un gran potencial para el análisis lingüístico y el desarrollo de aplicaciones en un contexto en que la dependencia de los corpus y de los avances tecnológicos está resultando ser claramente bidireccional. Las características de estos corpus requieren de un esfuerzo en la anotación tanto si se parte de la reutilización de sistemas como si supone la creación de otros nuevos. Sin embargo y aunque la cantidad de corpus sea aún escasa, ya existen experiencias suficientes en el panorama internacional como para acometer la tarea con garantías.

La lingüística aplicada ha sido pionera en el uso de colecciones de textos para el tratamiento lingüístico y cuenta con metodologías afianzadas para el análisis de corpus. Esta experiencia previa está siendo básica para afrontar los nuevos retos, pero las exigencias actuales imponen cambios importantes. En este artículo se han mostrado algunos problemas representativos para cada nivel de análisis, pero es importante señalar en estas conclusiones que la solución a muchos de ellos deberá venir en el futuro también a través de la combinación de rasgos de diferentes niveles. Sin embargo, primero necesitamos corpus bien anotados para que eso pueda ocurrir, condición que, por suerte, cada vez tienen en mente más grupos dedicados a la lingüística aplicada.

AGRADECIMIENTOS

El presente artículo ha sido realizado con una beca postdoctoral del M.E.C. y como parte del trabajo del autor dentro de la red europea de excelencia K-Space (FP6-027026) y del proyecto RILARIM (TIN2004-07588-C03-02).

BIBLIOGRAFÍA

Alcántara, M. 2007. *Introducción al análisis de estructuras lingüísticas en corpus.* Madrid: UAM-Ediciones.

Alcántara, M. y T. Declerck. 2007. "Shallow Semantic Analysis of ASR Transcripts Associated with Video Shots". *Proceedings of the IWCS-7*, Tilburg.

ANNOTATE http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html

BNC British National Corpus, http://www-dev.natcorp.ox.ac.uk/

CES (Corpus Encoding Standard of EAGLES), http://www.cs.vassar.edu/CES/

CGN Corpus Gesproken Nederlands, http://lands.let.kun.nl/cgn/ehome.htm

Church, K., y R. Mercer. 1993. "Introduction to the Special Issue on Computational Linguistics Using Large Corpora". *Computational Linguistics* 19.

CORALROM Integrated reference corpora for spoken romance languages, http://lablita.dit.unifi.it/coralrom/

CORLEC Corpus de Referencia de la Lengua Española Contemporánea, http://www.lllf.uam.es/ESP/proyectos/corpus/corpus_oral.html

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

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

CSJ Corpus of Spontaneous Japanese, http://www2.kokken.go.jp/~csj/public/

Hoekstra H., M. Moortgat, B. Renmans, M. Schouppe, I. Schuurman y T. van der Wouden. 2003. *CGN Syntactische annotatie*.

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

MAP Task Corpus http://www.hcrc.ed.ac.uk/maptask/

MATE Multilevel Annotation Tools Engineering, http://mate.nis.sdu.dk/

McEnery, T. y A. Wilson. 1996. *Corpus Linguistics.* Edimburgo: University Press.

Nakatani, C. H. y D. R. Traum. 1999. *Coding discourse structure in dialogue (version 1.0). Technical Report UMIACS-TR-99-03*, University of Maryland.

de Rocha, M. 1997. "Corpus-Based Study of Anaphora in English and Portuguese". *Corpus-Based and Computational Approaches to Discourse Anaphora.* Eds. S.P. Botley y A .M. McEnery. UCL Press.

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

Ozaku, Hiromi itoh, A. Abe, N. Kuwahara, F. Naya, K. Kogure, y K. Sagara. 2005. "Building Dialogue Corpora for Nursing Activity Analysis", *Proceedings of LINC-2005*.
http://www.lllf.uam.es/ESP/proyectos/corpus/corpus_lee.html

UAM C-ORAL-ROM Corpus of Spoken Spanish, http://www.lllf.uam.es/

UCREL, http://www.comp.lancs.ac.uk/computing/research/ucrel/annotation.html

USAS UCREL Semantic Analysis System, http://www.comp.lancs.ac.uk/ucrel/usas/

xCES (Corpus Encoding Standard for XML), http://www.cs.vassar.edu/XCES/