АВТОМАТИЧЕСКАЯ ИДЕНТИФИКАЦИЯ ОБЩИХ АРГУМЕНТОВ СОЧИНЕННЫХ ГЛАГОЛОВ

Бердичевский А. (aleksandrs.berdicevskis@uit.no),
Экхофф Х. (hanne. m.eckhoff@uit.no)

Университет Тромсё—Норвежский арктический университет, Тромсё, Норвегия

Ключевые слова: синтаксический корпус, сочинение, синтаксис зависимостей, общие зависимые, общие аргументы, русский

AUTOMATIC IDENTIFICATION OF SHARED ARGUMENTS IN VERBAL COORDINATIONS

Aleksandrs Berdičevskis (aleksandrs.berdicevskis@uit.no),
Hanne Eckhoff (hanne. m.eckhoff@uit.no)

UiT The Arctic University of Norway, Tromsø, Norway

We describe automatic conversion of the SynTagRus dependency treebank of Russian to the PROIEL format (with the ultimate purpose of obtaining a single-format diachronic treebank spanning more than a thousand years), focusing on analysis of shared arguments in verbal coordinations. Whether arguments are shared or private is not marked in the SynTagRus native format, but the PROIEL format indicates sharing by means of secondary dependencies. In order to recover missing information and insert secondary dependencies into the converted SynTagRus, we create a simple guessing algorithm based on four probabilistic features: how likely a given argument type is to be shared; how likely an argument in a given position is to be shared; how likely a given verb is to have a given argument; how likely a given verb is to have a given argument frame. Boosted with a few deterministic rules and trained on a small manually annotated sample (346 sentences), the guesser very successfully inserts shared subjects (F-score 0.97), which results in excellent overall performance (F-score 0.92). Non-subject arguments are shared much more rarely, and for them the results are poorer (0.31 for objects; 0.22 for obliques). We show, however, that there are strong reasons to believe that performance can be increased if a larger training sample is used and the guesser gets to see enough positive examples. Apart from describing a useful practical solution, the paper also provides quantitative data about and offers non-trivial insights into Russian verbal coordination.

Key words: treebank, coordination, dependency syntax, shared dependents, shared modifiers, shared arguments, Russian

0. Introduction

This paper reports on an experiment where we use various cues to identify shared arguments in verbal coordinations. The experiment is a part of the full-scale conversion of the SynTagRus dependency treebank[1] of Contemporary Standard Russian (CSR) into the dependency format used in a large family of treebanks of ancient languages originating in the PROIEL corpus.[2] The PROIEL family of treebanks also includes the Tromsø Old Russian and OCS Treebank (TOROT, nestor.uit.no), which contains approximately 300,000 syntactically annotated word tokens divided between canonical Old Church Slavonic (OCS), Old Russian and Middle Russian. By converting the SynTagRus treebank to the PROIEL format, we add a compatible final stage to the TOROT treebank, in effect obtaining a diachronic treebank spanning more than a thousand years, from Late Common Slavic to CSR.

Although the conversion thus has an obvious practical purpose, it also has interesting theoretical implications. In this paper we focus on coordinations and, more specifically, on the analysis of omitted arguments in verbal coordinations. In this area, the PROIEL annotation contains more information than the SynTagRus annotation does: The PROIEL format uses secondary dependencies to mark shared dependents, while the SynTagRus format does not indicate them.

1. SynTagRus and TOROT/PROIEL

Both PROIEL and SynTagRus are dependency formats, both have close links to a particular syntactic framework. The SynTagRus format (Apresyan et al. 2005) is based on the Meaning–Text model (Mel'čuk 1995), which makes it strongly dependent on lexical semantics and gives it a highly granular argument structure representation based on ranked valencies. In other respects, it is the more traditional of the two formats, in that it only allows primary dependencies and only to a limited extent allows empty nodes. The PROIEL format is inspired by Lexical-Functional Grammar (LFG), and the dependency trees are convertible to LFG F-structures (Haug 2010, Haug et al. 2009). This is the origin of several departures from more traditional dependency formats. Structure sharing is indicated by way of secondary dependencies—for external subjects in control and raising structures, but also to indicate shared arguments and predicate identity. The format also systematically uses empty verb and conjunction nodes to account for ellipsis, gapping and asyndetic coordination. Argument representation is less granular than in SynTagRus, and the labels are largely based on morphosyntactic features: transitive objects (OBJ) are distinguished from oblique objects (OBL). In addition, complement clauses, passive agents, and arguments with

[1] Developed by the Laboratory of Computational Linguistics at the Institute for Information Transmission Problems (to whom we are grateful for the use of the offline version of SynTagRus and advice on its usage), found at http://ruscorpora.ru/search-syntax.html

[2] Developed by the members of the project *Pragmatic Resources in Old Indo-European Languages*, found at foni.uio.no:3000

external subjects, e. g. control infinitives, have separate labels, resp. COMP, AG and XOBJ. The two formats also take different approaches to coordination, as illustrated in Fig. 1 and 2, and further discussed in Section 2.

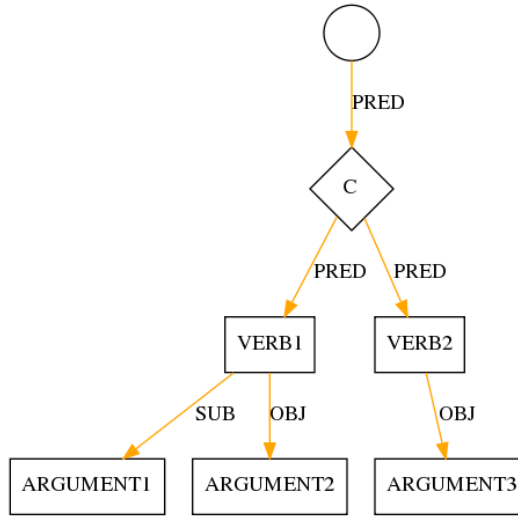

Fig. 1. PROIEL-style coordination. The conjunction is the head of the coordinated nodes, and its outgoing relations are the same as the ingoing one

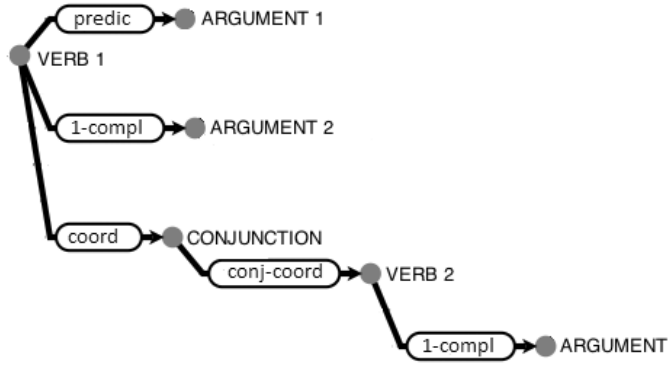

Fig. 2. SynTagRus-style coordination. The first conjunct is the head of the conjunction, which again is the head of the second conjunct, they are connected through a set of special "coordinative" relations

We managed to convert SynTagRus to the PROIEL format with high accuracy using a rule-based algorithm (excluding 528 sentences where the algorithm identified a structure that would most likely not be converted correctly, which leaves us 58,712

sentences). A spot check of 50 random converted sentences (763 words, including empty tokens) gives an unlabeled attachment score (UAS) of 0.96 and a labeled attachment score (LAS) of 0.93. The main issues encountered in the conversion process had to do with coordination and null verb insertion, since the necessary information was not always recoverable in the SynTagRus data. In many other cases, however, it was possible to recover missing information using lexical, morphological and part-of-speech cues. For instance, secondary dependencies indicating external subjects (e. g. of control infinitives and predicative complements) were inserted with 0.96 accuracy. In the current experiment, we insert secondary dependencies indicating shared arguments in coordinations. We are not aware of any highly successful previous attempts at this task. Zeman et al. (2012:2738), converting various treebanks to the Prague style, note that they "apply a few heuristics", but cannot recover shared dependents reliably. Mareček and Kljueva (2009:28–29), converting SynTagRus to the Prague style, note that they apply "a couple of rules", but "by far not all cases can be solved."

2. Coordination: theoretical and practical importance

Dependency grammar is fundamentally based on asymmetric relationships: one node is the head and another node is its dependent, i. e. it is based on subordination. Coordination, therefore, constitutes a problem to all dependency annotation, since it is symmetrical in nature: in paratactic relationships, there is no head, rather, nodes are equal and have the same status. Any dependency format will therefore have to coerce coordinations into hierarchical structures in some way, and it is not obvious what the best choice is. Popel et al. 2013 offer a comprehensive typology of coordination styles, recognising three major approaches: 1) Moscow-style, where the first conjunct is the head, the second conjunct is a dependent on the first, the third is a dependent on the second etc. As regards conjunctions, the policies can vary. In SynTagRus (Fig. 2), they are typically placed between the conjuncts according to the word order: the first conjunction is a dependent on the first conjunct, the second conjunct is a dependent on the first conjunction, etc. 2) Prague-style, where the conjunction is the head, and all conjuncts are its dependents. 3) Stanford-style, where the first conjunct is the head, and the following conjuncts and conjunctions are its dependents. There are different advantages and drawbacks to each solution. The Moscow-style solution has considerable advantages when it comes to simplicity of potential syntactic queries (the first conjunct will be the direct dependent of its true head), whereas the Prague-style solution makes it possible to render complicated stacked structures better, and also allows encoding of shared dependents.

The PROIEL format (Fig. 1) employs a variety of the Prague-style solution, also including the use of null verbs in the case of gapping and other types of verb ellipsis, and null conjunctions in the case of asyndetic coordination. The PROIEL format also uses secondary dependencies to indicate predicate identity and shared dependents. The latter feature is of particular importance for large-scale studies of argument structure, since it makes it possible to extract more complete argument frames. For some possible uses of such data, see Berdičevskis and Eckhoff 2014.

3. Materials and methods

3.1. Objective, key notions and limitations

Our objective is to insert correct shared-argument secondary dependencies in verbal coordinations. We assume that such dependencies can potentially be present when and only when one of the two coordinated verbs has an argument slot filled, while the other verb has the same slot unfilled. The empty-slot criterion is illustrated in Fig. 1: both Verb1 and Verb2 have the object slot filled, and thus none of the objects can be shared. Verb2, however, does not have a subject, while Verb1 does, and thus it is possible that Argument1 is a subject for Verb2, too.

Obviously, coordinated groups can contain more than two verbs, and our samples do include such cases, but at any given point in time we will be looking at two verbs only. More specifically, a *datapoint* in our analysis will be a pair {Argument; Verb}. In Fig. 1, the pair is {Argument1; Verb2}; in an example such as

(1) *Потом он посмотрел на нее и обрадовался*[3]
'Then he looked at her and was pleased'

the available pairs are {он (SUB) 'he'; обрадовался 'was_pleased'} and {на (OBL) 'at'; обрадовался 'was_pleased'}. For every datapoint, we will have to make a decision whether a secondary dependency should be inserted or not (it should in the first of the two pairs mentioned in the previous sentence, but should not in the second). For brevity's sake, we will call drawing a secondary dependency an *adoption* of an argument by a verb (extending the parent-and-children metaphor that is widely used in syntax).

In this experiment, we attempt to identify cases of argument adoption only, i. e. we focus on the relations SUB, OBJ, OBL, XOBJ, COMP, AG (Section 1).[4] In addition, we impose several other limitations on our dataset: we exclude empty verbal nodes[5], participles (since they are expected to display more limited argument frames) and some complex infrequent sharing cases not covered by Fig.1, such as arguments shared by one conjunct verb and a verb in a dependent clause daughter of another conjunct verb.

---

[3] Ju. Kazakov, *Dvoe v dekabre*.

[4] For the relation OBL, the empty-slot criterion is more relaxed: the adopting verb is not required to have the OBL slot unfilled. In other words, since one verb can have more than one outgoing OBL (which is forbidden for the other five relations), verbs are permitted to adopt OBLs even if they already have OBLs as their primary dependents. For the other relations, if there are several candidates that satisfy the empty-slot criterion, only one of them will be selected (the most plausible one, as determined by the algorithm described in Section 3.4).

[5] Elided copulas and cases of verbal gapping

3.2. Data

We used a number of probabilistic features to identify adoption (Section 3.3). For some of the features, we calculated values using the whole converted version of SynTagRus. For some of the features, however, a training set with correct secondary dependencies already inserted was required. For that purpose, we manually annotated a small sample.

At the first stage, the sample included 297 sentences. The sentences were randomly drawn from the whole corpus, the only requirement being that each sentence should contain at least one potential adoption case, i. e. at least one construction satisfying the criteria described in Section 3.1.[6] If sentences contained critical errors that distorted coordination (mostly as a result of conversion, sometimes SynTagRus misannotations), we excluded them, which resulted in 267 annotated sentences.

The analysis of this sample, however, showed that different arguments behave very differently with respect to how often they occur as potential adoptees and with respect to how often they actually are adopted, i. e. to their adoption rate (cf. Table 1). SUBs are frequent and likely to be shared, which gives us a lot of both positive and negative examples. OBJs and OBLs are frequent, but much less likely to be shared, which gives us a lot of negative examples, but very few positive ones. XOBJs are relatively frequent, but very unlikely to be shared, and are thus even more problematic. COMPs and especially AGs are infrequent, and it is not even possible to calculate their expected adoption rate.

To estimate the adoption rate for infrequent arguments, we turned to the TOROT data, which are tagged for secondary dependencies, restricted to the Codex Marianus (OCS), which has been proofread and submitted to comprehensive consistency checks (Table 1, columns 3 and 4). We assume that tendencies in argument sharing are approximately the same in OCS and CSR.

Given the extremely low frequency of positive examples for XOBJ (1), COMP and AG (0) and low estimated adoption rates, we focus mostly on inserting secondary dependencies for the other three arguments. In order to get more positive examples for OBJ and OBL, we extracted another 91 SynTagRus sentences (12 were excluded as containing critical errors), this time requiring that they contain a potential adoption case of one of these relations. The total sample size was 346 sentences, 1103 datapoints (Table 1). The increase in positive OBJ and OBL examples after annotating the extra sample was very small, the total numbers being resp. 16 and 8. Still, we decided to investigate whether it is possible to insert secondary dependencies for these arguments, relying on the sparse data available.

---

[6] For technical reasons, the empty-slot criterion was applied in its strong form to all six relations (cf. footnote 4). This means that some potential cases of OBL adoption might not have made it into the sample, but their number is most probably negligible and could not have affected the results significantly.

Table 1. Number of potential adoption cases and adoption rate (#real adoptions/#potential adoptions) for SynTagRus (estimated on a sample of 346 sentences, see main text) and the Codex Marianus (6350 sentences)

Argument type	Adoption rate (SynTagRus sample)	Potential adoptions (SynTagRus sample)	Adoption rate (Marianus)	Potential adoptions (Marianus)
SUB	0.80	261	0.59	901
OBJ	0.06	251	0.06	895
OBL	0.02	393	0.01	1,409
XOBJ	0.01	158	0.01	472
COMP	0.00	36	0.02	94
AG	0.00	4	0.00	9

3.3. Probabilistic features

By combining our experience in tagging shared dependents manually, theoretical reasoning about argument frames, and the trial-and-error method, we identified four features expected to be the most informative about the adoption status of a given datapoint.

The first feature is the probability of a potential adopter having a potential adoptee as an argument. For the pair {он; обрадовался} in example (1), that would be the probability of the lemma обрадоваться having a subject (as opposed to it being subjectless), which is 0.70. The second one is the probability of a potential adopter having an argument frame that would consist of its own primary arguments *and* a potential adoptee. For example (1), that would be the probability of обрадоваться having a frame V+sub+obl (0.10). These data were calculated using the whole SynTagRus corpus and did not require any manual tagging of secondary dependencies.[7]

It would seem reasonable to expect that for low-frequency verbs these data can be misleading rather than useful. However, when we tried excluding these features for low-frequency potential adopters, that led to decrease in performance. In other words, very little information turns out to be better than no information (see Berdičevskis and Eckhoff 2014:11 for a similar conclusion about argument frames as an information source).

The third feature is the probability of a particular argument being an adoptee, see Table 1, column 1. We tried excluding this feature as disfavouring non-subjects too strongly (or excluding it for non-subjects only), but that led to decrease in performance, both for subjects and non-subjects.

[7] Full data on all the verbs and other data not reported in the paper can be found at the TROLLing Dataverse (http://opendata.uit.no/), hdl: 10037.1/10174.

Table 2. Adoption rate for different positions of a potential adoptee, calculated for all arguments together; subjects only; non-subjects only

Position of a potential adoptee	Adoption rate (all arguments)	Adoption rate (subjects only)	Adoption rate (non-subjects only)
FL (on the first conjunct and to the left of it)	0.75	0.93	0.17
FR (on the first conjunct and to the right of it)	0.06	0.65	0.02
M (on a middle conjunct)	0.03	0.21	0.01
LL (on the last conjunct and to the left of it)	0.01	0.06	0.00
LR (on the last conjunct and to the right of it)	0.03	0.25	0.02

The fourth feature is the probability of a potential adoptee being adopted if it is found in a given *position*. We distinguished between five positions: FL, FR, M, LL and LR (Table 2). The subjects, however, again create a strong bias: first, they are the most frequent adoptees, second, they occur most frequently to the left of the verb, third, according to SynTagRus guidelines for annotators, shared subjects have to be placed on the *first* conjunct, while other arguments have to be placed on the nearest conjunct regardless of whether they are shared or private (Leonid Iomdin 2015, p. c.). Given all that, it seemed reasonable to calculate position adoption rate separately for subjects and non-subjects, which we did (Table 2, columns 3 and 4), but using separate scores, again, led to decrease in performance, so we used data from column 2 in our final evaluation.

Values for these two features were calculated using the manually tagged sample described in Section 3.2. We tried including some other features (such as, for instance, the probability of a given conjunction allowing its children to adopt its grandchildren), but that did not lead to any increase in performance.

To sum up, for every datapoint four features were identified, whose values are probabilities between 0 and 1.

3.4. Algorithm

We devised a very simple guessing algorithm. For every datapoint in a training set, it calculates an average of the four probabilities (note the result is not a probability in itself). All datapoints with an average higher than a certain cutoff *c* are considered to be cases of real adoption, all datapoints with the average lower than c are not. The algorithm finds the optimal value of c for the training set (by trial and error, i. e. by trying all possible values of c and selecting the one which gives the highest accuracy for the training set), and then applies the calculated feature values and c to the test set.

We also trained a support vector machine using SVMlight (Joachims 1999) with a radial basis function kernel (other parameters default for SVMlight). SVM provides a slightly higher accuracy than our average-based guesser per se, but when both outputs are corrected using deterministic rules (Section 3.5), our guesser outperforms SVM, especially as regards infrequent arguments types. For this reason, SVM performance is not reported here.

3.5. Deterministic Rules

In addition to the statistical algorithms, we also implemented several deterministic rules. All rules concern only subjects and can only predict negative outcomes (i. e. the absence of adoption) that override statistical guesses. The rules are as follows: a first-person potential adopter cannot adopt subjects that have lemmas different from *я* 'I', *мы* 'we' and *сам* 'self'; a second-person potential adopter cannot adopt subjects that have lemmas different from *ты* 'thou', *вы* 'you' and *сам* 'self', unless it is in the imperative; if a potential adopter has person, number, gender or mood (relevant values of the PROIEL mood category are indicative, imperative and infinitive) different from the potential subject adoptee's real parent, then the adoption is impossible. Thus, we avoid adoptions with obvious person agreement clashes. In addition, there are some rules specific to the PROIEL format.

These rules never fail on our dataset and prevent a small number of false positives.

4. Results and discussion

The guesser's performance was evaluated using 5-fold cross-section validation. Since the classes are highly skewed (adoptions are rare), we report not only accuracy, but also F-score, precision and recall for every argument type (Table 3).

Table 3. Performance of the average-based guesser (with rule-based correction)

	Accuracy	F-score	Precision	Recall	Datapoints	Real adoptions
Overall	0.97	0.92	0.95	0.89	1,100	234
SUB	0.95	0.97	0.97	0.96	261	209
OBJ	0.96	0.31	0.40	0.28	250	16
OBL	0.97	0.21	0.20	0.25	392	8

The overall good results are mostly achieved through excellent performance on SUBs. For OBJs and OBLs, the total number of positive examples is extremely small, and hence the guesser is providing many false negatives. This results in high accuracy, but low F-scores.

Manual error analysis confirms that most false negatives are non-subjects. Typical reasons why they get low average scores are low scores for the "relation"

and "position" features. In example (2), the guesser misses a case of OBJ adoption: *дискредитировать* 'discredit' should adopt *нас* 'us'. The reason is the low adoption rate for OBJs and a low position score for FR.

(2) *Им важнее выдавить нас из страны, дискредитировать, доказать, что мы воруем деньги наблюдателей*[8]
'It is more important to them to squeeze us out of the country, to discredit [us], to prove that we are stealing the observers' money.'

When false negatives are subjects, they are almost always in a non-typical position (postverbal, or depending on a non-first conjunct, or both). For instance, the guesser misses the adoption of the postverbal SUB *она* 'she' of the first conjunct in example (3) by the verb *опустила* 'lowered':

(3) *Ты не сердись,—торопливо сказала она и опустила глаза.*[9]
'Don't be angry, she said hastily, and lowered her eyes.'

As mentioned above, excluding the "position" feature did not lead to increase in performance: even non-subjects tend to be shared more often when they are in FL position (Table 2). A possible reason is that *topical* elements are more likely to be shared.

False positives are less numerous (which is good: for linguistic uses of the corpus, it is better to miss real adoptions than to insert false ones). Interestingly, more than half of the cases identified by the guesser as false positives at the intermediate work stage turned out to be human annotation errors (i. e. cases where we should have inserted secondary dependencies when tagging the extracted sample, but failed to do so). This means that the algorithm can have a practical application as an error identification device in a manually annotated treebank, and an experimental application of the algorithm to the OCS part of TOROT has already confirmed this.

Given the high performance for SUBs, it is reasonable to expect that similar results could be achieved for OBJs and OBLs, if several hundred sentences containing potential adoption cases were annotated and thus at least several dozens positive examples for each relation were collected. With more data, there would also be more possibilities to fine-tune the features to avoid the excessive punishment of non-subject relations.

5. Conclusions

We have described a simple algorithm which allows us to identify shared arguments with high accuracy and F-score. While the performance is excellent on subjects, F-scores are low for other relations. For objects and obliques, the algorithm has a clear

[8] E. Masjuk. "Lilija Shibanova: Vladimir Vladimirovich, vy bol'ny shpionomaniej", *Novaja gazeta,* 45, 2013.

[9] Ju. Kazakov, *Dvoe v dekabre*.

potential of achieving much better results, if more sentences are manually tagged in order to collect more positive examples.

Our solution is a practical contribution, useful both for our specific purposes (the conversion of the SynTagRus to the PROIEL format and subsequent diachronic studies of verbal argument frames) and more general applications (shared subject identification can, for instance, be important for agreement studies). It can potentially contribute to theoretical linguistics, too, by providing quantitative data about some tendencies in Russian coordination. It can, for instance, be tested whether our observation that topical arguments are more likely to be shared holds.

References

1. *Apresyan Yu. D., I. M. Boguslavskij, B. L. Iomdin, L. L. Iomdin, A. V. Sannikov, V. Z. Sannikov, V. G. Sizov, L. L. Tsinman.* (2005), Syntactically and semantically annotated Russian corpus: state of the art and perspectives [Sintaksicheski i semanticheski annotirovannyj korpus russkogo yazyka: sovremennoe sostoyanie i perspektivy], Russian National Corpus [Nacional'nyj korpus russkogo jazyka: 2003–2005]. Indrik, Moscow, pp. 193–214.
2. *Berdičevskis A., H. Eckhoff.* (2014), Verbal constructional profiles: reliability, distinction power and practical applications, Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13), Tübingen, pp. 2–13.
3. *Haug D.* (2010), PROIEL guidelines for annotation. http://folk.uio.no/daghaug/syntactic_guidelines.pdf
4. *Haug D., M. Jøhndal, H. Eckhoff, E. Welo, M. Hertzenberg, A. Müth.* (2009), Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages, Traitement Automatique des Langues, Vol. 50:2, pp. 17–45.
5. *Joachims T.* (1999), Making large-Scale SVM Learning Practical, Advances in Kernel Methods—Support Vector Learning, MIT-Press, Cambridge, pp. 41–56.
6. *Mareček D., N. Kljueva.* (2009), Converting Russian Treebank SynTagRus into Praguian PDT Style, Proceedings of the Workshop on Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages, pp. 26–31.
7. *Mel'čuk I.A.* (1995), Russian Language in the Meaning-Text Model [Russkij jazyk v modeli "Smysl–Tekst"], Škola "Jazyki russkoj kul'tury", Vienna.
8. *Popel M., D. Mareček, J. Štěpánek, D. Zeman, Z. Žabokrtský.* (2013), Coordination Structures in Dependency Treebanks, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Sofia, pp. 517–527.
9. *Zeman D., D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, J. Hajič.* (2012), HamleDT: To Parse or Not to Parse?, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC›12), Istanbul, pp. 2735–2741.