Universalizing BulTreeBank: a Linguistic Tale about Glocalization

Petya Osenova
Linguistic Modeling Department
IICT-BAS
petya@bultreebank.org

Kiril Simov
Linguistic Modeling Department
IICT-BAS
kivs@bultreebank.org

Abstract

The paper presents the strategies and conversion principles of BulTreeBank into Universal Dependencies annotation scheme. The mappings are discussed from linguistic and technical point of view. The mapping from the original resource to the new one has been done on morphological and syntactic level. The first release of the treebank was issued in May 2015. It contains 125 000 tokens, which cover roughly half of the corpus data.

1 Introduction

The efforts within the NLP community towards universalized language datasets for getting comparable, objective and scalable results in parsing and other tasks are not so recent. Concerning syntax, some shared representations have been proposed and used at CoNLL contests on dependency parsing in 2006 (Buchholz and Marsi, 2006) and 2007 (Nivre et al., 2007). Another stream of sharing the same annotation framework was the adoption of the schemes of already existing treebanks. For example, a number of syntactic annotation works followed the style of Prague Dependency Treebank (Bejček et al., 2013) (i.e., Slovene (Džeroski et al., 2006), Croatian (Berovic et al., 2012), Tamil (Ramasamy and Žabokrtsky, 2012) etc.); many other treebanks followed the Penn Treebank style (i.e., Arabic (Maamouri et al., 2008), Chinese (Xue et al., 2005), etc.). An alternative way of pursuing a common annotation architecture is the pre-shared core deep grammar, such as the Matrix Grammar (Bender et al., 2002) in DELPH-IN initiative,[1] which helps to develop the language specific part further. However, all shared annotation schemes face the same challenges, namely what model might ensure maximum coverage of language specific phenomena and then, how to deal with the phenomena that are easy to universalize, and with those that are hard to incorporate.

The most recent initiatives which refer to Stanford typed dependencies (de Marneffe and Manning, 2008) and Universal Dependencies (de Marneffe et al., 2014) are not an exception to the above presented situation. They build on the existing treebanks and aim at universal parts-of-speech (POS) and dependency relations. With more and more languages coming on board, new issues are raised and considered. For that reason, the Universal Dependencies initiative has taken a dynamic approach. This means that there are regular releases of the treebanks in accordance to some current annotation model. Each release is frozen to its agreed annotation model. Then the model is enriched, changed, reconsidered, and the follow-up release takes into account the revised one. It seems that versioning is indeed the only fair way to tackle the diversity of language phenomena.

BulTreeBank did not participate in the first release of universalized treebanks (UD v1.0 (Nivre et al., 2015)). However, part of it was delivered in the second one – UD v1.1 (Agić et al., 2015) together with other 17 languages. Its size is 125 000 tokens, which constitute half of the data.

In this paper we present the strategies of converting BulTreeBank into Universal Dependency format with respect to morphology and syntax. The undertaken conversion steps and various linguistic issues are discussed in the context of manual/automated work and universal/specific language features.

The structure of the paper is as follows: Section 2 focuses on related work. Section 3 highlights the BulTreeBank resource in a nutshell. Section 4 outlines the universalizing principles of morphology and syntax. Section 5 describes the conversion procedure. Section 6 reports on some

[1] http://www.delph-in.net/matrix/

Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pages 81–89,
Hissar, Bulgaria, 10–11 September 2015.

preliminary results from training MATE Tools on the converted treebank. Section 7 concludes the paper.

2 Related Work

The Universal Dependency initiative evolved mainly from the Stanford Type Dependency efforts and Google attempts (Petrov et al., 2012) in universalizing parts-of-speech. However, it is also ideologically related to CoNLL contests (2006 and 2007).

The universalizing activities started with two main directions of research. The first can be illustrated by the work of Rosa et al. (2014) where 30 treebanks have been harmonized into a common Prague Dependency style, and then converted into Stanford Dependencies.[2] It does not handle language specific features. BulTreeBank was also among the harmonized treebanks. The second can be exemplified by the work of Sanguinetti and Bosco (2014) and Bosco and Sanguinetti (2014). The authors describe the conversion of the parallel treebank ParTUT (Italian, English, French) into Stanford dependencies. In the same context is the work of Lipenkova and Souček (2014) on Russian dependency treebank.

Later on came also work on the conversion of the treebanks into Universal Dependencies. These include the conversion of the Swedish treebank (Nivre, 2014) and the Finnish treebank (Pyysalo et al., 2015). The experiments with the converted Finnish treebank showed that the parsing results are better with the Universal Dependencies (UD).

3 BulTreeBank Resource in a Nutshell

The original BulTreeBank (Simov et al., 2004; Simov and Osenova, 2003) that has been used in the conversion to the universal format comprises 214,000 tokens, which form a little more than 15,000 sentences. Each token has been annotated with elaborate morphosyntactic information. The original XML format of the BulTreeBank is based on HPSG. The syntactic structure is presented through a set of constituents with head-dependant markings. The phrasal constituents contain two types of information: the domain of the constituent (*NP, VP* etc.) and the type of the phrase (head-complement (*NPC, VPC* etc.), head-subject (*VPS*), head-adjunct (*NPA, VPA* etc.). The treebank provides also functional nodes, such as clausal ones – *CLDA* (subordinate clause introduced by the auxiliary particle да to), *CLCHE* (subordinate clause introduced by the subordinator че that), etc.

Tracing back to the developments of BulTreeBank, its first 'glocalization' happened in 2006, when it was converted into the shared CoNLL dependency format – (Chanev et al., 2006), (Chanev et al., 2007). The rich structure was flattened to a set of 18 relations.[3] This part consists of 196 000 tokens, because the sentences with ellipses were not considered.

Alternative versions of BulTreeBank exist in two other popular formats: PennTreebank (Ghayoomi et al., 2014) and Stanford Dependencies (Rosa et al., 2014). The former was used for constituent parsing of Bulgarian, while the latter was part of a bigger endevour towards universalizing syntactic annotation schemes of many languages.

Now, BulTreeBank is part of the common efforts that evolved from the previous initiatives towards the creation of comparable syntactically annotated multilingual datasets. For the Universal Dependencies initiative we used the original BulTreeBank constituent-based format, because in the previous conversions to dependency format some important information was either lost, or underspecified.

4 Universalizing Morphology and Syntax

At this stage our conversion adheres fully to the universal annotation schemes. This means that we postponed the addition of language specific features for the next stage. The only language specific feature considered in this version is the morphologically marked count form – remnant of the old Slavic dual form within the category of Number. The morphological mapping includes parts-of-speech and their lexical as well as inflectional features. The syntactic mapping focuses on dependency relations.

In this section we do not aim at exhaustive description of the mappings, but rather at illustrating the varieties between the models.

4.1 Morphology

In morphology the following mapping cases occurred from the direction of the original tagset to

[2] This initiative as well as the Universal Dependencies stream build on the idea of interset, proposed by Zeman (2008).

[3] http://www.bultreebank.org/dpbtb/

the UD tagset: identical parts-of-speech, division of one POS into more parts-of-speech and changing the POS. It should be noted however that all the processes are interrelated.

1. **Direct Mapping.** The first case refers to subordinators and conjunctions, adjectives, prepositions.

2. **Division of one POS into more parts-of-speech.** The BulTreeBank original POS tagset[4] respects the morphological nature of the parts-of-speech, i.e., their origin. The UD tagset, however, is more syntactically oriented. It considers the syntactic function at the cost of parts-of-speech partitioning into several other groups. For example, in our original tagset the group of pronouns is homogeneous in spite of their differing functions. However, in UD this group is split into the groups DET, PRON and ADV. The category DET (determiner) is syntactic for Bulgarian, since the definite article is a phrasal affix and part of the word (масата 'table.DET' the table; високата маса 'tall.DEF table' the tall table). Thus, to this category belong the appropriate pronouns that are used attributively (definite, indefinite, collective, etc.). The pronouns that are used substantively, remain in the group PRON (pronoun). The pronouns that are used adverbially, are considered in the group ADV (adverb). Another division applies to nouns. The common ones map the group NOUN, while the proper nouns go to the specific group PROPN. Numerals also divide between the groups of ADJ, ADV and NUM. The verbs are divided into the groups VERB (main verbs, copulas and modals, participles that are part of verb forms), AUX (auxiliaries), ADJ (participles with attributive usages).

3. **Changing the POS**. One case of changing the original POS is the transition of the affirmative and negative particles to the group of INTJ (interjections). Also, all the pronouns that went to DET group, also changed their POS label.

Concerning the UD set of accompanying features, three of them were not specifically encoded in the original tagset: animacy, degree and passive forms. Concerning animacy, in Bulgarian the grammar-related dichotomy is more specific – Person vs. Non-Person. Thus, it is derivable from some explicit grammatical features, such as the case in some pronouns, the count form of the masculine nouns and the masculine form of the numerals. Concerning degree, the original tagset does not differentiate among positive, comparative and superlative forms. Concerning passive, active voice is considered a default, and passive form is handled at the syntactic level, since both ways of its formation are analytical (participial forms and se-forms).

4.2 Syntax

The transfer of the syntactic relations faces the following situations: direct transfer relations; non-direct relations; 'floating' relations and non-handled relations.

1. **Direct transfer relations.** Direct mappings are those that provide the necessary information on phrasal level. They include relations like *dobj*, *iobj*, *nsubj*, *csubj*, etc.[5] Also the distinction between the relations *aux* and *cop* is directly derived from the original annotation. The former being annotated lexically with V(erb) and the latter being annotated syntactically with a head-complement relation (VPC).

2. **Non-direct relations.** Indirect mappings are those that provide the necessary information in a more underspecified way. One example of such relations is the division of our original complement clauses (CLDA, CLCHE, etc.) into control (*xcomp*) and non-control ones (*ccomp*) within UD. Another example is the division of our head – adjunct nominal phrase (NPA) into several relations depending on the non-head sister: *nummod* (the non-head sister is numeral), *amod* (the non-head sister is adjective), *det* (the non-head sister is determiner). The division of complement clauses and head-adjunct nominal phrases into more specific structures is linguistically sound with respect to semantics. Our original style introduces preferences to generalization over structural analyses. In our opinion, these two approaches exhibit two different models

[4] http://www.bultreebank.org/TechRep/BTB-TR03.pdf

[5] The UD labels are given in footnote 6.

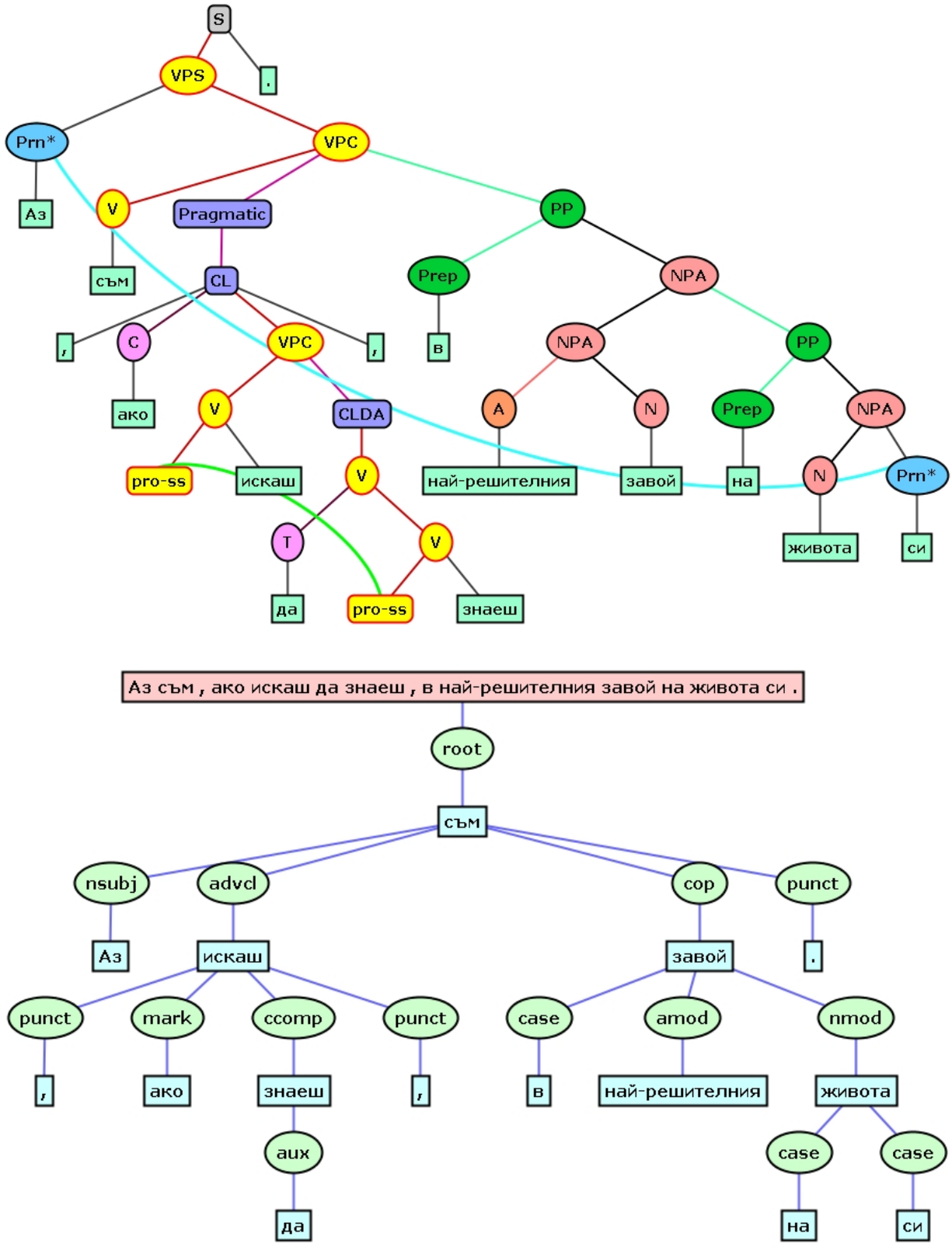

Figure 1: An HPSG-based tree and a Universal Dependency tree for the sentence: Аз съм, ако искаш да знаеш, в най-решителния завой на живота си. 'I am, if want.2PER.SG to know.2PER.SG, in most-crucial turn of life.DEF my.REFL' If you would like to know, I am in the most important turn of my life.

which might be useful for various tasks in NLP. Also, in the original treebank the passive constructions together with their participants were not marked explicitly. Hence, additional work was needed for annotating relations such as *nsubjpass* (nominal subject of a passive verb form) and *csubjpass* (clausal subject of a passive verb form). Thus, the specific auxiliary relation *auxpass* (relation between an auxiliary verb and the main verb form) is handled manually (see in Table 2 that at the moment only one such relation is available in the data).

3. **'Floating' relations.** There are mappings that have selected one alternative among several appropriate possibilities in the tagset. Such decisions might be temporary, since they are likely to be reconsidered in the future. Such a case is the encoding of the question particle ли 'li' in Bulgarian, which is used in yes – no questions. At the moment it is annotated with the relation *discourse*, but there are also other options, such as *aux*, *expl* or *mark*.

   Here also belongs the phenomenon of clitic doubling. In the original annotation we consider argument-like clitics at lexical level, while their counterparts (long pronoun forms or nouns) – at syntactic level. Here is an example: На него му се падна труден въпрос на изпита. 'To **him.LONG-PRON him.SHORT-PRON** REFL happened difficult question at exam.DEF' He got a difficult question at his exam.

   In UD, however, at the moment clitics receive two different relations depending on whether they are part of clitic doubling (then they are marked as *expl*) or not (then they are marked as *dobj* or *iobj*)).

4. **Non-handled relations.** We still have to analyze the elliptical phenomena in the remaining sentences of the treebank. Another thing to be reflected in the next release is the secondary predication, since this phenomenon requires also some co-reference information. Here is an example: Тя влезе тъжна в стаята. 'She entered **sad.FEM-SG** in room.DEF' She entered the room sad.

5 Conversion Procedure

Since in our original resource some multiword expressions were analyzed as one unit (especially those that matched one POS), for the UD scheme they had to be syntactically analyzed. In cases where it was not obvious what the head and dependencies are, the expressions were processed manually.

The parts-of-speech together with the relevant grammatical features were converted automatically through pre-defined mappings.

The syntactic relations required more work. Part of them were converted automatically, while part of them needed human intervention. For that reason all sentences with at least one unsolved mapping have been left for the next release.

In almost every constituent the head daughter could be determined unambiguously. However, more specific rules are needed in some combinations of constituents. For example, in *NP*s of type *NN* the head might be the first or the second noun depending on the semantics of the phrase. In such cases manual annotation of the head is necessary. Coordinations originally have been considered to be non-headed phrases, where the grammatical function overrides the syntactic labels. Thus, they also needed some special conversion treatment.

The procedure for the conversion of the BulTreeBank to Universal Dependencies is rule-based. The rules are of two kinds: (1) lexical head identifier moving up the constituent tree; and (2) relation assignment for a constituent node of the dependent child when all children of the parent node have lexical identifiers.

For example, let us have the following constituent, whose lexicalized example might be this one: твърде висок зелен стол. 'too tall green chair' [**NPA** [**APA** too tall] [**NPA** green chair]].

$$\texttt{NPA} \rightarrow \texttt{APA}_{id_1}\ \texttt{NPA}_{id_2}\ ,$$

where id_1 is a lexical head identifier for the adjectival phrase APA and id_2 is a lexical head identifier for the noun phrase NPA. Then we establish the relation amod from $\texttt{APA}_{id_1}$ to $\texttt{NPA}_{id_2}$ and the identifier for the child NPA is moved up, because the lexical head of the child NPA is the lexical head for the whole phrase. After the application of these two rules we have the constituent tree annotated with lexical identifiers and dependency relations in this way:

$$\texttt{NPA}_{id_2} \rightarrow \texttt{APA}_{id_1,amod}\ \texttt{NPA}_{id_2}\ .$$

Through the recursive application of such rules for the different types of phrases we annotated the whole constituent trees with lexical identifiers and universal dependency relations. When the root node receives an identifier, then the process stops and the constituent tree is converted to universal dependency tree.

In this way, we keep the original constituent annotation, while constructing the universal dependency annotation on top of it.

Some constructions like coordination, as mentioned above, require more complicated rules, since the necessary information was not directly encoded but it is trackable via the morphological annotation. However, the basic principle is the same.

Label	Num	Label	Num
A	9922	M	2436
APA	681	N	31513
APC	247	ND-Elip	27
Adv	5197	NPA	27664
AdvPA	381	NPC	67
AdvPC	52	Nomin	17
C	5407	PP	17478
CL	1479	Participle	3883
CLCHE	722	Prep	17286
CLDA	1965	Pron	9315
CLQ	166	Subst	497
CLR	1084	T	4817
CLZADA	147	V	22431
Conj	5465	VPA	8576
ConjArg	8958	VPC	11291
CoordP	4387	VPF	203
Gerund	15	VPS	9579
H	1037	Verbalised	4
I	25		

Table 1: Statistics over the HPSG Labels.

Table 1[6] summarizes the statistics of the syntactic labels in the original HPSG-based BulTreeBank, while Table 2[7] gives an overview of the converted BulTreeBank-UD. As it can be seen, direct comparisons cannot be made due to the fact that most often one original relation has been divided into more relations, or some UD relation combines material from two or more original ones. But even in such a setting, it can be observed that the most frequent type of relation is the one, in which a noun is connected to another noun via preposition (see relation PP in Table 1 and relations *case* and *nmod* in Table 2).

Label	Num	Label	Num
acl	1051	discourse	591
advcl	1258	dobj	5332
advmod	4437	expl	2790
amod	9528	iobj	2655
appos	38	mark	1410
aux	4839	mwe	671
auxpass	1	name	1110
case	18362	neg	1137
cc	3992	nmod	17293
ccomp	2428	nsubj	8506
conj	4573	nsubjpass	789
cop	1944	nummod	1460
csubj	368	punct	18013
csubjpass	16	root	9405
det	1586	vocative	6

Table 2: Statistics over the Universal Dependency Labels.

Additionally, in Fig. 1 an original treebank sentence is shown together with its UD conversion. Definitely, the new presentation flattens the tree,

[6]**A** – lexical adjective; **APA** – head-adjunct adjective phrase; **APC** – head-complement adjective phrase; **Adv** – lexical adverb; **AdvPA** – head-adjunct adverb phrase; **AdvPC** – head-complement adverb phrase; **C** – lexical conjunction; **CL** – clause that is outside the specific classes of clauses; **CLCHE** – clause introduced via "che" conjunction; **CLDA** – clause introduced via "da" verbal form; **CLQ** – interrogative clause; **CLR** – relative clause; **CLZADA** – adjuct clause for purpose; **Conj** – conjuction in a coordination phrase; **ConjArg** – argument of a coordination phrase; **CoordP** – coordination phrase; **Gerund** – lexical gerund form; **H** – lexical family name; **I** – lexical interjection; **M** – lexical numeral; **N** – lexical noun; **ND-Elip** – elliptical noun defined in the discourse; **NPA** – head-adjunct noun phrase; **NPC** – head-complement noun phrase; **Nomin** – nominalization of a phrase; **PP** – prepositional phrase; **Participle** – lexical participle; **Prep** – lexical preposition; **Pron** – lexical pronoun; **Subst** – substantive usage; **T** – lexical particle; **V** – lexical finite verb form; **VPA** – head-adjunct verb phrase; **VPC** – head-complement verb phrase; **VPF** – head-filler verb phrase; **VPS** – head-subject verb phrase; **Verbalised** – verbalization of a phrase.

[7]**acl** – clausal modifier of noun; **advcl** – adverbial clause modifier; **advmod** – adverbial modifier; **amod** – adjectival modifier; **appos** – appositional modifier; **aux** – auxiliary; **auxpass** – passive auxiliary; **case** – case marking; **cc** – coordinating conjunction; **ccomp** – clausal complement; **conj** – conjunct; **cop** – copula; **csubj** – clausal subject; **csubjpass** – clausal passive subject; **det** – determiner; **discourse** – discourse element; **dobj** – direct object; **expl** – expletive; **iobj** – indirect object; **mark** – marker; **mwe** – multi-word expression; **name** – name; **neg** – negation modifier; **nmod** – nominal modifier; **nsubj** – nominal subject; **nsubjpass** – passive nominal subject; **nummod** – numeric modifier; **punct** – punctuation; **root** – root; **vocative** – vocative

but it also adds more specific relations to it. It should be noted that the two lines in the HPSG-based tree in Fig. 1 connect the coreferences in the sentence (between the subject 'I' and the reflexive pronoun; and between the unexpressed subjects of the verbs 'want' and 'know').

6 Preliminary Experiments for POS Tagging and Dependency Parsing

We performed some preliminary experiments with the BulTreeBank-UD to train existing tools for POS tagging and Dependency Parsing. The 10-fold cross validation approach was used. We selected MATE tools[8] for the experiments, because they provide all the necessary components in one framework. The results are surprisingly good for the POS and Morphological tagging, while the dependency parsing performs somewhat sub-optimally. As background information it should be noted that the state-of-the-art results achieved in our previous work, with different data and different settings are as follows: in POS tagging (13 tags) – 99.30 % accuracy; in morphological tagging (680 tags) – 97.98 % accuracy (Georgiev et al., 2012), and in dependency parsing on BulTreeBank (ConLL-2006): LAS – 89.14 % and UAS – 92.45 % (Simova et al., 2014), using an ensemble model.

The current results are presented in Table 3 below:

Task	Accuracy	LAS	UAS
POS Tagging	96.89 %	–	–
Mor. Tagging	98.50 %	–	–
Dep. Parsing	–	83.50 %	88.08 %

Table 3: Evaluation. LAS = Labeled Accuracy Score, UAS = Unlabeled Accuracy Score.

However, we consider these results preliminary, because, as it was mentioned above, only part of the original treebank has been transformed into the universal representation and thus, only this part was used for the training. Additionally, many complex phenomena have not been represented within the current version yet.

It is worth noting that at the moment the original BulTreeBank tagset consists of 680 tags, while the UD one has 535 tags as combinations between POS and the respective grammatical features. This situation will change when more language specific features are added.

[8] http://code.google.com/p/mate-tools/

7 Conclusion

In this paper we describe the conversion of the original HPSG-based BulTreeBank into the Universal Dependencies format. The process included assigning Universal POS and Universal Morphological Features to the original annotations as well as conversion of the tree structures.

The conversion and the label assignments were done mainly automatically with a high level of certainty because the dependent elements in the original treebank were easy to track. At the same time, some phenomena will be detailed and handled in the next release of the treebank due to the need of human intervention in the language or annotation model specific cases.

The reported effort is part of a wider initiative that includes many languages and working groups. As such it faces similar challenges and shares similar perspectives. The main challenge is the proper handling of the language universal and language specific phenomena at a minimal linguistic and data model loss. The most important perspective is the ultimate goal of having comparable syntactically annotated resources for many languages that would serve better for various NLP tasks.

Acknowledgements

This research has received partial support by the EC's FP7 (FP7/2007 – 2013) project under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches" and FP7 grant 316087 AComIn "Advanced Computing for Innovation", funded by the European Commission in 2012 – 2016.

We are grateful to the three anonymous reviewers, whose remarks, comments, suggestions and encouragement helped us to improve the initial variant of the paper. All errors remain our own responsibility.

References

Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Jan Hajič, Anders Trærup Johannsen, Jenna Kanerva, Juha Kuokkala, Veronika Laippala, Alessandro Lenci,

Krister Lindén, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Héctor Alonso Martínez, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Joakim Nivre, Hanna Nurmi, Petya Osenova, Slav Petrov, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Sampo Pyysalo, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Kiril Simov, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal Dependencies 1.1. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague Dependency Treebank 3.0.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

Dasa Berovic, Zeljko Agic, and Marko Tadić. 2012. Croatian Dependency Treebank: Recent development and initial experiments. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Cristina Bosco and Manuela Sanguinetti. 2014. Towards a Universal Stanford Dependencies parallel treebank. In Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, and Adam Przepiórkowski, editors, *Proceedings of TLT-13*, pages 14–25, Tübingen, Germany. European Language Resources Association (ELRA).

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 149–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Atanas Chanev, Kiril Simov, Petya Osenova, and Svetoslav Marinov. 2006. Dependency conversion and parsing of the BulTreeBank. In *Proceedings of the LREC workshop Merging and Layering Linguistic Information*, pages 16–23.

Atanas Chanev, Kiril Simov, Petya Osenova, and Svetoslav Marinov. 2007. The BulTreeBank: Parsing and Conversion. In *Proceedings of the Recent Advances in Natural Language Processing Conference*, pages 114–120.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford Typed Dependencies Representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: a cross-linguistic typology. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtsky, and Andreja Žele. 2006. Towards a Slovene dependency treebank. In *In Proc. Int. Conf. on Language Resources and Evaluation (LREC*.

Georgi Georgiev, Valentin Zhikov, Kiril Simov, Petya Osenova, and Preslav Nakov. 2012. Feature-rich part-of-speech tagging for morphologically complex languages: Application to Bulgarian. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 492–502. Association for Computational Linguistics.

Masood Ghayoomi, Kiril Simov, and Petya Osenova. 2014. Constituency parsing of Bulgarian: Word- vs class-based parsing. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Janna Lipenkova and Milan Souček. 2014. Converting Russian Dependency Treebank to Stanford Typed Dependencies Representation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, page 143–147, Gothenburg, Sweden.

Mohamed Maamouri, Ann Bies, and Seth Kulick. 2008. Enhanced annotation and parsing of the Arabic treebank. In *In 6th International Conference on Computers and Informatics, INFOS2008*.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz

Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic, June. Association for Computational Linguistics.

Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Hajič, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Maria Simi, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal Dependencies 1.0. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Joakim Nivre. 2014. Universal dependencies for Swedish. In *SLTC Conference 2014*, Uppsala, Sweden.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal Dependencies for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, page 163–172.

Loganathan Ramasamy and Zdenek Žabokrtsky. 2012. Prague Dependency Style Treebank for Tamil. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 23–25, Istanbul, Turkey. European Language Resources Association.

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, and Zdeněk Žabokrtský. 2014. Hamledt 2.0: Thirty Dependency Treebanks Stanfordized. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Manuela Sanguinetti and Cristina Bosco. 2014. Converting the parallel treebank ParTUT in Universal Stanford Dependencies. In *Proceedings of CLiC-it 2014*, pages 316–321, Pisa, Italy.

Kiril Simov and Petya Osenova. 2003. Practical annotation scheme for an HPSG treebank of Bulgarian. In *Proceedings of the 4th International Workshop on Linguistically Interpreteted Corpora (LINC-2003)*, Budapest, Hungary.

Kiril Simov, Petya Osenova, Alexander Simov, and Milen Kouylekov. 2004. Design and implementation of the Bulgarian HPSG-based treebank. In *Journal of Research on Language and Computation, Special Issue*, pages 495–522, Kluwer Academic Publishers.

Iliana Simova, Dimitar Vasilev, Alexander Popov, Kiril Simov, and Petya Osenova. 2014. Joint ensemble model for POS tagging and dependency parsing. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 15–25, Dublin, Ireland, August. Dublin City University.

Naiwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Nat. Lang. Eng.*, 11(2):207–238, June.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.