La anotación de los corpus *CREA* y *CORDE*

Fernando Sánchez*, Jordi Porta, José Luis Sancho, Amalio Nieto,
Almudena Ballester, Adelaida Fernández, Javier Gómez,
Laura Gómez, Encarnación Raigal, Rafael Ruiz

fsanchez@rae.es
Departamento de Lingüística Computacional
Real Academia Española
*y Departamento de Lingüística
Universidad Autónoma de Madrid

Área temática: Lingüística de Corpus

Resumen

Este artículo describe las herramientas y recursos desarrollados en el Departamento de Lingüística Computacional de la Real Academia Española para la anotación lingüística de los corpus *CREA* y *CORDE*. Además de abundar sobre el enfoque clásico del procesamiento lingüístico a bajo nivel de textos de muy diversa índole y procedencia, el artículo aporta soluciones lingüísticamente motivadas para el análisis morfológico y la desambiguación, incidiendo, de paso, en la imposibilidad de esta última tarea con el único concurso de la información morfológica. Por este motivo, los autores se plantean como solución un adecuado compromiso entre cobertura y precisión guiado por los objetivos del proyecto.

1 Introducción

La Real Academia Española (RAE) está trabajando, desde 1995, en el diseño y construcción de dos corpus del español, uno de carácter histórico y otro del español actual. El primero de ellos, el Corpus Diacrónico del Español (*CORDE*), estructurado en tres grandes épocas (Edad Media, Siglos de Oro y Época Contemporánea), pretende ser una muestra representativa del español a lo largo de su historia. Por otra parte, el Corpus de Referencia del Español Actual (*CREA*) abarcará los últimos veinticinco años de la historia del español en este siglo, desde 1975 hasta 1999. Ambos proyectos terminarán su segunda fase a finales del año 2000.

El volumen de formas textuales contenido en cada uno de los corpus será de 125 millones de palabras[1]. Asimismo, se ha cuidado especialmente el diseño de ambos corpus, de forma que queden representados todos los territorios de habla hispana, tanto peninsulares como extrapeninsulares. Se han incluido en ambos corpus obras completas y, para su codificación, se utiliza *SGML*, siguiendo las normas propuestas por la *TEI* y por el *CES*[2].

El proceso de anotación de estos corpus sigue un modelo clásico secuencial en el que segmentación, análisis morfológico y desambiguación se llevan a cabo como procesos de transformación del texto original independientes. Sin embargo, existen herramientas que permiten la reinterpretación de algunos términos (por

[1] Actualmente, el *CREA* contiene 110 millones y 70 el *CORDE*.

[2] El lector interesado puede encontrar más información sobre el proyecto en las páginas Web de la RAE: http://www.rae.es.

resegmentación o reanálisis). Este ejercicio permite pensar en una integración futura del análisis lingüístico y la segmentación.

2 Segmentación

El objetivo de la segmentación es identificar y clasificar los elementos de análisis y estructurales que contiene un texto. En una primera fase del proyecto se utilizó el segmentador del proyecto *MULTEXT*, MtSeg, por su concepción (e implementación) modular. Sin embargo, por razones de eficiencia y adecuación, se ha desarrollado un segmentador en flex, que, sin renunciar a la modularidad lógica del segmentador original, permite integrar en un único componente software todos los módulos de MtSeg y algunos nuevos. Es más, el segmentador combina código y documentación, facilitando el mantenimiento y mejora del mismo[3].

El segmentador identifica y clasifica tanto los elementos estructurales de la codificación *SGML* como unidades atómicas objeto de análisis. Esta clasificación taxonómica se apoya en características que van desde lo estrictamente tipográfico a lo lingüístico, de forma que se facilite el proceso posterior de análisis lingüístico, y las clases se articulan como estructuras de rasgos. La taxonomía aparece como un compromiso entre varios criterios clasificatorios que tienen en cuenta adecuación, norma, uso, complejidad de reconocimiento, etc.

La segmentación, básicamente determinista, se apoya en gramáticas regulares, pero también en gramáticas de transducción que permiten una segmentación contextual. El inventario de elementos textuales identificados incluye palabras, elementos multipalabra, signos de puntuación, números, enumeraciones, abreviaturas, fechas, símbolos, referencias electrónicas, nombres propios, etc. Asimismo, se reconocen oraciones tipográficas.

[3]Para este fin, se emplea la herramienta de programación documentada noweb.

3 Anotación

3.1 Léxico

Se dispone de un lexicón, *LEX-CREA*, que incluye la nomenclatura completa del *DRAE*. Además, se ha codificado un conjunto de palabras cuya frecuencia es alta en los corpus y que no figuran en dicho diccionario, bien porque se encuentran en fase de estudio o aprobación, bien porque se ha rechazado su inclusión en el diccionario. Los criterios de categorización de *LEX-CREA* son *grosso modo* coincidentes con los del *DRAE*[4].

El léxico procedente del *DRAE* se ha obtenido automáticamente, por medio de un programa que interpreta la información presente en este diccionario y la transforma en rasgos formalizados de *LEX-CREA*. Este lexicón ha sido revisado manualmente, por medio de una aplicación gráfica desarrollada en Tcl/Tk que actúa como cliente de un servidor que se encarga de atender las peticiones de actualización de la base de datos, que utiliza como gestor la librería db de la Universidad de Berkeley.

3.2 Análisis morfológico

El análisis morfológico se realiza por medio de una consulta léxica a un lexicón de formas plenas. Este léxico se crea con ayuda del generador morfológico mmorph, desarrollado en el proyecto *MULTEXT* [PR95]. Este programa está basado en una tecnología mixta que combina morfología de dos niveles (para los aspectos morfografémicos) y gramáticas de unificación (para los morfosintácticos).

Con objeto de mejorar la eficiencia de la anotación, parte de este lexicón reside en memoria. El criterio para la selección de los elementos léxicos que se deben introducir en esta porción del lexicón ha sido exclusivamente su frecuencia en el *CREA*. Se han probado distintos ta-

[4]El esquema de anotación de los corpus es más fino en la clasificación categorial de ciertas clases cerradas que el *DRAE*, que los considera normalmente adjetivos y pronombres. Asimismo, el esquema 'esconde' la función —pronombre/determinante— de algunas categorías como rasgo secundario (cf. *infra*).

maños y se ha optado finalmente por guardar las 20.000 formas más frecuentes en memoria, pues esta cifra representa un compromiso razonable entre ocupación de memoria y cobertura del corpus[5]. Además, durante el análisis se van guardando en una memoria intermedia las formas que aparecen en el texto y se han debido buscar en disco, pues su probabilidad de reaparición en el mismo texto se supone más alta.

La implementación morfológica, originalmente inspirada en el correspondiente componente del proyecto EUROTRA [SL90], utiliza una estrategia no alomórfica en la que todos los procesos involucrados en la flexión (fonología, ortografía, alomorfía léxica y gramaticalmente condicionadas, e, incluso, supleción) se tratan a través de una adecuada interacción entre el componente de unificación y el de dos niveles [SL97]. El bloqueo entre tipos de irregularidad en la flexión verbal se implementa siguiendo los *órdenes* propuestos por Bello [Bel81], demostrando así la vigencia de las aportaciones de la gramática tradicional en el paradigma computacional.

Junto al tratamiento de la flexión, existen componentes especializados encargados de analizar algunos procesos derivativos (morfología apreciativa regular, adverbios en *-mente*), superlativos, formas verbales con enclíticos[6], además de un estimador categorial para la anotación sin léxico. Asimismo, existen componentes morfológicos especializados, como una gramática verbal que implementa las formas de voseo y, consecuentemente, su coaparición con formas enclíticas. Actualmente, se está desarrollando otro componente capaz de tratar la prefijación y la sufijación.

[5]Las 20.000 formas más frecuentes representan algo más del 90% de los elementos textuales del corpus; por tanto, solo una de cada diez palabras debe buscarse en disco o descomponerse.

[6]Debido a las características de los corpus, especialmente del *CORDE*, este componente analiza formas enclíticas sobre cualquier forma verbal, y no solo aquellas que las llevan en el uso moderno.

3.3 Esquema de anotación

En la definición del esquema de anotación se han tenido en cuenta diferentes aspectos que van desde la pertinencia lingüística de las distinciones propuestas, en el plano teórico, a la facilidad para aplicar dichas distinciones, en el pragmático. Así, frente a un esquema "ideal", propuesto en fases iniciales del proyecto, este se ha ido modificando, de forma que acomoda un conjunto de clases y rasgos morfoléxicos que se han demostrado adecuados sobre un **prototipo** del *CREA*[7].

En principio, buena parte de las decisiones adoptadas favorecen la simplicidad de aplicación de los criterios de desambiguación en casos de baja rentabilidad y alta complejidad de la desambiguación, pero obligando a realizar esta en otros casos en los que el contexto local lo permite[8]. Como consecuencia, la *aplicabilidad* de estas decisiones ha sido un aspecto fundamental para fijar las categorías léxicas del esquema de anotación. En contadas ocasiones, y siguiendo en parte la propuesta de EAGLES [VV.94], se ha preferido considerar ciertas categorías tradicionales, como los pronombres y determinantes, valores del rasgo función y habilitar la subclasificación en tipos de estas clases de palabras como valor de la categoría, con objeto de manejar la ambigüedad distribucional como ambigüedad *interna* a las descripciones categoriales.

Este esquema se ha aplicado sobre el prototipo, manualmente, en dos fases diferentes. La separación entre ambigüedades triviales/fortuitas (como ADJ|V no participio, N|V no participio o ART|PRON pers) y ambigüedades que presentan una mayor dificultad teórica (como ADJ|V participio, N común|N propio, N sing|V infinitivo), establecida de forma

[7]En realidad, un subcorpus del *CREA* de un millón de palabras que mantiene los criterios de distribución geográfica, temática y, en definitiva, de variedades lingüísticas del corpus completo.

[8]Esta afirmación, como el resto de lo expuesto en este epígrafe, debe entenderse respecto de un desambiguador o posteditor humano, y no respecto de las herramientas de desambiguación, para las que puede seguir siendo muy difícil, si no imposible, aplicar el esquema correctamente.

apriorística, ha permitido tomar datos del segundo tipo de ambigüedad para su resolución durante la segunda fase. En cualquier caso, el grueso de la desambiguación no ha planteado demasiados problemas, como lo prueba el hecho de que el 88% de las formas del prototipo quedaron totalmente desambiguadas tras aplicar la primera fase. La desambiguación se ha realizado en un editor controlado especializado que permite relacionar descripciones morfosintácticas con colores cuya semántica, arbitrariamente determinada, permite distinguir entre asignaciones correctas e irrelevantes en un contexto. Igualmente, permite controlar el conjunto de elementos textuales sobre los que debe aplicarse la desambiguación en una determinada fase. La herramienta de desambiguación se controla por medio del ratón y su uso es muy intuitivo.

Esta metodología incremental ha permitido la fijación de criterios de desambiguación (que se espera poder mimetizar en la desambiguación automática) al tiempo que se valoraba la conveniencia o no de ciertas distinciones. El grado de fijación se ha verificado realizando la desambiguación doble de los 32 textos que componen el prototipo de trabajo, lo que ha permitido comprobar el porcentaje de desambiguaciones no coincidentes entre parejas de posteditores sobre las clases de ambigüedad de la segunda fase, significativamente menor conforme se iban aplicando los criterios a nuevos contextos[9].

Por otra parte, el esquema de anotación acomoda perfectamente la no desambiguación total del corpus. En efecto, ciertas distinciones categoriales que sobrepasan el universo morfosintáctico para tocar el semántico, cuando no la esfera de lo subjetivo, han quedado reflejadas de forma ambigua en este prototipo. El porcentaje de ambigüedad residual es del 1,1%, porcentaje del que casi el 40% está representado por las distinciones ADJ|V participio y ADV rel|CONJ sub[10].

Finalmente, no se ha cuantificado la tasa de error residual contenida en este prototipo, si bien se sabe que el 0,09% de las unidades de análisis contienen alguna errata, fortuita o simple copia del original.

4 Desambiguación

Para la desambiguación de los corpus *CREA* y *CORDE* se ha rechazado una aproximación estadística. Los motivos principales para esta decisión radican en los siguientes aspectos:

- Esta aproximación parece haber alcanzado unas cotas no superadas en la última década [KVHA95]. Por otra parte, generalmente favorece la precisión en detrimento de la cobertura, aspecto este menos interesante para los objetivos marcados en el proyecto que aquí se presenta.
- Su comportamiento es razonablemente bueno cuando se dispone de un corpus homogéneo, pero se produce una caída de la precisión cuando el corpus, como en este caso, es heterogéneo y carece de una división apriorística satisfactoria.
- Las aproximaciones simbólicas favorecen una mejor relación entre el esquema de anotación y el proceso de desambiguación, pues es posible contrastar el concepto de 'distinción lingüísticamente relevante' con el de 'distinción formalizable' o 'abordable'.
- Finalmente, en contra de la creencia generalizada del profano, los tiempos de desarrollo de sistemas que utilicen una y otra aproximaciones no son tan dispares, como ya se ha demostrado para distintas lenguas [CT95, SV97, SL97].

Con objeto de probar los argumentos anteriores, el Departamento de Lingüística Computacional ha desarrollado su propio sistema de desambiguación reduccionista.

[9] Estos datos se encuentran en fase de análisis actualmente, por lo que se incorporarán en la versión final de este artículo.

El prototipo se podrá consultar a través de las páginas del proyecto en la RAE antes de que finalice este año.

[10] En el segundo caso, se trata de la palabra *cuando*, que se decidió no desambiguar.

4.1 Rtag

Rtag [Por96] es una herramienta para la desambiguación basada en principios lingüísticos que se enmarca dentro de una estrategia reduccionista. Permite, mediante expresiones regulares, representar la información léxica (referente a forma, lema y descripción morfosintáctica) y dar cuenta de relaciones lineales mediante operadores de contexto para contigüidad, no contigüidad y acotación, este último potencialmente útil para la aplicación de reglas que permitan el reconocimiento de lindes sintagmáticos. El pivote de la regla es una secuencia de elementos léxicos sobre los que se puntúan sus interpretaciones morfosintácticas, siendo la puntuación restrictiva cuando es negativa y promovedora cuando es positiva. Se permite una precondición o guarda en el pivote de la regla que permite atacar subclases de ambigüedad. Las reglas pueden aplicarse hasta saturación, si se desea, resolviendo las ambigüedades encontradas como 'cortes' en la puntuación de estas utilizando un umbral antes de iniciar un nuevo ciclo.

La herramienta, inspirada en *Constraint Grammars* [KVHA95], se asemeja más a la estrategia de *restricciones ponderadas* (*voting constraints*) propuesta por Oflazer y Tür [OT97], aunque se ha desarrollado independientemente de esta. Al formalismo, de bajo nivel, se le ha añadido un mecanismo de procesamiento de macros que permite aumentar la expresividad de las reglas. Asimismo, se han creado macrodefiniciones para poder definir un vocabulario para la expresión de descripciones morfosintácticas; con estas macros se pueden expresar conjuntos, subconjuntos y superconjuntos que permiten representar clases de palabras (clases de ambigüedad), así como ciertas expansiones de sintagmas. Finalmente, el formalismo es capaz de simular la unificación de rasgos mediante esquemas de reglas. Con estos recursos se pretende aumentar la 'legibilidad' de las gramáticas (y, por ende, mejorar su mantenimiento) y desvincularlas de un esquema/formato de anotación concreto.

4.2 Gramáticas

Una *gramática* de desambiguación no es exactamente una gramática, en el sentido tradicional ni computacional, en tanto que no contiene un conjunto de reglas positivas de producción. Las reglas, como instancias del mecanismo básico de restricción, suelen enunciarse más fácilmente con una lógica negativa en lugar de positiva [Roc92, Vou95]. Por otra parte, el hecho de que deban tratar tanto ambigüedades productivas como fortuitas y de que ciertas ambigüedades sean "resueltas" por aplicación de más de una regla o de que ciertos 'principios' deban expresarse por medio de varias reglas de restricción, debido a sus límites expresivos, ha servido, en no pocas ocasiones, para criticar sus condiciones de mantenimiento.

Por estos motivos, se ha hecho especial hincapié en la ordenación lógica de las reglas. Esta se ha realizado atendiendo a diferentes criterios, articulados en dos dimensiones: una vertical y otra horizontal. En el plano vertical, las reglas se organizan en una gradación según su *universalidad*, de forma que actualmente existen dos gramáticas que contienen reglas fiables, una, y reglas con un cierto componente heurístico, la otra. La segunda gramática enuncia principios universales que es difícil expresar con un formalismo lineal[11] o bien otros que siendo 'rentables' pueden introducir errores en ciertos contextos. La estrategia da cabida a otras gramáticas netamente heurísticas, que este proyecto no ha considerado al favorecer la cobertura sobre la precisión en aras de un mejor uso del corpus en las tareas lexicográficas de la Corporación.

Por su parte, el plano horizontal permite una ordenación de las reglas en capítulos que atacan la ambigüedad que afecta a clases homogéneas o la que aparece en contextos que permiten una desambiguación precisa. De este

[11] O con un universo informativo a todas luces insuficiente para realizar una desambiguación completa de un texto. El techo de un sistema de desambiguación morfosintáctico es consecuencia no sólo de su deficiente poder expresivo, sino especialmente de esta falta de información. Sin embargo, nadie parece haber comentado hasta ahora la imposibilidad de desambiguar un texto con tan poca información lingüística.

modo, el primer capítulo intenta la desambiguación de artículos y pronombres personales homógrafos mientras que el segundo capítulo trata de discriminar entre verbos y sustantivos homógrafos. Un tercer capítulo intenta explotar la información de concordancia para asociar modificadores a núcleos sintagmáticos. El cuarto capítulo ataca exclusivamente la ambigüedad de que, mientras que el quinto ataca la distinción funcional entre pronombres y determinantes. Hay, finalmente, tres capítulos que, tratando de rentabilizar toda la desambiguación previa, intentan tomar decisiones derivando de las relaciones lineales estructura para el sintagma norminal, el sintagma verbal y el sintagma adjetival. El proceso de prueba iterativa de estas gramáticas, junto con la introducción del mecanismo de saturación en rtag, ha permitido pasar de una disposición secuencial de estos capítulos, que se aplicaban en distintas configuraciones, comprobando así su grado de interdependencia, a una sola gramática que permite una aplicación indeterminista de las reglas. La pérdida de cobertura con esta gramática es pequeña, como se muestra a continuación, mientras que esta disposición mejora el mantenimiento de la gramática, al evitar la creación de reglas interdependientes o condicionadas. La gramática contiene unas 270 reglas.

El segundo conjunto de reglas tiene la misma estructuración, siendo su única diferencia con el primero la *relajación* de algunos contextos de aplicación de reglas. Es decir, su cariz heurístico es relativamente débil y, de momento, no se está utilizando en las pruebas de desambiguación al favorecer, como se ha dicho, la cobertura sobre la precisión. Esta gramática contiene unas 90 reglas. La tabla siguiente muestra algunos datos relativos a este prototipo y al comportamiento de la primera gramática sobre él.

5 *CORDE*

El corpus diacrónico incluye textos que abarcan desde el siglo XIII al XX e incluso colecciones de documentos en latín tardío de siglos anteriores (siglos X, XI y XII). Por tanto, *CORDE* representa todos los estadios de la evolución léxica del castellano desde sus orígenes. Estas características plantean problemas que van desde la segmentación a la desambiguación, pasando, como es lógico por la cobertura léxica y la fijación ortográfica.

No abordaremos en este artículo los problemas especiales de segmentación que plantea el *CORDE*, que tienen que ver no solo con las convenciones ortotipográficas empleadas en la época o por el responsable de la edición que se ha seleccionado para el corpus, sino también con el esquema de codificación empleado en este proyecto, que se muestra especialmente respetuoso con las intervenciones de autores, copistas y editores sobre el texto y que, en consecuencia, dificulta la identificación de las unidades de análisis lingüístico. Asimismo, poco puede decirse actualmente sobre la validez de las gramáticas de desambiguación diseñadas para el *CREA* para los textos del *CORDE*, aunque las primeras pruebas realizadas sobre textos del XIX y del XVIII permiten contemplar la posibilidad de utilizar las mismas gramáticas sobre textos de algunas épocas representadas en el *CORDE*.

En relación con el léxico, el *CORDE* plantea, básicamente, tres tipos de problemas: a) piezas léxicas que no han tenido continuidad a lo largo de la evolución de la lengua, b) piezas léxicas que han sufrido cambios en dicha evolución, y c) problemas relacionados con la falta de fijación ortográfica. En la enunciación de estos tres problemas, siempre se tiene en cuenta una visión del léxico actual. Por este motivo, se ha decidido aproximar el léxico del *CORDE* lo más posible al del *CREA*. En este sentido, para resolver el problema planteado en a), existe un lexicón especial para este corpus, que contiene lemas y formas del español medieval y de los siglos de Oro. Este lexicón cuenta, de momento, con todas las categorías que pertenecen a clases cerradas y algunos paradigmas verbales irregulares completos. El número de entradas es ligeramente inferior a 1.000[12].

Por su parte, b) y c), que son los más intere-

[12]En un futuro, contendrá una morfología flexiva completa, especialmente para el verbo.

Tabla 1: Prototipo y primera gramática

Ambigüedad inicial media:	1,77
Ambigüedad residual media tras gramática 1:	1,22
Cobertura media inicial:	$\approx$ 100%
Cobertura media tras gramática 1:	99,04%
Precisión media tras gramática 1:	81,58%

santes desde el punto de vista computacional, se han tratado por medio de los llamados *filtros de modernización grafémica*. Este recurso es, básicamente, una función de conversión de grafemas en otros según unas reglas expresadas declarativamente, que aprovechan la regularidad de algunas equivalencias entre distintos estados de la lengua. En su enunciación actual, es posible definir un cambio grafémico y contextualizar la región de la palabra original en la que este está permitido: solo al principio, solo al final, solo en posición media o en cualquier posición. Asimismo, es posible establecer restricciones sobre el contexto grafémico[13]. Tras la aplicación de todas las combinaciones de filtros se vuelve a consultar el lexicón, que devolverá, de forma no determinista, todos los análisis posibles. Este mecanismo, ideado para el problema b), funciona con éxito en algunos casos de c), pues los filtros, heterogéneos en su implementación, incluyen tanto casos de evolución como de falta de fijación. Los verdaderos casos de falta de fijación de la ortografía, especialmente cuando resultan en palabras 'posibles', y su relación con la desambiguación del texto, no tienen un tratamiento adecuado en este nivel de procesamiento.

Como primera aproximación, se ha trabajado con un conjunto de formas del corpus que representan el 91-92% del mismo. La estrategia comentada en este epígrafe eleva la cobertura (es decir, el número de formas que entre sus análisis reciben el análisis correcto) al 96-97%, sin un aumento significativo de la verbosidad inicial en relación con el *CREA*. Los errores se deben, principalmente, a excepciones en los filtros o problemas de homografía fortuitos, que se tratarán introduciendo tales elementos léxicos en el lexicón específico.

[13]Las reglas tienen el aspecto siguiente: "sustitúyase nb por mb en todo el contexto" o "sustitúyase ç por c ante e o i".

6 Conclusiones

En este artículo, se han descrito brevemente las herramientas y recursos desarrollados para la anotación morfosintáctica de los corpus *CREA* y *CORDE*, actualmente en fase de construcción en la Real Academia Española. El esfuerzo realizado permite pronosticar unos buenos resultados en esta tarea, si bien pone de manifiesto no pocas limitaciones de las aproximaciones tradicionales al Procesamiento del Lenguaje Natural, incluso en las (mal llamadas) *aplicaciones de bajo nivel*. El análisis lingüístico de estos corpus permitirá no solo extraer conclusiones sobre el uso de la lengua (en distintos periodos históricos) sino también una valiosa experiencia que permita ofrecer técnicas, herramientas y recursos más sólidos para el tratamiento automático del español.

Referencias

[Bel81] A. Bello. *Gramática de la Lengua Castellana Destinada al Uso de los Americanos*. Instituto Universitario de Lingüística Andrés Bello, Tenerife, 1981. 4ª, 1ª edición 1847.

[CT95] J.-P. Chanod and P. Tapanainen. Tagging French – comparing statistical and constraint-based methods. In *Proceedings of the EACL-95*, University College, Belfield, Dublin, Ireland, 1995.

[KVHA95] F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Anttila, editors. *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin, 1995.

[OT97] K. Oflazer and G. Tür. Morphological Disambiguation by Voting Constraints. In *Proceedings of the ACL/EACL'97*, Madrid, 1997.

[Por96] J. Porta. Rtag. Technical report, Grupo de Investigación en Lingüística Computacional, Universidad de Barcelona, 1996.

[PR95] D. Petitpierre and G. Russell. MMORPH - The MULTEXT Morphology Program. MULTEXT deliverable report for the task 2.3.1, ISSCO, University of Geneva, February 1995.

[Roc92] E. Roche. Text disambiguation by finite state automata, an algorithm and experiments on corpora. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING-92*, pages 993–997, Nantes, 1992.

[SL90] F. Sánchez León. Spanish Morphology Implementation Report. Documento interno de Eurotra, Comisión de las Comunidades Europeas, Abril 1990.

[SL97] F. Sánchez León. *Análisis Morfosintáctico y Desambiguación en Castellano.* PhD thesis, Facultad de Filosofía y Letras, Septiembre 1997.

[SV97] C. Samuelsson and A. Voutilainen. Comparing a Linguistic and a Stochastic Tagger. In *Proceedings of the ACL/EACL'97*, Madrid, 1997.

[Vou95] A. Voutilainen. Morphological disambiguation. In F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Anttila, editors, *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*, pages 165–284. Mouton de Gruyter, Berlin, 1995.

[VV.94] VV.AA. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. a common proposal and applications to european languages. Technical report, EAGLES document, October 1994.