L I N G U A X
Revista de Lenguas Aplicadas
S E P A R A T A

Carmen Gálvez

El diccionario electrónico: un instrumento para la unificación de términos en la indización automática

UNIVERSIDAD ALFONSO X EL SABIO
Facultad de Lenguas Aplicadas
Villanueva de la Cañada, MMVI

Junio de 2006
https//www.uax.es/publicaciones/linguax/lincom001-06

Universidad Alfonso X el Sabio
28691 - Villanueva de la Cañada - Madrid
ISSN: 1695-632X
Editor: J. Ramón Trujillo - linguax@uax.es
Última actualización: 15 de julio de 2006

EL DICCIONARIO ELECTRÓNICO: UN INSTRUMENTO PARA LA UNIFICACIÓN DE TÉRMINOS EN LA INDIZACIÓN AUTOMÁTICA

Carmen Gálvez[1]
Universidad de Granada

RESUMEN: Se presenta una aplicación basada en técnicas de estado-finito con el objetivo de construir diccionarios electrónicos que se puedan aplicar a los procesos de unificación de términos en la indización automática. El método utilizado para la construcción de las analizadores léxicos consiste en la implementación de una herramienta lingüística que permite elaborar diccionarios electrónicos representados internamente en transductores de estado-finito.

PALABRAS CLAVE: Unificación de términos – Lematización – Reducción de variantes léxicas

ABSTRACT: An application based on state-finite techniques is presented with the objective to build electronic dictionaries that can apply to the processes of term unificatio in the automatic indexation.The method utilized for the construction of the lexicalanalysis tools consists of the implementation of a linguistic enviroment that permits to build electronic dictionaries represented internally in finite-state transducers (FST).

KEY-WORDS: Term unification – Lemmatisation – Conflation of term variants

1. Introducción / Planteamiento

El procesamiento del lenguaje natural (PLN) es un área de investigación que estudia cómo los computadores se pueden utilizar para entender y manipular textos en lenguaje natural. Los investigadores de sistemas PLN necesitan reunir conocimiento suficiente sobre cómo los humanos entienden y utilizan el idioma que, una vez representado y formalizado con las técnicas adecuadas, para implementar ese conocimiento en el desarrollo de *softwares* que realicen las tareas deseadas. La informática, la lingüística, las matemáticas, la inteligencia artíficial, la ingeniería electrónica, o la psicología son

[1] Profesora de Técnicas Automáticas de Explotación Documental en la Universidad de Granada.

disciplinas en las que se apoya este área de investigación. Las aplicaciones de sistemas PLN incluyen varios campos de estudios relacionados con la tecnología de la información y la comunicación, tales como traducción automática, aprendizaje de las lenguas extranjeras, reconocimiento del habla, inteligencia artificial, resumen documental, indización automática, o recuperación de la información. Los avances en la investigación de PLN se publican regularmente en varias actas de conferencias anuales, como ACL (*Association of Computational Linguistics*), COLING (*International Conference on Computational Linguistics*), MUCs (*Message Understanding Conference),* o TRECs (*Text Retrieval Conferences*).

La indización automática, dentro de los procesos documentales, es una de las mayores áreas de aplicación del PLN. Stevens (1965) definió la indización automática como el uso de máquinas para extraer o asignar términos de indización sin intervención humana. Dentro de este proceso, las variantes de los términos de indización es uno de sus grandes obstáculos, que afecta a la posterior recuperación de información en las bases de datos documentales. La reducción de las palabras que tienen una raíz común bajo el mismo término de indización podría incrementar la eficacia en la correspondencia entre los términos de las consultas y los términos almacenados en los registros. Este proceso es especialmente relevante en aquellos idiomas con una morfología compleja (Pirkola, 2001), como es el idioma español

Para evitar la perdida de documentos relevantes, muchos sistemas agrupan las variantes de términos por medio de los denominados métodos de unificación, o *métodos de conflación*, definidos como procedimientos computacionales encargados de la agrupación de variantes de términos, semánticamente equivalentes, a una forma normalizada. Se han propuesto diferentes clasificaciones de estos métodos en recuperación de información (Lennon et al, 1981; Frakes, 1992; Gálvez et al., 2005). En general, los programas que realizan esta función se denominan:

- *stemmers*, cuando se aplican procedimientos para agrupar las formas variantes de un término a un *stem*, definido como la forma base o radical de una palabra.
- l*ematizadores*, cuando se aplican procedimientos para agrupar las formas variantes de un término a un *lema*, definido como el conjunto de palabras con la misma *raíz* y la misma *categoría léxico-gramatical*, o etiqueta *part-of-speech* (*POS*).

En los procesos de lematización, los diccionarios electrónicos constituyen una herramienta básica de análisis. La tarea de construir diccionarios electrónicos completos para una lengua natural es enorme, y requiere una demanda constante de información léxica y detallada sobre amplias áreas de vocabulario. Sin embargo, si el tratamiento morfológico se orienta a una tarea muy concreta, dentro de los sistemas de recuperación de información –y a un vocabulario en un dominio especializado– su diseño no es tan costoso.

1.1. Antecedentes

El método de unificación basado en técnicas de *stemming* implica la eliminación de afijos de acuerdo a un diccionario, que contiene listas de terminaciones de palabra, y un conjunto de reglas. Dentro de estas técnicas, los *stemmers* más conocidos son los algoritmos de Lovins (1968), Dawson (1974), Porter (1980), y Paice (1990) y se suelen aplicar al idioma inglés. Aunque dentro de estos métodos también se encuentran algoritmos específicamente creados para otros idiomas (Popovic y Willett, 1992; Savoy, 1999). El algoritmo de Porter, disponible en *Snowball Web Site* (2006), ha sido implementado para el francés, español, italiano, portugués y alemán, entre otras lenguas.

Frente a las técnicas anteriores, otros métodos se enfrentan al problema de la variabilidad del lenguaje desde una aproximación lingüística, por medio de técnicas cuyo objetivo es la reducción de las variantes léxicas a lemas. En esta línea, una de las implementaciones computaciones más importantes la constituyó el analizador *PC-KIMMO* (Karttunen, 1994) basado en tecnología de estado-finito, posteriormente utilizado en el *Analizador Morfológico de Xerox* desarrollado por el *Multi-Lingual Theory and Technology Group* (MLTT).

Una de las aplicaciones más relevantes de la herramienta diseñado por Xerox es la reducción de variantes léxicas en los sistemas de recuperación de información. Este analizador se ha aplicado, entre otras lenguas, al idioma inglés, danés, alemán, francés, italiano, portugués, o español. Otra herramienta basada en métodos de estado-finito es el analizador morfológico para la lengua inglesa *ENGTWOL* (Voutilainen, 1995). Para el idioma español contamos con la herramienta *COES* (Rodríguez y Carretero, 1996), o el analizador *MACO+* (Carmona et al., 1998). Otros etiquetadotes morfosintácticos para el español son *SPOST* (Farwell et al., 1995), *SMORPH* (Aït-Mokhtar y Rodrigo Mateos, 1995) o el de Gala (1999). En la Universidad

del País Vaso se ha desarrollado *MORFEUS* para el euskera (Alegría, 1995) y en la Universidad Pompeu Fabra se ha desarrollado *CATMORF* para el catalán (Badía, 1997).

2. Análisis léxico por medio de técnicas de estado-finito

La *Teoría de los Lenguajes Formales* se dirige a aquellas expresiones que pueden ser descritas de forma muy precisa, como son los lenguajes de programación. Los lenguajes naturales no son lenguajes formales, y, por tanto, no hay un límite claramente definido entre una sentencia correcta de otra que no lo es. Sin embargo, se pueden adoptar algunas aproximaciones formales a ciertos fenómenos del lenguaje natural susceptibles de una codificación similar a la realizada en los lenguajes de programación. Estas descripciones formales se utilizan por los lingüistas computacionales para expresar teorías sobre aspectos específicos de los lenguajes naturales, tales como el análisis morfológico.

Johnson (1972) fue el primero en observar que determinadas reglas fonológicas y morfológicas se podrían representar por mecanismos de estado-finito, denominando a su formalismo *'two level model'*. La idea del modelo de dos-niveles fue clave para el progreso del formalismo computacional sobre la morfología propuesto por Koskenniemi (1983). El modelo de Koskenniemi estableció una correspondencia entre la forma canónica, o forma léxica, y la forma superficial de las palabras. Esta relación la representó usando Transductores de Estado-Finito, *Finite-State Transducers* (FST).

De forma sintetizada, un transductor es un sistema de representación computacional que comprende un conjunto de estados y una función de transición, que define el cambio de estado. La función de transición se etiqueta con un par de símbolos que constituyen el alfabeto del *input* y el alfabeto de *output*. Este mecanismo se puede representar en la forma de un diagrama, o gráfico de estado-finito. El transductor tomaría cadenas en el *input* y las relacionaría con cadenas en el *output*. Formalmente un FST se define como una tupla de cinco elementos (Roche y Schabes, 1995) que se expresa de la forma siguiente:

$$FST= (\Sigma, Q, i, F, E)$$

donde

Σ=alfabeto de input y output
Q=número de estado
i=estado inicial, i∈Q
F=estado final, F∈Q
E=número de relaciones de transición

En la Figura 1 se muestra la representación gráfica de un transductor cuyos arcos están etiquetados con pares de símbolos que constituyen el alfabeto de *input* y *output*. Por ejemplo, "*a*" denota el símbolo superior y "*b*" el símbolo inferior. Este transductor podría establecer una relación entre: el lenguaje superior y el inferior. Así, este mecanismo podría reconocer la cadena representada por "*ac*" y la podría transformar en la cadena "*bd*". La equiparación es bidireccional, y una cadena de un lenguaje se podría corresponder a una, o más, cadenas de otro lenguaje. Las transducciones son posibles si la cadena en la parte del *input* lleva al transductor a un estado final. Hopcroft y Ullman (1979), Roche y Schabes (1995) y Mohri (1996) proporcionan una explicación más completa de este tipo de sistemas de estado-finito.

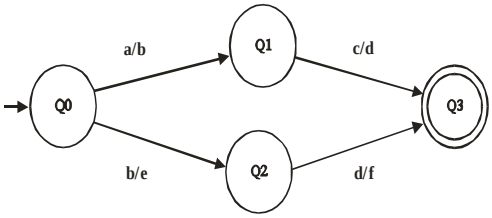

Figura 1. Transductor de estado-finito

La aplicación del formalismo de estado-finito a la unificación de términos parte básicamente de que se puede establecer una relación de equivalencia entre las distintas formas superficiales y la forma normalizada, o lema, a la que se le puede añadir una etiqueta de la categoría gramatical correspondiente, o etiqueta *POS*. Esta correspondencia se puede implementar computacionalmente por medio de transductores (Karttunen et al., 1992). Una analizador de dos-niveles o lematizador desarrollado con tecnología de estado-finito se encargaría de equiparar variantes de términos a formas unificadas, tal y como se representa en la Figura 2.

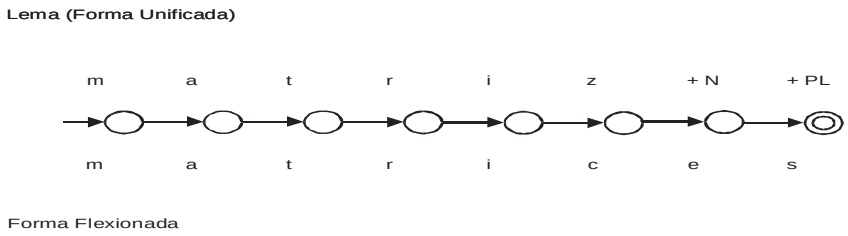

Figura 2. Relación entre una forma variante y una forma unificada (adaptado de Karttunen et al., 1992)

3. Construcción de diccionarios electrónicos

El planteamiento general para la identificación y agrupación de variantes flexionales que hemos adoptado parte del establecimiento de una *relación de equivalencia* entre el lema y la descripción flexional (Gálvez, 2003). Para manipular formalmente esta relación se ha utilizado tecnología de estado finito El procedimiento que se ha seguido para la construcción de los recursos de análisis léxico es la implementación de una herramienta lingüística basada en Transductores de Estado-Finito (Silberztein, 1999). El desarrollo de los analizadores léxicos nos permitirá distinguir las formas flexivas y las irregularidades que se producen en los distintos tipos de flexión y asignar las distintas categorías *POS* a las unidades lingüísticas.

En relación con lo anterior, el análisis flexional se rige por reglas, que aportan los mecanismos para poder relacionar los distintos elementos en el contexto de la oración. En el proceso de flexión se va a tomar como base una unidad genérica denominada *lema*. Por ejemplo, el lema de la palabra {*usuario*} estaría compuesto por el conjunto {*usuario, usuarios, usuaria, usuarias*} formado por todas las cadenas con la misma raíz y la misma categoría general de N (*Nombre*). Frente a esto, otro tipo de análisis, como es el derivacional, no se somete tan claramente a la regularidad de las reglas y muchas palabras compuestas, o derivadas de otras, llegan a transformarse totalmente y se alejan de la palabra origen. La variabilidad en la derivación, y el hecho de que en las palabras derivadas los afijos formen parte de la *raíz*, hace que sea muy difícil fijar cualquier tipo de regularidad en su representación, por esta razón este análisis queda excluido del presente trabajo.

En general, el desarrollo de las herramientas de análisis léxicos ha estado guiado por la representación de la flexión de las palabras en sistemas cerrados o *paradigmas flexivos* (Matthews, 1974). Por otra parte, la estructura del

paradigma hace referencia al número de categorías que puedan aparecer en el interior de los paradigmas. Dependiendo de la variación de ese número se puede hablar de estructuras simples, cuando intervenga sólo una dimensión o categoría, y estructuras complejas, cuando intervengan varias dimensiones o categorías.

La información sobre la morfología flexiva se va a representar en transductores gráficos, por medio de una interfaz. Este recurso nos va a permitir la construcción de tres tipos de diccionarios representados internamente en transductores, que son los que vamos a utilizar para la unificación de términos: *Diccionario de Formas Unificadas*, o lemas (DELAS), *Diccionario Expandido de Formas Flexionadas* (DELAF) y *Diccionario de Formas Compuestas, Nombres Propios, Acrónimos y Abreviaturas* (DELACF).

Básicamente, el procedimiento seguido se sintetiza como sigue: hemos creado manualmente un diccionario (DELAS), que contiene una entrada para cada lema con su etiqueta *POS* correspondiente y un código numérico. Todos los lemas, que pertenecen al mismo paradigma flexivo, se relacionarán con el mismo transductor gráfico. Las entradas del diccionario DELAS se han seleccionado según la oposición binaria 'término marcado/término no marcado'. Por ejemplo, dentro de la categoría N (*Nombre*) y A (*Adjetivo*) se selecciona el término no marcado masculino/singular, y dentro de categoría V (*Verbo*) se elige el infinitivo. Un ejemplo de las entradas en el diccionario de lemas (DELAS) se muestran en la Tabla 1.

DICCIONARIO DELAS
autor, N1
investigador, N1
experimental, A2
documental, A2
documental, N2
éste, PRODE1
iniciar, V1
añadir, V1
citar, V1
solucionar, V3
acceder, V2
alguno, CUANT3

Tabla 1. Entradas del diccionario de lemas (DELAS)

Cada entrada del diccionario DELAS está vinculada a un transductor gráfico que describe el trayecto que el analizador morfológico debe seguir para producir todas las formas flexionadas de un término. Los nombres de los transductores gráficos se corresponderán con las etiquetas *POS* de los lemas y con los códigos numéricos, tal y como se muestra en la Figura 2.

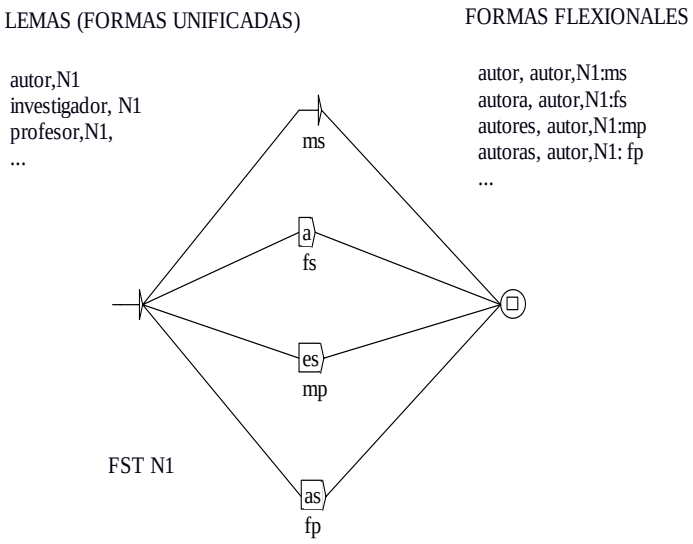

Figura 2. Transductor gráfico

Una vez compilados los gráficos de estado-finito en transductores se proyectan sobre las formas canónicas produciendo de forma automática el *Diccionario Expandido de Formas Flexionadas* (DELAF). Las entradas de este diccionario contienen:

- *Forma canónica*, o *lema*, establecida básicamente a partir de la oposición binaria de los términos *no-marcados*, o negativos, y *marcados*, o positivos. Dentro de la categoría general N (*Nombre*) y A (*Adjetivo*) se seleccionan los términos no-marcados, que son el *masculino/singular* y dentro de la categoría verbo se selecciona el *infinitivo*.
- Categorías léxico-gramaticales, categorías POS, representadas por

los códigos siguientes: A (Adjetivo), ADV (Adverbio), ADVIN (Adverbio Interrogativo), ADVRE (Adverbio Relativo),AIN (Adjetivo Interrogativo), ARE (Adjetivo Relativo), CARD (Cardinal), CUANT (Cuantificador), CONJC (Conjunción de Coordinación), CONJS (Conjunción de Subordinación), DEM (Determinante Demostrativo), DET (Determinante), ORD (Ordinal), PA (Participio Adjetival), POS (Posesivo), PREP (Preposición), PRO (Pronombre), PRODE (Pronombre Demostrativo), PROIN (Pronombre Interrogativo), PRORE (Pronombre Relativo), V (Verbo).

- Información flexional: s (singular), p (plural), m (masculino), f (femenino), n (neutro), W (Infinitivo), P (Participio), G (Gerundio), P1s (1ª persona del singular del Presente de Indicativo), J1s (1ª persona del singular del Pretérito Indefinido), I1s (1ª persona del singular del Pretérito Imperfecto), F1s (1ª persona del singular del Futuro), C1s (1ª persona del singular del Condicional), Y2s (2ª persona del singular del Imperativo), S1s (1ª persona del singular del Presente de Subjuntivo), IS1s (1ª persona del singular del Pretérito Imperfecto del Subjuntivo),...

El tipo de entradas del diccionario de formas flexionadas (DELAF) se muestra en la Tabla 2.

DICCIONARIO DELAF
autor, autor. N1: ms
autora, autor. N1 : fs
autores, autor. N1: mp
autoras, autor. N1: fp
investigador, investigador. N1: ms
investigadora, investigador. N1: fs
investigadores, investigador. N1: ms
investigadoras, investigador. N1: fs
experimentales, experimental. A2 : mp : fp
documental, documental. A2 : ms : fs
documentales, documental. A2 : mp : fp
documental, documental. N2 : ms
documentales, documental. N2 : mp
inicio, iniciar. V1 : P1s

Tabla 2. Entradas del diccionario de formas flexionadas (DELAF)

Por último, el diccionario DELACF contiene términos compuestos. Cada entrada de este diccionario está asociada a su correspondiente forma canónica y a su etiqueta de categoría léxica. Este tipo de diccionario nos aporta los instrumentos necesarios para reconocer términos especializados en un dominio de aplicación, abreviaturas, topónimos, locativos, o siglas. Las entradas de estos diccionarios, tal y como se muestran en la Tabla 3, están integradas igualmente por formas simples, o compuestas, y van seguidas de las categorías morfo-sintácticas vinculadas a los FST gráficos correspondientes.

DICCIONARIO DELACF
a causa de que, a causa de que. CONJS
a disposición de, a disposición de. PREP
al igual, al igual. ADV
Unión Europea, Unión Europea. N10+Loc : fs
Países Bajos, Países Bajos. N102+Top : ms
Asoc,asociación. N : ms
Dep,departamento. N : mp
Chem,chemistry. N : fs
Google, Google. N+PR
Univ. Granada, Universidad de Granada. N : fs
Univ. Carlos III, Universidad Carlos III. N : fs
Univ. Salamanca, Universidad de Salamanca. N : fs

Tabla 3. Entradas del diccionario de formas compuestas (DELACF)

Siguiendo el procedimiento anterior, hemos elaborado un total de *205 paradigmas flexivos* representados en FST gráficos que incluyen las variantes flexivas de palabras simples, compuestas, términos especializados, topónimos, locativos, abreviaturas, términos en latín, anglicismos y siglas.

5. Aplicación de los diccionarios electrónicos

A través del proceso de vinculación de los diccionarios con los FST gráficos hemos obtenido el diccionario de formas flexionadas (DELACF), con 70.500 entradas en total, que es el que realmente vamos a utilizar en el proceso de reconocimiento y unificación de expresiones léxicas. El resultado de la aplicación de esta herramienta se puede presentar de tres modos distintos en el etiquetado lineal:

- Análisis de las unidades léxicas en *{lemas}*

  el acceso mediante browsing a los registro contenido en una base de dato. con el fin de representar las relaciones existente entre las distinto materia que conformar el área de las ciencia social y las humanidad, se haber formado conjunto de documento a partir de los código de clasificación utilizado en la base de dato ISOC. dicha relaciones se representar mediante matriz de co-ocurrencia de número de clasificación.

- Etiquetado de las unidades en *{lemas+etiquetas POS}*

  {el,.DET} {acceso,.N} {mediante,.A} {browsing,.N} {a,.PREP} los {registro,.N} contenidos {en,.PREP} una base {de,.PREP} {dato,.N}.{S} {con el fin de,.PREP} {representar,.V} las relaciones {existente,.A} entre las {distinto,.A} {materia,.N} que {conformar,.V} {el,.DET} {área,.N} {de,.PREP} las {ciencia,.N} {social,.A} {y,.CONJC} las {humanidad,.N}, {se,.PRO} {haber,.V} formado {conjunto,.N} {de,.PREP} {documento,.N} {a partir de,.PREP} los {código,.N} {de,.PREP} {clasificación,.N} {utilizado,.PA} {en,.PREP} la base {de,.PREP} {dato,.N} {ISOC,.N}.{S} {dicha,.N} relaciones {se,.PRO} {representar,.V} {mediante,.A} {matriz,.N} {de,.PREP} {co-ocurrencia,.N} {de,.PREP} {número,.N} {de,.PREP} {clasificación,.N}.{S} Las {matriz,.N} {se,.PRO} {formar,.V} {seguir,.V} la estructura {jerárquico,.A} {de,.PREP} la {propio,.A} {clasificación,.N}.

- Etiquetado de las unidades léxicas en *{formas flexivas+etiquetas POS}*

  {el,.DET} {acceso,.N} {mediante,.A} {browsing,.N} {a,.PREP} los {registros,.N} contenidos {en,.PREP} una base {de,.PREP} {datos,.N}.{S} {con el fin de,.PREP} {representar,.V} las relaciones {existentes,.A} entre las {distintas,.A} {materias,.N} que {conforman,.V} {el,.DET} {área,.N} {de,.PREP} las {ciencias,.N} {sociales,.A} {y,.CONJC} las {humanidades,.N}, {se,.PRO} {han,.V} formado {conjuntos,.N} {de,.PREP} {documentos,.N} {a partir de,.PREP} los {códigos,.N} {de,.PREP} {clasificación,.N} {utilizados,.PA} {en,.PREP} la base {de,.PREP} {datos,.N} {ISOC,.N}.{S} {dichas,.N} relaciones {se,.PRO} {representan,.V} {mediante,.A} {matrices,.N} {de,.PREP} {co-ocurrencia,.N} {de,.PREP} {números,.N} {de,.PREP} {clasificación,.N}.

Es necesario indicar que los recursos de análisis que se han construido han estado orientados a los datos de un dominio específico, con el propósito de poder tratar determinas expresiones propias de la ciencia de la información y documentación. Esta restricción, adoptada por los objetivos prácticos de este trabajo, no afecta a los resultados y, sin embargo, sí evita que muchas palabras no puedan ser analizadas por no estar incluidas en los diccionarios electrónicos.

6. Consideraciones finales

En este trabajo se ha comprobado que los analizadores léxicos desarrollados sólo unifican las variantes que se corresponden a un sólo lema, es decir cuando una mismas variante se puede agrupar a lemas distintos, como *{base,.N}* y *{base,.V},* el analizador no lematiza. Todo esto nos lleva a constatar que la principal limitación de los diccionarios electrónicos basados en técnicas de estado-finito a los procesos de normalización de términos es el *infraanálisis*. Este obstáculo no se puede superar en el etiquetado lineal porque se trata de una limitación inherente de este procedimiento: en los casos de ambigüedad, los analizadores léxicos no son capaces de reducir los términos que se pueden fusionar a distintas formas normalizadas. Sin embargo, se puede considerar que si estas herramientas se diseñan para analizar términos en dominios restringidos de conocimiento, con un vocabulario delimitado, el problema del infraanálisis se podría atenuar, porque disminuirían los casos de ambigüedad (al restringir también los términos a analizar).

Para finalizar, se puede concluir que el procesamiento morfológico es una tarea básica para sistemas PLN como paso previo a otros tratamientos más complejos (como sintaxis o semántica). El proceso de lematización, sobre todo en lenguas de flexión compleja, es útil para aplicaciones de recuperación de información, así como otras aplicaciones lexicográficas. Pero no sólo eso, para cualquier otro tratamiento lingüístico automatizado, el uso de diccionarios completos, o léxicos, es totalmente inviable en lenguas de flexión rica, y sobre todo en lenguas aglutinantes. Por esta razón, el análisis léxico, por medio de transductores, en sí también es útil para desarrollar diccionarios dirigidas a la generación de textos, el análisis de diálogos, o la enseñanza de idiomas asistida por ordenador.

BIBLOGRAFÍA CITADA

AÏT-MOKHTAR, S., RODRIGO MATEOS, J. L. (1995): «Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH», en *Procesamiento del Lenguaje Natural*, vol. 17, pp. 29-41.

ALEGRÍA, I. (1995): *Euskal morfologiaren tratamendu automatirako tresnak*, Tesis doctoral, Universidad del País Vasco.

BADÍA, T. (1997): «CATMORF: multi two-level steps for Catalan morphology», en *Applied Natural Language Proceedings,* ANLP 97, Washington.

CARMONA, J., CERVELL, S., MARQUEZ, L.: MARTI, M. A., PADRO, L., PLACER, R., RODRIGUEZ, H., TAULE, M. y TURMO, J. (1998): «An environment for morphosyntactic processing of Spanish unrestricted text», en *First Internacional Conference on Language Resources and Evaluation,* TREC 98, Granada.

DAWSON, J. L. (1974): «Suffix removal for word conflation», en *Bulletin of the Association for Literary & Linguistic Computing*, vol. 2, n. 3, pp. 33-46.

FARWELL, D., HELMREICH, S., CASPER, M. (1995): «SPOST: a Spanish part of speech tagger», en *Procesamiento del Lenguaje Natural*, vol. 17, pp. 42-53.

FRAKES, W. B. (1992): «Stemming algorithms», en FRAKES, W. B. y BAEZA-YATES, R. (eds.): *Information retrieval: data structures and algorithms*, Prentice-Hall, Englewood Cliffs, NJ.

GALA, N. (1999): «Using the incremental finite-state architecture to create a Spanish shallow parser», en *Procesamiento del Lenguaje Natural*, vol. 25, pp. 75-823.

GALVEZ, C. (2003): *Reconocimiento y normalización de expresiones lingüísticas por medio de transductores de estado-finito*, Tesis doctoral, Universidad de Granada.

GALVEZ, C., MOYA-ANEGON, F. y SOLANA, V. H. (2005): «Term conflation methods in information retrieval: non-linguistic and linguistic approaches», en *Journal of Documentation*, vol. 61, n. 4, pp. 520-547.

HOPCROFT, J. E. y ULLMAN, J. D. (1979): *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley: Reading, MA.

JOHNSON, C. D. (1972): *Formal Aspects of Phonological Description*, Mouton, La Haya.

KARTTUNEN, L. (1994): «Constructing lexical transducers», en *Proceedings of the Fifteenth International Conference on Computational Linguistics*, COLING 94, Kyoto.

KARTTUNEN, L., KAPLAN, R. M. y ZAENEN, A. (1992): «Two-level morphology with composition», en *Proceedings of the 15th International Conference on Computational Linguistics, COLING 92*, Nantes, France.

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

LENNON, M., PIERCE, D. S., TARRY, B. D. y WILLETT, P. (1981): «An evaluation of some conflation algorithms for information retrieval», en *Journal of Information Science*, vol. 3, n. 4, pp. 177-183.

LOVINS, J. B. (1968): Development of a stemming algorithm, en *Mechanical Translation and Computational Linguistics*, vol. 11, pp. 22-31.

MATTHEWS, P. H. (1974): *Morphology. An introduction to the theory of word-structure*, Cambridge University Press.

MOHRI, M. (1996): «On some applications of finite-state automata theory to natural language processing», en *Journal of Natural Language Engineering*, vol. 2, n. 1, pp. 61-80.

PAICE, C. D. (1990): «Another Stemmer», en *ACM SIGIR Forum*, vol. 24, n. 3, pp. 56-61.

PIRKOLA, A. (2001): «Morphological typology of languages for IR», en *Journal of Documentation,* vol. 57, n. 3, pp. 330-348.

POPOVIC, M. y WILLET, P. (1992): «The effectiveness of stemming for natural-language access to Slovene textual data», en *Journal of the American Society for Information Science*, vol. 43, n. 5, pp. 384-90.

PORTER, M. F. (1980): «An algorithm for suffix stripping», en *Program*, vol. 14, pp. 130-137.

ROCHE, E. y SCHABES, Y. (1995): «Deterministic part-of-speech tagging with finite state transducers», en *Computational Linguistics*, vol. 21, n. 2, pp. 227-253.

RODRIGUEZ, S. y CARRETERO, J. (1996): «A formal approach to Spanish morphology: the COES tools», en *XII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, SEPLN, Sevilla, p. 118-126.

STEVENS, M. E. (1965): *Automatic indexing: a state of the art report*, National Bureau of Standards, Washington.

SAVOY, J. (1993): «Stemming of French words based on grammatical categories», en *Journal of the American Society for Information Science*, vol. 44, n. 1, pp. 1-9.

SILBERZTEIN, M. (1999): «Text indexation with INTEX», en *Computers and the Humanities*, vol. 33, n. 3, pp. 265-80.

SONWBALL WEB SITE, accesible en línea [*http://snowball.tartarus.org*]

VOUTILAINEN, A. (1995): «Morphological disambiguation», en KARLSSON, F., VOUTILAINEN, A. y HEIKKILA, J. (eds.): *Constraint grammar: a language-independent system for parsing unrestricted Text*, Mouton de Gruyter, New York, pp. 165-284.