Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 87-94 recibido 10-04-14 revisado 15-07-14 aceptado 15-07-14

ParTes. Test Suite for Parsing Evaluation *

ParTes: Test suite para evaluación de analizadores sintácticos

Marina Lloberes
Irene Castellón
GRIAL-UB
Gran Via Corts Catalanes 585
08007 Barcelona
marina.lloberes@ub.edu
icastellon@ub.edu

Lluís Padró
TALP-UPC
Jordi Girona 1-3
08034 Barcelona
padro@lsi.upc.edu

Edgar Gonzàlez
Google Research
1600 Amphitheatre Parkway
94043 Mountain View - CA
edgargip@google.com

Resumen: En este artículo se presenta ParTes, el primer test suite en español y catalán para la evaluación cualitativa de analizadores sintácticos automáticos. Este recurso es una jerarquía de los fenómenos representativos acerca de la estructura sintáctica y el orden de argumentos. ParTes propone una simplificación de la evaluación cualitativa contribuyendo a la automatización de esta tarea.
Palabras clave: test suite, evaluación cualitativa, analizador sintáctico, español, catalán

Abstract: This paper presents ParTes, the first test suite in Spanish and Catalan for parsing qualitative evaluation. This resource is a hierarchical test suite of the representative syntactic structure and argument order phenomena. ParTes proposes a simplification of the qualitative evaluation by contributing to the automatization of this task.
Keywords: test suite, qualitative evaluation, parsing, Spanish, Catalan

1 Introduction

Qualitative evaluation in Natural Language Processing (NLP) is usually excluded in evaluation tasks because it requires a human effort and time cost. Generally, NLP evaluation is performed with corpora that are built over random language samples and that correspond to real language utterances. These evaluations are based on frequencies of the syntactic phenomena and, thus, on their representativity, but they usually exclude low-frequency syntactic phenomena. Consequently, current evaluation methods tend to focus on the accuracy of the most frequent linguistic phenomena rather than the accuracy of both high-frequent and low-frequent linguistic phenomena.

This paper takes as a starting point these issues related to qualitative evaluation. It presents ParTes, the first parsing test suite in Spanish and Catalan, to allow automatic qualitative evaluation as a complementary task of quantitative evaluation. This resource is designed to simplify the issues related to qualitative analysis reducing the human effort and time cost. Furthermore, ParTes provides a set of representative linguistic utterances based on syntax. The final result is a hierarchical test suite of syntactic structure and argument order phenomena defined by means of syntactic features.

2 Evaluation databases

Traditionally, two analysis methods have been defined: the quantitative analysis and the qualitative analysis. Both approaches are complementary and they can contribute to a global interpretation.

The main difference is that quantitative analysis relies on statistically informative data, while qualitative analysis talks about richness and precision of the data (McEnery and Wilson, 1996).

Representativeness by means of frequency is the main feature of quantitative studies. That is, the observed data cover the most frequent phenomena of the data set. Rare phenomena are considered irrelevant for a quantitative explanation. Thus, quantitative descriptions provide a close approximation of

* The resource presented in this paper arises from the research project SKATeR (Ministry of Economy and Competitiveness, TIN2012-38584-C06-06 and TIN2012-38584-C06-01). Edgar Gonzàlez collaborated in the ParTes automatization process. We thank Marta Recasens for her suggestions.

ISSN 1135-5948
© 2014 Sociedad Española para el Procesamiento del Lenguaje Natural

the real spectrum.

Qualitative studies offer an in-depth description rather than a quantification of the data (McEnery and Wilson, 1996). Frequent phenomena and marginal phenomena are considered items of the same condition because the focus is on providing an exhaustive description of the data.

In terms of analysis methods and databases, two resources have been widely used: corpora and test suites. Language technologies find these resources a reliable evaluation test because they are coherent and they are built over guidelines.

A corpus contains a finite collection of representative real linguistic utterances that are machine readable and that are a standard reference of the language variety represented in the resource itself (McEnery and Wilson, 1996). From this naive conceptualization, Corpus Linguistics takes the notion of representativeness as a presence in a large population of linguistic utterances, where the most frequent utterances are represented as a simulation of the reality and they are annotated according to the resource goals. That is why corpora are appropriate test data for quantitative studies.

On the other hand, test suites are structured and robust annotated databases which store an exhaustive collection of linguistic utterances according to a set of linguistic features. They are built over a delimited group of linguistic utterances where every utterance is detailed and classified according to rich linguistic and non-linguistic annotations (Lehmann et al., 1996). Thus, the control over test data and their detailed annotations make test suites a perfect guidance for qualitative studies.

Corpora have also been used in qualitative analysis, but they collect representative linguistic utterances by means of frequency rather than the representative linguistic utterances by means of exhaustiveness. Then, they are not the most appropriate tool for qualitative studies.

3 Existing test suites

Traditional test suites were simple collections of linguistic test cases or interesting examples. However, with the success of the NLP technologies, there was a real need for developing test suites based on pre-defined guidelines, with a deep structure, richly annotated and not necessarily developed for a particular tool (Flickinger, Nerbonne, and Sag, 1987). For this reason, the new generation of test suites are databases that cover the real needs of the NLP software evaluation (Lehmann et al., 1996).

The HP test suite (Flickinger, Nerbonne, and Sag, 1987) is an English and general purpose resource developed to diagnose and monitor the progress of NLP software development. The main goal of this test suite is to evaluate the performance of heuristic-based parsers under development. The suite contains a wide-range collection of linguistic examples that refer to syntactic phenomena such as argument structure verbs and verbal subcategorization among others. It also includes some basic anaphora-related phenomena. Furthermore, these phenomena are represented by a set of artificially constructed sentences and the annotations are shallow. This resource has a minimal internal classification since the suite organizes the test data under headings and sub-headings.

In order to step further, subsequent test suites have been developed as in-depth resources with rich structure and annotations. One of the groups of EAGLES proposes a set of guidelines for evaluating grammar checkers based on test suites (EAGLES, 1994). The test suite is a collection of attributes that allow to validate the quality of the functions of the evaluated tool. It is derived from a taxonomy of errors, where each error class is translated into a feature which is collected in the test suite. The final result is a classification of sentences containing an error, the corresponding sentence without the error, the name of the error and the guidelines for the correction process.

The TSNLP (Lehmann et al., 1996) is a multilingual test suite (English, French and German) richly annotated with linguistic and meta-linguistic features. This test suite is a collection of test items with general, categorial and structural information. Every test item is classified according to linguistic and extra-linguistic features (e.g. number and type of arguments, word order, etc.). These test items are also included in test sets by means of positive and negative examples. Furthermore, the TSNLP includes information about frequency or relevance for a particular domain.

In Spanish, a previous test suite exists

for NLP software evaluation, the SPARTE test suite (Peñas, Álvaro, and Verdejo, 2006). Specifically, it has been developed to validate Recognizing Textual Entailment systems and it is a collection of text and hypothesis pairs with true/false annotations. Although SPARTE and the presented ParTes in Spanish (ParTesEs) are resources for the same language, both test suites have been developed for different purposes which make both resources unique. With respect to the Catalan language, the version of ParTes in Catalan (ParTesCa) is the first test suite for this language.

4 The construction of ParTes

ParTes is a new test suite in Spanish and Catalan for qualitatively evaluating parsing systems. This test suite follows the main trends on test suite design, so that it shares some features with the EAGLES test suite (EAGLES, 1994) and the TSNLP (Lehmann et al., 1996).

Additionally, ParTes adds two new concepts in test suite design concerning how the data are classified and which data are encoded. The test suite is seen as a hierarchy where the phenomenon data are explicitly connected. Furthermore, representativeness is the key-concept in ParTes to select the phenomenon-testing data that configure the test suite.

The ParTes guidelines are created to ensure the coherence, the robustness and the easy implementation of this resource.

Specific purpose. While some test suites are general purpose like TSNLP, ParTes is a specific purpose test suite. Particularly, it is focused to validate the accuracy of the syntactic representations generated by parsers. For this reason, the test cases are related to syntactic phenomena and the test suite has been annotated with several syntactic features.

Test suite of syntactic phenomena. ParTes is not a simple collection of linguistic test cases nor a set of linguistic features, actually. This resource lists the syntactic phenomena that configure a language by a set of syntactic features.

For example, ParTes collects syntactic structures based on head-child relation. It also contains several features that syntactically define every phenomenon (e.g. the syntactic category of the head or the child, the syntactic relation with the node that governs it, etc.). Complementarily, every phenomenon is associated with a test case that corresponds to the linguistic utterance of the actual phenomenon described and that is used to evaluate the accuracy of the performance of the parser.

Hierarchy of syntactic phenomena. Previous test suites were a collection of test sentences, optionally structured (EAGLES and TSNLP). ParTes proposes a hierarchically-structured set of syntactic phenomena to which tests are associated.

Polyhedral hierarchy. Test suites can define linguistic phenomena from several perspectives (e.g. morphologic features, syntactic structures, semantic information, etc.). Because ParTes is built as a global test suite, it defines syntactic phenomena from two major syntactic concepts: syntactic structure and argument order (Section 5).

Exhaustive test suite. In order to evaluate NLP tools qualitatively, test suites list exhaustively a set of linguistic samples that describe in detail the language(s) of the resource, as discussed in Section 2. ParTes is not an exception and it contains an exhaustive list of the covered syntactic phenomena of the considered languages. However, some restrictions are applied to this list. Otherwise, listing the whole set of syntactic phenomena of a language is not feasible, and it is not one of the goals of the test suite's design.

Representative syntactic phenomena. As mentioned, lists of test cases need to be delimited because test suites are controlled data sets. Similarly to corpora development, the syntactic phenomena to be included in the test suite can be selected according to a certain notion of representativeness. Consequently, representative syntactic phenomena are relevant for testing purposes and they should be added in the test suite, whereas peripheral syntactic phenomena can be excluded. The next section (Section 5) details the definition of representativeness in ParTes and how it is implemented.

Rich annotations. Every syntactic phenomenon of ParTes is annotated with precise information that provides a detailed description and that allows the qualitative interpretation of the data. The annotations refer to

several linguistic and extra-linguistic features that determine the syntactic phenomena.

Controlled data. As argued in Section 2, there is a direct relation between qualitative evaluation, test suites and controlled test data. Because ParTes is a test suite for qualitative evaluation, there is a strong control over the test data and, specifically, the control is applied in a double way. The number of test cases is limited to human-processing size. The sentences of the test cases are controlled to avoid ambiguities and interactions with other linguistic utterances. For this reason, test cases are artificially created.

Semi-automatically generated. Linguistic resources usually have a high cost in terms of human effort and time. For this reason, automatic methods have been implemented whenever it has been possible. Manual linguistic description of the syntactic structure has been the main method to annotate the syntactic phenomena related to the structure. On the other hand, argument order annotations have been automatically generated and manually reviewed, using the automatization process of the SenSem corpus (Fernández and Vàzquez, 2012).

Multilingual. The architecture of this resource allows it to be developed in any language. The current version of ParTes includes the Spanish version of the test suite (ParTesEs) and the Catalan version (ParTesCa).

5 The results of ParTes

The final result of ParTes is an XML hierarchically and richly annotated test suite of the representative syntactic phenomena of the Spanish (ParTesEs) and Catalan (ParTesCa) languages. This resource is the first test suite for the evaluation of parsing software in the considered languages. It is freely available[1] and distributed under the Creative Commons Attribution-ShareAlike 3.0 Unported License.

ParTes is built over two kinds of information: the test suite module with the syntactic phenomena to be evaluated and the test data module with the linguistic samples to evaluate over. Since it is a polyhedral test suite, it is organized according to two major concepts in Syntax: structure and order. Table 1 gives the size of the current version of ParTes.

[1] http://grial.uab.es/descarregues.php

Section	ParTesEs	ParTesCa
Structure	99	101
Order	62	46
Total	**161**	**147**

Table 1: ParTes in numbers

5.1 Syntactic structure

The structure section is a hierarchy of syntactic levels where each level receives a tag and it is associated to a set of attributes that define several aspects about the syntactic structure. This section is placed between the `<structure></structure>` tags and it is organized into the following parts:

`<level>` It can be intrachunk (i.e. any structure inside a chunk) or intraclause (i.e. any connection between a clause marker and a grammatical category, phrase or clause).

`<constituent>` Phrase or clause that determines the nature of the constituent (e.g. noun phrase, verb phrase, infinitive clause, etc.). The head of the constituent corresponds to the parent node.

`<hierarchy>` Given two connected constituents, it defines which one occurs in the parent position and which other one in the child position.

`<realization>` Definition of the attributes of the head or child:

- `id`: Numerical code that identifies every `<realization>`.
- `name`: Name of the gramatical category, phrase or clause that occurs in head or child position (e.g. noun, pronoun, etc., as heads of noun phrase).
- `class`: Specifications about the gramatical category, the phrase or the clause that occurs in head or child position (e.g. a nominal head can be a common noun or a proper noun).
- `subclass`: Sub-specifications about the gramatical category, the phrase or the clause that occur in head or child position (e.g. a nominal head can be a bare noun).
- `link`: Arch between parent and child expressed by Part of Speech tags (e.g. the link between a nominal head and a modifying adjective is 'n-a').

```
<constituent name="verbphrase">
   <hierarchy name="head">
      <realization id="0001" name="verb" class="finite" subclass="default" link="null"
                   parent="salir" child="null" freq="null"
                   test="Saldrán"/>
      <realization id="0002" name="verb" class="nonfinite" subclass="default" link="null"
                   parent="viajar" child="null" freq="null"
                   test="Hubiesen viajado"/>
   </hierarchy>
   <hierarchy name="child">
      <realization id="0003" name="verb" class="auxiliar" subclass="haber" link="v-v"
                   parent="vender" child="haber" freq="0.010655" test="Habrán vendido la casa"/>
      <realization id="0004" name="verb" class="auxiliar" subclass="ser" link="v-v"
                   parent="acusar" child="ser" freq="0.010655"
                   test="Es acusada de robo"/>
      ...
      <realization id="0009" name="noun" class="null" subclass="default" link="v-n"
                   parent="romper" child="taza" freq="0.131629"
                   test="La taza se rompió"/>
      <realization id="0010" name="adjective" class="null" subclass="default" link="v-a"
                   parent="considerar" child="innovador" freq="0.010373"
                   test="Se considera una propuesta innovadora"/>
            ...
   </hierarchy>
</constituent>
```

Figure 1: Syntactic structure of the verb phrase in ParTesEs

- `parent`: Lemma of the upper level between the two nodes defined in `link` (e.g. in 'casa cara' - 'expensive house', the parent is 'casa').
- `child`: Lemma of the lower level between the two nodes defined in `link` (e.g. in 'casa cara' - 'expensive house', the child is 'caro').
- `freq`: Relative frequency in the AnCora corpus of the link between the two nodes defined in `link`.
- `test`: Linguistic test data that illustrates the syntactic structure.

For example, in the definition of verb phrase as `<constituent name="verbphrase">` (Figure 1), the possible grammatical categories, phrases and clauses that can form a verb phrase are detected and classified into two categories: those pieces that can be the head of the verb phrase (`<hierarchy name="head">`) and those that occur in child position (`<hierarchy name="child">`).

Next, the set of the possible heads of the verb phrase are listed in the several instances of `<realization>`. Furthermore, all the candidates of the child position are identified.

Every realization is defined by the previous set of attributes. In the Figure 1, in the case where the realization of one of the verb phrase children is a noun (`<realization ... name="noun".../>`), the frequency of occurrence of this link (i.e. the link of a verbal head and a nominal child, `link="v-n"`) is 0.131629 (in a scale between 0 and 1) and the test case to represent this structure is 'La taza se rompió' ('The cup broke'). Furthermore, the parent of the link 'v-n' of the test case is the lemma of the finite verb form 'rompió' (`parent="romper"`, 'to break') and the child of this link is the substantive 'taza' (`child="taza"`, 'cup'). The rest of this realization's attributes are empty.

As mentioned in Section 4, the most representative syntactic structure phenomena have been manually collected. In order to determine which phenomena are relevant to be included in ParTes, linguistic descriptive grammars have been used as a resource in the decision process. Thus, the syntactic phenomena that receive a special attention in the descriptive grammars can be considered candidates in terms of representativeness. In particular, the constructions described in *Gramática Descriptiva de la Lengua Española* (Bosque and Demonte, 1999) and in *Gramàtica del Català Contemporani* (Solà et al., 2002), for Spanish and Catalan respectively, have been included.

In addition, the representativeness of the selected syntactic phenomena is supported by the frequencies of the syntactic head-child re-

lations of the AnCora corpus (Taulé, Martí, and Recasens, 2008). These frequencies are automatically extracted and they are generalizations of the Part of Speech tag of both head and child given a link: all the main verb instances are grouped together, the auxiliaries are recognized into the same class, etc. Some frequencies are not extracted due to the complexity of certain constructions. For example, comparisons are excluded because it is not possible to reliable detect them by automatic means in the corpus.

The representation of the syntactic structures in ParTes follows the linguistic proposal implemented in FreeLing Dependency Grammars (Lloberes, Castellón, and Padró, 2010). This proposal states that the nature of the lexical unit determines the nature of the head and it determines the list of syntactic categories that can occur in the head position.

5.2 Argument order

Similarly to the syntactic structure section, the argument order schemas are also a hierarchy of the most representative argument structures that occur in the SenSem corpus. This section is organized in ParTes as follows:

`<class>` Number and type of arguments in which an order schema is classified. Three classes have been identified: monoargumental with subject expressed (`subj#V`), biargumental where subject and object are expressed (`subj#V#obj`), and monoargumental with object expressed (`V#obj`).

`<schema>` Sub-class of `<class>` where the argument order and the specific number of arguments are defined. For example, ditransitive verbs with an enclitic argument (e.g. '[El col·leccionista$_{subj}$] no [li$_{iobj}$] [ven$_{v}$] [el llibre$_{dobj}$]' - 'The collector to him do not sell the book') are expressed by the schema `subj#obj#V#obj` (Figure 2).

`<realization>` Specifications of the argument order schema, which are defined by the following set of attributes (Figure 2):

- `id`: Numerical code that identifies every `<realization>`.
- `func`: Syntactic functions that define every argument of the argument order schema. In Figure 2, the argument schema is composed by subject (`subj`), preverbal indirect object (`iobj`) and postverbal direct object (`dobj`).
- `cat`: Grammatical categories, phrases or clauses that define every argument of the argument order schema. For example, the three arguments of Figure 2 are realized as noun phrases (`np`).
- `parent`: Lemma of the upper level node of the argument order schema. In the case illustrated in Figure 2, the parent corresponds to the lemma of the verbal form of the test case (i.e. 'vendre'-'to sell').
- `children`: Lemmas of the lower level nodes of the argument order schema. In the test case of Figure 2, the children are the head of every argument (i.e. 'col·leccionista'-'collector', 'ell'-'him', 'llibre'-'book').
- `constr`: Construction type where a particular argument order schema occurs (active, passive, pronominal passive, impersonal, pronominal impersonal). In Figure 2, the construction is in active voice.
- `sbjtype`: Subject type of a particular argument order schema (semantically full or empty and lexically full or empty). The subject type of Figure 2 is semantically and lexically full so the value is `full`.
- `freq`: Relative frequency of the argument order schema in the SenSem corpus (Fernández and Vàzquez, 2012). The frequency of the ditransitive argument schema in Figure 2 is 0.005176, which means that the realization `subj#iobj#V#dobj` occurs 0.005176 times (in a scale between 0 and 1) in the SenSem corpus.
- `idsensem`: Three random SenSem id sentences have been linked to every ParTes argument order schema.
- `test`: Linguistic test data of the described realization of the argument order schema (in Figure 2, 'El col·leccionista no li ven el llibre'-'The collector to him do not sell the book').

The ParTes argument order schemas have been automatically generated from the syntactic patterns of the annotations of the SenSem corpus (Fernández and Vàzquez, 2012). Specifically, for every annotated verb

```
<class name="subj#V#obj">
   <schema name="subj#obj#V#obj">
      <realization id="0140" func="subj#iobj#v#dobj" cat="np#np#v#np" parent="vendre"
                   children="col·leccionista#ell#llibre" constr="active" sbjtype="full"
                   freq="0.005176" idsensem="43177#45210#52053"
                   test="El col·leccionista no li ven el llibre"/>
   </schema>
</class>
```

Figure 2: Argument order of ditransitive verbs in ParTesCa

in the corpus, the argument structure has been recognized. This information has been classified into the ParTes argument order schemas. Finally, the most frequent schemas have been filtered and manually reviewed, considering those schemas above the average. The total set of candidates is 62 argument order schemas for Spanish and 46 for Catalan.

5.3 Test data module

ParTes contains a test data set module to evaluate a syntactic tool over the phenomena included in the test suite. For the sentences in the data set, both plain text and syntactic annotations are available. The test data set is controlled in size: ParTesEs contains 94 sentences and ParTesCa is 99 sentences long. It is also controlled in terms of linguistic phenomena to prevent the interaction with other linguistic phenomena that may cause incorrect analysis. For this reason, test cases are artificially created.

A semi-automated process has been implemented to annotate ParTesEs and ParTesCa data sets. Both data sets have been automatically analyzed by the FreeLing Dependency Parser (Lloberes, Castellón, and Padró, 2010). The dependency trees have been mapped to the CoNLL format (Figure 3) proposed for the shared task on multilingual dependency parsing (Buchholz and Marsi, 2006). Finally, two annotators have reviewed and corrected the FreeLing Dependency Parser mapped outputs.

6 ParTes evaluation

To validate that ParTes is a useful evaluation parsing test suite, an evaluation task has been done. ParTes test sentences have been used to evaluate the performance of Spanish and Catalan FreeLing Dependency Grammars (Lloberes, Castellón, and Padró, 2010). The accuracy metrics have been provided by the CoNLL-X Shared Task 2007 script (Buchholz and Marsi, 2006), in which the syntactic analysis generated by the FreeLing Dependency Grammars (*system output*) are compared to ParTes data sets (*gold standard*).

The global scores of the Spanish Dependency Grammar are 82.71% for LAS[2], 88.38% for UAS and 85.39% for LAS2. Concerning to the Catalan FreeLing Dependency Grammar, the global results are 76.33% for LAS, 83.38% for UAS and 80.98% LAS2.

A detailed observation of the ParTes syntactic phenomena shows that FreeLing Dependency Grammars recognize successfuly the root of the main clause (Spanish: 96.8%; Catalan: 85.86%). On the other hand, subordinate clause recognition is not perfomed as precise as main clause recognition (Spanish: 11%; Catalan: 20%) because there are some limitations to determine the boundaries of the clause, and the node where it should be attached to.

Noun phrase is one of the most stable phrases because it is formed and attached right most of times (Spanish: 83%-100%; Catalan: 62%-100%). On the contrary, prepositional phrase is very unstable (Spanish: 66%; Catalan: 49%) because the current version of the grammars deals with this syntactic phenomenon shallowly.

This evaluation has allowed to determine which FreeLing Dependency Grammars syntactic phenomena are also covered in ParTes (*coverage*), how these syntactic phenomena are performed (*accuracy*) and why these phenomena are performed right/wrong (*qualitative analysis*).

7 Conclusions

The resource presented in this paper is the first test suite in Spanish and Catalan for parsing evaluation. ParTes has been de-

[2] *Labeled Attachment Score* (LAS): the percentage of tokens with correct head and syntactic function label; *Unlabeled Attachment Score* (UAS): the percentage of tokens with correct head; *Label Accuracy Score* (LAS2): the percentage of tokens with correct syntactic function label.

```
1       Habrán          haber       VAIF3P0     _       _       2       aux
2       vendido         vender      VMP00SM     _       _       0       top
3       la              el          DA0FS0      _       _       4       espec
4       casa            casa        NCFS000     _       _       2       dobj
5       .               .           Fp          _       _       2       term
```

Figure 3: Annotation of the sentence 'Habrán vendido la casa' ('[They] will have sold the house')

signed to evaluate qualitatively the accuracy of parsers.

This test suite has been built following the main trends in test suite design. However, it also adds some new functionalities. ParTes has been conceptualized as a complex structured test suite where every test case is classified in a hierarchy of syntactic phenomena. Furthermore, it is exhaustive, but exhaustiveness of syntactic phenomena is defined in this resource as representativity in corpora and descriptive grammars.

Despite the fact that ParTes is a polyhedral test suite based on the notions of structure and order, there are more foundations in Syntax, such as syntactic functions that currently are being included to make ParTes a more robust resource and to allow more precise evaluation tasks.

In addition, the current ParTes version contains the test data set annotated with syntactic dependencies. Future versions of ParTes may be distributed with other grammatical formalisms (e.g. constituents) in order to open ParTes to more parsing evaluation tasks.

References

Bosque, I. and V. Demonte. 1999. *Gramática Descriptiva de la Lengua Española*. Espasa Calpe, Madrid.

Buchholz, S. and E. Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164.

EAGLES. 1994. Draft Interim Report EAGLES. Technical report.

Fernández, A. and G. Vàzquez. 2012. Análisis cuantitativo del corpus SenSem. In I. Elorza, O. Carbonell i Cortés, R. Albarrán, B. García Riaza, and M. Pérez-Veneros, editors, *Empiricism and Analytical Tools For 21st Century Applied Linguistics*. Ediciones Universidad Salamanca, pages 157–170.

Flickinger, D., J. Nerbonne, and I.A. Sag. 1987. Toward Evaluation of NLP Systems. Technical report, Hewlett Packard Laboratories, Cambridge, England.

Lehmann, S., S. Oepen, S. Regnier-Prost, K. Netter, V. Lux, J. Klein, K. Falkedal, F. Fouvy, D. Estival, E. Dauphin, H. Compagnion, J. Baur, L. Balkan, and D. Arnold. 1996. TSNLP – Test Suites for Natural Language Processing. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 2, pages 711–716.

Lloberes, M., I. Castellón, and L. Padró. 2010. Spanish FreeLing Dependency Grammar. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 693–699.

McEnery, T. and A. Wilson. 1996. *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Peñas, A., R. Álvaro, and F. Verdejo. 2006. SPARTE, a Test Suite for Recognising Textual Entailment in Spanish. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 3878 of *Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, pages 275–286.

Solà, J., M.R. Lloret, J. Mascaró, and M. Pérez-Saldanya. 2002. *Gramàtica del Català Contemporani*. Empúries, Barcelona.

Taulé, M., M.A. Martí, and M. Recasens. 2008. AnCora: Multi level annotated corpora for Catalan and Spanish. In *6th International Conference on Language Resources and Evaluation*, pages 96–101.