Algoritmos de derivación de palabras con ortografía irregulares en el análisis morfológico automático del inglés antiguo

JAVIER CALLE MARTÍN - JOSÉ LUIS TRIVIÑO RODRÍGUEZ
UNIVERSIDAD DE MÁLAGA

1. INTRODUCCIÓN

El diseño e implementación de un analizador morfológico automático del inglés antiguo (Miranda et al. 1997) supone una ardua tarea de programación a tenor de los inconvenientes que el inglés antiguo plantea fundamentalmente desde una perspectiva ortográfica. De hecho, estas dificultades vienen determinadas principalmente por la falta de consistencia del sistema ortográfico del inglés antiguo. Muchas palabras, debido a la escasa regularidad ortográfica característica del período, no pueden ser etiquetadas y analizadas automáticamente puesto que el diccionario de lexemas carece de alguna de estas fórmulas ortográficas alternativas. A modo ilustrativo, pueden mencionarse, entre otros, los contrastes *fellan/fyllan, elde/ielde* o *clipian/cliopian.* Para dar solución a este inconveniente, por tanto, optamos por la implementación de unos algoritmos de derivación de palabras con el fin de conseguir un análisis morfológico automático para estas palabras con ortografía irregular.

Así pues, a lo largo del presente artículo pretendemos (1) presentar las variedades gráficas vocálicas más usuales que se producen en el inglés antiguo y (2) describir la implementación de nuestro algoritmo de derivación y su actual nivel de efectividad.

2. INCONSISTENCIA ORTOGRÁFICA DEL INGLÉS ANTIGUO

Puede afirmarse que la falta de regularización ortográfica del inglés antiguo está especialmente motivada por los siguientes factores:

a) En primer lugar, los textos escritos del inglés antiguo se transmitían principalmente gracias a la incansable labor de los escribas, los cuales dedicaban su

esfuerzo a la copia de manuscritos. En muchos casos, estos escribas producían alteraciones gráficas respecto al texto original, bien por error, por voluntad propia para adecuarlos a su propio dialecto o bien porque encontraban un mayor grado de evolución, tanto ortográfica como fonética, al trabajar principalmente con textos del período antiguo temprano.

b) En segundo lugar, el período antiguo se caracteriza por la presencia de varios dialectos extendidos a lo largo de toda la isla, el sajón occidental, el kéntico, el mércico y el northumbrio (Baugh y Cable 1978: 52-54) existiendo, en algunos casos, notables diferencias interdialectales en la escritura de una misma palabra. Obsérvense, por ejemplo, los casos de *æfter/efter, weorc/werc, man/mon, hand/hond,* etc. (Cf. de la Cruz 1986: 166-69).

c) En tercer lugar, deben mencionarse también las diferencias intradialectales existentes entre palabras pertenecientes a los distintos períodos del inglés antiguo (inglés antiguo temprano frente al clásico o tardío). Algunos ejemplos representativos son *hie/hi, cyning/cining, neah/neh,* etc. (Cf. de la Cruz 1986: 169-71).

Ateniéndonos a estos factores que intervienen de manera decisiva en la ortografía de las vocales del inglés antiguo, hemos llevado a cabo un exhaustivo estudio de los textos del período con el fin de obtener las diversas realizaciones ortográficas que se producen en las vocales del inglés antiguo, entre las cuales ilustramos las siguientes[1]:

<a>	<æ>	acer	æcer
<a>	<e>	andswarian	andswerian
<i>	<y>	Drihten	Dryhten
<i>	<eo>	cniht	cneoht

3. EL ALGORITMO DE DERIVACIÓN DE PALABRAS

Este algoritmo ha sido especialmente diseñado para realizar una transformación de cada palabra previa al análisis morfológico siguiendo, en cierta medida, el modelo de dos niveles descrito por Koskenniemi (1983: 83). Esta derivación permite transformar una palabra desconocida (cuyo lexema no se encuentra en el diccionario de raíces) en una forma conocida, la cual puede, por tanto, ser analizada correctamente mediante su descomposición en lexema y morfema.

1. Por razones tipográficas no ilustramos ejemplos con vocal larga o macro.

La información necesaria para realizar esta derivación de la palabra se almacena en una base de conocimiento compuesta por un conjunto de reglas de reescritura de la palabra. Estas reglas están almacenadas en forma de tabla ordenada en pares (cadena origen y cadena destino) con el objeto de aumentar la eficacia del algoritmo. Por tanto, al contrario que en el proceso seguido por Koskenniemi, antes de realizar la procedimiento de derivación de la palabra en cuestión se realiza un proceso de análisis morfológico inicial de forma que, si la palabra puede ser analizada directamente (es una palabra conocida), no es necesario realizar el proceso de derivación. De igual forma, no basta con realizar todas las posibles derivaciones de la palabra y posteriormente analizarla, sino que es necesario comprobar si la palabra puede ser analizada tras aplicar una derivación antes de continuar con la siguiente.

Este algoritmo de derivación ha sido implementado para obviar la última vocal de la palabra puesto que ésta normalmente pertenece a la inflexión de la misma lo cual podría generar resultados erróneos en el etiquetado y análisis morfológico de las palabras.

A través del siguiente ejemplo puede observarse el proceso seguido para analizar la palabra *bewendan,* cuya raíz *bewend-* no se encuentra en el diccionario de lexemas, encontrándose en cambio la raíz *bewænd.*

1. Análisis morfológico de la palabra bewendan → La palabra no puede ser analizada.

2. Transformar la palabra aplicando todas las derivaciones posibles a una sola vocal → Obtención de las formas alternativas: *bo̲wendan, bæ̲wendan, by̲wendan, bewo̲ndan, bewæ̲ndan,* etc.

3. Análisis morfológico de las formas alternativas obtenidas → La forma alternativa *bewæ̲ndan* es analizada.

4. Transformar la palabra aplicando todas las derivaciones posibles a dos vocales de manera simultánea → Obtención de las formas alternativas: *bo̲wo̲ndan, bæ̲wo̲ndan, by̲wo̲ndan, bæ̲wo̲ndan, bæ̲wæ̲ndan,* etc.

5. Análisis de las formas alternativas generadas → Ninguna de las formas alternativas puede ser analizada.

6. Tomar como análisis válido para la palabra todos aquellos encontrados para las formas alternativas de la misma.

4. CONCLUSIÓN

Este algoritmo ha sido probado en un analizador morfológico automático de inglés antiguo (Miranda et al. 1997) y, de hecho, ha resultado ser una herra-

mienta lingüística de gran utilidad, no sólo por el grado de efectividad obtenido en el etiquetado morfológico que genera sino también porque permite disminuir considerablemente el número de palabras en el diccionario de lexemas, dando la posibilidad de incluir una sola forma de entre todas las posibles realizaciones ortográficas de cada palabra.

BIBLIOGRAFÍA

BAUGH, A.C. y CABLE, Th. (1978): *A history of the English language,* Londres: Routledge Kegan Paul.

CAMPBELL, A. (1991): *Old English grammar,* Oxford: Oxford University Press.

DE LA CRUZ, J. (1986): *Iniciación práctica al inglés antiguo.* Madrid: Alhambra.

DE LA CRUZ, J., CAÑETE, A., y MIRANDA, A. (1995), *Introducción histórica a la lengua inglesa,* Málaga: Ágora.

DE LA CRUZ, J., CAÑETE, A., y MIRANDA, A. (1995): *Textos y vocabularios para la introducción histórica a la lengua inglesa,* Málaga, Ágora.

KOSKENNIEMI, K. (1983): *Two-level morphology: a general computational model for word-form recognition and production,* Helsinki: University of Helsinki Department of General Linguistics.

LASS, R. (1994): *Old English. A historical companion,* Cambridge: Cambridge University Press.

McCRAE-GIBSON, O.D. (1974): *Learning old English,* Aberdeen.

MIRANDA GARCÍA, A., TRIVIÑO RODRÍGUEZ, J.L. y CALLE MARTÍN, J. (1997): "MAOET: Morphological Analyser of Old English Texts". *Actas del X Congreso Internacional de SELIM,* Zaragoza 16-18 de Octubre de 1997. En prensa.