Suitability of ParTes Test Suite for Parsing Evaluation

Marina Lloberes
U. de Barcelona
Barcelona, Spain
mllobesa8@alumnes.ub.edu

Irene Castellón
U. de Barcelona
Barcelona, Spain
icastellon@ub.edu

Lluís Padró
U. Politècnica de Catalunya
Barcelona, Spain
padro@cs.upc.edu

Abstract

Parsers have evolved significantly in the last decades, but currently big and accurate improvements are needed to enhance their performance. ParTes, a test suite in Spanish and Catalan for parsing evaluation, aims to contribute to this situation by pointing to the main factors that can decisively improve the parser performance.

1 Introduction

Parsing has been a very active area, so that parsers have progressed significantly over the recent years (Klein and Manning, 2003; Collins and Koo, 2005; Nivre et al., 2006; Ballesteros and Nivre, 2012; Bohnet and Nivre, 2012; Ballesteros and Carreras, 2015). However, nowadays significant improvement in parser performance needs extra effort.

A deeper and detailed analysis of the parsers performance can provide the keys to exceed the current accuracy. Tests suites are a linguistic resource which makes it possible this kind of analysis and which can contribute to highlight the key issues to improve decisively the Natural Language Processing (NLP) tools (Flickinger et al., 1987; EAGLES, 1994; Lehmann et al., 1996).

This paper presents ParTes 15.02, a test suite of syntactic phenomena for parsing evaluation. This resource contains an exhaustive and representative set of structure and word order phenomena for Spanish and Catalan languages (Lloberes et al., 2014). The new version adds a development data set and a test data set.

The rest of the paper describes the main contributions in test suite development (Section 2). Section 3 shows the characteristics and the specifications of ParTes. The results of an evaluation task of the FreeLing Dependency Grammars (FDGs) with verb subcategorization information added (Lloberes et al., 2010) using ParTes are discussed in Section 4. Finally, the main conclusions and future work are exposed (Section 5).

Features	HP	EAGLES	TSNLP
Domain	general	specific	general
Goal	parsing	grammar checkers	NLP software
Languages	English	English	English, German, French
Annotation	minimal	minimal	robust
Content	syntax	taxonomy of errors	(extra-)linguistic

Table 1: HP, EAGLES & TSNLP features

2 Test suite development

The main aim of qualitative studies is to offer empirical evidence about the richness and precision of the data, in comparison with quantitative studies which provide a view of the actual spectrum (McEnery and Wilson, 1996). For this reason, qualitative analysis are deep and detail-oriented, while quantitative analysis focus on statistically informative data. In the qualitative approach, representativeness of the studied phenomena focuses on exhaustiveness rather than frequency, which is the base of the quantitative approach. Both approaches are not exclusive because they contribute to build a global interpretation.

While corpora are a large databases of the most frequent linguistic utterances (McEnery and Wilson, 1996), test suites are controlled and exhaustive databases of linguistic utterances classified by linguistic features. These collections of cases are internally organized and richly annotated (Lehmann et al., 1996). Controlledness, exhaustiveness and detailedness properties allow these databases to provide qualitatively analyzed data.

They were developed in parallel with the NLP technologies. The more sophisticated the software became, the more complex the test suites evolved

Proceedings of the 14th International Conference on Parsing Technologies, pages 61–65,
Bilbao, Spain; July 22–24, 2015.

to be (Lehmann et al., 1996). From a collection of interesting examples, they transformed into deeply structured and richly annotated databases (Table 1), such as the HP test suite (Flickinger et al., 1987), the test suite developed by one of the groups of EAGLES (EAGLES, 1994), the TSNLP (Lehmann et al., 1996) and the corpus of unbounded depdendencies (Rimell et al., 2009).

Concerning the languages of this study, a test suite for Spanish was developed by Marimon et al. (2007). The goal of this test suite is to assess the development of a Spanish Head-Driven Phrase Structure Grammar and it offers grammatical and agrammatical test cases.

```
<level name="intrachunk">
   <constituent name="nounphrase">
      <hierarchy name="child">
         <realization id="0037"
            name="prepositionalphrase"
            class="noun" subclass="prepobj"
            link="n-s" freq="0.084357"
            parent_devel="recurso"
            child_devel="para"
            parent_test="libro"
            child_test="para"
            devel="Es un recurso para los
            alumnos"
            test="Los alumnos tienen un
            libro para la lectura"/>
      </hierarchy>
   </constituent>
</level>
```

Figure 1: Structure in ParTes. Example of the PP-attachment in the noun phrase

3 The ParTes test suite

This test suite is a hierarchically structured and richly annotated set of of syntactic phenomena for qualitative parsing evaluation available in Spanish (ParTesEs) and Catalan (ParTesCa) and freely distributed under the Creative Commons Attribution-ShareAlike 3.0 Unported License.[1]

The new release of ParTes (15.02) consists in the improvement of the linguistic data sets. Initially, ParTes included a test data module formed by sentences illustrating the syntactic phenomena of the test suite (Lloberes et al., 2014). The current version incorporates a set of linguistic data for development purposes that extends the capabilities of the test suite by allowing the parser development monitoring and a second iteration of the evaluation task.

This resource has been created following the main contributions in test suite design (Flickinger et al., 1987; EAGLES, 1994; Lehmann et al., 1996). The main feature shared with the existent test suites is the control over the data, which makes it possible to work as a qualitative evaluation tool. Furthermore, ParTes adds the concepts of complexity of the resource organization, exhaustiveness of the phenomena descriptions and representativity of the phenomena included.

ParTes is a test suite of syntactic phenomena annotated with syntactic and meta-linguistic information. The content has been hierarchically structured by means of syntactic features and over two major syntactic concepts (Figures 1 and 2): structure and word order.

It provides an exhaustive description of the syntactic phenomena, offering a detailed view of their features and their behavior. A selection of the representative phenomena has been performed, which allowed to delimit the number of cases preserving the control over the data.

The test suite has been semi-automatically generated, extracting automatically data from computational resources when available. Otherwise, written linguistic resources have been used to populate manually the resource. Its architecture makes it possible to extend the test suite to new languages, although the current version is available in two languages.

3.1 Test suite specifications

The current version contains a total of 161 syntactic phenomena in ParTesEs (99 relate to syntactic structure and 62 to word order) and a total of 145 syntactic phenomena in ParTesCa (99 concern to syntactic structure and 46 to word order).

The structure phenomena have been manually collected from descriptive grammars (Bosque and Demonte, 1999; Solà et al., 2002) and represented following the criteria of the FDGs (Lloberes et al., 2010). The selection of phenomena has been validated by the dependency links frequency of the AnCora Corpus (Taulé et al., 2008).

As Figure 1 shows, the first level of the hierarchy determines the *level* of the syntactic phenomenon (inside a chunk or between a marker and the subordinate verb). The second level expresses the phrase or the clause involved in the syntactic phenomenon (*constituent*) and the third level describes the position (*head* or *child*) in the *hierarchy*. Finally, a set of syntactic features describes the type of constituent observed (*realization*).

Specifically, the syntactic features of the *realization* concern to the grammatical category, the

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

```
<class name="subj#V">
   <schema name="subj#V">
      <realization id="0104"
         func="subj#v"
         cat="pron#v"
         parent="perdre"
         children="tot"
         constr="passive-pron"
         sbjtype="full"
         freq="0.001875"
         idsensem="45074#45239#48770"
         test="Tot s'ha perdut"/>
   </schema>
</class>
```

Figure 2: Word order in ParTes. Example of pronominal passive with particle 'se'

phrase or the clause that defines the structure phenomenon (*name*), its syntactic specifications (*class*, *subclass*), the arch between the parent and the child (*link*), the occurrence frequency of the link (*freq*) in the AnCora Corpus. Additionally, every phenomenon is identified with a numeric *id*.

For every syntactic structure phenomenon, two linguistic examples have been manually defined, one of them to be used for development purposes (*devel*) and the other one for testing purposes (*test*). The lemmas of the parent and the child of the exemplified phenomenon are also provided (*parent_devel*, *parent_test*, *child_devel*, *child_test*).

Word order in ParTes is semi-automatically built from the most frequent argument structure frames of the SenSem Corpus (Fernández and Vàzquez, 2014).

The hierarchy about the word order is structured firstly by the number and the type of arguments of the word order schema (*class*), as Figure 2 illustrates. Every class is defined by a set of *schemas* about the number of arguments and their order. The most concrete level (*realization*) describes the properties of the schema.

These properties refer to the syntactic function (*func*)[2] and the grammatical category (*cat*) of every argument of the schema. Furthermore, the type of construction (*constr*) where the schema occurs in and the type of subject (*sbjtype*) are provided. The occurrence frequency of the schema in the SenSem Corpus is associated (*freq*). In addition, a numeric *id* is assigned to every schema and a link to SenSem Corpus sentences with the same schema is created (*idsensem*).

Every schema recorded is exemplified with a sentence for testing purposes (*test*). For every test sentence, the lemmas of the *parent* and the *children* corresponding to the head of the arguments of the schema are added.

3.2 Description of the data sets

The development and the test data are built over the manually defined linguistic examples of the syntactic phenomena of ParTes.

The sentences have been automatically annotated by using the FDGs, so that a complete dependency analysis of the whole sentence is offered. The output has been reviewed manually by two annotators: a native in Spanish responsible for the annotation of ParTesEs and a native in Catalan who annotated the ParTesCa. A second manual revision has been performed: the Catalan annotator reviewed the ParTesEs annotated and the Spanish annotator reviewed the ParTesCa annotated guaranteeing the agreement between the annotations in both languages and preserving the quality of the annotation according to the criteria.

Up to the current version, the number of sentences referring to the syntactic structure are: 95 sentences in the ParTesEs development data set, 99 sentences in the ParTesEs test data set, 98 sentences in the ParTesCa development data set and 99 sentences in the ParTesCa test data set. The data sets are distributed in plain text format and in the CoNLL annotation format (Nivre et al., 2007).

4 Evaluation task

In order to test the usability of ParTes for parsing evaluation, it has been applied as a gold standard in an evaluation task of the FDGs. Particularly, the capabilities of the test suite have been tested for explaining the performance of FDG as regards the argument recognition since it still remains to be solved successfully (Carroll et al., 1998; Zeman, 2002; Mirroshandel et al., 2013).

The FDGs are the core part of the rule-based FreeLing Dependency Parser (Padró and Stanilovsky, 2012). They provide a deep and complete syntactic analysis in the form of dependencies. The grammars are a set of manually-defined rules that comple the structure of the tree (*linking rules*) and assign a syntactic function to every link of the tree (*labelling rules*) by means of a system of priorities and a set of conditions.

Two FDGs versions for both languages have been evaluated: a version without verb subcategorization classes (*Bare*) and a version with verb sub-

[2]Tagset: *adjt* - adjunct; *attr* - attribute; *dobj* - direct object; *iobj* - indirect object; *pobj* - prepositional object; *pred* - predicative; *subj* - subject.

Metric	ParTesEs		ParTesCa	
	Bare	Subcat	Bare	Subcat
LAS	77.57	79.66	79.41	81.80
UAS	88.21	88.21	88.24	88.24
LA	78.90	81.94	80.88	83.64

Table 2: Label Accuracy of FDG on ParTes

categorization classes (*Subcat*) extracted from the verbal frames of the SenSem Corpus (Fernández and Vàzquez, 2014). The system analysis built for every version of the grammars is compared to the ParTes analysis using the evaluation metrics of the CoNLL-X Shared Task (Nivre et al., 2007).[3]

According to the accuracy results (Table 2), the evaluation with ParTes shows that FDGs performance is medium-accuracy (near or above 80% in LAS). Both versions of the grammar in both languages perform in high-accuracy in terms of attachment (UAS), whereas they obtain medium accuracy on syntactic function labelling (LA). ParTes data highlight that the *Subcat* grammar scores better than the *Bare* grammar in LA, which is directly related to the addition of subcategorization classes, as stated in the following discussion.

A detailed observation reveals that ParTes sentences related to subcategorization are performed better in precision by *Subcat* rather than *Bare* (Table 3). Furthermore, the test data allows to show that subcategorization has more impact in the recognition of the majority of arguments (*dobj*, *pobj*, *pred*) and the subject (*subj*) than in the adjuncts (*adjt*) because the precision scores increment is higher. Subcategorization do not have an effect on the attribute (*attr*) because it can be solved lexically. The indirect objects (*iobj*) correspond to cases of dative clitic, which are solved by morphological information.

The integration of subcategorization information bounds the rules to the verbs included in the classes. Consequently, some cases may be not captured if the verb is not expected by the subcategorization classes as it happens in the prepositional object (*pobj*). For example, the prepositional argument of the sentence 'Ha creido en sí mismo' ('He has believed in himself') should be labelled as *pobj*, but the *adjt* tag is assigned because the verb 'creer' is not in any of the prepositional argument classes of the grammar. However, in the majority of types of arguments and the adjuncts the recall is maintained or increased (Table 4).

[3] *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* (LA): the percentage of tokens with correct syntactic function label; *Precision* (P): the ratio between the system correct tokens and the system tokens; *Recall* (R): the ratio between the system correct tokens and the gold standard tokens.

Tag	ParTesEs			ParTesCa		
	#	Bare	Subcat	#	Bare	Subcat
adjt	39	53.85	65.96	30	60.00	61.90
attr	28	88.89	83.87	20	90.00	78.26
dobj	39	65.31	73.81	42	74.51	86.96
iobj	7	100.00	100.00	3	100.00	75.00
pobj	11	23.68	37.50	13	45.83	60.00
pred	2	25.00	100.00	2	22.22	100.00
subj	51	93.02	93.02	43	87.88	90.62

Table 3: Precision scores of FDG on ParTes

Tag	ParTesEs			ParTesCa		
	#	Bare	Subcat	#	Bare	Subcat
adjt	39	35.90	79.49	30	50.00	86.67
attr	28	85.71	92.86	20	90.00	90.00
dobj	39	82.05	79.49	42	90.48	95.24
iobj	7	28.57	28.57	3	66.67	100.00
pobj	11	81.82	54.55	13	84.62	69.23
pred	2	50.00	50.00	2	100.00	50.00
subj	51	78.43	78.43	43	67.44	67.44

Table 4: Recall scores of FDG on ParTes

5 Conclusions

The new version of the ParTes test suite for parsing evaluation has been presented. The main features and the data sets have been described. In addition, the results of an evaluation task of the FDGs with ParTes data have been exposed.

The characteristics of the test suite made it possible to analyze in detail the causes of the performance improvement on the argument recognition of the FDGs including subcategorization information. Therefore, these results show that ParTes is an appropriate resource for parsing evaluation.

Currently, ParTes is extended to English following the methodology explained in this paper. In the upcoming releases, test and development sentences belonging to the word order will be incorporated in the ParTes data sets. Furthermore, we are exploring a systematic methodology to generate agrammatical variants of the existent sentences.

Acknowledgments

This work has been funded by the SKATeR project (Spanish Ministry of Economy and Competitiveness, TIN2012-38584-C06-06 and TIN2012-38584-C06-01).

References

M. Ballesteros and X. Carreras. 2015. Transition-based Spinal Parsing. In *Proceedings of CoNLL-2015*.

M. Ballesteros and J. Nivre. 2012. MaltOptimizer: A System for MaltParser Optimization. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

B. Bohnet and J. Nivre. 2012. A Transition-based System for Joint Part-of-speech Tagging and Labeled Non-projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

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

J. Carroll, G. Minnen, and T. Briscoe. 1998. Can Subcategorisation Probabilities Help a Statistical Parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*.

M. Collins and T. Koo. 2005. Discriminative Reranking for Natural Language Parsing. *Computational Linguistics*, 31(1).

EAGLES. 1994. Draft Interim Report EAGLES. Technical report, Expert Advisory Group on Language Engineering Standards.

A. Fernández and G. Vàzquez. 2014. The SenSem Corpus: an annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, 10(2).

D. Flickinger, J. Nerbonne, and I.A. Sag. 1987. Toward Evaluation of NLP Systems. Technical report, Hewlett Packard Laboratories. Distributed at the 24th Annual Meeting of the Association for Computational Linguistics (ACL).

D. Klein and C.D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*.

S. Lehmann, 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.

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

M. Lloberes, I. Castellón, L. Padró, and E. Gonzàlez. 2014. ParTes. Test Suite for Parsing Evaluation. *Procesamiento del Lenguaje Natural*, 53.

M. Marimon, N. Bel, and N. Seghezzi. 2007. Test-suite Construction for a Spanish Grammar. In *Proceedings of the GEAF 2007 Workshop*.

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

S.A. Mirroshandel, A. Nasr, and B. Sagot. 2013. Enforcing Subcategorization Constraints in a Parser Using Sub-parses Recombining. In *NAACL 2013 - Conference of the North American Chapter of the Association for Computational Linguistics*.

J. Nivre, J. Hall, J. Nilsson, G. Eryiğit, and S. Marinov. 2006. Labeled Pseudo-projective Dependency Parsing with Support Vector Machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.

L. Padró and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

L. Rimell, S. Clark, and M. Steedman. 2009. Unbounded Dependency Recovery for Parser Evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

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

M. Taulé, 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, Marrakesh*.

D. Zeman. 2002. Can Subcategorization Help a Statistical Dependency Parser? In *19th International Conference on Computational Linguistics*.