Procesamiento del Lenguaje Natural, num. 29 (2002), pp.97-104 recibido 25-05-2002; aceptado 02-07-2002

Adquisición automática de información léxica y morfosintáctica a partir de corpus sin anotar: aplicación al serbo-croata y ruso *

Antoni Oliver
D. Lingüística General
Universitat de Barcelona
aoliverg@clic.fil.ub.es

Irene Castellón
D. Lingüística General
Universitat de Barcelona
castel@lingua.fil.ub.es

Lluís Màrquez
Centre de recerca TALP
LSI, UPC
lluism@lsi.upc.es

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 a 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. Se trata de un sistema en desarrollo por lo que las evaluaciones que incluimos son preliminares.
Palabras clave: morfología computacional, análisis morfológico, lenguas eslavas

Abstract: This paper presents a methodology for the automatic acquisition of lexical and morpho-syntactic information from raw corpora. The system uses information about the inflectional morphology of the language and lexical and morpho-syntactic information of the words belonging to non-inflectional categories and of the words not belonging to regular paradigms.
Keywords: computational morphology, morphological analysis, Slavonic languages

1. *Introducción*

Uno de los principales inconvenientes para el desarrollo de analizadores morfológicos es la necesidad de recursos léxicos extensos 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; Martí et al., 2001; Badia, 1997) a partir de diccionarios de la lengua en formato electrónico con procesos automáticos (aprovechando la información morfológica presente en los diccionarios) pero que 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 a partir del 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 serbo-croata y el ruso. Estas lenguas se caracterizan por poseer una morfología flexiva muy rica. Ambas lenguas son declinables (6 casos el ruso: nominativo, genitivo, dativo, acusativo, instrumental y prepositivo; 7 casos el serbo-croata: 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]

El primer sistema experimental se desarrolló en Prolog para el serbo-croata (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

* Esta investigación se ha llevado a cabo gracias a la ayuda del proyecto HERMES (TIC 2000-0335-C03-02)

[1] Los ejemplos los presentamos en serbo-croata, para facilitar su lectura, ya que utiliza alfabeto latino.

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

Cuadro 1: Declinación de los sustantivos masculinos acabados en consonante (*jelen*, ciervo)

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

Cuadro 2: Declinación de los sustantivos femeninos acabados en -a (*srna*, corza)

de sus componentes. En la sección 3 trataremos la metodología básica seguida para la adquisición. La sección 4 tratará de la ambigüedad de las reglas y de la solución propuesta en el mecanismo de adquisición, 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	S-croata	Ruso
Pronombres	2049	1134
Numerales	649	706
Preposiciones	67	122
Conjunciones	34	87
Interjecciones	50	185
Partículas	9	105
Adverbios	224	1389

Cuadro 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 cierta forma, 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 serbo-croata 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 Tadić (Tadić, 1994) o del ruso de Mikheev (Mikheev y Liubushkina, 1996). Estos analizadores no contemplan ninguna palabra irregular, sino que todas corresponden a algún paradigma, hasta el punto que algunos paradigmas son aplicables a una única palabra. De esta manera el número de

paradigmas se llega a hacer muy elevado (el analizador de Tadić trabaja con 404 modelos para los sustantivos, 51 para los adjetivos y 154 para los verbos). En cambio nuestro sistema experimental para el serbo-croata 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, 11 modelos para los adjetivos y 74 para los verbos. 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 —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 morfológicos 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, facilitando de esta manera la implementación.

Las reglas morfológicas son de la forma:

```
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 etiquetas morfosintácticas cada símbolo es informativo, por ejemplo NCMSN indica 'Nombre común masculino singular nominativo' [2].

Estas reglas operan mediante substituciones que utilizan expresiones regulares de Perl

[2]para ver una explicación del sistema de etiquetas Multext http://www.lpl.univ-aix.fr/projects/multext/

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

Cuadro 4: Reglas correspondiente 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*

que se evalúan para llevar a cabo la transformación. Por ejemplo, la regla

```
u:a:NCFSA
```

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 en la regla:

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

  Esta regla se transforma en la substitución

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

Categoría	S-croata	Ruso
Substantivos	557	817
Adjetivos	1374	236
Verbos	1582	2108

Cuadro 5: Número de reglas desarrolladas

Donde `\1` significa el caracter substituido de la parte de la expresión que está entre paréntesis, es decir `[^aeiou]`, que a su vez significa un caracter de los no incluidos en la lista, es decir, cualquier consonante. Esta substitución 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 locativo singular de los sustantivos femeninos (*majka* NCFSN $\longrightarrow$ *majci* NCFSD, NCFSL). La expresión de este cambio fonético se puede hacer de manera fácil 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 se ha de modificar y debe quedar

```
\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

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

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 serbo-croata y el ruso.

2.4. Corpus

Se ha recopilado un corpus del serbocroata de 9.000.000 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 y Hercegovina y Serbia.
- Las obras literarias corresponden en su totalidad al siglo XX y por 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.

Se ha recopilado también un corpus de similares características para el ruso de 16.000.000 de palabras. El número total de apariciones diferentes en este caso es de 540.193.

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.

Veamos un ejemplo de la metodología de adquisición. Consideremos que tenemos un corpus que contiene todas las formas de las tablas 1 y 2, es decir, que contiene las formas: *jelen, jelena, jelenu, jelene, jelenom, jeleni,*

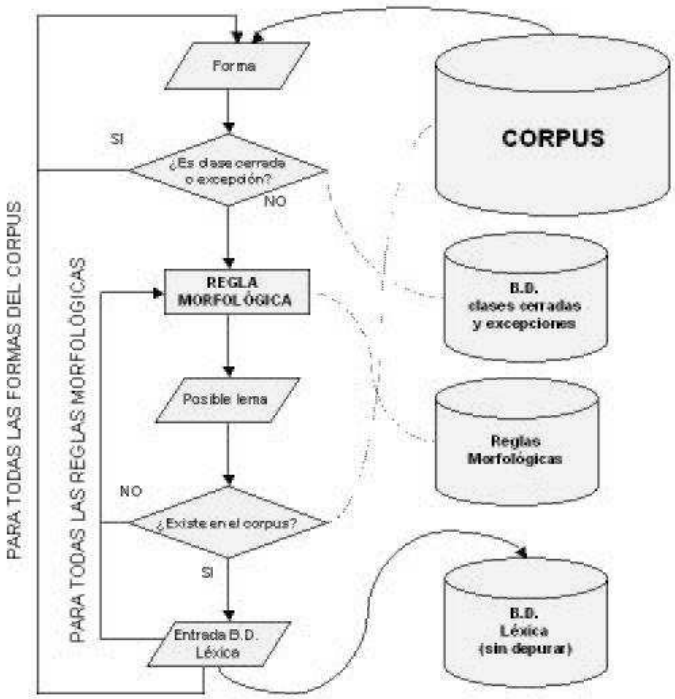

Figura 1: Funcionamiento de la metodología de adquisición

jelenima, srna, srne, srni, srnu, srno, srnom, srnama. Consideremos la forma de entrada *srnom.* Al aplicar la regla morfológica

`om::NCMSI`

se genera la pseudoraiz *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. Por lo tanto 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 que contiene todas las formas de las tablas 1 y 2. El sistema de adquisición, ante una forma como, *jelenom* aplicará la regla

`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, si no que es el acusativo y el genitivo singular de *jelen.*

El sistema, pues, ha confundido una forma de un paradigma con el lema de otro paradigma. El sistema añadiría a la base de datos una entrada incorrecta `jelenom:jelena:NCFSI`.

Por lo tanto 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.

La tarea consiste en discernir qué lemas hipotetizados (con su correspondiente categoría gramatical y subcategorización) son correctos. A partir de esta lista de lemas correctos se podrá depurar la base de datos léxica adquirida. Ésta se podría depurar fácilmente a partir de un diccionario en formato electrónico (lista de lemas con información morfológica mínima), pero no hemos considerado oportuno realizar el filtrado de esta manera porque no siempre está disponible este recurso y porque este filtrado eliminaría también lemas correctos pero no contenidos en el diccionario.

El sistema experimental de serbo-croata utiliza un sistema basado 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 3 el número de reglas mínimo para validar un lema de un paradigma. Los resultados conseguidos con esta aproximación, y que exponemos en el siguiente apartado, fueros satisfactorios. No obstante, se evidenciaba la necesidad de idear un sistema de análisis más sofisticado para detectar y solucionar la ambigüedad de las reglas.

Conscientes de esta necesidad, actualmente estamos estudiando la metodología más adecuada para detectar y solucionar la ambigüedad entre reglas. Para ello nos basamos en los siguientes principios:

- dos paradigmas son ambiguos si la terminación de lema de uno de los paradigmas coincide con alguna terminación de forma del otro paradigma, y al menos tengan otra terminación de forma en común
- 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 se produce interferencia entre las reglas porque la terminación de lema del paradigma femenino *a* coincide con la terminación de acusativo y genitivo singular del paradigma masculino. Ahora bien, las formas acabadas en *-o* y *-ama* indican que se trata del paradigma femenino, y la existencia de la forma en *-ima* indica que se trata del paradigma masculino.

Actualmente estamos estudiando dos posibles vías para el análisis de las reglas morfológicas y la discriminación de las adquisiciones correctas:

- algoritmo de análisis de las reglas morfológicas y de creación de reglas de desambiguación de las adquisiciones.
- división de las reglas en inequívocas y ambiguas y adquisición por separado con cada conjunto de reglas.

La elección entre uno de estos métodos se realizará atendiendo a criterios de precisión y cobertura. Esta última nos preocupa especialmente y no descartamos utilizar el contexto para desambiguar los casos más problemáticos.

4.1. Algoritmo de análisis de las reglas morfológicas

Se ha desarrollado un algoritmo que detecta las ambigüedades que se producen entre las reglas de diferentes paradigmas y que propone un conjunto de reglas de desambiguación de las adquisiciones. 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* indican que se trata del paradigma femenino, y la existencia de la forma en *-ima* indica que se trata del paradigma masculino.

Para poder analizar las reglas con esta metodología se han numerado todas las reglas con un código que identifica el paradigma y un número de regla dentro de este paradigma (ver tabla 6). El algoritmo de análisis de reglas utiliza esta numeración para identificar las ambigüedades entre paradigmas y a

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

Cuadro 6: Reglas correspondiente al paradigma de los sustantivos masculinos acabados en consonante y al paradigma de los sustantivos femeninos acabados en *-a*, incluyendo numeración de reglas

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,N2-13,N2-14)*. 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, N2-13, N2-14 validan el paradigma N2)[3]

El procedimiento que se seguirá para realizar la extracción 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 la implementación del sistema 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, N2-13 y N2-14, que se simplifican en ama:a:NCFPD,NCFPL,NCFPI.

Categoría	**Adquiridos**	**Precisión**
Substantivos	11839	88.0 %
Adjetivos	4511	90.9 %
Verbos	3104	89.0 %
TOTAL	19454	88.9 %

Cuadro 7: Resultados obtenidos con el sistema experimental de serbo-croata

4.2. Adquisición por separado con las reglas inequívocas y con las ambiguas

Un segundo procedimiento que se está evaluando actualmente para el análisis de las ambigüedades y la discriminación de las adquisiciones correctas consiste en analizar las reglas con el objeto de dividirlas en dos grupos, uno de reglas inequívocas y otro de reglas ambiguas. Inicialmente se realiza el proceso de adquisición automática utilizando únicamente las reglas inequívocas. Las adquisiciones obtenidas de esta manera serán válidas, por lo que tanto la forma como el lema serán correctos. Posteriormente se realiza el proceso de adquisición automática utilizando las reglas ambiguas. Esta adquisición presentará el problema que algunos de los lemas hipotetizados no serán correctos. Utilizaremos los lemas adquiridos utilizando las reglas inequívocas para validar estas últimas adquisiciones.

5. Evaluación del sistema de adquisición

Por el momento se ha evaluado el sistema de adquisición experimental del serbo-croata. 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 validó manualmente un total de 9100 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 3 reglas morfológicas, o dicho de otra manera, si existen tres formas en el corpus para dicho lema.

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.

Para el 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 serbo-croata son alentadores aunque se debe mejorar el sistema de análisis de ambigüedad entre las reglas y de discriminación de las adquisiciones correctas. Por otro lado, la falta de recursos del serbo-croata 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 utilizan 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 en base al estudio de la ambigüedad de las reglas, de forma 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.

Bibliografía

Alshawi, H., editor. 1992. *The Core Language.* MIT Press.

Badia, A. 1997. The derivation of a large computational lexicon for a two-level morphological analyzer for catalan from a machine readable dictionary. En *SEPLN 97.*

Badia, T., A. Egea, y A. Tuells. 1997. Catmorf: Multi two-levels steps for catalan morphology. En *Demo proceedings 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*, 27(2):153–198.

Koskenniemi, K. 1983. *Two-level morphology: a general computational model for word recognition and production.* Publications n 11. University of Helsinki.

Martí, M.A. 1988. *Processament informàtic del llenguatge natural: un sistema d'anàlisi morfològica per ordinador.* Ph.D. tesis, Departament de Filologia Romànica. Facultat de Filologia de la Universitat de Barcelona.

Martí, M.A., J.J. López, M. Arévalo, M.J. Simón, I. Castellón, M. Civit, G. Vázquez, A. Fernández, L. Padró, L. Márquez, H. Rodriguez, y G. Rigau. 2001. Recursos de ingeniería lingüística. En *Actas del Segundo Taller Internacional de procesamiento computacional del español y tecnologías del lenguaje*, páginas 193–197, Jaén.

Mikheev, A. y L. Liubushkina. 1996. Russian morphology: An engineering approach. *Natural Language Engineering*, 1(3):235–260.

Oliver, A. 2001. Processament morfològic de les llengües eslaves: el serbo-croat. Trabajo para la obtención de la D.E.A. Universitat de Barcelona.

Tadić, M. 1994. *Računalna obradba morfologije hrvatskoga književnog jezika.* Ph.D. tesis, Sveučilište u Zagrebu, Filozofski fakultet. Zagreb.

Zaliznjak, A.A. 1977. *Grammatitxeskii slovar russkogo jazika. Slovoizmenenie.* Russkii jazik. Moskva.