Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks

Kira Droganova[1], *Olga Lyashevskaya*[2,3], *Daniel Zeman*[1]

(1) Charles University, Faculty of Mathematics and Physics, Prague
(2) National Research University Higher School of Economics, Moscow
(3) Vinogradov Institute of the Russian Language RAS, Moscow

{droganova,zeman}@ufal.mff.cuni.cz, olesar@yandex.ru

Abstract

In this paper we focus on syntactic annotation consistency within Universal Dependencies (UD) treebanks for Russian: UD_Russian-SynTagRus, UD_Russian-GSD, UD_Russian-Taiga, and UD_Russian-PUD. We describe the four treebanks, their distinctive features and development. In order to test and improve consistency within the treebanks, we reconsidered the experiments by Martínez Alonso and Zeman; our parsing experiments were conducted using a state-of-the-art parser that took part in the CoNLL 2017 Shared Task. We analyze error classes in functional and content relations and discuss a method to separate the errors induced by annotation inconsistency and those caused by syntactic complexity and other factors.

Keywords: Annotation consistency, Universal dependencies, Russian treebanks, Dependency parsing.

1 Introduction

The co-existence of several treebanks for one language, made by different teams and converted from different sources, within the Universal Dependencies (UD) platform (Nivre et al., 2016) is a key factor that makes data grow and attracts new contributors. A user can choose a treebank more appropriate for their needs, or combine data into one training set. However, the heterogeneous nature of treebanks in terms of their annotation policies, genres, etc. means that combining data is not straightforward.

During the last two decades surprisingly little attention has been paid to syntactic annotation consistency. Brants and Skut (1998) proposed using a statistical parser for the detection and labeling of phrase structures and specifying the functions of the words during annotation. The method is more of an annotation strategy rather than a mechanism for checking consistency within an already annotated treebank. Dickinson and Meurers (2003) described an approach of using n-grams for identifying errors in phrase structure trees in context: n-grams with annotations that vary in the same context signal inconsistency. Longer contexts are preferable. Kaljurand (2004) presented a tool for checking consistency by restructuring a treebank in a way that the focus is given to groups which can be formed by either words, POS-tags or syntactic labels. Statistics are extracted from the treebank for each group. After that, consistency is estimated using the context: the annotation of the group varies less in a wider context. The tool is restricted to the NEGRA format (Brants, 1997). A method independent of dependency representation was proposed by Boyd et al. (2008), who reused the concept of variation nuclei (Dickinson and Meurers, 2003) and extended it with context restrictions. Kulick et al. (2013) combined the concept of variation nuclei with the idea of decomposing full syntactic trees into smaller units, and applied it to the task of the evaluation of inter-annotator agreement. Another approach to consistency checking is the use of heuristic patterns. De Smedt et al. (2015) proposed this strategy to check the consistency of multi-word expressions within the Universal Dependencies (UD) treebanks with INESS-Search. Martínez Alonso and Zeman (2016) conducted a series of parsing experiments to determine the consistency of the AnCora Spanish and Catalan treebanks after their conversion to UD. One of the most recent works (de Marneffe et al., 2017) adapts the method proposed by Boyd et al. (2008) to the UD treebanks, tests the approach on English, French and Finnish UD treebanks and proposes an annotation tool to organize the manual effort required by the method. Another recent work by Alzetta et al. (2018) describes a method aimed at detecting erroneously annotated arcs in gold standard dependency treebanks by measuring the reliability of automatically produced dependency relations.

For the purpose of the analysis of syntactic annotation consistency within the four Russian treebanks in UD, we have chosen the method by Martínez Alonso and Zeman (2016). They conducted a series of monolingual, cross-lingual and cross-domain parsing experiments using attachment accuracies as the means to estimate consistency with other UD treebanks and between two treebanks of the same language (Spanish AnCora and 'Web' treebank, the latter now being officially labeled GSD). We consider the method useful not only for treebank contributors but also for external users, who need a simple measure to make a decision about which corpora can be jointly used in their experiments. Labeled attachment score (LAS) is a good candidate to serve as a simple criterion to find an optimal corpus mix for parsing experiments. On top of that, the method is suitable for languages with rich morphology and is context-independent; the latter is very important because the four Russian corpora comprise texts of different genres. The method proposed by de Marneffe et al. (2017)

is mostly designed for UD contributors and is not fully suitable for morphologically-rich languages[1]. However, we plan to conduct such experiment in the future.

2 UD Russian Treebanks

The UD release 2.2 includes four Russian treebanks: **SynTagRus, GSD** (Google Stanford Dependencies), **PUD** (Parallel Universal Dependencies) and **Taiga.** Only Taiga was annotated directly in the UD style; the other three treebanks were manually annotated under a different scheme and then automatically converted to UD. The original scheme of GSD and PUD (both annotated at Google) is arguably more similar to UD than SynTagRus.

2.1 UD_Russian-SynTagRus

The Russian dependency treebank, SynTagRus, was taken as the source. Until the other UD treebanks emerged, SynTagRus was the only human-corrected corpus of Russian supplied with comprehensive morphological annotation and syntactic annotation in the form of a complete dependency tree provided for every sentence (Boguslavsky et al., 2009; Dyachenko et al., 2015).

The treebank is built upon Melčuk's Meaning-Text Theory (Mel'čuk, 1981) and specifies a set of 67 dependency relations, 11 coarse-grained part-of-speech tags and a set of morphological features.

Currently the treebank contains over 1,000,000 tokens (over 66,000 sentences) belonging to texts from a variety of genres (contemporary fiction, popular science, texts of online news, newspaper and journal articles, dated between 1960 and 2016).

2.1.1 Key features of corpus development

We developed the conversion procedure using the original corpus statistics and the corpus description[2]. Clearly, the original SynTagRus annotation principles and the UD guidelines have important differences that should be explained in detail.

1. Following the UD principle that dependency relations hold primarily between content words, the tree structure has been transformed. The following nodes have been moved to dependent positions:
   (a) Prepositions (heads of prepositional phrases in SynTagRus);
   (b) Copulas and auxiliary verbs;
   (c) Coordinating conjunctions (coordinate clauses are connected via conjunctions in SynTagRus);
   (d) Subordinating conjunctions (subordinate clauses are attached via subordinating conjunctions in SynTagRus).
2. SynTagRus currently provides a set of 67 dependency relations. However, only 16 relations can be directly mapped to appropriate relations within the set of 37 UD dependency relations. The remaining majority of relations require additional information

[1] Their lemma-based approach fails on morphologically-rich languages. The results for the wordform-based approach seem promising even for morphologically-rich languages.

[2] `http://www.ruscorpora.ru/instruction-syntax.html`

concerning morphological information of the node, morphosyntactic information of the head and the dependents of that node due to differences between the original annotation and the UD guidelines:

(a) The original relations do not distinguish between core arguments and oblique dependents. The rule of thumb in Russian UD is that core arguments are bare noun phrases (without a preposition) in nominative, accusative, dative or genitive (although the latter two cases are slightly controversial and may be reconsidered in the future);

(b) The original relations do not encode information about clauses;

(c) Some of the relations in UD depend on part-of-speech tags, which is not always the case for original SynTagRus dependency relations.

3. Elliptic constructions are represented in SynTagRus by reconstructed tokens, which contain all the information a normal token would, except the word form. Regarding dependency structure, these tokens behave like normal nodes. This representation allows the development of conversion rules to generate both the basic and the enhanced UD layer.

   In the basic representation, reconstructed nodes are omitted and dependency paths that traverse them are contracted. Then the trees are rearranged according to the obliqueness hierarchy specified in the UD guidelines (Figure 1). In the enhanced representation, the reconstructed nodes are preserved.

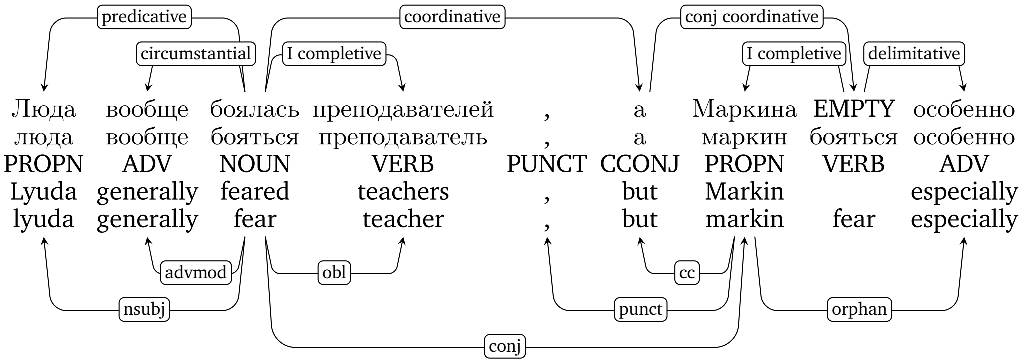
predicative
circumstantial
I completive
coordinative
conj coordinative
I completive
delimitative
Люда вообще боялась преподавателей , а Маркина EMPTY особенно
люда вообще бояться преподаватель , а маркин бояться особенно
PROPN ADV NOUN VERB PUNCT CCONJ PROPN VERB ADV
Lyuda generally feared teachers , but Markin especially
lyuda generally fear teacher , but markin fear especially
advmod
obl
cc
nsubj
punct
orphan
conj

Figure 1: An example of a Russian elliptic sentence. The upper part shows the original SynTagRus dependency tree annotated with original dependency relations. The bottom part shows the UD style converted tree. Translation: "Lyuda generally feared teachers, but Markin especially."

4. Multiword expressions are mainly re-analyzed using a manually created MWE dictionary, which includes POS tags, features and dependency relations. In SynTagRus, a multiword expression is usually placed into one node where words are separated by spaces. Therefore, it has just one part-of-speech tag, one set of features and one dependency label. Additionally, we use a rule that makes sure that the first node of the MWE is technically the head and all other member words are attached to it.

5. The UD guidelines specify a set of 21 morphological features, which can be extended by language-specific features. However, only 13 features were involved in the conversion process. The rest either do not apply to Russian or could not be reliably derived from the features in SynTagRus.

6. For most of the 11 part-of-speech tags in SynTagRus the conversion mapping is pretty straightforward. However, several heuristics were developed for tricky cases:

   (a) SynTagRus does not distinguish between content verbs and auxiliary verbs. The AUX tag is always assigned to the lemma "быть" (to be);

   (b) Determiners are converted using closed-class lists since they are annotated as adjectives in SynTagRus;

   (c) Proper nouns do not have a special tag in SynTagRus. Therefore they were separated from nouns based on orthography: a capitalized noun that does not start a sentence is likely to be a proper noun. Then a list of proper nouns was collected and manually analyzed. Finally the sentence-initial words were processed using this list;

   (d) A set of prefixoides, for instance само 'self-', полу 'half-', пост 'post-' etc., is annotated as a compound (COM pos-tag) in SynTagRus and usually stored in a separate node. For example, пол года 'half a year' would form two separate tokens. Such tokens must be concatenated into one token;

   (e) Punctuation marks are preserved in XML at the positions where they occurred in the raw text but they do not have their own token-level XML elements (Iomdin and Sizov, 2009) in SynTagRus. Therefore, we extracted punctuation from XML separately, created nodes for all punctuation symbols and assigned them the PUNCT tag directly.

2.2 UD_Russian-GSD

UD_Russian-GSD contains excerpts from Russian Wikipedia (100K tokens, 5K sentences). Its manual annotation at Google was not done directly in the UD annotation scheme; however, both the UD and the Google scheme are based on Stanford Dependencies and thus fairly close. The first set of transformations was already done by Google researchers.[3] The dependency structures and labels were checked manually. The biggest changes were in the annotations of verb patterns, parentheticals of different kind, appositives and other flat syntactic relations in nominal groups, and, predictably, long-distance relations. Since it is not uncommon that some words and text fragments are missing (as a result of fast web-crawling), we treated such cases as a special type of ellipsis and re-annotated the corresponding clauses almost in full. A few 'second order' dependency relations got simplified in UD, for example, `nmod:gobj` and `nmod:tmod` were converted into `nmod`. In contrast, the semantic subjects in passive constructions (marked by the Instrumental case) were labeled as `obl:agent`.

The treebank was not lemmatized, so lemmas were added and manually checked in subsequent releases. The feature annotation was made more consistent with the Russian National Corpus tagset and UD 2.x guidelines.

[3] Ryan McDonald and Vitaly Nikolaev

2.3 UD_Russian-PUD

The treebank is a part of the Parallel Universal Dependencies (PUD) treebanks created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies (Zeman et al., 2017).

There are 1000 sentences that are taken from the news domain and from Wikipedia. All the sentences are translations from other languages—mostly English, partially also German, French, Italian and Spanish. The same sentences have been translated to and annotated in 18 languages, hence the name *Parallel* UD. Like with GSD, the manual annotation was provided by Google and conformed to their version of the Stanford dependencies. We then converted it to the UD scheme. The treebank has not been lemmatized until v.2.3.

The entire treebank is labeled as a test set and was used for testing in the shared task. In our experiments we use the treebank as a test set as well.

2.4 UD_Russian-Taiga

UD_Russian-Taiga is a new treebank that focuses on registers not covered by the other corpora: blogs, social media, and poetry (21K tokens, 1.7K sentences). It presents constructions specific for web communication, cf. relations between root and emojis, topic hashtags, and nicknames; nominalizations, infinitive roots, and web-specific kinds of ellipsis occurring frequently in the treebank. Less standard word order such as postposed adjectives, and non-projective constructions are characteristic of poetic texts. Overall, the number of words spelled in non-standard orthography and non-standard grammatical forms is higher than in SynTagRus and other Russian UD corpora. UD_Russian-Taiga was annotated directly in UD 2.x style (specifically, it was preprocessed by UDPipe (Straka and Straková, 2017) trained on SynTagRus, then manually corrected). At the time of annotation, the three other Russian treebanks were available in UD version 2.1, so the developers of Taiga could query them for annotation of specific constructions and lexemes and make consistent decisions.

2.5 Comparison of the Treebanks

As the four treebanks come from different sources and were originally created by different teams, there are differences in how they interpret the UD guidelines. We list some of the more significant differences in this section. This is meant as a warning to the users of UD 2.2 and earlier releases, it may also help with understanding the results of the 2017 and 2018 CoNLL shared tasks. We intend to contribute better harmonized versions of the treebanks, which will hopefully become part of the UD 2.3 release in November 2018.

- There are no lemmas in PUD, and the lemmas in GSD are all-uppercase (thus not compatible with SynTagRus and Taiga). Both GSD and PUD will be automatically lemmatized with a model compatible with SynTagRus and Taiga.

- There are discrepancies in how the morphological features are used (see the Notes column in Table 1 for details).

- There are many different auxiliary verbs in GSD (which correspond to auxiliaries in English UD). The only approved auxiliary verb is быть (to be).

Feature	Values	Notes
Gender	Fem Masc Neut	
Animacy	Anim Inan	
Number	Plur Sing	
Case	Acc Dat Gen Ins Loc Nom Par Voc	No Par and Voc in PUD
Degree	Cmp Pos Sup	No Degree in PUD, no Sup in Taiga
Polarity	Neg	
Variant	Long Short	Long appears only in PUD
VerbForm	Conv Fin Inf Part	Inconsistent in PUD
Mood	Cnd Imp Ind	No Cnd in GSD and PUD
Aspect	Imp Perf	
Tense	Fut Past Pres	
Voice	Act Mid Pass	No Mid in PUD
PronType	Prs	Only in PUD and only this value
NumType	Card	Only in GSD and only this value
Reflex	Yes	Only in GSD and PUD
Person	1 2 3	
Gender[psor]	Fem Masc	Only in PUD
Number[psor]	Plur Sing	Only in PUD
Abbr	Yes	Only in Taiga
Foreign	Yes	Not in GSD, insufficient in PUD

Table 1: Features and their values in the Russian treebanks as of UD release 2.2.

- Some language-specific extensions of relation types used in PUD reflect distinctions made in the Google Stanford Dependencies, but they are unimportant in the context of Russian UD and should be reduced to the universal relation label: `cc:preconj`, `nmod:gmod`, `nmod:poss`, `obl:tmod`.

- In contrast, some language-specific relations are present in the other treebanks but missing from PUD: `nummod:entity`, `nummod:gov`, `flat:foreign`, `flat:name`. For all four there are sentences in the treebank where they could and should be used.

In summary, GSD bears some similarity with PUD (both converted from the same source annotation style) and Taiga to SynTagRus (maintained by the same people, although not of same origin: SynTagRus has its own source annotation while Taiga is natively UD). GSD has been harmonized with SynTagRus to a higher degree than PUD, effectively making PUD an outlier in many aspects. This can be probably attributed to the fact that PUD entered the UD family later and there have been no harmonization efforts beyond the initial quick-fix conversion, done hastily before the deadline for the CoNLL 2017 Shared Task.

3 Parsing Experiments

To assess the discrepancies among the four UD Russian treebanks we conducted parsing experiments that were previously described by Martínez Alonso and Zeman (2016). Martínez Alonso and Zeman describe the conversion of the Catalan and Spanish AnCora treebanks to the Universal Dependencies formalism and assess the quality of the resulting treebanks by conducting a series of monolingual, cross-lingual and cross-domain parsing experiments. Since the four Russian treebanks contain not only different genres, but also were created in principally different ways, we decided to re-use the idea of the monolingual cross-domain parsing experiment, where two models were trained for the two UD Spanish treebanks and attachment accuracies were measured using one Spanish treebank to parse the other. The idea behind this experiment is as follows. If the two treebanks were as similar as possible, the differences in parsing accuracy would be due to dataset size and domain change, and not to differences in dependency convention.

We believe that not only does this approach help to reveal annotation errors and dissimilarities, but also highlights the variety of equivalent parses.

3.1 Design

Like Martínez Alonso and Zeman, for all parsing experiments we use a single parser not to obtain state-of-the-art scores, but rather to assess the relative consistency of the Russian UD treebanks of different genres and origins. For the purpose of the parsing experiments we chose the neural graph-based Stanford parser (Dozat et al., 2017), the winner system from the CoNLL 2017 Shared Task (Zeman et al., 2017). We trained models on UD_Russian-SynTagRus and UD_Russian-GSD using the hyperparameters that were used in the CoNLL 2017 Shared Task. Due to their size, UD_Russian-Taiga and UD_Russian-PUD were used only to measure attachment accuracy. We trained one model on GSD and two models on SynTagRus. The first SynTagRus model was trained on the standard training set. The training set for the second model was made by reducing the standard SynTagRus training data to the size of the GSD training set (10% of the full SynTagRus training set). The reduced model was used to show the influence of the size of the data. Then we measured labeled attachment scores on the four parsed test sets and identified an exemplary model – the model that yields the highest scores on the four test files. Finally, for the outputs of the exemplary model, we examined pairs of parsed and gold standard files for every treebank and manually corrected the discrepancies.

Additionally, considering the variety of genres presented in the four corpora, we calculated statistics on out-of-vocabulary words[4] for every test set to ensure that the LAS scores are not affected by genre-specific or topic-specific vocabulary, slang or terminology.

3.2 Results

Table 2 shows labeled attachment scores measured on the four Russian treebanks. No model is decisively the best, but Table 2 clearly demonstrates the benefits of the size of the full SynTagRus training set. There is a great drop in performance when substituting the standard model with the reduced one. The domain change is probably not the main reason for the performance drop. The *Fixed* row shows LAS scores on manually corrected gold standard files. After manual correction, the full SynTagRus model performs better on corrected GSD/PUD test sets than the GSD model, although these test sets are out-of-domain for SynTagRus, and in-domain for GSD.

Table3 shows Content-word Labeled Attachment Score (CLAS), a parsing metric that consid-

[4] By out-of-vocabulary words we mean the words that are present in a test set, but not in the training set.

ers only relations between content words, for the model trained on the full standard training set.

Table 4 shows statistics on out-of-vocabulary words calculated by word form and by lemma[5]. The numbers show that each test set has almost the same number of unknown words. On top of that, we analyzed the top frequent unknown words for each test set and discovered that the unknown words are either different symbols for punctuation marks or proper names. Thereby the results confirm that the LAS scores are not affected by vocabulary.

Test Model	SynTagRus	Taiga	GSD	PUD
SynTagRus	**92.14%**	**69.55%**	65.01%	76.73%
– *Fixed*	–	*70.90%*	*87.41%*	*79.04%*
SynTagRus 10%	86.60%	65.35%	63.57%	74.71%
GSD	67.16%	60.29%	**81.84%**	**78.29%**

Table 2: LAS for UD_Russian-SynTagRus full, UD_Russian-SynTagRus 10% and UD_Russian-GSD models; Fixed: LAS after fixing errors in the test data.

Test Model	SynTagRus	Taiga	GSD	PUD
SynTagRus	90.61%	69.17%	65.57%	76.97%
– *Fixed*	-	70.40%	87.15%	79.49%

Table 3: CLAS for UD_Russian-SynTagRus full model.

Parameter	SynTagRus	Taiga	GSD	PUD
word form	30.18%	30.07%	30.83%	24.73%
lemma	20.49%	22.01%	**26.10%**	21.98%

Table 4: Statistics on out-of-vocabulary tokens (word forms and lemmas) in the testing sets.

3.3 Analysis

This section focuses on the mismatches in the dependency annotation of the gold standard Taiga (gold) and Taiga parsed by SynTagRus full model (predicted). The test and train treebanks represent two poles of the annotation strategy (conversion from another dependency representation vs. annotation from scratch).

It is interesting to compare the annotation consistency of the corpora in terms of functional and content relations. While the SynTagRus model fails to predict the functional relations only 5% of the time in Taiga (cf. 30% on all relations), labeling of `aux` relations is still poor (14%) due to the fact that the subjunctive marker бы 'would' is not treated as auxiliary in SynTagRus. Similarly, a number of words do not fall under the category of `case`, `det`, and `cc` (eg. как 'as', всякий 'any'; also in specific patterns: '–' in *2–4* '2 to 4', то... то... 'now...

[5] UD_Russian-PUD did not contain lemmas in UD 2.2. Therefore we added lemmas using the part of the pipeline(Mediankin and Droganova, 2016) designed specifically for Russian and manually checked them.

then...'). On the contrary, errors in `mark` and `cop` are due to the parser's failure to recognize the structure and usually more distant head.

Of content relations, the cross-parsing model misclassifies 32% inter-clausal, `conj`, and `parataxis` relations; 25% verb arguments; and 25% NP arguments. We can assume that the frequent errors in `parataxis` (55%) are due to complicated and irregular structures of the sentences; however, in some cases, it is a consequence of differing strategies in the annotation of different kinds of parenthesis (cf. `parataxis` vs. `appos` vs. `discourse`). Besides that, there are also genre-specific paratactic patterns in Taiga not seen or underrepresented in SynTagRus such as numbered and bulleted lists, or emoticons that follow or precede the utterances. As for typical verb-argument relations, corpora are inconsistent in the annotation of the 2nd and 3rd arguments, as expected, and in labeling depictive constructions (`xcomp`), whereas the high rate of unlabeled attachment (UA) errors in `nsubj` is an indicator that the model performs poorly on verb-less sentences (underrepresented in SynTagRus). Moreover, errors in `nmod` labeling reveal two problematic constructions: nouns with infinitive and bare Dative complement. Both SynTagRus and Taiga seem to be consistent in their annotation, but in SynTagRus only one noun (смысл 'sense') shows up with infinitives, and Dative cannot be seen without a preposition (cf. Флаг тебе.Dat в руки 'do just as you like', lit. 'flag to you.Dat in hands' typical of net communication genres in Taiga).

Finally, the relations `fixed` and `flat` are not predicted in 40% and 47% cases, respectively. We can see that the list of multi-word expressions is larger in Taiga, as it was updated during the manual annotation (cf. the conjunction не только 'not only', parenthetic в принципе 'in principle', preposition что до 'as for', among others). In Taiga, `flat` is applied to words repeated two or more times (e.g. всех всех 'all', не не не 'no no no'; also in the idiomatic construction уж кто кто, а... 'I don't know about anyone else, but...'), the pattern which the SynTagRus model fails to parse the same way. The error rate for punctuation is considerably low (2%) compared to those for flat relations, but analysis shows a difference in the attachment of terminal symbols and commas in nested clauses in train and test corpora.

It should not be forgotten, however, that the parser's errors can be caused not only by the inconsistency in corpus annotation per se but also by different distribution of text genres, syntactic structures, and lexicon (cf. abundance of abbreviated words and misspellings in Taiga not presented SynTagRus). The classification of cross-parsing errors by relation types suggests some heuristics to distinguish annotation inconsistency among other reasons and reveal the patterns tagged differently in different corpora. If we consider the ratio of UA errors and LA errors (i.e. cases in which the arcs are attached correctly but the labels are predicted incorrectly), then numbers close to zero usually signal inconsistency in annotation and mismatches in lexical inventories. The predominance of UA errors over LA errors may suggest that the target construction is unlikely to have been present in the training corpus. Nevertheless, in the case of the flat relation `fixed`, the mismatch in patterns covered by the annotation guidelines leads to the parser's inability to guess the unlabeled structure and thus the ratio of UA to LA errors is high.

The limitation of the method is that the UA/LA error ratio seems to be different for different relations and depends on the total LAS score in a particular pair of train and test treebanks. The complexity of parsed structures and the degree of overlap of lexical units labeled by a particular relation may also affect the UA/LA ratio. Overall, more work should be done to estimate the impact of the training and test corpus size on the findings.

Last but not the least, analysis of cross-parsing mismatches helps to reveal a set of ambiguous patterns in the treebank, i.e. those that can be parsed in different ways, sometimes with certain difference in interpretation. As an example, cf. the label `ccomp` (gold) and `acl` (parsed) on the verb платили 'pay' in главное[head], чтобы платили[ccomp vs. acl], и рейтинг рос... 'the main thing[head] is to be paid[ccomp vs. acl], and to get a higher rating...'.

4 Future Work

According to our observations the following changes need to be made to improve consistency across the four treebanks:

- fix punctuation attachment; revise the `parataxis` vs `ccomp` distinction in reported speech patterns; revise the annotation of adverbs in UD_Russian-SynTagRus;
- revise the annotation of auxiliary verbs; revise lemmatization; revise the morphological features in UD_Russian-GSD;
- revise the annotation of a number of specific patterns in UD_Russian-Taiga;
- add lemmas; revise language-specific syntactic relations; revise the morphological features in UD_Russian-PUD.

5 Conclusion

We have presented Russian UD treebanks: UD_Russian-SynTagRus, UD_Russian-GSD, UD_Russian-Taiga, and UD_Russian-PUD and described a series of experiments aimed at checking syntactic annotation consistency within the four treebanks. We have extended the method previously proposed by Martínez Alonso and Zeman (2016) to confirm that the LAS scores are not affected by genre-specific or topic-specific vocabulary. We have presented the LAS scores which can serve as criteria for deciding on the optimal corpus mix for parsing experiments. The analysis of mismatches in the test and predicted relations reveals a list of patterns that could be annotated more consistently in the four corpora.

Acknowledgments

The work by Kira Droganova and Daniel Zeman was partially supported by the GA UK grant 794417, the SVV project number 260 453, the grant 15-10472S of the Czech Science Foundation (GAČR), and FP7-ICT-2009-4-249119 (MŠMT 7E11040). Research by Olga Lyashevskaya was done within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2018 (grant 18-05-0047) and by the Russian Academic Excellence Project «5-100».

References

Alzetta, C., Dell'Orletta, F., Montemagni, S., and Venturi, G. (2018). Dangerous relations in dependency treebanks. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 201–210, Praha, Czechia.

Boguslavsky, I., Iomdin, L., Timoshenko, S., and Frolova, T. (2009). Development of the russian tagged corpus with lexical and functional annotation. In *Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop. Proceedings. Bratislava, Slovakia*, pages 83–90.

Boyd, A., Dickinson, M., and Meurers, W. D. (2008). On detecting errors in dependency treebanks. *Research on Language and Computation*, 6(2):113–137.

Brants, T. (1997). The negra export format for annotated corpora. *University of Saarbrücken, Germany*.

Brants, T. and Skut, W. (1998). Automation of treebank annotation. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pages 49–57. Association for Computational Linguistics.

de Marneffe, M.-C., Grioni, M., Kanerva, J., and Ginter, F. (2017). Assessing the annotation consistency of the universal dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 108–115.

De Smedt, K., Rosén, V., and Meurer, P. (2015). Studying consistency in ud treebanks with iness-search. In *Proceedings of the Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 258–267.

Dickinson, M. and Meurers, W. D. (2003). Detecting inconsistencies in treebanks. In *Proceedings of TLT*, volume 3, pages 45–56.

Dozat, T., Qi, P., and Manning, C. D. (2017). Stanford's graph-based neural dependency parser at the conll 2017 shared task. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

Dyachenko, P., Iomdin, L., Lazursky, A., Mityushin, L., Podlesskaya, O., Sizov, V., Frolova, T., and Tsinman, L. (2015). Sovremennoe sostoyanie gluboko annotirovannogo korpusa tekstov russkogo yazyka (syntagrus). *Trudy Instituta Russkogo Yazyka im. V. V. Vinogradova*, (6):272–300.

Iomdin, L. and Sizov, V. (2009). Structure editor: a powerful environment for tagged corpora. *Research Infrastructure for Digital Lexicography*, page 1.

Kaljurand, K. (2004). Checking treebank consistency to find annotation errors. Technical report at ResearchGate, `https://www.researchgate.net/publication/265628472_Checking_treebank_consistency_to_find_annotation_errors`.

Kulick, S., Bies, A., Mott, J., Maamouri, M., Santorini, B., and Kroch, A. (2013). Using derivation trees for informative treebank inter-annotator agreement evaluation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 550–555.

Martínez Alonso, H. and Zeman, D. (2016). Universal dependencies for the ancora treebanks. *Procesamiento del Lenguaje Natural*, (57):91–98.

Mediankin, N. and Droganova, K. (2016). Building NLP pipeline for russian with a handful of linguistic knowledge. In Chernyak, E., Ilvovsky, D., Skorinkin, D., and Vybornova, A., editors, *Proceedings of the Workshop on Computational Linguistics and Language Science*, pages 48–56, Aachen, Germany. NRU HSE, CEUR-WS.

Mel'čuk, I. A. (1981). Meaning-text models: a recent trend in soviet linguistics. *Annual Review of Anthropology*, 10(1):27–62.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökırmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Fernandez Alcalde, H., Strnadova, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.