UOC
IN3 Internet Interdisciplinary Institute
www.uoc.edu

Documento de trabajo

Adquisición automática de información léxica y morfosintáctica a partir de corpus sin anotar: aplicación al serbocroata y al ruso

Antoni Oliver (aoliverg@uoc.edu)
Investigador (IN3-UOC)

Irene Castellón (castel@lingua.fil.ub.es)
Departamento de Lingüística General (Universidad de Barcelona)

Lluís Màrquez (lluism@lsi.upc.es)
Centro de investigación TALP (Universidad Politécnica de Cataluña)

Working Paper Series WP03-003

Proyecto de investigación: Desarrollo, evaluación, experimentación y aplicación de técnicas para la traducción no supervisada y el tratamiento de la información textual (INTERLINGUA)

Fecha de publicación: julio de 2003

Internet Interdisciplinary Institute (IN3): http://www.uoc.edu/in3

RESUMEN

En este artículo presentamos una metodología para la adquisición automática de información léxica y morfosintáctica a partir de corpus sin anotar. El sistema utiliza información sobre la morfología flexiva de la lengua que se desea tratar, así como información léxica y morfosintáctica de las palabras pertenecientes a clases no flexivas y de las palabras cuya flexión no responde a paradigmas regulares. Es un sistema en desarrollo, por lo que las evaluaciones que incluimos son preliminares.

PALABRAS CLAVE

morfología computacional, análisis morfológico, lenguas eslavas

SUMARIO

Para citar este documento, puedes utilizar la siguiente referencia:

OLIVER, Antoni; CASTELLÓN, Irene; MÀRQUEZ, Lluís (2003). *Adquisición automática de información léxica y morfosintáctica a partir de corpus sin anotar: aplicación al serbocroata y al ruso* [documento de trabajo en línea]. IN3 : UOC. (Working Paper Series; WP03-004) [Fecha de consulta: dd/mm/aa].
<http://www.uoc.edu/in3/dt/20255/index.html>

Adquisición automática de información léxica y morfosintáctica a partir de corpus sin anotar: aplicación al serbocroata y al ruso[*]

1. Introducción

Uno de los principales inconvenientes para el desarrollo de analizadores morfológicos es la necesidad de recursos léxicos extensos y con información morfosintáctica suficiente. Este hecho afecta especialmente a lenguas que no disponen de recursos a gran escala. Se han desarrollado diversos analizadores morfológicos, por ejemplo Martí (1988) y Badia (1997), a partir de diccionarios de la lengua en formato electrónico con procesos automáticos (aprovechando la información morfológica presente en los diccionarios), pero a menudo precisan una importante revisión manual. También existen metodologías para la extracción automática de información morfológica sin ningún conocimiento previo sobre la lengua a partir de corpus sin anotar (Goldsmith, 2001). Esta aproximación consigue extraer información morfológica interesante pero no suficiente para aplicarla directamente en un analizador morfológico. En este artículo presentamos una metodología para la extracción automática de información léxica y morfosintáctica a partir de corpus sin anotar que trabaja tomando como base el conocimiento de la morfología de la lengua expresada en forma de reglas. Así, el proceso de adquisición se resume en una tarea de clasificación automática de las formas de un corpus en clases morfológicas.

Se han desarrollado prototipos experimentales para dos lenguas eslavas: el serbocroata y el ruso. Estas lenguas se caracterizan por poseer una morfología flexiva muy rica. Ambas son declinables (6 casos el ruso: nominativo, genitivo, dativo, acusativo, instrumental y prepositivo; 7 casos el serbocroata: los ya mencionados más el vocativo). El sistema verbal presenta también numerosas formas para cada lema. Esta característica se ha aprovechado para idear el sistema de adquisición automática. En las tablas 1 y 2 presentamos ejemplos de paradigmas flexivos.[1]

Caso	Singular	Plural
Nominativo	*jelen*	*jeleni*
Genitivo	*jelena*	*jelena*
Dativo	*jelenu*	*jelenima*
Acusativo	*jelena*	*jelene*
Vocativo	*jelene*	*jeleni*
Locativo	*jelenu*	*jelenima*
Instrumental	*jelenom*	*jelenima*

Tabla 1. **Declinación de sustantivos masculinos en consonante**

* Este artículo fue presentado en forma de ponencia en el XVIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), celebrado en la Universidad de Valladolid (Valladolid, 11, 12 y 13 de septiembre de 2002)
1. Los ejemplos los presentamos en serbocroata, para facilitar su lectura, ya que utiliza alfabeto latino.

Caso	Singular	Plural
Nominativo	*srna*	*srne*
Genitivo	*srne*	*srna*
Dativo	*srni*	*srnama*
Acusativo	*srnu*	*srne*
Vocativo	*srno*	*srne*
Locativo	*srni*	*srnama*
Instrumental	*srnom*	*srnama*

Tabla 2. **Declinación de sustantivos femeninos en *-a***

El primer sistema experimental se desarrolló en Prolog para el serbocroata (Oliver, 2001), y los resultados obtenidos mostraron, por un lado, la utilidad del método y, por otro, ciertas carencias que intentamos solventar en un nuevo prototipo aplicado al ruso. Al tratarse de la expresión de reglas morfológicas flexivas, la implementación del nuevo prototipo se está realizando en Perl, ya que permite el uso de expresiones regulares de una forma cómoda, además de mejorar notablemente la velocidad.

En la siguiente sección (2) presentamos la organización general del sistema y cada uno de sus componentes. En la sección 3 trataremos la metodología básica seguida para la adquisición. El siguiente apartado (4) tratará la ambigüedad de las reglas y la solución propuesta en el mecanismo de adquisición, mientras que en la sección 5 presentaremos la evaluación preliminar del sistema y, por último, expondremos las conclusiones y las líneas futuras.

2. Componentes del sistema de adquisición

El sistema de adquisición automática está formado por los siguientes componentes:

- Lista de clases no flexivas y de clases cerradas.
- Lista de palabras irregulares.
- Reglas morfológicas.
- Corpus.

A continuación presentamos cada uno de estos componentes.

2.1. Lista de clases no flexivas y de clases cerradas

Las palabras pertenecientes a clases no flexivas y a clases cerradas quedan excluidas del proceso de adquisición automática. Se han confeccionado manualmente listas de palabras correspondientes a las clases enumeradas en la tabla 3.

Categoría	Serbocroata	Ruso
Pronombres	2.049	1.134
Numerales	649	en desarrollo
Preposiciones	67	122
Conjunciones	34	87
Interjecciones	50	185
Partículas	9	105
Adverbios	224	1.389

Tabla 3. **Número de formas en las listas de clases no flexivas**

La inclusión de los adverbios en estas listas es provisional; en próximas versiones del sistema de adquisición se incluirán reglas de morfología derivativa de formación de adverbios a partir de otras categorías gramaticales (por ejemplo, a partir del adjetivo *brz* se forma el adverbio *brzo*). De esta manera se creará, en cierto modo, un paradigma derivativo que se podrá aprovechar para la adquisición automática.

2.2. Lista de palabras irregulares

Las palabras irregulares quedan también excluidas del proceso de adquisición. Estas palabras se incluyen en una lista que contiene todas sus formas y sus correspondientes descripciones morfosintácticas. En el sistema experimental de serbocroata se confeccionó manualmente una lista de 2.050 formas correspondientes a 93 sustantivos y 19 verbos. Queda por desarrollar la lista de excepciones del ruso.

El concepto de irregularidad está relacionado con la no adscripción de un lema a un paradigma determinado. Por lo tanto, la regularidad o irregularidad irá asociada al número de paradigmas presentes en nuestro modelo, lo que desde el punto de vista computacional equivale a decir el número de paradigmas implementados en nuestro sistema. Un caso extremo lo constituye el analizador morfológico del croata de Tadic (Tadic, 1994). Este analizador no prevé ninguna palabra irregular, sino que todas corresponden a algún paradigma, hasta el punto de que algunos paradigmas son aplicables a una única palabra. De esta manera, el número de paradigmas llega a ser muy elevado (el analizador de Tadic trabaja con 404 modelos para los sustantivos, 51 para los adjetivos y 154 para los verbos). En cambio, nuestro sistema experimental para el serbocroata trabaja con 18 modelos para los sustantivos, 2 para los adjetivos y 16 para los verbos. El sistema para el ruso, actualmente en desarrollo, cuenta con 41 modelos para los sustantivos y 11 modelos para los adjetivos. Este nivel de granularidad en la descripción lingüística nos permite un menor esfuerzo en la implementación de las reglas y asegura que la adquisición se realizará sobre los modelos más productivos (que corresponden a modelos más regulares).

2.3. Reglas morfológicas

Las reglas morfológicas se han implementado siguiendo un formalismo de descomposición morfológica, o *morphological stripping* (Alshawi, 1992). No se ha escogido una aproximación de dos niveles (Koskenniemi, 1983) porque los fenómenos de cambios fonéticos para las lenguas tratadas se producen en contextos morfografémicos y morfotácticos determinados. En Badia, Egea y Tuells (1997), se propone una variación del formalismo de dos niveles que permite expresar contextos morfografémicos y morfotácticos que restringen la aplicación de las reglas, pero nosotros hemos optado por utilizar unas reglas que trabajan en un único nivel, lo que facilita la implementación.

Las reglas morfológicas son de la forma siguiente:

TL:TF:Desc

donde *TL* significa terminación del lema, *TF* significa terminación de la forma y *Desc* contiene la descripción morfológica.

La descripción morfológica se expresa mediante una serie de etiquetas similares a las expuestas en Multext East (Erjavec, 2001). En la tabla 4 podemos ver las reglas correspondientes a los paradigmas de las tablas 1 y 2, respectivamente. En las categorías cada dígito es informativo, por ejemplo *NCMSN* indica "nombre común masculino singular nominativo".[2]

2. Para ver una explicación de todos los dígitos: <http://www.lpl.univ-aix.fr/projects/multext/[url1]>.

Masculinos	Femeninos
::NCMSN	a:a:NCFSN
a::NCMSG	e:a:NCFSG
u::NCMSD	i:a:NCFSD
a::NCMSA	u:a:NCFSA
e::NCMSV	o:a:NCFSV
u::NCMSL	i:a:NCFSL
om::NCMSI	om:a:NCFSI
i::NCMPN	e:a:NCFPN
a::NCMPG	a:a:NCFPG
ima::NCMPD	ama:a:NCFPD
e::NCMPA	e:a:NCFPA
i::NCMPV	e:a:NCFPV
ima::NCMPL	ama:a:NCFPL
ima::NCMPI	ama:a:NCFPI

Tabla 4. **Reglas correspondientes al paradigma de los sustantivos masculinos acabados en consonante (para este paradigma la terminación de forma es nula) y al paradigma de los sustantivos femeninos acabados en *-a***

Estas reglas operan mediante substituciones que utilizan expresiones regulares de Perl que se evalúan para llevar a cabo la transformación. Por ejemplo:

u:a:NCFSA

Esta regla se transforma en la substitución

s/a$/u/;

para formar el acusativo singular a partir del lema,

donde *s* indica substitución de *a$*, es decir, una *a* final por *u*. Por ejemplo, esta regla trataría la formación del acusativo *knjigu* a partir del nominativo *knjiga*.

El uso de expresiones regulares facilita mucho la escritura de reglas de aquellos paradigmas que presentan particularidades. Veamos algunos ejemplos:

- A *móvil*: la vocal *a* en ciertas posiciones se pierde en diferentes casos del paradigma. Por ejemplo, el sustantivo *centar* tiene el genitivo singular *centra*, y no *centara*. Esta particularidad se puede expresar fácilmente con la regla siguiente:

  \1a:a([^aeiou]):NCMSG

  Esta regla se transforma en la substitución

  s/a([^aeiou])$/\1a/;

  lo que da el resultado deseado (a partir del nominativo *centar* se forma el genitivo *centra*).

- *Sibilarización*: en algunas formas del paradigma las consonantes *k*, *g*, *h* delante de *i* cambian a *c*, *z*, *s*. Este cambio, por ejemplo, tiene lugar en la formación del dativo y el locativo singular de los sustantivos femeninos (*majka* NCFSN a *majci* NCFSD, NCFSL). La expresión de este cambio fonético se puede hacer fácilmente mediante tres reglas:

```
ci:ka:NCFSD,NCFSL
zi:ga:NCFSD,NCFSL
si:ha:NCFSD,NCFSL
```

Para evitar aplicar la regla correspondiente a los sustantivos femeninos que no presentan este cambio, la regla general debe modificarse y ha de quedar así:

```
\1i:[^kgh]a:NCFSD,NCFSL
```

- En la formación del comparativo de los adjetivos acabados en *p, b, v* aparece una *l* entre esta consonante y la terminación de adjetivo. Por ejemplo, el comparativo de *skup* es *skuplji*. Este fenómeno se puede expresar con la regla siguiente:

```
\1lji:([pbv]):ADCMSN
```

Esta regla se transforma en la substitución

```
s/([pbv])$/\1lji/;
```

En la tabla 5 podemos ver el número total de reglas desarrolladas manualmente para el serbocroata y el ruso.

Categoría	Serbocroata	Ruso
Sustantivos	557	822
Adjetivos	1.374	228
Verbos	1.582	en desarrollo[*]

Tabla 5. **Número de reglas desarrolladas**

[*] Se han desarrollado 105 reglas correspondientes al infinito; 516, al presente, y 633, al futuro.

2.4. Corpus

Se ha recopilado un corpus del serbocroata de nueve millones de palabras a partir de obras literarias, diarios y revistas. Para la recopilación se ha atendido a los siguientes criterios:

- Los textos están en dialecto *štokavski* y corresponden tanto a la variante occidental como a la oriental.

- En el caso de diarios y revistas se han incluido publicaciones de Croacia, Bosnia-Herzegovina y Serbia.

- Las obras literarias corresponden en su totalidad al siglo XX y, por lo tanto, están escritas en la ortografía actual.

Se ha procesado este corpus para obtener una lista de ocurrencias de formas. El número

total de apariciones diferentes es de 375.489.

Actualmente se está recopilando un corpus de características similares para el ruso. En estos momentos el corpus del ruso es de unos diez millones de palabras.

3. Metodología de adquisición

En la figura 1 podemos ver el esquema básico de funcionamiento de la metodología de adquisición. El corpus de entrada se ha reducido previamente a una lista de las diferentes formas que aparecen en él con su frecuencia de aparición. Para cada forma de entrada se verifica si corresponde a una clase no flexiva o a una excepción. Si alguna forma regular coincide con una clase no flexiva o excepción deberá también aparecer en las listas de clases no flexivas y excepciones. Para cada forma se aplican las reglas morfológicas para intentar deducir el lema correspondiente. Si el lema hipotetizado existe en el corpus, se introduce una nueva entrada, constituida por la forma, el lema asociado y la descripción morfosintáctica dada por la regla, en la base de datos léxica. Como veremos en el siguiente apartado esta base de datos léxica se tendrá que depurar, ya que no todas las entradas adquiridas de esta manera serán correctas.

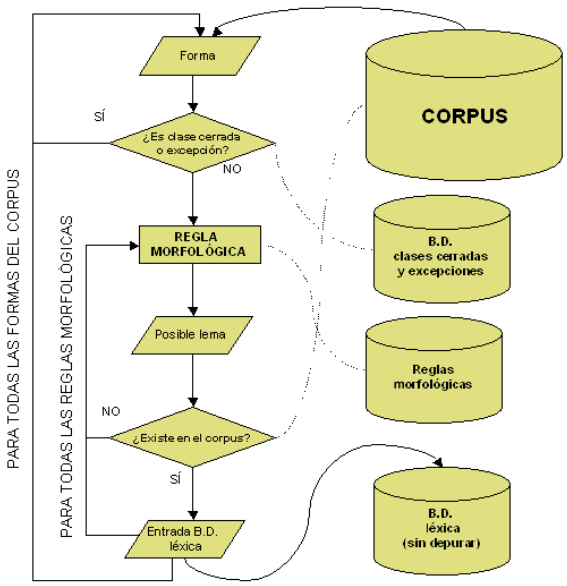

Figura 1. **Funcionamiento de la metodología de adquisición**

Veamos un ejemplo de la metodología de adquisición. Consideremos que tenemos un corpus compuesto por todas las formas de las tablas 1 y 2, es decir, constituido por las formas *jelen*, *jelena*, *jelenu*, *jelene*, *jelenom*, *jeleni*, *jelenima*, *srna*, *srne*, *srni*, *srnu*, *srno*, *srnom*, *srnama*. Consideremos la forma de entrada *srnom*. Al aplicar la regla morfológica

om::NCMSI

se genera la pseudoraíz *srn*, y concatenando la terminación de lema (en este caso nula) se hipotetiza el lema *srn*. Este lema no existe en el corpus, por lo que no se generaría una nueva entrada de la base de datos léxica. En cambio, al aplicar la regla morfológica

om:a:NCFSI

se hipotetiza el lema *srna*, en este caso existente en el corpus. En consecuencia, se generaría una nueva entrada de la base de datos léxica *srnom:srna:NCFSI* (forma:lema:descripción).

4. Análisis de la ambigüedad de las reglas y discriminación de las adquisiciones correctas

En el apartado anterior se ha explicado el procedimiento básico de adquisición: a partir de una forma y una regla se hipotetiza un lema; la existencia de este lema se verifica en el propio corpus. La existencia de esta forma (la correspondiente al lema hipotetizado) no garantiza que se trate de un lema, ya que puede tratarse de una forma de otro paradigma. Ilustraremos este hecho con un ejemplo. Consideremos nuevamente que tenemos un corpus constituido por todas las formas de las tablas 1 y 2. El sistema de adquisición ante una forma como *jelenom* aplicará la regla siguiente:

om::NCMSI

e hipotetizará el lema *jelen*, que al existir en el corpus formará la entrada *jelenom:jelen:NCMSI*. Pero el sistema continuará aplicando las reglas morfológicas y al aplicar la regla

om:a:NCFSI

hipotetizará el lema *jelena*. Esta forma existe en el corpus, pero no es un lema, sino que es el acusativo y el genitivo singular de *jelen*. El sistema, por tanto, ha confundido una forma de un paradigma con el lema de otro paradigma, por lo que añadiría a la base de datos una entrada incorrecta *jelenom:jelena:NCFSI*.

Así pues, se producen interferencias entre las reglas. Es necesario detectar estas interferencias e idear un método para depurar la base de datos léxica de manera que queden en ella únicamente las entradas deducidas correctamente.

El sistema experimental de serbocroata se basa simplemente en establecer un número de reglas mínimo que deben ser aplicadas para dar por válido un lema hipotetizado. Esto es equivalente a hablar de número de formas adquiridas para un mismo lema y un mismo paradigma.

En el sistema experimental se estableció de forma empírica en tres el número de reglas mínimo para validar un lema de un paradigma. Los resultados conseguidos con esta aproximación, que exponemos en el siguiente apartado, fueros satisfactorios. No obstante, se evidenciaba la necesidad de idear un sistema de análisis más desarrollado para detectar y solucionar la ambigüedad de las reglas.

Conscientes de esta necesidad, se ha implementado un algoritmo de detección y resolución de la ambigüedad entre reglas. El algoritmo funciona basándose en los siguientes principios:

- Dos reglas morfológicas son ambiguas si la terminación de lema de una coincide con la terminación de forma de la otra.

- La ambigüedad se puede resolver observando si existe alguna terminación de forma de uno de los paradigmas que no esté presente en el otro paradigma.

En el caso del ejemplo anterior el algoritmo detecta que existe una ambigüedad entre los dos paradigmas (la terminación de lema del paradigma femenino *a* coincide con la terminación de acusativo y genitivo singular del paradigma masculino), e indica que la existencia de las formas acabadas en *-o* y *-ama* muestran que se trata del paradigma femenino, mientras que la existencia de la forma en *-ima* indica que se trata del paradigma masculino.

Para poder analizar las reglas con esta metodología han sido numeradas todas ellas con un código que identifica el paradigma y un número de regla dentro de este paradigma (véase la tabla 6). El algoritmo de análisis de reglas utiliza esta numeración para identificar las ambigüedades entre paradigmas y, a la vez, las formas que servirán para desambiguar entre uno y otro paradigma. El algoritmo da como resultado (siguiendo el mismo ejemplo): N1:N2;(N1-10, N1-13, N1-14 | N2-5, N2-10, N2-14, N2-15). Esta nomenclatura indica que entre los paradigmas 1 y 2 existe ambigüedad y que las formas obtenidas a partir de las reglas N1-10, N1-13, N1-14 validan el paradigma N1, y que las formas obtenidas a partir de las reglas N2-5, N2-10, N2-14, N2-15 validan el paradigma N2.[3]

Masculinos	Femeninos
::NCMSN:N1-1	a:a:NCFSN:N2-1
a::NCMSG:N1-2	e:a:NCFSG:N2-2
u::NCMSD:N1-3	i:a:NCFSD:N2-3
a::NCMSA:N1-4	u:a:NCFSA:N2-4
e::NCMSV:N1-5	o:a:NCFSV:N2-5
u::NCMSL:N1-6	i:a:NCFSL:N2-6
om::NCMSI:N1-7	om:a:NCFSI:N2-7
i::NCMPN:N1-8	e:a:NCFPN:N2-8
a::NCMPG:N1-9	a:a:NCFPG:N2-9
ima::NCMPD:N1-10	ama:a:NCFPD:N2-10
e::NCMPA:N1-11	e:a:NCFPA:N2-11
i::NCMPV:N1-12	e:a:NCFPV:N2-12
ima::NCMPL:N1-13	ama:a:NCFPL:N2-13
ima::NCMPI:N1-14	ama:a:NCFPI:N2-14

Tabla 6. **Reglas correspondientes al paradigma de los sustantivos masculinos acabados en consonante y al paradigma de los sustantivos femeninos acabados en *-a*, incluyendo numeración de reglas**

El procedimiento que se seguirá será el siguiente:

- Realizar la extracción automática de todas las formas aplicando todas las reglas. El resultado de la extracción para cada forma dará la información de forma:lema:descripción:regla-utilizada (por ejemplo, *jelenom:jelen:NCMSI:N1-7*).

- Utilizar la información obtenida a partir del algoritmo de análisis de reglas para validar aquellos lemas asociados a paradigmas que sean correctos y eliminar los incorrectos.

- Depurar el formario obtenido eliminando las formas correspondientes a lemas y paradigmas incorrectos.

3. En el sistema real, las reglas N1-10, N1-13, N1-14 se simplifican en una sola regla: *ima::NCMPD,NCMPL,NCMPI*. Lo mismo ocurre con las reglas N2-10, N2-14 y N2-15, que se simplifican en *ama:a:NCFPD,NCFPL,NCFPI*.

5. Evaluación del sistema de adquisición

Por el momento se ha evaluado el sistema de adquisición experimental del serbocroata. La evaluación de los resultados constituyó un problema importante, ya que no se disponía de un formario de referencia para esta lengua, por lo que la evaluación fue manual y muy costosa. Se validaron manualmente un total de 9.100 lemas.

La tarea concreta que se evaluó fue la correcta extracción de lemas y su clasificación en un determinado paradigma. Recordemos que para este sistema un lema asociado a un paradigma se valida si cumple un mínimo de tres reglas morfológicas o, dicho de otra manera, si existen tres formas en el corpus para dicho lema.

Categoría	Adquiridos	Precisión
Sustantivos	11.839	88,0%
Adjetivos	4.511	90,9%
Verbos	3.104	89,0%
TOTAL	19.454	88,9%

Tabla 7. **Resultados obtenidos con el sistema experimental de serbocroata**

Los resultados de precisión (número de lemas correctos y correctamente clasificados sobre el total de lemas extraídos) son satisfactorios, aunque evidencian la necesidad de realizar un análisis más exhaustivo de la ambigüedad de las reglas y de discriminación de las adquisiciones correctas.

En el caso del ruso se está desarrollando un formario de referencia a partir de un diccionario morfológico (Zaliznjak, 1977). Este formario se utilizará para evaluar de una manera rápida y fiable el sistema de adquisición para el ruso.

6. Conclusiones y líneas futuras

Hemos presentado un sistema de adquisición automática de información léxica y morfosintáctica a partir de corpus sin anotar. Este método se basa en la coaparición en el corpus de diferentes formas de un mismo paradigma. Los resultados obtenidos con el sistema experimental de serbocroata son alentadores, aunque se debe mejorar el sistema de análisis de ambigüedad entre las reglas y discriminación de las adquisiciones correctas. Por otro lado, la falta de recursos del serbocroata hizo muy costosa la evaluación del método de adquisición.

Se está desarrollando un sistema de adquisición para el ruso, siguiendo la misma metodología, al mismo tiempo que se elabora un formario suficientemente completo para la evaluación del método. En este caso, las reglas que se utilizan para la generación del formario a partir del diccionario morfológico son las mismas que se emplean para la adquisición, de forma que este formario sea un punto de referencia para evaluar el nuevo sistema de adquisición.

Uno de los aspectos que queremos desarrollar, relacionado con lo mencionado anteriormente, es la validación de los lemas hipotetizados. Una mejora prevista es la ponderación de los lemas basándose en el estudio de la ambigüedad de las reglas, de modo que sea posible la adquisición de un subconjunto válido y la validación manual del subconjunto de lemas ponderados por debajo de un determinado índice de bondad.

Lista de URL:

[url1]:http://www.lpl.univ-aix.fr/projects/multext/

Bibliografía:

ALSHAWI, H. (ed.) (1992). *The core lnguage*. MIT Press.

BADIA, T.; EGEA, A.; TUELLS, A. (1997). "Catmorf: Multi two-levels steps for Catalan morphology". En: *Demo proceedins of ANLP 97.*

ERJAVEC, T. (2001). *Specifications and notation for multext-east lexicon encoding*. Informe técnico. Multext-East/Concede.

GOLDSMITH, J. (2001). "Unsupervised learning of the morphology of a natural language". *Computational Linguistics*. Vol. 27, núm. 2, págs. 153-198.

KOSKENNIEMI, K. (1983). *Two-level morphology: ageneral computational model for word recognition and production*. University of Helsinki. (Publications, núm. 11).

MARTÍ, M.A. (1998). *Processament informàtic del llenguatge natural: un sistema d'anàlisi morfològica per ordinador.* Tesis doctoral. Departamento de Filología Románica. Facultad de Filología de la Universidad de Barcelona.

OLIVER, A. (2001). *Processament morfològic de les llengües eslaves: el serbocroat*. Trabajo para la obtención de la D.E.A. Universidad de Barcelona.

TADIC, M. *Racunalna obradba morfologije hrvatskoga kjiževnog jezika*. Tesis doctoral. Zagreb: Sveucilište u Zagrebu, Filozofski Fakultet.

ZALIZNJAK, A.A. (1977). *Grammatitxeskii slovar russkogo jazika*. Moskva: Russkii jazik.

Fecha de publicación: julio de 2003