NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Olga Lyashevskaya, Kira Droganova, Daniel Zeman, Maria Alexeeva, Tatiana Gavrilova, Nina Mustafina, Elena Shakurova

UNIVERSAL DEPENDENCIES FOR RUSSIAN: A NEW SYNTACTIC DEPENDENCIES TAGSET

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS
WP BRP 44/LNG/2016

This Working Paper is an output of a research project implemented at the National Research University Higher School of Economics (HSE). Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.

Olga Lyashevskaya[1], *Kira Droganova*[2], *Daniel Zeman*[3],
Maria Alexeeva[4], *Tatiana Gavrilova*[5], *Nina Mustafina*[6], *Elena Shakurova*[7]

UNIVERSAL DEPENDENCIES FOR RUSSIAN: A NEW SYNTACTIC DEPENDENCIES TAGSET

This paper presents the Universal Dependencies tagset (UD v1) as a new annotation scheme for Russian treebanks. The universal list of dependency relations was adopted and extended to comply with certain language-specific syntactic constructions. The tagset was validated, converting two Russian treebanks into the UD format, UD-Russian-SynTagRus and UD-Russian-Google.

JEL Classification: Z.

Keywords: natural language processing, dependency parsing, universal dependencies, Russian language

[1] National Research University Higher School of Economics, Moscow, Russia; Vinogradov Institute of the Russian Langiage RAS, Moscow, Russia; olesar@yandex.ru

[2] Charles University, Prague, Czech Republic; kira.droganova@gmail.com

[3] Charles University, Prague, Czech Republic; zeman@ufal.mff.cuni.cz

[4] National Research University Higher School of Economics, Moscow, Russia; alexeevamary@rambler.ru

[5] National Research University Higher School of Economics, Moscow, Russia; tanya96gavrilova@yandex.ru

[6] National Research University Higher School of Economics, Moscow, Russia; mus_scor@mail.ru

[7] National Research University Higher School of Economics, Moscow, Russia; lenashakurova@yandex.ru

I. Introduction[*]

The Universal Dependencies (UD) project integrates previous efforts and provides guidelines for cross-linguistically consistent treebank annotation for typologically different languages (Nivre 2015; Nivre et al. 2016). UD has considerable practical value for modern linguistic applications such as machine translation and information retrieval. It can be employed as a universal grammar for natural language processing, facilitating multilingual parser development, cross-language parsing, and the evaluation of parsing results.

The UD layout is based on the Google Universal part-of-speech (POS) tagset (Petrov et al. 2012), the Interset interlingua of morphosyntactic features (Zeman 2008), and Stanford Dependencies (Tsarfaty 2013, de Marneffe et al. 2014). The storage format of the treebanks is CoNLL-U. UD v.1.3 (released in May 2016), which covers 40 languages and includes 52 treebanks.

At present, only one branch of the Russian National Corpus, the rather small 1M word SynTagRus treebank, is annotated with dependencies via a manual correction of the parsing results (Boguslavsky et al. 2009). The 1.3B token ruWaC (Sharoff and Nivre 2011) was parsed automatically by the Malt dependency parser trained on SynTagRus (Nivre 2007). The 14.5B word ruTenTen corpus (Jakubicek et al. 2013) is commercial and only partially (automatically) annotated for a limited list of constructions. RU-EVAL 2012, a Russian parsing task (Gareyshina et al. 2012), produced a number of treebank resources including the small gold standard and 1M word treebank with parallel annotation by four parsers. Thus, it would be no exaggeration to say that Russian still lacks large open resources with full-fledged and high-quality syntactic annotation.

The Russian UD project was launched in 2015 with three objectives: (i) to provide the POS, morphosyntactic and dependencies tagsets and guidelines for Russian, (ii) to provide freely available treebanks including manually tagged gold standards, converted treebanks, and automatically annotated larger corpora, and (iii) to develop Russian UD parsers trained on these data. In this paper, we focus on the dependency label set and annotation scheme developed for Russian.

II. Russian UD corpora

UD defines the sentence split and tokenization rules, lemmatization rules (language-specific), and the universal inventories of POS, morphosyntactic features, and dependency labels which can, in principle, be extended by the language-specific tags. In addition, the layer of POS tags can be divided into two layers: coarse-grained and fine-grained.

UD v.1.3 includes two automatically converted Russian treebanks: UD-Russian-SynTagRus and UD-Russian-Google (http://universaldependencies.org/#ru). UD-Russian-SynTagRus contains news, non-fiction and fiction prose (1M tokens), UD-Russian-Google (UD-Russian) contains wiki texts (0.1M tokens). Both resources were manually checked in their original dependency layout.

The genuine SynTagRus scheme is more compatible with the UD standard in the structure of POS tags (the major distinction is pronouns which are treated as NOUN, ADJ, ADV in SynTagRus while they are treated as either PRON or DET in UD; proper nouns have a distinct tag in UD). In contrast, the number of dependency tags differs greatly: 67 tags in SynTagRus and 40 tags in UD (on mapping from one tagset to another, see Droganova and Zeman 2016). Multi-word expressions are treated as one token in SynTagRus and as separate tokens in UD.

[*] This work was partly supported by grant No.15-07-09306 of the Russian Basic Research Foundation (RBRF). We are grateful to Andrey Kutuzov who suggested the very idea of the Russian UD project and Francis Tyers for his generous help and encouragement. We would like to thank the Google team and especially Ryan McDonald and Vitaly Nikolaev who provided the converted version of the UD-Russian-Google treebank. We also thank the Laboratory of Computational Linguistics of the Kharkevich Institute for Information Transmission Problems RAS (IPPI RAN) in Moscow and especially Leonid Iomdin for providing original SynTagRus data and generous feedback. Finally, we express our gratitude to the UD support team and members of many UD groups working with other languages.

The Google treebank is compatible with UD in syntax rather than in POS/features tags since it is based on the same Stanford scheme. Some dependency relations were simplified, for example, “nmod:gobj” and “nmod:tmod” were converted into “nmod”.

Figure 1 illustrates the CoNLL-U format for the Russian sentence *Мариано закончил Национальную академию* ‘Mariano graduated from the National academy’ as it appears in the current version of the Google UD treebank. The fields include:

1) ID: Indicates the position of the token in a sentence (integer, starting at 1 for each new sentence).
2) FORM: Word form or punctuation mark.
3) LEMMA: The result of lemmatization.
4) UPOSTAG: Universal POS tag (coarse-grained).
5) XPOSTAG: Fine-grained POS tag.
6) FEATS: List of morphological features (formatted as Feature=Value and separated with |).
7) 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.
8) DEPREL: Universal dependency relation to the HEAD (root if HEAD = 0).
9) DEPS: List of secondary dependencies (not used).
10) MISC: Any other additional information (not used).

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Мариано	Мариано	PROPN	NNP	Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing	2	nsubj	_	_
2	закончил	закончить	VERB	VBC	Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past	0	root	_	_
3	Национальную	национальный	ADJ	JJL	Animacy=Inan|Case=Acc|Gender=Fem|Number=Sing	4	amod	_	_
4	академию	академия	NOUN	NN	Animacy=Inan|Case=Acc|Gender=Fem|Number=Sing	2	dobj	_	_
5	.	.	PUNCT	.	_	2	punct	_	_

Fig. 1. CONLL-U storage format.

III. General principles of UD dependencies

The head of the sentence in UD (marked as “root”) is usually a finite verb or other predicate. The punctuation mark is attached to the root, or the head of the clause, or the next/previous word (for brackets and quotation marks). There are three general principles of dependencies annotation:

1) the primacy of content words: the content words are linked directly, so prepositions, conjunctions, auxiliaries and other function words are treated as dependents of a content word; the head of the subordinate clause is also linked directly to the content word in the next higher clause;

2) the centrality principle: in a coordinate group, all coordinated content words, conjunctions, and punctuation marks are linked to the first content word; the same applies to NPs with multiple adjectives (all adjectives are linked to the noun), VPs with multiple adjectives, multi-word expressions, names, etc. The only two exceptions from the principle of centrality are lists and compounds, see below;

Figure 2 gives an example of the UD dependency annotation for Russian.

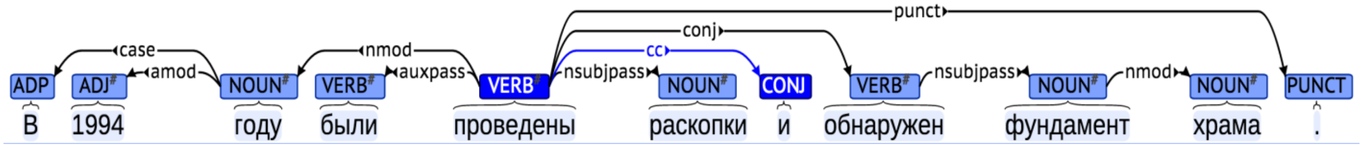


Fig. 2. A sample of the UD annotation (English translation: ‘In 1994, excavations were carried out and the foundations of the cathedral were discovered.’).

The sentence structure is close to a disassortative network. It has very few nodes with degree 2 and a lot of nodes with degree 3 to 6; in comparison with the SynTagRus structure, the average path length from the root node to the content node is almost half as long in the UD structure;

3) promotion by head elision: if the (natural) head of a phrase is omitted, one of its dependents will be "promoted": the new head will be assumed to bear the function of the omitted content word head. This way of analysis aims to disrupt the structure to a lesser degree.

Example (1) illustrates the promotion of the auxiliary verb in the case of ellipsis.

(1) *Если.*SCONJ *ты.*PRON *не*.PART *будешь.*VERB ←aux– *обманывать.*VERB.Inf*,* *я*.PRON ←nsubj– *буду.*VERB_root.
'lit. If you will not cheat, I will'.

IV. UD dependency labels for Russian

The set of UD dependency labels includes the following 45 tags:
root – sentence head
punct – punctuation marks

The nominal dependencies of the verb:
nsubj – subject
nsubjpass – syntactic subject in a passive clause
dobj – direct object (the second argument, usually in the Accusative, but also in the Genitive, Dative, and Instrumental cases)
iobj – indirect object (the third, fourth, etc. arguments of the verb)
nmod – nominal modifier (e.g. a prepositional phrase)
nmod:agent – semantic subject in the Instrumental case in a passive clause

Other modifiers of the verb:
advmod – adverbial modifier
neg – negation
aux – auxiliary verb or grammatical marker attached to the content verb
auxpass – auxiliary verb etc. attached to the content verb in a passive clause

The clausal dependencies of the verb:
ccomp – clausal complement
xcomp – clausal complement (in the Infinitive, short passive participle) with obligatory control of its (omitted) subject (normally coreferent with the subject of the higher clause)
advcl – clausal adverbial (e.g. the head of a gerund phrase or subordinate clause)

The dependencies of the nominal heads:
case – preposition attached to a nominal head (also a conjunction in a comparative or explanatory function)
amod – adjectival modifier
acl – clausal adjectival modifier (e.g. a participle)
acl:relcl – relative clausal modifier of a noun
appos – appositive modifier
det – determiner (pronominal quantifier); applied to demonstrative, possessive, relative, indefinite and universal pronouns which tend to occur in the leftmost periphery of the nominal phrase
nummod – numeric modifier
nummod:gov – numeric modifier governing the case of a noun (including the pronominals *сколько* 'how many, how much' and *столько* 'so many, so much')
nummod:entity – numeric appositive modifier (e.g. *астероид 697* 'asteroid 697')
cop – copula; attached to a nominal predicate

(see above nmod, dobj, iobj, advmod, clausal dependencies, etc. which can also be attached to nominal heads)

Arguments without shallow morphosyntactic effects:
csubj – clausal subject
csubjpass – clausal subject in a passive clause
vocative – vocative, address

Coordination, subordination, parataxis:
conj – links the first conjunct with other items in a coordinate group
cc – coordinating conjunction; attached to the first conjunct
cc:preconj – coordinating conjunction that precedes the first conjunct
mark – subordinate conjunction; usually attached to the head of a subordinated finite clause
parataxis – links the heads of clauses and phrases placed side by side without any explicit coordination, subordination, or argument relation
discourse – discourse element (e.g. particles and parentheticals)
expl – expletive (the expletive marker *это,* e.g. *Это Ваня пришел* 'It is Vanja (who) came')

Compounding:
compound – relation within a compound numeral (the numerals are attached to the rightmost numeral)
mwe – sequentially joins the words in multi-word expressions
goeswith – links the parts of a hyphenated word or two parts of a word that are separated in a text that is not well edited
name – links the first name of a person with the patronymic and the last name
foreign – links the first foreign word with other words in a quoted foreign text incorporated into a Russian sentence
list – links the first item with other items in a list without evident syntactic structure (e.g. the list name, phone, email)

Other joining relations:
remnant – remnants in elliptic constructions
reparandum – links disfluencies overridden in a speech repair
dep – other dependencies (unspecified)
One more universal UD label, **dislocated** (for fronted or postposed elements that do not fulfil the usual core grammatical relations of a sentence), is not included in the Russian tagset.

V. Universal and language-specific dependency tags

Of 45 tag labels listed above, 39 are universal while "acl:relcl", "cc:preconj", "nmod:agent", "nummod:gov", and "nummod:entity" are language-specific. The tags "acl:relcl" and "cc:preconj" are also used in many other treebanks, and "nmod:agent" is used in the Swedish and Romanian treebanks. Compared to Czech (which has the most developed UD standard within Slavic languages), the Russian UD standard lacks such relations as "auxpass:reflex" (since the reflexive marker *ся* is always attached to the verb) and "advmod:emph".

The label "csubjpass" has been suggested as part of the standard; however it is not attested in Russian treebanks v.1.3. The label "reparandum" does not occur in the treebanks either since the data is only written language.

In addition, the following tags are not in use in the UD-Russian-SynTagRus treebank v.1.3: "cc:preconj", "discourse", "foreign", "goeswith", "list", "remnant", and "vocative". The label "nummod:gov" is of limited use in the treebank due to some simplification in conversion. At the same time, the following tags lack in the UD-Russian-Google treebank v.1.3: "compound", "nmod:agent", "nummod:entity", "nummod:gov".

VI. Language-specific Russian constructions

The whole range of syntactic constructions specific for Russian is beyond the scope of this paper. Therefore, in this section, we will outline some of them in order to illustrate how various syntactic phenomena are treated under the UD scheme.

A. Constructions with copula

A copula is the relation between the nominal predicate and the copular verb *быть/бывать* ʻto be’, cf. Fig. 3. Note that the verb *становиться* ʻto become’ is not analysed as “cop”.

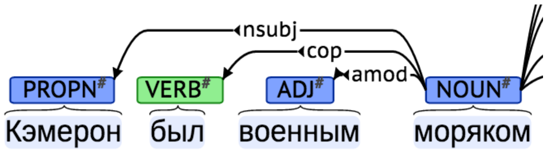


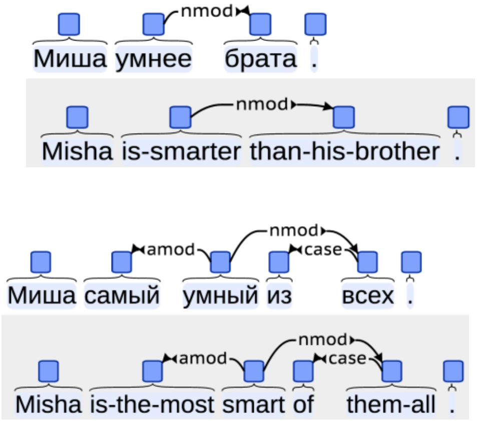
Fig. 3. The “cop” relation (English translation of example: ʻCameron was a sailor on a warship’).

B. Comparative constructions

The most frequently used comparative constructions are the following:

(2a) *Миша*.Nom *умнее брата*.Gen. ʻMisha is smarter than his brother.’ (only with synthetic comparatives)
(2b) *Миша*.Nom *более умный* / *более умен* / *умнее*, *чем брат*.Nom. ʻMisha is smarter than his brother.’ (with both synthetic and analytic comparatives)
(2c) *Миша*.Nom *самый умный* / *умнейший* из.PREP *всех*.Gen. ʻMisha is the smartest of them all.’ (with both synthetic and analytic superlatives)
(2d) *Миша*.Nom *такой же умный* / *так же умен* / *столь же умен*, *как* (*и*) *его брат*.Nom. ʻMisha is as smart as his brother.’ (equality comparison)
The ʻlesser degree’ comparison (expressed periphrastically) is encoded the same way:
(3a) *Миша*.Nom *менее умный* / *менее умен*, *чем его брат*.Nom. ʻMisha is not as smart as his brother.’ (with both types of comparatives)
(3a) *Миша*.Nom *наименее глупый из*.PREP *всех*.Gen. ʻMisha is the least stupid of them all.’ (with both types of superlatives)

Fig. 4 illustrates the annotation of such constructions in UD.


nmod
Миша умнее брата .
nmod
Misha is-smarter than-his-brother .
amod
nmod
case
Миша самый умный из всех .
amod
nmod
case
Misha is-the-most smart of them-all .

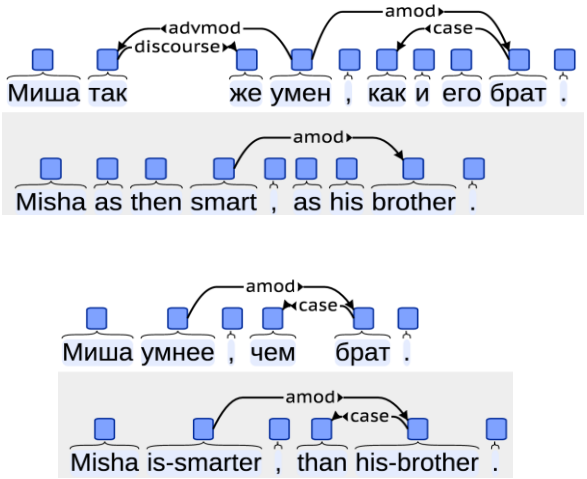

amod
advmod
case
discourse
Миша так же умен , как и его брат .
amod
Misha as then smart , as his brother .
amod
case
Миша умнее , чем брат .
amod
case
Misha is-smarter , than his-brother .

Fig. 4 (a-d). Annotation of Russian comparative constructions in UD. (English translation of examples: (a) ‘Misha is smarter than his brother’; (b) ‘Misha is the smartest of them all’; (c) ‘Misha is just as smart as his brother’; (a) ‘Misha is smarter than his brother’).

C. Noun phrases with quantifiers

In UD the numeral is annotated as dependent of the noun (as “nummod”). The “nummod:gov” label is used to preserve the information that the cardinal numeral governs the noun in certain cases, cf. (4a-b):

(4a) *две*.Nom ←mod:gov– *жены*.Gen ‘two wives’,
(4b) *пять*.Nom ←nummod:gov– *жен*.Gen ‘five wives’.

The pronominal quantifier governing the case of the noun is also labelled “nummod:gov”, cf. (5):

(5) *сколько*.Nom ←nummod:gov– *жен*.Gen ‘how many wives’.

The numeral and pronominal quantifier agreeing in case with the noun is labelled “nummod”, cf. (6):

(5) *со сколькими*.Ins ←nummod– *женами*.Ins ‘with how many wives’.

The following example shows the structure of more complex numerical phrases, cf. Fig. 5.

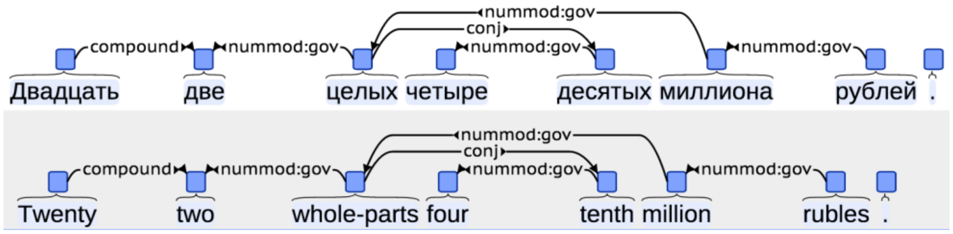


Fig. 5. Annotation of the phrase with complex numeric expression *Двадцать две целых четыре десятых миллиона рублей* ‘22.4 million rubles; lit. Twenty two whole-parts four tenth million rubles’.

A phrase with the postposition of cardinal numerals refers to approximate quantity (usually used with simplex numerals 2-10, 20, 30… etc.). The rules of agreement and case government in such QPs are preserved, cf. Fig. 6.

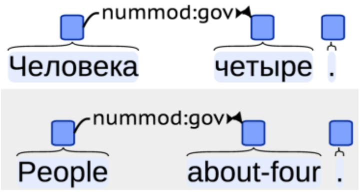
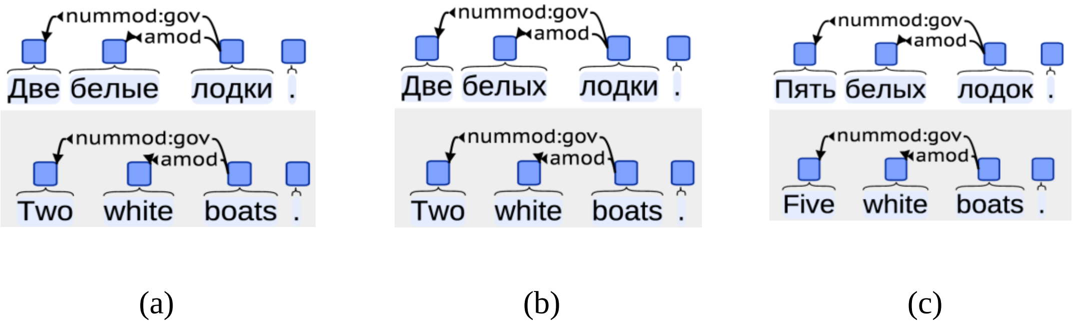


Fig. 6. Postposition of cardinal numerals.

If a phrase with a paucal numeral (*два*, *три*, *четыре*, *оба*, *полтора* ‘two, three, four, both, one and a half’) takes the Nominative or Accusative case, the adjective modifying the noun takes either Nominative (Accusative) plural or Genitive plural, e.g. *две белые лодки* ‘two.NUM.Nom/Acc white.ADJ.Pl.Nom/Acc boats.NOUN.Sg.Gen’ vs. *две белых лодки* ‘two.NUM.Nom/Acc white.ADJ.Pl.Gen boats.NOUN.Sg.Gen’; see. Fig 7 (a-b). With non-paucal numerals (which refer to five objects and more; also *половина*, *четверть* ‘a half, a quarter’ etc.), the adjective is always in Genitive plural, see Fig 7 (c).



(a) (b) (c)

Fig. 7. Annotation of the constructions (a) *Две*.Fem.Nom *белые*.Nom.Pl *лодки*.Fem.Gen.Sg and (b) *Две*.Fem.Nom *белых*.Nom.Pl *лодки*.Fem.Gen.Sg ‘Two white boats’ with the paucal numerals and adjectives; (c) *Пять*.Fem.Nom *белых*.Gen.Pl *лодок*.Fem.Gen.Pl ‘Five white boats’ with the non-paucal numeral and adjective.

The comparative forms *более*, *больше*, *менее*, *меньше* ‘more than, less than’ are used in constructions like (6):

(6) *более двухсот человек* ‘more than 200 people’, (*не*) *меньше пяти машин* ‘(no) less than five cars’.

These comparative forms govern the Genitive case of the cardinal numeral, however they are treated as dependents of the numerals (labeled “advmod”), see Fig. 8.

If QP is a subject, the finite predicate takes either singular (3rd person in present tense, neutral in past tense) or plural depending the information structure and some other factors, cf. (7a-b):

(7a) *Пришло*.Neut.Sg *более двухсот человек* ‘More than 200 people came.Neut.Sg’;

(7b) *Более двухсот человек пришли*.Pl *к памятнику* ‘More than 200 people came.Pl to the monument’.

The distribution of singular and plural is similar but not the same as with cardinal numerals.

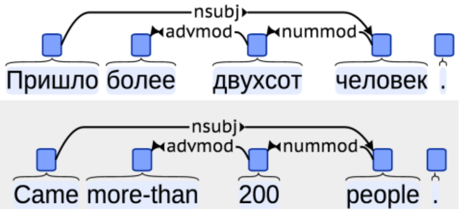


Fig. 8. Annotation of the constructions of more and less quantity.

The collective numerals like *двое*, *трое*, *четверо* ‘a group of two/three/four’, etc. govern the Genitive case of the noun in Nominative and Accusative, cf. (8):

(8) *двое*.NUM.Nom *студентов*.NOUN.Gen.Pl ‘two.Nom students.Nom’, see Fig. 9;

and agree in case with the noun in all other grammatical cases, cf. (9):

(9) *с двумя*.NUM.Ins *студентами*.NOUN.Ins.Pl ‘with two.Ins students.Ins’.

The noun is always in the plural. If this QP is a subject, the finite predicate takes either singular (3rd person in present tense, neutral in past tense) or plural depending the information structure and some other factors (e.g. *Пришло* / *Пришли* ‘Two students came.Neut.Sg / came.pl’). The distribution of singular and plural is similar but not the same as in the case of cardinal numerals and comparative forms.

The choice between cardinal and collective numerals in such constructions depends on animacy, (semantic) gender, semantic class, declination type, and the case of QP (Mel'chuk 1985, Sichinava 2012), collective numerals are usually used with animate masculine nouns or pluralia tantum nouns (e.g. *семеро друзей* ‘a group of seven friends’, *двое саней* ‘two sledges’).

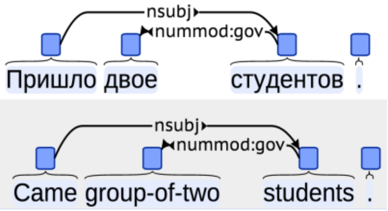


Fig. 9. Annotation of constructions with collective numerals.

D. Sentences with multi-word, multi-part conjunctions

The conjunction preceding the first conjunct clause is labelled as “cc:preconj” and attached to the head of this clause; the second part of the complex conjunction and other conjuncts are attached to the same head (with “cc” and “conj”), see Fig. 10. Only adjacent words in the multi-word conjunctions are treated as multi-word expressions:

(10) *не* ←mwe– *только* ‘not only’; *но* ←mwe– *и* ‘but also’.

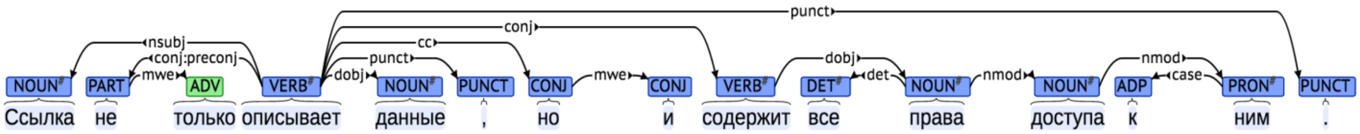


Fig. 10. The annotation of the sentence with the multi-word conjunction *не только... но и...* ‘not only... but also...’ (English translation: ‘The link not only describes the data, but also contains all access rights to them’).

E. Constructions with a clausal subject

The tag “csubj” labels the clausal syntactic subject of a clause, i.e. when the subject is itself a clause. The governor of this relation might not always be a verb: when the verb is a copular verb, the root of the clause is the complement of the copular verb. The dependent is the main content verb or other predicate of the subject clause, see Fig. 11.

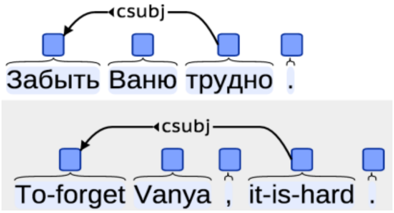


Fig. 11. Annotation of the clausal subject

A clausal passive subject (“csubjpass”) is a clausal syntactic subject of a passive clause, see Fig. 12.

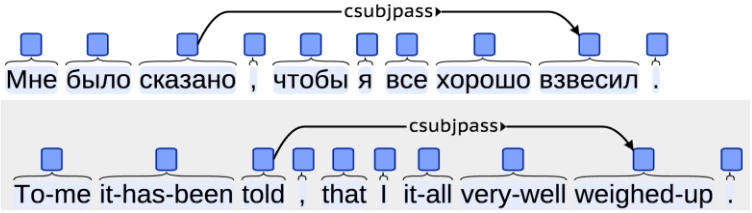

csubjpass
Мне было сказано , чтобы я все хорошо взвесил .
csubjpass
To-me it-has-been told , that I it-all very-well weighed-up .

Fig. 12. Annotation of the passive clausal subject

F. Verb constructions with non canonical case marking

In Russian, the standard case pattern of a predicate-argument construction is as follows: the first argument (subject) is in the Nominative case; the second argument (direct object) is in the Accusative case; all other arguments are coded in other cases or in the prepositional phrase depending on their semantics. However, there is a number of constructions which have non canonical case patterns (Kustova 2011).

As in other Indo-European languages, there are verbs in Russian which have two or more arguments but the second argument is not marked by a case other than Accusative. The most frequent cases include:

* Instrumental pattern: *Старики управляют страной* ‘The old people govern.VERB the country.INS’;

* Genitive pattern: *Он избегал наших встреч* ‘He avoided.VERB our meetings.GEN’;

* Dative pattern: *Он препятствовал отправлению правосудия* ‘He obstructed.VERB the administration.DAT of justice’.

In such cases, the second argument is labelled “dobj” and not “iobj”.

Constructions with a Dative subject are evoked by an infinitive verb (often under negation) or a predicative, the first argument of which is in the Dative case, cf. (11a-b). The Dative argument is labelled “iobj”, which corresponds to the label “дат-субъект” (dative subject) in the original SynTagRus tagset.

(11a) *Как девчонке*.Dat ←iobj– *найти*.Inf –dobj→ *мужа без всего этого ?*

ʹHow-do a-Girl.DAT find.VERB.Inf a husband without all this?’

(11b) *Мне* ←iobj– *стыдно за вас*.

‛I.Dat feel-ashamed.ADV-PREDIC for you.’

Therefore, this structure is parallel to the constructions of experiential verbs, such as in (12):

(12) *Мне* ←iobj– *хочется* –xcomp→ *пить* ʹI.Dat feel-like.DAT to drink’.

Another option would be to set up a new relation, a subtype of “nsubj”, or label the Dative subject with “nsubj”; however, the latter would bias the tendency for “nsubj” to be used predominantly in the Nominative case.

The construction called “Genitive of negation” involves the alternation of an NP’s case between Genitive and Nominative (or Accusative) when the NP is within the scope of the sentential negation. The alternation is sometimes optional and may be affected by certain differences in syntactic structure and/or in semantics or pragmatics (Partee & Borschev 2004). If the subject is under negation (“the Genitive of subject”) and takes Genitive, then the verb becomes impersonal (i.e. takes the 3rd person singular in present tense and neuter singular in past tense), e.g. *Писем*.Gen *не*.NEG *пришло*.VERB.Neut.Sg. ʹNo letters.Gen came’. If the direct object is under negation (“the Genitive of object”), only the case of the direct object NP may change, e.g. *Я не читал их*.Gen *писем*.Gen. ʹI did not read their.Gen letters.Gen’.

H. Constructions with nominal modifiers

The “nmod” relation is used for nominal modifiers. They depend either on another noun or on a predicate:

(13) *карта*.Nom –nmod→ *студента*.Gen ‛the card of student (student card)’

When attached to a noun, it usually corresponds to a non-agreeing attribute in the Genitive case. When it attaches to a verb, adjective or other adverb, “nmod” labels a noun that functions as an oblique argument or adjunct. This means that it functionally corresponds to an adverbial. The head of the prepositional group is always labelled “nmod” [n/a in SynTagRus v.1.3].

The tag “nmod” is also used for temporal nominal modifiers, cf. Fig. 13.

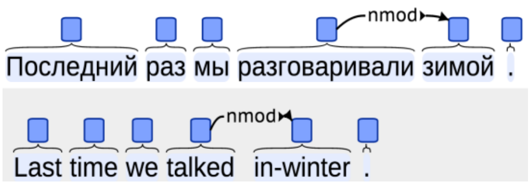

nmod
Последний раз мы разговаривали зимой .
nmod
Last time we talked in-winter .

Fig. 13. Annotation of “nmod” in the Instrumental case.

I. Names and named entities

The common patterns for the Russian personal names are the following: “name + patronymic + last name”, “name + last name”, “name + patronymic”, “last name + name (+ patronymic)”. The name and the patronymic can be expressed by their initial letters. The leftmost name is always the head of the group and the other name(s) are attached to it with the “name” relation.

We distinguish between two relations, “nmod” and “appos” (appositive), which code the attachment of titles and position names. If the title (position name) precedes the name of a person, it is labelled “nmod”, and is labelled “appos” otherwise, see Fig. 14 (a-b).

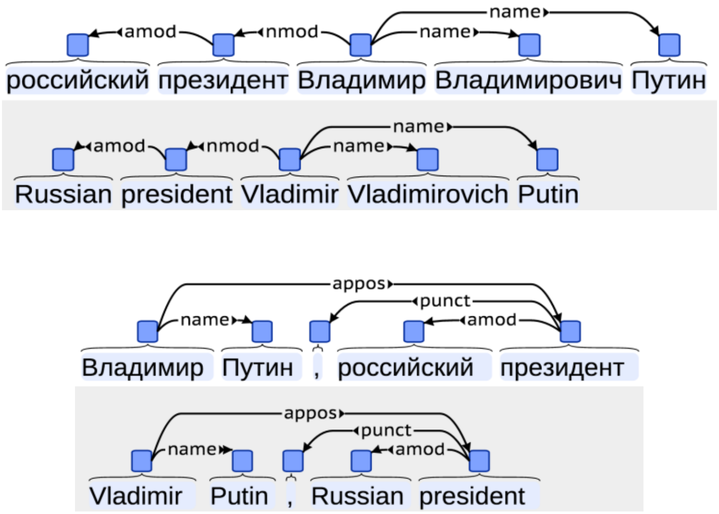

name
amod
nmod
name
российский президент Владимир Владимирович Путин
name
amod
nmod
name
Russian president Vladimir Vladimirovich Putin
appos
punct
name
amod
Владимир Путин , российский президент
appos
punct
name
amod
Vladimir Putin , Russian president

Fig. 14 (a-b). Annotation of personal names and title / position names.

Within multi-word names of places, organizations, etc., the tag “appos” is often used to attach adjacent nouns:

(14a) *банк*.NOM –appos→ *Прогресс*.NOM ‘bank “Progress”’,

(14b) *из банка*.GEN –appos→ *Прогресс*.NOM ‘from bank “Progress”’.

Note that the relation “nmod” is not used to attach adjectives to nouns in the named entities.

VII. UD RELATIONS AT WORK: TWO TREEBANKS

Figure 15 shows the statistics of the UD labels used across the SynTagRus and Google treebanks (see also Appendix). These data have to be considered preliminary until the quality of annotation improves. Nevertheless, this gives an idea about more and less exploited tags.

There are considerably more “nmod” and “appos” relations in the Google treebank than in SynTagRus. By contrast, there are more “nsubj”, “dobj”, “advmod”, “parataxis”, and “advcl” in SynTagRus than in the Google treebank. On the one hand, this disproportion can be explained by differences in genre: the sentences from Wikipedia (UD Google) are often nominative, with a lot of ellipses, appositives, parataxis, as well as participle and gerund groups (usually labelled as “acl” and “advcl”). News and fiction (SynTagRus) presuppose more finite verb clauses and active transitive constructions; the coordinate and subordinate clauses are connected with conjunctions labelled as “cc” and “mark”. Nevertheless, the proportion of passives is almost the same in the two treebanks (as the distribution of the “nsubjpass” suggests), despite genre differences.

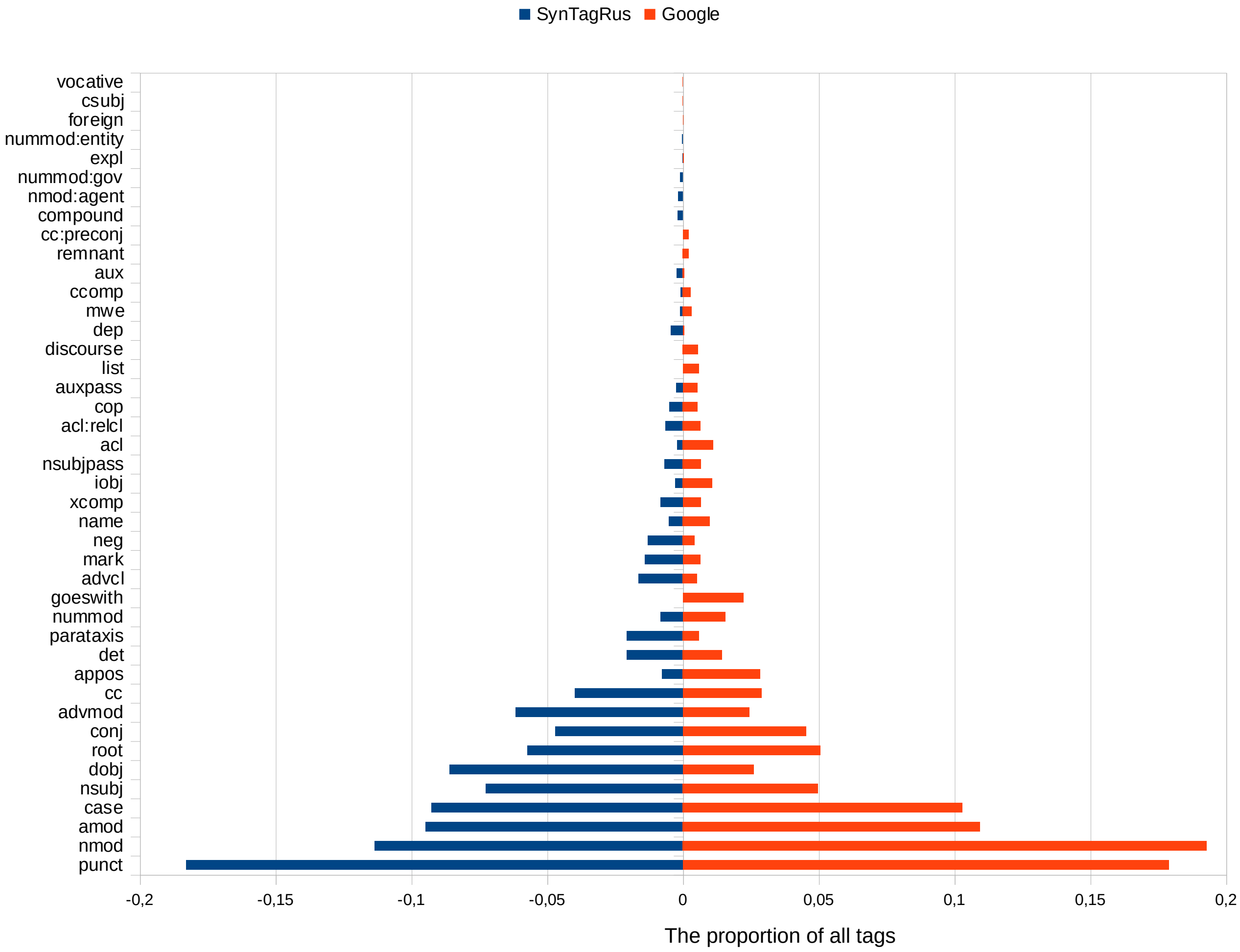


Fig. 15. Comparison of dependency tags occurrences in the UD-Russian-SynTagRus (to the left) and UD-Russian-Google treebanks (to the right).

On the other hand, the statistics show that the annotation schemes vary in details. For example, while the head of a prepositional phrase is always labelled “nmod” in the Google UD (except coordination, etc.), this is not the case in the UD-SynTagRus annotation. The non-canonical second arguments of the verb in SynTagRus can be labelled “iobj” and not “dobj”. The presence/absence of the “goeswith” tag can be explained by different rules of tokenization. There is some difference at the lexical level: for example, in Google UD v.1.3, restrictors like *только*, *лишь* ‘only’, *даже* ‘even’ are linked with the “discourse” relation while in SynTagRus v.1.3, their relations are tagged “advmod”.

VIII. Conclusion

At present, the two UD Russian treebanks differ slightly in the inventory of dependency tags. The conversion from their original annotation to UD usually required not only relabelling the relation types, but also restructuring the tree, which is a potential source of errors. Fortunately, both treebanks are not frozen projects. More work will be done to make them converge in terms of morphosyntax and dependency annotation in the future. In addition, a new Russian UD gold standard, annotated manually, is under development.

So far, many decisions regarding non-core Russian constructions are merely inherited from the original schemes of SynTagRus and Google, being the product of more general conversion rules. Our objective here is to constantly extend the list of such constructions under revision and provide annotation schemes that would be more consistent with the UD annotation of (semi-)parallel structures in other languages.

The resources described in this work and guidelines are available under open licenses from:

https://github.com/UniversalDependencies/UD_Russian-SynTagRus,

https://github.com/UniversalDependencies/UD_Russian,

http:// universaldependencies.org/ru/dep/index.html.

The online access to the Russian treebanks is available using SETS treebank search maintained by the University of Turku:

http://bionlp-www.utu.fi/dep_search/.

PML Tree Query maintained by the Charles University in Prague:

http://lindat.mff.cuni.cz/services/pmltq/.

References

I. Boguslavsky, L. Iomdin, T. Frolova, S. Timoshenko, “Development of a Russian Tagged Corpus with Lexical and Functional Annotation”, in *Proc. MONDILEX Third Open Workshop, Bratislava, Slovakia, Apr. 2009.*

K. Droganova, D. Zeman (2016): Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies (technical report). ÚFAL MFF UK, Praha, Czechia.

A. Gareyshina, M. Ionov, O. Lyashevskaya, D. Privoznov, E. Sokolova, and S. Toldova, “RU-EVAL-2012: Evaluating Dependency Parsers for Russian”, in *Proc. of COLING: Posters*, 2012, pp. 349-360.

M. Jakubicek, A. Kilgarriff, V. Kovar, P. Rychly, and V. Suchomel, “The tenten corpus family”, in *Proc. Corpus Linguistics* 2013.

G. Kustova, “Padezh: Materialy dlja proekta korpusnogo opisanija russkoj grammatiki (http://rusgram.ru) [Case: Materials for the project of corpus-based Russian grammar (http://rusgram.ru)]”, Ms. Moscow, 2011.

M.-C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and Ch. Manning, “Universal Stanford Dependencies: A cross-linguistic typology”, in *Proc.of the 9th International Conference on Language Resources and Evaluation* (*LREC 2014*), Vol. 14, pp. 4585–4592.

I. A. Mel'chuk, *Poverkhnostnyj sintaksis russkikh chislovykh vyrazhenij* [The shallow syntax of Russian numeric expressions]. Wien, 1985.

J. Nivre, “Towards a Universal Grammar for Natural Language Processing”, *Computational Linguistics and Intelligent Text Processing*, 2015, 3–16.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi, “MaltParser: A language independent system for data-driven dependency parsing”, *Natural Language Engineering,* 13 (2007), 95-135.

J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, Ch. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman, “Universal Dependencies v1: A Multilingual Treebank Collection”, in *Proc. of LREC 2016*.

B. H. Partee, and V. Borschev, “The semantics of Russian Genitive of Negation: The nature and role of Perspectival Structure”, *Semantics and Linguistic Theory* 14 (2004).

S. Petrov, D. Das, R. McDonald, “A universal part-of-speech tagset”, in *Proc. of LREC 2012,* pp. 2089–2096.

S. Sharoff, J. Nivre, “The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge”, in *Proc. of Dialogue 2011 International Conference on Computational Linguistics and Intellectual Technologies.*

D. Sichinava, “Numeral: Materials for the project of corpus-based Russian grammar (http://rusgram.ru) [RU: Chislitel'noje: Materialy dlja proekta korpusnogo opisanija russkoj grammatiki (http://rusgram.ru)”, Ms. Moscow, 2012.

R. Tsarfaty, “A unified morpho-syntactic scheme of Stanford dependencies”, in *Proc. of ACL 2013*, pp. 578–584.

D. Zeman, “Reusable tagset conversion using tagset drivers”, in *Proc. of LREC 2008*, pp. 213-218.

APPENDIX

Distribution of dependency tags in the UD-Russian-SynTagRus and UD-Russian-Google treebanks.

UD-Russian-SynTagRus			UD-Russian-Google		
DEPREL	links attested	% corpus	DEPREL	links attested	% corpus
acl	2071	0.20%	acl	1121	1.13%
acl:relcl	6627	0.64%	acl:relcl	653	0.66%
advcl	16742	1.62%	advcl	527	0.53%
advmod	63530	6.15%	advmod	2431	2.45%
amod	97892	9.48%	amod	10882	10.95%
appos	7922	0.77%	appos	2835	2.85%
aux	2244	0.22%	aux	61	0.06%
auxpass	2568	0.25%	auxpass	535	0.54%
case	95548	9.25%	case	10234	10.30%
cc	41123	3.98%	cc	2881	2.90%
cc:preconj			cc:preconj	212	0.21%
ccomp	892	0.09%	ccomp	295	0.30%
compound	1929	0.19%	compound		
conj	48527	4.70%	conj	4517	4.54%
cop	5083	0.49%	cop	544	0.55%
csubj			csubj	9	0.01%
csubjpass			csubjpass		
dep	4518	0.44%	dep	50	0.05%
det	21227	2.06%	det	1445	1.45%
discourse			discourse	565	0.57%
dobj	88674	8.59%	dobj	2597	2.61%
expl	33	0.00%	expl	29	0.03%
foreign			foreign	11	0.01%

UD-Russian-SynTagRus			UD-Russian-Google		
goeswith			goeswith	2221	2.23%
iobj	2948	0.29%	iobj	1079	1.09%
list			list	588	0.59%
mark	14440	1.40%	mark	639	0.64%
mwe	1090	0.11%	mwe	330	0.33%
name	5273	0.51%	name	994	1.00%
neg	13290	1.29%	neg	434	0.44%
nmod	117253	11.35%	nmod	19181	19.30%
nmod:agent	1835	0.18%	nmod:agent		
nsubj	74945	7.26%	nsubj	4955	4.99%
nsubjpass	6958	0.67%	nsubjpass	671	0.68%
nummod	8514	0.82%	nummod	1554	1.56%
nummod:entity	309	0.03%	nummod:entity		
nummod:gov	1004	0.10%	nummod:gov		
parataxis	21201	2.05%	parataxis	601	0.60%
punct	188918	18.29%	punct	17791	17.90%
remnant			remnant	220	0.22%
reparandum			reparandum		
root	59130	5.73%	root	5030	5.06%
vocative			vocative	7	0.01%
xcomp	8386	0.81%	xcomp	660	0.66%
Total	1032644	100.00%	Total	99389	100.00%

Olga Lyashevskaya
National Research University Higher School of Economics, Moscow, Russia; Vinogradov Institute of the Russian Langiage RAS, Moscow, Russia; olesar@yandex.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.