Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 335-343 recibido 1-05-2009; aceptado 5-06-2009

Construcción y extensión de un léxico morfológico y sintáctico para el español: el Le*ff*e *

Building and extending a morphological and syntactic lexicon for Spanish: the Le*ff*e

Miguel A. Molinero
Grupo LYS
Universidade da Coruña
A Coruña, Spain
mmolinero@udc.es

Benoît Sagot
Project ALPAGE
INRIA
Paris, France
benoit.sagot@inria.fr

Lionel Nicolas
Laboratoire I3S (Equipe RL)
Université de Nice-Sophia Antipolis
Sophia Antipolis, France
lnicolas@i3s.unice.fr

Resumen: Un léxico con información morfológica y sintáctica de amplia cobertura puede ser construido de forma eficiente reutilizando otros recursos existentes y mejorado usando técnicas semi-automáticas para detectar y corregir sus errores. Presentamos aquí un léxico español creado siguiendo esta estrategia: el Le*ff*e
Palabras clave: Léxico morfológico y sintáctico, recursos lingüísticos

Abstract: A morphological and syntactic wide coverage lexicon can be developed by using other existing resources and improved by using semi-automatic techniques which enables errors to be detected and fixed. We present here a Spanish lexicon developed using such an approach: the Le*ff*e
Keywords: Morphological and syntactic lexicon, linguistic resources

1. Introducción

Los recursos lingüísticos, como léxicos y gramáticas, son necesarios para construir muchas aplicaciones de Procesamiento del Lenguaje Natural (PLN) de alto nivel. Por ejemplo aquellas que requieren análisis sintáctico profundo para tareas como traducción automática, buscadores avanzados, etc. La situación actual para muchas lenguas es que existen varios de estos recursos, con diferentes niveles de cobertura, diferentes modelos lingüísticos y escritos en diferentes formalismos lexicales. Sin embargo, ninguno de ellos combina de un modo satisfactorio una amplia cobertura (incluyendo un gran número de palabras en todas sus categorías), alta calidad (ausencia de errores) y riqueza de la información (morfológica y sintáctica).

Aun así, los recursos existentes para una lengua contienen información valiosa que puede ser reutilizada. Por lo tanto, fusionar esos recursos y ampliarlos usando técnicas semi-automáticas es una interesante forma de construir nuevos recursos o incluso mejorar otros existentes. Para ello es necesario ser capaz de interpretar la información contenida en los recursos a pesar de que sus formalismos sean parcialmente incompatibles, convertirlos a un formato común, y por último fusionarlos en un recurso único. Ninguno de estos pasos es trivial.

En este artículo confirmamos la validez de esta aproximación aplicándola al español. Hemos construido un léxico morfológico y sintáctico de amplia cobertura y libre (bajo licencia LGPL-LR[1], el Le*ff*e (*Léxico de formas flexionadas del Español*). Este léxico puede ser usado directamente en aplicaciones de PLN de alto nivel, especialmente en aquellas que requieren un análisis sintáctico profundo. El Le*ff*e está desarrollado usando Alexina (Sagot et al., 2006; Sagot y Danlos, 2008; Danlos y Sagot, 2008), un formalismo lexical previamente usado en el desarrollo del Le*fff* (*Lexique des formes fléchies du Français - Léxico de formas flexionadas del francés*).

* Parcialmente financiado por el Ministerio de Educación y Ciencia (HUM2007-66607-C04-02), la Xunta de Galicia (INCITE08PXIB302179PR, INCITE08E1R104022ES, PGIDIT07SIN005206PR) y la 'Red Gallega para el procesamiento del lenguaje y la recuperación de información' 2006-2009

[1] Lesser General Public License for Linguistic Resources, http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/lgpllr.html, june 2009

ISSN: 1135-5948

La flexibilidad y la calidad de Alexina permiten su uso directo con varios formalismos gramaticales (LFG,LTAG, etc.) que requieren información sintáctica detallada para todas las palabras.

El trabajo descrito aquí se enmarca dentro del Proyecto Victoria[2], que tiene como objetivo el desarrollo de técnicas y herramientas para la adquisición y corrección eficiente de recursos lingüísticos de amplia cobertura. El hecho de desarrollar recursos de esta manera y usando el mismo formato los hace especialmente apropiados para conexiones multilingües y todas sus aplicaciones derivadas. En su primera fase, se centra en español, gallego y francés.

Este artículo está organizado como sigue: primero en la sección 2, presentamos el formalismo lexical Alexina. La sección 3 describe otros recursos lingüísticos para el español que hemos usado. En la sección 4 mostramos cómo esos recursos fueron fusionados, y en la sección 5 describimos una técnica para extender un léxico. Por último en la sección 6 evaluamos el léxico obtenido en nuestros experimentos, y presentamos nuestras conclusiones en la sección 7.

2. El formalismo lexical: Alexina

Podemos definir un léxico como una lista exhaustiva de las palabras que componen una lengua acompañadas de cierta información morfológica y/o sintáctica. Para desarrollar tareas de PLN de alto nivel, como análisis sintáctico profundo, es necesario disponer de léxicos que efectivamente describan el comportamiento sintáctico de sus entradas.

Alexina es un modelo que permite describir información morfológica y sintáctica de manera fácilmente legible, completa y eficiente (Sagot, 2005; Danlos y Sagot, 2008). Es capaz de representar un gran número de fenómenos a través de un formato sencillo que puede ser usado por herramientas de PLN que se basen en diferentes formalismos gramaticales. Los conceptos lingüísticos en los que se basa Alexina son compatibles con el estándar Lexical Markup Framework.[3]

El formato de Alexina ha evolucionado durante los últimos cinco años junto al Le*fff* y otros recursos para otras lenguas (polaco, eslovaco y otros). Es por tanto un formato capaz de representar una gran variedad de fenómenos lingüísticos. Esto, unido a la proximidad lingüística entre el francés y el español como Lenguas Romances que son, ha permitido describir el español sin tener que modificar el formato original de Alexina.

Alexina está basado en dos niveles de representación. Un *nivel intensional* que factoriza la información léxica, de modo que a cada lema se le asocia una clase morfológica (que permite construir toda la familia de formas asociada a dicho lema) e informaciónes sintácticas detalladas (marco de subcategorización, posibles restructuraciones, atributos, etc.), permitiendo una gestión más rápida y sencilla. Y un *nivel extensional*, que se genera automáticamente *compilando* el léxico intensional, en el que se asocia cada forma flexionada con toda su información morfológica y sintáctica: etiqueta morfológica, el marco de subcategorización de su correspondiente redistribución, etc.

Cuando el léxico intensional es compilado en un léxico extensional, se construyen todas las palabras pertenecientes a la familia de cada lema, usando para ello su clase morfológica. Las clases morfológicas están definidas bajo un formato descrito en (Sagot, 2005) que cubre la mayor parte de las entradas de léxico. Tan solo los lemas que se flexionan de una forma especial (irregular) son descritos de forma manual en un fichero adicional.

Por ejemplo, esta es la entrada intensional simplificada[4] en Le*fff*e para el lema *destacar*$_1$ en el sentido de *resaltar algo*:

```
destacar_1
        V4
        Lemma;v;
        <arg_0:Suj:cln|scompl|sinf|sn,
        arg_1:Obj:cla|scompl|sn>;
         %actif,%passif,%ppp_employé_comme_adj
```

Se trata de un verbo transitivo cuya clase morfológica es `V4`, que se corresponde con verbos de la primera conjugación que cambian su raíz (se cambia la *c* por *qu*) al formar el presente de subjuntivo. Su predicado semántico se representa directamente con el lema. Su categoría es *verbo* (`v`). Tiene dos argumentos canónicamente realizados por las funciones sintácticas `Suj` (sujeto) y `Obj` (objeto directo).[5] Cada función sintácti-

[2] `http://www.victoria-project.org` (Abril 2009).

[3] Lexical Markup Framework, el estándar ISO/TC37 para léxicos usados en PLN.

[4] Se han eliminado algunas informaciones sintácticas por motivos de claridad

[5] Las posibles funciones sintácticas usadas en

ca está asociada a una lista de posibles realizaciones[6] que aparecen entre paréntesis para indicar que la realización de la función es opcional. Esta entrada permite además tres redistribuciones: activa (`%actif`), pasiva (`%passif`) y participio empleado como adjetivo (`%ppp_employé_comme_adj`).

El proceso de compilación construye una entrada extensional para cada una de las formas flexionadas del lema y cada redistribución compatible mediante definiciones formales de esas redistribuciones. Por ejemplo, la única forma flexionada del verbo *destacar* que es compatible con la redistribución pasiva es el participio. La redistribución `%ppp_employé_comme_adj` indica que el participio de este verbo puede ser usado como adjetivo y provocará la generación de la correspondiente entrada adjetiva extensional.

La entrada extensional (simplificada) correspondiente a la redistribución pasiva para `destacar` es la siguiente (donde `MP00SM` es la etiqueta para el participio sigular masculino y en el que se crea un complemento agente opcional derivado de la transitividad del verbo):

```
destacado    v
[pred='destacar_1<arg_1:Suj:cln|scompl|sn,
arg_0:Obl2:(por-sn|por-scompl)>',
@passive,@pers,@MP00SM];
%passif
```

3. Recursos lingüísticos del español

Hoy en día es posible encontrar varios recursos lingüísticos disponibles para el español. Sin embargo, ninguno de ellos cumple con nuestros requisitos:

- Amplia cobertura, buena calidad y riqueza de los datos (morfológicos y sintácticos).
- Separación completa entre las informaciones lexicales y gramáticales.
- Formato claro, compacto y legible para los humanos.
- Disponible libremente en términos de acceso, modificación y distribución.
- Fácilmente enlazable con recursos en otros idiomas.

Aun así, hay mucha información de gran valor en los recursos existentes y sería un error ignorarla. En nuestro caso, los siguientes recursos han sido usados en algún momento para el desarrollo del Le*ff*e:

Multext es un proyecto internacional (Ide y Véronis, 1994) cuyos objetivos son desarrollar estándares y especificaciones para representar y procesar copus lingüísticos, así como desarrollar herramientas y recursos lingüísticos de acuerdo a dichos estándares. Como resultado han construido un conjunto de léxicos que contiene información morfológica para varios idiomas, incluido el español, usando un juego de etiquetas ampliamente extendido ya en la comundidad del PLN.

El léxico de la USC contiene información morfológica de un gran número de formas. Fue creado para realizar tareas de etiquetación en el Departamento de Gramática española de la Universidad de Santiago de Compostela (Alvarez et al., 1998);

ADESSE es una versión ampliada de la Base de Datos Sintácticos del español actual desarrollada en la Universidad de Vigo (García-Miguel y Albertuz, 2005). Es un trabajo de gran extensión que incluye los marcos de subcategorización de más 4000 verbos.

La *Spanish Resource Grammar* (SRG) es una gramática *open-source* del español (Marimon, Seghezzi, y Bel, 2007) basada en el marco teórico *Head-driven Phrase Structure Grammar* (HPSG) que incluye un léxico con información sintáctica organizada en una jerarquía de clases.

4. Reutilización de recursos existentes

Construir el Le*ff*e ha supuesto interpretar todos los recursos mencionados en la sección anterior (a pesar de que sus modelos

Le*ff*e son las siguientes: `Suj` (sujeto), `Obj` (objeto directo), `Objde` (Objeto indirecto introducido por la preposición *de*), `Objа` (objeto indirecto introducido por *a*), `Loc` (locativo), `Dloc` (delocativo), `Att` (atributo), `Obl` and `Obl2` (oblicuos).

[6] Pueden ser realizaciones clíticas `cln`, `cla` y `cld` para los casos nominativo, acusativo y dativo. Las realizaciones directas son *sn*, *sinf*, *scompl*, *sa* y *qcompl* para los sintagmas nominal, infinitivo, completivo, adjetival y preguntas indirectas. Las realicaciones preposicionales se construyen de la forma *prep-real*, donde *prep* es una preposición y *real* una realización directa (Ej. *con-sn*)

léxicos eran parcialmente incompatibles), convertirlos en el formato de Alexina y finalmente fusionar todos los léxicos en uno solo. Multext y el léxico de la USC contienen solamente información morfológica, mientras que SRG y ADESSE incluyen información sintáctica. Por tanto, decidimos proceder de la siguiente manera:

1. Construir un lexico morfológico inicial tomando como base Multext y transformándolo en el formato Alexina. Se le añadieron además algunas entradas específicas del modelo Alexina (prefijos, sufijos, signos de puntuación, etc.);
2. Convertir el léxico de la USC al formato de Alexina y fusionarlo con el léxico inicial extraído de Multext. De este modo obtenemos un léxico que contiene la información morfológica de Le*ff*e;
3. Convertir la información sintáctica de ADESSE y del léxico SRGen el formato de Alexina;
4. Fusionar el Le*ff*e morfológico del paso 2 y los dos léxicos sintácticos obtenidos en el paso 3;

4.1. Conversión de léxicos morfológicos al formato Alexina

Un léxico morfológico puede representarse mediante una lista de tripletas de la forma *(forma,lema,etiqueta)*. Sin embargo, en una arquitectura como Alexina, donde cada entrada contiene también información sintáctica, cada entrada (intensional) se corresponde con un único lema. Como se explicó en la sección 2, cada lema se asocia a una clase morfológica y las clases se crean mediante una descripción formalizada de la morfología de una lengua. Por tanto, para convertir un léxico morfológico al formato de Alexina, es necesario extraer dicha descripción morfológica de una lista de triplas *(forma,lema,etiqueta)*.

Para ello, hemos desarrollado una sencilla técnica capaz de obtener de forma totalmente automática un conjunto completo de clases morfológicas desde estas triplas. Además, las formas contenidas en el léxico utilizado quedan también automáticamente clasificadas en base a esas clases, por lo que esta técnica tiene un interés doble.

Para cada lema en el léxico, se extrae el prefijo más largo común a todas sus formas flexionadas, que tomaremos como raíz de la familia, y se construye una lista ordenada de pares *(sufijo,etiqueta)*. Si al menos 3 lemas distintos conducen a la misma lista de parejas *(sufijo,etiqueta)*, esa lista se convierte en la definición de una clase morfológica, y todos los lemas asociados a ella son asociados a dicha clase. Además, la raíz de todos los lemas de la clase son analizados en busca del más largo, común a todos sus miembros, para así contruir el patrón regular más específico posible. Esto evita que más tarde otros lemas se añadan a la clase de forma erronea. Por ejemplo, dentro del léxico de Multext, se estableció una clase morfológica usando la lista de pares *(sufijo,etiqueta)* que incluían la terminación *-ar* para el infinitivo, *-a* para la tercera persona del singular del presente de indicativo, y *-ué* para la primera persona del del singular del pretérito perfecto de indicativo. Un ejemplo de verbo perteneciente a esta clase es *halagar*, que conjuga las formas *halaga* y *halagué*. Dado que la raíz de todos los lemas en esta clase terminan en *-g*, el patrón regular `.*g` es asociado a dicha clase morfológica y determina que sólo otros verbos cuya raíz termine en *-g* puedan ser añadidos.

Aquellas clases en las que solo se incluyen uno o dos lemas no se construyen automáticamente. Sus componentes se consideran irregulares y por tanto son definidas manualmente en un fichero externo.

Esta técnica ha sido usada en la práctica tomando como entrada el Multext español para constuir nuestro léxico inicial que incluye ya una descripción morfológica del español en el formato de Alexina. Del mismo modo hemos aplicado dicha técnica sobre el léxico de la USC para transformarlo al formato de Alexina y, tal como esperábamos, ha dado lugar a un conjunto de clases morfológicas distintas. Esto se debe a que el listado de lemas, el juego de etiquetas e incluso a veces el conjunto de formas que genera un mismo lema son diferentes entre un léxico y otro. Se generan por tanto discrepancias que deben ser resueltas si se prentende fusionar dos léxicos morfológicos como estos (ver sección 4.3).

4.2. Conversión de ADESSE y SRG al formato Alexina

La fuente de información sintáctica más importante que hemos usado en nuestros experimentos es ADESSE. Hemos extraído y convertido la información que contiene al formato de Alexina de la forma siguiente:

Cada verbo en ADESSE fue transformado en una o más entradas del L*eff*e obviando por el momento su información morfológica (se asignó una clase morfológica por defecto tan solo a efectos de mantener consistente el formato) pero transformando la estructura argumental descrita en ADESSE en marcos de subcategorización del formato de Alexina. El resultado fue un léxico con una gran cantidad de información sintáctica para un importante número de verbos españoles (en concreto, se creo información para 3.427 lemas únicos).

Dado que algunos verbos incluidos en Multext o en el léxico de la USC no estaban recogidos en ADESSE y teniendo en cuenta también que utilizar varias fuentes de información siempre es interesante para comprobar su validez, hemos extraído también información sintáctica desde SRG. Sin embargo, como veremos a continuación, la técnica que hemos utilizado no es completamente fiable y el léxico SRG tiene una precisión menor que la de ADESSE. Por lo tanto, hemos dado una prioridad menor a la información obtenida desde SRG (ver sección 4.4).

SRG clasifica los verbos usando una jerarquía de clases sintácticas. De este modo, mapear una clase de SRG a L*eff*e significa poder extraer todos los lemas que pertenecen a dicha clase. En nuestro caso, hemos usado L*efff* como puente para establecer la transformación de una clase de SRG al formato de información sintáctica de Alexina. La similitud en términos de comportamientos sintácticos existentes entre el español y el francés permite reutilizar las descripciones sintácticas de L*efff* en el léxico español realizando modificaciones mínimas.

Hemos establecido la transferencia de información sintáctica mapeando las clases[7] a su información correspondiente en L*efff*. Para ello hemos seleccionado un lema representante de cada clase de SRG, tomado su traducción al francés y obtenido su descripción sintáctica en L*efff*. Tan sólo fueron necesarias modificaciones mínimas (traducir preposiciones) para adecuar esa información al L*eff*e. De este modo, pudimos asignar una descripción sintáctica a gran parte de los lemas de SRG en base a su clase sintáctica.

Evidentemente, este proceso puede generar algunas entradas en las que la información sintáctica es incompleta o incluso incorrecta. Para minimizar este problema decidimos ignorar la información extraída en caso de duda.

Aun así, podría no haberse encontrado la información sintáctica para algunas entradas del léxico inicial. Sin embargo la situación contraria es muy poco probable (disponer de la información sintáctica pero no de la información morfolótica) ya que la información morfológica es mucho más común y fácil de encontrar. Por lo tanto hemos establecido como condición necesaria para adquirir entradas de otros recursos el disponer al menos de la información morfológica de dicha entrada. Esto es, conocer su clase morfológica para permitirnos construir sus entradas extensionales.

4.3. Fusión de recursos morfológicos

Una vez transformado al formato de Alexina, un léxico morfológico puede ser visto como un conjunto de pares *(lema,clase)*, donde *clase* denota la clase morfológica de la entrada. Por lo tanto, fusionar un léxico morfológico principal L con un léxico morfológico adicional L' consiste en convertir de algún modo las clases morfológicas de L' en las clases morfológicas de L. Este proceso de fusión ha sido realizado de forma independiente para cada categoría gramatical (verbo, adjetivo, etc.) para evitar problemas relacionados con homónimos.

Para establecer esta conversión, hemos estudiado las clases morfológicas asignadas a los lemas comunes a los dos léxicos. Dada una clase de L', hemos extraido de L' todos los lemas que aparecían también en L y obtuvimos la clase o clases que tenían asignadas en L. Normalmente, la gran mayoría de lemas obtenidos tienen asignada la misma (única) clase en L, pero puede haber excepciones que constituyen incoherencias entre L y L', que además apuntan a errores

[7] En la práctica, hemos extraído las 40 clases más comunes en SRG, que cubren más de 3.000 lemas verbales.

en L y/o L'. Estas incoherencias pueden ser resueltas automáticamente dándole prioridad al contenido de L (o al de L'), o chequeándolas manualmente.

En la práctica hemos aplicado esta técnica siendo L el léxico extraido de Multext (de manera que preservamos el juego de etiquetas de Multext) y L' el resultado de la conversión del léxico de la USC al formato Alexina. El resultado final es el léxico correspondiente a la parte morfológica de Le*ff*e. En la sección 6 mostramos los datos correspondientes a dicho léxico y lo comparamos con otros léxicos morfológicos existentes.

4.4. Fusión de recursos sintácticos

Una vez que hemos construido la parte morfológica del Le*ff*e, debemos completar su información sintáctica. Para los verbos, esta información fue obtenida uniendo los léxicos en formato Alexina obtenidos usando ADESSE y SRG, es decir, dos léxicos intensionales. Para otras categorías, no cubiertas por ADESSE, hemos usado directamente la información sintáctica extraída de las clases sintácticas del léxico SRG. Finalmente, algunas entradas (preposiciones, verbos auxiliares y algunos verbos muy específicos) fueron completadas manualmente.

Contrariamente a (Danlos y Sagot, 2008), los dos léxicos de entrada no usaban el mismo criterio para distinguir entre dos entradas diferentes del mismo lema. Por lo tanto, no era posible mezclar las entradas intensionales directamente. En lugar de eso, el proceso de fusión que hemos utilizado se basa en la noción de *léxico intensional expandido*. Como ya hemos explicado, una entrada intensional incluye un marco de subcategorización con información factorizada de modo que puede haber funciones sintácticas opcionales y realizaciones alternativas de las mismas. Cada una de esas entradas intensionales factorizadas podría transformarse en un conjunto de *entradas intensionales expandidas* simplemente expandiendo la información de tal modo que el nuevo conjunto de entradas generado cubre el mismo grupo de casos que cubría la entrada factorizada original sin elementos opcionales ni alternativos. Por ejemplo, una entrada intensional con el marco de subcategorización $< Suj : cln \mid sn, Obj : (sn) >$ se correspondería con 4 entradas intensionales expandidas: $< Suj : sn >$, $< Suj : cln >$, $< Suj : sn, Obj : sn >$ and $< Suj : cln, Obj : sn >$.

La idea para realizar el proceso de fusión es la siguiente: primero se expanden los dos léxicos intensionales de entrada (las versiones en formato Alexina extraídas de ADESSE y SRG); esos dos léxicos intensionales expandidos son fusionados; finalmente se refactoriza la información sintáctica en el léxico intensional resultante de la fusión. Los dos primeros pasos (expansión y fusión) son simples: desfactorizar la información y hacer una unión de los dos léxicos. El proceso de refactorización cacula la factorización óptima a partir de todas las entradas expandidas de un lema concreto y sin tener en cuenta ninguna información lingüística.

El resultado final es un léxico únicamente con información sintáctica, que pudo ser fusionado directamente con la parte morfológica previamente construida. A aquellas entradas morfológicas cuya información sintáctica no fue adquirida (y que por tanto permanece vacía) se le asignó una información sintáctica por defecto. [8] Por ejemplo, a todos los lemas verbales no cubiertos por ADESSE ni SRG se les asignó el siguiente marco de subcategorización: $< Suj : sn \mid cln, Obj : (sn \mid cla) >$ (verbo transitivo con objeto directo opcional). Evidentemente, esta infomación debe ser completada. Para ello utilizaremos técnicas semi-automáticas de corrección y extensión de léxicos como las descritas en (Nicolas et al., 2008).

5. Extensión del léxico

Tras combinar varios recursos lingüísticos para obtener una primera versión de Le*ff*e, hemos obtenido un léxico con una cobertura significativa (ver sección 6). El siguiente paso es continuar la mejora del léxico encontrando y añadiendo entradas que falten en el mismo. A continuación presentamos un técnica semi-automática, simple y eficicente, que ayuda a encontrar deficiencias en un léxico. Dicha técnica , que presentamos a continuación, es capaz de detectar tanto entradas completamente nuevas, como homónimos de otras existentes.

Para detectar entradas ausentes hemos utilizado un etiquetador morfosintáctico (Graña, 2000; Molinero et al., 2007). Este tipo de etiquetadores tiene la capacidad de

[8] Se asignó la información más común entre todas las otras entradas del léxico pertenecientes a la misma categoría.

establecer (adivinar) etiquetas incluso para palabras que no aparecen en sus léxicos. Dicho etiquetador, entrenado con un corpus español de aproximadamente 500.000 palabras extraido de Ancora (Taulé, Martí, y Recasens, 2008) y usando Le*ff*e como léxico externo, puede ser usado de dos formas distintas en función del tipo de entradas que estemos intentando identificar.

Al intentar encontrar entradas completamente nuevas para el léxico, simplemente confiamos en la habilidad de etiquetador para encontrar etiquetas correctas para palabras desconocidas.

Al buscar homónimos se debe permitir asignar a palabras ya conocidas etiquetas distintas de las que están incluidas en el léxico. Es decir, esas palabras deberían ser consideradas como desconocidas, ya que en otro caso otras posibles etiquetas ni siquiera se considerarían. Para ello, hemos modificado el etiquetador para que en ciertos casos ignore la información del léxico y que por tanto intente adivinar nuevas etiquetas para algunas palabras en base a su morfología y su contexto.

Obviamente, esta estrategia introduce ambigüedad de forma deliberada. Para minimizarla, tan solo se fuerza a tomar como desconocida una única palabra por frase cada vez. Es decir, cada frase se etiqueta varias veces tratando de adivinar nuevas etiquetas para una sola palabra cada vez. Además, dado que las categorías cerradas[9] son bien conocidas y suelen estar bien descritas, solo las palabras pertencientes a categorías abiertas[10] son forzadas a desconocidas.

Por supuesto, los etiquetadores cometen errores y encontrar una nueva etiqueta para una palabra en una sola frase no es suficiente como para garantizar su relevancia. Sin embargo, si se etiqueta un texto de gran tamaño de la forma descrita, es posible obtener conclusiones e incluso clasificar los sospechosos.

Teniendo esto en cuenta, hemos suavizado la aparición de falsos positivos usando la precisión del etiquetador para cada categoría en forma de un índice $prec_{etiq}$, evaluado sobre el corpus de entrenamiento utilizado, y n_{form_etiq}, el número de ocurrencias de la forma $form$ etiquetada como $etiq$ en todo el texto. Concretamente, a cada pareja *(forma, etiqueta)* candidata le asignamos un valor $S_{\text{sc}}(form, etiq)$ calculado de la siguiente forma:

$$S_{\text{sc}}(form, etiq) = prec_{etiq} \cdot \log(n_{form/etiq}) \quad (1)$$

Usando esta medida pudimos generar una lista ordenada de candidatos lo suficientemente buena como para ser evaluada manualmente en muy poco tiempo (ver sección 6).

6. Evaluación preliminar

El Le*ff*e ha sido evaluado mediante las siguientes pruebas: por un lado, hemos comparado el Le*ff*e con otros léxicos de español en términos de cobertura; por otro, hemos medido la mejora obtenida sobre el léxico inicial después de añadir la información extraída de otras fuentes.

En relación a la cobertura, el Le*ff*e beta contiene más de 165.000 pares únicos *(lema,etiqueta)*, que se corresponden con aprox. 1.590.000 entradas extensionales (flexionadas) que asocian a cada forma su correspondiente información morfológica y sintáctica (aprox. 680.000 pares únicos *(forma,etiqueta)*). Otros léxicos presentan los siguientes datos:

- SRG: 76.000 pares únicos *(lema,etiqueta)*[11] (53,9 % menos que Le*ff*e);
- Multext: 510.710 pares únicos *(forma,etiqueta)*[12] (24,9 % menos que Le*ff*e), y no incluye información sintáctica;
- Diccionario español gilcUB-M: 70.000 lemas[12](57,6 % menos que Le*ff*e), y no incluye información sintáctica;
- Léxico de la USC[13]: 490.000 pares únicos *(forma,etiqueta)* (27,95 % menos que Le*ff*e), y no incluye información sintáctica.

Hemos testeado además la cobertura morfológica de nuestro léxico en el contexto de una aplicación real: un preprocesador morfológico (Graña, Barcala, y Vilares, 2002)

[9] Preposiciones, pronombres, determinantes.

[10] Adverbios, nombres comunes, nombres propios, verbos, adjetivos.

[11] Dato obtenido de Freeling (`http://garraf.epsevg.upc.es/freeling/`, abril 2009).

[12] Dato obtenido de ELRA (`http://catalog.elra.info`, abril 2009).

[13] Departamento de Gramática española de la Universidad de Santiago de Compostela

	Palabras desconocidas	Palabras desconocidas únicas
Léxico USC	70.026	25.888
Léxico inicial	86.521	27.234
Le*ff*e beta	69.756	24.703

Cuadro 1: Palabras desconocidas al aplicar el preprocesador morfológico usando distintos léxicos.

desarrollado en el *Centro Ramón Piñeiro para a Investigación en Humanidades* [14] por los grupos COLE[15] y LYS [16]. Hemos realizado un primer test usando los léxicos de la USC, y otros dos con nuestro léxico inicial, y con el Le*ff*e beta.

Hemos utilizado un corpus obtenido de Wikipedia Sources[17] de aprox. 4.320.000 palabras como entrada para estas pruebas. La evaluación del resultado consistió en determinar cuantas palabras no fueron etiquetadas por el preprocesador y que por tanto eran desconocidas para el léxico usado. Conviene destacar la importancia de reducir al máximo el número de palabras desconocidas, ya que estas son la principal causa de errores de etiquetación (Graña, 2000). Evidentemente, esto puede conseguirse utilizando un corpus de gran cobertura.

Como se puede ver en la Tabla 1, el proceso que hemos desarrollado proporciona notables beneficios. El Le*ff*e beta ha superado a otros léxicos en la tarea del preprocesamiento morfológico y puede verse claramente cómo su cobertura morfológica ha aumentado como consecuencia de la reutilización de otros recursos, demostrando el interés y la utilidad del proceso descrito aquí.

Para medir la cobertura sintáctica del Le*ff*e en todos los estados del proceso de fusión hemos usado el concepto de *entradas intensionales expandidas* que describen un comportamiento sintáctico de forma completamente explícita (ver sección 4.4).

El léxico intensional expandido adquirido desde SRG contenía 42.689 entradas únicas, es decir, marcos de subcategorización completamente especificados, mientras que el extraido de ADESSE contenía 39.040. Después de fusionar estos léxicos, le número de entradas únicas ascendía a 66.028. Finalmente, el Le*ff*e beta, que asocia una información sintáctica por defecto a todos los verbos no cubiertos por el resultado de la fusión, contiene 91.507 entradas únicas expandidas. Después de la refactorización, el Le*ff*e contiene 16.311 entradas verbales.

Una vez construida la primera versión de Le*ff*e, hemos usado la técnica descrita en la sección 5 para mejorar su cobertura. Para ello hemos usado un corpus español construido con un subconjunto de la parte española de Europarl[18] que contenía aproximadamente 6 millones de palabras.

Con ello se obtuvo una lista de pares *(forma, etiqueta)* candidatos a ser añadidos al léxico. La calidad de esta lista no era excepcional, ya que el léxico tenía ya una gran cobertura y por tanto el porcentaje de falsos positivos era muy alto. Aun así, esta lista permitió añadir más de 1.800 lemas (88 adjetivos, 54 adverbios, 26 verbos, 117 nombres comunes y 1.518 nombres propios), correspondientes a más de 3.700 formas, en un periodo de tiempo mínimo (fue hecho por una única persona en dos días). Algunos ejemplos de lemas añadidos son *documentar* (verbo), *abstraer* (verbo), biocarburante (nombre común), luxemburgués (adjetivo), Niza (nombre propio), así como un buen número de advervios terminados en *-mente*. Además, permitió la detección de carencias sistemáticas como las relacionadas con diminutivos y aumentativos.

7. Conclusiones

Para muchas lenguas es posible encontrar varios recursos lingüísticos dispersos, pero normalmente ninguno de ellos es suficientemente satisfactorio en términos de cobertura, calidad o riqueza. Aun así, la cantidad de trabajo invertido en su desarrollo no debería ser ignorado. De hecho, reutilizar conocimiento lingüístico ya formalizado es una manera práctica y productiva para construir y/o mejorar otros recursos lingüísticos y entendemos que esta será la estrategia habitual en el futuro próximo.

En este trabajo hemos descrito como construir un léxico morfológico y sintáctico

[14] `http://cirp.es`,abril 2009
[15] `http://www.grupocole.org`, abril 2009
[16] `http://www.grupolys.org`, abril 2009
[17] `http://download.wikimedia.org`, enero 2009
[18] Un corpus paralelo de las actas del Parlamento Europeo.

para el español usando recursos existentes y como extenderlo usando una técnica semi-automática. El contexto teórico y práctico descrito aquí podría ser usado para realizar un proceso similar en otras lenguas.

El léxico resultante, el Le*ff*e, se encuentra actualmente en versión beta y estará pronto disponible bajo licencia LGPL-LR[19]. Aunque evidentemente todavía puede mejorarse en muchos aspectos y debe ser evaluado en mayor profundidad, lo cierto es que hemos mostrado que el Le*ff*e ha superado ya a otros léxicos bien conocidos del español en términos de cobertura morfológica y sintáctica.

Bibliografía

Alvarez, Concepción, Pilar Alvariño, Adelaida Gil, Teresa Romero, María Paula Santalla, y Susana Sotelo. 1998. Avalon, una gramática formal basada en corpus. En *Procesamiento del Lenguaje Natural (Actas del XIV CONGRESO de la SEPLN)*, páginas 132–139, Alicante, Spain.

Danlos, Laurence y Benoît Sagot. 2008. Constructions pronominales dans dicovalence et le lexique-grammaire – intégration dans le Le*fff*. En *Proceedings of the 27th Lexicon-Grammar Conference*, L'Aquila, Italy.

García-Miguel, José M. y Francisco J. Albertuz. 2005. Verbs, semantic classes and semantic roles in the ADESSE project. En *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbrücken, Germany.

Graña, Jorge, Fco. Mario Barcala, y Jesús Vilares. 2002. Formal methods of tokenization for part-of-speech tagging. *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*.

Graña, Jorge. 2000. *Técnicas de Análisis Sintáctico Robusto para la Etiquetación del Lenguaje Natural (robust syntactic analysis methods for natural language tagging)*. Doctoral thesis, Universidad de La Coruña, Spain.

Ide, Nancy y Jean Véronis. 1994. Multext: Multilingual text tools and corpora. En *Proceedings of COLING'94*, Kyoto, Japan.

Marimon, Montserrat, Natalia Seghezzi, y Núria Bel. 2007. An open-source lexicon for Spanish. En *Sociedad Española para el Procesamiento del Lenguaje Natural, n. 39*.

Molinero, Miguel A., Fco. Mario Barcala, Juan Otero, y Jorge Graña. 2007. Practical application of one-pass viterbi algorithm in tokenization and pos tagging. *Recent Advances in Natural Language Processing (RANLP). Proceedings, pp. 35-40*.

Nicolas, Lionel, Benoît Sagot, Miguel A. Molinero, Jacques Farré, y Éric Villemonte de La Clergerie. 2008. Computer aided correction and extension of a syntactic wide-coverage lexicon. En *Proceedings of COLING'08*, Manchester, UK.

Sagot, Benoît. 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. En *Lecture Notes in Artificial Intelligence 3658 (© Springer-Verlag), Proceedings of TSD'05*, páginas 156–163, Karlovy Vary, Czech Republic.

Sagot, Benoît, Lionel Clément, Éric Villemonte de La Clergerie, y Pierre Boullier. 2006. The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use. En *Proceedings of LREC'06*.

Sagot, Benoît y Laurence Danlos. 2008. Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français. En *Proceedings of the workshop "Lexicographie et informatique : bilan et perspectives"*, Nancy, France.

Taulé, M., M.A. Martí, y M. Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. En *Proceedings of 6th International Conference on Language Resources and Evaluation*, Marrakesh, Morocco.

[19]Como hemos explicado en este artículo, hemos utilizado el léxico español desarrollado dentro del proyecto Multext, que es de libre uso para tareas de investigación, en el inicio de la contrucción del Le*ff*e. El Le*ff*e beta es el resultado del trabajo de investigación descrito aquí. Se ha fusionado información lexical proviniente de varios recursos, algunos de ellos con coberturas similares o mayores que las del léxico español de Multext. Por esta razón, consideramos apropiado publicar el Le*ff*e beta bajo licencia LGPL-LR.