Las entradas adverbiales de diversos léxicos del francés

Elsa Tolone

FaMAF, Universidad Nacional de Córdoba, Argentina

elsa.tolone@univ-paris-est.fr

Introducción

Los léxicos sintácticos son recursos básicos en la mayoría de las tareas avanzadas de Procesamiento del Lenguaje Natural, ya que la mayoría de los sistemas que disponen de cierta capacidad de comprensión del Lenguaje Natural requieren conocimiento sintáctico y semántico a nivel de predicativo (verbo, nombre o adjetivo).

En sintáctica, muchos trabajos buscan como identificar y formalizar reglas generales que se aplican a una clase de palabras. En particular, las reglas de transformación de Chomsky describen las relaciones sistemáticas entre diversas estructuras sintácticas. Pero estas generalizaciones no se aplican a todas las palabras. La descripción completa de la sintáctica de una lengua implica la identificación de reglas generales, pero también la determinación de que palabra exige, autoriza o prohibe, la aplicación de que regla en particular. Este fue el trabajo iniciado por Maurice Gross en los años 1970.

Su objetivo ha sido acumular datos empíricos los más completos y detallados posibles, lo que ha permitido formar una gran base de datos lexicales, sintácticos y semánticos : las tablas del Léxico-Gramática.

Tablas del Léxico-Gramática

Las tablas del Léxico-Gramática constituyen hoy en día uno de los principales léxicos sintácticos del francés (Gross 1975). Esas tablas han sido compiladas originalmente en los años 1970 por Maurice Gross, en el laboratorio LADL, y luego en el Laboratorio de Informática Gaspard-Monge (LIGM, Universidad de Paris-Est) en Francia. Cada tabla reúne elementos lexicales de una categoría gramatical (verbos, nombres predicativos, adverbios, expresiones estereotipadas, etc.) que comparten propiedades.

Las tablas del francés han sido mejoradas y evaluadas en un analizador sintáctico por (Tolone 2011). Contienen información útil para distintas tareas de Procesamiento del Lenguaje Natural, como el análisis sintáctico o la busqueda y la extracción de información. Por eso, se siguen mejorando añadiendo o completando las entradas y sobre todo, convirtiendolas en diferentes formatos que permiten usarlas directamente en dicha tarea.

Esta presentación se enfoca en los adverbios de las tablas del Léxico-Gramática de dos categorías: los adverbios simples que terminan en *-ment* (*-mente,* por ejemplo, *difficilement* - *dificilmente*) (Molinier & Levrier 2000) y los adverbios estereotipados (como por ejemplo, *du jour au lendemain* - *de un día para otro*) (Gross 1986b ; Gross 1986a). Las tablas contienen 3 203 adverbios simples en *-ment* y 7 285 adverbios simples y estereotipados.

La disponibilidad de léxicos lexicales, sintácticos y semanticos es indispensable para mejorar el reconocimiento y el análisis de los adverbios, incluyendo los problemas duales de variabilidad y de ambigüedad (Laporte & Voyatzi 2008). También puede ayudar a resolver el *PP-attachment* durante una análisis sintáctica profunda o superficial (Agirre *et al.* 2008).

Léxico sintáctico *LGLex*

En las tablas están descritas relaciones de supresión, permutación, intensificación y parafrases que asocian por un lado, adverbios simples a adverbios estereotipados, y por otro lado, diferentes tipos de adverbios estereotipados. Todos esos nuevos adverbios han sido extraídos en el léxico sintáctico *LGLex* del francés (Tolone & Voyatzi 2011) con la herramienta *LGExtract* (Constant & Tolone 2010). Ha permitido enriquecer el léxico de 11 351 entradas (+108%), por lo cual el léxico está compuesto de 21 843 entradas adverbiales en total.

Léxico morfológico DELA

Esos adverbios han sido convertidos al léxico morfológico DELA (Diccionarios Electrónicos del LADL) (Courtois & Silberztein 1990) del francés (Tolone *et al.* 2012a). El objectivo es integrarlos en el DELA inicial (el que ya existía), que está compuesto de 9 036 entradas. Ha permitido crear 20 761 entradas directamente utilizables, más 830 entradas que contienen variables y que les hace falta un tratamiento (remplazar las variables) para un reconocimiento inmediato en los textos.

Sin tener en cuenta las entradas con variables, y después de fusión y supresión de los duplicados, el DELA del francés ha sido enriquecido de 13 445 entradas (+149%), el DELA extendido está ahora compuesto de 22 481 entradas en total. Para las 830 entradas adverbiales conteniendo variables se han creado grafos diccionarios para poder reconocerlas tal cual aparecen en los textos. Un grafo diccionario es una especie de transductor que llama a sub-grafos y es capaz de producir dinamicamente nuevas entradas del diccionario al formato DELA.

Evaluación

Se han evaluado los datos obtenidos en el corpus de referencia del francés anotado por las expresiones estereotipadas con sus funciones adverbiales (Laporte *et al.* 2008). El corpus anotado contiene 8 794 frases o 168 846 palabras y está compuesto de:

– la transcripción de las sesiones de la Asamblea Nacional Francesa del 3 y 4 de octubre 2006;

– la novela de Jules Vernes *Le Tour du monde en quatre-vingt jours*, escrita en 1873.

Hemos producido y analizado la anotación automática del corpus por el DELA inicial, el DELA extendido y los grafos diccionarios. La referencia anotó 3 239 adverbios estereotipados (la referencia no anotó los adverbios simples), el DELA inicial anotó 3 690 adverbios (simples y estereotipados), el DELA extendido anotó 4 706 adverbios, los grafos diccionarios anotaron 727 adverbios y el DELA extendido con los grafos diccionarios anotaron 5 209 adverbios.

Conclusión

Todos los recursos lingüísticos estan disponibles en *http://infolingu.univ-mlv.fr/english* (*Languages Ressources* > *Lexicon-Grammar* > *Download*).

La anotación con el DELA extendido y los grafos diccionarios permite reconocer adverbios válidos, incluyendo los adverbios simples que no están reconocidos por la referencia. Pero esa nueva anotación aumenta también el ruido. Hace falta filtrar las nuevas entradas añadidas en el DELA para sacar o modificar algunas de ellas (añadiendo más restricciones), en particular las que son ambiguas.

Queremos luego convertir las nuevas entradas adverbiales al formato Le*fff* (Sagot 2010), con el objetivo de integrarlas en un analizador sintáctico, como lo han hecho (Tolone & Sagot 2011) y (Tolone *et al.* 2012a). Además, podemos también mejorar el Wordnet del francés con esas entradas adverbiales (Sagot *et al.* 2009).

Referencias

Agirre E., Baldwin T., & Martinez D. (2008), Improving parsing and PP attachment performance with sense information, in: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, Columbus, Ohio.

Constant M. & Tolone E. (2010), A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables, in: De Gioia M. (ed.), *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008), Seconde partie,* Volume 1 of *Lingue d'Europa e del Mediterraneo, Grammatica comparata*, Aracne, p. 79-93.

Courtois B., Silberztein M., (1990), *Dictionnaires électroniques du français*, Langue Française 87, Larousse, Paris, Francia.

Gross M. (1975), *Méthodes en syntaxe*, Hermann, Paris, Francia.

Gross M. (1986a), *Grammaire transformationnelle du francais : Syntaxe de l'adverbe*, Volume 3, ASSTRIL, Paris, Francia.

Gross M. (1986b), Lexicon-grammar: The representation of compound words, in: *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn, Alemania.

Laporte E., Nakamura, T. & Voyatzi, S. (2008), A French Corpus Annotated for Multiword Expressions with Adverbial Function, in: *Proceedings of the 6th Language Ressources and Evaluation Conference (LREC'08), Workshop on Linguistic Annotation Conference (LAW II)*, Marrakecj, Morocco, p. 48-51.

Molinier C. & Levrier F. (2000), *Grammaire des adverbes : description des formes en -ment*, Droz, Genève, Suiza.

Sagot B., Fort K. & Venant F. (2009), Extending the adverbial coverage of a French Wordnet, in: *Proceedings of the NODALIDA 2009 workshop on WordNets and other Lexical Semantic Resources*, Odense, Dinamarca.

Sagot B. (2010), The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French, in: *Proceedings of the 7th Language Resources and Evaluation Conference*, La Valette, Malta.

Tolone E. (2011), *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français,* Thèse de doctorat, Université Paris-Est, Francia, 340 pp.

Tolone E. & Sagot B. (2011), Using Lexicon-Grammar tables for French verbs in a large-coverage parser, in: Vetulani Z. (ed.), *Human Language Technology, Challenges for Computer Science and Linguistics, 4th Language and Technology Conference, LTC 2009, Poznán, Poland, November 6-8, 2009, Revised Selected Papers*, Volume 6562 of *Lecture Notes in Artificial Intelligence (LNAI)*, Springer Verlag, p. 183-191.

Tolone E. & Voyatzi S (2011), Extending the adverbial coverage of a NLP oriented resource for French, in: *Proceedings of poster session of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, Chiang Mai, Thailandia, p. 1225-1233,

Tolone E., Voyatzi S. Martineau C. & Constant M. (2012a), Extending the adverbial coverage of a French moprhological lexicon, in: *Proceedings of the 8th Language Ressources and Evaluation Conference (poster LREC'12)*, Istanbul, Turquia.

Tolone E., Sagot B. & de La Clergerie E. (2012b), Evaluating and improving syntactic lexica by plugging them within a parser., in: *Proceedings of the 8th Language Ressources and Evaluation Conference (LREC'12)*, Istanbul, Turquia.