DEL ANALIZADOR MORFOLÓGICO AL ETIQUETADOR/LEMATIZADOR: UNIDADES LÉXICAS COMPLEJAS Y DESAMBIGUACIÓN

Aldezabal I., Alegria I., Ezeiza N.,Urizar R. (*)
Aduriz I. (**)

(*) Informatika Fakultatea. 649 P.K. 20080 DONOSTIA (Euskal Herria).
jibecran@si.ehu.es (Nerea Ezeiza)
(**) UZEI. Aldapeta, 20. 20009 DONOSTIA (Euskal Herria)
uzei0005@sarenet.es (Itziar Aduriz)

Abstract

In this paper, we present two of the modules of the lemmatiser/tagger for Basque EUSLEM: the former handles the treatment of MultiWord Terms (MWT) and the latter makes the disambiguation of the source text. For this disambiguation process, we are working separately on two strategies —a method based on linguistic knowledge and a statistical approach— that we will try to combine in the future. We propose a formal description for MWTs in Basque and describe an implementation for their processing. Finally we expound the initial results of the statistical disambiguation process.

Keywords: Analysis of corpora, tagging, statistical disambiguation, multiword terms, basque language.

1. Introducción

Para aplicaciones tales como indexación automática, bases de datos documentales, análisis semántico, análisis de corpus etc., es básica una herramienta que ofrezca automáticamente la etiqueta de las unidades léxicas textuales. Sin embargo, en las lenguas aglutinantes y morfológicamente complejas, como es el caso del euskara, tan importante como la etiqueta es el lema de dicha unidad. Para el inglés y el castellano ya se han diseñado herramientas de etiquetado, pero la relativa simplicidad de su morfología, les ha permitido en muchos casos obviar la información sobre el lema de la palabra.

EUSLEM es un proyecto que surgió con el fin de lograr la lematización y etiquetado automático para textos en euskara. Su función inicial era la de servir de ayuda a los lexicógrafos y, en concreto, para agilizar el proceso de lematización que se lleva a cabo en UZEI de cara a la segunda fase del proyecto EEBS [Urkia et al., 91]. Actualmente, además de dicha función inicial, hemos ido detectando gran cantidad de utilidades que lo hacen especialmente interesante como herramienta básica en futuras aplicaciones.

Consta de cinco módulos básicos: un procesador para la detección y etiquetado de números, signos de puntuación etc.; un analizador morfológico basado en la morfología de dos niveles [Koskenniemi, 83] con tratamiento de variantes dialectales y errores lingüísticos —debidos al desconocimiento del idioma estándar—; lematización sin léxico [Alegria et al., 96]; tratamiento de unidades léxicas complejas; y por último, desambiguación basada, por un lado, en estadística y, por otro, en conocimiento lingüístico. En la fase de desambiguación estadística se han utilizado las herramientas desarrolladas en el contexto del proyecto Multext [Amstrong et al., 95], mientras que para la desambiguación basada en conocimiento lingüístico se está utilizando Constraint Grammar [Karlsson et al., 95]. Aunque por ahora estos métodos de desambiguación se están probando por separado, en una fase posterior se procederá a la integración de ambos.

En el presente artículo nos centraremos concretamente en el trabajo realizado para tratar las unidades complejas y explicaremos el proceso de desambiguación estadística. Para ello, nos basaremos en el análisis morfológico resultante de los primeros tres módulos mencionados anteriormente. Dicho análisis ofrece por cada token todos los posibles lemas con su información morfológica. Partiremos de este resultado para obtener, finalmente, un único lema y análisis morfológico por cada token del texto.

2. El tratamiento de Unidades Léxicas Complejas (ULCs)

No resulta fácil dar una definición exacta de *palabra*. A nivel de texto, podría definirse como "la secuencia de caracteres entre dos espacios" [Fontenelle et al., 94]. Muchas de las

unidades léxicas responden a esta definición (*etxe* 'casa', *zuri* 'blanco/a, *txakur* 'perro/a'), incluso un gran número de expresiones que en lenguas no flexivas son ULCs, en euskara (lengua con un nivel de flexión muy alto) constituyen una unidad tipográfica (*aurrerantzean* 'de aquí en adelante', *aldiz* 'no obstante'). Pero, obviamente, se precisa de otra definición que dé cuenta del resto de expresiones idiomáticas.

Nuestro criterio sobre qué considerar una unidad léxica compleja se basa en la experiencia lexicográfica de UZEI en el proyecto EEBS, donde un amplio espectro de términos de más de una palabra se lematizan como una unidad. De hecho, hemos utilizado esas unidades complejas ya lematizadas, como base de nuestro análisis. A continuación veremos lo que en nuestro trabajo se ha considerado ULC.

2.1. Palabras compuestas

En euskara, las palabras compuestas pueden aparecer escritas básicamente de cuatro formas [Euskaltzaindia, 92]:

- constituyendo una unidad tipográfica (*idazmakina* 'máquina de escribir').
- unidas por un guión (*datu-base* 'base de datos', *begi-nini* 'pupila del ojo').
- separadas por un espacio, cuando el primer término del compuesto ha sufrido alguna transformación fonológica (*Euskal Herria* 'País Vasco', *itsas portu* 'puerto marítimo', donde *euskal* e *itsas* son las variantes en composición y derivación de *euskara* 'lengua vasca' e *itsaso* ´mar´respectivamente).
- separadas por un espacio, sin que ninguno de los componentes haya sufrido transformación alguna (*bake ituna* 'tratado de paz').

Tanto los compuestos lexicalizados que constituyen una unidad tipográfica como los que se escriben con guión son tratados como términos simples, es decir, tienen su entrada en la base de datos. El analizador morfológico es también capaz de detectar los compuestos de libre generación separados por guión (*mahai-hanka* 'pata de mesa'), ya que el guión de composición es tratado como un elemento léxico.

La detección de los compuestos cuyo primer elemento ha sufrido alguna transformación tampoco supone gran dificultad ya que las palabras susceptibles de transformación constituyen un grupo bastante reducido y son fácilmente localizables.

Para detectar los compuestos lexicalizados que se escriben separados sin guión y en los que el primer término no ha sufrido ninguna transformación, la única manera es tratarlos en la base de datos como ULCs.

2.2. Unidades Léxicas Complejas

En la distinción entre colocaciones léxicas y expresiones idiomáticas se ha utilizado tradicionalmente un criterio semántico [Heid, 94]. Resulta sumamente difícil interpretar una

expresión idiomática partiendo del significado de sus componentes (*adarra jo* 'tomar el pelo' ≠ *adarra* 'cuerno' + *jo* 'tocar'), mientras que en las colocaciones léxicas, sus componentes (o alguno de ellos) se utilizan en su sentido literal (*zarata atera* 'meter ruido', donde *zarata* significa 'ruido' y *atera* 'sacar').

Sin embargo, es muy difícil marcar una linea divisoria clara entre colocaciones léxicas y expresiones idiomáticas, ya que entre las colocaciones léxicas libres y las expresiones idiomáticas más opacas, existe una gran gama de combinaciones de palabras dentro de una progresión o continuum [Cowie, 90]:

- *Expresiones idiomáticas puras* (opacas). En este grupo podemos encontrar tanto las expresiones idiomáticas léxicas (*ahuntzaren gauerdiko eztula* 'nadería'), como las que denominamos "gramaticales" (*harik eta* 'hasta que', *hala eta guztiz ere* 'no obstante').
- *Expresiones idiomáticas figurativas*. Estas combinaciones son idiomáticas en el sentido de que difícilmente admiten variaciones, pero para el hablante, la referencia literal primitiva de estas expresiones no se encuentra tan lejana como en las puras (*hutsaren hurrengoa* 'insignificante').
- *Colocaciones restringidas*. También llamadas semi-idiomáticas, ya que uno de sus componentes es utilizado en su sentido literal y el otro (u otros) se utiliza(n) en un sentido figurativo, que no se hallará más que en este contexto (*eskerrak eman* 'agradecer').
- *Colocaciones abiertas*. Cada elemento de la combinación es utilizado en su sentido literal (*hego haizea* 'viento sur').

Como hemos mencionado anteriormente, nos hemos basado en el criterio de UZEI a la hora de limitar las ULCs. De este modo, y con un criterio abierto, se han tomado en cuenta, tanto las expresiones idiomáticas, opacas y figurativas, como las colocaciones restringidas. En cuanto a las colocaciones abiertas, sólo se han considerado aquellas que expresan un concepto concreto (*Euskal Herria* 'País Vasco', *Amerikako Estatu Batuak* 'Estados Unidos de América').

2.3. Descripción de las Unidades Léxicas Complejas

A la hora de describir las unidades léxicas complejas del euskara para su procesamiento automático, hemos establecido las diferentes características funcionales a considerar, agrupándolas de la siguiente manera: 1) las que hacen referencia al término en su totalidad, 2) las que describen las relaciones de co-ocurrencia de sus componentes y 3) las relacionadas con cada componente de la unidad léxica compleja. A continuación describiremos las características de cada uno de los grupos mencionados:

- *Características de la unidad en su totalidad*: la característica más importante es la seguridad, aunque también incluiremos en este grupo la información morfológica de la ULC.

- Seguridad: diremos que son ULCs seguras aquellas que eliminan los análisis de los componentes como elementos independientes. De lo contrario, diremos que son ambiguas.
- Información morfológica: en el momento que se detecte una ULC, asignaremos a sus componentes el análisis correspondiente a dicha unidad. Por lo general, será suficiente con la categoría y la subcategoría de la unidad.

• *Características de co-ocurrencia de los componentes*: este grupo engloba las características de continuidad y orden.
- Continuidad: en algunas unidades los componentes no aparecen contiguos, por lo que su proceso se complica ya que habrá que buscar en un contexto más amplio que las palabras contiguas al posible componente de la ULC (***ezin*** *ikusi* ***izan*** *dugu* 'no lo hemos podido ver'). Si hay más de dos elementos, puede ser que algunos de ellos sean contiguos y otros no.
- Orden: vayan contiguos o dispersos puede ser que los componentes no mantengan un orden. Un ejemplo es el de las perífrasis verbales (*korrika egin* 'correr', *negar egin* 'llorar') en oraciones negativas. Así decimos ***negar egin*** *dut*, 'he llorado', pero *ez dut* ***egin negarrik***, 'no he llorado'.

• *Características de los componentes*: la característica que describe cada componente de una ULC es su posibilidad de flexión.
- flexión: hay componentes que no flexionan y otros que admiten distintas flexiones. A los términos que admiten un reducido número de formas flexionadas es preciso aplicar restricciones. Para representar dichas restricciones utilizamos un formalismo lógico simple.

2.4. Tratamiento de las Unidades Léxicas Complejas

Respecto al tratamiento de las ULCs, y aunque en muchos proyectos el tratamiento de estos términos es postergado o no explicado, existen algunas referencias interesantes entre las que se pueden distinguir las siguientes:

- Extensión de las clases de continuación entre palabras. La descripción se hace de la misma forma que la morfotáctica.
- Reglas compilables en autómatas (una por cada ULC) que se pueden componer con los autómatas morfosintácticos [Segond et al., 95].
- Modelos conectivistas [van der Linden et al., 90].

Este es un tema que está abierto y donde ninguna de las propuestas ha conseguido el éxito necesario para ser propuestas como método general ya que suelen ser soluciones demasiado particulares o de una gran complejidad. Por ejemplo, muchas de las citadas anteriormente no son capaces de tratar términos cuyos componentes pueden aparecer en desorden.

En EUSLEM hemos optado por un tratamiento propio pero muy flexible y generalizable de los términos compuestos, manteniendo su información en la mencionada base de datos léxica EDBL. Sin embargo, dichos términos no son codificados posteriormente como el resto del léxico, ya que no se procesarán morfológicamente por el mecanismo de dos niveles. El lingüista actualiza de forma semiautomática la base datos que será la fuente de información para el tratamiento de las ULCs que aparecen en los corpus. Así se distingue claramente la información lingüística del tratamiento.

El tratamiento ha sido implementado en *C++*, utilizando *flex* y *yacc* para el tratamiento de la entrada. El esquema general del proceso es el que se describe en la figura 1. Se puede observar como el texto de entrada es analizado léxica y sintácticamente —módulos *flex* y *yacc* respectivamente— para comprobar la corrección de su formato y al mismo tiempo ir obteniendo la información necesaria para decidir si existen ULCs en dicho texto. Lo mismo ocurre con la información relativa a las ULCs a detectar.

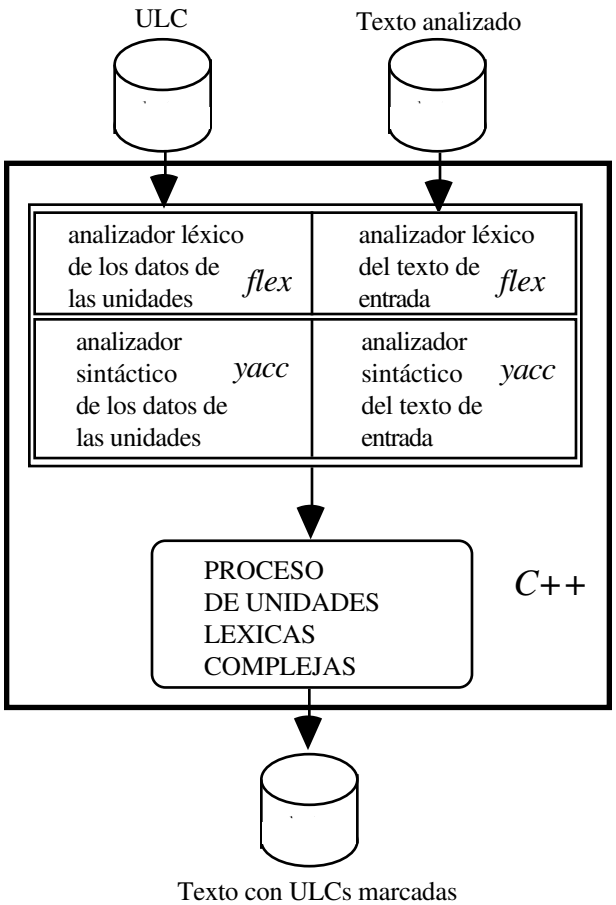

Fig. 1.- Esquema del tratamiento de las ULCs.

Hasta el momento hemos probado la utilidad y corrección de las herramientas con un subconjunto representativo de ULCs —unas 500— y hemos procedido a la depuración de los problemas encontrados. El principal obstáculo es el relativo a la característica de seguridad, ya que, aunque en un principio su definición pueda parecer clara, al confrontarlo con ejemplos del

corpus aparecen inconsistencias. Otro problema a reseñar es limitar la unidad de contexto donde las ULCs no contiguas pueden aparecer, obteniéndose resultados más precisos con contextos más restringidos que con el de oración completa.

3. Desambiguación

A diferencia de otras herramientas EUSLEM debe obtener, además de la etiqueta, el lema correspondiente a cada forma. Para ello lo que hemos hecho es dividir el proceso de desambiguación en dos fases: una primera, donde para cada categoría se elige el lema más probable y una segunda, clásica en la bibliografía, donde en función del contexto se elige la etiqueta.

La primera fase, la de desambiguación de lemas, la hemos desarrollado sin tener en cuenta el contexto y, basándonos en estadísticas que relacionan terminaciones de lemas con las etiquetas, obtenemos un solo lema por cada análisis posible. En el módulo de lematización sin léxico, mencionado en la introducción, esta desambiguación adquiere mayor importancia si cabe, ya que sin dicho proceso la ambigüedad entre lemas posibles se dispararía.

La desambiguación de etiquetas es una tarea fundamental para conseguir un 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, 92].

Aunque algunos de los desambiguadores basados en 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 [Cutting et al., 92] se ha convertido en clásico y sobre el mismo se han hecho distintas implementaciones —por ejemplo [Sánchez León, 95]— y variaciones. Una de estas variaciones es el desambiguador de *Multext* [Amstrong et al., 95] que hemos utilizado; seguidamente citamos algunas de las características que lo hacen muy interesante para nosotros:

- es de libre distribución
- es modular y está preparado para ser integrado detrás de un módulo morfológico
- no necesita una lista de todas las formas posibles y sus correspondientes etiquetas
- es muy flexible y ofrece herramientas para la recopilación de estadísticas.

Como paso previo a la desambiguación basada en estadísticas se realizó un proceso de desambiguación manual. Un corpus de alrededor de 25.000 palabras, elegido al azar del corpus EEBS, fue manualmente desambiguado desde el punto de vista morfológico.

La desambiguación manual de textos sirve a varios propósitos:

- la depuración del sistema de etiquetas y la mejora de la representación gramatical en los análisis
- como texto para la evaluación de los resultados obtenidos con el tagger automático basado en el conocimiento lingüístico
- y principalmente como base para el aprendizaje en la desambiguación tanto de lemas como de etiquetas.

Aunque en este momento nos hallamos en fase de pruebas y mejora de los resultados, podemos hacer una primera evaluación de éstos. Utilizando en la fase de entrenamiento un corpus de alrededor de 22.000 palabras y con un texto que presentaba las tasas de ambigüedad que se indican en la tabla 1, los resultados de la desambiguación estadística de etiquetas suponen un error del 7.25%.

TIPO DE PALABRA	**NUM.**	**TOTAL ANALISIS**	**AMBIGÜEDAD**
FORMAS ESTANDAR	1163	3087	2.65434
VARIANTES LINGÜISTICAS	16	34	2.125
SIN LEMA EN EL LEXICO	51	293	5.7451
TOTAL	1230	3414	2.77561

Tabla 1.- Características y ambigüedad del texto etiquetado

Hay que resaltar que estos son unos primeros resultados en los que se debe tener en cuenta que:

- Hemos trabajado con etiquetas a nivel de categoría (20). Con mayor número de etiquetas la ambigüedad de la entrada aumenta pero la precisión de la información que se maneja en la desambiguación es mayor.
- No hemos realizado ningún tipo de corrección por medio de *biases*.
- Es un texto abierto donde, como se ve en la tabla 1, hay formas a las que no se pueden asignar análisis morfológicos directamente, sino por medio de un tratamiento de prefijos y sufijos, lo que añade mucha ambigüedad e imprecisión a la entrada.

Por lo dicho anteriormente esperamos ofrecer a corto plazo unos resultados más exhaustivos y mejorados.

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 (CG) [Karlsson et al., 95], cuyo objetivo final es el análisis sintáctico pero que en una primera fase aborda el problema de la ambigüedad morfológica. En este momento el trabajo está centrado en la elaboración de las reglas de desambiguación morfológica [Aduriz et al., 96].

En una fase posterior procederemos a la integración de ambos métodos de desambiguación.

4. Conclusiones

En este artículo hemos descrito dos importantes módulos del lematizador/etiquetador para el euskara, el módulo de tratamiento de ULCs y el de desambiguación morfológica.

Hemos realizado una descripción de lo que consideramos ULCs en euskara y una propuesta para su tratamientos automático. En dicha propuesta se resuelven problemas, como la aparición de los componentes de una ULC en desorden, que en otras aproximaciones no se tenían en cuenta. También hemos descrito los métodos de desambiguación que estamos utilizando para seleccionar tanto los lemas como las etiquetas.

Una cuestión abierta y que esperamos experimentar a corto plazo es la relación entre la ambigüedad y el tratamiento de ULCs ya que no está demasiado claro el orden de tratamiento: primero desambiguar y posteriormente tratar las ULCs o a la inversa.

Bibliografía

Aduriz I, Aldezabal I, Alegria I, Artola X, Diaz de Illarraza A, Ezeiza N, Gojenola K., Urkia M. *EUSLEM: Un lematizador/etiquetador de textos en euskara*. Actas de la SEPLN, 1994.

Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M. *Different issues in the design of a lemmatizer/tagger for Basque*. "From text to tag" Workshop, SIGDAT, EACL. 1995.

Aduriz I, Alegria I., Arriola J. M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. *A Corpus Based Morphological Disambiguation Tool for Basque*. SEPLN 1996.

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.

Alegria I., Artola X., Ezeiza N., Gojenola K., Sarasola K. *A trade-off between robustness and overgeneration in morphology*. Proc. of Natural Language Processing + Industrial Applications. Moncton, Canada. 1996.

Amstrong, S, Bouillon P., Robert, G. *Tools for Part-of-Speech Tagging*, Tech. Report ISSCO, Geneva. 1995.

Amstrong, S, Russel, G., Petitpierre, D., Robert, G. *An open Architecture for Multilingual Text Processing*. Proceedings of the EACL SIGDAT workshop, Dublin. pp. 30—34. 1995.

Brill, E., *A simple rule-basedpart of speech tagger*. Proceeding of Third Conference ANLP (ACL), 133–140, 1992.

Cowie, A.P.; Mackin R.; McCaig I.R. *Oxford Dictionary of Current Idiomatic English.* v2. 1990.

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, *Hitz elkartuen osaera eta idazkera*. Hitz-elkarketa /4. LEF batzordea. 1992.

Farwell, D.; Helmreich, S.; Casper, M. *SPOST: a Spanish Part-of-Speech Tagger*, actas del XI Congreso *SEPLN*, nº17, 42–53, Deustua. 1995.

Fontenelle, T.; Adriaens, G.; De Braekeleer, G. *The Lexical Unit in the Metal® MT System, MT*.The Netherlands. 1-19. v9. 1994.

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

Heid, U. *On Ways Words Work Together - Topics In Lexical Combinatorics*, The way words work together / combinatorics, *Euralex'94*. Amsterdam. 226-257. 1994.

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*.. 1995.

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.

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

Sánchez Leon, F .*Desarrollo de un etiquetador morfosintáctico para el español*, actas del XI CongresoSEPLN, nº17, 14–28, Deustua. 1995

Segond F., Tapanainen P. *Using a finite-state based formalism to identify and generate multiword expressions*. Technical Report MLTT-019. Ran Xerox Research Centre. 1995.

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