M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A

P R A H A

Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies

Kira Droganova, Daniel Zeman

ÚFAL Technical Report
TR-2016-60

ISSN 1214-5521

U N I V E R S I T A S C A R O L I N A P R A G E N S I S

Copies of ÚFAL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czechia

or can be obtained via the Web: http://ufal.mff.cuni.cz/techrep

Abstract

This report presents the Universal Dependency (UD) annotated corpus for Russian and a conversion process which was developed to transform SynTagRus, the Russian dependency treebank, into a UD-style annotated corpus. The aim of this work was to create a UD-style annotated corpus for Russian since no such corpus was available prior to UD release 1.3.

The conversion rules were based on manually analyzed examples and statistics that were extracted from the corpus. The conversion itself was made automatically.

Firstly, we provide a brief description of the SynTagRus treebank. Next, we present a detailed description of the conversion process and some ideas of our future work.

Acknowledgements

We are grateful to Leonid Iomdin for assistance with the license for the UD-style corpus that made the corpus available to the UD community.

Contents

Introduction

Substantial attention has been paid to cross-linguistic research in recent years. This has encouraged the development of multilingual parsers and dependency parsing systems which can be easily applied to a wide range of languages.

The first considerable results on multilingual parsing were presented at the Conference on Computational Natural Language Learning in 2006 (Sabine Buchholz and Erwin Marsi, 2006 [1]) and in 2007 (Nivre et al., 2007 [2]). At the same time, special attention was paid to the annotation and structure of treebanks which were chosen for shared tasks. Both papers describe additional efforts that were made to convert the data into a similar format. Due to the variety of annotation principles, similar structures of closely related languages can look contrasting in different treebanks: "Linguistic diversity makes our life harder"[1].

Simultaneously, a number of independent projects on developing universal annotation were begun. The idea that a set of syntactic POS categories exist in similar forms across languages underlay the Google Universal Part-of-Speech Tagset that was developed to standardize best-practices (Petrov et al., 2011 [3]). Another remarkable project, the Interset, stressed the importance of unified annotation, both for humans and NLP frameworks, and proposed a universal feature set (Zeman, 2008 [4]). Later, this approach was integrated into the HamletDT project, which is a compilation of transformed dependency treebanks that share the same annotation style (Zeman, 2014 [5]). Another project that should be mentioned is The Stanford Dependencies, which was originally developed as a representation of English syntax for better processing of natural language understanding (NLU) applications and was transformed into a taxonomy that has a set of grammatical relations as its basis. These relations can be extended to language specific relations, which capture distinctive features of individual languages or language families (de Marneffe, 2014 [6]).

The most promising initiative, the Universal Dependencies (UD), integrates previous efforts and provides guidelines for cross-linguistically consistent grammatical annotation (Nivre et al., 2016 [7]). UD has a huge practical value for modern linguistic applications and can be employed as a universal grammar for natural language processing. Starting from the point of 10 UD-style annotated treebanks, the project shows significant growth: the latest release (1.3) contains 54 treebanks in 40 languages [8].

The aim of this work was to create the UD-style annotated corpus for Russian. The corpus is publicly available starting from release 1.3.

[1] J. Nivre. Towards a Universal Grammar for Natural Language Processing. Talk in Yandex, Moscow, April 2016

SynTagRus Treebank

The UD style annotated corpus was converted from the SynTagRus dependency treebank.

This treebank is being developed by the Computational Linguistics Laboratory of the Kharkevich Institute of Information Transmission Problems in the Russian Academy of Sciences, located in Moscow.

It is an integrated but fully autonomous part of the Russian National Corpus developed through a nationwide research project and can be freely consulted on the internet [9].

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, dating between 1960 and 2016).

To date, the treebank is 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., [10], [11]).

Conversion

Data Format

The original SynTagRus markup is XML that contains the following elements[2] (Leonid Iomdin and Victor Sizov, 2009 [12]):

S element indicates sentence borders (<S> </S>)
S element contains ID attribute (sentence number)
W element contains a word form (<W> </W>).
Punctuation is stored in-between W elements. Morpho-syntactic information for certain word form is stored in 5 attributes:

- **ID:** word number in the sentence.
- **Lemma:** dictionary entry form.
- **Feat:** set of morphological features.
- **Link:** syntactic relation, which can be omitted when the word is the head of the whole sentence.
- **Dom:** ID of the head (parent) token. The value of this attribute equals "_root" if the word is the head of the sentence.
- **Nodetype:** this attribute, if present, has the value "FANTOM", indicating an empty node.

Here is an example of an original SynTagRus annotation for the Russian sentence "Маша бегает"/"Masha is running around":

```
<S ID="1" >
<W DOM="2" FEAT="S ЕД ЖЕН ИМ ОД" ID="1" LEMMA="МАША“ LINK="предик">Маша</W>
<W DOM="_root" FEAT="V НЕСОВ НЕПРОШ ИЗЪЯВ 3-Л ЕД" ID="2"
LEMMA="БЕГАТЬ">бегает</W>
</S>.
```

The UD project uses a revised version of the CoNLL-X format (CoNLL-U) [13]. The CoNLL-U format contains following information:

ID: Indicates the position of a token in a sentence (integer, starting at 1 for each new sentence).
FORM: Word form or punctuation mark.
LEMMA: Lemma or stem of word form.
UPOSTAG: Universal part-of-speech tag.
XPOSTAG: Language-specific part-of-speech tag.
FEATS: List of morphological features separated with |.
HEAD: Head of the current token, which is either a value of the head token ID or zero (0) if the current token is the root of the whole sentence.
DEPREL: Universal dependency relation to the HEAD (root if HEAD = 0) or a defined language-specific subtype of the relation.
DEPS: List of secondary dependencies (head-deprel pairs).

[2] The description of the whole xml tree structure and elements that contain metadata were omitted.

MISC: Any other additional information.

Here is an example of the CoNLL-U format for the Russian sentence “Маша бегает”/“Masha is running around”:

```
1	Маша	МАША	NOUN	_	Animacy=Anim|Case=Nom|Gender=Fem|Number=Sing	2	nsubj	_	_
2	бегает	БЕГАТЬ	VERB	_	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act	0	root	_	_
3	.	.	PUNCT	.	_	2	punct	_	_
```

The transformation of the syntactic tree structures and replacements of labels were made in the original SynTagRus XML markup. The final conversion into the CoNLL-U format was made at the end of conversion process after the corpus had been fully annotated with UD dependency labels, part of speech tags and features.

Conversion scripts were written in Python 3.

Algorithm

The main idea that underlies the conversion procedure is the language patterns principle. We assumed that, using corpus statistics, it is possible to discover basic conversion rules that cover the majority of sentences. The remaining minority of sentences are treated as annotation inaccuracies or non-frequent sentence patterns that can be deferred until a future release.

Accordingly, we omitted sentences with foreign (non-Russian) words and elliptic constructions, which amounted to approximately 2800 sentences.

Elliptic constructions are marked with empty “FANTOM” nodes in SynTagRus. The UD guidelines disallow empty nodes. However, the guidelines are under revision on this issue now. The conversion of elliptic constructions can be deferred until the next version of the guidelines.

Details on foreign words are provided in the section entitled “Composites and Foreign Words”.

In addition, we marked relations that do not correspond to basic patterns or descriptions [9] as “dep”. Details on syntactic relations are provided in the section “Syntax”.

The conversion process is represented by 4 stages.

The first stage is the preparation of the corpus for the conversion. Prepositions are marked as the heads of the nouns in the SynTagRus corpus. Contrary, according to the UD guidelines, the heads need to be switched to the nouns and the prepositions need to be shifted to the dependent positions. Conjunctions that are marked as heads of the sentences need to be shifted to dependent positions.

The second stage is designed to perform the conversion of syntactic relations. The conversion of the syntactic annotation is conducted in two steps. The first step is designed to assign correct head-dependent positions and to provide tree structural transformations if necessary[3]. The second step of the stage assigns UD dependency labels. Details on conversion patterns and mappings are provided in the section "Syntax".

The third stage performs the conversion of morphological annotation. Details on part-of-speech tags and feature mappings are provided in the section "Morphological Annotation".

The final stage provides the conversion from the original SynTagRus markup into the CoNLL-U format. At this stage, punctuation marks are extracted from the original markup and inserted into the target format.

Some additional checks were implemented to control the conversion process. For instance, there is a check on the "mwe" relation direction. Another check provides correct "det" relation assignment.

[3] The UD guidelines emphasize the role of content words and that idea sometimes goes against the SynTagRus annotation scheme (prepositions, coordination, etc.). In these cases, the transformation of syntactic tree structures is required.

Morphological Annotation

The morphological description of a word consists of POS tag and a set of features. Currently, 17 coarse grained part of speech tags are specified in the UD guidelines. SynTagRus treebank uses a set of 11 coarse grained part of speech tags. For the most part, POS tags were converted from the SynTagRus tagset into the UD tagset without any additional transformation (Table 1).

The symbol "–" indicates that it is impossible to conduct a direct conversion. In these cases, additional rules were used.

Conjunctions

SynTagRus does not distinguish between coordinating conjunctions and subordinating conjunctions on the morphology layer. This information is stored in dependency relations. Two lists were created to distinguish between these two types of conjunction during the conversion process:

> Coordinating conjunctions: *и, да, или, либо, тоже, также, притом, причём, а, но, зато, однако, же;*
> Subordinating conjunctions: *что, чтобы, как, когда, лишь, едва, где, куда, откуда, столько, настолько, так, словно, будто, точно, если, хотя.*

However, this solution is used only for the first release. It is planned to improve the rules by adding the dependency relation labels.

Pronouns

Pronouns are marked as nouns, adjectives and adverbs in SynTagRus, thus they were converted using the list as well: *я, ты, он, она, оно, они, мы, вы, себе, кто, что, никто, ничто, некто, нечто*

Determiners

Usually, determiners are marked as adjectives in SynTagRus. Two rules were created to detect determiners.

Firstly, the lemma or the part of the lemma that starts from "-" must be in the list: *который, такой, какой, весь, каждый, всякий, некоторый, никакой, некий, сей, чей, -либо, -нибудь, -кое-, этот, тот, мой, твой, ваш, наш, его, ее, их, свой.*

Secondly, a noun must be the head of the supposed determiner. This rule distinguishes between "Он увидел ее." / "He saw her." and "Он увидел ее дом." / "He saw her house."

Proper Nouns

Proper nouns were recognized by upper case letters and the part of speech tag, which must be S (noun).

The rule is not applied to the first word of a sentence.

This solution is used only for the first release. It is planned to improve the rule by adding the feature which tries to detect a word tagged as “PROPN” that is equivalent to the first word of a sentence and assign “PROPN” to the first word of the sentence as well.

Auxiliary Verbs

SynTagRus does not distinguish between verbs and auxiliary verbs. A set of rules was developed to assign AUX POS tag.

Lemma must be “БЫТЬ” and the head of the supposed auxiliary verb must be one from the following list:

> a verb in the infinitive form (“Я не буду акцентировать внимание на этом.”/“I will not emphasize that.”);
> a passive participle (“Сундук был заброшен в море.”/“A chest was thrown into the sea.”);
> an adjective (“Его сон был не глубок.”/“His sleep was shallow.”).

Copulas are not tagged as “AUX” under the current UD guidelines. This solution is made with the reference to the future changes of the guidelines: it has been proposed that copulas become “AUX” in UD 2.

Punctuation

Punctuation was extracted from XML tree at the final stage of the conversion.

Punctuation marks are: “.” (fullstop), “,” (comma), “:” (colon), “;” (semicolon), “!” (exclamation mark), “?” (question mark), “()” (parenthesis), “«» (quotation marks), “–” (hyphen) and any combination of “.”, “!” and “?”.

For instance, “?!” or “...” are recognized as one token.

Symbols

Symbols were converted using the list: “%”, “$”, “№”, “°”, “€”, “+”, “=”, “№№”

Composites and Foreign Words

COM and NID tags do not have equivalent tags in UD POS tags. COM represents a word that cannot be used independently, it is always used in a compound. NID represents a foreign word or non-word expression, for instance, “ТУ-134”.

The UD guidelines place this information into syntactic labels. For this reason 790 sentences that contain COM and NID tags were omitted.

Foreign words will be included in the next release. They well be marked as “X”. The proper tags will be extracted for words tagged as “COM” as well.

UD	SynTagRus	Type
ADJ	A	adjective
ADV	ADV	adverb
ADP	PR	preposition
AUX	–	auxiliary verb
CONJ	CONJ	coordinating conjunction
DET	–	determiner
INTJ	INTJ	interjection
NOUN	S	noun
NUM	NUM	numeral
PART	PART	particle
PRON	–	pronoun
PROPN	–	proper noun
PUNCT	–	punctuation
VERB	V	verb
SCONJ	–	subordinating conjunction
SYM	–	symbol
X	–	other or unknown
–	COM	composite
–	NID	foreign text or non-word expression

Table 1. Part-of-speech tag mapping from SynTagRus to UD.

Features

The UD guidelines specify a set of 17 morphological features which can be extended by language-specific features. However, only 12 features were involved in the conversion process of SynTagRus, including one language-specific (Variant). Features were converted from the SynTagRus tagset into the UD tagset without any additional transformation (Table 2).

Symbol "–" indicates a lack of a feature in the SynTagRus treebank. At the same time, it indicates the presence of a feature according to the UD guidelines. For instance, if a verb does not have "страд" in its feature set, "Voice=Act" need to be assigned.

UD Feature	UD Feature value	Syntagrus
Animacy	Anim: animate	од
	Inan: inanimate	неод
Aspect	Imp: imperfect aspect	несов
	Perf: perfect aspect	сов
Case	Nom: nominative	им
	Gen: genitive	род
	Dat: dative	дат
	Acc: accusative	вин
	Ins: instrumental	твор

UD Feature	UD Feature value	Syntagrus
	Loc: locative	пр
		местн
	Voc: vocative	зв
	Par: partitive	парт
Degree	Pos: positive, first degree	–
	Cmp: comparative, second degree	срав
		смяг
	Sup: superlative, third degree	прев
Gender	Masc: masculine gender	муж
	Fem: feminine gender	жен
	Neut: neuter gender	сред
Mood	Ind: indicative	изъяв
	Imp: imperative	пов
Number	Sing: singular number	ед
	Plur: plural number	мн
Person	1: first person	1-л
	2: second person	2-л
	3: third person	3-л
Tense	Past: past tense	прош
	Pres: present tense	наст
	Fut: future tense	непрош
Variant	Short: short form of adjectives	кр
VerbForm	Fin: finite verb	–
	Inf: infinitive	инф
	Part: participle	прич
	Trans: transgressive	деепр
Voice	Act: active voice	–
	Pass: passive voice	страд

Table 2. Morphological feature mapping

Syntax

UD defines a set of 40 dependency relations that can be extended to language specific relations as well. Currently, the SynTagRus treebank provides a set of 67 dependency relations. To convert the dependency relations, we developed a set of rules that can be split into three types.

Simple rules

Simple rules provide the direct mapping from one set to another without any additional efforts (Table 3).

It should be mentioned that “neg” is converted at the last stage, if the token has “НЕ” in the lemma xml attribute.

#	Full Syntagrus name	Syntagrus label	UD label
1	Агентивное	агент	nmod:agent
2	Квазиагентивное	квазиагент	nmod
3	Несобственно-агентивное	несобст-агент	nmod
4	Элективное	электив	nmod
5	Сентенциально-предикативное	сент-предик	expl
6	Адресатно-присвязочное	адр-присв	nmod
7	Определительное	опред	amod
8	Описательно-определительное	оп-опред	acl
9	Релятивное	релят	acl:relcl
10	Атрибутивное	атриб	nmod
11	Композитное	композ	compound
12	Обособленно-аппозитивное	об-аппоз	appos
13	Нумеративно-аппозитивное	нум-аппоз	nummod:appos
14	Количественное	количест	nummod
15	Аппроксимативно-количественное	аппрокс-колич	nummod
16	Распределительное	распред	nmod
17	Аддитивное	аддит	nmod
18	Кратно-длительное	кратно-длительн	nmod
19	Дистанционное	дистанц	nmod
20	Обстоятельственно-тавтологическое	обст-тавт	nmod
21	Субъектно-обстоятельственное	суб-обст	nmod
22	Объектно-обстоятельственное	об-обст	nmod
23	Субъектно-копредикативное	суб-копр	acl
24	Объектно-копредикативное	об-копр	acl
25	Ограничительное	огранич	advmod
26	Вводное	вводн	parataxis
27	Изъяснительное	изъясн	parataxis
28	Разъяснительное	разъяснит	parataxis

#	Full Syntagrus name	Syntagrus label	UD label
29	Количественно-вспомогательное	колич-вспом	compound
30			neg

Table 3. Simple rules

Medium-level rules

Medium-level rules provide the mapping according to the POS tag. These rules do not involve tree structure transformations, but sometimes rely on a dependency relation of a token's head or dependent.

The description of medium-level rules is provided below (Table 4):

1. Relation 1 “предик” is converted into “nsubjpass” if the token or the head of this token has “СТРАД” in feats. Otherwise, the correct label is “nsubj”.
2. Relation 2 and 3 are converted into “iobj” if the token has “ДАТ” in feats. Otherwise, the correct label is “nmod”.
3. Relations 4–6 are converted into “ccomp” if the token and the head of this token have “V” in feats and the head of this token does not have “ПРИЧ” and the token has a dependent that has “предик” as the relation.
   If the token does not have such a dependent, the correct dependency label is “xcomp” (Figure 1). Otherwise, the correct label is “nmod”.
4. Relation 7 “примыкат” is converted into “appos” if the head of this token has the same feature set. Otherwise, the correct label is “parataxis”.
5. Relation 8 “аппоз” is converted into “name” if the token has “ОД” in feats and the word form starts with a capital letter. Otherwise, the correct label is “appos”. “Appos” is always assigned if the token is the first token of a sentence.
6. Relation 9 “ном-аппоз” is converted into “appos” if the word form starts with a capital letter. Otherwise, the correct label is “nmod”.
7. Relations 10–14 are converted into UD dependency labels according to the POS tags. For instance, “кратн” is converted into “xcomp” if the token has “V” in feats.

#	Full Syntagrus name	Syntagrus label	UD label
1	Предикативное	предик	nsubjpass/nsubj
2	Дательно-субъектное	дат-субъект	iobj/nmod
3	Неактантно-комплетивное	неакт-компл	iobj/nmod
4	1-е несобственно-комплетивное	1-несобст-компл	nmod/ccomp/xcomp
5	2-е несобственно-комплетивное	2-несобст-компл	nmod/ccomp/xcomp
6	3-е несобственно-комплетивное	3-несобст-компл	nmod/ccomp/xcomp
7	Примыкательное	примыкат	appos/parataxis
8	Аппозитивное	аппоз	appos/name
9	Номинативно-аппозитивное	ном-аппоз	appos/nmod
10	Количественно-копредикативное	колич-копред	nummod/nmod/advmod
11	Количественно-ограничительное	колич-огран	advmod/nmod
12	Длительное	длитель	nmod/advcl

#	Full Syntagrus name	Syntagrus label	UD label
13	Комплетивно-аппозитивное	компл-аппоз	amod/nummod/nmod/acl
14	Кратное	кратн	amod/nummod/nmod/advmod/xcomp

Table 4. Medium-level rules

Example: ...ее созидатели поставили себе целью максимально запутать... / "her creators set themselves a goal to maximally complicate…"

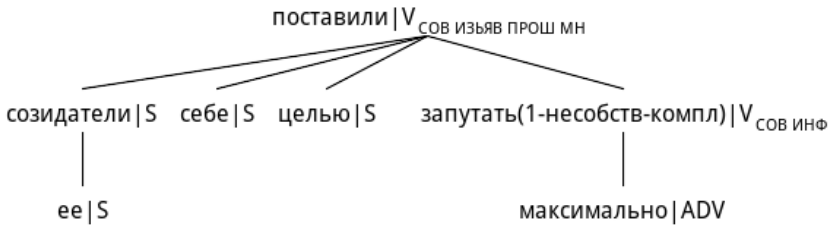

Figure 1. An example of 1-несобств-компл (original structure)

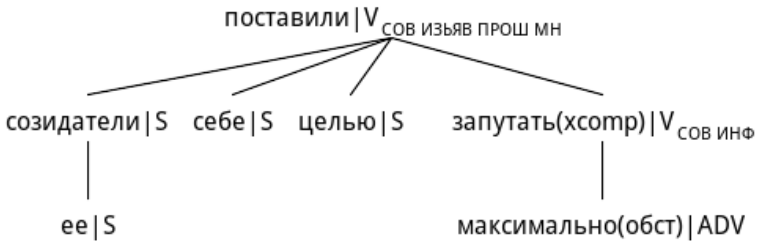

Figure 2. An example of the transformation of 1-несобств-компл into xcomp

Complex rules

Complex rules perform a transformation of a sentence structure taking into account the POS tag and the dependency label of the head or the dependent (Table 5). The conversion is conducted in two steps: a transformation of a sentence structure and a label modification. The order of the processing is important.

1. Relations сочин, сент-соч and соч-союзн are converted into "cc" or "conj" (Figure 3, 4, 5, 6 and 7).

Example: ... *одет в малиновую ливрею с галунами, белые чулки и туфли с пряжками* "dressed in crimson livery with gold braid, white stockings and shoes with buckles"

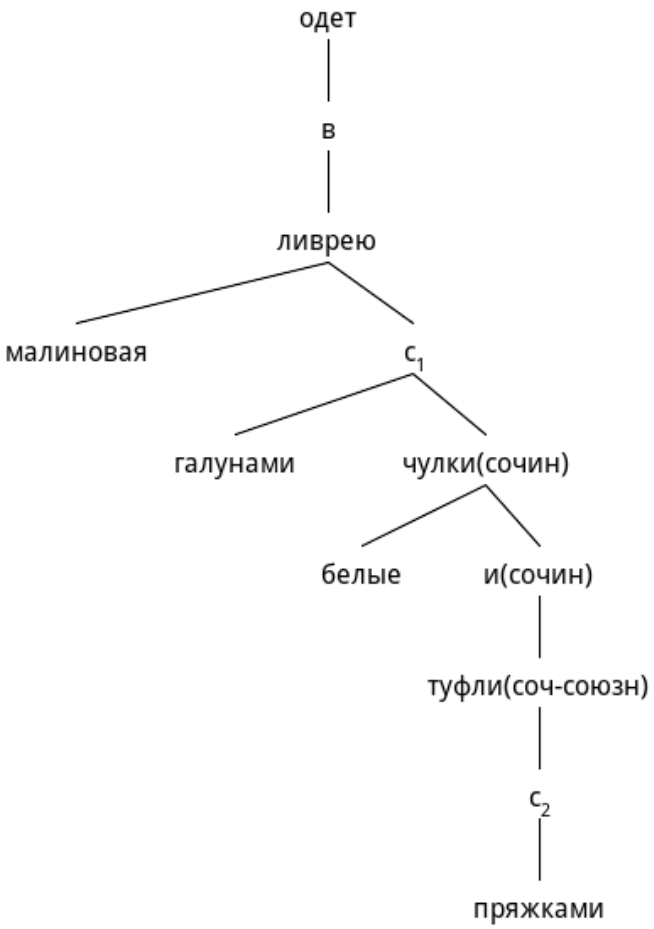

Figure 3. An example of сочин and соч-союзн (original structure)

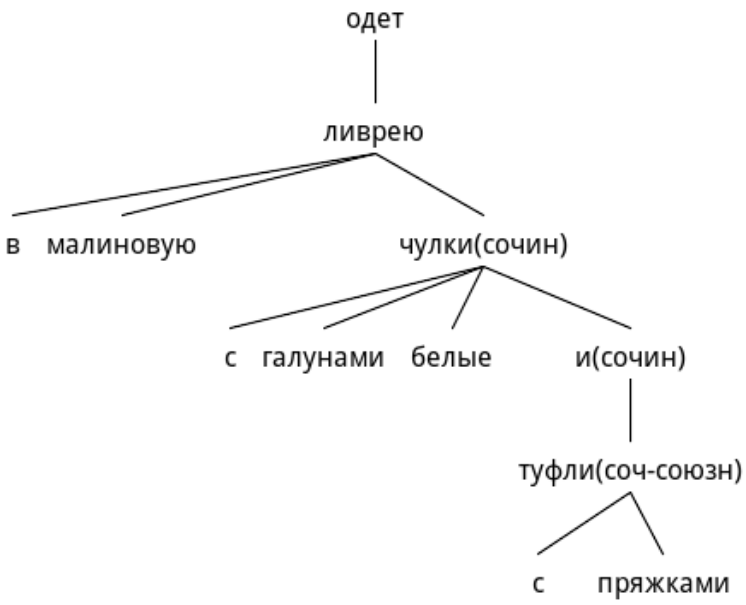

Figure 4. An example of the processed structure

Figure 5. An example of the transformation of союзн/соч-союзн into cc/conj

Example: *Кулиса была его, он всегда здесь стоял.* "The side scene was his, he always stood there."

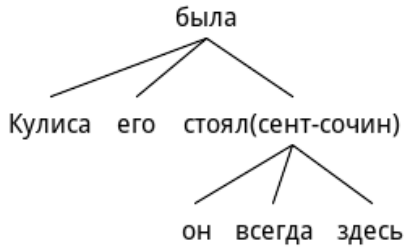

Figure 6. An example of сент-сочин (original structure)

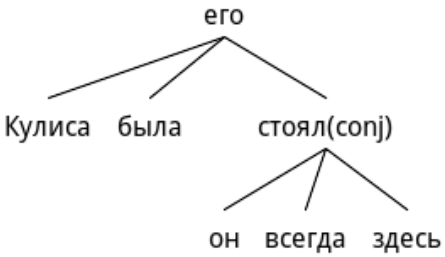

Figure 7. An example of the transformation of сент-сочин into conj

First, the sub-tree of coordination relations is detected. The root of the sub-tree is the token that does not have "сочин", "сент-соч" or "соч-союзн" as the relation and the dependents of this token have one of these relations.

Next, the chains are detected. The chain is the complete path from the start token to the leaf (the last token that has one of coordination relations).

Finally, the elements of the chains, starting from the second element, are moved to the first element of the chain which is, apparently, the start token. The dependency label is assigned according to the POS tag of the moved element. Conjunctions are converted into "cc". Otherwise, the correct relation is "conj".

2. Relations подч-союзн and инф-союзн are converted into “advcl”. If the head token is the head of the sentence, the sentence does not require additional transformations. If the head token has “cc” or “conj” as the relation, the head of the token is switched to the parent of the head token.

Example: *Сколько раз я говорил, чтобы запирали дверь.* “How many times have I said that the door was to be locked.” (Figure 8, Figure 9)

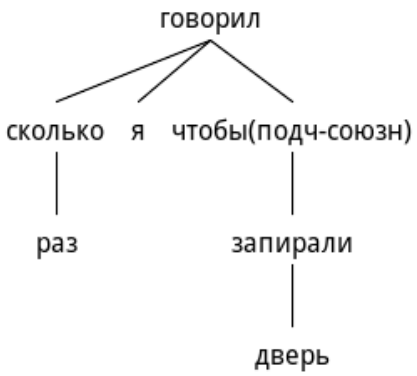

Figure 8. An example of подч-союзн (original structure)

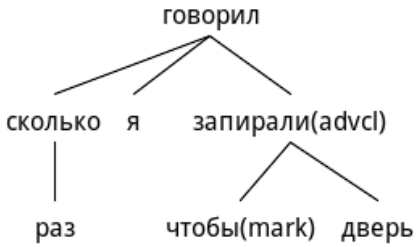

Figure 9. An example of the transformation of подч-союзн

Example: *Только так, чтобы вечерами быть дома.* “Only to be at home in the evenings.” (Figure 10, Figure 11)

Figure 10. An example of инф-союзн (original structure)

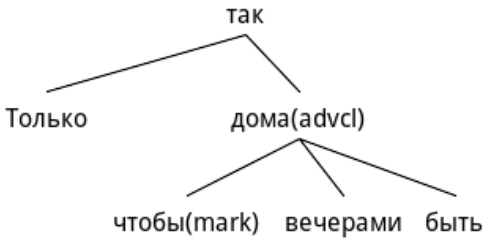

Figure 11. An example of the transformation of инф-союзн

Otherwise, the head of the token is switched to the parent of the head token, the head token and its dependents are shifted to the token and the (ex) head token is converted into "mark".

3. Relation соотнос is converted into "cc" if the token and the head token have "CONJ" in feats. In this case, the transformation is required: the head of the token is switched to the parent of the head token.

Otherwise, the dependency relation is assigned according to the POS tag of the token.

Example: *Кроме того правительство должно быть компьютеризировано с головы до ног.* "In addition the government should be computerized from head to toe." (Figure 12, Figure 13)

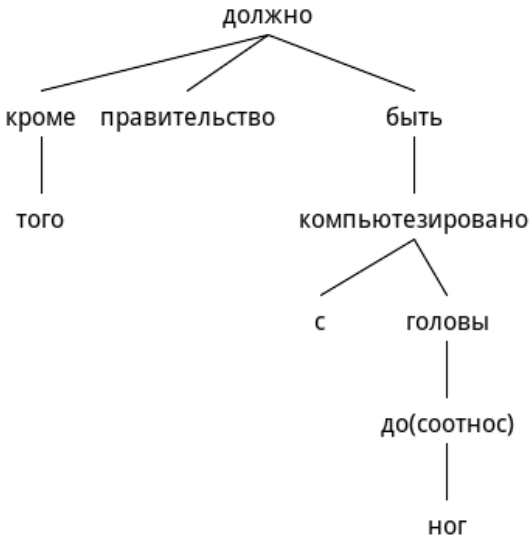

Figure 12. An example of соотнос (original structure)

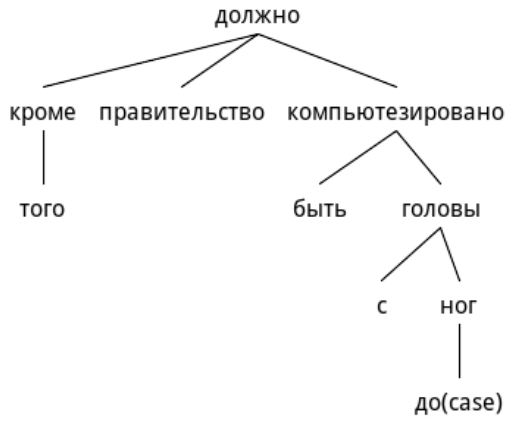

Figure 13. An example of the transformation of соотнос

Example: *За это их били, и Эренбурга и Померанцева.* "For this they were beaten, both Erenburg and Pomerantsev."

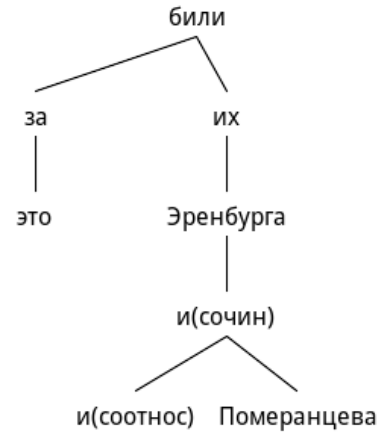

Figure 14. An example of соотнос (original structure)

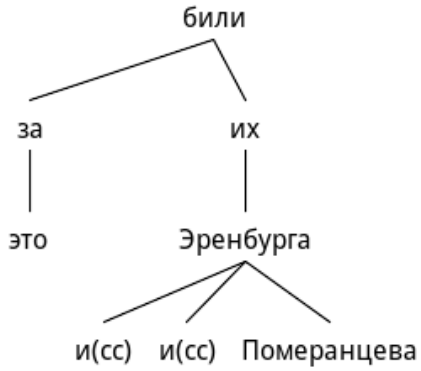

Figure 15. An example of the transformation of соотнос

4. Relation аналит is converted into “aux” or “auxpass”. If the token has “V НЕСОВ СТРАД ИНФ” or “V НЕСОВ ИНФ”, the head of the token is switched to the parent of the head token, the head token and its dependents are shifted to the dependents of the token. The relation label of the head token is assigned to the token instead of its original label. If the head token has “СТРАД” in feats, the relation is converted into “auxpass”, if not, the correct label is “aux”.

If the token has “PART” in feats and the lemma is “БЫ”, the relation is converted into “auxpass”.

If the token has “PART” in feats and the lemma is “БЫЛО”, the relation is converted into “aux”.

Example: *он будет располагаться* “he will be located” (Figure 16, Figure 17)

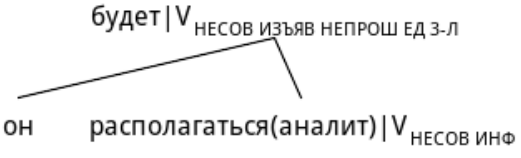

Figure 16. An example of аналит (original structure)

Figure 17. An example of the transformation of аналит

Example: *Это могло бы случиться.* "It could happen." (Figure 18, Figure 19)

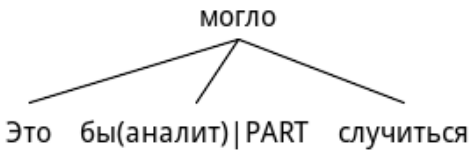

Figure 18. An example of аналит (original structure)

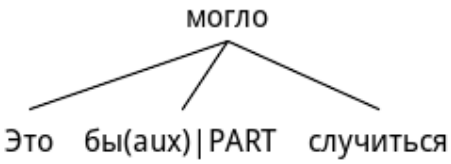

Figure 19. An example of the conversion of аналит without any transformation

5. Relation пасс-анал is converted into "auxpass". The head of the token is switched to the parent of the head token, the head token and its dependents are shifted to the token. The head token relation label is assigned to the token instead of its original label.

Example: *Женщина должна быть подготовлена к тому, что...* "A woman should be prepared for the fact that..."

Figure 20. An example of пасс-анал (original structure)

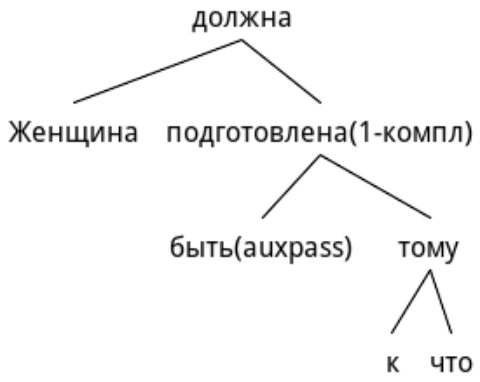

Figure 21. An example of the transformation of пасс-анал

6. Relation присвяз is converted into “cop”. If the lemma of the head token is “БЫТЬ”, the head of the token is switched to the parent of the head token, the head token and its dependents are shifted to the token. Otherwise, the dependency relation is assigned according to the POS tag of the token without any structure transformations.

Example: *Под Москвой есть прекрасное Озернинское водохранилище.* “Near Moscow there is the wonderful Ozerninskoe reservoir.”

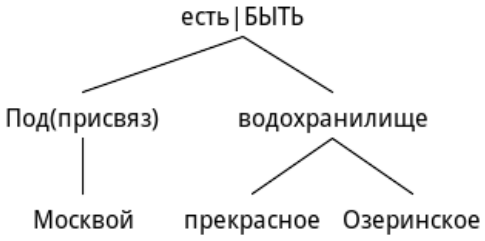

Figure 22. An example of присвяз (original structure)

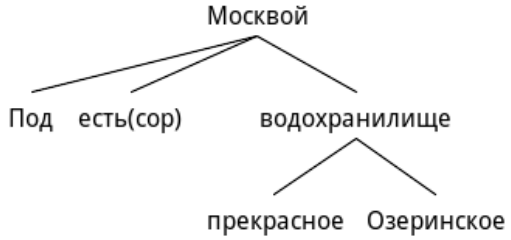

Figure 23. An example of the transformation of присвяз

Example: *Экологическая программа должна стать основой Государственной программы охраны природы.* “The Environmental Program must become the basis of the State Program of Nature Conservation.”

Figure 24. An example of присвяз (original structure)

Figure 25. An example of the conversion of присвяз without any transformation

7. Relations 1-компл, 2-компл, 3-компл, 4-компл and 5-компл are converted according to these rules:

Example: *[образовался] человек переменивший свой пол с мужского на женский* “a man who changed his gender from male to female”

Figure 26. An example of 1-компл, 2-компл and 3-компл (original structure)

Figure 27. An example of 1-компл, 2-компл фтв 3-компл (converted structure)

Example: *Он обязан был выйти на сцену* “he had to go on stage”

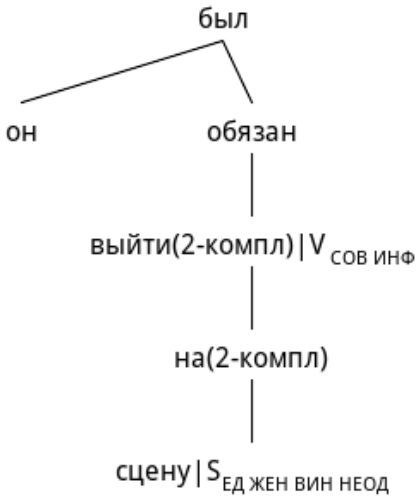

Figure 28. An example of 2-компл (original structure)

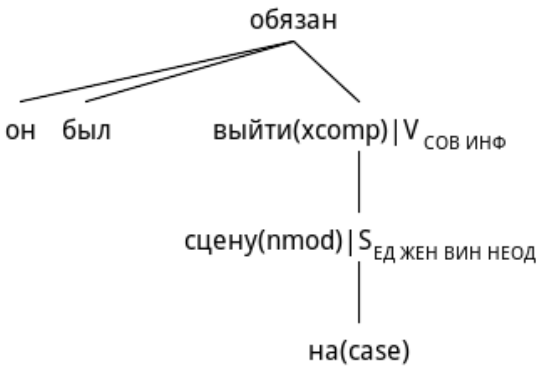

Figure 29. An example of 2-компл (converted structure)

Example: *кто-то провел по векам тончайшей кисточкой* "someone outlined the eyelids with the finest brush"

Figure 30. An example of 3-компл and 4-компл (original structure)

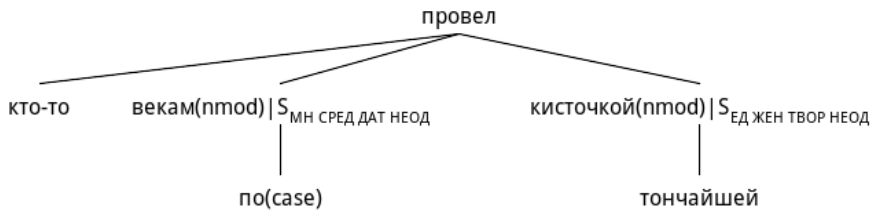

Figure 31. An example of 3-компл and 4-компл (converted structure)

If the token has "CONJ" in feats and it has a verb in dependents, the token and its dependents are shifted to that dependent which is a verb. The token relation is converted into "mark". The head token becomes the head of the verb. The token dependency label is assigned to the verb instead of its original label.

If the token has "CONJ" in feats and it does not have a verb as the dependent, the rule tries to detect a noun, than an adjective, than an adverb, than a numeral and to shift the token and its dependents to that "detected" dependent. The token relation is converted into "mark". The head token becomes the head of the "detected" dependent. The token dependency label is assigned to that dependent instead of its original label.

The additional rule is applied here: if the token has a noun as the dependent and "2-компл" as the relation, the correct label for this dependent is "iobj".

Step 2 assigns correct dependency labels according to these rules:

If the token has "V" in feats and does not have "ПРИЧ" in feats and one of its dependents is marked as "nsubjpass" or "nsubj", the correct label is "ccomp", otherwise, it is "xcomp".

If the token has "S" and "ПР" in feats, the correct label is "nmod". If the token has "S" and "ВИН" in feats and does not have "2-компл" as the relation and a preposition in the dependent position, the correct label is "dobj". If the token has "S" and "РОД" in feats and "2-компл" as the relation, the correct label is "iobj". Otherwise "nmod" is assigned.

If the token has "A" and "ПР" in feats, the correct label is "nmod". Otherwise, "amod" is assigned. If the token has "ADV" in feats, the correct label is "advmod". The correct label for conjunctions is "mark".

Here is an example of the conversion process of the dependency relation "1-компл":

Figure 32 shows the original sentence structure.

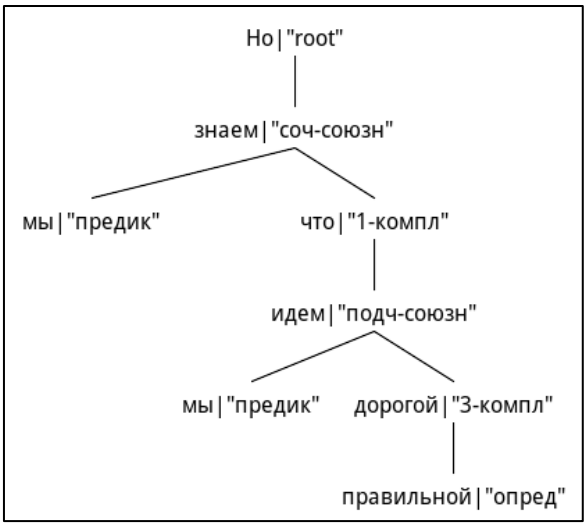

Figure 32. Original structure

Figure 33 shows the transformed structure after the first (preparation) stage.

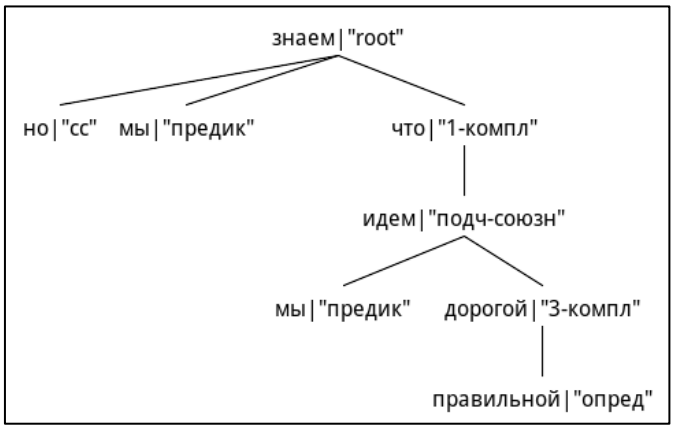

Figure 33. "Prepared" structure

Figure 34 shows the transformed structure after conversion of syntactic relations, step1.

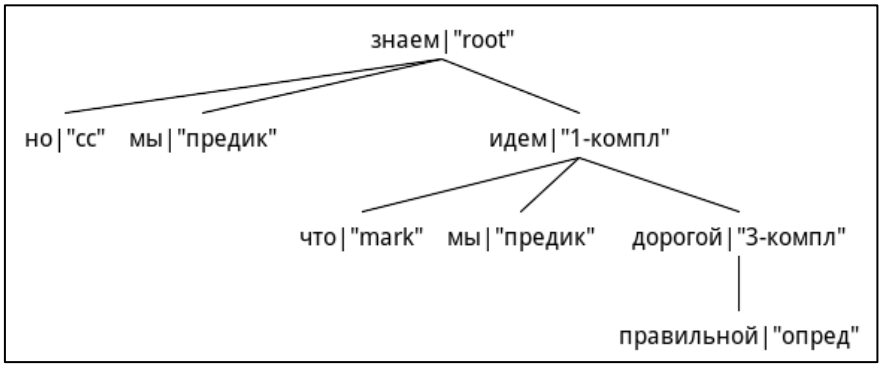

Figure 34. Transformed structure. Step 1

Figure 35 shows the transformed structure after conversion of syntactic relations step2.

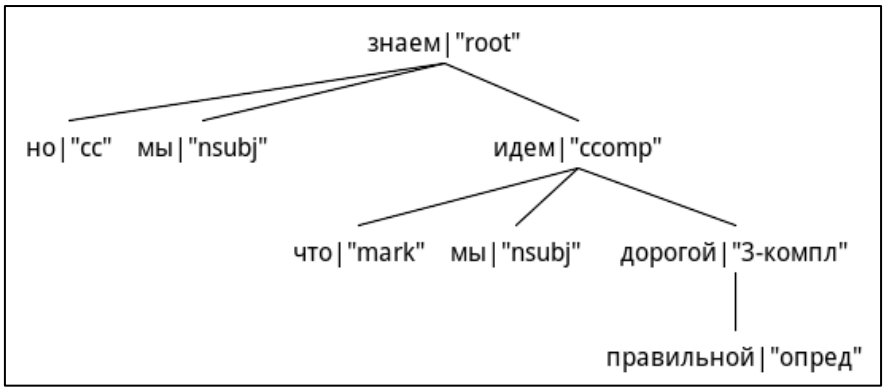

Figure 35. Transformed structure. Step 2

8. Relation эксплет is converted into “acl”. If the token has dependents and “CONJ” in feats, the token and its dependents are shifted to the “closest” dependent. The “closest” is the dependent which does not have “CONJ” or “PART” in feats and the ID of this dependent has minimal positive difference with the ID of the token (typically, the next word in the sentence, but not a particle or a conjunction). The minimal positive difference runs over all the dependents.

Example: *Дело в том, что совершался переход от кратковременных к длительным полетам.* “The fact is that the transition has been completed from short to long flights.”

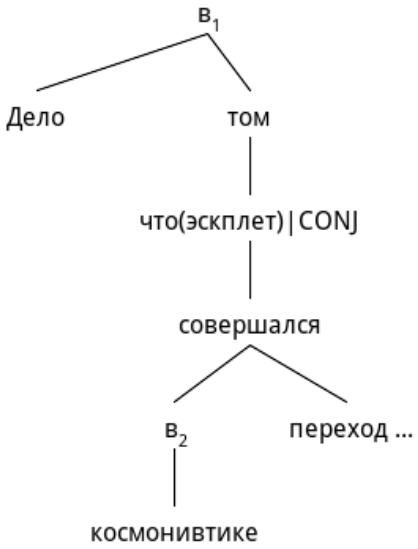

Figure 36. An example of эксплет (original structure)

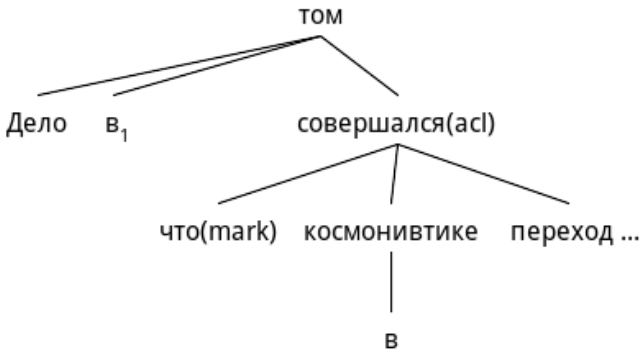

Figure 37. An example of эксплет (converted structure)

The token relation converted into “cc”. The “acl” relation is assigned to the “closest” dependent instead of its original label.

9. Relation обст is converted into various tags according to the rules: If the token has “CONJ” in feats, “mark” is assigned to the token as the relation. The transformation is required. If the token has only one dependent, that dependent becomes the head of the phrase, the “advcl” relation is assigned to the dependent. If there is more than one dependent, the rule tries to detect a verb, then a noun, then an adjective, then an adverb, then a numeral and move it to the head of the phrase. Other dependents of the token and the token itself are shifted to that detected dependent. The “advcl” relation is assigned to that dependent.

Otherwise, the relation is assigned according to the POS tag: “nmod” for nouns, adjectives and numerals, “advmod” for adverbs and particles, “mark” for conjunctions, “advcl” for adverbial participles and for infinitives, but only if the infinitive does not have a verb as the head, in this case “xcomp” is assigned, and “acl” is assigned to other verb forms.

Example: *Это случилось вечером в третьем акте.* “It happened in the evening in the third act.”

Figure 38. An example of обст (original structure)

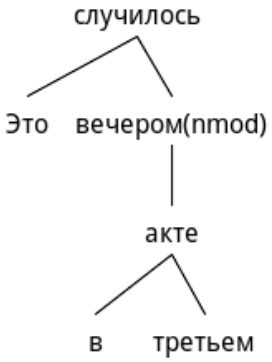

Figure 39. An example of обст (converted structure)

10. Relation сравнит is converted into “nmod”, “advmod” or “advcl”. If the token has “CONJ” in feats, “cc” is assigned as the relation. If the token has only one dependent, the correct label for that dependent is “advmod”. If there is more than one dependent, the rule tries to detect a verb, then a noun, then an adjective, then an adverb, then a numeral and move it to the head of the phrase. The detected dependent loses the link to the token. Other dependents of the token and the token itself are shifted to that detected dependent. The correct dependency label is “advmod”.

Example: *таких изданий как ваш журнал* “publications such as your magazine” …

Figure 40. An example of сравнит (original structure)

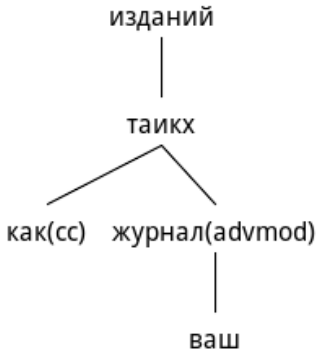

Figure 41. An example of сравнит (converted structure)

Otherwise, the relation is assigned according to the part of speech tag: “nmod” for nouns and adjectives, “advmod” for adverbs and “advcl” for verbs.

11. Relation уточн is converted into “nmod”, “advmod” or “cc”. If the token has “CONJ” in feats, the token is shifted to the first dependent[4]. The relation is assigned according to the POS tag: “nmod” for nouns, “advmod” for adverbs and “cc” for conjunctions.

Example: *За калиткой на улице узловатый дуб.* “There is a knotted oak outside the gate in the street.”

[4] The first dependent is not just a random choice. The statistics shows that usually, the token has only one dependent, but if there is more than one dependent, the first is the correct choice.

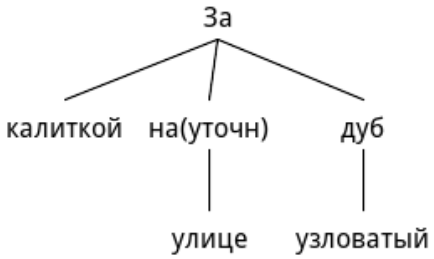

Figure 42. An example of уточн (original structure)

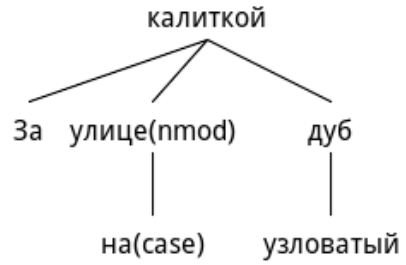

Figure 43. An example of уточн (converted structure)

12. Relation пролепт is converted into “appos” or “cop”. If the head token has one of the lemmas “ЭТО” or “ВОТ”, the head token is analyzed as a copula and the пролепт token is promoted as its sibling, inheriting its original dependency label.

Example: *Большая часть иммигрантов — это турецкие курды.* “Most of the immigrants are Turkish Kurds.”

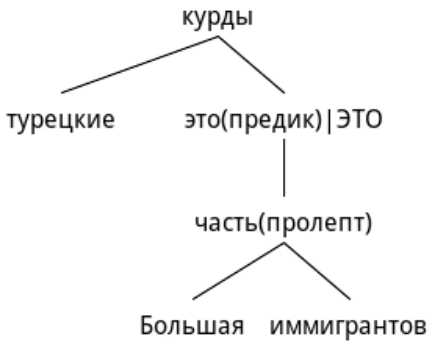

Figure 44. An example of пролепт (original structure)

Figure 45. An example of пролепт (converted structure)

Otherwise, the correct token is “appos”.

13. Relation вспом is converted into “name” or “mwe”. If the token has “NUM” in feats and the word form of the head token starts with a capital letter, the correct label is “name”. If the word form of the head token starts with a capital letter and the token goes after the head token in the tree structure (the token has the greater value of the ID attribute than the head token), the correct label is “name”. If the token goes before the head token in the tree structure, the transformation is required: the head token is shifted to the token and the token is moved to the head of the phrase.

Example: *мы представим себе, что...* “we imagine that...”

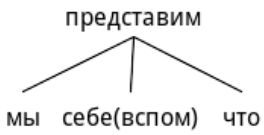

Figure 46. An example of вспом (original structure)

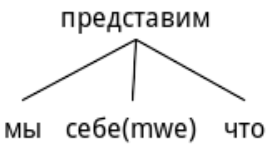

Figure 47. An example of вспом (converted structure)

Example: *Ф. Габер пытался получать золото* “F Gaber was trying to obtain gold”

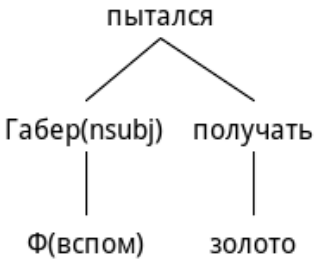

Figure 48. An example of вспом (original structure)

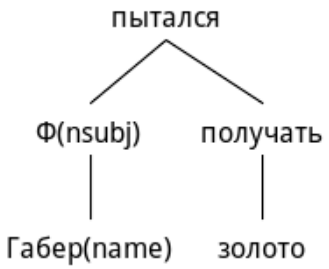

Figure 49. An example of вспом (converted structure)

Otherwise, the correct token is “mwe”.

вспом(а, то) … multi-word coordinating conjunction
вспом(как, бы) … multi-word subordinating conjunction

14. Relation сравн-союзн is converted into “nmod”, “advmod” or “advcl” according to the POS tag. If the token is a noun or an adjective, the correct label is “nmod”. If the token is an adverb – “advmod”.

Example: *намного шире и интереснее чем в прежние годы* “much broader and more interesting than in previous years”

Figure 50. An example of сравн-союзн (original structure)

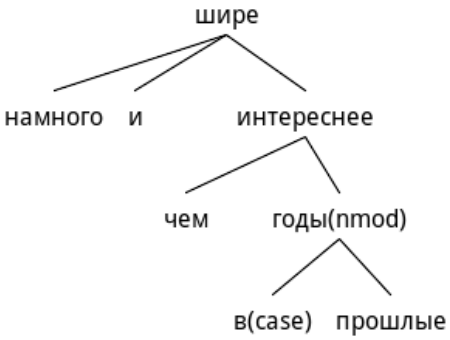

Figure 51. An example of сравн-союзн (converted structure)

If the token is a verb and the head token has “parataxis” as the relation, the head of the token is switched to the parent of the head token, the head token is shifted to this token as the dependent. “Mark” is assigned to the (ex) head token and “parataxis” is assigned to the token. If the token does not have “parataxis” as the relation, the same transformation is required, but the assigned labels are “mark” and “advcl”, correspondingly.

15. Relation предл is converted into “case” or “mark” according to the POS tag. Firstly, the preposition and its dependents are shifted to the dependent that has “предл” as the relation. If the new head is a verb, the correct dependency label for preposition is “mark”. Otherwise, “case” is assigned. The original dependency label of the preposition is assigned to the “new head” instead of “предл”.

Example: *Это случилось вечером в третьем акте.* “It happened in the evening in the third act.”

Figure 52. An example of сочин (original structure)

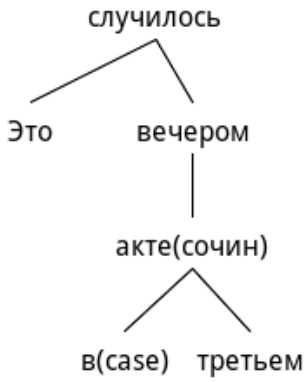

Figure 53. An example of сочин (converted structure)

16. The rule that produces “nummod:gov” employs lemmas “МИЛЛИАРД”, “МИЛЛИОН”, “ТРИЛЛИОН”, “ТЫСЯЧА” and “БИЛЛИОН”.

The original sentence structure treats these lemmas as the local head: “Тут живут 20 миллионов человек”/“20 million people live here”(Figure 54)

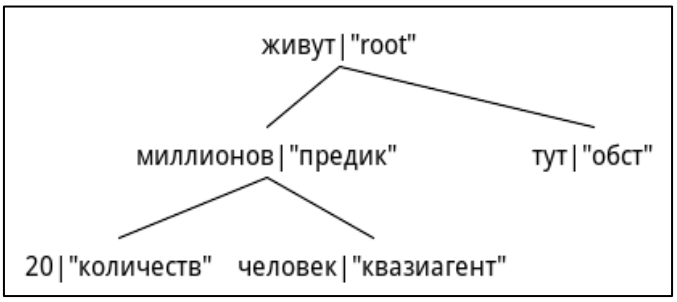

Figure 54. An example of “nummod:gov” transformation. Step 1

If the token has only one dependent and it is a numeral, the correct label for that dependent is “compound” and “nummod:gov” is assigned to the token.

If the token has a numeral and a noun in the genitive case among its dependents, the sentence structure is transformed (Figure 55): the token is shifted to the noun dependent, the noun dependent becomes the new head and it collects the dependency relation of a token. The correct label for the token is “nummod:gov” and “compound” is assigned to the numeral dependent.

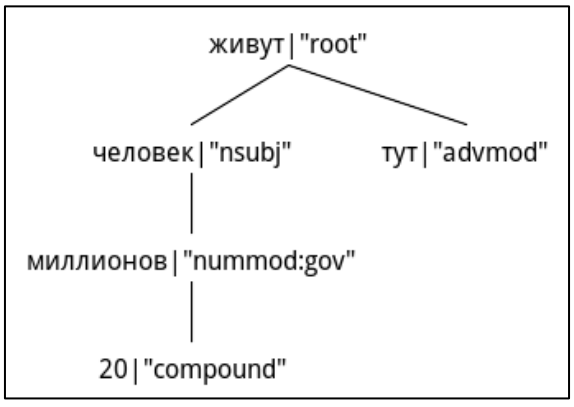

Figure 55. An example of "nummod:gov" transformation. Step 2

If the token has a noun in the genitive case, but does not have a numeral among its dependents, the sentence structure is transformed in a similar way (Figure 55), but "nummod:gov" relation is not assigned.

#	Full Syntagrus name	Syntagrus label	UD label
1	Сочинительное	сочин	cc/conj
2	Сентенциально-сочинительное	сент-соч	cc/conj
3	Сочинительно-союзное	соч-союзн	conj/cc
4	Подчинительно-союзное	подч-союзн	advcl
5	Инфинитивно-союзное	инф-союзн	advcl
6	Соотносительное	соотнос	cc/nmod/advmod/nummod
7	Аналитическое	аналит	aux/auxpass
8	Пассивно-аналитическое	пасс-анал	auxpass
9	Присвязочное	присвяз	cop
10	1-е комплетивное	1-компл	dobj/iobj/ccomp/xcomp
11	2-е комплетивное	2-компл	dobj/iobj/ccomp/xcomp
12	3-е комплетивное	3-компл	dobj/iobj/ccomp/xcomp
13	4-е комплетивное	4-компл	dobj/iobj/ccomp/xcomp
14	5-е комплетивное	5-компл	dobj/iobj/ccomp/xcomp
15	Эксплетивное	эксплет	acl
16	Обстоятельственное	обст	advmod/nmod/advcl/acl/xcomp
17	Сравнительное	сравнит	cc/nmod/advmod/advcl
18	Уточнительное	уточн	nmod/advmod/cc
19	Пролептическое	пролепт	appos/cop (это, вот)
20	Вспомогательное	вспом	name/mwe
21	Сравнительно-союзное	сравн-союзн	nmod/advcl/advmod
22	Предложное	предл	case/mark
23			nummod:gov

Table 5. Complex rules

Conclusions

We have presented the UD corpus for Russian and provided the detailed description on the conversion process. The corpus has been released as a part of the fourth UD release.

However, we still have features which should be implemented in the next release.

In the SynTagRus corpus, certain multiword expressions are represented as single tokens. The UD-style corpus contains these tokens as well, but spaces have been replaced with '_' ("что_ни_на_есть", "хотя_бы"). This is a temporary solution. Such tokens will be replaced with series of separate tokens connected with the mwe relation.

COM and NID part of speech tags need to be analyzed more precisely. The conjunction rules and rules for dobj and iobj relations require modification.

The proper noun detection rules need to be implemented as well.

References

[1] Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In Proc. of the Tenth Conference on Computational Natural Language Learning, New York, USA.

[2] J. Nivre, J. Hall, S. Kubler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In Proc. of the CoNLL 2007 Shared Task. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLPCoNLL).

[3] Slav Petrov, Dipanjan Das, Ryan McDonald. 2011. A Universal Part-of-Speech Tagset. ArXiv:1104.2086

[4] Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, Proceedings of the Sixth International Language Resources and Evaluation Conference, LREC 2008, pages 28–30, Marrakech, Morocco, May. European Language Resources Association (ELRA).

[5] Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2014. HamleDT: Harmonized multi-language dependency treebank. In Language Resources and Evaluation, DOI 10.1007/s10579-014-9275-2. (Extended version of paper from LREC 2012.)

[6] Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In Proceedings of LREC.

[7] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, ISBN 978-2-9517408-9-1, pp. 1659-1666

[8] Universal Dependencies URL:
http://universaldependencies.org/

[9] Russian National Corpus URL:
http://www.ruscorpora.ru/instruction-syntax.html

[10] Igor Boguslavsky, Leonid Iomdin, Tatyana Frolova, Svetlana Timoshenko. Development of a Russian Tagged Corpus with Lexical and Functional Annotation. In Proceedings of the MONDILEX Third Open Workshop. Bratislava, Slovakia, 15–16 April, 2009

[11] Дяченко П.В., Иомдин Л.Л., Лазурский А.В., Митюшин Л.Г., Подлесская О.Ю., Сизов В.Г., Фролова Т.И., Цинман Л.Л. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Сборник "Национальный корпус русского языка: 10 лет проекту". Труды Института русского языка им. В.В. Виноградова. М., 2015. Вып. 6. С. 272-299.

[12] Leonid Iomdin, Victor Sizov. Structure Editor: a Powerful Environment for Tagged Corpora. MONDILEX Fifth Open Workshop, Ljubljana, Slovenia, October 14–15, 2009. Ljubljana, 2009. pp 1-12. ISBN 978-961-264-012-5

[13] CoNLL-U Format URL:
http://universaldependencies.org/format.html

Appendix A. Dependency relations mapping

Full Syntagrus name	Syntagrus label	UD label
Предикативное	предик	nsubjpass/nsubj
Дательно-субъектное	дат-субъект	iobj/nmod
Агентивное	агент	nmod:agent
Квазиагентивное	квазиагент	nmod
Несобственно-агентивное	несобст-агент	nmod
1-е комплетивное	1-компл	dobj/iobj/ccomp/xcomp
2-е комплетивное	2-компл	dobj/iobj/ccomp/xcomp
3-е комплетивное	3-компл	dobj/iobj/ccomp/xcomp
4-е комплетивное	4-компл	dobj/iobj/ccomp/xcomp
5-е комплетивное	5-компл	dobj/iobj/ccomp/xcomp
Присвязочное	присвяз	cop
1-е несобственно-комплетивное	1-несобст-компл	nmod/ccomp/xcomp
2-е несобственно-комплетивное	2-несобст-компл	nmod/ccomp/xcomp
3-е несобственно-комплетивное	3-несобст-компл	nmod/ccomp/xcomp
Неактантно-комплетивное	неакт-компл	iobj/nmod
Комплетивно-аппозитивное	компл-аппоз	amod/nummod/nmod/acl
Предложное	предл	case/mark
Подчинительно-союзное	подч-союзн	advcl
Инфинитивно-союзное	инф-союзн	advcl
Сравнительное	сравнит	cc/nmod/advmod/advcl
Сравнительно-союзное	сравн-союзн	nmod/advcl/advmod
Элективное	электив	nmod
Сентенциально-предикативное	сент-предик	expl
Адресатно-присвязочное	адр-присв	nmod
Определительное	опред	amod
Описательно-определительное	оп-опред	acl
Релятивное	релят	acl:relcl
Атрибутивное	атриб	nmod
Композитное	композ	compound
Аппозитивное	аппоз	appos/name
Обособленно-аппозитивное	об-аппоз	appos
Номинативно-аппозитивное	ном-аппоз	appos/nmod
Нумеративно-аппозитивное	нум-аппоз	nummod:appos
Количественное	количест	nummod
Аппроксимативно-количественное	аппрокс-колич	nummod
Количественно-копредикативное	колич-копред	nummod/nmod/advmod
Количественно-ограничительное	колич-огран	advmod/nmod
Распределительное	распред	nmod
Аддитивное	аддит	nmod

Full Syntagrus name	Syntagrus label	UD label
Обстоятельственное	обст	advmod/nmod/advcl/acl/xcomp
Длительное	длительн	nmod/advcl
Кратно-длительное	кратно-длительн	nmod
Дистанционное	дистанц	nmod
Обстоятельственно-тавтологическое	обст-тавт	nmod
Субъектно-обстоятельственное	суб-обст	nmod
Объектно-обстоятельственное	об-обст	nmod
Субъектно-копредикативное	суб-копр	acl
Объектно-копредикативное	об-копр	acl
Ограничительное	огранич	advmod
Вводное	вводн	parataxis
Изъяснительное	изъясн	parataxis
Разъяснительное	разъяснит	parataxis
Примыкательное	примыкат	appos/parataxis
Уточнительное	уточн	nmod/advmod/cc
Сочинительное	сочин	cc/conj
Сентенциально-сочинительное	сент-соч	cc/conj
Сочинительно-союзное	соч-союзн	conj/cc
Коммуникативно-сочинительное	ком-сочин	no data with elliptical constructions in release 1.3
Кратное	кратн	amod/nummod/nmod/advmod/xcomp
Аналитическое	аналит	aux/auxpass
Пассивно-аналитическое	пасс-анал	auxpass
Вспомогательное	вспом	name/mwe
Количественно-вспомогательное	колич-вспом	compound
Соотносительное	соотнос	cc/nmod/advmod/nummod
Эксплетивное	эксплет	acl
Пролептическое	пролепт	appos/cop (это,вот)

Appendix B. Dependency relations in the UD corpus of Russian

Dependency relations, which are represented in the UD corpus of Russian, are listed below:

acl	adnominal clause
acl:relcl	relative clause
advcl	adverbial clause
advmod	adverbial modifier
amod	adjectival modifier
appos	apposition
aux	auxiliary verb (not passive)
auxpass	auxiliary verb in periphrastic passive
case	preposition
cc	coordinating conjunction
ccomp	complement clause
compound	compound numeral
conj	non-head conjunct
cop	copula
dep	other or unknown dependency
det	determiner
dobj	direct object
expl	expletive
iobj	indirect object
mark	subordinating conjunction
mwe	non-head part of a frozen multi-word expression
name	non-head part of a personal name
neg	negative particle “не”
nmod	nominal modifier
nmod:agent	agent in passive constructions
nsubj	nominal subject of active clause
nsubjpass	nominal subject of passive clause
nummod	numeric modifier
nummod:appos	numeric modifier governed by a noun (“room 30”)
nummod:gov	numeric modifier governing the case of the counted n.
parataxis	loosely attached phrase
xcomp	controlled clausal complement

The ÚFAL/CKL Technical Report Series

ÚFAL

ÚFAL (Ústav formální a aplikované lingvistiky; http://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

CKL

As of 1 June 2000 the Center for Computational Linguistics (Centrum komputační lingvistiky; http://ckl.mff.cuni.cz) was established as one of the centers of excellence within the governmental program for support of research in the Czech Republic. The center is attached to the Faculty of Mathematics and Physics of Charles University in Prague.

Technical Reports

The ÚFAL/CKL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. The technical reports published in this Series are results of the research carried out in the research projects supported by the Grant Agency of the Czech Republic, GAČR 405/96/K214 ("Komplexní program"), GAČR 405/96/0198 (Treebank project), grant of the Ministry of Education of the Czech Republic VS 96151, and project of the Ministry of Education of the Czech Republic LN00A063 (Center for Computational Linguistics). Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*

ÚFAL/CKL TR-2006-34 Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová (in cooperation with Karolína Skwarska, Klára Hrstková, Michaela Nová, Eduard Bejček, Miroslav Tichý) *Valency Lexicon of Czech Verbs. VALLEX 2.0*

ÚFAL/CKL TR-2006-35 Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský, *Annotation of English on the tectogrammatical level*

ÚFAL/CKL TR-2007-36 Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza, *Zpracování pojmenovaných entit v českých textech*

ÚFAL/CKL TR-2008-37 Silvie Cinková, Marie Mikulová, *Spontaneous speech reconstruction for the syntactic and semantic analysis of the NAP corpus*

ÚFAL/CKL TR-2008-38 Marie Mikulová, *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory*

ÚFAL/CKL TR-2008-39 Zdeněk Žabokrtský, Ondřej Bojar, *TectoMT, Developer's Guide*

ÚFAL/CKL TR-2008-40 Lucie Mladová, *Diskurzní vztahy v češtině a jejich zachycení v Pražském závislostním korpusu 2.0*

ÚFAL/CKL TR-2009-41 Marie Mikulová, *Pokyny k překladu určené překladatelům, revizorům a korektorům textů z Wall Street Journal pro projekt PCEDT*

ÚFAL/CKL TR-2011-42 Loganathan Ramasamy, Zdeněk Žabokrtský, *Tamil Dependency Treebank (TamilTB) – 0.1 Annotation Manual*

ÚFAL/CKL TR-2011-43 Ngụy Giang Linh, Michal Novák, Anna Nedoluzhko, *Coreference Resolution in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-44 Anna Nedoluzhko, Jiří Mírovský, *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-45 David Mareček, Zdeněk Žabokrtský, *Unsupervised Dependency Parsing*

ÚFAL/CKL TR-2011-46 Martin Majliš, Zdeněk Žabokrtský, *W2C – Large Multilingual Corpus*

ÚFAL TR-2012-47 Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, Eva Hajičová, *Manual for annotation of discourse relations in the Prague Dependency Treebank*

ÚFAL TR-2012-48 Nathan Green, Zdeněk Žabokrtský, *Ensemble Parsing and its Effect on Machine Translation*

ÚFAL TR-2013-49 David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Daniel Zemana, Zdeněk Žabokrtský, Jan Hajič *Cross-language Study on Influence of Coordination Style on Dependency Parsing Performance*

ÚFAL TR-2013-50 Jan Berka, Ondřej Bojar, Mark Fishel, Maja Popović, Daniel Zeman, *Tools for Machine Translation Quality Inspection*

ÚFAL TR-2013-51 Marie Mikulová, *Anotace na tektogramatické rovině. Dodatky k anotátorské příručce (s ohledem na anotování PDTSC a PCEDT)*

ÚFAL TR-2013-52 Marie Mikulová, *Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT)*

ÚFAL TR-2013-53 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *Úpravy a doplňky Pražského závislostního korpusu (Od PDT 2.0 k PDT 3.0)*

ÚFAL TR-2013-54 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *From PDT 2.0 to PDT 3.0 (Modifications and Complements)*

ÚFAL TR-2014-55 Rudolf Rosa, *Depfix Manual*

ÚFAL TR-2014-56 Veronika Kolářová, *Valence vybraných typů deverbativních substantiv ve valenčním slovníku PDT-Vallex*

ÚFAL TR-2014-57 Anna Nedoluzhko, Eva Fučíková, Jiří Mírovský, Jiří Pergler, Lenka Šíková, *Annotation of coreference in Prague Czech-English Dependency Treebank*

ÚFAL TR-2015-58 Zdeňka Urešová, Eva Fučíková, Jana Šindlerová, *CzEngVallex: Mapping Valency between Languages*

ÚFAL TR-2015-59 Kateřina Rysová, Magdaléna Rysová, Eva Hajičová, *Topic–Focus Articulation in English Texts on the Basis of Functional Generative Description*

ÚFAL TR-2016-60 Kira Droganova, Daniel Zeman, *Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies*