

Test-suite Construction for a Spanish Grammar

Montserrat Marimon, Núria Bel and Natalia Seghezzi
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra

Proceedings of the GEAF 2007 Workshop

Tracy Holloway King and Emily M. Bender (Editors)

CSLI Studies in Computational Linguistics ONLINE

Ann Copestake (Series Editor)

2007

CSLI Publications

<http://csli-publications.stanford.edu/>

Abstract

This paper describes the testing component we use for the development and maintenance of the Spanish Resource Grammar, an open-source multi-purpose broad-coverage HPSG grammar for Spanish implemented within the LKB system. Following a brief description of the main features of the grammar, we describe the set of test suites we have manually constructed and the way we have extended them with publicly available data with the aim of producing better resources for testing our grammar.

1 Introduction

Natural language is a system of rather complex interactions. Grammar writing, thus, requires broad and systematic testing to be successful in both research and industrial contexts.

This paper describes work on the development and maintenance of the testing component for a multi-purpose broad-coverage precise grammar for Spanish implemented within the LKB system, the Spanish Resource Grammar. On the one hand, for the development of a multi-purpose grammar, the linguistic phenomena included in the testing component should be abstracted away from any particular application.¹ On the other hand, broad-coverage grammar writing requires testing material which not only includes all variations of a particular phenomenon, but also reflects the real world language complexity; thus, in addition to traditional controlled test items, test data should include examples that present combinations of different phenomena.

We see the development and maintenance of the testing module as part of the process of grammar writing. The idea behind evaluation is to determine its usability for different applications. The grammar can also be evaluated in terms of recall (i.e. coverage) and precision (i.e. overgeneration) with large natural language corpora.

The rest of the paper is organized as follows. Section 2 presents the main features of the grammar. Then, section 3 describes the set of test suites we have manually constructed. In section 4 we explain the way we have extended them with publicly available data with the aim of producing better resources for testing our grammar.²

[†]This research was supported by the Spanish *Ministerio de Educación y Ciencia* under the programmes *Ramón y Cajal* and *Juan de la Cierva*.

¹In previous experience in grammar writing – in the *Advanced Linguistic Engineering Platform* platform – during the European projects LS-GRAM (LRE-61029), MELISSA (ESPRIT-22252) and IMAGINE (IST-2000-29490), the test data we used to define (and refine) the coverage of the grammar was designed according to the user need analysis; consequently, breadth and depth of grammar coverage was defined by the applications.

²The test data we describe may be downloaded from: <http://www.upf.edu/pdi/iula/montserrat.marimon/>

2 The Spanish Resource Grammar

The Spanish Resource Grammar (SRG) is an open-source³ multi-purpose large-coverage grammar for Spanish.

The grammar is grounded in the theoretical framework of HPSG (*Head-driven Phrase Structure Grammar* (Pollard and Sag, 1987, 1994)). HPSG is a constraint-based, lexicalist approach to grammatical theory where all linguistic objects are represented as typed-feature structures. The grammar uses *Minimal Recursion Semantics* (MRS) for the semantic representation. MRS is a flat approach to semantic representation for large-coverage linguistically-motivated computational grammars of natural language that can be used for both parsing and generation (Copestake et al., 2006). The SRG is implemented within the *Linguistic Knowledge Builder* (LKB) system (Copestake, 2002), based on the basic components of the grammar Matrix,⁴ an open-source starter-kit for the development of HPSG grammars developed as part of the LinGO consortium's multilingual grammar engineering (Bender et al., 2002; Bender and Flickinger, 2005).

The SRG has a full coverage lexicon of closed word classes (pronouns, determiners, prepositions and conjunctions) and it contains about 50,000 lexical entries for open classes (6,600 verbs, 28,000 nouns, 11,200 adjectives and 4,000 adverbs). These lexical entries are organized into a type hierarchy of about 400 leaf types (defined by a type hierarchy of around 5,500 types). The grammar also has 50 lexical rules to perform valence changing operations on lexical items (e.g. movement and removal of complements), and 150 phrase structure rules to combine words and phrases into larger constituents and to compositionally build up the semantic representation.

The range of linguistic phenomena that the grammar handles includes: all types of subcategorization structures, surface word order variation and valence alternations, subordinate clauses, raising and control, determination, null-subjects and impersonal constructions, compound tenses, modification, passive constructions, comparatives and superlatives, cliticization, relative and interrogative clauses, sentential adjuncts, negation, and coordination among others. Appendix A includes a more detailed (though not complete) list of the phenomena covered by the grammar.

Following previous experiments within the *Advanced Linguistic Engineering Platform* (ALEP) platform (Marimon, 2002), we have integrated a shallow processing tool, the FreeLing tool, as a preprocessing module of the grammar with the aim of improving both coverage and robustness. The FreeLing tool is an open-source⁵ language analysis tool suite performing shallow processing functionalities which include: text tokenization (includ-

³The Spanish Resource Grammar may be downloaded from: <http://www.upf.edu/pdi/iula/montserrat.marimon/>

⁴<http://www.delph-in.net/matrix/>

⁵The FreeLing tool may be downloaded from <http://www.lsi.upc.edu/nlp/freeling>

ing MWU and contraction splitting), sentence splitting, morpho-syntactic analysis and disambiguation, proper name detection and classification, date / number / currency / ratios / physical magnitude (speed, weight, temperature, density, etc.) recognition, chart-based shallow parsing, WordNet-based sense annotation and dependency parsing⁶ (Atserias et al., 2006). The integration of FreeLing allows us to release the parser from certain tasks (i.e. morphological analysis and recognition and classification of special text expressions, e.g. numbers, dates, percentages, currencies, proper names, etc.) that may be reliably dealt with by shallow external components. Our hybrid architecture also permits the implementation of default lexical entry templates for unknown words for virtually unlimited lexical coverage (Marimon et al., 2007).

We are also investigating Machine Learning (ML) methods applied to the acquisition of the information contained in the lexicon of the SRG (Bel et al., 2007; Marimon et al., 2007). The automatic acquisition of lexical information is a very active area of research. It is specially important for deep linguistic analysis due to the central role that lexical information has in lexicalized grammars and the cost of hand-crafting them (Korhonen, 2002; Carroll and Fang, 2004; Baldwin, 2005; Blunsom and Baldwin, 2006; Zhang and Kordoni, 2006). The most successful systems of lexical acquisition are based on the linguistic idea that the contexts where words occur are indicative of the particular types of words. Although the methods used are different, most of the systems work upon the syntactic information of words as collected from a corpus, and they develop different techniques to decide whether this information is relevant for type assignment or whether it is noise. In the LKB system, lexical types are defined as a combination of grammatical features. For our research, we have worked with morpho-syntactically motivated features. Thus, words are assigned a number of features, the combination of which will help in defining the particular lexical type the word belongs to.

The SRG is part of the DELPH-IN open-source repository of linguistic resources and tools for writing (the LKB system), testing and benchmarking (the [incr tsbd()] competence and performance profiler (Oepen and Carroll, 2000)) and efficiently processing HPSG grammars (the PET system (Callmeier, 2000)), as well as an architecture for integrating deep and shallow natural language processing components to increase robustness of HPSG grammars (the Heart of Gold (Schäfer, 2007)). Further linguistic resources that are available in the DELPH-IN repository include broad-coverage grammars for English, German and Japanese as well as smaller

⁶FreeLing also includes a guesser to deal with words which are not found in the lexicon by computing the probability of each possible PoS tag given the longest observed termination string for that word. Smoothing using probabilities of shorter termination strings is also performed. Details can be found in (Brants, 2000; Samuelsson, 1993).

grammars for French, Korean, modern Greek, Norwegian and Portuguese.⁷

3 Hand-built Test Suites

Together with the linguistic resources (grammar and lexicon) the SRG includes a set of test suites. A test suite is a hand-constructed collection of test cases, e.g. sentences, that exemplify the grammatical – and related ungrammatical – constructions that the grammar should parse, or not, in the case of the ungrammatical ones.

The construction and maintenance of the test suites plays a major role in the development of the SRG.⁸ Test suites provide a fine-grained diagnosis of the grammar behaviour in terms of coverage, overgeneration and efficiency when we change and/or extend the grammar components. To determine that the output produced is correct we have to inspect it manually. Test suites also allow us to compare the SRG with other DELPH-IN grammars. Comparison with other DELPH-IN grammars, e.g. English Resource Grammar and *La Grenouille* (i.e. the French Resource Grammar), is done by producing parallel test data (i.e. data that covers the same phenomena) and comparing the outputs at the MRS level.

In building the test suites, we followed the guidelines for test suite construction and maintenance of the TSNLP project (LRE-62-089) to meet the demands for systematicity and exhaustivity (i.e. systematic increase in depth of coverage), and control over data (i.e. control of interaction of phenomena and ambiguity). Thus,

- We test linguistic phenomena in isolation or in controlled interactions. Most of our test cases include a single grammatical phenomenon in each test sentence.
- Starting from simple test items and increasing their complexity progressively (e.g. in (1), where we show the positive test items we have created to test the non-universal quantifier/adjective *poco* (few)), we provide test cases which show systematic and exhaustive variations over each phenomenon, including infrequent phenomena and variations recognized as linguistically interesting but which do not occur commonly in corpora.

⁷See <http://www.delph-in.net/>

⁸Note that there are no standard general purpose test suites publicly available for Spanish – like the Hewlett Packard (HP) test suite for English (Flickinger et al., 1987), the DITO test suite for German (Nerbonne et al., 1991), or the TSNLP for English, French and German (Lehmann et al., 1996; Oepen et al., 1997) – we could use. Our test suites have been primarily aimed at the SRG, in that some of the test data has been designed to test its linguistic modules. Nevertheless, most test data reflect central language phenomena, and this makes them adequate and reusable in other parsing systems. It is hoped that the availability of this testing material will be of value to the NLP community.

(1)

- a. *Pocos muchachos lloran.* (A few boys cry.)
- b. *Otros pocos muchachos lloran.* (Other few boys cry.)
- c. *Pocos otros muchachos lloran.* (Few other boys cry.)
- d. *Los pocos muchachos lloran.* (The few boys cry.)
- e. *Los pocos otros muchachos lloran.* (The few other boys cry.)
- f. *Todos los pocos otros muchachos lloran.* (All the few other boys cry.)
- g. *Casi todos los pocos otros muchachos lloran.* (Almost all the few other boys cry.)
- h. *Muy pocos otros muchachos lloran.* (Very few other boys cry.)

- We avoid irrelevant variation (i.e. different instances of the same lexical type or same syntactic structure) and both structural and lexical ambiguity. Note that even the simplest sentences may turn to be ambiguous as the grammar coverage increases. For example, when testing the Spanish definite articles, a sentence like (2.a) becomes ambiguous when dealing with elliptical constructions, as we show in (2.b) and (2.c).

(2)

- a. *Los chicos lloran.*
- b. DET NOUN VERB (The boys cry.)
- c. DET ADJ VERB (The small (ones) cry.)

- We include negative or ungrammatical data to check both overgeneration and coverage. Following the TSNLP, negative cases are derived from well-formed ones by one of the following operations:
 - replacement, e.g. change of agreement features (**ambos **muchacho** lloran* (both boys cry)), change of mood (**quiero que **vienes*** ((I) want that (you) come)), change of marking preposition (**el muchacho desertó **desde** su regimiento* (the boy deserted from his regiment)), change of copular verb (**los muchachos **son** contentos* (the boys are happy)).
 - addition, e.g. of subject in impersonal constructions (****el cielo** llueve* (the sky rains)).
 - deletion, e.g. of an obligatory complement (**los muchachos **fabrican** 0* (the boys produce)), of an obligatory complementizer (**las muchachas **intentaron** 0 los muchachos lloraran* (the girls tried the boys cried)).

- permutation, e.g. inversion of word order (**tres unos muchachos lloran* (three about boys cry), **los todos muchachos lloran* (the all boys cry)).

Test cases have been divided by linguistic phenomena. We currently use 16 files testing linguistic structures plus one file for special text constructions. Test cases include a short annotation describing the phenomenon that we are actually testing and the number of expected results when ambiguity cannot be avoided (e.g. when we test optionality). Note that even ungrammaticality may be due to different reasons, for example the ungrammatical sentence in (3) may be derived by removing the definite article under the reading where *todo* (all) is a definite quantifier, or by changing the agreement features under the reading where *todo* is an indefinite quantifier, in which case it co-occurs with singular nouns.

(3) **todos muchachos lloran.* (all boys cry.)

Table 1 shows the set of test suites of the grammar with the number of test items that each contains.

Test suites	Phenomenon	Number of items
t01_basic_subcat	basic subcategorization structures for verbs, nouns (and pronouns) and adjectives	99
t02_null_subj	pro-drop and impersonal verbs	6
t03_det	determiners/quantifiers	32
t04_val_alt	surface word order variation and valence alternations	10
t05_cl_comp	finite/non-finite completive clauses and indirect questions	77
t06_rais_cntrl	raising and control	26
t07_aux	compound tenses	7
t08_pass	passive constructions	7
t09_mod	basic modifiers	58
t10_compar	comparatives and superlatives	13
t11_sent_mod	sentential adjuncts	3
t12_rel_cl	relative clauses	77
t13_ques	interrogative clauses	30
t15_se_constr	impersonal and passive constructions with <i>se</i>	20
t16_clitics	clitics	71
t17_coord	coordination	60
txx_messy_details	special text constructions	53

Table 1: Hand-built test suites for the SRG.

4 Extending the Test Suites

Test suites have traditionally been used to test linguistically-motivated computational grammars.⁹ Though simple, test cases included in hand-built test suites are crucial to determine progress in grammar development.

Controlled hand-built test suites are certainly necessary for incremental grammar maintenance and development to detect unintended interactions of extensions and/or changes in the linguistic resources that cause the treatment of some phenomena already covered to deteriorate. However, from the point of view of building a large-coverage grammar, test data that shows the real world language complexity is also necessary. Therefore, test cases that present combination of different phenomena should also be included in the testing module.

Combining all different phenomena could lead to a combinatorial explosion; besides, not every combination of phenomena produces grammatical sentences or shows interesting cases. Instead, we have re-used available data reflecting natural combination of phenomena.

As the coverage of our grammar increased, hand-constructed sentences were complemented by real corpus cases we took from:

- a. the Spanish questions from the Question Answering track at CLEF (CLEF-2003, CLEF-2004, CLEF-2005 and CLEF-2006). We built up a test suite with 619 test items we took from the 800 available sentences. We only removed those sentences which only differed in a proper name.
- b. the general sub-corpus of the Corpus Tècnic de l'IULA (IULA's Technical Corpus; (Cabr e and Bach, 2004)); this sub-corpus includes newspaper articles and it has been set up for contrastive studies. We built up a test suite with some of the articles that we chose randomly.

CLEF cases include short queries, sentences and a few NPs showing none or very little combination of phenomena, and an average of 9.2 words. Most of these test cases include core linguistic phenomena, e.g. verbs with only one complement (DO, attribute), passives with *ser* and *estar*, impersonal constructions, comparatives and superlatives, and basic nominal and verbal modifiers (i.e. APs, PPs, temporal NPs), and we find just a few examples of more complex structures such as relatives clauses, coordination or ellipsis. Very rarely more than two or three different phenomena appear in the same sentence.

Newspaper articles include more interesting and challenging linguistic structures showing a high level of syntactic complexity due to the combination of several phenomena in a sentence. Sentences are longer, ranging up to 35 words.

⁹Other testing mechanisms are briefly described in (Butt and King, 2003).

The parsing of this new data displayed unanticipated analyses showing errors/deficiencies not only in our linguistic modules (grammar rules and lexical entries were not restrictive enough to exclude some ungrammatical examples which had not been considered), but also in the FreeLing tool (and, for example, we realized that the FreeLing tool allowed enclitics to appear on all verbal forms; in Spanish, clitics can only be attached to imperatives, gerunds and infinitives).

We are currently shifting to much more varied corpus data, and we are extending the test suites with more specialized tests (these have also been chosen randomly) from the *Corpus Tècnic de l'IULA*. This includes specialized corpora of written text in the areas of computer science, environment, law, medicine and economics, collected from several sources, such as legal texts, textbooks, research reports, user manuals, etc. In these texts sentence length may range up to 100 words. In addition, this corpus contains highly specialized words which must be added to the lexicon.

5 Conclusions

We have presented the set of test suites we have manually constructed for the development and maintenance of an open-source multi-purpose broad-coverage HPSG grammar for Spanish, and the way we have extended them with publicly available data with the aim of producing better resources for testing our grammar. Even though our test suites have been primarily aimed at the SRG, most test data reflect central language phenomena, and this makes them adequate and reusable in other parsing systems. It is hoped that the availability of this testing material will be of value to the NLP community.

A Grammar Coverage

List of linguistic phenomena that the SRG handles. Note that we have not included all variants of the phenomena.

- main clauses with canonical word order, i.e. SVO.
- subcategorization structures: unaccusative verbs (*nacieron* ((they) were born), *viven en la ciudad* ((they) live in the town)), intransitive verbs (*claudicaron* ((they) gave in), *desertó del regimiento* ((s/he) deserted from the regiment), *me gusta el muchacho* (I like the boy)), transitive verbs (*fabrican juguetes* ((they) make toys), *abastecieron la ciudad de víveres* ((they) provided the town with provisions), *colgaron los cuadros en el salón* ((they) hang the paintings in the living-room)), *acercó la sal a la muchacha* ((s/he) brought over the salt to the girl)), quantifying nouns (*el grupo de los muchachos* (the group of boys)),

argument taking nouns (*el padre de la muchacha* (the father of the girl), *el apoyo de los muchachos a la muchacha* (the boys' support of the girl)), transitive adjectives (*es atento con los muchachos* ((s/he) is kind to the boys)).

- null-subjects: pro-drop (*lloraremos* ((we) will cry)) and impersonal verbs (*llueve* ((it) rains), *hay muchos muchachos* (there are a lot of boys)).
- determination: def/indef. articles (*el/un muchacho llora* (the/a boy cries)), demonstratives (*estos muchachos lloran* (these boys cry)), possessives (*mis muchachos lloran* (my boys cry)), quantifiers (*todos los/algunos/muchos muchachos lloran* (all of the/some/many boys cry)).
- surface word order variation: subject-predicate inversion (*en ese parque anidan pájaros* (in this park nest birds)), complement permutation (*el muchacho acercó a la muchacha la sal* (the boy brought the girl the salt)).
- subordinate clauses (finite and non-finite), e.g. *quieren que lloren* ((they) want that (they) cry), *la ventaja de claudicar/que claudicaran* (the advantage of giving in/that (they) gave in), *está segura de que lloraron* ((she) is sure that (they) cried)).
- indirect questions (finite and non-finite), e.g. *preguntó cómo ir/que cuándo claudicaron* ((s/he) asked how to go/when (they) gave in), *la incógnita de si claudicarán* (the question of whether (they) will give in), *está seguro de dónde fue/ir* ((he) is sure about where (s/he) went/to go).
- raising and control verbs (subj-control (*intentaremos claudicar* ((we) will try giving in)), obj-control (*me gusta llorar* (I like crying)), subj-to-subj raising (*deberían claudicar* ((they) should give in)), perception (*vieron a la muchacha llorar* ((they) saw the girl crying)), and adjectives (subj-control (*es capaz de claudicar* ((s/he) is able to give in)), subj-raising (*es libre de claudicar* ((s/he) is free to give in)), and obj-raising (*el es fácil de querer* ((s/he) is easy to love))).
- compound tenses, e.g. *hemos llorado* ((we) have cried), *estamos llorando* ((we) are crying), *hemos estado llorando* ((we) have been crying), *fue invadido* ((it) was invaded), *ha sido invadido* ((it) has been invaded), *está siendo invadido* ((it) is being invaded), *ha estado siendo invadido* ((it) has been being invaded).
- passive constructions: with *ser*, with and without optional PP_{por} (i.e. by-agent) (see example above); and with *estar* (*el país está invadido* (the country is invaded)).

- constructions with *se*: impersonal constructions (*se invadió el país* (the country has been invaded)), passive constructions (*los países se invaden* (the countries are invaded)) and unaccusative constructions (*el tren se abre* (the train opens)).
- modification
 - verbal modifiers: PPs (*corrió desde el parque/ahí* ((s/he) ran from the park/there)), adverbs (*quizás claudicarán* (maybe (they) will give in), *claudicarán pronto* ((they) will give in soon)), temporal NPs (*correrá el lunes*, ((s/he) will run on Monday)).
 - nominal modifiers: APs (*el muchacho alto murió* (the tall boy died)), PPs (*el muchacho de ahí llora* (the boy from there cries)), participles (*el país invadido* (the invaded country)), proper names (*mi amigo Juan llora* (my friend Juan cries)), adverbs (*fabrican sólo juguetes* ((they) make only toys)).
 - adjectival modifiers: adverbs (*es muy guapo* ((he) is very pretty), PPs (*fiel hasta la muerte* (loyal to death)).
 - adverbial modifiers: adverbs (*corre sólo muy raramente* ((s/he) runs only very occasionally)).
 - prepositional modifiers: adverbs (*corrió sólo por el parque* ((s/he) ran only along the park)), PPs (*lloró desde el aeropuerto hasta la ciudad* ((s/he) cried from the airport to the town)).
- comparatives, e.g. *es mejor* ((s/he) is better), *es más listo* ((he) is cleverer), *tiene tantos libros como revistas* ((s/he) has as many books as journals); and superlatives, e.g. *es el mejor muchacho* ((he) is the best boy).
- clitics: complement cliticization (*los abrió* ((s/he) opened them)), clitic doubling (*a él le gusta el muchacho* (he likes the boy)), clitic climbing (*los ha comido* ((s/he) has eaten them)).
- relative clauses: restrictive RC (*el hombre que claudicó llora* (the man that gave away is crying)), non-restrictive RC (*el chico, con cuyos padres cuento, claudicará* (the boy, whose parents I count on, will give in)), free RC (*yo vivo donde tú vives* (I live where you live)), semi-free RC (*el que claudica claudica* (who gives in gives in)).
- interrogative clauses: polar-questions (*Quieres libros?* (do you want books?)), wh-questions (*con quién cuentas?* (whom do you count with?), *quién tiene qué?* (who has what?)).
- sentential adjuncts, e.g. *las muchachas lloraron porque los muchachos claudicaron* (the girls cried because the boys gave in), *el muchacho corrió hasta morir* (the boy ran to die).

- negation, e.g. *los muchachos no claudicaron* (the boys didn't give in).
- coordination of all major categories: binary (*fabrica coches y juguetes* ((s/he) makes cars and toys), *es guapo y listo* ((he) is pretty and clever)), multiple (*es guapo, guapo y guapo* ((he) is pretty, pretty and pretty)), doubled conjunctions (*no sólo guapo sino listo* (not only pretty, but also clever)), unlike categories (*habló alta y claramente* ((s/he) talked loud and clearly), *habló claramente y sin parar* ((s/he) talked clearly and without stopping)).

References

- Atserias, Jordi, Casas, Bernardino, Comelles, Elisabet, González, Meritxell, Padró, Lluís and Padró, Muntsa. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC'06*, Genoa, Italy.
- Baldwin, Timothy. 2005. Bootstrapping Deep Lexical Resources: Resources for Courses. In *Proceedings of Workshop on Deep Lexical Acquisition, ACL-SIGLEX 2005*, Ann Arbor, Michigan, USA.
- Bel, Núria, Espeja, Sergio and Marimon, Montserrat. 2007. Automatic Acquisition of Grammatical Types for Nouns. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, USA.
- Bender, Emily M. and Flickinger, Dan. 2005. Rapid Prototyping of Scalable Grammars: Towards Modularity in Extensions to a Language-Independent Core. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing IJCNLP-05 (Posters/Demos)*, Jeju Island, Korea.
- Bender, Emily M., Flickinger, Dan and Oepen, Stephan. 2002. The grammar Matrix. An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics COLING-02*, Taipei, Taiwan.
- Blunsom, Phil and Baldwin, Timothy. 2006. Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

- Brants, Thorsten. 2000. TnT: A statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, Seattle, USA.
- Butt, Miriam and King, Tracy H. 2003. Grammar Writing, Testing, and Evaluation. In A. Farghaly (ed.), *Handbook for Language Engineers*, pages 129–179, CSLI Publications.
- Cabré, Teresa and Bach, Carme. 2004. El corpus tècnic del IULA: corpus textual especializado plurilingüe. *Panacea*, V. 16 pages 173–176.
- Callmeier, Ulrich. 2000. Pet a platform for experimentation with efficient HPSG processing. In D. Flickinger, S. Oepen, J. Tsujii and H. Uszkor-eit (eds.), *Natural Language Engineering (6)1 —Special Issue: Efficiency Processing with HPSG: Methods, Systems, Evaluation*, pages 99–108, Cambridge University Press.
- Carroll, John and Fang, Alex C. 2004. The automatic acquisition of verb sub-categorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing IJCNLP-04*, Sanya City, China.
- Copestake, Ann. 2002. *Implementing Typed Feature Structure Grammars*. Chicago: CSLI Publications, CSLI lecture notes, number 110.
- Copestake, Ann, Flickinger, Dan, Pollard, Carl J. and Sag, Ivan A. 2006. Minimal Recursion Semantics: an Introduction. *Research on Language and Computation* 3(4), 281–332.
- Flickinger, Dan, Nerbonne, John, Sag, Ivan A. and Wassow, Thomas. 1987. Toward evaluation of NLP systems. In *Technical Report, Hewlett-Packard Laboratories. Distributed at the 24th Annual Meeting of the Association for Computational Linguistics ACL-87*, Stanford, USA.
- Korhonen, Anna. 2002. *Subcategorization acquisition. A Technical Report UCAM-CL-TR-530*. UK: University of Cambridge.
- Lehmann, Sabine, Oepen, Stephan, Regnier-Prost, Sylvie, Netter, Klaus, Lux, Veronika, Klein, Judith, Falkedal, Kirsten, Fouvry, Frederik, Estival, Dominique, Dauphin, Eva, Compagnion, Herve, Baur, Judith, Balkan, Lorna and Arnold, Doug. 1996. TSNLP: test suites for natural language processing. In *Proceedings of the 16th International Conference on Computational Linguistics COLING-96*, Copenhagen, Denmark.
- Marimon, Montserrat. 2002. Integrating shallow linguistic processing into a unification-based Spanish grammar. In *Proceedings of the 19th International Conference on Computational Linguistics COLING-02*, Taipei, Taiwan.

- Marimon, Montserrat, Bel, Núria, Espeja, Sergio and Seghezzi, Natalia. 2007. The Spanish Resource Grammar: Pre-processing Strategy and Lexical Acquisition. In *Proceedings of the Workshop on Deep Linguistic Processing, ACL-2007*, Prague, Czech Republic.
- Nerbonne, John, Netter, Klaus, Diagne, Abdel K., Dickmann, Ludwig and Klein, Judith. 1991. A Diagnostic Tool for German Syntax. In J. G. Neal and S. M. Walter (eds.), *Natural Language Processing Systems Workshop: Final Technical Report RL-TR-91-362*, New York: Rome Laboratory.
- Oepen, Stephan and Carroll, John. 2000. Performance Profiling for Parser Engineering. In D. Flickinger, S. Oepen, J. Tsujii and H. Uszkoreit (eds.), *Natural Language Engineering (6)1 —Special Issue: Efficiency Processing with HPSG: Methods, Systems, Evaluation*, pages 81–97, Cambridge University Press.
- Oepen, Stephan, Netter, Klaus and Klein, Judith. 1997. TSNLP: test suites for natural language processing. In J. Nerbonne (ed.), *Linguistic Databases*, CSLI Publications, CSLI lecture notes.
- Pollard, Carl J. and Sag, Ivan A. 1987. *Information-Based Syntax and Semantics. Volume 1: Fundamentals*. Stanford: CSLI Publications.
- Pollard, Carl J. and Sag, Ivan A. 1994. *Head-driven Phrase Structure Grammar*. Chicago: The University of Chicago Press and CSLI Publications.
- Samuelsson, Christer. 1993. Morphological tagging based entirely on Bayesian inference. In *Proceedings of 9th Nordic Conference on Computational Linguistics*, Stockholm, Sweden.
- Schäfer, Ulrich. 2007. *Integrating Deep and Shallow Natural Language Processing Components – Representations and Hybrid Architectures*. Ph.D.thesis, Faculty of Mathematics and Computer Science, Saarland University, Saarbrücken, Germany.
- Zhang, Yi and Kordoni, Valia. 2006. Automated deep lexical acquisition for robust open text processing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC'06*, Genoa, Italy.