Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 103-106 recibido 23-11-14 revisado 27-01-15 aceptado 10-02-15

SSG: Simplified Spanish Grammar. An HPSG Grammar of Spanish with a reduced computational cost

SSG: Simplified Spanish Grammar. Una gramática del español de tipo HPSG de coste computacional reducido

Benjamín Ramírez González
Qindel Group
Príncipe de Vergara, 204, 28002 Madrid
bramirez@qindel.com/benjaminramirezg@gmail.com

Abstract: PhD Thesis written by Benjamín Ramírez González at the Universidad Complutense de Madrid, under the supervision of Dr. Fernando Sánchez León (Real Academia Española, Technology Department). It was defended on February 25th, 2014 at the Instituto Universitario Ortega y Gasset, and it was awarded Summa Cum Laude. The members of the committee were José Lázaro Rodrigo (Universidad Complutense de Madrid), Guadalupe Aguado de Cea (Universidad Politécnica de Madrid), Montserrat Marimón Felipe (Universidad de Barcelona), Olga Fernández Soriano (Universidad Autónoma de Madrid) and Cristina Sánchez López (Universidad Complutense de Madrid).
Keywords: HPSG, computational grammar, Spanish grammar, computational complexity, reduction of computational cost, lexical rules reduction, diathesis alternations, clitics, word order.

Resumen: Tesis escrita por Benjamín Ramírez González en la Universidad Complutense de Madrid, bajo la dirección del doctor Fernando Sánchez León (Departamento de Tecnología de la Real Academia Española). La tesis fue defendida el 25 de febrero de 2014 en el Instituto Universitario Ortega y Gasset y obtuvo una calificación de sobresaliente cum laude. El tribunal lo formaron los doctores José Lázaro Rodrigo (Universidad Complutense de Madrid), Guadalupe Aguado de Cea (Universidad Politécnica de Madrid), Montserrat Marimón Felipe (Universidad de Barcelona), Olga Fernández Soriano (Universidad Autónoma de Madrid) y Cristina Sánchez López (Universidad Complutense de Madrid).
Palabras clave: HPSG, gramática computacioanal, gramática del español, complejidad computacional, reducción de coste computacional, reducción de reglas léxicas, alternancias de diátesis, clíticos, orden de palabras.

1 Objectives and motivation

This PhD Thesis presented SSG (Simplified Spanish Grammar), an HPSG (Head-driven Phrase Structure Grammar) Spanish Grammar.

Every computational grammar of a natural language must face the challenging problem of ambiguity. In order to analyze a sentence in a natural language, an HPSG grammar must generate all possible behavioral patterns of every word in the sentence in the first stages of the process, and then try all possible combinations. In fact, the result in non-trivial cases is a combinational explosion of hypothetical behavioral patterns.

This thesis aims to develop the core of an HPSG grammar of Spanish with a really small amount of lexical rules, which has been named Simplified Spanish Grammar (SSG). It is claimed that SSG analysis are elegant and theoretically motivated, and such analysis significantly reduces the computational cost of grammar and improves analysis times.

2 Structure of the thesis

Three main groups of central phenomena in Spanish have been implemented in SSG.

The first phenomenon is diathesis alternations. From a computational point of

ISSN 1135-5948

view, this is one of the most challenging phenomena in natural languages as verbs can usually behave in very different ways: they may have both active and passive versions, they may accept certain optional complements, and so on. HPSG lexical rules are meant to deal with these alternations.

Traditional computational grammars usually deal with this diversity by means of specialized lexical rules or lexical units to: transitive verbs with nominal object, transitive verbs with nominal object and dative, transitive verbs with clausal object, transitive verbs with clausal object and dative, and so on. This traditional approach fails to capture due generalizations. Every grammatical reality (transitivity, passive, and a certain kind of dative complement) should be implemented just once. Moreover, argumental positions can be filled with different types of phrases, which mean that both clausal and nominal objects should be considered different fillers available to the same argumental position in the same pattern. This thesis develops a system in which every intuitive verbal pattern is implemented with a unique lexical rule.

The second central grammatical phenomenon implemented in SSG is the Spanish clitics system. Clitization in HPSG has always been formalized by means of lexical rules. By following this approach, many lexical rules and clitization patterns can be added to grammar, which can become a great source of complexity. In Spanish, both accusative and dative arguments can suffer clitization. Moreover, depending on the context, a clitic can appear instead of its canonical object or beside it. Therefore, this thesis develops an analysis of clitics that avoids using any rule or lexical unit intended to deal with clitics.

The last grammatical phenomenon implemented in an innovative way in SSG is word order. The possibilities of word order are a great source of complexity in every Spanish computational grammar. First of all, canonical preverbal subjects can be inverted in several contexts. That inversion has been implemented in traditional HPSG grammars by means of a lexical rule, which leads to a bigger combinational explosion of patterns. At the same time, post-verbal complements can switch their canonical positions, maybe only in a specific context, with certain intonation patterns and with different informational purposes. SSG proposes an analysis of subjects as postverval complements. This proposal is plausible in a theoretical way and contributes to reduce the combinational explosion of grammar. At the same time, in SSG, post-verbal linearization of complements is implemented, according to the classical Linearization Theory in HPSG, as non-continuous constituents.

Finally, it has been added a compared analysis of the same test suit both with SSG and NSSG (Non Simplified Spanish Grammar). NSSG is a traditional grammar whose analysis of diathesis alternation, clitics and word order use the traditional lexical rules. In order to analyze this test suite, as a part of this thesis, SGP (Simplified Grammars Parser) has been developed. SGP is a bunch of libraries written in Perl. SGP provides all the needed tools to analyze written text with HPSG grammars. Moreover, it provides all the needed tools to analyze with SSG, such as a library that joins clitics and verb, as well as a parser compatible with discontinuous constituents.

3 Contributions and future work

It is claimed that SSG analysis are elegant and theoretically motivated, and such analysis significantly reduces the computational cost of grammar and its analysis times. Specifically, these are the main contributions of SSG.

3.1 Theoretical contributions: non-destructive lexical rules

In this thesis it has been coined the term non-destructive rule. Usually, in HPSG, all verbs are supposed to have a canonical characterization, and lexical rules are intended to change that canonical pattern into another. These rules destruct a feature structure and create another one. Crucially, input and output are not supposed to be necessarily compatible. The result is that an HPSG rule is able to change its input in almost every way: it can add or remove an argument, change its category, its case, its position and so on. Unlike previous grammars, lexical rules used by SSG are non-destructive rules. Non-destructive rules never change their input structure, they only specify them. In a non-destructive rule, input and output must share their feature structure and both structures must be identical. Those rules take an underspecified verb and specify it by adding information compatible with their original characterization. The non-destructive rule system is easier to implement and maintain than

a traditional system. This approach has theoretical significance. Every science aims to explain as much data as possible with a theoretical system in the simplest way possible. HPSG lexical rules can operate almost every conceivable change in input and this power reduces HPSG's explanatory capacity. A non-destructive lexical rules system can entirely solve this problem. All non-destructive rules can be reduced, in fact, to a single universal operation: specification, application of an independently-legitimated behavioral pattern.

3.2 A drastic reduction of lexical rules by means of a linguistically motivated analysis

SSG deduces syntactic behavior of verbs from their semantic characterization. Verbs in SSG are really under-specified in a syntactic sense, but they feature a rich semantic characterization. It has been assumed that syntactic alternatives share a common semantic background. A classic semantic characterization has been used: verbs can be accomplishments, achievements, activities or states. According to this main classification, the semantic feature structure of verbs informs about the possible presence of an external argument, an inner argument, and the ability of the verb to receive a certain kind of dative complements or certain controlled predicates. Verbs are also crucially characterized by relevant syntactic features: their ability to assign accusative case or government idiosyncrasies. All these features are well-known verbal characterization criteria, so it is safe to say that they are natural and linguistically motivated. The interesting point is that, just by means of a system of several simple, classic notions, it is possible to develop a general grammar of diathesis alternations of Spanish verbs in a non-destructive fashion. On the other hand, lexical rules restring the nature of their arguments in an interesting way. SSG has a general description of the general notion of argument and it also has a description of case: nominative, accusative, dative and obliq cases. The confluence of all these notions, as well as several semantic idiosyncrasies of certain verbs, successfully regulates the nature of the fillers of every argumental position.

Moreover, in SSG clitics are verbal affixes. Thanks to this morphological approach, SSG avoids using a grammatical rule to merge clitics and verb. In SSG, clitics information is added to the verb by means of an inflectional rule. Note that inflectional rules do not trigger combinational explosion, because they are applied separately and only if pre-syntactic analysis (tokenization) has found actual clitics in the verb. In SSG, clitics are not considered fillers available to an argumental position. Rather, they are only the morphological mark that certain words have left in the verb when they have filled their accusative or dative position. These words are personal pronouns, elliptic pronouns and traces left in topicalization processes. This thesis claims that these words exist in grammar independently of clitics. The outcome is a system of clitics that does not add complexity to the grammar.

Finally, SSG features innovative analysis of Spanish word order. In Spanish, subjects are typically pre-verbal arguments. But a grammar with canonical preverbal subjects features a systematic ambiguity between local and topicalized subjects. In order to reach a simplified and computationally efficient analysis of subject linearization, SSG regards subjects as originally post-verbal arguments where pre-verbal subjects are the result of a topicalization. It is claimed that this approach is plausible in theoretical terms, it solves ambiguity (all preverbal subjects are topics) and reduces the computational cost of grammar. Post-verbal complements in Spanish can be sorted in many ways (scrambling). SSG analysis of scrambling leads to a great simplification of grammar. This solution is a technical application for Spanish of a well known theoretical proposal in HPSG. The key idea is to use discontinuous constituents: all arguments are always listed in the same order in the verb. However, the parser is able to merge two constituents no matter if they are adjacent. In that case, all these arguments, which are always listed in the same order, can be found in different relative positions. This approach has not been applied to traditional computational grammars because traditional parsers cannot deal with this kind of discontinuous constituents. In this thesis, it has been implemented a parser able to do that. For this reason, SSG does not need any rule to deal with scrambling as all complements are always listed in the verb according to a unique increasing order of obliquity.

3.3 A drastic reduction of analysis time

SSG proposals significantly reduces the computational cost of grammars and its analysis times, as it was proved by this work in an empirical way.

For a future work, it would be interesting to improve the current version of SGP (Simplified Grammars Parser). It should include a wrapper to a C library in order to perform feature structures unification tasks in an efficient way, and it should include support to statistical information. On the other hand, SSG coverage should be expanded.

4 References

Bender, E., D. Flickinger and S. Oepen. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. CSLI, Stanford University.

Chomsky, N. 1956. Three Models for the Description of Language. *IRE Transactions PGIT* 2, pp. 113–124.

Donohue, C. and I. Sag. 2006. Domains in Warlpiri. Stanford University.

Fernández Soriano, O. 1993. Sobre el orden de lapabras en español. *Cuadernos de Filología Hispánica* II, pp. 113–152.

Flickinger, D. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6, pp. 15–28.

Gazdar, G., E. Klein, G. Pullum and I. Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press.

Levin, B. and M. Rappaport. 1995. *Unaccusativity at the Syntax-Lexical Semantics Interface*. Massachusetts Institute of Technology (MIT Press).

Marimon, M. 2013. The Spanish DELPH-IN Grammar. *Languages Resources and Evaluation* 47.

Monachesi, P. 1998. Decomposing Italian Clitics. In *Romance in HPSG*, pp. 305–357. CSLI Publications.

Müller, S. 2004. Continuous or Discontinuous Constituents? A Comparison between Syntactic Analyses for Constituents Order and Their Processing System. *Research on Language and Computation* 2, pp. 209-257.

Pineda, L. and I. Meza 2005. The Spanish Pronominal Clitic System. *Procesamiento del lenguaje natural* 34, pp. 67–104.

Pollard, C., R. Kasper and R. Levine 1993. Studies in Constituent Ordering: Toward a Theory of Linearization in HPSG. Grant Proposal to the National Science Foundation, Ohio State University.

Ramírez González, B. 2014. Hacia un modelo computacional unificado del lenguaje natural. *Linguamática* 5:2.

Sánchez León, F. 2006. Gramáticas y Lenguajes Formales. Departamento de Lingüística Computacional de la Real Academia Espñola.

Wall, L. and R. Schwartz. 1991. *Programming Perl*. O'Reilly Media.