EUSLEM: Un lematizador/etiquetador de textos en euskara

Aldezabal I., Alegria I., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K. (*)
Aduriz I., Urkia M. (**)

(*) Informatika Fakultatea. 649 P.K. 20080 DONOSTIA (Euskal Herria).
acpalloi@si.ehu.es
(**) UZEI. Aldapeta, 20. 20009 DONOSTIA (Euskal Herria)

Resumen

Un lematizador/etiquetador automático es una herramienta básica para aplicaciones tales como indexación automática, bases de datos documentales, análisis sintáctico y semántico, análisis de corpus, etc.

Las características con las que se ha diseñado el lematizador/etiquetador EUSLEM han sido las siguientes: extenso, robusto, de propósito general, modular y eficiente.

Los módulos básicos en los que se divide son: preprocesador para la detección y etiquetado de números, signos de puntuación, etc.; analizador morfológico general basado en la morfología de dos niveles con tratamiento de formas dialectales, errores debidos al desconocimiento del idioma estándar (tratamiento de variantes lingüísticas) y gestión de léxicos específicos; lematización sin léxico para que el sistema sea robusto; tratamiento de unidades léxicas complejas (términos de más de una palabra) y desambiguación basada en estadísticas. En una fase posterior se complementará con la desambiguación basada en conocimiento lingüístico.

A la hora de abordar este proyecto un inconveniente que hemos tenido que superar es la falta de un sistema de etiquetas para el euskara. Esto combinado con que el euskara es una lengua aglutinante que puede acumular gran cantidad de información en una palabra (varios casos, elipsis, etc.) hace que la definición del sistema de etiquetas no sea evidente.

Temas: Análisis de corpus, análisis morfológico.

1. Introducción

Un lematizador/etiquetador(*tagger*) automático es una herramienta básica para aplicaciones tales como indexación automática, bases de datos documentales, análisis sintáctico y semántico, análisis de corpus, etc. Existe gran cantidad de bibliografía sobre *taggers* y entre la que hemos utilizado podemos destacar los siguientes proyectos: CLAWS [Garside et al., 87] para el inglés y SWETWOL [Karlsson, 92] para el sueco.

El antecedente de este proyecto es el proyecto EEBS (Recopilación Sistemática del Euskara Actual) patrocinado por Euskaltzaindia (Academia de la Lengua Vasca) y realizado por UZEI [Urkia, Sagarna, 91], que consistía en la recopilación y lematización semiautomática de un corpus de tres millones de palabras del euskara escrito del siglo XX. Utilizando ese corpus y el analizador morfológico automático para el euskara [Agirre et al., 89] hemos diseñado y estamos desarrollando un lematizador/etiquetador automático con las siguientes características: extenso, robusto, de propósito general, modular y eficiente.*

A la hora de abordar este proyecto un inconveniente que hemos tenido que superar es la falta de un sistema de etiquetas para el euskara. Esto combinado con que el euskara es una lengua aglutinante que puede acumular gran cantidad de información en una palabra (concatenación de casos, elipsis, etc.) hace que la definición de dicho sistema de etiquetas no sea evidente y a ello le dedicaremos el apartado 2.

Posteriormente se irán detallando los módulos básicos que componen EUSLEM: analizador morfológico general basado en la morfología de dos niveles, con tratamiento de formas dialectales y errores debidos al desconocimiento del idioma unificado (tratamiento de variantes lingüísticas), y gestión de léxicos específicos; lematización sin léxico para que el sistema sea robusto; tratamiento de unidades léxicas complejas y desambiguación basada en estadísticas, además del preprocesador para la detección y etiquetado de números, signos de puntuación, etc.

2. Las etiquetas

A la hora de definir el sistema de etiquetas para el euskara nos hemos encontrado con dos grandes dificultades:

- No había definido previamente ningún sistema exhaustivo de etiquetas para su uso automático, ya que en el mencionado proyecto EEBS el corpus se lematizó pero no se

* Este proyecto cuenta con una subvención parcial de la Universidad del País Vasco y del Departamento de Industria del Gobierno Vasco.

etiquetó sistematicamente, y los diccionarios impresos adolecen de una falta de sistematización en cuanto a las categorías utilizadas, en el caso de que las utilicen.

- La salida del analizador morfológico automático no ofrece un sistema de etiquetas viable, ya que al ser el euskara una lengua aglutinante, declinada y de morfología muy compleja (concatenación de casos, elipsis, derivación, etc.) [Euskaltzaindia, 85] la información morfológica (morfosintáctica en la práctica) cruda necesita ser tratada previamente. Así *etxekoarenak* (los de la de la casa (sic)) es una palabra no demasiado extraña en euskara, compuesta por el lema *etxe* (casa) seguido de tres sufijos de declinación: genitivo de lugar (*-ko*), genitivo (*-aren*) y nominativo (*-ak*), cada uno de los cuales tiene su determinación y su número, suponiendo los dos genitivos sendas elipsis. *Helarazteak* (el hacer llegar) es la composición del infinitivo *hel* (llegar) y el infinitivo *araz* (factitivo), seguida de varios sufijos: de nominalización (*te*), y *ak* que es ambiguo ya que puede ser nominativo plural o ergativo singular (lo más probable en este caso).

La elección del sistema de etiquetas es una tarea fundamental a la hora de diseñar un etiquetador ya que de él depende, además de la utilidad del producto, el grado de ambigüedad. Las etiquetas se utilizan tanto como resultado del proceso como para establecer las tablas que permitirán la desambiguación basada en estadísticas. En [Garside et al., 87] apéndice B hay una interesante descripción de diversos sistemas de etiquetas.

La mayoría de los sistemas manejan un conjunto de etiquetas de alrededor de 100 elementos. Las diferencias principales vienen del tratamiento o no de etiquetas compuestas (elipsis, lexías, etc.) aunque hay sistemas que utilizan solamente las categorías elementales (alrededor de 10). El nivel de ambigüedad, además de depender del sistema de etiquetas, también depende del idioma. Por ejemplo, el nivel de ambigüedad que presentan CLAWS (para el inglés) distinguiendo alrededor de 100 etiquetas y SWETWOL (para el sueco) con 12 etiquetas es similar (alrededor del 40%).

El sistema de etiquetas que hemos elegido para el euskara es un sistema de tres niveles que el usuario puede parametrizar cuando utiliza el programa. El primero corresponde a la categoría (nombre, adjetivo, verbo, etc.), el segundo a la subcategoría y el último a cierta información morfológica que se considera interesante. En el proyecto London-Lub Corpus [Svartvik & Eeg-Olofsson, 82] se hace algo similar para un corpus del inglés hablado con un total de 197 etiquetas.

También tratamos etiquetas compuestas ya que es vital tanto para las derivaciones como para las expresiones compuestas o idiomáticas, siglas, etc. El sistema de etiquetas todavía es abierto, pero se han definido un total de 17 categorías en el primer nivel (dos derivadas), con una media de 3 subcategorías por nivel. El nivel de ambigüedad del euskara en el primer nivel de etiquetas es de aproximadamente el 30%. La desambiguación semiautomática para el

proceso de aprendizaje se realiza en el tercer nivel, aunque en función de los resultados no descartamos que las tablas se establezcan en el segundo o primer nivel.

3. El analizador morfológico general

Aunque hay etiquetadores que funcionan sin léxico y, por lo tanto, sin análisis morfológico, lo habitual, sobre todo si ya se dispone de uno, es hacerlo con un léxico y, cuando no es posible el análisis de una palabra, para que el sistema sea robusto, combinarlo con un etiquetado adicional sin utilizar el léxico.

En [Agirre et al., 89] se describe la parte básica de nuestro analizador morfológico consistente en la aplicación de la morfología de dos niveles [Koskenniemi, 83] al euskara. Nuestro sistema se compone de un conjunto de 24 reglas morfofonológicas que describen los cambios que se producen entre los niveles léxicos y de superficie y un sistema léxico compuesto de unos 60.000 elementos agrupados en alrededor de 120 subléxicos mantenido en una base de datos léxica general (EDBL - Base de Datos Léxica para el euskara, implementada en ORACLE) y posteriormente estructurada en árbol. Las secuencias de morfemas que pueden encadenarse se expresan por medio de las denominadas clases de continuación que definen el conjunto de morfemas que pueden seguir a un morfema dado. Debido a que la clase de continuación no puede condicionar morfemas posteriores al inmediatamente siguiente, este mecanismo no soluciona el problema de las dependencias a distancia. Como solución hemos generalizado el concepto de clase de continuación, extendiéndolo para poder expresar conjuntos de subléxicos que condicionan morfemas subsiguientes [Agirre et al., 92].

El sistema también dispone de un módulo de gestión de léxico particular con lo que se puede enriquecer con léxicos dependientes de la aplicación.

El formalismo de dos niveles en su implementación original no es demasiado eficiente. De todas formas nuevas implementaciones, principalmente la de Xerox (herramientas *lexc* y *twolc*) [Karttunen, 93], permiten un análisis muy eficiente ya que evitan el backtracking. Estamos estudiando la posibilidad de utilizar estas herramientas en nuestro sistema, aunque los problemas de las dependencias a distancia y de la gran cantidad de información morfológica a gestionar son un obstáculo.

El analizador obtiene la descomposición en morfemas de cada forma, pero, al ser el euskara una lengua morfológicamente compleja, esta información debe ser tratada para resolver los problemas de elipsis, aglutinación de casos, lemas compuestos, etc. Este tratamiento lo realizaremos por medio de un sistema de reglas que describe cómo se deducen los rasgos correspondientes a la forma analizada a partir de los de sus morfemas. Un tratamiento similar se hace en el proyecto ANLT [Ritchie et al., 87].

4. El tratamiento de variantes lingüísticas

Debido a la reciente normalización del euskara (en 1968 se establecen las primeras normas) y a que la misma no está finalizada, además del uso de abundantes variantes dialectales, en el mencionado proyecto EEBS estaba prevista una doble lematización compuesta por el lema normalizado y el lema de la variante lingüística (lema no normalizado).

EUSLEM recoge esta idea, y para implementarla usa el mismo mecanismo usado por nuestro corrector ortográfico Xuxen para el tratamiento de errores típicos, consistente en el uso de un conjunto adicional de reglas y subléxicos que añadido al analizador de dos niveles básico es capaz de reconocer formas no normalizadas y obtener su correspondiente lema normalizado [Aduriz et al., 93].

Esto hace que la ambigüedad aumente, ya que hay formas que pueden tener interpretación como normalizadas pero también como variante lingüística. Así, *hasi* se puede interpretar como forma normalizada del infinitivo *has* (empezar), pero también como variante lingüística del infinitivo *haz* (sembrar), ya que en zonas de Euskal Herria no se distingue la s y la z. Aunque en principio si hay interpretación normalizada se rechaza la interpretación como variante, no se descarta la opción de obtener todos los posibles análisis para una utilización de EUSLEM con postedición.

5. La lematización sin léxico

Si en los análisis anteriores no se ha producido ningún análisis válido, para que el sistema sea robusto hace falta una lematización sin léxico que permita que no queden elementos sin lematizar ni etiquetar.

Los distintos sistemas estudiados dan en general soluciones diferentes a este problema. Lo más común es la implementación de un sistema *ad hoc* de tratamiento de sufijos (a veces también de prefijos), aunque algunos sistemas asignan a estas formas todas o algunas de las etiquetas posibles y dejan en manos del desambiguador la tarea de la asignación. Esta última solución es posible para un etiquetador pero no para un lematizador, ya que obtiene alguna etiqueta pero nunca un lema.

Nuestra solución para EUSLEM se basa en una idea ya utilizada para síntesis de voz [Black et al., 91] consistente en utilizar la morfología de dos niveles para reconocer palabras cuyo lema no está en el léxico. En esta solución se utiliza un sistema léxico alternativo compuesto solamente por los morfemas que no son lemas (prefijos, afijos y sufijos); además de unos lemas genéricos *$$* para cada una de las categorías/subcategorías posibles. Dos reglas adicionales expresan la correspondencia entre *$$* a nivel léxico y cualquier cadena de

caracteres plausible en el idioma. Una vez localizados los morfemas constituyentes, salvo el lema que es sustituido por *$$*, un procedimiento ad-hoc permite reconocer el lema correspondiente a *$$* (siempre existe comprobación por medio de la síntesis ya que el mecanismo de dos niveles es bidireccional).

Esta solución tiene una doble ventaja: homogeneidad con el análisis morfológico general, y obtención siempre de alguna solución. Este último punto es a su vez un inconveniente ya que normalmente se obtienen muchas posibilidades que hay que reducir. En el artículo reseñado se mencionan como criterio de reducción la longitud del lema (cuanto más corto mejor), un orden de aplicación de las reglas y la frecuencia de los afijos. De todas formas, en aquel sistema había necesidad de obtener una sola interpretación, mientras que EUSLEM lo que hace es una preselección que no debe obtener un único candidato y dejará en manos de la desambiguación la selección definitiva.

6. El tratamiento de unidades léxicas complejas

La decisión de cuándo un elemento debe ser lematizado/etiquetado como complejo es controvertida. Mientras en algunos casos, como en las expresiones idiomáticas, es bastante claro su tratamiento como unidad separada, en otros es discutible. Nuestra decisión sobre qué considerar una unidad léxica compleja se basa en la experiencia de UZEI en el proyecto EEBS, y no opta por una vía restrictiva sino que hay un amplio espectro de casos en que términos de más de una palabra se lematizan y etiquetan como una unidad.

A la hora de clasificar las que consideramos unidades léxicas complejas del euskara, funcionalmente hemos establecido las siguientes características:

- seguros/ambiguos: los seguros son aquellos que eliminan los análisis de los componentes como elementos independientes.
- declinables/fijos: hay términos que no se declinan y otros que se declinan. En estos habrá que marcar el elemento declinable y el tratamiento debe ser precedido del análisis morfológico de los elementos.
- contiguos/dispersos: en algunos términos los elementos no aparecen seguidos por lo que su proceso se complica ya que habrá que buscar en palabras posteriores. En este caso, si hay más de dos elementos, puede ser que algunos elementos deban ser contiguos y otros no.
- ordenados/desordenados: vayan contiguos o dispersos puede ser posible que no mantengan un orden. Un caso de esto son los verbos perifrásticos (*korrika egin* -correr-, *negar egin* -llorar-, etc.) en oraciones negativas.

A la hora del tratamiento las opciones van desde el no tratamiento, a tratamientos complejos como son la extensión de las clases de continuación entre palabras, o los modelos conectivistas [van der Linden & Kraaij, 90].

En EUSLEM hemos optado por un tratamiento *ad hoc* de los términos compuestos, manteniendo su información en la mencionada base de datos léxica EDBL, pero sin codificarlos como el resto del léxico, ya que no se procesarán por el mecanismo de dos niveles. Una vez lematizadas las palabras se consultará la base de datos para saber si la forma o el lema pueden ser parte de un término compuesto, y si es así se obtendrán sus características y se verificará la presencia del resto de elementos, eliminándose el resto de análisis si se trata de un término no ambiguo.

7. La desambiguación

La desambiguación es una tarea fundamental para conseguir un lematizador/etiquetador automático o semiautomático con una buena selección y ordenación de las propuestas. En los últimos años un gran número de proyectos se han centrado en esta tarea, dividiéndose los mismos en dos grandes grupos:

- *taggers* con desambiguación por medio de métodos estadísticos con o sin aprendizaje [Garside et al., 87] [De Rose, 88] [Cutting et al., 92].
- *taggers* con desambiguación basada en conocimiento lingüístico [Brill, 92] [Karlsson et al., 92].

Aunque algunos de los desambiguadores por medio de reglas han conseguido buenos resultados es comúnmente aceptado que los basados en métodos estadísticos ofrecen muy buenas prestaciones para desambiguación de etiquetas. El método propuesto en CLAWS [Garside et al., 87] basado en probabilidades de lemas combinadas con probabilidades condicionadas al elemento anterior (bigramas) se ha convertido en clásico. Existe una implementación para castellano [Moreno, 94]. La eficiencia del algoritmo es mejorada por De Rose [De Rose, 88] y este último algoritmo es el que se usa en el proyecto SUC que desambigua la salida del analizador SWETWOL.

Hemos pensado que EUSLEM utilice el algoritmo de De Rose. Actualmente estamos procediendo a desambiguar manualmente un corpus, para poder establecer las primeras tablas en las que se base la desambiguación automática.

Además de la desambiguación basada en estadísticas, nuestro grupo está trabajando en la definición de un conjunto de reglas para una gramática de restricciones, Constraint Grammar [Karlsson et al., 1992], cuyo objetivo final es el análisis sintáctico pero que en una primera

fase aborda el problema de la ambigüedad morfológica. En una fase posterior procederemos a la integración de ambos modos de desambiguación.

8. El preprocesador

Todo tratamiento de textos escritos necesita de un preproceso para unificar la entrada y diferenciar los elementos básicos de proceso: párrafo, frase, palabra, etc. En el caso de EUSLEM, el preprocesador se encarga de la búsqueda y etiquetado de signos de puntuación y separadores, guarismos, siglas, etc.

El etiquetado de signos de puntuación y separadores, además de ser interesante para el etiquetado de corpus, es fundamental para la desambiguación estadística ya que son elementos que no producen etiquetas ambiguas y por lo tanto reducen las cadenas de elementos ambiguos.

En el caso de los guarismos, siglas, etc., al ser susceptibles de declinación, no es suficiente el preproceso sino que deberán procesarse morfológicamente con lo que se obtendrá una etiqueta compuesta.

9. Los resultados

EUSLEM está diseñado como un programa general no ligado a ninguna utilización concreta. Recibiendo un fichero de entrada obtiene el correspondiente fichero etiquetado que posteriormente podrá ser utilizado para indexación automática, establecimiento de bases de textos etiquetadas, análisis de corpus, lexicografía, concordancias, etc.

Para ello EUSLEM es parametrizable en función de los siguientes conceptos:

- nivel de etiquetado: primer, segundo o tercer nivel del sistema de etiquetas, o información morfológica completa
- etiquetado automático o semiautomático: es decir, presencia o no de ambigüedad en la salida

Debido a la posibilidad de ambigüedad en la interpretación de formas como correspondientes a términos simples o partes de términos multipalabra, y a que estos últimos no son siempre contiguos, el etiquetado y lematización de los términos multipalabra se realizará en cada uno de sus componentes, especificándose, además del lema completo y su categoría compuesta, el número correlativo de componente y el total de ellos.

Vamos a ver como ejemplo el análisis de la frase "Eta gauza aundirik ekartzerik ez zuen izan" (*Y no pudo traer gran cosa*).

El analizador morfológico devuelve el siguiente resultado:

```
((forma "*eta")
  ((anal 1)
   ((lema "etA")((KAT JNT))))
)
((forma "gauza")
  ((anal 1)
   ((lema "gauza")((KAT ADI))))
  ((anal 2)
   ((lema "gauzA")((KAT IZE))))
  ((anal 3)
   ((lema "gauzA")((KAT IZE)))
   ((morf "a")((KAT DEK)(KAS NOM)(NUM S)(MUG M))))
)
((forma "aundirik")
)
((forma "ekartzerik")
  ((anal 1)
   ((lema "ekaR")((KAT ADI)))
   ((morf "tzerik")((KAT ERL)(ERL KONP))))
  ((anal 2)
   ((lema "ekaR")((KAT ADI)))
   ((lema "tze")((KAT ASP)(KER IZE)))
   ((morf "Rik")((KAT DEK)(KAS PAR)(MUG MG))))
)
((forma "ez")
  ((anal 1)
   ((lema "ez")((KAT ADB))))
  ((anal 2)
   ((lema "ez")((KAT IZE))))
)
((forma "zuen")
  ((anal 1)
   ((lema "zuen")((KAT ADL)(MDN B1)(NOR 3)(NRK 3)(ERR *edun))))
  ((anal 2)
   ((lema "zu")((KAT IOR)))
   ((morf "eM")((KAT DEK)(KAS GEN)(NUM P)(MUG M))))
  ((anal 3)
   ((lema "zuen")((KAT ADL)(MDN B1)(NOR 3)(NRK 3)(ERR *edun)))
   ((morf "En")((KAT ERL)(ERL ERLT))))
  ((anal 4)
   ((lema "zuen")((KAT ADL)(MDN B1)(NOR 3)(NRK 3)(ERR *edun)))
   ((morf "En")((KAT ERL)(ERL ZHG))))
)
((forma "izan")
  ((anal 1)
   ((lema "izaN")((KAT ADI))))
  ((anal 2)
   ((lema "izaN")((KAT ADI)))
   ((morf "0")((KAT ASP)(ADM PART))))
)
```

Esta descomposición morfológica es tratada componiendo la información morfosintáctica y obteniendo las etiquetas en el nivel requerido (suponemos que al primer nivel), consiguiéndose el siguiente resultado:

```
("*eta" (etA/JNT))
("gauza" (gauza/ADI)(gauzA/IZE))
("aundirik" )
("ekartzerik" (ekaR/ADI)(ekaR+tze/IZE))
("ez" (ez/ADB)(ez/IZE))
("zuen" (*edun/ADL)(zu/IOR))
("izan" (izaN/ADI))
```

Como se ve la palabra "aundirik" no tiene análisis ya que es una variante lingüística de "handirik". Para estos problemas está el tratamiento de dichas variantes. Con dicho tratamiento obtenemos el siguiente resultado, donde el símbolo "$" indica variante lingüística:

```
("*eta" (etA/JNT))
("gauza" (gauza/ADI)(gauzA/IZE))
("aundirik" ($handi/ADJ))
("ekartzerik" (ekaR/ADI)(ekaR+tze/IZE))
("ez" (ez/ADB)(ez/IZE)($hes/ADI))
("zuen" (*edun/ADL)(zu/IOR)($su/IZE))
("izan" (izaN/ADI))
```

Tras la desambigüación el resultado que se obtendrá será el siguiente:

```
*eta/eta:JNT
gauza/gauza:IZE
aundirik/$handi:ADJ
ekartzerik/ekar:ADI
ez/ez:ADB
zuen/*edun:ADL
izan/izan:ADI
```

Conclusiones

Se han expuesto las características generales del lematizador/etiquetador automático para textos en euskara EUSLEM, además de la descripción de cada uno de los módulos que lo componen. El proyecto está en fase de ejecución teniendo ya realizados los módulos de preproceso, análisis morfológico general, tratamiento de variantes lingüísticas y lematización sin léxico. El módulo de tratamiento de unidades léxicas complejas está programándose, a la vez que se están introduciendo los datos en la base de datos léxica. Respecto al módulo de desambiguación basada en estadística, estamos en la fase inicial de desambiguación manual de un pequeño corpus, cuyos resultados serán utilizados para el aprendizaje del sistema.

Bibliografía

Aduriz I., Agirre E., Alegria I., Arregi X., Arriola J.M., Artola X., Diaz de Illarraza A., Ezeiza N., Maritxalar M., Sarasola K., Urkia M. *A Morphological Analysis Based Method for Spelling Correction*. Proceedings of the 6th Conference of the EACL 1993, p.463.

Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K., Urkia M. *Aplicación de la morfología de dos niveles al euskara*. S.E.P.L.N, vol. 8, 87-102. 1989.

Agirre E, Alegria I, Arregi X, Artola X, Diaz de Illarraza A, Maritxalar M, Sarasola K, Urkia M. *XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology*. Proceedings of the Third Conference ANLP (ACL), 119-125, 1992.

Black A., van de Plassche,J., Williams B. *Analysis of Unkown Words through Morphological Descomposition*. Proceedings of 5th Conference of the EACL 1991, vol. 1 pp.101-106.

Brill E. *A simple rule-based part of speech tagger.* Proceedings of the Third Conference ANLP (ACL), 152-155, 1992.

Cutting D., Kupiec J., Pedersen J., Sibun P. *A practical part-of-speech tagger*. Proceedings of the Third Conference ANLP (ACL), 133-140, 1992.

De Rose S. *Grammatical category disambiguation by statistical optimization*. Computational Linguistics, 14, 31-39. 1988.

Euskaltzaindia. *Euskal Gramatika: Lehen urratsak (I eta II)*. Euskaltzaindia, Bilbo 1985.

Euskaltzaindia. *Euskal Gramatika: Lehen urratsak (IIIeta IV)*. Euskaltzaindia, Bilbo 1990.

Garside R., Leech G., Sampson G. *The Computational Analysis of English: A corpus-based approach*. Longman. London, 1987.

Karlsson F. *SWETWOL: A comprehensive morphological analyser for Swedish*. Nordic Journal of Linguistics, 15, 1-45. 1992.

Karlsson F., Voutilainen A., Heikkila J,. Anttila A. *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*.

Karttunen L. *Finite-State Lexicon Compiler*. Xerox ISTL-NLTT-1993-04-02. 1993.

Koskenniemi K. *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production,* University of Helsinki, Department of General Linguistics. Publications no. 11, 1983.

Moreno-Torres I. *A Morphological Disambiguation Tool (MDT): Aplication to Spanish*. Universidad de Málaga, Febrero 1994.

Ritchie G.D., Pullman S.G., Black A.W., Russell G.J. *A computational framework for lexical description*. Computational Linguistics, 13, 3-4, 290-307, 1987.

Svartvik J., Eeg-Olofsson M. *Tagging the London-Lund Corpus of Spoken English*. In Johanson (1982) pp. 85-109.

Urkia M, Sagarna A. *Terminología y Lexicografía asistida por ordenador. La experiencia de UZEI*. SEPLN, vol 8, 1991.

Van der Linden E., Kraaij W. *Ambiguity resolution and the retrieval of idioms: two approaches*. COLING-90, vol 2, pp. 245-249. 1990.