# Extensión y corrección semi-automática de léxicos morfo-sintácticos

Lionel Nicolas, Benoît Sagot, Miguel A. Molinero, Jacques Farré, Éric Villemonte de La Clergerie

# Extensión y corrección semi-automática de léxicos morfo-sintácticos*

***Semi-automatic extension and correction of morpho-syntactic lexicons***

Lionel Nicolas♢ Benoît Sagot♣ Miguel A. Molinero♠
Jacques Farré♢ Éric de La Clergerie♣

♢Team RL, Laboratory I3S - UNSA + CNRS, 2000 routes des lucioles B.P. 121
06903 Sophia Antipolis, France
{lnicolas, jf}@i3s.unice.fr
♣Project ALPAGE, INRIA Rocquencourt + París 7, Domaine de Voluceau B.P. 105
78153 Le Chesnay, France
{benoit.sagot, Eric.De_La_Clergerie}@inria.fr
♠Grupo LYS, Univ. de A Coruña, Dpto. de Computación, Fac. de Informática
Campus de Elviña S/N, 15071 A Coruña, España
mmolinero@udc.es

**Resumen:** En este artículo describimos un conjunto de técnicas para la extensión y corrección de léxicos de amplia cobertura. Se basan en la detección de entradas erróneas y la generación automática de hipótesis de corrección mediante el uso del contexto sintáctico. Exponemos los resultados alcanzados sobre un léxico francés y planteamos su aplicación en el desarrollo de un léxico español.
**Palabras clave:** Adquisición de recursos lingüísticos, análisis sintáctico, léxicos morfo-sintácticos, análisis de errores

**Abstract:** This paper describes a set of techniques for the extension and correction of wide-coverage lexicons based on detection of erroneous entries and automatic generation of correction hypotheses using the syntactical context. We report the results achieved on a French lexicon and we consider the application of our techniques on a Spanish lexicon.
**Keywords:** Linguistic resource acquisition, parsing, morpho-syntactic lexicons, error-minning

## *1. Introducción*

El incremento de la cobertura y la precisión de los analizadores sintácticos no entrenados depende fundamentalmente de la mejora de los léxicos y gramáticas que utilizan.

La construcción manual de recursos lingüísticos de amplia cobertura es un trabajo laborioso, complejo y causante de errores, que requiere la intervención de personal experto. Con el objetivo de minimizar la intervención humana, simplificar el proceso y aumentar la calidad de los resultados, es posible usar herramientas automáticas o semi-automáticas. En el presente trabajo presentamos un conjunto de herramientas que permiten detectar defectos en léxicos morfo-sintácticos y proponer correcciones a los mismos. Todo ello tomando texto plano como entrada del proceso.

La extensión y corrección de un léxico puede dividirse en dos fases: Primero identificar entradas erróneas o incompletas en el léxico, y segundo proponer correcciones para dichas entradas.

Afrontamos el primer paso usando dos técnicas que permiten descubrir *formas sospechosas*, es decir, aquellas que parecen causar errores de análisis sintáctico en un conjunto de frases.

La solución al segundo paso se basa en el siguiente principio: podemos encontrar patrones de uso para una forma sospechosa estu-

* Parcialmente financiado por el Ministerio de Educación y Ciencia (HUM2007-66607-C04-02) y la Xunta de Galicia ("Red gallega para el procesamiento del lenguaje y recuperación de información" 2006-2009). Damos también las gracias al grupo COLE de la Univ. de Vigo por permitirnos utilizar sus sistemas de cálculo.

diando varias frases que no han podido ser analizadas y viendo que información hubiera necesitado la gramática para poder realizar análisis completos. Estos esquemas pueden entonces ser planteados como hipótesis de corrección. En cierto modo, podríamos decir que sabemos que el problema se debe al léxico, y le pedimos a la gramática que exprese qué información hubiese aceptado para una forma sospechosa.

El conjunto de técnicas presentado es complemente independiente del lenguaje y de la plataforma. Puede ser aplicado a cualquier a cualquier analizador sintáctico. La única condición es garantizar que el texto usado como entrada es lexical y gramaticalmente correcto. Esto asegura que el rechazo de una frase se debe solamente a errores en algun componente (tipicamente el lexico y/o la grammatica).

Este artículo está organizado de la siguiente manera. Primero introduciremos los conceptos teóricos en los que se basan nuestras técnicas (Sec. 2). Después detallaremos en Sec. 3 y Sec. 4 las técnicas usadas para detectar informaciones erróneas en el léxico. A continuación explicaremos como generar (Sec. 5) y ordenar hipótesis de corrección (Sec. 6). En Sec. 7 comentaremos las diferencias y similitudes con trabajos previos. Después, presentaremos los resultados alcanzados (Sec. 8). Finalmente, hablaremos de trabajo futuro (Sec. 9), justo antes de concluir (Sec. 10).

## 2. *Conceptos teóricos*

Las formas de una lengua suelen describirse en un léxico mediante una o más entradas que incluyen varios tipos de información: la categoría gramatical, información morfológica, información sintáctica (marcos de subcategorización) y información semántica.

Una forma concreta provocará un error de análisis sintáctico si su descripción en el léxico conduce a un conflicto con la gramática. Es decir, cuando la gramática y el léxico no coinciden en el patrón de uso de una forma.

Por razones prácticas diferenciaremos entre conflictos relacionados con categorías gramaticales, que llamaremos **defectos de categorización**, y conflictos relacionados con marcos de subcategorización, que llamaremos **conflictos de rasgos**.

Los defectos de categorización hacen referencia al hecho de que una forma concreta no tenga todas sus posibles categorías gramaticales representadas en las entradas del léxico. Por ejemplo, la forma "ficha"podría aparecer como verbo (fichar) y no como sustantivo. Este tipo de errores suele estar asociada a la homonimia. Se trata de lemas que pueden desempeñar varias categorías gramaticales y alguna de las menos habituales ha sido olvidada.

Los conflictos de rasgos reflejan incoherencias en la descripción del marco de subcategorización de alguna entrada del léxico. Resultan de la dificultad de describir exahustivamente el comportamiento sintáctico de una forma. Si el uso más común es también el más restrictivo, conduce a la sobre especificación, es decir, el marco sintáctico no permite todas las funciones que esa forma puede desempeñar en la práctica.

Tomemos una forma sospechosa cualquiera asociada a un conjunto de frases no analizables, donde dicha forma es la principal sospechosa de causar el fallo de análisis. La generación de correcciones léxicas para esta forma requiere obtener datos de la gramática para cada una de las frases asociadas. Es decir, obtener análisis de frases no analizables. Buscamos el conjunto de análisis sintácticos que la gramática hubiese generado para esas frases con un léxico carente de errores.

Conseguiremos este objetivo eliminando las restricciones sintácticas de la forma sospechosa, es decir, incrementando el conjunto de posibles categorías gramaticales (esto es, añadiendo, de forma virtual, nuevas entradas al léxico) y/o relajaremos las restricciones sintácticas de una entrada del léxico. Aunque a veces la forma sospechosa no es la única razón de todos los errores de análisis, este proceso habitualmente incrementa el porcentaje de análisis completados.

La supresión de restricciones puede verse de la siguiente forma: durante el proceso de análisis sintáctico, cada vez que se accede a la información lexical de una forma sospechosa, el léxico es ignorado y todas las restricciones sintácticas se consideran cumplidas. De este modo, la forma se convierte en lo que la gramática quiera que sea, es decir, encaja con cualquier patrón morfológico y sintáctico que la gramática necesitase para hacer un análisis completo. Estos patrones son los datos que usaremos para generar las correcciones.

Suprimimos las restricciones de las formas sospechosas sustituyéndolas en las frases

por una forma especial que llamaremos **comodín**.

## 3. Detección de defectos de categorización

Con el objetivo de descubrir defectos de categorización en el léxico, hemos desarrollado una técnica que se basa en el uso de un etiquetador estocástico (Graña, Chappelier, y Vilares, 2001; Molinero et al., 2007). La idea es intentar adivinar nuevas categorías gramaticales para las formas del corpus de entrada usando un etiquetador configurado de forma especial. Este etiquetador considerará como desconocidas todas aquellas palabras que pertenecen a las categorías abiertas[1]. Como consecuencia el etiquetador propondrá etiquetas candidatas para cada una de estas palabras y las más probables de ser correctas son escogidas por el propio proceso de etiquetación estocástica. De este modo, nuevas categorías gramaticales surgen para algunas formas del corpus de entrada.

Para obtener este etiquetador hemos usando dos corpus de entrenamiento. El primero es un corpus de oraciones (330K palabras) etiquetado manualmente y extraido del Treebank de la Universidad de París 7 (Abeillé, 2003). El segundo está compuesto por una lista de formas pertenecientes a las clases cerradas[2]. El etiquetador fue modificado para considerar como conocidas las formas pertenecientes al segundo corpus. El resto son consideradas desconocidas.

Hemos pasado el corpus de entrada al etiquetador y extraido los pares forma/etiqueta. Aquellos pares que no existían en el léxico fueron propuestos como candidatos de defectos de categorización. La aparición de falsos positivos ha sido atenuada ordenando los candidatos según la siguiente medida:

$$(n_{wt}/n_w) * log(n_{wt}),$$

Donde $n_{wt}$ es el número de apariciones de la forma $w$ etiquetada como $t$ y $n_w$ es número total de apariciones de la forma $w$.

## 4. Detección de conflictos de rasgos

La técnica descrita aquí amplía las ideas descritas en Sagot y Villemonte de La Clergerie (2006), donde los autores detectan formas sospechosas mediante el análisis estadístico de los resultados de un analizador sintáctico. Esta técnica permite obtener una lista de formas, cada una con un coeficiente de sospecha y un conjunto de frases asociadas en las que dicha forma es la principal sospechosa de ser la causante del fallo de análisis.

Dado que no hay un modo automático e inequívoco para decidir si un fallo de análisis se debe a un error en el léxico o en otro componente del analizador, la técnica de análisis de errores (*error mining*) para detectar formas sospechosas se basa en la siguiente idea: estudiando los resultados del análisis sintáctico de un corpus suficientemente amplio de frases correctas, cuanto menos aparece una forma en frases analizables y más lo hace en frases no analizables, más probable es que las entradas lexicales de esa forma sean incorrectas; sobre todo si dicha forma aparece en frases no analizables junto con otras formas que aparecen en frases analizables.

La principal desventaja es que los resultados dependen en gran medida de la calidad de la gramática usada. De hecho, si una forma concreta está asociada con ciertas construcciones sintácticas no manejadas por la gramática, esta forma aparecerá en frases no analizables y será considerada, incorrectamente, como sospechosa. Se puede limitar este inconveniente aplicando dos mejoras:

- Usar varios analizadores, como se describe en Sagot y Villemonte de La Clergerie (2006), basados en diferentes gramáticas y combinar sus resultados para evitar los errores sistemáticos de cada una de ellas.
- Buscar patrones sintácticos no cubiertos en la gramática y filtrar las frases no analizables donde aparecen. Para hacer esto, se puede reducir cada frase de la entrada a una secuencia de categorías gramaticales mediante un etiquetador, y luego entrenar un clasificador de máxima entropía (Daumé III, 2004) usando los posibles trigramas. Este clasificador permite identificar cada frase, a priori, como analizable o no analizable. Aunque el resultado no sea perfecto (el etiquetador o el clasificador pueden equivocarse), este filtrado permite incrementar notablemente la calidad de los sospechosos que se obtienen mediante el análisis de errores.

[1] Adjetivos, sustantivos, adverbios, verbos y nombres propios.

[2] Preposiciones, determinantes, pronombres y signos de puntuación.

## 5. *Generación de correcciones*

Una vez que las formas sospechosas han sido detectadas y ordenadas, el siguiente paso es sugerir automáticamente correcciones. La manera más simple de generar hipótesis de corrección sería usar comodines que no contengan ningún tipo de restricción. Así se evitarían todo tipo de conflictos y aumentaría notablemente la cobertura del analizador.

Sin embargo, como se explica en Fouvry (2003), esto genera una ambigüedad innecesaria y conduce a una explosión del número de análisis posibles o incluso a ningún análisis por falta de memoria o de tiempo. De modo metafórico, como hemos dicho antes, buscamos que la gramática nos proporcione la información léxical que hubiera aceptado para las formas sospechosas. Introduciendo comodines sin restricciones, la gramática generaría tanta información que no sabríamos cuál tomar como correcta, o incluso podría ser que tenga tantas cosas que decir que no pueda expresar ninguna.

Por lo tanto refinamos los comodines introduciendo datos para restringir su uso y disminuir la ambiguidad. Por razones prácticas, usamos comodines con una categoría gramatical definida.

Para obtener hipótesis sobre defectos de categorización necesitamos que el analizador explore reglas gramaticales distintas a las visitadas cuando el análisis falló. Por lo tanto, para cada forma sospechosa generamos comodines con categorías gramaticales diferentes a las presentes en el léxico.

Para obtener hipótesis sobre conflictos de rasgos, necesitamos que el analizador explore de nuevo las mismas reglas de la gramática pero sin detenerse por fallos de unificación de los rasgos. Por lo tanto generamos comodines con la misma categoría gramatical que aquellos ya presentes en el léxico.

Los análisis obtenidos tras la introducción de los comodines son proporcionados a un módulo de conversión, desarrollado para cada analizador, que extrae la entrada lexical instanciada de cada comodín en el formato del léxico. Esta forma de proceder tiene tres ventajas:

- No se necesita comprender el formato de salida del analizador para estudiar las correcciones;
- Las correcciones propuestas están compuestas exclusivamente de datos relativos al léxico;
- Se pueden combinar los resultados producidos por varios analizadores, lo cual es una solución eficiente para solventar algunas limitaciones del proceso (Ver Sec. 6).

## 6. *Ordenación de las hipótesis*

Los lenguajes naturales son ambiguos, y por tanto lo son las gramáticas que los modelan. Por ejemplo, en algunas lenguas romances, un adjetivo puede ser usado como sustantivo y un sustantivo como adjetivo. En consecuencia, un comodín con una categoría gramatical incorrecta puede conducir a análisis completos y ofrecer correcciones incorrectas. Para paliar este problema clasificamos primero las hipótesis de corrección de acuerdo a sus correspondientes comodines categorizados. Estudiando el porcentaje de análisis completos producidos por cada tipo de comodín y las frases que son analizables gracias a ellos, resulta simple para un humano identificar el comodín válido.

Cuando se usa un solo analizador ordenar las correcciones es una tarea simple, pero los resultados dependen completamente de la calidad de la gramática. Utilizar las hipótesis de corrección provenientes de varios analizadores alivia este problema, pero requiere técnicas de ordenación más sofisticadas.

### 6.1. Ordenación simple con un solo analizador

Las hipótesis de corrección obtenidas después de introducir un comodín son generalmente irrelevantes, es decir, muchas de ellas son correcciones parásitas que provienen de la ambigüedad introducida por el comodín. Sin embargo, entre todas las correcciones, algunas son válidas, o al menos útiles para descubrir las verdaderas. En el ámbito de una sola frase, no hay un modo fiable de determinar cuáles son parásitas y cuáles válidas. Pero si consideramos simultaneamente muchas frases que incluyen la misma forma sospechosa en diferentes construcciones sintácticas reconocidas por diferentes reglas gramaticales, podremos observar una gran dispersión de las hipótesis parásitas. Al contrario, las correcciones correctas que representan el verdadero

sentido de la palabra según la gramática, aparecerán de forma recurrente. Por tanto, ordenaremos las hipótesis de corrección en función del número de frases que la producen.

### 6.2. Ordenación avanzada con varios analizadores

Usar más de un analizador no sólo mejora la detección de formas sospechosas sino que también permite combinar hipótesis de corrección para reducir al máximo la influencia de cada gramática. Cuando alguna forma está relacionada con una construcción sintáctica que no está correctamente cubierta por una gramática, esta forma aparece en frases no analizables y por tanto será sospechosa. Reemplazarla por comodines solo conducirá a correcciones incorrectas porque el problema no se encuentra en el léxico.

Por tanto, usar varios analizadores permite obtener varios conjuntos de frases no analizables y varios conjuntos de hipótesis de corrección. Las hipótesis pueden descartarse (o considerarse menos relevantes) según tres principios:

- Si una forma sospechosa realmente se corresponde con un error en el léxico, ninguna frase que la contenga desempeñando la función sintáctica asociada al error podrá ser analizada. Las hipótesis producidas por frases que son analizables por al menos uno de los analizadores pueden ser descartadas, ya que generalmente el error no proviene del léxico sino de las gramáticas.
- Por la misma razón, las hipótesis de corrección producidas a partir de frases en las que sólo un analizador ha identificado la forma como sospechosa deben ser también eliminadas.
- Finalmente, las hipótesis de corrección propuestas sólo por uno de los analizadores (o propuestas muchas más veces por uno de los analizadores que por los otros) pueden ser simplemente consecuencia de la ambigüedad de la gramática. Al fin y al cabo, las gramáticas describen el mismo lenguaje, por lo que deberían de ofrecer resultados comunes en el uso de una forma.

Entonces, usamos el siguiente esquema de ordenación: dada una forma sospechosa, solo guardamos las hipótesis de corrección que son obtenidas de conjuntos de frases que eran originalmente no analizables, pero que pasan a serlo por todos los analizadores con la introducción de un mismo comodín. A continuación, ordenamos las hipótesis de cada uno de los analizadores por separado y finalmente combinamos los resultados.

## *7. Trabajos relacionados*

Una vez expuestas nuestras técnicas, discutimos las similitudes y diferencias entre nuestras investigaciones y las ya publicadas.

La adquisición/extensión/corrección de léxicos ha sido un tema de investigación muy activo durante los últimos años. Sobre todo desde que formalismos lexicales y gramaticales adecuados para representar conocimiento lingüístico profundo han sido desarrollados.

La idea de inspirarse en el contexto sintáctico para adquirir datos lexicales comenzó en 1990 (Erbach, 1990). La técnica de identificación de formas sospechosas descrita en van Noord (2004), se combinó con esta idea a partir de 2006 (van de Cruys, 2006; Yi y Kordoni, 2006). Salvo en Nicolas, Farré, y Villemonte de La Clergerie (2007), no se ha usado la mejora descrita en Sagot y Villemonte de La Clergerie (2006). Hasta el momento tampoco se ha intentado filtrar las frases de la entrada (Sec. 4) para mejorar la identificación.

La generación de comodines empezó a afinarse a partir del año 1998 (Barg y Walther, 1998). Desde entonces se suelen construir comodines parciales para las clases abiertas. En Yi y Kordoni (2006) se utiliza una elegante técnica de clasificación por entropía para elegir los comodines más adecuados antes de introducirlos.

La forma de clasificar las hipótesis suele ser mediante el uso de una herramienta entrenada (van de Cruys, 2006; Yi y Kordoni, 2006), como un clasificador de entropía, pero nunca se ha intentado evaluar las hipótesis sobre varias frases para discriminar las parásitas.

En definitiva, no se obtuvo ningún resultado concreto en la correción de léxicos hasta el año 2005. van de Cruys (2006) y sobre todo Yi y Kordoni (2006) exponen resultados aceptables basándose en frases extraidas de un Treebank van de Cruys (2006) separa los resultados según la categoría sintáctica y se puede observar claramente, especialmente en lemas complejos como

los verbos, la imposibilidad de aplicar este tipo de técnicas de forma automática sin perjudicar la calidad del léxico. Salvo Nicolas, Farré, y Villemonte de La Clergerie (2007), ningún trabajo expone de forma explícita la dependencia hacia la calidad de las gramáticas usadas, que representa el umbral de esta corriente y explica por qué procedemos de forma semi-automática y no automática.

## *8. Resultados*

A continuación, presentamos los resultados alcanzados al aplicar las técnicas descritas en este artículo al léxico francés Le*fff*[3]. Describiremos primero el contexto práctico y mediremos la efectividad del proceso de corrección.

### 8.1. Contexto práctico

El léxico Le*fff* es un léxico morfo-sintáctico de amplia cobertura que ha sido parcialmente construido usando técnicas de adquisición automática (Sagot et al., 2006). En el momento de escribir el presente artículo, contiene mas de 520.000 formas.

Hemos usado dos analizadores basados en sendas gramáticas:

- FRMG (***Fr**ench **M**eta-**G**rammar*) es una meta-gramática (Thomasset y Villemonte de La Clergerie, 2005) que compilamos en un analizador híbrido TAG/TIG.
- SxLFG-Fr (Boullier y Sagot, 2005; Boullier y Sagot, 2006) es una gramática LFG profunda no-probabilística.

El corpus de entrada usado proviene de un periódico de noticias políticas *Le monde diplomatique* y está formado por más de 280.000 frases de menos de 25 palabras. En total, consta de 4,3 millones de palabras.

### 8.2. Eficiencia de las correcciones

Existen varias formas de medir la calidad de un conjunto de correcciones. En nuestro caso, hemos escogido medir la eficiencia del proceso estudiando el aumento del porcentaje de frases analizables alcanzado durante nuestros experimentos. En cualquier caso, debemos tener presente que las correcciones son validadas y añadidas manualmente, por

[3]Lexique des formes fléchies du français/Léxico de formas flexionadas del francés.

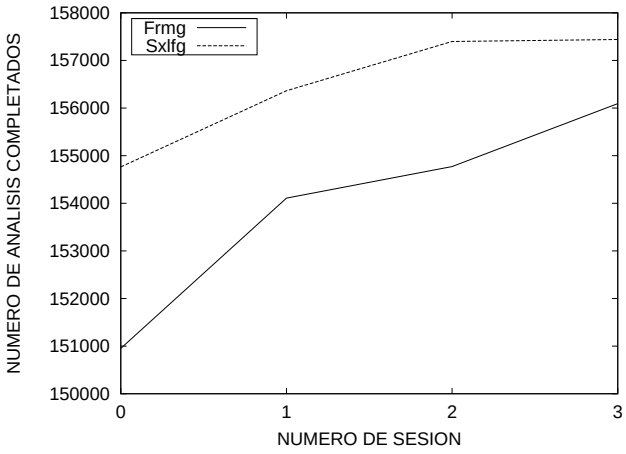


Figura 1: Número de frases analizadas después de cada sesión de corrección.

| Sesión | 1 | 2 | 3 | total |
|---|---|---|---|---|
| nc | 30 | 99 | 1 | 130 |
| adj | 66 | 694 | 27 | 787 |
| verb | 1183 | 0 | 385 | 1568 |
| adv | 1 | 7 | 0 | 8 |
| total | 1280 | 800 | 413 | 2493 |

Cuadro 1: Formas actualizadas en el léxico en cada sesión de corrección

tanto el notable incremento experimentado en la cobertura del analizador se debe globalmente a la mejora del léxico.

La Figura 1 muestra esta ganancia como el número de frases analizables con cada analizador después de cada sesión de corrección.

El cuadro 1 muestra el número de formas actualizadas en el léxico en cada sesión.

Todas las sesiones de corrección han sido realizadas usando las técnicas de detección de errores y generación de hipótesis excepto la segunda sesión. En ella solo ha sido aplicada la técnica de detección de defectos de categorización, que todavía no ha sido conectada con el módulo de generación automática de hipótesis por falta de tiempo. En cualquier caso, la lista de formas sospechosas producida por esta técnica era suficientemente simple como para ser revisada sin la ayuda del módulo de generación de hipótesis.

Como temíamos, los resultados alcanzados han sido rápidamente limitados por la calidad de las gramáticas y del corpus. De hecho,

el léxico y las gramáticas usados han sido desarrollados conjuntamente durante los últimos años usando el mismo corpus como campo de pruebas. Esto hace que la técnica de detección de errores dé lugar a correcciones irrelevantes después de unas pocas sesiones. Además, la técnica de detección de defectos de categorización sólo puede ser usada una sola vez para cada corpus de entrada. Para realizar nuevas sesiones es necesario mejorar o cambiar las gramáticas usadas u obtener nuevos corpora de entrada.

Aun así, en este experimento hemos corregido 254 lemas correspondientes a 2493 formas. El porcentaje de frases analizables ha aumentado un 3,41 % (5141 frases) para FRMG y un 1,73 % (2677 frases) para SXLFG. Cabe destacar que gracias a la eficiencia de las técnicas de detección de errores y generación de hipótesis aquí presentadas, estos resultados fueron alcanzados con tan solo unas pocas horas de trabajo humano.

## *9. Trabajo futuro*

Nuestros esfuerzos se focalizarán en dos tareas.

### 9.1. Aplicación al español

La Universitat Pompeu Fabra[4] ha sido pionera en el desarrollo de un léxico morfosintáctico de amplia cobertura para el español: SRG (Spanish Resource Grammar) (Marimon, Seghezzi, y Bel, 2006), que a día de hoy es el más extenso y desarrollado. En Yi y Kordoni (2006), los autores apuntan a los fallos del léxico como causantes de la mayor parte de los errores de análisis sintáctico de textos generalistas escritos en inglés: alrededor del 70 % de los análisis se detienen por no disponer de información léxica de alguna palabra. La lejanía entre el inglés y el español impide extender esta conclusión. Pero si pensamos en el francés, un idioma mucho más cercano al español en términos lingüísticos, vemos que el Le*fff* describe más de 110.000 lemas, y el SRG tan sólo 50.000. Parece razonable considerar que este recurso todavía ha de ser ampliado.

Consideramos su extensión aplicando la metodología siguiente:

- Ampliaremos el número de lemas aplicando una técnica semi-automática de adquisición (Clément, Sagot, y Lang, 2004) que ha desmostrado su eficacia en varios idiomas tan diferentes como el francés, el eslovaco y el checo. Obtendremos así nuevos lemas con informaciones morfológicas.

- A continuación, aplicaremos la técnica descrita en este articulo para obtener su información sintáctica.

En teoría, esta metodología se puede aplicar incluso a idiomas con léxicos muy pequeños. Pero es necesario que el léxico permita encontrar en el corpus de entrada un buen número de frases con una sola forma sospechosa. SRG es lo suficientemente extenso como para obtener muchas frases que cumplen esta condición, lo cual hace viable el uso esta metodología.

### 9.2. Extensión de las técnicas

Aunque la técnica de detección y corrección de defectos de categorización ha ofrecido resultados aceptables, se encuentra todavía en un estado preliminar. Es necesario disminuir la ambigüedad introducida por el alto número de palabras desconocidas que induce nuestra técnica. Nos planteamos modificarla para considerar, cuando sea posible, una sola palabra desconocida en cada frase. También es necesario conectarla con el módulo de generación de hipótesis de corrección para constituir una herramienta integrada.

Una ventaja del processo está relacionada con su principal desventaja : la dependencia hacia la gramática usada. Si en una frase no analizable no se ha podido validar ninguna de las correcciones propuestas para las formas sospechosas, entonces esta frase puede considerarse léxicamente correcta para el estado actual de la gramática. Es decir, esa frase representa un error de la gramática. Por lo tanto, mejorar sucesivamente el léxico hasta que no dé lugar a nuevas hipótesis de corrección correctas, permitirá obtener un corpus representativo de las carencias de la gramática. Este corpus podría ser la base de otra herramienta que permita mejorar la gramática. Hecho esto, podría usarse de nuevo el mismo corpus en la detección de errores lexicales. De esta forma se podría realizar un proceso alternativo e incremental para la mejora conjunta de léxicos y gramáticas.

[4] http://www.iula.upf.edu/

## 10. Conclusiones

En conclusión, el conjunto de técnicas presentadas han probado ser relevantes y eficientes en la práctica sobre un léxico francés. Su aplicación a un léxico español nos permitirá, por un lado, mejorar los recursos lingüísticos disponibles en español y, por otro, detectar carencias en nuestras técnicas que todavía no hayan sido identificadas.

El punto alcanzado en el desarrollo de las técnicas presentadas no constituye un final. Todavía existen mejoras que podemos implementar pero, sobre todo, es el objetivo de la corrección gramatical el que llama nuestra atención. En efecto, las técnicas presentadas en este trabajo constituyen un sistema efectivo para la extensión y corrección de léxicos morfo-sintácticos. Pero también permiten construir un corpus representativo de las carencias de la gramática, lo cual abre un camino hacia la extensión y corrección de la gramática usada.

## Bibliografía


Abeillé, Anne. 2003. Annotation morpho-syntaxique. Paper available at http://www.llf.cnrs.fr/Gens/Abeille/guide-morpho-synt.02.pdf, January.

Barg, Petra y Markus Walther. 1998. Processing unkonwn words in hpsg. En *Proceedings of the 36th Conference of the ACL and the 17th International Conference on Computational Linguistics.*

Boullier, Pierre y Benoît Sagot. 2005. Efficient and robust LFG parsing: SxLfg. En *Proceedings of IWPT'05*, páginas 1–10.

Boullier, Pierre y Benoît Sagot. 2006. Efficient parsing of large corpora with a deep LFG parser. En *Proceedings of LREC'06.*

Clément, Lionel, Benoît Sagot, y Bernard Lang. 2004. Morphology based automatic acquisition of large-coverage lexica. En *Proceedings of the LREC'04.*

Daumé III, Hal. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at http://pub.hal3.name/daume04cg-bfgs, implementation available at http://hal3.name/megam/, August.

Erbach, Gregor. 1990. Syntactic processing of unknown words. En *IWBS Report 131.*

Fouvry, Frederik. 2003. Lexicon acquisition with a large coverage unification-based grammar. En *Companion to the 10th of EACL.*

Graña, Jorge, Jean-Cédric Chappelier, y Manuel Vilares. 2001. Integrating external dictionaries into stochastic part-of-speech taggers. *EuroConference Recent Advances in Natural Language Processing (RANLP). Proceedings, pp. 122-128.*

Marimon, Montserrat, Natalia Seghezzi, y Núria Bel. 2006. An open-source lexicon for spanish. En *XXIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural.*

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, Jacques Farré, y Éric Villemonte de La Clergerie. 2007. Correction mining in parsing results. En *Proceedings of LTC'07.*

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

Sagot, Benoît y Éric Villemonte de La Clergerie. 2006. Error mining in parsing results. En *Proceedings of ACL/COLING'06*, páginas 329–336. Association for Computational Linguistics.

Thomasset, François y Éric Villemonte de La Clergerie. 2005. Comment obtenir plus des méta-grammaires. En *Proceedings of TALN'05.*

van de Cruys, Tim. 2006. Automatically extending the lexicon for parsing. En *Proceedings of the eleventh ESSLLI student session.*

van Noord, Gertjan. 2004. Error mining for wide-coverage grammar engineering. En *Proceedings of ACL 2004.*

Yi, Zhang y Valia Kordoni. 2006. Automated deep lexical acquisition for robust open texts processing. En *Proceedings of LREC-2006.*