Uso de Internet para aumentar la cobertura de un sistema de adquisición léxica del ruso*

Antoni Oliver
IN3
U. Oberta de Catalunya
aoliverg@uoc.edu

Irene Castellón
Grupo GRIAL
Lingüística General - UB
castel@lingua.fil.ub.es

Lluís Màrquez
TALP Research Center
LSI, UPC
lluism@lsi.upc.es

Resumen: En este artículo presentamos una metodología para la adquisición de recursos léxicos a partir de corpus sin anotar. Esta metodología está demostrando ser de una gran eficacia para lenguas que, como el ruso, presentan una morfología rica y de tipo predominantemente concatenativa. La metodología puede aplicarse tanto a la creación de nuevos recursos léxicos como en la ampliación de recursos léxicos ya existentes. Presentamos asimismo una extensión de la metodología que realiza consultas automáticas a Internet para adquirir aquellas entradas para las cuales no existe suficiente información en nuestro corpus.
Palabras clave: adquisición léxica, morfología computacional, filología eslava, ruso

Abstract: This paper presents a methodology for the automatic acquisition of lexical resources from raw corpora. This methodology has proved to be efficient for those languages that, like Russian, present a rich and mainly concatenative morphology. This method can be applied in the creation of new resources, as well as in the enrichment of existing resources. We also present an extension of the system that uses automatic querying to Internet to acquire those entries for which we have not enough information in our corpus.
Keywords: lexical acquisition, computational morphology, Slavonic philology, Russian

1. Introducción

La implementación de diferentes herramientas de Procesamiento del Lenguaje Natural necesita de una gran cantidad de información léxica. La creación de estos recursos normalmente requiere una gran cantidad de esfuerzo humano. En este artículo presentamos una metodología para la adquisición de recursos léxicos a partir de corpus sin anotar eficaz para lenguas que presenten una morfología rica y de tipo predominantemente concatenativa. La evaluación del sistema se ha realizado para el ruso, que se caracteriza por tener una morfología muy rica (es una lengua declinable con 6 casos: nominativo, genitivo, dativo, acusativo, instrumental y prepositivo).

El primer sistema experimental de adquisición léxica que desarrollamos se aplicó al serbo-croata (Oliver, 2001) y (Oliver, Castellón, y Màrquez, 2002). Esta metodología se basaba en la coaparición de diversas formas del paradigma en el corpus y necesitaba la presencia del lema para validar la adquisición. El principal problema con el que nos encontramos fue la ambigüedad que presentan las reglas de adquisición, que provocaba la confusión entre una forma de un paradigma con el lema de otro paradigma, y que conducía a resultados de precisión reducidos (34,65 % de precisión y 85,25 % de cobertura, F_1=49,27, para el ruso). Para solventar este problema se desarrolló un algoritmo de análisis de las reglas que permitía a priori distinguir las reglas ambiguas de las que no lo son (Oliver, Castellón, y Màrquez, 2003). Aplicando el análisis de reglas antes del proceso de adquisición se obtenían unos resultados de precisión muy aceptables (93,49 %) pero una disminución dramática de la cobertura (38,52 %, F_1=54,55). Los resultados de diversos experimentos nos demostraron que cuanto más pequeño fuese el corpus más alta era la cobertura obtenida. Para evaluar el efecto de tratar el corpus como un conjunto de diversos corpus de menor tamaño realizamos nuevamente los experimentos dividi-

* Esta investigación se ha llevado a cabo con el apoyo de los proyectos INTERLINGUA (Universitat Oberta de Catalunya e IN3 – IR266) y HERMES (TIC 2000–0335–C03–02)

endo la lista de formas de manera alfabética según la primera letra, las dos primeras y las tres primeras. De esta manera el número de reglas ambiguas se reducía notablemente. Se debe recordar que en esta metodología el corpus se transforma en una lista de todas las formas que aparecen en él. Los mejores resultados se obtuvieron para el experimento alfabético con tres letras iniciales (91,93 % de precisión y 67,19 % de cobertura, F_1=77,64), lo que supone una mejora significativa respecto los resultados iniciales.

La metodología que presentamos en este artículo presenta diversas ventajas respecto a la que acabamos de comentar. En primer lugar, no precisa de la presencia del lema para validar la adquisición, es decir, es capaz de dar la información morfosintáctica correcta y el lema asociado de una forma determinada del corpus incluso si su lema no está presente. Nuestro sistema, además, nos ofrece información de todas las posibles interpretaciones de las formas que no han podido ser adquiridas. Esta información será muy útil para intentar desambiguarla utilizando un corpus de mayor tamaño o, como en nuestros experimentos, utilizando Internet como corpus de gran tamaño.

2. *Objetivo del sistema de adquisición*

El objetivo principal del sistema de adquisición es adquirir automáticamente una lista lo más completa posible de formas con su lema e información morfosintáctica asociados. La información morfosintáctica se expresa mediante etiquetas que siguen las recomendaciones de Multext (Véronis y Khouri, 1995; Erjavec, 2001) (por ejemplo, forma: мостом; lema: мост; POS/TAG: NCMSI; todo ello expresado como мостом:мост:NCMSI). En el cuadro 1 y en el cuadro 2 podemos observar la declinación completa, expresadas con la notación explicada, de las palabra masculina мост y de la palabra femenina карта.

3. *Componentes del sistema de adquisición*

El sistema de adquisición necesitará cierto conocimiento lingüístico que estará contenido básicamente en el corpus sin anotar y en el conjunto de reglas morfológicas. No obstante, ya que el sistema se aplicará únicamente a las formas flexivas regulares, precisaremos también unas listas de palabras pertenecientes a clases no flexivas y clases cerradas y una lista de palabras irregulares.

мост:мост:NCMSN,NCMSAI
моста:мост:NCMSG,NCMSAA
мосту:мост:NCMSD
мостом:мост:NCMSI
мосте:мост:NCMSP
мосты:мост:NCMPN,NCMPAI
мостов:мост:NCMPG,NCMPAA
мостам:мост:NCMPD
мостами:мост:NCMPI
мостах:мост:NCMPP

Cuadro 1: Declinación del sustantivo masculino de tipo A мост

карта:карта:NCFSN
карты:карта:NCFSG,NCFPN,NCFPAI
карте:карта:NCFSD,NCFSP
карту:карта:NCFSA
картой:карта:NCFSI
карт:карта:NCFPG,NCFPAA
картам:карта:NCFPD
картами:карта:NCFPI
картах:карта:NCFPP

Cuadro 2: Declinación del sustantivo femenino de tipo A карта

3.1. Listas de clases no flexivas y de clases cerradas

Las palabras que pertenecen a categorías cerradas o no flexivas se excluyen del proceso de adquisición. Se han construido manualmente listas de palabras de dichas categorías. 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.

3.2. 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. Esta lista está actualmente bajo desarrollo. El criterio que adoptamos es escribir todas las formas de palabras irregulares incluidas en la lista de las 5.000 palabras rusas más frecuentes según (Sharov, 2001).

El concepto de irregularidad está relacionado con la no adscripción de un lema a un paradigma determinado. Por lo tanto la regu-

laridad 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. Casos extremos lo constituyen el analizador morfológico del ruso de Mikheev (Mikheev y Liubushkina, 1995) o del croata de Tadić (Tadić, 1994), que 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, y aunque este hecho no supone ningún problema para el análisis o la generación de formas, sí que resulta problemático para la adquisición automática de información léxica.

3.3. Reglas morfológicas

Las reglas morfológicas se han implementado siguiendo un formalismo de descomposición morfológica (Alshawi, 1992). En tiempo de ejecución estas reglas morfológicas se convierten en expresiones regulares de Perl. Las reglas tienen la forma `TF:TL:Desc`, donde: `TF` significa terminación de forma, `TL` significa terminación de lema y `Desc` significa descripción morfosintáctica. Por ejemplo, la regla:

```
ом::NCMSI
```

puede expresar la entrada:

```
мостом:мост:NCMSI
```

La regla del ejemplo tiene una terminación de lema nula. El uso de expresiones regulares de Perl nos permite describir la terminación de lema con una mayor precisión, como por ejemplo en la siguiente regla:

```
([^аяеэиыоуюькгхжшщчцй])ом:\1:NCMSI
```

donde '`^`' significa el conjunto de caracteres complementario del expresado entre corchetes y '`\1`' es una variable que contiene el carácter que satisface la expresión regular indicada entre paréntesis. Las expresiones regulares nos permiten describir fenómenos morfológicos complejos con facilidad, como por ejemplo la alternancia vocálica. Así la regla:

```
ль([^аяеэиыоуюькгхжшщчцй])а:ле\1:NCMSG
```

puede expresar una entrada del tipo

```
льва:лев:NCMSG.
```

Las reglas se han desarrollado siguiendo los modelos más productivos de (Zaliznjak, 1977). Concretamente se han desarrollado 565 reglas correspondientes a sustantivos, 219 a adjetivos y 12.038 a verbos. El elevado número de reglas correspondientes a los verbos se debe al hecho que ciertas formas, como los participios, son declinables. Las reglas correspondientes a formas declinadas se han derivado automáticamente a partir de las reglas que expresan la forma base.

En el proceso de adquisición no se utilizan todas las reglas. Aquellas reglas que expresan terminaciones alternativas que coinciden con otras terminaciones del mismo paradigma se excluyen del proceso de adquisición. Por ejemplo, la regla:

```
([^аяеэиыоуюькгхжшщчцй])а:\1:NCMPN
```

que expresa el nominativo plural alternativo en '`а`' para los sustantivos masculinos se excluye porque la `TF` '`а`' es igual a la del genitivo singular. En las reglas también se ha eliminado la información aspectual de los verbos.

En tiempo de ejecución las reglas se transforman en substituciones de Perl, por ejemplo:

```
([^кгхжшщчц])у:\1а:NCFSA
```

se transforma en la substitución

```
s/([^кгхжшщчц])у$/\1а/
```

Esta expresión significa "cambia la `у` final si la letra precedente no es una de la lista `{к,г,х,ж,ш,щ,ч,ц}` por una `а`". Esta substitución permite la formación del lema de un sustantivo femenino a partir de su acusativo singular.

3.4. Corpus

Se ha compilado un corpus del ruso de 16.000.000 de palabras a partir de diarios, revistas y textos literarios[1]. El corpus se ha segmentado automáticamente en frases y no se ha añadido ningún otro tipo de información lingüística.

4. Metodología de adquisición

La metodología de adquisición se puede dividir en siete pasos. Los tres primeros pasos (expansión, filtrado y reorganización de las

[1] Agradecemos a Огонек, Правда and Библиотека Максима Мошкова el permiso desinteresado para utilizar sus textos en la compilación del corpus

reglas) están relacionados con la adaptación de la notación de las reglas al proceso de adquisición. El cuarto paso se refiere al tratamiento de las formas del corpus. El quinto paso es la adquisición propiamente dicha, mientras que los dos últimos pasos están relacionados con la generación del fichero de dudas y su resolución mediante consultas a Internet.

4.1. Expansión de las reglas

Las reglas que expresan contextos morfológicos se expanden en todas sus posibilidades. Por ejemplo la regla:

```
([^аяеэиыоуюькгхжшщчцй])ом:\1:NCMSI
```

se expande a todas las reglas siguientes:

```
ъом:ъ:NCMSI  фом:ф:NCMSI  том:т:NCMSI
сом:с:NCMSI  ром:р:NCMSI  пом:п:NCMSI
ном:н:NCMSI  мом:м:NCMSI  лом:л:NCMSI
зом:з:NCMSI  дом:д:NCMSI  вом:в:NCMSI
бом:б:NCMSI
```

4.2. Filtrado de las reglas

La expansión de las reglas morfológicas conduce a una gran multiplicación del número de reglas. Al expandir nuestro conjunto de 9821 reglas obtenemos 74.956 reglas. Algunas de estas reglas no son aplicables ya que la combinación de letras expresada por la regla no existe en la lengua (o al menos no está presente en nuestro corpus). Después de la expansión de las reglas estas se filtran con el conjunto de n–gramas finales calculados con las palabras de nuestro corpus. Mediante este procedimiento se eliminan una gran cantidad de reglas que no son aplicables, con lo que se agiliza el resto del proceso. Concretamente, nuestro conjunto de reglas expandidas una vez filtrada, se reduce a 22.033 reglas.

Siguiendo el ejemplo anterior, en el proceso de filtrado se eliminaría la regla `ъом:ъ:NCMSI` ya que la terminación `ъом` no existe en nuestro corpus. De hecho, no importa si la combinación es posible o no en la lengua, ya que es suficiente que no exista en el corpus para que la regla no sea aplicable en nuestro proceso de adquisición. Ahora bien, en caso de modificar el corpus, se debería repetir el proceso de filtrado.

4.3. Reorganización de las reglas

Una vez expandidas y filtradas las reglas morfológicas se reorganizan realizando una lista de terminaciones que comparten la misma categoría gramatical y modelo flexivo, así como la terminación de lema. Estos grupos se guardan indexados por el grupo flexivo y la terminación de lema, como en el siguiente ejemplo:

```
NCMA:т    т,та,ту,том,те,ты,тов,там,тами,тах
NCFA:та   та,ты,те,ту,той,т,там,тами,тах
```

4.4. Agrupación por posibles paradigmas

Las formas presentes en el corpus se intentan dividir en raíz y terminación por todas las terminaciones posibles y se agrupan por paradigmas. Para explicar este paso consideremos que nuestro corpus está formado por las formas: мост, моста, мостом, мостах, карта, карты, картой y картах. Si realizamos la división por todas las posibles terminaciones y realizamos la agrupación, obtendremos el siguiente resultado:

```
NCMA:мост    мост,моста,мостом,мостах
NCFA:моста   моста,мост,мостах
NCMA:карт    карта,карты,картах
NCFA:карта   карта,карты,картой,картах
```

4.5. Adquisición mediante comparación

Una vez que hemos realizado la agrupación podemos iniciar el proceso de adquisición. Para ello tomamos cada una de las formas de la lista de formas presentes en el corpus y observamos en qué divisiones se puede clasificar y en caso que sea en más de una escogemos la división que posee más formas. Por ejemplo, si tomamos la forma мостах observamos que la podemos asociar al grupo NCMA:мост o bien al grupo NCFA:моста, pero que el primero de ellos tiene cuatro formas asociadas y en cambio la segunda opción solo tiene tres, por lo que escogemos la primera opción. Con esta información podemos relacionar la forma мостах con el lema мост. A partir de toda esta información podemos recuperar la información morfosintáctica asociada de las reglas sin agrupar y crear la nueva entrada мостах:мост:NCMPP. De la misma manera si consideramos la forma карты podemos asociarla al grupo NCMA:карт, con tres formas asociadas, o bien al grupo NCFA:карта, con cuatro formas asociadas. Esta última será la interpretación correcta.

Es interesante observar que no es necesario que la forma correspondiente al lema esté presente en el corpus. Incluso si no está presente

la adquisición se realizará correctamente. Por ejemplo, si ahora nuestro corpus está formado por las formas: моста, мостом, мостах, карта, карты, картой y картах, es decir que no está presente el lema мост, la agrupación de posibles paradigmas se realizará según se muestra a continuación:

NCMA:мост	моста,мостом,мостах
NCFA:моста	моста,мостах

Si tomamos nuevamente la forma мостах observamos que la podemos asociar igualmente al grupo NCMA:мост o bien al grupo NCFA:моста, pero que el primero de ellos tiene ahora tres formas asociadas y la segunda opción únicamente dos, por lo que continuaremos escogiendo la primera opción. Con esta información podemos relacionar la forma мостах con el lema мост aunque este no esté presente en el corpus.

4.6. Generación del fichero de dudas

En el ejemplo que hemos planteado todo ha funcionado correctamente porque en el corpus considerado estaban presentes las formas cuyas terminaciones no eran comunes entre ambos paradigmas (мостом y картой). Supongamos ahora que nuestro corpus está formado por las formas: мост, моста, мосту, мостах, карта, карты, карту y картах, es decir que ya no están presentes las formas discriminantes мостом y картой. Al realizar la división y la agrupación obtendríamos los resultados que se pueden observar a continuación:

NCMA:мост	мост,моста,мосту,мостах
NCFA:моста	моста,мост,мосту,мостах
NCMA:карт	карта,карты,карту,картах
NCFA:карта	карта,карты,карту,картах

Si intentamos adquirir el lema y la información morfosintáctica asociados a la forma мостах tenemos dos posibles opciones, NCMA:мост y NCFA:моста, y ambas con cuatro formas asociadas, por lo que es imposible determinar la opción correcta. Lo mismo sucede si intentamos deducir la información asociada por ejemplo a la palabra карты ya que tiene las opciones NCMA:карт y NCFA:карта con el mismo número de formas asociadas.

En este caso nuestro sistema genera un fichero de dudas. En este fichero se especifican las posibles opciones para cada palabra de entrada no resuelta. En los ejemplos presentados el fichero de dudas nos presentaría las siguientes opciones:

forma: мосту	forma: карты
opción 1: мос,NCMA,т	opción 1:кар,NCMA,т
opción 2: мос,NCFA,та	opción 2:кар,NCFA,та

Es decir, para cada una de las opciones el fichero nos da la raíz asociada, el grupo flexivo y la terminación de lema. Este fichero de dudas es interesante ya que se podrá utilizar en etapas posteriores para intentar determinar la interpretación correcta.

4.7. Resolución de las dudas mediante consulta a Internet

Los casos no resueltos por el sistema de adquisición se deben a la falta de algunas formas del paradigma en nuestro corpus. Si aumentásemos suficientemente el tamaño del corpus es de suponer que las formas que nos faltan para discriminar la entradas dudosas aparecerían en el corpus. No es posible determinar a priori el tamaño deseable para poder encontrar todas las formas necesarias de manera que no nos quede ninguna forma regular sin adquirir. El aumento del tamaño del corpus nos llevaría a encontrar nuevas formas que no presentan suficientes formas asociadas para ser adquiridas y para las cuales necesitaríamos nuevamente aumentar el tamaño del corpus. Por este motivo hemos decidido utilizar Internet como corpus de gran tamaño. Para la utilización de Internet como corpus utilizaremos buscadores de Internet para verificar la existencia de las formas necesarias que nos ayuden a discriminar entre las diferentes opciones. Para determinar las formas discriminantes se flexionan las palabras según los modelos dados por las opciones y las formas discriminantes serán las que estén presentes en un modelo pero no en el resto. Siguiendo el ejemplo anterior, para discriminar las opciones de la forma мостах generaríamos todas las formas correspondientes a la opción мос:NCMA:т (мост, моста, мосту, мостом, мосте, мосты, мостов, мостам, мостами, мостах) y todas las formas correspondientes a la opción мос:NCFA:та (моста, мосты, мосте, мосту, мостой, мост, мостам, мостами, мостах). Las consultas las generaremos a partir de las formas no comunes entre las diversas opciones, es decir, en nuestro ejemplo generaríamos las consultas a partir de las formas 'мостом' y 'мостов' correspondientes a la

primera opción y de la forma 'мостой' correspondiente a la segunda. Las consultas a Internet correspondientes a opciones que supongan más de una forma se lanzan en una sola consulta mediante el operador OR (que representaremos como "||"). En nuestro ejemplo lanzaremos la consulta мостом||мостов y la consulta мостой. Lanzando la primera consulta al buscador, este devolverá un número mayor de documentos encontrados que al lanzar la segunda, que es una forma no existente en la lengua. De este modo se validará la opción мос:NCMA:т. De la misma manera, para discriminar entre las opciones de la palabra карты generaremos las consultas картом||картов y картой, y en este caso el número de documentos devueltos será mucho mayor en el segundo caso, validando la opción кар:NCFA:та. Otro aspecto que se debe tener en cuenta es que en el caso de los verbos, para los que existen una gran cantidad de formas por lema, se ha limitado el número de formas a consultar, ya que de los contrario se generaban consultas muy complejas que provocaban errores en el buscador.

5. *Evaluación experimental*

Se ha constituido un subconjunto del corpus compuesto únicamente por formas conocidas y regulares. Para ello hemos seleccionado todas las formas presentes en el corpus que también lo están en un formario generado previamente a partir de un diccionario morfológico (Zaliznjak, 1977) y las reglas desarrolladas. De esta manera hemos conseguido obtener una lista de formas regulares pero con una distribución de formas y lemas igual que en el corpus. El resultado es una lista de 232.543 formas correspondientes a 43.543 lemas. El proceso de adquisición se realizará a partir de esta lista de formas.

Para solucionar el fichero de dudas utilizaremos dos buscadores de Internet: *Yahoo*[2] y *Yandex*[3]

Adicionalmente hemos realizado una simulación de consulta a Internet utilizando un analizador morfológico del ruso. Concretamente hemos utilizado el analizador *mystem* de la empresa rusa Yandex[4]. La simulación de consulta consiste en enviar al analizador morfológico cada una de las formas que conforman una consulta generada por el sistema. Asignamos un número arbitrario, por ejemplo 1000, si el analizador ha sido capaz de analizar la forma y un número menor, por ejemplo 0, si no ha sido capaz. De este modo el comportamiento es similar al de una consulta a Internet. Hemos realizado esta simulación por dos motivos principales. En primer lugar porque el proceso de consultas a Internet es lento, ya que se deben realizar numerosas consultas y existe un número máximo de consultas diarias. Y en segundo lugar porque nos interesa verificar si la consulta libre a Internet nos ofrece resultados similares a los obtenidos utilizando herramientas existentes.

Debido a la limitación del número de consultas diarias a Internet no podemos ofrecer resultados completos para los dos buscadores. Presentaremos los resultados completos para la simulación mediante el analizador morfológico y para el buscador Yahoo. Para el buscador Yandex, nos limitaremos a ofrecer los resultados para subconjuntos del corpus, concretamente para las palabras que empiezan por las letras а, б, в y г.

5.1. Resultados

En este apartado presentaremos los diferentes resultados obtenidos en nuestros experimentos. Los parámetros que ofreceremos son la precisión (P), cobertura (C), i la media armónica entre precisión y cobertura ($F_1 = 2PC/(P+C)$).

En el cuadro 3 podemos observar los resultados obtenidos en el tratamiento de todo el corpus. En la fila *Adqui.* se presentan los resultados de la adquisición sin resolución de dudas. En la fila *Simul.* se presentan los resultados de la adquisición más la desambiguación de dudas mediante simulación de consulta con el analizador morfológico. Finalmente las filas etiquetadas con los diferentes valores de n muestran los resultados obtenidos mediante consulta a Internet utilizando el buscador Yahoo para diferentes valores de n. El parámetro n hace referencia al número mínimo de documentos que debe devolver el buscador para que se considere la consulta como válida. Mediante este parámetro podemos evitar dar como válidas consultas con un número bajo de resultados, ya que se pueden tratar de documentos que presenten faltas de ortografía o palabras erróneas. Este parámetro nos servirá

[2] http://www.yahoo.com

[3] htpp://yandex.ru

[4] Agradecemos a Yandex la posibilidad de utilizar su analizador morfológico en nuestras investigaciones.

	GLOBAL			N			A			V		
	P	C	F_1	P	C	F_1	P	C	F_1	P	C	F_1
Adqui.	95,53	62,32	75,43	97,76	50,92	66,96	99,46	64,82	78,49	90,49	81,57	85,8
Simul.	94,47	78,26	**85,6**	97,35	67,27	**79,59**	99,07	83,72	**90,75**	87,18	93,05	**90,02**
n=1	89,62	78,80	83,86	90,65	69,30	78,55	94,99	82,10	**88,08**	83,87	93,34	88,35
n=10	92,02	77,47	**84,12**	94,58	68,17	**79,23**	96,54	79,85	87,41	85,25	92,74	**88,84**
n=100	92,78	75,42	83,53	95,30	66,10	78,06	97,61	77,31	86,28	85,94	91,37	88,57
n=1000	93,42	68,84	79,27	95,85	58,76	72,86	98,69	70,30	82,11	86,77	86,79	86,78

Cuadro 3: Resultados obtenidos con todo el corpus mediante simulación y el buscador Yahoo

	n=1			n=10			n=100			n=1000		
	P	C	F_1	P	C	F_1	P	C	F_1	P	C	F_1
а adqui.	99,22	70,07	82,14									
а simul.	98,51	89,24	**93,65**									
а yahoo	97,87	90,18	**93,87**	98,62	89,40	93,78	98,75	86,89	92,44	98,84	79,03	87,83
а yandex	97,84	90,21	93,87	98,44	89,87	**93,96**	98,68	87,59	92,80	98,78	80,40	88,65
б adqui.	94,87	61,84	74,87									
б simul.	94,78	78,35	**85,79**									
б yahoo	90,65	79,70	84,82	92,45	78,64	**84,99**	92,90	75,70	83,42	93,53	68,84	79,31
б yandex	90,13	79,48	84,47	91,49	79,06	**84,82**	92,45	76,49	83,72	93,24	70,19	80,09
в adqui.	95,42	69,07	80,13									
в simul,	95,14	81,80	**87,97**									
в yahoo	91,39	82,35	86,63	93,53	81,01	**86,82**	94,27	79,86	86,47	94,89	75,46	84,07
в yandex	90,84	82,47	86,45	92,63	81,76	**86,86**	93,96	80,27	86,58	94,65	76,07	84,35
г adqui.	94,55	59,37	72,94									
г simul.	94,22	78,60	**85,70**									
г yahoo	89,38	72,93	80,32	91,67	74,95	**82,47**	92,92	73,48	82,06	94,26	67,88	78,92
г yandex	89,19	79,44	84,03	91,06	78,74	**84,45**	92,5	76,73	83,88	93,71	70,32	80,35

Cuadro 4: Resultados parciales de la adquisición y de la resolución de dudas mediante simulación y dos buscadores: Yahoo y Yandex

para ajustar el compromiso entre precisión y cobertura.

Los resultados obtenidos mediante el proceso de adquisición sin resolución de dudas son mucho mejores que los obtenidos con el método anterior (Oliver, Castellón, y Màrquez, 2003) sin división en subcorpus (F_1 de 75,43 frente a 54,55). Respecto al mejor de los resultados obtenidos en el método anterior, que era el correspondiente a un proceso con división en subcorpus de forma alfabética con tres letras iniciales, los resultados son ligeramente inferiores pero muy comparables (F_1 de 75,43 frente a 77,64). La ventaja del nuevo método es la velocidad de proceso mucho mayor y la posibilidad de solucionar el fichero de dudas. Mediante la simulación de consulta se obtiene una F_1 de 85,6 lo que supone una mejora de 10 puntos sobre el resultado de la adquisición, con una aumento de 16 puntos en la cobertura y una disminución de sólo un punto en la precisión. Mediante las consultas al buscador Yahoo obtenemos resultados muy similares ($F_1 = 84{,}12$ para $n = 10$).

En el cuadro 4 podemos observar una comparación de los resultados parciales para las letras а, б, в y г. En este caso presentamos los resultados globales y no para cada categoría gramatical, aunque las tendencias se mantienen para las diferentes categorías. Comparando los resultados de la desambiguación simulada con los de la consulta a Internet, observamos que, en general y para $n = 1$, se obtienen mejores resultados de precisión con el proceso simulado, pero mejores resultados de cobertura mediante las consultas a Internet. También se puede observar que los resultados obtenidos para los dos buscadores son muy similares para todos los valores de n. Asimismo se evidencia que para cualquier n los valores de F_1 obtenidos mediante consulta a Internet son significativamente superiores a los obtenidos únicamente

mediante el proceso de adquisición. De los resultados también se observa que en general para $n = 10$ se obtinenen los mejores resultados. En los cuadros 3 y 4 se han marcado en negrilla los mejores resultados de F_1.

6. Conclusiones y líneas futuras

En este artículo hemos presentado una metodología de adquisición léxica a partir de corpus sin anotar que utiliza búsquedas a Internet para adquirir aquellas entradas para las cuales no existe suficiente información en el corpus.

Los resultados obtenidos son notables y suponen una mejora sustancial respecto sistemas desarrollados anteriormente. El sistema de adquisición se ha utilizado para la lengua rusa, que se caracteriza por poseer una morfología muy rica. En el futuro se pretende probar esta metodología para otras lenguas como el croata, el castellano y el catalán.

La metodología puede ser de gran utilidad para la creación de nuevos recursos léxicos así como para la ampliación de recursos existentes. La gran ventaja que ofrece es que, mediante consultas a Internet, podemos evitar la compilación de corpus de gran tamaño.

En estos experimentos se han utilizado un conjunto de reglas desarrolladas manualmente a partir de gramáticas tradicionales. Actualmente, estamos desarrollando un sistema de asistencia a la creación de las reglas a partir de corpus sin anotar, similar al presentado en (Goldsmith, 2001). Este sistema será de gran utilidad para desarrollar las reglas morfológicas flexivas más productivas y para extender el conjunto de reglas a los procesos derivativos más frecuentes.

Bibliografía

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

Erjavec, T. 2001. Specifications and notation for multext-east lexicon encoding. Informe técnico, Multext-East/Concede.(http://nl.ijs.si/MTE/V2).

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Mikheev, A. y L. Liubushkina. 1995. Russian morphology: An engineering approach. *Natural Language Engineering*, 1(3):235–260.

Oliver, A. 2001. Processament morfològic de les llengües eslaves: el serbo-croat. Master's thesis, Treball per a l'obtenció de la D.E.A: Universitat de Barcelona.

Oliver, A., I. Castellón, y L. Màrquez. 2002. Adquisición automática de información léxica y morfosintáctica a partir de corpus sin anotar: aplicación al serbo-croata y ruso. *Procesamiento del Lenguaje Natural*, 29:97–104.

Oliver, A., I. Castellón, y L. Màrquez. 2003. Automatic lexical acquisition from raw corpora: An application to russian. En *Workshop on Morphological Processing of Slavic Languages*, páginas 17–24. Association for Computational Linguistics. 10th Conference of The European Chapter of the EACL.

Sharov, S.A. 2001. Chastotnyi slovar. www.artint.ru/projects/frqlist.asp.

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

Véronis, J. y L. Khouri. 1995. Etiquetage grammatical multilingue: modèle. Informe técnico, MULTEXT Project, http://www.lpl.univ-aix.fr/projects/multext/LEX/LEX2.html.

Zaliznjak, A.A. 1977. *Grammaticheskii slovar russkogo jazyka. Slovoizmenenie*. Izdatelstvo Russkii jazyk. Moskva.