Procesamiento del lenguaje Natural, nº 41 (2008), pp. 21-28 recibido 7-05-2008; aceptado 16-06-2008

Dependency Grammars in Freeling

Gramáticas de Dependencia en Freeling

Jordi Carrera
Univ. Politécnica de Cataluña
Dep. Lenguajes y Sistemas
Campus Nord UPC, C/ Jordi Girona 1-3
jcarrera@lsi.upc.edu

Irene Castellón
Universidad Barcelona
Departament de Lingüistica General,
Gran Via de les Corts Catalanes 585
icastellon@ub.edu

Marina Lloberes
Universidad Barcelona
Dep. de Lingüistica General,
Gran Via de les Corts
Catalanes 585
marina.lloberes@ub.edu

Lluís Padró
Univ. Politécnica de Cataluña
Dep. Lenguajes y Sistemas
Campus Nord UPC,
C/ Jordi Girona 1-3
padro@lsi.upc.edu

Nevena Tinkova
Universidad Barcelona
Dep. de Lingüistica General,
Gran Via de les Corts
Catalanes 585
nevenatinkova@ub.edu

Resumen: En el marco del área del PLN, obtener análisis sintácticos profundos de manera automática es indispensable de cara a desarrollar aplicaciones que puedan hacer uso de representaciones semánticas de cualquier nivel. Uno de los objetivos del proyecto KNOW es poner a disposición de la comunidad científica gramáticas de segmentación profunda de amplia cobertura. En este artículo presentamos la implementación en el entorno FreeLing de las gramáticas del castellano, catalán e inglés, lenguas que, junto con el vasco, constituyen las lenguas objeto de interés del proyecto KNOW.

Palabras clave: PLN, análisis automático, análisis profundo, gramática de análisis, representación semántica, catalán, castellano, español, inglés

Abstract: Automatic deep parsing is necessary for any NLP applications requiring a certain level of semantic representation. One of the goals of the KNOW project is the development of wide-coverage deep parsing grammars whose outcome will be open to the scientific community. In this article we present a implementation of Spanish, Catalan and English grammars in the FreeLing environment. These three languages, together with Basque, are those we work on in KNOW.

Keywords: NLP, automatic parsing, deep parsing, parsing grammar, semantic representation, Catalan, Spanish, English

ISSN: 1135-5948

1. Some Words on Dependency Parsing

Automatic deep parsing is necessary for any NLP applications requiring some level of semantic representation. Although for some languages, such as English, there are several resources, such as Minipar (Lin, D. 1998), VISL (Bick, E. 2006), Connexor (Jarvinen, T. et al 1998) or Link Parser (Sleator, D. et al 1993), few broad-coverage grammars exist for Spanish and Catalan that deliver consistently good quality and can be efficiently embedded in NLP applications.

One of the goals of the KNOW project is the development of wide-coverage, deep parsing grammars whose outcome will be open to the scientific community. FreeLing (Atserias, J. et al 2006) includes a module for rule-based dependency parsing, named TXALA (Atserias, J. et al 2005). This module has been developed in the framework of OpenTrad, an Open-Source Machine Translation project funded by the Spanish Industry Ministry which aims at developing transfer translators for all official languages in Spain (Spanish, Catalan, Galician, and Basque), as well as English.

Observing the results of extensive coverage analysers for Spanish (Bick, E., 2006; Ferrández, A. et al 2000; Marimon, M. et al 2007; Tapanainen, P., 1996,), although in many cases the analysis is correct, there are some shortcomings such as the treatment of discontinuous constituents, infinitive clauses, the doubling of arguments in syntactic realization and the detection of multiword expressions.

On the other hand, these analysers are not open-source: Connexor grants a licence to researchers, but Hispal, which is the most refined, provides only parsed texts. This is why we believe it both a good idea and a necessary endeavor to create wide-coverage, open-source grammars for English, Catalan and Spanish.

In this article we present the parsing grammar implemented for each of these three languages which, together with Euskera (Aranzabe, M. et al 2004; Bengoetxea, K. et al 2007), are those we are working on in the framework of the KNOW project.

The rest of the article is structured as follows: in Section 2, a brief description is given regarding recent improvements in the TXALA analyser. In Section 3, problems posed by deep syntactic analysis and resources needed to deal with them are succinctly described. Each of the grammars is also broadly examined in this section. Section 4 includes some comments concerning evaluation aspects and, finally, in section 5 we draw conclusions and trace out some ideas for further research.

2. Dependency Parsing with FreeLing

The TXALA parser is the last step in the FreeLing processing chain, and is preceded by:

- Sentence splitting
- Morphological analysis
- Shallow parsing

After the shallow parser produces sequences of subtrees (one for each chunk in the sentence), the dependency parser performs three actions:

1. *Completion of the tree sequence into a full parsing tree.*

This is done by means of manually defined rules. Each rule applies to a pair of consecutive subtrees, and is assigned a priority value. At each step, the rule with higher priority is applied, and the affected pair of consecutive subtrees is fused into a single subtree.

The linguist defining the rules can specify conditions on each subtree head regarding its form, lemma, PoS, or word class (word classes may be defined by the grammarian as lemmata lists). Conditions on the context where the pair of chunks appears can also be specified such that the rule does not apply if conditions are not met.

2. *Conversion of syntax tree to dependency tree.*

At each level, the head node (marked as such in the manually defined rules) is set as the parent of all the trees below it.

3. *Functional labelling of dependencies.*

After the parse has been converted to a dependency tree, each dependency is then labelled with its syntactic function. Another set of rules is applied where conditions are stated on both head and dependent nodes. Conditions range from morphosyntactic checks (v.gr. lemma, relative position) to semantic properties (v.gr. predefined classes, WordNet

semantic files, EuroWordNet top-ontology features).

The version of the parser presented in this paper contains several improvements with respect to the version described in (Atserias, J. et al 2005). As regards tree completion rules:

- Extension of the repertory of subtree-fusion operations.
- Possibility of specifying form, lemma, PoS or word class conditions on subtrees.
- Possibility of specifying context conditions (stated as labels corresponding to subtrees).
- Defining word classes via lists in external files.

Regarding dependency labelling rules, new conditions on headwords bounded by dependencies are allowed, including:

- EWN Top Ontology properties
- WN semantic file
- Synonyms
- Hypernyms' synonyms

3. Deep parsing

When carrying out full syntatic analysis, sentences must be assigned some sort of semantic representation (more than one if ambiguous). A study carried out on data from Spanish concerning difficulties stemming from deep analysis (Tinkova, N. et al 2007) showed that the most complex phenomena to be solved were coordination, prepositional phrase attachment, inversion or constituent displacement, distinguishing between arguments and adjuncts and parsing subordinate clauses.

From a lexicalist standpoint, and as regards prepositional phrase attachment, crucial knowledge is provided by lexical heads. This kind of knowledge can be integrated in the form of a repertoire of syntactico-semantic structures (i.e. diathesis schemes) containing possible combinations of lexical heads with satellites.

Concerning coordination, this is a syntactic phenomenon with which we have dealt only partly and which requires extreme inter-rule synchronization. Complex situations arise in which coordinations must be resolved either before noun phrases (e.g. to create a compound subject) or after noun phrases and verb phrases and before sentence rules (e.g. not to create a compound subject but to coordinate two continguous sentences). Coordinated elements must be abstracted from, and context of the conjunction taken into account in order to prioritize some rule over the others.

3.1. Catalan dependency grammar

Catalan dependency grammar consists of a set of 2,914 rules, of which 2,565 complete the parse tree by creating dependencies and the remaining 349 label these dependencies.

Catalan grammar treats dependency recursion and dependency relations between a) phrases, b) clauses headed by conjunctions or relative pronouns, c) non-finite clauses and d) punctuation marks.

Verb subcategorization frames created on the basis of the Volem Multilingüe database (Fernández et al 2002) determine chunk selection and chunk labelling conditions for transitive verbs, verbs with a wh- clause as an argument, ditransitive verbs, intransitive verbs, verbs modified by one prepositional phrase argument, verbs modified by two prepositional phrase arguments, impersonal verbs, copulative verbs, verbs with a second predicate and motion verbs.

One problem arose during deep parsing regarding prepositional phrase attachment and, specifically, preposition *de* ('of' or 'from') attachment. In Catalan, *de*-headed prepositional phrases can modify either a noun phrase or a verb phrase. Adding information about both verb behaviour and context allowed to partly account for these problematic cases.

Sometimes, motion verbs code the source of the movement, which is expressed with a prepositional phrase headed by *de*. Therefore, defining a class of motion verbs allows dependency rules to be more fine-grained. However, given that *de*-phrases appear mostly after noun phrases, dependency rules for motion verbs yield only a partial solution. In this case, context conditions become essential to discriminate prepositional phrase attachment (a).

(a) Rule for attaching prepositional phrases to verb phrases:

grup-verb[mov] $_sn_$_grup-sp	sp-de	-	top_left
	17		

```
grup-verb[mov]    sp-de    -    top_left
$$_grup-sp    17
```

Rule for attaching prepositional phrases to nominal phrases:

```
sn    sp-de    -    top_left    -    20
```

Although *de*-phrases with a verbal head have a higher priority than *de*-phrases with a noun head (a), there is one exception to this rule: it is possible to attach a *de*-phrase to a nominal head after a motion verb. Thus, the rule accounting for this case ((b), below) has a higher priority than rules dealing with prepositional phrases attached to verb phrases (the first rule in (a)):

```
(b) grup-verb[mov] sp-de -        top_left
    $_sn_sp-de_$_grup-sp 21
    sn             sp-de -        top_left
    $$_sp-de_grup-sp    13
```

This way, prepositional attachment is solved in a wide range of cases. Figure 1 shows the analysis of sentence (c).

(c) Els operaris pugen les caixes de les eines del soterrani a la terrassa. [Catalan]

Workers are taking the toolboxes up from the cellar to the balcony. [English]

Another troublesome analysis obtained regarding wh- particles having multiple values. Some wh-particles introducing indirect questions can also appear as adverbial clauses, but whereas in the former case they must receive a direct object tag, in the latter case they must be labelled as verbal modifiers. In order to distinguish between both structures, a feasible solution consisted in listing verbs which usually take clausal direct objects (e.g. *verba dicendi*) and to create specific labelling rules for this type of verbs (d).

(d) Rules that assign direct object tag to wh-chunks:

```
grup-verb    dobj              d.label=subord
             d.side=right      p.class=que
inf          dobj              d.label=subord
             d.side=right      p.class=que
infinitiu    dobj              d.label=subord
             d.side=right      p.class=que
subord-ger   dobj              d.label=subord
             d.side=right      p.class=que
subord-part  dobj              d.label=subord
             d.side=right      p.class=que
```

Rules that assign verbal modifier tag to wh-chunks:

```
grup-verb    cc                d.label=subord
             d.lemma!=que|qui  p.class!=que
verb-pass    cc                d.label=subord
             d.lemma!=que|qui  p.class!=que
inf          cc                d.label=subord
             d.lemma!=que|qui  p.class!=que
infinitiu    cc                d.label=subord
             d.lemma!=que|qui  p.class!=que
subord-ger   cc                d.label=subord
             d.lemma!=que|qui  p.class!=que
subord-part  cc                d.label=subord
             d.lemma!=que|qui  p.class!=que
```

```
grup-verb/top/(pugen pujar VMIP3P0 -)
    [sn/ncsubj-subjecte/(opreraris opreraris NCMP000 -)
        [espec-mp/det/(Els el DA0MP0 -)]
    sn/dobj-objecte_directe/(caixes caixa NCFP000 -)
        [espec-fp/det/(les el DA0FP0 -)
        sp-de/ncmod/(de de SPS00 -)
        [sn/dobj-prep/(eines eina NCFP000 -)
        [espec-fp/det/(les el DA0FP0 -)]]]
    sp-de/iobj-prep/(de de SPS00 -)
        [sn/dobj-prep/(soterrani soterrani AQ0MS0 -)
            [j-ms/det/(el el DA0MS0 -)]]
    grup-sp/iobj-prep/(a a SPS00 -)
        [sn/dobj-prep/(terrassa terrassa NCFS000 -)
            [espec-fs/det/(la el DA0FS0 -)]]
F-no-c/ta/(. . Fp -)]
```

Figure 1. Textual output of example (c)

These rules allow indirect speech to be labelled with direct object tags (d) and adverbial clauses with wh-particles to be labelled with verbal modifier tags (e), as can be seen in Figure 2.

(e) El consell econòmic assenyala quan va començar la recessió econòmica. [Catalan]

Economic council points when economic recession began. [English]

3.2. Spanish dependency grammar

As for Spanish, TXALA dependency parser consists of 9,600 rules (9,245 parsing rules and 355 dependency rules) acting on a number of categories, such as noun, verb and prepositional phrases, pronouns, coordination, passive voice, punctuation and subordination. A rules applying to noun phrases is shown in (f). There can be seen, in order, the head, a modifier, a label denoting one child of the head, the function applied, no conditions and, finally, a priority index:

```
(f) sn    grup-sp sn    last_left    -    200
```

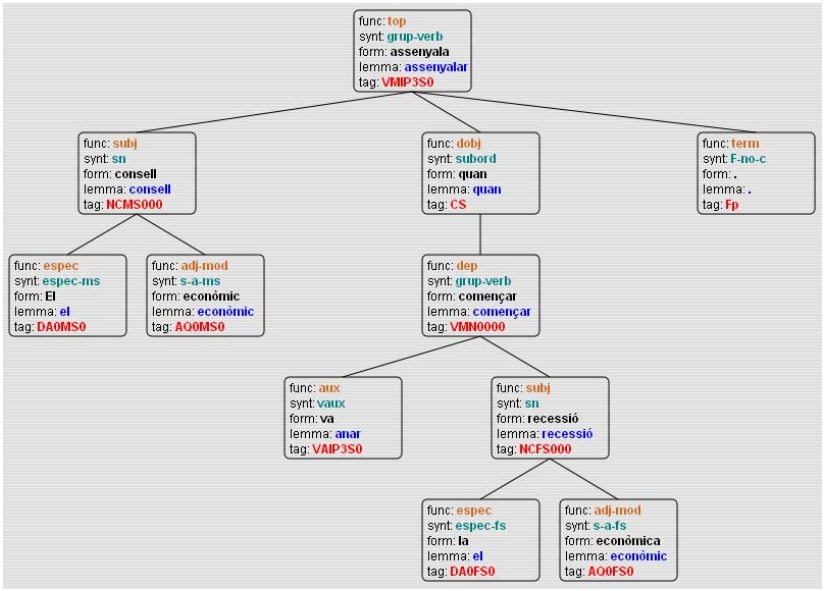

Figure 2. Parsing output of example (e)

Tag assignment is also carried out in Spanish grammar using labelling rules:

(g) grup-verb sp-obj
d.label=grup-sp| grup-sp-inf
d.side=right
d.lemma=a|al|para|hacia
p.class=mov

(h) grup-verb iobj
d.label=grup-sp|grup-sp-inf
d.side=right
d.lemma=a|para
p.class=ditr

(g) and (h) state that any prepositional phrase following a verb and including either of the prepositions *a, al* or *para*, be assigned *prepositional objec*t label (g) or *indirect object* label (h). Before this distinction was set up, whenever TXALA found a prepositional phrase introduced by any of the aforementioned prepositions, it invariably labelled it as an indirect object.

The Spanish grammar is being constantly updated. Incorporation of more complex subordination rules and verb subcategorization frames will result in increased coverage. Taking as a departure point the SenSem databank (Fernández, A. et al 2004), a ninefold typology of verbs was described (i.e. impersonal, intransitive, transitive, ditransitive, predicative, copulative, verbs followed by an argument wh-clause and verbs followed by either one or two argument prepositional phrases).

Solving prepositional phrase attachment is utterly necessary, for it is the cause of most syntactic misanalyses. As was the case for Catalan, attachment of *de*-phrases is of particular concern for Spanish as well, for these are able to act both as noun or as verb modifiers. Subcategorization information, together with context information, is expected to rule out wrong parses. In (i) and (j), PP-attachment rules are shown which have been enriched with contextual information. One screenshot of the output of the rule in (i) is given in Figure 3:

(i) sn $$_grup-verb sp-de 34 - top_left
(j) grup-verb[mov] $_sp-de_$ coor-sp 741 - top_left

As for function assignment, the parse in Figure 3 resulted from applying the rule in (k).

(k) sp-de prepos d.label=sn*

This rule labels the relation between the prepositional head and the head of the noun phrase immediately to its right.

3.3. English dependency grammar

Dependency rules for the English grammar amount to circa 1,340. They proceed in the following way: <*noun chunk, verb*> pairs are combined first. After that, rules apply recursively until another such pair is found,

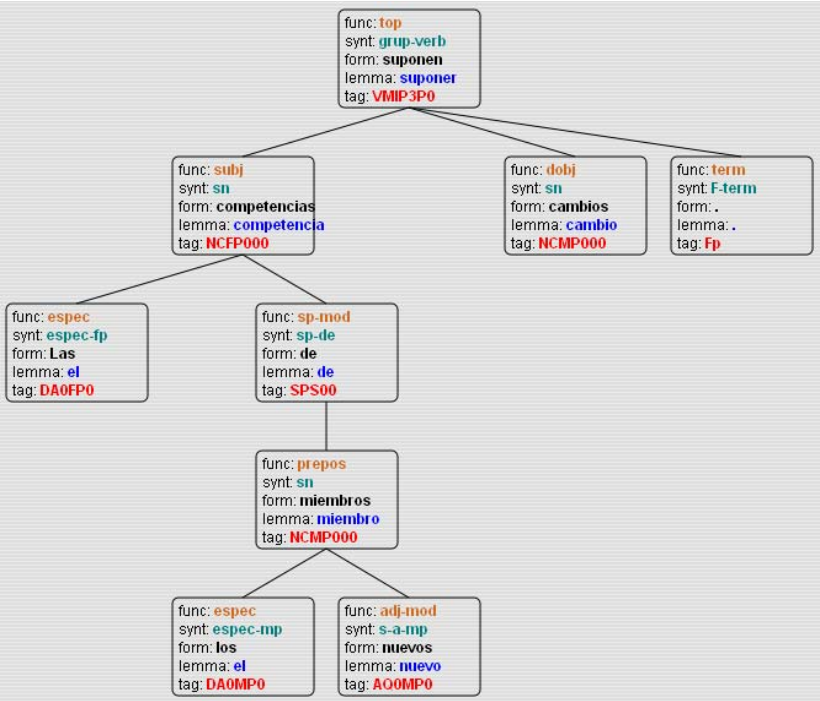

Figure 3. Graphical output of rule (i) application

and the process goes on iteratively until a full stop is found. Rules have been provided for all major kinds of clauses: declaratives, imperatives, interrogatives, completives, relatives, adverbial and existential. Analogously, separate verb phrase rules have been provided for intransitive, transitive and ditransitive sentences, including specific sets of rules for dealing with completive sentences. Sentences with ditransitive or higher valencies are treated formally as a subtype of transitive sentences.

A section was included in the grammar which contained a special kind of default dependency rules. These consisted broadly in heuristics intended to deal with relatively widespread cases of relatively unsystematic phenomena, i.e.:

- adjoining adverbs, prepositional phrases, etc. to their potential heads when in ambiguous syntactic positions, e.g. *I_x approached the man$_y$ on the chariot$_{x/y}$;*
- preventing main verbs from taking other clauses' direct objects as their subjects whenever they took as their subjects either other clauses having direct objects, or nominal subjects with embedded clauses, as in (l) and (m) (potentially mistakenly combined terms appear in bold):

l) The man who brought **the book was** interesting.

m) That he saw **the man was** uninteresting.

Besides, rules more often than not had to be multiplied. Since one given set of dependency rules would apply to a pair of chunks with a given priority, the same rules would not apply to plausible candidate expressions embedded in those chunks.

For instance, consider the example in (n), taken from Google:

n) The Astrakhan Region is capable of **making** products **having** an assured solvent demand in external market.

In (n), each *-ing* verb form takes its own direct object. The first two chunks, however *<making, products>*, should be grouped *after* the second pair of chunks *<having, demand>* has been grouped in turn. With a single set of rules, nonetheless, and since our algorithm proceeds from left to right, the leftmost *<participle, noun chunk>* pair is combined first, which results in the second modifier's being left behind.

This forced us to use several sets of multiplier rules performing virtually identical operations at different priorities, thus causing a remarkable grammar redundancy.

Another distinctive feature of English grammar as opposed to Spanish and Catalan grammars consisted in subordinate clauses'

lacking subordinating connectors for either completive clauses (e.g. *I've said he broke the car*) or relative clauses (e.g. *The man you saw was tall)*.

For these cases, long range rules were created that swept for series of concatenated <*noun chunk, verb*> pairs along with any noun phrases intervening in between (including null events). This yielded more reliable <*subject, verb*> dependencies extraction and, when conditioned on verbs taking completive sentences (v.gr. say, think, etc.), this heuristic proved to solve a fairly large number of ambiguities, which is remarkable taking into account its simplicity.

4. Evaluation

As yet, we have just finished the version 1.1 of Spanish, Catalan and English grammars, which we now intend to evaluate.

As for qualitative evaluation, a corpus has been created for each language. Text was extracted from newspaper articles and Internet websites. The corpora thus created vary in size: 50 sentences for Catalan, 100 for Spanish and 120 for English (this is dependent on the concept of *sentence* used). All of them contain a number of syntactic phenomena, v.gr. clausal embedding, coordination, different subcategorization frames, different phrase structures, etc. During development, corpora have been regularly analyzed as a testbed for the grammars, with analysis results guiding subsequent implementations.

At the time being, grammars are unable to tackle the following phenomena:

- Lexical coordination. Only some coordinations have been dealt with. We will keep expanding the number of cases covered with each successive update.
- Function assignment. When dependencies are assigned functional labels, information is necessary that the system is currently not sensitive to (e.g. PoS and morphological information for pronouns). New versions of the analyzer able to use this kind of data will have to be developed parallel to newer versions of the grammars.
- Constituent displacement has not been dealt with.
- Neither adverbial phrases or adverbial sentences have been treated (i.e. the system is unable to tell either adjuncts or arguments one from the other).

As for quantitative evaluation, so far we have been unable to carry out any such complete evaluation.

One of the main problems we face lies in the fact that analyses can differ substantially despite all of them being descriptively adequate.

In order to overcome this problem, corpora annotated according to the same formalism, in the same language and following the same grammatical criteria are required, which are usually unavailable.

Another problem lies in the fact that syntactic analysis takes as input the output of several previous processes (v.gr. multiword detection, named entity recognition, morphological labelling, etc.) Since none of these is completely error free, mistakes may take place at some point and keep then passing on to each subsequent step, all of which require an evaluation of their own prior to grammar evaluation proper.

For the languages we have been currently working with, there exist several corpora that we intend to use (3LB, WSJ, CONLL corpora). Our goal is to carry out evaluation using some subset of each of these, but we must still study whether the formalism and the criteria can be adapted to those utilized in the grammars presented here.

5. Conclusions and future work

In this article we have presented the version 1.1 of the Spanish, Catalan and English grammars to be used in the framework of the KNOW project in order to develop a broad-coverage deep parser to be distributed open-source. We have also presented the most recent update of the TXALA parser, which features a number of improvements over its predecessor.

There is ample room for improvement, however, specially as regards coordinations and constituent displacement for all three languages, and subcategorization frames for Spanish and English in particular. Likewise, subsequent improvement on the databases grammars rely on will also lead to performance increase.

On the other hand, coming up with evaluation metrics resting on a well-founded evaluation methodology constitutes another appealing line to deepen our present research.

Acknowledgements

This research has been funded by the Spanish Industry Ministry with the projects: KNOW (TIN MEC 2006-1549-C03-02), and OpenTrad (PROFIT FIT-350401-2006-5) and by a Predoctoral Scholarship FI-IQUC granted by the Generalitat de Catalunya to Nevena Tinkova (2004FI-IQUC1/00084).

References

Aranzabe M., J.M. Arriola, and A. Díaz de Ilarraza. 2004. Towards a Dependency Parser of Basque. Proceedings of the *Coling 2004 Workshop on Recent Advances in Dependency Grammar*.

Atserias, J., E. Comelles and A. Mayor. 2005. TXALA un analizador libre de dependencias para el castellano. Procesamiento del Lenguaje Natural, n. 35, p. 455-456.

Atserias, J., 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 Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06).

Bengoetxea K. and K. Gojenola. 2007. Desarrollo de un analizador sintáctico estadístico basado en dependencias para el euskera. *XXIII Congreso de la SEPLN.*.

Bick, Eckhard. 2006. A Constraint Grammar-Based Parser for Spanish. In: *Proceedings of TIL 2006 - 4th Workshop on Information and Human Language Technology*.

Briscoe, E., J. Carroll, and R. Watson. 2006. The Second Release of the RASP System. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, Sydney.

Fernández, A., P. Saint-Dizier, G. Vázquez, F. Benamara and M. Kamel. 2002. The VOLEM Project: a Framework for the Construction of Advanced Multilingual Lexic. *Proceedings of the Language Engineering Conference*

Fernández, A., G. Vázquez, I. Castellón (2004) "Sensem: base de datos verbal del español". G. de Ita, O. Fuentes, M. Osorio (ed.), *IX Ibero-American Workshop on Artificial Intelligence, IBERAMIA.*

Ferrández, A., M. Palomar and L. Moreno. 2000. "Slot Unification Grammar and anaphora resolution". In: *Recent Advances in Natural Language Processing*. Nicolas Nicolov & Ruslan Mitkov (eds). John Benjamins: Amsterdam & Philadelphia, pp. 155-166.

Jarvinen T. and P. Tapanainen. 1998. Towards an implementable dependency grammar. CoLing-ACL'98 workshop *'Processing of Dependence-Based Grammars'*, Kahane and Polguere (eds), p. 1-10, Montreal, Canada.

Lin D. 1998. Dependence-based Evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*.

Marimon,M. N.Bel and N.Seghezzi. 2007. Test Suite Construction for a Spanish Grammar, in Tracy Holloway King and Emily M. Bender (eds.) *Proceedings of the Grammar Engineering Across Frameworks (GEAF-2007) Workshop "CSLI Studies in Computational Linguistics ONLINE"* pp. 250-264

Sleator, D. and D.Temperley. 1993. Parsing English with a Link Grammar. *Third International Workshop on Parsing Technologies*.

Tinkova, N. and I. Castellón. 2007. A Comparative Study of Parsers Outputs for Spanish. *International Conference'07 Recent Advances in Natural Language (RANLP 2007).*