# The Spanish Resource Grammar

**Montserrat Marimon**

Universitat de Barcelona
Gran Via de les Corts Catalanes 585
08007-Barcelona
montserrat.marimon@ub.edu

**Abstract**

This paper describes the Spanish Resource Grammar, an open-source multi-purpose broad-coverage precise grammar for Spanish. The grammar is implemented on the Linguistic Knowledge Builder (LKB) system, it is grounded in the theoretical framework of Head-driven Phrase Structure Grammar (HPSG), and it uses Minimal Recursion Semantics (MRS) for the semantic representation. We have developed a hybrid architecture which integrates shallow processing functionalities – morphological analysis, and Named Entity recognition and classification – into the parsing process. The SRG has a full coverage lexicon of closed word classes and it contains 50,852 lexical entries for open word classes. The grammar also has 64 lexical rules to perform valence changing operations on lexical items, and 191 phrase structure rules that combine words and phrases into larger constituents and compositionally build up their semantic representation. The annotation of each parsed sentence in an LKB grammar simultaneously represents a traditional phrase structure tree, and a MRS semantic representation. We provide evaluation results on sentences from newspaper texts and discuss future work.

## 1. Introduction

This paper describes the Spanish Resource Grammar (SRG). This grammar is designed as multi-purpose (abstracted away from any particular application), and broad-coverage (aiming to cover not only all variations of the phenomena that have been implemented, but also the combinations of different phenomena).

The grammar is implemented on the *Linguistic Knowledge Builder* (LKB) system (Copestake, 2002), an interactive grammar development environment for typed feature structure grammars, which includes a parser and generation, visualization tools for all relevant data structures, and a set of specialized debugging facilities, and it is grounded in the theoretical framework of *Head-driven Phrase Structure Grammar* (HPSG) (Pollard and Sag, 1987; Pollard and Sag, 1994), a constraint-based, lexicalist approach to grammatical theory where all linguistic objects (i.e. words and phrases) are represented as typed feature structures. The SRG uses *Minimal Recursion Semantics* (MRS) (Copestake et al., 2006) for the semantic representation. MRS is not a semantic theory in itself, but a kind of meta-level which has been defined for describing semantic structures. Using unification of typed features structures, MRS assigns a syntactically flat semantic representation to linguistic expressions.

The basis of the development of the SRG is the LinGO Grammar Matrix, an open-source starter-kit for rapid development of broad-coverage HPSG grammars compatible with the LKB system which supplies (1) the necessary configuration files for an LKB grammar development environment, and (2) the basic grammar types and rules (Bender et al., 2002; Bender and Flickinger, 2005).[1]

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 (Flickinger, 2002), German (Crysmann, 2005), and Japanese (Siegel and Bender, 2002), as well as smaller grammars for French, Korean (Kim and Yangs, 2003), modern Greek (Kordoni and Neu, 2005), Norwegian (Hellan and Haugereid, 2005), and Portuguese (Branco and Costa, 2008).[2]

## 2. Architecture

We have developed a hybrid architecture which integrates shallow processing functionalities – morphological analysis, and Named Entity (i.e. proper names, dates, numbers, ratios, currency, and physical magnitudes) recognition and classification – into the parsing process. See Figure 1.

Before parsing input sentences with the LKB system, raw text is pre-processed by the FreeLing toolkit, an open-source language analysis tool suite performing shallow processing functionalities (Atserias et al., 2006).[3] Our system plugs the FreeLing tool into the system by means of the LKB *Simple PreProcessor Protocol* (SPPP),[4] which assumes that a preprocessor runs as an external process to the LKB system. (1) is the output from the SPPP for the input *"gato"* (cat).

(1)
```
<segment>
 <token form="gato" from="0" to="1">
  <analysis stem="gato">
   <rule id="NCMS" form="gato"/>
  </analysis>
 </token>
```

[1] The Grammar Matrix is accessible through a web-based customization system: http://www.delph-in.net/matrix/customize/matrix.cgi.

[2] See http://www.delph-in.net/.

[3] The FreeLing toolkit may be downloaded from http://www.lsi.upc.edu/˜nlp/freeling.

[4] See http://wiki.delph-in.net/moin/LkbSppp.

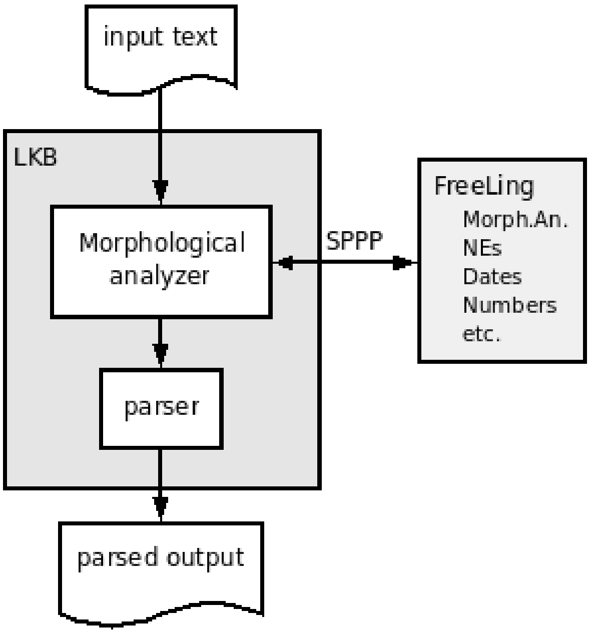


Figure 1: System architecture.

The advantage of our hybrid architecture is that it allows us to release the parser from certain tasks that may be robustly, efficiently, and reliably dealt with by shallow external components, thus making the whole system more adequate to deal with real world text we find in application contexts.

## 3. The Spanish Resource Grammar

### 3.1. Linguistic components and coverage

To parse a sentence, an LKB grammar requires three basic components: inflectional rules, a lexicon, and syntactic rules. This section describes these components of the SRG.

#### 3.1.1. Inflectional rules

The inflectional rules in the LKB system perform the morphological analysis of the words in the input sentences. Since we use an external morphological analyzer, the SRG does not need a morphology component, instead we use the inflectional rule component to convert the PoS tags associated to full-forms and produced by the FreeLing tool into feature structures. (2) shows the rule which converts the FreeLing PoS tag NCMS (Noun Common Masculine Singular) into a feature structure.

```
(2) ncms :=
    %suffix ()
    [ SYNSEM.LOCAL[ CAT.HEAD noun,
                    AGR.PNG[ PN 3sg,
                             GEN masc ]]].
```

We also use the inflectional component to integrate into the grammar the output of the NE recognition and classification module of FreeLing, which identifies the different instances of a given NE type and assigns them a tag. We assume that all instances of the same type can be dealt with a single lexical entry in the lexicon of the grammar. In order to identify the appropriate entry for each NE type, we use the inflectional rules to identify the tags associated to type instances with the attribute STEM, as we show in (3) with the rule dealing with the tag 'W' that FreeLing assigns to dates. This strategy provides a complete coverage of NEs to the grammar.

```
(3) w :=
    %suffix ()
    [ STEM < "w" > ].
```

#### 3.1.2. The lexicon

The lexicon component in the LKB system contains the lexical entries of the grammar. In the SRG, each lexical entry consists of a unique identifier, a lexical type (one among about 500 types,defined by a type hierarchy of around 5,000 types), an orthography and a semantic relation.[5] (4) shows the lexical entry for the noun *"gato"* (cat).

```
(4) gato_n := n_-_c_le &
    [ STEM < "gato" >
      SYNSEM.LKEYS.KEYREL "_gato_n_rel" ].
```

Lexical types represent the type of words that we have in our lexicon and they are defined by a multiple inheritance type hierarchy. Lexical subtypes for nouns are basically distinguished on the basis of valence information and the mass/countable/uncountable distinction. Adjectives are cross-classified according to their position within the NP, whether they are predicative or non-predicative, whether they are gradable or not, whether they are intersective or scopal, whether they are positive, comparative, or superlative, and according to their subcategorization restrictions. Leaving apart closed classes of adverbs (i.e. deictic, relative, interrogative and degree adverbs), we distinguish scopal and intersective adverbs, which in turn have subtypes specifying whether they may co-occur with degree adverbs and their position. Finally, main verbs are distiguished by their valence features SUBJ(ect) and COMP(lement)S, so we have subtypes for impersonal verbs taking no subject, verbs taking verbal subject, and verbs taking nominal subjects, and subtypes according to the number and category of the elements in the COMPS-list.[6]. Table 1 shows the number of types we have for each open word type and for closed word classes.

| | Number of types |
|---|---|
| verbs | 210 |
| nouns | 72 |
| adjectives | 33 |
| adverbs | 45 |
| closed categories | 119 |

Table 1: Number of lexical types

The SRG has a full coverage lexicon of closed word classes (pronouns, determiners, prepositions and conjunctions) and it contains 50,852 lexical entries for open word classes. Table 2 shows the number of lexical entries we have for each

[5] The attribute SYNSEM.LKEYS.KEYREL provides a shortcut to the semantic relation in RELS with highest scope (see Section 3.2) and it is only used in the lexicon.

[6] Details on the lexical types of the SRG may be found in (Marimon et al., 2007).

open word type and for closed word classes.[7] The grammar also has 64 lexical rules to perform valence changing operations on lexical items that include, for instance, movement and removal of complements, and, in that, they reduce the number of lexical entries to be manually encoded in the lexicon.

| | Number of entries |
|---:|---:|
| verbs | 7,871 |
| nouns | 27,950 |
| adjectives | 10,420 |
| adverbs | 4,131 |
| closed categories | 1,075 |

Table 2: Number of lexical entries

### 3.1.3. Syntactic rules

The syntactic rules in the LKB system are phrase structure rules that combine words and phrases into larger constituents and compositionally build up their semantic representation. The SRG has 191 phrase structure rules.

With these linguistic resources, the range of linguistic phenomena that the SRG 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, noun ellipsis, and coordination, among others.

## 3.2. Grammar output

The annotation of each parsed sentence in an LKB grammar simultaneously represents two different descriptive levels: (i) a traditional phrase structure tree, and (ii) an MRS semantic representation. Thus, for an input sentence such as *"el gato come pescado"* (the cat eats fish), the output of the SRG is as we show in Figure 2 and Figure 3.

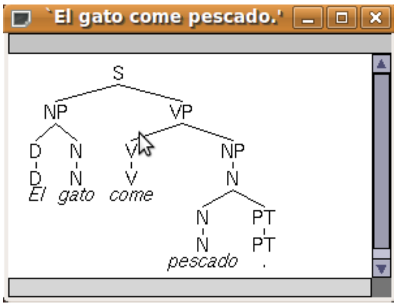


Figure 2: Phrase structure tree.

In the phrase structure tree each node is labeled with a set of atomic labels of the type 'S', 'VP', 'V', 'NP', etc. In addition, each tree node has an identifier of the grammar rule or lexical entry which has been used to build up that node tree. These identifiers, and the feature structures that define them, may be displayed by placing the cursor on the tree nodes. Figure 4 partially shows the identifier and the feature structures of the grammar rule `sp-hd_constr`, which attaches the specifier *"el"* to the noun *"gato"* and builds the NP node *"el gato"* (the cat).

The MRS semantic representation consists of: 1) a list of semantic relations (RELS), each with a "handle" (LBL) (used to express scope relations) and one or more roles (ARG0, ARG1,...). Relations are classified according to the number and type of arguments; 2) a set of handle constraints (HCONS), reflecting syntactic limitations on possible scope relations among the semantic relations, and 3) a group of distinguished semantic attributes of a linguistic sign. These attributes are: LTOP – the local top handle, and INDEX – the salient nominal instance or event variable introduced by the lexical semantic head.

# 4. Evaluation

Hand-built test suites are traditionally used to test and evaluate linguistically-motivated computational grammars.

When test suites meet the demands for systematicity and exhaustivity, they provide a fine-grained diagnosis of the grammar behavior in terms of coverage, overgeneration, and efficiency, and, for example, they are crucial to detect unintended interactions in the linguistic resources. However, hand-built test suites hardly present combinations of different phenomena showing the real world language complexity.

From the point of view of writing a large-coverage grammar, it is needed to evaluate its behavior against the combinations of different phenomena. Combining all different phenomena could lead to a combinatorial explosion; besides, not every combination of phenomena produces grammatical sentences or shows interesting cases.

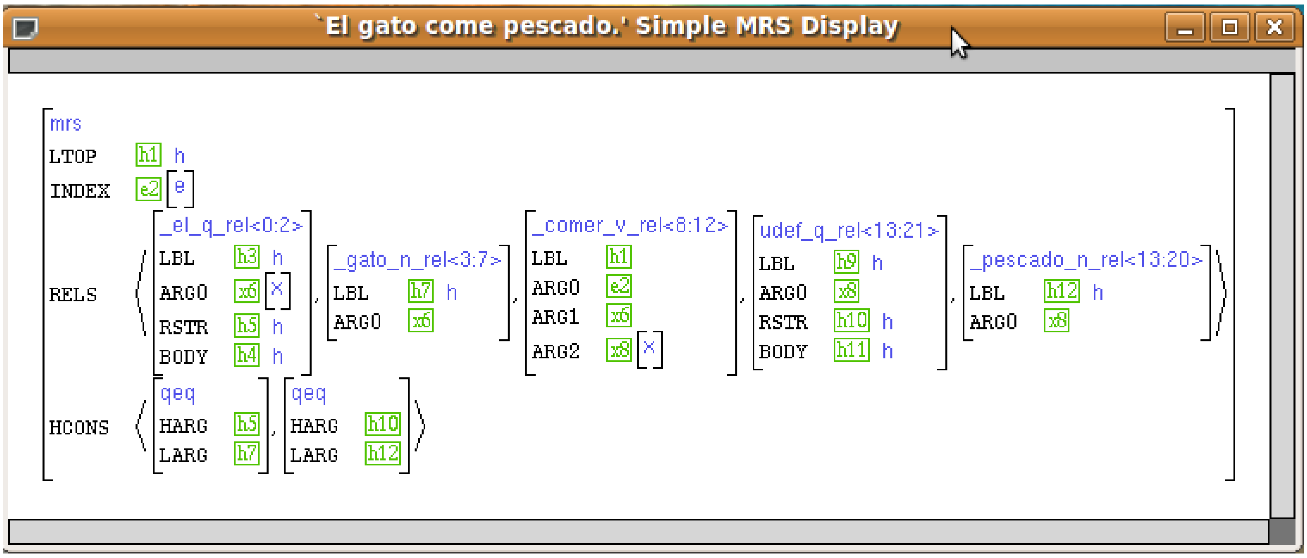
To evaluate the coverage of the SRG, we have used a sentences from newspaper texts from the AnCora corpus (Taulé et al., 2008). Table 3 shows the average of sentences we successfully parsed.

| sentence length | Number of sentences | parsed sentences |
|---:|---:|---:|
| 1-10 | 2,359 | 1,751 |
| 11-20 | 4,117 | 2,236 |
| 21-30 | 4,064 | 1,574 |
| 31-40 | 3,475 | 614 |
| 41-50 | 2,323 | 94 |
| 51-60 | 915 | 8 |
| +60 | 426 | 0 |
| total | 17,679 | 6,277 |

Table 3: Grammar results.

Parsing failures are due to several reason. First, even though the grammar has a set of generic lexical entry templates for open classes to deal with unknown words, we decided to evaluate the SRG without them, therefore, many parsing failures are due to missing lexical entries (many of them are foreign words). Second, although we have developed indeed a large-coverage grammar, there are still some

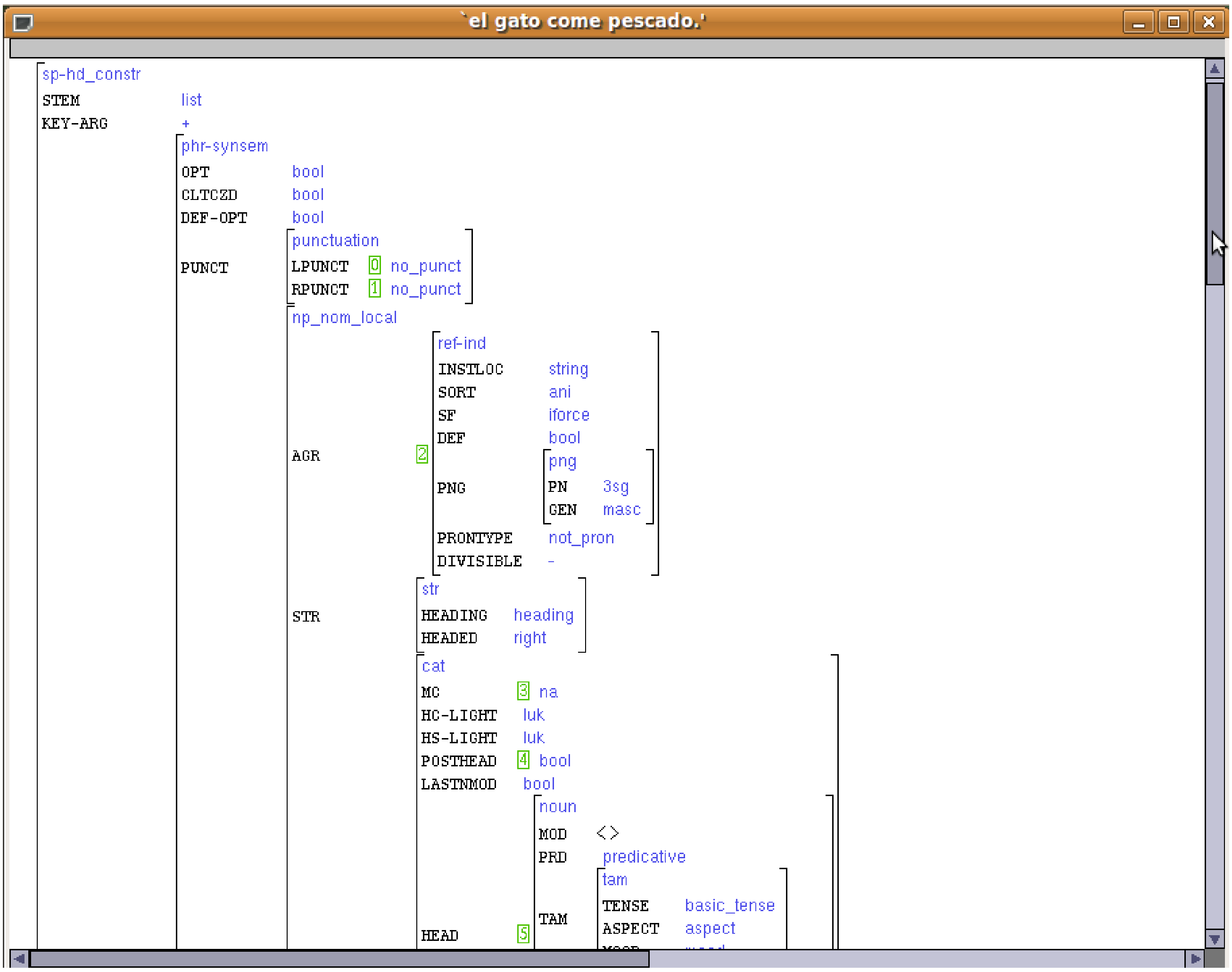
[7] The grammar also includes a set of generic lexical entry templates for open classes to deal with unknown words for virtually unlimited lexical coverage.



```
`El gato come pescado.' Simple MRS Display

mrs
LTOP    h1  h
INDEX   e2  e
              _el_q_rel<0:2>      _gato_n_rel<3:7>     _comer_v_rel<8:12>     udef_q_rel<13:21>      _pescado_n_rel<13:20>
RELS     <    LBL   h3  h       , LBL   h7  h        , LBL   h1             , LBL   h9  h          , LBL   h12  h           >
              ARG0  x6  x         ARG0  x6             ARG0  e2               ARG0  x8               ARG0  x8
              RSTR  h5  h                              ARG1  x6               RSTR  h10 h
              BODY  h4  h                              ARG2  x8  x            BODY  h11 h
              qeq           qeq
HCONS    <    HARG  h5   ,  HARG  h10  >
              LARG  h7      LARG  h12
```

Figure 3: MRS semantic representation.



```
`el gato come pescado.'

sp-hd_constr
STEM           list
KEY-ARG        +
               phr-synsem
               OPT         bool
               CLTCZD      bool
               DEF-OPT     bool
                           punctuation
               PUNCT       LPUNCT   0 no_punct
                           RPUNCT   1 no_punct
                           np_nom_local
                                       ref-ind
                                       INSTLOC     string
                                       SORT        ani
                                       SF          iforce
                                       DEF         bool
                           AGR       2             png
                                       PNG         PN     3sg
                                                   GEN    masc
                                       PRONTYPE    not_pron
                                       DIVISIBLE   -
                                     str
                           STR       HEADING    heading
                                     HEADED     right
                                     cat
                                     MC           3 na
                                     HC-LIGHT     luk
                                     HS-LIGHT     luk
                                     POSTHEAD     4 bool
                                     LASTNMOD     bool
                                                  noun
                                                  MOD    < >
                                                  PRD    predicative
                                                         tam
                                                  TAM    TENSE     basic_tense
                                     HEAD         5      ASPECT    aspect
```

Figure 4: sp-hd_constr grammar rule.

phenomena that need to be incorporated into the SRG, e.g. VP ellipsis. Third, the SRG certainly contains some errors, deficiencies, and unanticipated interactions in our linguistic modules. Fourth, some of the sentences, the long ones, reach the timeout. Finally, some of the parsed sentences are ungrammatical.

## 5. Conclusion and Future work

In this paper, we have described the basic linguistic components and linguistic coverage of an HPSG-based grammar for Spanish. We have also shown the results of the grammar on newspaper texts in terms of recall (i.e. coverage).

Besides improving the linguistic modules by extending the coverage of the grammar to deal with the phenomena which have not been implemented so far, and debugging the grammar to avoid errors and deficiencies, future work includes the development of a parse selection model over the hand-built grammar and to evaluate the SRG in terms of precision (i.e. the ratio assigned a unique and correct analysis) (Toutanova et al., 2005).

To reduce the ratio of parsing failures because of reaching the timeout, we are currently investigating the benefits of integrating further shallow processing functionalities into the parsing process that may reduce the ambiguity and, in that, avoid the construction of partial parses that do not contribute to the final parse.

## Acknowledgments

This work was funded by the *Ramón y Cajal* program of the Spanish *Ministerio de Ciencia e Innovación.*

## 6. References

J. Atserias, B. Casas, E. Comelles, M. González, L. Padró, and M. Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of LREC'06*, Genoa, Italy.

E.M. Bender and D. Flickinger. 2005. Rapid prototyping of scalable grammars: Towards modularity in extensions to a language-independent core. In *Proceedings of IJCNLP-05 (Posters/Demos)*, Jeju Island, Korea.

E.M. Bender, D. Flickinger, and SC.J. C.J. C.J. . Oepen. 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, COLING-02*, Tapei, Taiwan.

A. Branco and F Costa. 2008. *A Computational Grammar for Deep Linguistic Processing of Portuguese: LXGram, version A.4.1. Research Report TR-2008-17*. Universidade de Lisboa, Faculdade de Ciłncias, Departamento de Informatica.

U. Callmeier. 2000. Pet a platform for experimentation with efficient hpsg processing. In D. Flickinger, S. Oepen, J. Tsujii, and H. Uszkoreit, editors, *Natural Language Engineering (6)1 —Special Issue: Efficiency Processing with HPSG: Methods, Systems, Evaluation*, pages 99–108. Cambridge University Press.

A. Copestake, D. Flickinger, C.J. Pollard, and I. A. Sag. 2006. Minimal recursion semantics: an introduction. *Research on Language and Computation*, 3(4):281–332.

A. Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, CSLI lecture notes, number 110, Chicago.

B. Crysmann. 2005. Syncretism in German: a unified approach to underspecification, indeterminacy, and likeness of case. In *Proceedings of the 12th International Conference on Head-driven Phrase Structure Grammar*, Lisbon, Portugal.

D. Flickinger. 2002. On building a more efficient grammar by exploiting types. In D. Flickinger, S. Oepen, J. Tsujii, and H. Uszkoreit, editors, *Collaborative Language Engineering*, pages 1–17. Stanford: CSLI Publications.

L. Hellan and P Haugereid. 2005. Norsource - an excercise in the matrix grammar building design. In Emily M. Bender, Dan Flickinger, Frederik Fouvry, and Melanie Siegel, editors, *A Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI*, Vienna, Austria.

J-B. Kim and J. Yangs. 2003. Korean phrase structure grammar and its implementations into the lkb system. In *Proceedings of the 17th Pacific Asia Conference on Language, Information, and Computation*.

V. Kordoni and J. Neu. 2005. Deep analysis of modern greek. In Jong-Hyeok Lee et al. Keh-Yih Su, Jun'ichi Tsujii, editor, *Lecture Notes in Computer Science, Vol 3248*, pages 674 – 683. Springer-Verlag Berlin Heidelberg.

M. Marimon, N. Seghezzi, and Nuria Bel. 2007. An open-source lexicon for spanish. *Procesamiento del Lenguaje Natural*, 39:131–137.

S. Oepen and J. Carroll. 2000. Performance profiling for parser engineering. In D. Flickinger, S. Oepen, J. Tsujii, and H. Uszkoreit, editors, *Natural Language Engineering (6)1 —Special Issue: Efficiency Processing with HPSG: Methods, Systems, Evaluation*, pages 81–97. Cambridge University Press.

C.J. Pollard and I.A. Sag. 1987. *Information-based Syntax and Semantics. Volume I: Fundamentals*. CSLI Lecture Notes, Stanford.

C.J. Pollard and I.A. Sag. 1994. *Head-driven Phrase Structure Grammar*. The University of Chicago Press and CSLI Publications, Chicago.

U. Schäfer. 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.

M. Siegel and E.M. Bender. 2002. Efficient deep processing of japanese. In *3rd Workshop on Asian Language Resources and International Standardization, COLING-02*, Tapei, Taiwan.

M. Taulé, M.A. Martí, and M. Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC-2008*, Marrakech, Morocco.

K. Toutanova, C.D. Manning, D. Flickinger, and S Oepen. 2005. Stochastic hpsg parse disambiguation using the redwoods corpus. In *Journal of Logic and Computation*.