ParDeepBank: Multiple Parallel Deep Treebanking

Dan Flickinger[1], Valia Kordoni[3,2], Yi Zhang[2],
António Branco[4], Kiril Simov[5], Petya Osenova[5],
Catarina Carvalheiro[4], Francisco Costa[4], Sérgio Castro[4]
[1]Stanford University, USA, [2]DFKI GmbH, Germany,
[3]Humboldt University of Berlin, Germany,
[4]University of Lisbon, Portugal,
[5]Bulgarian Academy of Sciences, Bulgaria
danf@stanford.edu, evangelia.kordoni@anglistik.hu-berlin.de,
yizhang@dfki.de, antonio.branco@di.fc.ul.pt,
kivs@bultreebank.org, petya@bultreebank.org,
catarina.carvalheiro@di.fc.ul.pt, f.costa@di.fc.ul.pt,
sergio.castro@di.fc.ul.pt

Abstract

This paper describes the creation of an innovative and highly parallel treebank of three languages from different language groups — English, Portuguese and Bulgarian. The linguistic analyses for the three languages are done by compatible parallel automatic HPSG grammars using the same formalism, tools and implementation strategy. The final analysis for each sentence in each language consists of (1) a detailed feature structure analysis by the corresponding grammar and (2) derivative information such as derivation trees, constituent trees, dependency trees, and Minimal Recursion Semantics structures. The parallel sentences are extracted from the Penn Treebank and translated into the other languages. The Parallel Deep Bank (ParDeepBank) has potentially many applications: for HPSG grammar development; machine translation; evaluation of parsers on comparable data; etc.

1 Introduction

In this paper we present the initial version of ParDeepBank — a parallel treebank for three languages: English, Portuguese, and Bulgarian. The annotation of each sentence in the treebank is automatically analyzed by a deep HPSG grammar (Head-driven Phrase Structure Grammar: [24]) for the corresponding language. These grammars are implemented within the same linguistic theory and formalism, following a similar approach to grammar implementation. The correct analysis is selected manually in all of the three cases. The HPSG grammars for the three languages have different levels of development. Hence, they differ in their coverage. In order to process the whole set of sentences in the parallel treebank, we employ all the available NLP tools for each language and produce a comparable analysis for each sentence. The sentences are selected from PTB (PennTreebank) data and then translated to Portuguese and Bulgarian. Our parallel treebanking approach takes a starting point and motivation similar to that already adopted for the Czech-English PennTreebank Corpus[1].

Recently, a number of initiatives have been observed for constructing parallel treebanks. For example, in [19] the construction of a parallel Czech-Russian Dependency Treebank required smoothing of the syntactic schemes before handling the alignments on various levels. Another project, among others, is SMULTRON: a parallel treebank of English, German and Swedish [32]. In this case the annotation scheme is different for each treebank. The alignments are done on the sentence, phrase and word level. There are attempts for construction of parallel treebanks on a large scale and in a fully automated way [31], where large multilingual corpora are parsed with the respective language tools, and then models are created for alignments using small amounts of aligned data as well as complex features.

We consider our efforts to be innovative in a number of directions, including performing multilingual treebanking in the same semantic formalism (MRS: [13]), which ensures a deep logical representation for the syntactic analyses and ensures automatic alignments, which would make possible the comparison of semantic structures among parallel data in different languages. Our approach would facilitate the incorporation of further languages to the ParDeepBank, since it allows for a multilayered addition of the parser as well as the analysed data. The English DeepBank and ERG play the role of state-of-the-art pivots with respect to the core language phenomena coverage and best parsing abilities. The other grammars and treebanks aim at the established standards, but their development is supported by additional robustness mechanisms, such as dependency analyses plus rule-based projection to MRS in the case of the Bulgarian Grammar. The sentences, analyzed via a dependency parser, will be processed additionally by the BURGER grammar when it is extended appropriately. The current manually checked dependency analyses will be used for the selection of the correct analyses, produced by BURGER. Similarly, such analyses updates might be expected for the other languages, too. Our design is as follows: we use parallel data from the Wall Street Journal portion of the Penn Treebank, then parse it, using common NLP components, with each of the three grammars, which are represented in an identical formalism on syntactic and semantic grounds. Our initial goal is each sentence of WSJ corpus to be analyzed syntactically and semantically in MRS.

The structure of the paper is as follows: Section 2 introduces the idea behind the ParDeepBank; Section 3 describes the Resource Grammars for three languages: English, Portuguese and Bulgarian; Section 4 focuses on the process of sentence selection and the details of treebanking in English, Portuguese and Bulgarian; Section 5 outlines some preliminary features of the dynamic parallel treebanking; Section 6 concludes the paper and presents insights on future work.

[1] http://ufal.mff.cuni.cz/pcedt1.0/doc/PCEDT_body.html

2 The ParDeepBank

The PTB has emerged as the de facto standard for evaluation of different NLP analyzers for English including POS taggers, chunkers, and parsers (both constituent and dependency). Even for non-English-oriented analyzers there is a need for comparable methods of evaluation. The PTB is also widely used for creation of new resources like the Penn Discourse Treebank [25]. It is a natural step to reuse the same text in the process of creation of deep processed corpora. The treebanking for the ParDeepBank is built on the same approach used for the soon-to-be-released English DeepBank, which adopts the Redwoods treebanking approach of [22]. The annotation of a sentence starts with producing all possible analyses with respect to the English Resource Grammar (ERG [16]). The system calculates the set of binary discriminants which disambiguate between the different analyses of each sentence. These discriminants are used by the annotators to select the correct analysis. There are two cases when a given sentence is not included in the DeepBank: (1) when the ERG fails to produce any analysis of the sentence at all, and (2) when the annotator cannot find a correct analysis among the candidates. In both cases a modification of the ERG would be necessary in order to produce the required analysis. In the current version of the English DeepBank development, some 92% of all sentences in the WSJ section of the PTB receive one or more candidate analyses, and 82% of all sentences are successfully annotated with the correct analysis in the DeepBank.

The creation of ParDeepBank extends the work on the English DeepBank in the multilingual dimension. This effort is necessary for several reasons:

- *Comparable evaluation of NLP tools for several languages.* In many cases, NLP systems exploit hybrid architectures which include language-specific components that are hard to transfer to other languages. Thus, application of the same system as the one used for English is expensive, if not impossible. In such cases, if a given work reports 97.83 % accuracy for a POS tagger of Bulgarian, it is not possible to compare it to an English POS tagger reporting 98.03 % accuracy, for the evaluation corpora are not comparable in any sense. The construction of ParDeepBank will overcome this problem by providing directly comparable analysis on several linguistic levels for several languages over parallel texts. The treebank might be used for defining comparable measures for various languages and different NLP tasks.

- *Comparable coverage of the resource grammars for several languages.* Although the development of most of the resource grammars for different languages follows similar scenarios, their coverage diverges in the process of development. ParDeepBank will provide a basis for measuring the coverage of such grammars over a large amount of parallel data. This is of great importance with respect to exploitation of such grammars in applications like machine translation (see [2]).

- *Linguistic research.* Parallel corpora annotated with such detailed linguistic analyses are valuable for language data in multilingual contexts. The selection of the three languages in different language families will also facilitate the comparison of language phenomena.

At the moment the idea behind ParDeepBank is demonstrated through an extension of DeepBank with two other languages: Portuguese and Bulgarian. For both languages a portion of the data present in the DeepBank has been translated, parsed with the two other grammars and parallelized. The details for each language effort are presented and commented on in the next sections.

3 Resource Grammars for Three Languages

3.1 English Resource Grammar

The ERG is an open-source, broad-coverage, declarative grammar implementation for English, designed within the HPSG framework. This linguistic framework places most of the burden of linguistic description on the lexicon, employing a relatively small number of highly schematic syntactic rules to combine words or phrases to form larger constituents. Each word or phrase (more generally, each sign) is defined in terms of feature structures where the values of attributes are governed by general principles such as the Head Feature Convention, which ensures the identity of a particular subset of feature-value pairs on a phrasal node and on one distinguished daughter, the head of the phrase. Many of these generalizations aim to be language-independent, constraining the space of possible analyses for linguistic phenomena. Central to the HPSG framework is the inclusion in each sign of constraints on multiple dimensions of representation, including at least syntactic and semantic properties, so that rules, lexical entries, and principles determine semantic well-formedness just as they do syntax. Under continuous development at CSLI since 1993, the ERG provides syntactic and semantic analyses for the large majority of common constructions in written English text (cf. [17]). The current grammar consists of a 35,000-word lexicon instantiating 980 leaf lexical types, as well as 70 derivational and inflection rules, and 220 syntactic rules.

3.2 Portuguese Resource Grammar

The development of the Portuguese part of the ParDeepBank was supported by the Portuguese Resource Grammar LXGram. This grammar was presented at length in [7], [8]. A brief overview is provided in the present section. LXGram is based on hand coded linguistic generalizations supplemented with a stochastic model for ambiguity resolution. Like the other grammars used for the English and Bulgarian parts of the ParDeepBank, it follows the grammatical framework of Head-Driven Phrase Structure Grammar (HPSG, [24]). As this is a linguistic framework for which there is a substantial amount of published work, this option allows for the straightforward implementation of grammatically principled analyses that have undergone extensive scientific scrutiny. Following this grammatical framework, LXGram associates grammatical representations to natural language expressions, including the formal representation of their meaning, thus providing for so called deep linguistic processing of the input sentences. It was developed on top of a cross-language seed computational grammar fostered by the Matrix project [1].

Like several other computational grammars, including the other two used for the construction of the present multilingual ParDeepBank, LXGram uses Minimal Recursion Semantics (MRS, [13]) for the representation of meaning. An MRS representation supports scope underspecification; i.e. it is a description of a set of possible logic formulas that differ only in the relative scope of the relations present in these formulas. This format of semantic representation is well defined in the sense that it is known how to map between MRS representations and formulas of second order logic, for which there is a set-theoretic interpretation.

LXGram is developed in the Linguistic Knowledge Builder (LKB) system [11], an open-source development environment for constraint-based grammars. This environment provides a GUI, debugging tools and very efficient algorithms for parsing and generation with the grammars developed there. Several broad coverage HPSGs have been developed in the LKB, of which the largest ones are for English [12], used in this paper, German [14] and Japanese [26].

LXGram is in active development, and it already encompasses a wide range of linguistic phenomena, such as long distance dependencies, coordination, subordination, modification and many subcategorization frames, with a lexicon containing around 25 000 entries. In its last stable version, it contains over 60 lexical rules, 100 syntax rules, and 850 lexical leaf types (determining syntactic and semantic properties of lexical entries). It resorts to a pre-processing step performed by a pipeline of the shallow processing tools handling tokenization, POS tagging, morphological analysis, lemmatization and named entity recognition [3], [4], [15], [10].

LXGram copes with the European and the Brazilian variants of Portuguese. It contains lexical entries that are specific to either of them, and it covers both European and Brazilian syntax, as more thoroughly described in [5], [6]. The LXGram operation is coupled with a statistical disambiguation model, in order to automatically select the most likely analysis of a sentence when the grammar produces multiple solutions. Using a maximum entropy algorithm, this model is trained from the CINTIL DeepBank [9]. At present, this dataset comprises over 10 000 sentences of newspaper text, and development continues. The linguistic analyses that are implemented in the grammar are documented in a report that is updated and expanded with each version of the grammar. The grammar is available for download at http://nlx.di.fc.ul.pt/lxgram, together with this documentation.

An experiment was conducted to assess the coverage of LXGram's version on spontaneous text at the time of the experiment. This experiment and its results are presented at length in [8]. In a nutshell, the grammar exhibited a parsing coverage of around one third (i.e. one third of the input sentences get at least one parse by the grammar), and a parsing accuracy in the range of 30-40% (i.e. from the sentences that got at least one parse, that was the proportion of sentences for which the grammar delivers the correct grammatical representation).[2]

3.3 Bulgarian Resource Grammar

In the development of the Bulgarian part of ParDeepBank we rely on the Bulgarian HPSG resource grammar BURGER [23], and on a dependency parser (Malt Parser — [20], trained on the BulTreeBank data. Both parsers produce semantic representations in terms of MRS. BURGER automatically constructs them, while the output of the Malt Parser is augmented with rules for construction of MRS-es from the dependency trees. The integration of both tools has several advantages, such as: in the first version of the Bulgarian part of the parallel treebank, all the translated from English sentences have a correct analysis on MRS level, which to be used for the alignment purposes. Later on, when BURGER is extended to cover also these sentences, the analyses will be substituted. BURGER covers the main constructions in the MRS dataset of ERG (translated into Bulgarian and augmented with some sentences from BulTreeBank). Then it has been extended by a verbal lexicon, containing about 15000 verb lemmas (more than 700000 wordforms) encoded on morphosyntactic level as well as about 3000 ones, encoded on valency level. We are working on the extension of the lexicon of BURGER with more valency information and other parts-of-speech entries.

The chosen procedure, as mentioned above, is as follows: first, the Bulgarian sentences are parsed with BURGER. If it succeeds, then the produced MRS-es are used for the alignment. In case BURGER has no coverage, the sentences are parsed with the Malt Parser, and then MRS-es are constructed over the dependency parses. The MRS-es are created via a set of transformation rules [27]. Here is an overview of the components of the Bulgarian Language Processing Pipeline, exploited within the work:

- POS tagging. POS tagging is performed by a cascade of several modules — including a guesser, linguistic knowledge (lexicon and rules) and a statistical tagger. The accuracy of the whole pipeline is 97.83% — [18]. In this pipeline the SVM POS Tagger plays the role of a guesser for the GTagger.

- Lemmatization. The lemmatization is based on the morphological lexicon. From the lexicon we extracted functions which convert each wordform into its basic form (as a representative of the lemma). The accuracy is 95.23%.

- Dependency parsing. MALTParser has been trained on the dependency version of BulTreeBank. The model achieves 87.6% labeled parsing accuracy.

- MRS analyzer. We exploit two types of rules over the dependency parses. The first constructs for each lexical node its elementary predication, while

[2] To put these results into perspective, it is worth mentioning [33], who report values of 80.4% parsing coverage on newspaper text for ERG, 42.7% for the Japanese grammar and 28.6% for the German grammar, which have been in development for over 15 years now, being older than LXGram, with around 5 years of development.

the second combines the structures for two nodes on the basis of the dependency relations.

4 Sentence Selection and Treebanking

English DeepBank The English DeepBank is a treebank created by application of the Redwoods treebank approach to the Wall Street Journal (WSJ) corpus included in the PTB. The process of DeepBank annotation of the WSJ corpus is organised into iterations of a cycle of parsing, treebanking, error analysis and grammar/treebank updates, with the goal of maximizing the accuracy of annotation through successive refinement.

Sentences from the WSJ are first parsed with the PET parser using the ERG. Up to 500 top readings are recorded for each sentence. The exact best-first parsing mode guarantees that these recorded readings are the ones that have “achieved" the highest disambiguation scores according to the currently in-use parse selection model, without enumerating through all possible analyses.

The parsing results are then manually disambiguated by the human annotators. However, instead of looking at individual trees, the annotators spend most of their effort making binary decisions on either accepting or rejecting constructions. Each of these decisions, called discriminants, reduces the number of the trees satisfying the constraints (here maybe an example is due). Every time a decision is made, the remaining set of trees and discriminants are updated simultaneously. This continues until one of the following conditions is met: i) if there is only one remaining tree and it represents a correct analysis of the sentence, the tree is marked as gold; ii) if none of the remaining trees represents a valid analysis, the sentence will be marked as “rejected”, indicating an error in the grammar[3]; iii) if the annotator is not sure about any further decision, a “low confidence” state will be marked on the sentence, saved together with the partial disambiguation decisions. Generally speaking, given n candidate trees, on average $\log_2 n$ decisions are needed in order to fully disambiguate. Given that we set a limit of 500 candidate readings per sentence, the whole process should require no more than 9 decisions. If both the syntactic and the semantic analyses look valid, the tree is recorded as the gold reading for the sentence.

While the grammar development is independent to the treebanking progress, we periodically incorporate the recent changes of the grammar into the treebank annotation cycle. When a grammar update is incorporated, the treebank also gets updated accordingly by i) parsing anew all the sentences with the new grammar; ii) re-applying the recorded annotation decisions; iii) re-annotating those sentences which are not fully disambiguated after step ii. The extra manual annotation effort in treebank update is usually small when compared to the first round annotation.

[3] In some cases, the grammar does produce a valid reading, but the disambiguation model fails to rank it among the top 500 candidates. In practice, we find such errors occuring frequently during the first annotation circle, but they diminish quickly when the disambiguation model gets updated.

Portuguese Component A first step in the construction of the Portuguese part of the ParDeepBank consisted in obtaining a corpus of raw sentences that is parallel to the WSJ corpus, on which the PTB is based. To this end, the WSJ text was translated from English into Portuguese. This translation was performed by two professional translators. Each portion of the corpus that was translated by one of the translators was subsequently reviewed by the other translator in order to double-check their outcome and enforce consistency among the translators.

Given that the original English corpus results from the gathering of newspaper texts, more specifically from the Wall Street Journal, a newspaper specialized on economic and business matters, the translators were instructed to perform the translation as if the result of their work was to be published in a Portuguese newspaper of a similar type, suitable to be read by native speakers of Portuguese. A second recommendation was that each English sentence should be translated into a Portuguese sentence if possible and if that would not clash with the first recommendation concerning the "naturalness" of the outcome.

As the Portuguese corpus was obtained, it entered a process of dynamic annotation, analogous to the one applied to the Redwoods treebank. With the support of the annotation environment [incr tsdb()] [21], the Portuguese Resource Grammar LXGram was used to support the association of sentences with their deep grammatical representation. For each sentence the grammar provides a parse forest; the correct parse if available, can be selected and stored by deciding on a number of binary semantic discriminants that differentiate the different parses in the respective parse forest.

The translation of the WSJ corpus into Portuguese is completed, and at the time of writing the present paper, only a portion of these sentences had been parsed and annotated. While this is an ongoing endeavor, at present the Portuguese part of the ParDeepBank includes more than 1,000 sentences.

The sentences are treebanked resorting to the annotation methodology that has been deemed in the literature as better ensuring the reliability of the dataset produced. They are submitted to a process of double blind annotation followed by adjudication. Two different annotators annotate each sentence in an independent way, without having access to each other's decisions. For those sentences over whose annotation they happen to disagree, a third element of the annotation team, the adjudicator, decides which one of the two different grammatical representations for the same sentence, if any, is the suitable one to be stored. The annotation team consists of experts graduated in Linguistics or Language studies, specifically hired to perform the annotation task on a full time basis.

Bulgarian Component Bulgarian part of ParDeepBank was produced in a similar way to the Portuguese part. First, translations of WSJ texts were performed in two steps. During the first step the text was translated by one translator. We could not afford professional translators. We have hired three students in translation studies(one PhD student and two master students studying at English Department of the

Sofia University). Each sentence was translated by one of them. The distribution of sentences included whole articles. The reason for this is the fact that one of the main problems during the translation turned out to be the named entities in the text. Bulgarian news tradition changed a lot in the last decades moving from transliteration of foreign names to acceptance of some names in their original form. Because of the absence of strict rules, we have asked the translators to do search over the web for existing transliteration of a given name in the original text. If such did not exist, the translator had two possibilities: (1) to transliterate the name according to its pronunciation in English; or (2) to keep the original form in English. The first option was mainly used for people and location names. The second was more appropriate for acronyms. Translating a whole article ensured that the names have been handled in the same way. Additionally, the translators had to translate each original sentence into just one sentence in Bulgarian, in spite of its complex structure. The second phase of the translation process is the editing of the translations by a professional editor. The idea is the editor to ensure the "naturalness" of the text. The editor also has a very good knowledge of English. At the moment we have translated section 00 to 03.

The treebanking is done in two complementary ways. First, the sentences are processed by BURGER. If the sentence is parsed, then the resulting parses are loaded in the environment [incrs tsdb()] and the selection of the correct analysis is done similarly to English and Portuguese cases. If the sentence cannot be processed by BURGER or all analyses produced by BURGER are not acceptable, then the sentence is processed by the Bulgarian language pipeline. It always produces some analysis, but in some cases it contains errors. All the results from the pipeline are manually checked via the visualization tool within the CLaRK system. After the corrections have been repaired, the MRS analyzer is applied. For the creation of the current version of ParDeepBank we have concentrated on the intersection between the English DeepBank and Portuguese DeepBank. At the moment we have processed 328 sentences from this intersection.

5 Dynamic Parallel Treebanking

Having dynamic treebanks as components of ParDeepBank we cannot rely on fixed parallelism on the level of syntactic and semantic structures. They are subject to change during the further development of the resource grammars for the three languages. Thus, similarly to [28] we rely on alignment done on the sentence and word levels. The sentence level alignment is ensured by the mechanisms of translation of the original data from English to Portuguese and Bulgarian. The word level could be done in different ways as described in this section.

For Bulgarian we are following the word level alignment guidelines presented in [29] and [30]. The rules follow the guidelines for segmentation of BulTreeBank. This ensures a good interaction between the word level alignment and the parsing mechanism for Bulgarian. The rules are tested by manual annotation of several parallel Bulgarian/English corpora which represent the gold standard corpus for word level alignment between Bulgarian and English. For the rest of the data we are following the approach, undertaken for the alignment of the Portuguese DeepBank to English DeepBank.

The word level alignment between Portuguese and English is done in two steps. First, GIZA++ tool[4] is trained on the parallel corpus resulting from the translation of WSJ into Portuguese. This training produces an alignment model which is tuned with respect to this particular project. After this automatic step, a manual checking with a correction procedure is performed. The manual step was performed by the alignment editing tool in the Sanchay collection of NLP tools[5].

The alignment on the syntactic and semantic levels is dynamically constructed from the sentence and word level alignment as well as the current analyses of the sentences in ParDeepBank. The syntactic and semantic analyses in all three treebanks are lexicalized, thus the word level alignment is a good starting point for establishing of alignment also on the upper two levels. As one can see in the guidelines for Bulgarian/English word level alignment, the non-compositional phrases (i.e. idioms and collocations) are aligned on word level. Thus, their special syntax and semantics is captured already on this first level of alignment. Then, considering larger phrases, we establish syntactic and semantic alignment between the corresponding analyses only if their lexical realizations in the sentence are aligned on word level.

This mechanism for alignment on different levels has at least two advantages: (1) it allows the exploitation of word level alignment which is easier for the annotators; and (2) it provides a flexible way for updating of the existing syntactic and semantic alignments when the DeepBank for one or more languages is updated after an improvement has been made in the corresponding grammar. In this way, we have adopted the idea of dynamic treebanking to the parallel treebanking. It provides an easy way for improving the quality of the linguistic knowledge encoded in the parallel treebanks. Also, this mechanism of alignment facilitates the additions of DeepBanks for other languages or additions of analyses in other formalisms.

6 Conclusion and Future Work

In this paper we presented the design and initial implementation of a scalable approach to parallel deep treebanking in a constraint-based formalism, beginning with three languages for which well-developed grammars already exist. The methods for translation and alignment at several levels of linguistic representation are viable, and our experience confirms that the monolingual deep treebanking methodology can be extended quite successfully in a parallel multi-lingual context, when using the same data and the same grammar architecture. The next steps of the ParDeepBank development are the expansion of the volume of aligned annotated

[4] http://code.google.com/p/giza-pp/

[5] http://sanchay.co.in/

data, and improvements in the infrastructure for supporting the creation, maintenance and exploitation of these dynamic Parallel DeepBanks.

References

[1] Bender, E. M., Flickinger, D., and Oepen, S. 2002. The Grammar Matrix. An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammar. In *Proc. of the Workshop on Grammar Engineering and Evaluation at COLING'02*.

[2] Bond, F., Oepen, S., Siegel, M., Copestake, A., and Flickinger, D. 2005. Open-Source Machine Translation with DELPH-IN. In *Proc. of the Open-Source Machine Translation Workshop at the 10th MTS*, pp. 15–22.

[3] Branco, A., and Silva, J. 2006a. Dedicated Nominal Featurization of Portuguese. In *Proc. of the 7th International Conference on Computational Processing of the Portuguese Language*, PROPOR'06, pp. 244–247.

[4] Branco, A., and Silva, J. 2006b. A Suite of Shallow Processing Tools for Portuguese: LX-Suite. In *Proc. of the Eleventh Conference of the EACL: Posters & Demonstrations*, EACL'06, pp. 179–182.

[5] Branco, A., and Costa, F. 2007a. Accommodating Language Variation in Deep Processing. In *Proc. of GEAF07 Workshop*, pp. 67–86.

[6] Branco, A., and Costa, F. 2007b Self- or Pre-Tuning? Deep Linguistic Processing of Language Variants. In *Proc. of the ACL Workshop on Deep Linguistic Processing*.

[7] Branco, A., and Costa, F. 2008. LXGram in the Shared Task "Comparing Semantic Representations" of STEP2008. In *Proc. of the 2008 Conference on Semantics in Text Processing*, pp. 299–310.

[8] Branco, A., and Costa, F. 2010. LXGram: A Deep Linguistic Processing Grammar for Portuguese. In *LNAI, 6001*, pp. 86–89.

[9] Branco, A., Costa, F., Silva, J., Silveira, S., Castro, S., Avelãs, M., Pinto, C., and Graça, J. 2010. Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: CINTIL DeepGramBank. In *Proc. of LREC'10.*

[10] Branco, A., and Nunes, F. 2012. Verb Analysis in a Highly Inflective Language with an MFF Algorithm. In *Proc. of the 10th international conference on Computational Processing of the Portuguese Language*, PROPOR'12, pp. 1–11.

[11] Copestake, A. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications.

[12] Copestake, A., and Flickinger, D. 2000. An Open-Source Grammar Development Environment and Broad-Coverage English Grammar Using HPSG. In *Proc. of the LREC00.*

[13] Copestake, A., Flickinger, D., Pollard, C., and Sag, I. 2005. Minimal Recursion Semantics: an Introduction. *Research on Language and Computation*, 3(4).

[14] Crysmann, B. 2007. Local Ambiguity Packing and Discontinuity in German. In *Proc. of the ACL Workshop on Deep Linguistic Processing*.

[15] Ferreira, E., Balsa, J., and Branco, A. 2007. Combining Rule-Based and Statistical Methods for Named Entity Recognition in Portuguese. In *Proc. of the 5th Workshop em Tecnologia da Informação e da Linguagem Humana*, pp. 1615–1624.

[16] Flickinger, D. 2002. On Building a More Efficient Grammar by Exploiting Types. In *Collaborative Language Engineering*, pp. 1–17. CSLI Publications.

[17] Flickinger, D. 2011. Accuracy vs. Robustness in Grammar Engineering. In *Language from a Cognitive Perspective: Grammar, Usage, and Processing*, pp. 31–50. CSLI Publications.

[18] Georgiev, G., Zhikov, V., Osenova, P., Simov, K., and Nakov, P. 2012. Feature-Rich Part-Of-Speech Tagging for Morphologically Complex Languages: Application to Bulgarian. In *EACL 2012*.

[19] Klyueva, N., and Mareček, D. 2010. Towards parallel czech-russian dependency treebank. In *Proc. of the Workshop on Annotation and Exploitation of Parallel Corpora*.

[20] Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, E., Kübler, S., Marinov, S., and Marsi, E. 2007. MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. In *Natural Language Engineering, 13(2)*, pp. 95–135.

[21] Oepen, S. 1999. *[incr tsdb()] - Competence and Performance Laboratory*. Saarland University.

[22] Oepen, S., Toutanova, K., Shieber, S., Manning, C., Flickinger, D., and Brants, T. 2002. The LinGO Redwoods Treebank: Motivation and Preliminary Applications. In *Proc. of COLING'02*, pp. 1–5.

[23] Osenova, P. 2010. *The Bulgarian Resource Grammar*. VDM.

[24] Pollard, C., and Sag, I. 1994. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. University of Chicago Press.

[25] Prasad, R., Dinesh, N., Lee, A., Miltsakaki E., Robaldo, L., Joshi, A., and Webber, B. 2008. The Penn Discourse Treebank 2.0. In *In Proc. of LREC'08*.

[26] Siegel, M., and Bender, E. 2002. Efficient Deep Processing of Japanese. In *Proc. of COLING'02*.

[27] Simov, K., and Osenova, P. 2011. Towards Minimal Recursion Semantics over Bulgarian Dependency Parsing. In *Proc. of the RANLP 2011*.

[28] Simov, K., and Osenova, O. 2012. Bulgarian-English Treebank: Design and Implementation. In *Proc. TLT10*.

[29] Simov, K., Osenova, P., Laskova, L., Kancheva, S., Savkov, A., and Wang, R. 2012. HPSG-Based Bulgarian-English Statistical Machine Translation. *Littera et Lingua*.

[30] Simov, K., Osenova, P., Laskova, L., Savkov, A., and Kancheva, S. 2011. Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment. In *Proc. of The Second Workshop on Annotation and Exploitation of Parallel Corpora*, pp. 29–38.

[31] Tiedemann, J., and Kotzé, G. 2009. Building a Large Machine-Aligned Parallel Treebank. In *Proc. of TLT08*, pp. 197–208.

[32] Volk, M., Göhring, A., Marek, T., and Samuelsson, Y. 2010. SMULTRON (version 3.0) — The Stockholm MULtilingual Parallel TReebank.

[33] Zhang, Y., Wang, R., and Oepen S. 2009. Hybrid Multilingual Parsing with HPSG for SRL. In *Proc. of CoNLL 2009*.