Russian Dependency Treebank

Dan Zeman (zeman@ufal.mff.cuni.cz), 2/2/2006

This file contains notes on a portion of the Russian Dependency Treebank I got from Igor Boguslavsky (Игор Богуславский, bogus@iitp.ru) for the purpose of testing a parser. If anyone wants to use the corpus for their research, please let Igor know.

The portion contains 20 files, 3263 sentences, 39726 words (excluding punctuation[1]). The original file names use Cyrillic letters (Igor would tell you they are using the Windows codepage cp1251 but in fact, Windows uses so-called OEM codepage for filenames, cp855 in this case[2]). I have re-named the files because not all file systems support non-ASCII filenames the same way (if at all) and it is also impossible to open such files in Perl 5.8 (as far as I know). The original names may be needed for reference purposes; therefore I indicate them in the following table. All the original files were dated 8/14/2005.

ASCII file name	Original file name	File size (KBytes)	Words	Sentences
A_on_mjatezhnyj_2.tgt	А_он_мятежный_2.tgt	85	52	818
Anketa.tgt	Анкета.tgt	297	321	2920
Armenija.tgt	Армения.tgt	249	160	2397
Artist_mimansa.tgt	Артист_миманса.tgt	629	534	6226
Ataka.tgt	Атака.tgt	81	48	777
Atlanty_i_atlantologi.tgt	Атланты_и_атлантологи.tgt	182	106	1741
Avtomatizacija.tgt	Автоматизация.tgt	120	77	1169
Baklanov.tgt	Бакланов.tgt	508	548	4993
Batarejka.tgt	Батарейка.tgt	79	45	744
Bessonnica.tgt	Бессонница.tgt	79	60	866
Bez_wpohi.tgt	Без_эпохи.tgt	190	128	1840
Biologija.tgt	Биология.tgt	141	75	1344
Bionika.tgt	Бионика.tgt	143	86	1393
Bol+shie_peremeny.tgt	Большие_перемены.tgt	203	152	2022
Dvoe_v_dekabre.tgt	Двое_в_декабре.tgt	225	167	2713
Gorodskoe_slonovodstvo.tgt	Городское_слоноводство.tgt	158	94	1548
V_zimu_bez_grippa.tgt	В_зиму_без_гриппа.tgt	50	35	485
Vash_personal+nyj_nimb.tgt	Ваш_персональный_нимб.tgt	182	123	1731
Vremja_torgovli.tgt	Время_торговли.tgt	52	50	542
Vyzhivshij_kamikadze.tgt	Выживший_камикадзе.tgt	343	402	3457

[1] Punctuation marks are not treated as words in the Russian Treebank. A word is surrounded by the XML tags <W> and </W>, respectively. If there is a punctuation mark (comma, period etc.) after the word in the original text, it immediately follows the closing tag ("</W>,").

[2] Some more technical details: 1. I received the files in a zipped archive. During extracting, WinZip wanted to use cp1250/cp852 which are native to my Windows system. These names were already damaged in some way. The original names could be found (and decoded from cp855) in the zip file comments. 2. The original encoding of the filenames should not affect their appearance in this doc file which uses Unicode (UTF-8 if exported to HTML).

Contents of the files

The file contents use the cp1251 encoding. The annotations shall be compatible with the TEI guidelines. For a quick start:

- A sentence is enclosed in <S ID="…">…</S>.
- A word is enclosed in <W …>…</W>. One word usually occupies one line.
- Punctuation marks are not considered words. They appear inside <S></S> but outside <W></W>.
- Most of morphology is captured by the LEMMA and FEAT attributes of <W> (see below).
- Most of syntax is captured by the DOM and LINK attributes of <W> (see below).

Morphology

The LEMMA attribute contains the base form of the word (e.g., infinitive of a verb or singular nominative of a noun).

The FEAT attribute contains the part-of-speech and the morphological features of the word. The original description of the values is in the pamjatka_corpus.doc file accompanying the corpus. I have translated some relevant parts from Russian and included them below.

Since my current goal is to convert the corpus into a format compatible with the Prague Dependency Treebank (PDT), I am including corresponding PDT tags when applicable.

The value of the FEAT attribute is a string of uppercase words (both in Latin and Cyrillic scripts), separated by spaces. Each word is a value of a particular morphological feature. The feature repertoire depends on the part of speech, which is always the first feature in the row.

Part of speech (Часть речи)

Value	**Meaning (Russian)**	**Meaning (English)**	**Example**	**Translation**	**PDT**
S	Существительное	Noun	*завод, я*	*company, I*	N, P
A	Прилагательное	Adjective	*новый, мой, второй*	*new, my, second*	A, P, C
V	Глагол	Verb			V
ADV	Наречие	Adverb	*плохо, отчасти*	*badly, partially*	D
NUM	Числительное	Numeral	*пять, 2*	*five, 2*	C
PR	Предлог	Preposition			R
COM	Композит	Compound (part)	*авиа, гидро, агро*	*avia, hydro, agro*	A
CONJ	Союз	Conjunction			J
PART	Частица	Particle			T
P	Слово-предложение	One-word sentence	Only: *нет, да*	Only: *no, yes*	T
INTJ	Междометие	Interjection			I
NID	Иноязычное слово	Foreign word	*Берлинер*	*Berliner*	(F), X

	или несловесная формула	or non-word string	*Цайтунг, Berliner Zeitung, Щ243*	*Cajtung, Berliner Zeitung, ŠČ243*	

Traditional pronouns (Местоимения) are spread among nouns, adjectives and adverbs according to their morphological and syntactic features. Ordinal numerals are considered adjectives. This is a difference from the PDT.

Number (Число)

Value	**Meaning (Russian)**	**Meaning (English)**	**PDT n**
ЕД	Единственное	Singular	S
МН	Множественное	Plural	P, D

Gender (Род)

Value	**Meaning (Russian)**	**Meaning (English)**	**PDT g**
МУЖ	Мужской	Masculine	M, I
ЖЕН	Женский	Feminine	F
СРЕД	Средний	Neuter	N

Case (Падеж)

Value	**Meaning (Russian)**	**Meaning (English)**	**PDT c**
ИМ	Именительный	Nominative	1
РОД	Родительный	Genitive	2
ПАРТ	Партитивный (*нет чаю*)	Partitive	2
ДАТ	Дательный	Dative	3
ВИН	Винительный	Accusative	4
ТВОР	Творительный	Instrumental	7
ПР	Предложный	Prepositional	6
МЕСТН	Местный (*в лесу*)	Locative	6

Note: the ПАРТ and МЕСТН cases apply only to nouns, where their forms differ from the forms of the РОД and ПР cases, respectively.

Animateness (Одушевленность)

Value	**Meaning (Russian)**	**Meaning (English)**	**PDT g**
ОД	Одушевленное	Animate	M, F, N
НЕОД	Неодушевленное	Inanimate	I, F, N

Degree of comparison (Степень сравнения)

Value	Meaning (Russian)	Meaning (English)	PDT
СРАВ	Сравнительная	Comparative	2
ПРЕВ	Превосходная	Superlative	3

Shortness (Краткость)

Some adjectives and adverbs have dual forms, long and short. Long is the default.

Value	Meaning (Russian)	Meaning (English)	PDT
КР	Краткое	Short	

Verb form (Репрезентация)

The default verb form is finite, present tense (личная форма, "personal form"). For other forms, one of the following has to be indicated.

Value	Meaning (Russian)	Meaning (English)	PDT
ИНФ	Инфинитив	Infinitive	Vf
ПРИЧ	Причастие	Participle	Vs
ДЕЕПР	Деепричастие	Transgressive	Ve, Vm

Mood (Наклонение)

Value	Meaning (Russian)	Meaning (English)	PDT
ИЗЪЯВ	Изъявительное	Indicative	
ПОВ	Повелительное	Imperative	Vi

Aspect (Вид)

Value	Meaning (Russian)	Meaning (English)	PDT
НЕСОВ	Несовершенный	Imperfective	lemma
СОВ	Совершенный	Perfective	lemma

Tense (Время)

Value	Meaning (Russian)	Meaning (English)	PDT
НЕПРОШ	Непрошедшее (настоящее/будущее) (*читаю; прочитаю*)	Non-past (present/future)	VB
ПРОШ	Прошедшее (*читал; прочитал; был*)	Past	Vp
НАСТ	Настоящее (*есть, суть*)	Present	VB

Person (Лицо)

Value	Meaning (Russian)	Meaning (English)	PDT
1-Л	Первое	First	1
2-Л	Второе	Second	2
3-Л	Третье	Third	3

Reflexive voice (Страдательный залог)

Value	Meaning (Russian)	Meaning (English)	PDT
СТРАД	Страдательный	Reflexive	

Additional characteristics (Дополнительные характеристики)

Value	Meaning (Russian)	Meaning (English)	PDT
СЛ	Форма, используемая в словосложении (*турецко, авиа*)	Form used in compounds	
СМЯГ	Смягченная сравнительная степень (*поумнее, пораньше*)	Softened comparative degree	

Syntax

The DOM attribute contains either a number or the string "_root". A number is to be interpreted as the index of the node governing the present word in the sentence. Governing is the word that has the same value in its ID attribute.

The LINK attribute of the dependent node classifies the type of the relation between the governing and the dependent nodes.

Актантные отношения (valency relations)

Value	Transliteration	Meaning (Russian)	Remarks
предик	predik	предикативное	predicate-subject
дат-субъект	dat-sub"ekt	дательно-субъектное	predicate-dative subject
агент	agent	агентивное	reflexive predicate or deverbative noun – instrumental noun
квазиагент	kvaziagent	квазиагентивное	predicative noun – word realizing its subject
несобст-агент	nesobst-agent	несобственно-агентивное	*Между* [dep] *Англией и Францией шла* [gov] *война.*
1-компл	1-kompl	комплетивное	predicate – arguments other than subject
2-компл	2-kompl		
3-компл	3-kompl		

4-компл	4-kompl		
5-компл	5-kompl		
присвяз	prisvjaz	присвязочное	copula – nominal predicate
1-несобст-компл	1-nesobst-kompl	несобственно-комплетивное	verb - dependent (?)
2-несобст-компл	2-nesobst-kompl		
3-несобст-компл	3-nesobst-kompl		
неакт-компл	neakt-kompl	неактантно-комплетивное	predicate – non-valent benefactive in dative case
компл-аппоз	kompl-appoz	комплетивно-аппозитивное	valent relation between two nouns
предл	predl	предложное	preposition – noun
подч-союзн	podč-sojuzn	подчинительно-союзное	subordinating conjunction – subordinated predicate
сравнит	sravnit	сравнительное	comparative relation *(better – than)*
сравн-союзн	sravn-sojuzn	сравнительно-союзное	comparative relation *(than – me)*
элект	èlekt	элективное	selected member – preposition starting the selection *(best – of)*

Атрибутивные отношения (attributive relations)

Value	**Transliteration**	**Meaning (Russian)**	**Remarks**
опред	opred	определительное	noun – attribute
оп-опред	op-opred	описательно-определительное	noun – attribute delimited by commas
аппрокс-порядк	approks-porjadk	аппроксимативно-порядковое	*числа* [gov] *первого* [dep]
релят	reljat	релятивное	noun – relative clause
атриб	atrib	атрибутивное	noun – attribute
композ	kompoz	композитное	last part of a compound – first part
аппоз	appoz	аппозитивное	left member of an apposition – right member
об-аппоз	ob-appoz	обособленно-аппозитивное	apposition (right member delimited by punctuation marks)
ном-аппоз	nom-appoz	номинатиовно-аппозитивное	apposition (right member in nominative regardless the case of the left member)
нум-аппоз	num-appoz	нумеративно-аппозитивное	apposition (right member is a number identifying the left member)
количест	količest	количественное	counted noun – number

аппрокс-колич	approks-količ	аппроксимативно-количественное	counted noun – number
колич-копред	količ-kopred	количественно-копредикативное	
колич-огран	količ-ogran	количественно-ограничительное	
распред	raspred	распределительное	parameter – its extent (miles – per hour)
аддит	addit	аддитивное	*it costs two* [gov] *fifty* [dep]
обст	obst	обстоятельственное	verb – adverb
длительн	dlitel'n	длительное	verb – duration modifier
кратно-длительн	kratno-dlitel'n	кратно-длительное	verb – duration modifier
дистанц	distanc	дистанционное	verb – spatial length modifier
обст-тавт	obst-tavt	обстоятельственно-тавтологическое	verb – locative noun overlapping the verb's meaning
суб-обст	sub-obst	субъектно-обстоятельственное	verb – modifier
об-обст	ob-obst	объектно-обстоятельственное	verb – modifier
суб-копр	sub-kopr	субъектно-копредикативное	verb – modifier
об-копр	ob-kopr	объектно-копредикативное	verb – modifier
огранич	ogranič	ограничительное	verb – modifier
вводн	vvodn	вводное	verb – modifier telling what the author things about the action
изъясн	iz"jasn	изъяснительное	verb – explaining modifier
разъяснит	raz"jasn	разъяснительное	term – enumeration
примыкат	primykat	примыкательное	multi-word expression – abbreviation
уточн	utočn	уточнительное	modifier – more precise modifier

Сочинительные отношения (Coordination)

Value	**Transliteration**	**Meaning (Russian)**	**Remarks**
сочин	sočin	сочинительное	coordination (the leftest member is the local root, all other members depend on their preceding members)
сент-соч	sent-soč	сентенциально-сочинительное	sentential coordination
соч-союзн	soč-sojuzn	сочинительно-союзное	coordinating conjunction – right member of coordination
ком-сочин	kom-sočin	коммуникативно-сочинительное	relation that is both coordinating and subordinating (?)

кратн	kratn	кратное	*3:4; годен – негоден; обязательные / факультативные позиции*

Служебные отношения (Auxiliary relations)

Value	Transliteration	Meaning (Russian)	Remarks
аналит	analit	аналитическое	connects auxiliary verbs in analytical verb forms
пасс-анал	pass-anal	пассивно-аналитическое	the verb *to be* – passive participle
вспом	vspom	вспомогательное	connects the two parts of a common multi-word expression (*слева* [gov] *направо* [dep]; *само* [gov] *собой* [dep]; *сам* [dep] *себя* [gov], *да* [dep] *здравствует* [gov], *друг* [dep]*за другом* [gov])
колич-вспом	količ-vspom	количественно-вспомогательное	from right to left, connects parts of a compound adjective, that is semantically an ordinal numeral
соотнос	sootnos	соотносительное	multi-word conjunctions, prepositions and particles
эксплет	èksplet	эксплетивное	relative clauses depending on demonstratives
пролепт	prolept	пролептическое	*Школа* [dep] – *это* [gov] *наш дом!*
эллипт	èllipt	эллиптическое	connects parts of a syntactically non-continuous sentence (ellipsis)

Selected syntactic features

Due to reintroducing deleted words (ellipsis), the actual number of nodes in a tree may be greater than the number of words in the original sentence. Reintroduced nodes would have the feature ФАНТОМ (FANTOM) in their FEAT attribute. I found no occurrences of the feature in the portion of data we have.

Other sort of "phantom" nodes have their LEMMA beginning with "НЕОПР-ГЛАГОЛ" (NEOPR-GLAGOL). Such nodes are not found in our data, either.

Coordination: the first member is the local root, the conjunction hangs on it and the second member hangs on the conjunction *(собирали → гриби → и → ягоды)*. (In PDT, the conjunction would be the local root and both the members would hang on it.)

Adverbial modifiers: a more specific modifier depends on the less specific one *(it is → in the room → on the desk)*. (In PDT, both would depend on the verb.)

References

Igor Boguslavsky, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Kreidlin, Nadezhda Frid: *Dependency Treebank for Russian: Concept, Tools, Types of Information.* In: Proceedings of the International Conference on Computational Linguistics (Coling 2000), Universität des Saarlandes, Saarbrücken, Germany, 2000

Igor Boguslavsky, Ivan Chardin, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Iomdin, Leonid Kreidlin, Nadezhda Frid: *Development of a Dependency Treebank for Russian and its Possible Applications in NLP.* In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), vol. III, pp. 852–856. Las Palmas, Spain, 2002

Игор М. Богуславский, Леонид Л. Иомдин, Виктор Г. Сизов, Иван С. Чардин: *Использование размеченного корпуса текстов при автоматическом синтаксическом анализе (Using a corpus of annotated texts for the automatic syntactic parsing).* In: Труды Международной конференции «Когнитивное моделирование в лингвистике-2003». Варна, Bulgary, 2003

Дмитрий В. Сичинава: *К задаче создания корпусов русского языка (On the problem of building Russian linguistic corpora for the Internet).* http://www.mccme.ru/ling/mitrius/article.html, 2001

Глубоко аннотированный корпус русских текстов, информация для пользователя (Deeply annotated corpus of Russian texts, information for the user). http://proling.iitp.ru/; pamjatka_corpus.doc