Spanish Morphosyntactic Disambiguator

Octavio Santana Suárez *(osatana@dis.ulpgc.es)*

Departamento de Informática y Sistemas.
Universidad de Las Palmas de Gran Canaria

José Rafael Pérez Aguiar *(jperez@dis.ulpgc.es)*

Departamento de Informática y Sistemas.
Universidad de Las Palmas de Gran Canaria.

Luis Javier Losada García *(losada@dis.ulpgc.es)*

Departamento de Informática y Sistemas.
Universidad de Las Palmas de Gran Canaria

Francisco Javier Carreras Riudavets *(fcarreras@dis.ulpgc.es)*

Departamento de Informática y Sistemas.
Universidad de Las Palmas de Gran Canaria

1. Introduction

The written expression of an idea is not achieved only through the simple combination of the different components of the grammar based on a given syntax. Other factors take part in the process, such as semantics and context. But it is obvious that a first approach requires at least a correct syntactic analysis, and for this it is necessary, from the computer-science point of view, to obtain results similar to those obtainable by human knowledge. In this work, a first approach is achieved by the identification and then disambiguation of the elements that are part of a sentence.

Traditionally, syntactic analysis requires a specialized knowledge of the language, all the more so in the case of Spanish, due to its wide range of variations which turn the syntactic analysis into a task only for experts. From the educational point of view, syntactic analysis is very useful to help learn to distinguish the different symbols implied: on the one hand, the correct combination of the elements by means of the application of grammar rules, and on the other hand, the incorporation of less tangible, although necessary aspects, like semantics and context. People usually perform an intuitive use that hides the true difficulty of the problem.

This system is intended to provide a close view of the Spanish grammar to researchers, enhancing their performance and reliability. This is a first step that will allow, with the addition of new features, to keep improving until reaching100% accuracy. Any automated processing of a text entails inevitably the syntactic analysis of its sentences, following the morphosyntactic disambiguation of the elements that compose it, allowing for different possible applications: a) to provide a precise synonymfor a given word, b) to analyze its literary style, c) toreveal its semantics, d) to extract information or summarize its contents, e) to make trustworthy translations to other languages, f) to answer to concrete questions on its content, etc.

2. Methodology

In this work, the number of erroneous syntactic representation trees, obtained by the application of the rules of the Spanish grammar by means of a set of structural disambiguation rules, is notably reduced. In spite of the remarkable amount of necessary combinations, this system does not limit itself to subgroups of the grammar like most of the other proposals, but instead it uses a system of rules which covers all the possible combinations of the Spanish grammar. In addition to being the starting point for an automated syntactic analysis system, it complements the local functional disambiguator developed by the Group of Data Structures and Computational Linguistics of the University of Las Palmas de Gran Canaria(<http://www.gedlc.ulpgc.es/investigacion/desambigua/desambigua.htm>). As an indicator of its performance, theaccuracy of the disambiguation is raised from 87% to 96%.

A solution is provided to the problem of the appearance of structural ambiguities that are generated during the process of construction of syntactic representation trees. The syntactic structures are combined to each other to allow for the syntactic representation trees. Many of these combinations generate erroneous trees. Direct conflicts between rules have been identified as one of the main causes of the problem. The characteristics of the different syntactic structures and how they must be considered at the time of accepting or not the construction of a representation symbol have been studied for the development of methods of structural disambiguation.

In view of the great number of possible combinations of the grammar elements (more evident in verb-phrase constructions which allow any number of elements and almost in any combination), the adequate representation mechanisms have been defined so that all the possibilities are covered, not leaving valid options unrepresented. When allowing any combination of possible elements in the verb-phrase, some combinations appear, which should not be allowed, and would be rejected in the structural disambiguation processes. In this way, all the possible combinations are represented, from a structural point of view, and those not allowed are rejected.

Groups of semantic identification oriented to the recognition of syntactic structures are catalogued. The processes of structural disambiguation include some rules that introduce semantic information. The generated lists have been obtained from the tables of the ideological dictionaries that can be related to certain syntactic structures.

3. Knowledge base

The grammar used is based mainly on the description made by Gili Gaya. To achieve maximum system completeness and include all the syntactic structures that can appear we followed Gutiérrez Araus. The examplescited by Gómez Torrego (2002a, 2002b), were useful to test the system and contributed mainly to illustrate the aspects relative to the compound sentences that remained to be refined.

For this work, the tagger developed by GEDLC was used (<http://www.gedlc.ulpgc.es/investigacion/scogeme02/lematiza.htm>) which gathers the main lexicographical repertoires of the Spanish language[1], and admits 151103 canonical forms and something more than 4900000 inflectioned and derived forms (without adding the inherent extension to the prefixes and the enclitic pronouns that have also been contemplated).

4. Related works

There are other authors that approach this problem for the Spanish language from diverse points of view. In the same way as our work, which can be used for free at discretion through the Internet (<http://www.gedlc.ulpgc.es/investigacion/desambigua/morfosintactico.htm>), we have only been able to find oneother operative tool of this kind on the network: the parser from the Center of Language and Computing of the University of Barcelona. Given the high complexity of the problem, they have chosen to write down exclusively those elements that are explicitly present in the sentence, which had led them to a simplified treatment of some syntactic aspects like coordination and some subordinated types that they leave unsolved. Also, they abandon the concept of sentence understood like noun-phrase and verb-phrase, optingfor a list of components instead.

Although the computer methodologies applied are different, they try to reach the same objectives. Our work is based on the real and complete study of: a) a Spanish grammar that includes all the possibilities available in the written language, b) the direct structural ambiguities that cause the appearance of multiple syntactic representation trees, c) the symbols that cannot cover all the sentence, d) the complex verbal form, e) other situations where ambiguities can be solved based on linguistic knowledge about words, grammar categories and objects involved, and f) the considerations for the generation of the predicate symbol. Nevertheless, other methodologies apply statistical criteria for the resolution of ambiguities, with the consequent loss of reliability for unfrequent cases. The richness of our language and, particularly, the writers' freedom in the construction of syntactic structuresmakes usreconsider the probabilistic methods as the only solution to this complex problem.

5 Conclusions

This work is not limited to subsets of the grammar, but is based instead on a system of rules for the Spanish grammar in spite of the remarkable quantity of necessary combinations.

It provides a solution to the problem of the appearance of functional ambiguities. First a disambiguation process is applied, based on local syntactic structures that reach an accuracy of 87%; and second, another disambiguation process is applied, based on trees of syntactic representation that improve the averageaccuracy level up to 96%.

The importance of this work lies on the fact that it fosters the development of future applications, because:

1. It accelerates the process of syntactic analysis when pruning incorrect structures.
2. It improves the precision in the results of advanced word searches.
3. It allows the discarding of non valid options in information extraction.
4. It detects grammatical errors in the written constructions.

[1] Alvar Ezquerra; Casares; García Márquez & Hernández; Diccionario General de la Lengua Española Vox; Gran Diccionario de la Lengua Española; Gran Diccionario de Sinónimos y Antónimos; Moline; Real Academia Española.

Bibliography

Bosque, I., V. Demonte, and F. Lázaro Carreter. *Gramática descriptiva de la lengua española.* Madrid: Espasa, 1999.

Casares, J. *Diccionario ideológico de la lengua española.* Barcelona: Gustavo Gili, 1994.

Diccionario General de la Lengua Española Vox, Edición en CD-ROM. Barcelona: Biblograf, S.A., 1997.

Ezquerra, Alvar. *M. Diccionario de voces de uso actual.* Madrid: Arco-Libros, 1994.

García Márquez, Gabriel, and Humberto Hernández. *Clave. Diccionario de Uso del Español Actual, Edición en CD-ROM.* Madrid: Ediciones SM, 1997.

Gili Gaya, S. *Curso Superior de Sintaxis Española (Higher Course on Spanish Syntax).* Barcelona: Biblograf S.A., 1998.

Gómez Torrego, L. *Análisis sintáctico: Teoría y práctica.* Madrid: Ediciones SM, 2002a.

Gómez Torrego, L. *Gramática didáctica del español.* Madrid: Ediciones SM, 2002b.

Gran Diccionario de la Lengua Española. Barcelona: Larousse Planeta, S.A., 1996.

Gran Diccionario de Sinónimos y Antónimos. Madrid: Espasa-Calpe, 1991.

Gutiérrez Araus, M.L. *Estructuras sintácticas del español actual (Syntactic Structures of Current Spanish).* Madrid: Sociedad General Española de Librería, S.A, 1978.

Moliner, M. *Diccionario de Uso del Español, Edición en CD-ROM.* Madrid: Gredos, 1996.

Quesada, J.F. *Un modelo robusto y eficiente para el análisis sintáctico de lenguajes naturales mediante árboles múltiples virtuales.* Sevilla: Centro Informático Científico de Andalucía (CICA), 1996.

Real Academia Española. *Esbozo de una nueva gramática de la lengua española.* Madrid: Espasa-Calpe, 1989.

Real Academia Española. *Diccionario de la Lengua Española, Edición electrónica.* Madrid: Espasa-Calpe, 1995.

Santana, O., J. Pérez, Z. Hernández, F. Carreras, and G. Rodríguez. "FLAVER: Flexionador y lematizador automático de formas verbales." *Lingüística Española Actual* XIX, 2 (1997): 229-282.

Santana, O., J. Pérez, F. Carreras, J. Duque, Z. Hernández, and G. Rodríguez. "FLANOM: Flexionador y lematizador automático de formas nominales." *Lingüística Española Actual* XXI, 2 (1999): 253 - 297.

Santana, O., J. Pérez, L. Losada, and F. Carreras. "Hacia la desambiguación funcional automática en Español." *Procesamiento del Lenguaje Natural* 28 (2002): 1-22.