APLICACION DE LA MORFOLOGIA DE DOS NIVELES AL EUSKARA

E. Agirre, I. Alegria, X. Arregi, X. Artola, A. Diaz de Ilarraza, K. Sarasola
Informatika Fakultatea - E.H.U.- U.P.V. - DONOSTIA
M. Urkia
U.Z.E.I. - Unibertsitate-Zerbitzuetarako Euskal Ikastetxea - DONOSTIA.

AREA : Análisis Morfológico

RESUMEN

Entre los diferentes formalismos propuestos para el análisis morfológico automático la morfología de dos niveles ha suscitado notable interés en los últimos años. Su principal característica es la clara diferenciación entre el nivel superficial (o textual) y el nivel léxico (o del diccionario). Los cambios entre ambos niveles son descritos por medio de reglas que se traducen para su implementación a autómatas de estados finitos. El programa es totalmente independiente de la lengua para la que se utiliza, siendo válido tanto en análisis como en generación.

En este artículo se describe este formalismo y su aplicación al euskara (para el que resulta muy adecuado debido a su alto grado de flexión) dentro de un proyecto de elaboración de herramientas lingüísticas. En este marco y como subproducto del analizador morfológico general se trabaja en la construcción de un corrector ortográfico para el euskara (*).

(*)Este trabajo está enmarcado dentro de un proyecto subvencionado por la Diputación Foral de Gipuzkoa.

APLICACION DE LA MORFOLOGIA DE DOS NIVELES AL EUSKARA

E. Agirre, I. Alegria, X. Arregi, X. Artola, A. Diaz de Ilarraza, K. Sarasola
Informatika Fakultatea - E.H.U.- U.P.V. - DONOSTIA
M. Urkia
U.Z.E.I. - Unibertsitate-Zerbitzuetarako Euskal Ikastetxea - DONOSTIA.

1. INTRODUCCION.

La simplicidad de las flexiones en inglés ha hecho que el interés por la investigación en análisis morfológico por ordenador haya sido bastante reducido. En inglés lo más usual es utilizar un léxico con todas las formas flexionadas o un mínimo conjunto de reglas morfológicas [Winograd 83] . Así, mientras se han desarrollado gran número de herramientas independientes del idioma para el análisis sintáctico y semántico no pasa lo mismo con las herramientas morfológicas.

De todas formas, para idiomas distintos al inglés se han ido desarrollando sistemas de análisis morfológico por ordenador como ATEF de GETA [GETA 82], sistemas para el finlandés [Broda et al. 80] y otras.

En el año 1981 Kaplan y Kay [Kaplan et al. 81] hacen una interesante aportación diseñando un formalismo para generación fonológica por medio de reglas que se compilan en un autómata. Esta idea será continuada por Koskenniemi en el formalismo de dos niveles que vamos a examinar.

En los ultimos años distintos formalismos para análisis morfológico se han desarrollado rapidamente. Concretamente, para el castellano, se han desarrollado MARS [Meya 87] que usa un autómata de estados finitos para la descomposición

morfológica y AM [Martí 87] que utiliza un autómata markoviano ampliado con condiciones.

A continuación se describen el formalismo de la morfología de dos niveles y su aplicación al euskara.

2. MORFOLOGIA DE DOS NIVELES.

En 1983 Koskenniemi [Koskenniemi 83] definió el modelo computacional de morfología de dos niveles. Este modelo ha tenido una gran aceptación en años posteriores y se distingue por las siguientes características :

- Es un modelo general aplicable a cualquier lengua aunque su primera aplicación fue para el finlandés.
- Es válido tanto para el análisis como para la síntesis.
- Separa claramente el conocimiento lingüístico y el algoritmo, como consecuencia la implementación para cualquier lengua es sencilla ya que el programa es el mismo.
- Separa claramente el nivel superficial de la palabra a analizar o generar y el nivel léxico o profundo que es el que se representa en el sistema de diccionario (sistema léxico). Asi se evita el almacenar distintas formas del mismo morfema debido a cambios morfofonológicos.
- Utiliza un sistema de reglas paralelas en lugar de los sistemas de reglas de reescritura utilizados en fonología generativa, con lo que el sistema es conceptual y computacionalmente más sencillo.

Los elementos básicos de la morfología de dos niveles son dos: las reglas y el sistema léxico.

2.1. LAS REGLAS**.**

El modelo de dos niveles maneja dos representaciones: la léxica y la superficial. El léxico contiene representaciones morfofonológicas de raices y afijos.

La diferencia esencial con la fonología generativa es que no hay estados intermedios entre las dos representaciones. Las reglas no ejecutan nada, sólo establecen correspondencias entre los niveles. Así el reconocimiento de palabras se reduce a encontrar una representación léxica válida correspondiente a una forma de superficie. Inversamente la generación parte de la representación léxica conocida y busca representaciones de superficie que se correspondan con ella.

Las reglas constan de las siguientes tres partes :

- La correspondencia, consiste en un par de caracteres (el primero de nivel léxico y el segundo su correspondiente a nivel superficial). Estos caracteres pueden ser concretos o abstractos; éstos últimos permiten la generalización de reglas. Ejemplos de caracteres abstractos: "C" cualquier carácter consonante, "V" cualquier carácter vocal.
- El contexto, restringe los casos en que se verifica la correspondencia. En función de los caracteres anteriores y posteriores.
- El operador indica el tipo de relación entre el contexto y el par expresado en la correspondencia. Puede ser de restricción de contexto ($\Rightarrow$), de coerción de superficie ($\Leftarrow$) o ambos a la vez ($\Leftrightarrow$).

La sintaxis primitiva de estas reglas sufrió pequeñas modificaciones [Koskenniemi 85] para poder implementar un compilador de estas reglas y no tenerlas que convertir a autómatas manualmente.

Por ejemplo $l{:}i \Rightarrow b{:}b \_ e{:}e$ indica que el par $l{:}i$ (esto es, la aparición del carácter "l" a nivel léxico y el carácter "i" a nivel de superficie) implica que el

contexto a la izquierda sea el par b:b y a la derecha el par e:e . La regla l:i ⇒ b:b _ e:e indica por su lado que dondequiera que aparezca el par b:b a la izquierda, el par e:e a la derecha y el carácter "l" a nivel léxico entonces el carácter superficial tiene que ser forzosamente "i ".

l:i	⇒	**b:b**	**_**	**e:e**
correspondencia	operador	contexto a izquierda		contexto a derecha

El uso de corchetes permite definir contextos alternativos para una misma correspondencia. Por ejemplo la regla

2:e ⇐ [C:C / 8:Ø / 6:r] _

es equivalente a las tres reglas siguientes:

2:e ⇐ C:C _

2:e ⇐ 8:Ø _

2:e ⇐ 6:r _

El uso de las reglas exige la definición de los siguientes elementos :

- El alfabeto, o conjunto de caracteres de superficie.
- Los subconjuntos del alfabeto de superficie que se usan en las reglas.

  Por ejemplo, C denota una consonante cualquiera y V cualquier vocal.
- El alfabeto del léxico. Consiste en el alfabeto de superficie más caracteres especiales llamadas marcas de selección ($, §, 4, 5,...)que controlan la aplicación de las reglas.
- Los subconjuntos del alfabeto del léxico. Son similares a los de superficie.
- Definiciones de abreviaturas o subexpresiones de las reglas.
- Las reglas propiamente dichas.

2.2. EL SISTEMA LEXICO.

Mientras las reglas sirven para describir las diferencias entre los dos niveles, el sistema léxico define el conjunto de morfemas, clasificándolos según los posibles encadenamientos entre ellos. Consiste en un conjunto de subléxicos y en las clases de continuación que regulan las secuencias posibles de raíces y afijos.

Los subléxicos sirven para agrupar elementos léxicos de las mismas características (sufijos, prefijos, lemas nominales, lemas verbales,...). Todos los subléxicos tienen la misma estructura, un nombre que los identifica y un conjunto de entradas. Cada entrada contiene tres campos :

- La representación léxica, es una secuencia de caracteres del alfabeto léxico. Estos caracteres pueden ser caracteres de superficie o marcas de selección. A estos últimos se les podrá hacer equivaler otros caracteres de superficie por medio de las reglas.
- La clase de continuación a la que pertenece. Es un nombre que agrupa a varios subléxicos y/o otras clases de continuación, indicando que los morfemas incluídos en ellos son los únicos que pueden seguir a la entrada que se define.
- La información morfológica correspondiente a la entrada.

Por tanto las clases de continuación son la base del mecanismo para definir cuáles son las combinaciones posibles de los distintos morfemas en una palabra. Su poder expresivo es mínimo por lo que en algunos casos es necesario duplicar descripciones que no serían necesarias, como en el caso de la dependencia a larga distancia entre morfemas.

2.3. EL PROGRAMA.

El programa utiliza dos módulos auxiliares principales Fsp y Lex. El primero hace la labor de autómata y va aceptando los pares de caracteres. Para ello en la inicialización construye un único autómata de estados finitos, a partir de los autómatas correspondientes a cada regla, lo alinea y obtiene el conjunto de pares posibles válidos.

El módulo léxico (Lex) realiza la función de acceso al léxico. Para ello está creado un fichero dividido en subléxicos, cada uno de los cuales esta organizado en forma de árbol en el que cada arco es un carácter con lo que se consigue acceso incremental. La unión entre subléxicos se realiza por medio de las clases de continuación.

Estos modulos son independientes del idioma y se referencian desde un programa principal que puede realizar análisis o síntesis.

Además de para el finlandés existen implementaciones del modelo de dos niveles para el inglés [Karttunen et al. 83], francés [Lun 83], japonés [Alam 83] y rumano [Khan 83].

La complejidad del modelo es estudiada a fondo en [Barton 85] y salvo algunas excepciones da la razón a Kartunnen [Karttunen 83] en que la complejidad de la lengua no tiene efectos significativos en la velocidad del análisis o la síntesis.

3. APLICACION AL EUSKARA

3.1. BREVE DESCRIPCION DE LA MORFOLOGIA DEL EUSKARA.

El euskara es una lengua aglutinante, es decir, para la formación de las palabras la entrada de diccionario toma de forma independiente cada uno de los elementos necesarios para las diversas funciones (el caso incluído) . Concretamente

los afijos correspondientes al determinante, número y caso de declinación, se toman en este orden e independientemente.

Una de las principales características del euskara es que funciona por medio de un sistema de declinación, con casos, lo cual la aleja de las lenguas que la rodean. Las flexiones de determinante, número y caso se hacen a nivel de todo el sintagma nominal, no a nivel de cada elemento en particular como ocurre en las lenguas románicas que hoy nos rodean. La declinación vasca es única, es decir, existe una sola tabla de declinación que se añade a todas las entradas.

El sistema verbal del euskara es rico, así como la facilidad para la composición y la derivación, lo cual posibilita la creación léxica partiendo de una única entrada.

3.2. ¿POR QUÉ LA MORFOLOGIA DE DOS NIVELES ?

Algo similar al euskara ocurre en finlandés, que también es una lengua aglutinante. Ambas lenguas pueden recibir un tratamiento similar, si bien el euskara es más simple en cuanto que cada unidad léxica tiene una sola raíz, lo cual no siempre ocurre en finlandés.

La declinación funciona de modo análogo en las dos lenguas, así como la composición y derivación.

Esta similitud es la razón que nos ha llevado a optar por la morfología de dos niveles propuesta por Koskenniemi.

3.3. APLICACION DE LA MORFOLOGIA DE DOS NIVELES.

Se ha aplicado al euskara el formalismo descrito anteriormente elaborando reglas y clasificando todos los morfemas según subléxicos y clases de continuación.

3.3.1. Las reglas.

Las correspondencias existentes entre el nivel léxico y el superficial debidas a transformaciones morfofonológicas se expresan por medio de las reglas. En el caso del euskara también las hemos aplicado con sus tres componentes.

Las reglas se aplican para expresar tres tipos de transformaciones: añadir, suprimir o cambiar un carácter del nivel léxico. Estas transformaciones básicas podrán combinarse. A continuación se presentan ejemplos de reglas que llevan a cabo los tres tipos de transformaciones básicas:

a) Añadir : Se muestran aquí la inclusión de vocales y consonantes epentéticas (e,r,...).

Regla nº 2. Añadir una "e" epentética.

2:e ⇔ [C:C / 8:Ø / 6: r] _

Entre otros fenómenos indica que cuando a nivel léxico nos encontramos con una consonante (C) seguida de la marca de selección 2 (que distingue a los morfemas que exigen esta "e" epentética), el 2 de nivel léxico se corresponde con una "e" a nivel de superficie. Ejemplo: sakoneko ("de lo profundo")

Nivel Léxico :	sakon **2**ko
Nivel Superficial :	sakon **e**ko

Regla nº 1. Añadir una "r" epentética.

1:r ⇔ V:V _

Indica que cuando a nivel léxico una marca de selección 1 sigue a una vocal ("1" distingue a los sufijos que exigen esta "r" epentética), entonces al "1" de nivel léxico le corresponde una "r" a nivel de superficie. Además, teniendo en cuenta el otro sentido del operador, siempre que aparezca el par **1:r** el carácter anterior será una vocal en

ambos niveles. Ejemplo: semeri ("a (algún, ningún....) hijo")

Nivel Léxico : seme **1**i

Nivel Superficial: seme **r**i

b) Suprimir : Desaparece un carácter al unirse un elemento a otro:

pérdida de la última consonante de la forma del infinitivo del verbo al tomar la marca del aspecto o la nominalización.

Regla nº 15. Pérdida de "n".

n:Ø ⇔ _ [4$te:ØØte / 5k:Øk / 45k:ØØk / $la:Øla / $n:Øn]

Entre otros fenómenos indica que cuando nos encontramos a nivel léxico un carácter "n" seguido de "4$te", a nivel de superficie a la "n" le corresponde un carácter vacio (Ø) y a la cadena "4$te" le corresponde la cadena "ØØte" (la marca de selección 4 corresponde a infinitvos que pierden su último carácter y la marca $ a los sufijos verbales que producen esa pérdida). Ejemplo: egiten ("haciendo")

NivelLéxico: egi**n**4 $ten

Nivel Superficial: egi**ØØØ**ten

Regla nº 12. Desaparición de la "r" del destinativo ante nasal, en demostrativos.

r:Ø ⇔ n!2:nØØ _

hona ("aquí" (destin.))

Nivel Léxico: hon! 2**r**a

Nivel Superficial: honØØ**Ø**a

c) Cambiar : Cambio de consonante sorda a sonora por influencia de una nasal, etc.; cambio entre vocales.

Regla nº 25. "k" sorda se transforma en "g" sonora.

k:g ⇒ [§2:ØØ / n4$:nØØ] _ o:o

Usurbilgo ("de Usurbil").

N.L.: Usurbil§ 2**k**o

N.S.: Usurbil ØØ**g**o

El operador "⇒" nos posibilita "Usurbileko" (con "e" epentética (regla nº 2)), también posible en los nombres propios de lugar.

Regla nº 5. Cambio de "a" a "e".

a:e ⇔ _ [7$[a/e/o]:ØØ[a/e/o] / %$[n/l]:ØØ[n,l]]

den ("que es")	aterea ("salido")
N.L.: d**a**% $n	N.L.: ater**a**7 $a
N.S.: d**e**Ø Øn	N.S.: ater**e**Ø Øa

La introducción de varios contextos en las reglas hará que las mismas sean capaces de abarcar un mayor número de casos.

3.3.2. El sistema léxico**.**

____La representación léxica se define asociando a cada una de las entradas el subléxico y la clase de continuación a la que pertenece.

a) Subléxicos : Distinguimos lemas (entradas del diccionario), auxiliares del verbo y las denominadas "estructuras morfotácticas". Estas últimas están compuestas por los prefijos y sufijos, donde se distinguen los casos de declinación, el plural, sufijos del verbo, etc.

El conjunto de todas las entradas se declara dividido en subléxicos.Todas las entradas en los subléxicos llevan su clase de continuación e información morfológica. La información morfológica contiene categoría, relación (de subordinación), cambio de

categoría que produce, caso de declinación, género, número, tiempo-modo, raíz (del verbo auxiliar), tipo de verbo, persona gramatical, más la información específica que cada entrada requiera.

b) Clases de continuación : No siempre son posibles las generalizaciones. Por ejemplo, mientras que con nombres y adjetivos si ha sido posible la asignación de una única clase de continuación a todos los elementos de cada categoría, el caso de adverbios y verbos ha requerido una solución más particularizada. Ej.: A la entrada "polit" (bonito-a) le corresponde la clase de continuación ADJ. Esta a su vez se define así:

ADJ = (I1, ago_egi_en)

I1 = (a_ms, IMG, IMS, IMP)

donde los identificadores en mayúscula son clases de continuación y los en minúsculas subléxicos.

El análisis morfológico, tal y como lo hemos descrito, permite acumular información a medida que vamos extrayendo datos del diccionario según la descomposición en elementos efectuada.

3.4. VENTAJAS DEL SISTEMA DESDE LA PERSPECTIVA LINGÜISTICA.

El modelo de Koskenniemi se basa en el nivel léxico y a partir de él aplica las reglas que van a controlar su realización en superficie. Esto significa que en los subléxicos se guarda la forma completa, sin alteraciones, a diferencia de lo que ocurre en otros sistemas.

Por ejemplo, "ama" ("madre") es la unidad léxica, tiene "a" orgánica, pero el dativo plural será "amei" ("a las madres"); esto hace que en muchos sistemas estemos obligados a tener en el léxico "ama" y "am" para un mismo lexema. En la morfología de dos niveles la clara diferenciación entre el nivel léxico y el de superficie permite

conservar la unidad completa y sin duplicaciones innecesarias. Desde el punto de vista lingüístico, consideramos de enorme importancia la claridad y el respeto de la unidad léxica que supone este enfoque para el análisis morfológico.

El tratamiento de las excepciones -tan frecuentes en una lengua, y de las cuales no se salva por supuesto el euskara- será como siempre el punto crítico a la hora de la aplicación generalizada de un sistema de este tipo.

4. CONCLUSIONES.

Este sistema se ha aplicado al finlandés, y posteriormente a otras lenguas, de diferentes tipos. El aspecto más interesante desde el punto de vista lingüístico es que no limita al lingüista a atenerse a un programa ya adecuado para una determinada lengua, sino que le deja la posibilidad de expresar libremente la morfología de la lengua que se está tratando, sin tener en cuenta el aspecto operativo.

En su primera implementacion en un Burroughs B7800 [Koskenniemi 83]el programa ocupó unas 2000 lineas de Pascal y la velocidad media de análisis por forma fue de 0.1 sg. En nuestro caso se ha hecho una implementación en C con vistas a una posible mejora para su utilización en PC´s y compatibles pero no disponemos todavía de resultados acerca de la eficiencia del programa.

Como se ha dicho más arriba la primera aplicación práctica del analizador en nuestro caso va a ser la construcción de un corrector ortográfico para el euskara: las lenguas con un alto grado de flexión presentan problemas de almacenamiento del diccionario que no se pueden resolver sin un tratamiento adecuado desde el punto de vista de la morfología de las palabras. El corrector dará por buena toda palabra que admita una descomposición morfológica correcta, mientras que la misión del analizador general será la de obtener todas las descomposiciones posibles.

Por otro lado consideramos el analizador-generador morfológico como una herramienta básica indispensable para futuros trabajos dentro del campo del tratamiento del lenguaje natural.

REFERENCIAS :

[Alam 83] Alam, Y. S. . *A two-level analysis of Japanese*. Texas Linguistic Forum, vol 22, Pp. 229-252, 1983.

[Barton 85] Barton, E. . *Computational Complexity in two-level Morphology*, 1985.

[Broda et al. 80] Brodda, B. and F. Karlsson. *An experiment with Automatic Morphological Analysis of Finnish*. Papers from the Institute of Linguistics, University of Stockholm. Publication 40, 1980.

[Euskaltzaindia 73] Euskaltzaindia. *Aditz laguntzaile batua*. Euskaltzaindia, Bilbo 1973.

[Euskaltzaindia 85] Euskaltzaindia. *Euskal Gramatika: Lehen urratsak (I eta II)*. Euskaltzaindia, Bilbo 1985.

[GETA 82] G.E.T.A. . *Le point sur ARIANE-78 debut 1982*. Vol.1, partie 1: Le logiciel. pp. 28-75, 1982.

[Kaplan et al. 81] Kaplan, Ronald M., and M. Kay. *Phonological rules and finite-state transducers*. Paper read at the annual meeting of the Linguistic Society of America in New York City, 1981.

[Karttunen et al. 83] Karttunen, L., and K.Wittenburg. *A two-level analysis of English*. Texas Linguistic Forum, vol 22, Pp. 217-228,1983.

[Karttunen 83] Karttunen, L. . *KIMMO : A two-level Morphological Analyzer*. Texas Linguistic Forum, Vol 22, Pp.165-186, 1983.

[Kay 73] Kay, Martin. *Morphological Analysis*.. A.Zampolli & N. Calzolari eds. (1980). Proc. of the Int. Conference on Computational Linguistics (Pisa), 1973.

[Khan 83] Khan, Robert. *A two-level analysis of Rumanian.* Texas Linguistic Forum, vol 22, Pp. 253-270, 1983.

[Koskenniemi 83] 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.

[Koskenniemi 85] Koskenniemi, K. *Compilation of Automata from Morphological Two-level Rules*. Pp. 143-149. Publication nº 15. University of Helsinki, 1985.

[Lun 83] Lun, S. *A two-level analysis of French*. Texas Linguistic Forum, vol 22, Pp. 271-278, 1983.

[Martí 87] M. A. Martí. *Un sistema de análisis morfológico por ordenado*r. SEPLN nº 4, 1987.

[Meya 87] M. Meya. *Análisis morfológico como ayuda a la recuperación de información.* SEPLN nº 4, 1987.

[Ritchie et al. 87] Ritchie, G.D., S.G. Pulman, A.W.Black and G.J. Russell. A Computational Framework for Lexical Description. Computational Linguistics, vol. 13, numbers 3-4, 1987.

[Sarasola 82] Sarasola, Ibon. *Gaurko euskara idatziaren maiztasun-hiztegia*. (3gn. liburukia), GAK, Donostia, 1982.

[Winograd 83] Winograd, Terry. *Language as a cognitive process*. Vol.1: Syntax, pp 544-549. Addison-Wesley, 1983.