A Positional Tagset for Russian

Jirka Hana and Anna Feldman

Charles University, Montclair State University
first.last@gmail.com, first.last@montclair.edu

Abstract

Fusional languages have rich inflection. As a consequence, tagsets capturing their morphological features are necessarily large. A natural way to make a tagset manageable is to use a structured system. In this paper, we present a positional tagset for describing morphological properties of Russian. The tagset was inspired by the Czech positional system (Hajič, 2004). We have used preliminary versions of this tagset in our previous work (e.g., Hana et al. (2004, 2006); Feldman (2006); Feldman and Hana (2010)). Here, we both systematize and extend these preliminary versions (by adding information about animacy, aspect and reflexivity); give a more detailed description of the tagset and provide comparisons with the Czech system.

1. Introduction

In this paper we present a positional tagset for capturing morphological properties of Russian words. The tagset was inspired by the Czech positional system (Hajič, 2004). We have used preliminary versions of this tagset in our previous work (e.g., Hana et al. (2004, 2006); Feldman (2006); Feldman and Hana (2010)). Here, we both systematize and extend these preliminary versions (by adding information about animacy, aspect and reflexivity); give a more detailed description of the tagset and provide comparisons between the two systems. Technical details about the tagset can be found at `http://purl.org/net/rutags`.

2. Positional tagset

Russian has rich inflection. Thus, any tagset capturing its morphological features is necessarily large. A natural way to make a tagset manageable is to use a structured system. In such a system, a tag is a composition of tags each coming from a much smaller and simpler atomic tagset associated with a particular morpho-syntactic property (e.g. gender or tense).

Examples of structured tagsets are what we call *positional tagsets* and *compact tagsets*. In both systems, the tags are sequences of values encoding individual morphological features. In a positional tagset, all tags have the same length, encoding all the features distinguished by the tagset. Features not applicable for a particular word have a N/A value. In a compact tagset, the N/A values are left out. Usually, part-of-speech or a similar category (e.g. the so-called SubPOS in the Czech Positional Tagset (Hajič, 2004)) determines which values are applicable and which are not. For example, `AAFS4----2A----` in the Czech Positional Tagset and `AFS42A` in the Czech Compact Tagset (now obsolete) both encode the same information: adjective (`A`), feminine gender (`F`), singular (`S`), accusative (`4`), comparative (`2`), not-negated (`A`).

For large tagsets, a structured system has many practical benefits. For example,

1. *Learnability:* It is much easier to link traditional linguistic categories to the corresponding structured tag than to an unstructured atomic tag. While it takes some time to learn the positions and the associated values of the Czech Positional Tagset, for most people, it is still far easier than learning the corresponding 4000+ tags as atomic symbols.
2. *Decomposability:* The fact that the tag can be decomposed into individual components has been used in various applications. For instance, the tagger of (Hajic and Hladká, 1998), for a long time the best Czech tagger, operates on the subtag level.
3. *Systematic evaluation:* The evaluation of tagging results can be done in a more systematic way. Each category can be evaluated separately on each morphological feature. Not only is it easy to see on which POS the tagger performs the best/worst, but it is also possible to determine which individual morphological features cause the most problems.

It is also worth noting that it is trivial to view a structured tagset as an atomic tagset (e.g. by assigning a unique natural number to each tag), while the opposite is not true.

3. Positions

Pos	Abbr	Name	Nr. of values
1	p	Part of Speech	12
2	s	SubPOS (Detailed Part of Speech)	42
3	g	Gender	4
4	y	Animacy	3
5	n	Number	3
6	c	Case	7
7	f	Possessor's Gender	4
8	m	Possessor's Number	2
9	e	Person	4
10	r	Reflexivity	2
11	t	Tense	4
12	b	Verbal aspect	3
13	d	Degree of comparison	3
14	a	Negation	2
15	v	Voice	2
16	i	Variant, Abbreviation	7

Table 1: Positions of the Russian tagset

Table 1 lists the positions of the tagset. In most cases, the positions correspond to traditional categories in an obvious way. Below we discuss several positions where this correspondence is not as obvious. This includes cases where the distinction between adjectives and participles is obscure; the gender features that capture both lexical gender of nouns and agreement gender of adjectives, pronouns, numerals, and verbs; the category of reflexivity for verbs and deverbal categories as well as reflexivity of personal and possessive pronouns. The number of values in the table excludes the N/A value, but does include the wildcard `X` value, if it exist. We list and describe all the values for each positions.

The positions are similar to the Czech tagset. There are three additional positions corresponding to three categories not encoded in by the Czech tagset: animacy used in most nominals, reflexivity used in verbs, participles and pronouns, and aspect used in verbs.

3.1. Part of Speech

Most POS values are traditional and include nouns, adjectives, verbs, prepositions, etc. However, there are some other distinctions as well. Participles in Russian behave similarly to adjectives. They agree with nouns they modify in gender, number, and case; their forms are identical to those of adjectives'. Since their syntactic distribution is different, they are distinguished from adjectives in the subPOS position. Gerunds ("verbal adverbs") are treated as a special form of verbs and are distinguished in the subPOS position as well.

3.2. Detailed Part of Speech

This position specifies the POS in more detail. For example, it distinguishes long and short adjectives, various types of pronouns, such as personal, possessive, demonstrative, interrogative pronouns; finite and infinite verbs and so on. One might criticize our distinctions here as the ones that do not happen at the same linguistic level. We do agree with this criticism. However, it was a compromise between the size of the tagset and the ability to identify fine-grained distinctions that are extremely rare in the language and perhaps do not deserve a separate slot in the tag. This is a price for using a positional tagset, which we are ready to pay.

3.3. Animacy

Animacy manifests itself only in accusative masculine singular and accusative plural of all genders. For nouns, we consider it a lexical feature, on par with gender, thus we marked it for all forms. It is also encoded for all noun modifiers that have different forms depending on the animacy of their the noun. Therefore certain adjectives, pronouns and numerals have animate and inanimate forms in acc.masc.sg. and acc.pl., and a single form (tagged with the wildcard `X` value) otherwise.

3.4. Gender

The gender position stands for grammatical gender, which captures both lexical gender of nouns and agreement gender of adjectives, pronouns, numerals and verbs.

3.5. Possessor's Gender & Number

The tagset distinguishes two number slots (similarly genders):

1. Slot 5: Agreement number for nouns, adjectives, verbs, ordinal numerals, possessive pronouns etc.
2. Slot 8: Number of the possessor for possessive pronouns (possessor's gender is distinguished for possessive adjectives)

The two different numbers are exemplified by the following example:

(1) On / kupil
he / bought
`PPM-S1--3I------` `VBM-S----IR-----`
našu / staruju
our$_{fem.sg}$ / old$_{fem.sg}$
`PSFIS4-P1I------` `AAFIS4------1A--`
fotografiju.
photograph$_{fem.sg}$
`NNFIS4-------A--`
'He bought the old photograph of ours'

'

3.6. Reflexivity

This position captures the traditional category of reflexivity:

- reflexivity of verbs and deverbatives, i.e. the presence of a reflexive suffix (*-sja*)
- reflexivity of personal and possessive pronouns

3.7. Tense

This position encodes *morphological* tense. For example, morphologically present verbs used to express past are annotated as being in the present tense (`P`), or infinitive in compound future tense is tagged as not distinguishing tense (`-`).

3.8. Verbal aspect

Aspect is encoded (at least to some extent) in the verb morphology of Russian, mostly by prefixes. Most linguists more or less confidently prefer to categorize Russian aspect as a derivational category (Karcevski, 1927; Ruzicka, 1952; Dahl, 1985; Bermel, 1997), only very few claim aspect to be an inflectional category (e.g. Isačenko, 1968).

3.9. Negation

The negation slot refers to the presence (value `N`) or absence (`A`) of the negative prefix *ne* for open class words. For pronouns the slot has always `-` value. Words that are not negated synchronically do not have `N` in this slot (they may still have negative semantics, but the initial *ne* is not a morphological prefix anymore), for example *nenavist'* 'hate' is tagged as `NNFIS1-------A--`.

All adjectives, including participles, allow such negation, at least in theory:

- *Ego nevolčij vzljad menja ispugal* 'His non-wolfish look scared me'.

- *Staršij syn byl bolee "nemamin"* 'The eldest son was more non-mother's' (unusual, but at least theoretically possible)

3.10. Variant, Abbreviation

The main function of the last slot is to enable unique generation of forms, i.e. ensure that a lemma with a tag corresponds to a single form. Therefore, if a particular combination of morphological categories can be expressed by more than one form of a single lemma, the variant slot *can* be used to distinguish between them. The values are assigned to forms based on their register (standard, colloquial and archaic) and frequency (common vs. rare). But unlike in the case of the other slots, these are just basic guidelines and strictly speaking, the assignment is arbitrary.

Specifying this position is optional. In applications where such distinction between forms is not needed or even desirable, all forms should be assigned the basic variant (`-`).

Note that this slot is used only to distinguish variants of forms of a single lemma, not to provide information about the register/frequency of lemmas. Therefore, forms of a colloquial/archaic lemma are assigned the basic variant value.

(2) a. *lodkoj* – 'boat$_{sg.inst}$' `NNFIS7-------A--`
    b. *lodkoju* – 'boat$_{sg.inst}$' `NNFIS7-------A-1`
    c. *kapusta* – 'big bucks' `NNFIS1-------A--`

The final slot serves one more function. It is used to mark abbreviations as such. In theory, we could have introduced a dedicated slot. However, because there is very little need for distinguishing variants of abbreviations (abbreviations rarely, if ever, inflect), this would make the tagset more complex without bringing much benefit. Also, this is the way abbreviations are marked in the Czech tagset.

An abbreviation could be seen as a form of a lemma (e.g. *gr.* being a form of *graždanin* 'citizen'). However, because the abbreviation is not really an inflection of the lemma and because many words can be abbreviated in several ways, we decided to use the abbreviation itself as its lemma.

4. Values

Table 3 summarizes possible values for each position. Not all combinations are possible, the set of possible tags is described in the following section.

5. Overview of possible tags

Table 4 provides an overview of the Russian tagset by POS. A template denotes a set of tags. Roman letters refer to particular values, while italics denote variables. Thus for example, to obtain the set of tags corresponding to the template `NN`*`gync`*`------`*`a`*`---`, one needs to instantiate all the possible combinations of the *g* (gender), *y* (animacy), *n* (number), *c* (case), and *a* (negation) variables. In this case, $g \in \{\texttt{F},\texttt{M},\texttt{N},\texttt{X}\}$, $y \in \{\texttt{A},\texttt{I},\texttt{X}\}$, $n \in \{\texttt{P},\texttt{S},\texttt{X}\}$, $c \in \{\texttt{1},\texttt{2},\texttt{3},\texttt{4},\texttt{6},\texttt{7},\texttt{X}\}$, $a \in \{\texttt{A},\texttt{N}\}$. A variable never stands for the `-` (N/A) value. If a single Sub-POS allows a particular position to have both the N/A value and other values, we list them as separate templates.

In some cases, there might be additional restrictions on the possible co-occurrences of values, we mention these restrictions below. Also, the templates are simplified somewhat by ignoring the possibility of having different variants of the same tag (the last slot).

Each template is accompanied by a sample word and a tag corresponding to this word. If the word can be tagged with more than one tag, one is arbitrary chosen.

5.1. Additional restrictions

1. Gender in plural is distinguished by nouns only (i.e. only lexical gender not agreement gender is distinguished).
2. The `X` wild-card values are used in the following cases only:

   (a) Gender: agreement gender in plural (adjectives, participles, determiners, etc.), pluraletantum nouns, non-declinable adjectives (e.g. non-Russian words and abbreviations), personal pronouns in 3rd person plural.
   (b) Animacy: Except for nouns, in all forms except accusative masculine singular and accusative plural of all genders. Non-declinable words in all forms.
   (c) Number: non-declinable nouns, adjectives and verbs, 3rd person possessive pronouns.
   (d) Case: non-declinable nouns and adjectives, 3rd person possessive pronouns.
   (e) Possessor's Gender: for the 3rd person plural possessive pronoun.
   (f) Person: for non-declinable verbs
   (g) Tense: for passive long participles (`AG`).
   (h) Aspect: bi-aspectual verbs, e.g. *ispol'zovat'* 'to use'.

6. Additional Notes

6.1. Numerals

1. *nol'*/*nul'* 'zero' and numerals above 999 (e.g. *tysjača* 'thousand', *milion*) are considered to be regular nouns.
2. Only *odinoždy* 'one time', *triždi* 'three times', etc. are considered to be multiplicative numerals; *šestikratnyj* 'sixfold' is annotated as a regular adjective.
3. Other words related to numerals are considered to be nouns or adjectives: number names (*dvojka* 'number two' – noun); *pjatok* 'five pieces' – noun; composites (*dvuxletnij* 'biannual' – adjective, *pjatiletka* 'five-year period/plan' – noun).

6.2. Participles

1. Participles are classified as adjectives:

   - *čitajuščij* – `AGMXS1---IPI-AA-` – active (A) present (P) participle
   - *čitavšij* – `AGMXS1---IRI-AA-` – active (A) past (R) participle
   - *pročitavšij* – `AGMXS1---IRP-AA-` – active (A) past (R) participle
   - *čitaemyj* – `AGMXS1---IXI-AP-` – passive (P) long (imperf/perf) participle

- *pročitan* – `AcM-S----I-P-AP-` – passive (P) perf. short participle

2. Similarly as in Czech, all *-nyj* (*ostavlennyj* 'deserted', *varenyj* 'cooked', *zadelannyj* 'clogged') participles/adjectives are considered to be general adjectives, because it is very hard to draw the line between their purely adjectival and participial use.

7. Set of tags, Tag abbreviations

7.1. Set of tags

Regular expressions can be used to capture sets of tags or abbreviate tags. For example:

1. `NN[MF]AS[1-3]-------A--` – masculine or feminine singular animate noun, in nominative, genitive or dative
2. `NN.AS[ˆ14]-------A--` – singular animate noun of any gender, not in nominative nor in accusative
3. `II.*` – `II--------------`
4. `NN(MI|FA)S1-------A--` – masculine inanimate or feminine animate noun

These expressions, especially the value `.` (dot), can also be used when certain categories are not predicted or annotated. For example an analyzer not capturing animacy, can output tags with a dot in the animacy position. Note that there is a difference between the dot value and the `X` value. The dot is a technical value as opposed to the `X` value, which is linguistically motivated. For example, in case of gender, the dot value stands for all possible gender values including `X`.

7.2. Tag abbreviations

In documents that are not intended for machine consumption (e.g. manuals for annotators), it is often convenient to abbreviate tags. We suggest the following conventions: The abbreviations are formed by:

1. omitting dashes, especially trailing dashes
2. omitting the most common values for certain positions: `1` for degree, `A` for negation, `I` for reflexivity.
3. adding non-basic variant to the abbreviated tag preceded by a dash (e.g. `-8` for abbreviations).

Examples:

1. `NN`*`gync`* – noun;
   `NNFIS1` = `NNFIS1-------A--`
2. `A`*`sgync`* – adjective;
   `AAXXXX` = `AAXXXX------1A--`
3. `D[bg](`*`d`*`)` – adverb;
   `Db` = `Db--------------`,
   `Dg` = `Dg----------1A--`,
   `Dg2` = `Dg----------2A--`
4. `J[ˆ,]` – conjunction;
   `Jˆ` = `Jˆ--------------`
5. `R[RV]`*`c`* – preposition;
   `RR7` = `RR---7----------`
6. `TT` – particle (similarly `II`, `Z#` `Z:`, `X0`, `XX`);
   `TT` = `TT--------------`
7. `NN`*`gy`*`XX-8` noun abbreviation;
   `NFIXX-8` = `NNFIXX-------A-8`

8. Differences from the Czech positional tagset

Rus	Cze	Abbr	Name
1	1	p	Part of Speech
2	2	s	SubPOS (Detailed Part of Speech)
3	3	g	Gender
4		y	Animacy
5	4	n	Number
6	5	c	Case
7	6	f	Possessor's Gender
8	7	m	Possessor's Number
9	8	e	Person
10		r	Reflexivity
11	9	t	Tense
12		b	Verbal aspect
13	10	d	Degree of comparison
14	11	a	Negation
15	12	v	Voice
	13		Not used
	14		Not used
16	15	i	Variant, Abbreviation

Table 2: Comparison with the Czech Positional Tagset

The Czech and the Russian tagsets encode similar morphological information in a similar manner (i.e. encoding the categories in the same order, and in in most cases the same symbol has the same meaning). However, there are some important differences. Many of them are a consequence of linguistic differences between the languages, but some are the result of us making different design decisions than the authors of the Czech tagset.

The differences can be categorized into four groups:

1. The sets of captured categories. The Russian tagset captures three categories not captured (directly) by the Czech tagset:

   (a) animacy: In Czech, animacy is distinguished only for masculine gender. The Czech tagset splits the traditional masculine gender into two genders: masculine animate and masculine inanimate, thus dispensing with the need of a separate position for animacy.

   (b) reflexivity: In Czech, reflexivity of verbs and deverbal categories is expressed analytically (by orthographically separate reflexive pronouns), therefore there is no need for a morphological reflexivity category. Reflexivity of personal and possessive pronouns is analogous to Russian, however the Czech tagset encodes their reflexivity by assigning them to a separate SubPOS.

   (c) aspect: Verbal aspect is very similar in both languages from a linguistic point of view. However, the standard Czech positional tagset does not capture this distinction.

2. The set of values for a particular category: For example, Russian has neither vocative nor dual, nor does it have auxiliary or pronominal clitics; and the difference between colloquial and official Russian is not as systematic and profound as in Czech.
3. Exact meaning of a particular value symbol.
4. Wildcards: The Russian tagset also uses far fewer wildcards (symbols representing a set of atomic values). Even though wildcards might lead to better tagging performance, we intentionally avoid them. The reason is that they provide less information about the word, which might be needed for linguistic analysis or an NLP application. In addition, it is trivial to translate atomic values to wildcards if needed.

   The Russian tagset contains only wildcards covering all atomic values (denoted by `X` for all applicable positions). There are no wildcards covering a subset of atomic values. Forms that would be tagged with a tag containing a partial wildcard in Czech are regarded as ambiguous.

   For example, where the Czech tagset uses `Z` (all genders except feminine), our Russian tagset uses `M` (masculine) or `N` (neuter) depending on the context. Thus, Czech *tomto* 'this$_{masc/neut.loc}$' is tagged as `PDZS6----------` in *v tomto domě* 'in this house$_{masc}$' and *v tomto místě* 'in this place$_{neut}$', while Russian *ètom* 'this$_{masc/neut.loc}$' is tagged as `PDMXS6----------` in *v ètom dome* 'in this house$_{masc}$' and `PDNXS6----------` in *v ètom meste* 'in this place$_{neut}$'.

9. Acknowledgement

The development of the tagset was supported by the GAČR grant P406/10/P328 and by the NSF grant # 0916280.

Table 3: Values of individual positions of the Russian tagset

Position 1 – POS	
A	Adjective
C	Numeral
D	Adverb
I	Interjection
J	Conjunction
N	Noun
P	Pronoun
V	Verb
R	Preposition
T	Particle
X	Unknown, not determined, unclassifiable
Z	Punctuation
Position 2 – SubPOS	
N	N: Noun
A	A: Adjective (long, non-participle) (*xorošij, ploxoj*)
C	A: Short adjective (non-participle) (*surov, krasiv*)
G	A: Participle, active or long passive (*čitajuščij, čitavšij, pročitavšij, čitaemyj*; but not *pročitannyj* (`AA`), *pročitan* (`Ac`)
c	A: Short passive participle (*pročitan*)
U	A: Possessive adjective (*mamin, oveč'ju*)
P	P: Personal pronoun (*ja, my, ty, vy, on, ona, ono, oni, sebja*)
5	P: 3rd person pronoun in prepositional forms (*nego, nej, . . .)*
S	P: Possessive pronoun (*moj, ego, svoj, ..*)
D	P: Pronoun demonstrative (*ètot, tot, sej, takoj, èkij, . . .*)
Q	P: Relative/interrogative pronoun with nominal declension (*kto, čto*)
q	P: Relative/interrogative pronoun with adjectival declension (*kakoj, kotoryj, čej, . . .*)
W	P: Negative pronoun with nominal declension (*ničto, nikto*)
w	P: Negative pronoun with adjectival declension (*nikakoj, ničej*)
Z	P: Indefinite pronoun with nominal declension (*kto-to, kto-nibud', čto-to, . . .*)
z	P: Indefinite pronoun with adjectival declension (*samyj, ves', . . .*)
=	C: Number written using digits
}	C: Number written using Roman numerals (*XIV*)
n	C: Cardinal numeral (*odin, tri, sorok*)
r	C: Ordinal numeral (*pervyj, tretij*)
j	C: Generic/collective numeral (*dvoje, četvero*)
u	C: Interrogative numeral (*skol'ko*)
a	C: Indefinite numeral (*mnogo, neskol'ko*)
v	C: Multiplicative numeral (*dvaždy, triždy*)
B	V: Verb in present or rarely future form (*čitaju, splju, pišu*)
f	V: Infinitive (*delat', spat'*)

i	V: Imperative (*spi, sdelaj, pročti*)
p	V: Past tense (*spal, ždal*)
e	V: Gerund (*delaja; pridja, otpisav*)
b	D: Adverb without a possibility to form negation and degrees of comparison (*vverxu, vnizu, potom*)
g	D: Adverb forming negation and degrees of comparison (*vysoko, daleko*)
F	R: Part of a preposition; never appears isolated (*nesmotrja*)
R	R: Nonvocalized preposition (*ob, pered, s, v, . . .*)
V	R: Vocalized preposition (*obo, peredo, so, vo, . . .*)
,	J: Subordinate conjunction (*esli, čto, kotoryj*)
ˆ	J: Non-subordinate conjunction (*i, a, xotja, pričem*)
I	I: Interjection (*oj, aga, m-da*)
T	T: Particle (*li*)
#	Z: Sentence boundary
:	Z: Punctuation
0	X: Part of a multiword foreign phrase
X	X: Unknown, Not Determined, Unclassifiable

Position 3 – Gender	Distinguished for: N, A{ACGUc}, P{P5DLwSq8}, C{nra}, VB
F	Feminine
M	Masculine
N	Neuter
X	Any gender

Position 4 – Animacy	Distinguished for: N, A{AGU}, P{SDwqz}, C{nrja}
A	Animate
I	Inanimate
X	Either

Position 5 – Number	Distinguished for: N, A{ACGUc}, P{P5DwSq}, C{nra}, V{Bp}
P	Plural
S	Singular
X	Any number

Position 6 – Case	Distinguished for: N, A{AGU}, P, C{nrjua}
1	Nominative
2	Genitive
3	Dative
4	Accusative
6	Locative
7	Instrumental
X	Any case

Position 7 – Possessor’s Gender	Distinguished for: PS, AU
F	Feminine possessor
M	Masculine possessor
N	Neuter possessor
X	Possessor of any gender

Position 8 – Possessor’s Number	Distinguished for: PP
P	Plural possessor
S	Singular possessor

Position 9 – Person	Distinguished for: P{P5S}, V{Bi}
1	1st person
2	2nd person
3	3rd person
X	Any person

Position 10 – Reflexivity	Distinguished for: AG, P{P5S}, V
R	Reflexive
I	Irreflexive

Position 11 – Tense	Distinguished for: A{G}, V{Bp}
F	Future
P	Present

R	Past
X	Any (Past, Present, or Future)

Position 12 – Aspect — Distinguished for: AG, V

P	perfective
I	imperfective
X	either aspect

Position 13 – Degree of comparison — Distinguished for: AA, Dg

1	Positive
2	Comparative
3	Superlative

Position 14 – Negation — Distinguished for: N, A, Dg

A	Affirmative (not negated)
N	Negated

Position 15 – Voice — Distinguished for: AG, Ac

A	Active
P	Passive

Position 16 – Variant — Distinguished for: As needed

-	Basic variant
1	Variant (generally less frequent)
2	Variant (generally rarely used, bookish, or archaic)
3	Variant (very archaic)
5	Variant (colloquial)
6	Variant (colloquial, generally less frequent)
7	Variant (colloquial, generally less frequent)
8	Abbreviations

Table 4: Overview of the Russian tagset

template	description	sample word	sample tag
N – Nouns			
NN*gync*-------*a*--	noun	*golos*	NNMIS4-------A--
A – Adjectives (incl. Participles)			
AA*gync*------*da*--	long adjective	*tjaželyj*	AAMIS4------1A--
AC*g*-*n*--------*a*--	short adjective	*krasiv*	ACM-S--------A--
AG*gync*---*rtb*-*av*-	long participle *tv* ∈ {PA, RA, XP} i.e. present/past active, passive	*čitajuščij* *smejuščajasja*	AGMXS1---IIP-AA- AGFXS1---RRP-AA-
AU*gyncf*------*a*--	possessive adjective	*mužnin*	AUMXS2M------A--
Ac*g*-*n*--------*a*P-	pass.perf.short participle	*pročitan*	AcM-S--------AP-
P – pronoun			
PP--*nc*--*e*I------	personal pronoun; *e* ∈ {1,2}	*nam*	PP--P3--1I------
PP*g*-*nc*--3I------	personal pronoun 3rd person	*on*	PPM-S1--3I------
PP---*c*---R------	personal reflexive sebja	*sebja*	PP---4---R------
P5*g*-*nc*--3I------	personal p. in prep. forms	*nego*	P5M-S2--3-------
PD*gync*----------	demonstrative	*ètu*	PDFXS4----------
PW---*c*----------	negative (nominal declension)	*ničto*	PW---1----------
Pw*gync*----------	negative (adj declension)	*nikakoj*	PwMXS1----------
PS*gync*-*me*I------	possessive	*moja*	PSFXS1-S1I------
PSXXXX*fm*3I------	possessive	*ego*	PSXXXXMS3I------
PS*gync*---R------	possessive reflexive	*svoj*	PSMXS1---R------
PQ---*c*----------	relative/interrogative (nom decl)	*što, kto*	PQ---1----------
Pq*gync*----------	relative/interrogative (adj decl)	*kakoj*	PqMXS1----------
PZ---*c*----------	indefinite (nom. decl.)	*kogo-to*	PZ---4----------
Pz*gync*----------	indefinite (adj. decl.)	*kakoj-to*	PzMXS1----------
C – Numeral			
C=--------------	numbers (using digits)	*3.14*	C=--------------
C}--------------	roman numeral	*XVII*	C}--------------
Cn*gync*----------	cardinal numeral 1	*odnomu*	CnMAS3----------

Cn*gy*-*c*----------	cardinal numeral 2, poltora	*dvux*	CnMA-2----------
Cn-*y*-*c*----------	cardinal numeral 3,4	*trx*	Cn-A-4----------
Cn--*yc*----------	cardinal numeral 5+	*pjati*	Cn-A-2----------
Cr*gync*----------	ordinal	*pervyj*	CrMXS1----------
Cj-*y*-*c*----------	generic/collective numeral	*dvoix*	Cj-A-3----------
Cu---*c*----------	interrogative	*skol’ko*	Cu---x----------
Ca---*c*----------	indefinite numeral	*neskol’ko*	Ca---1----------
Ca*gync*----------	indefinite num. (adj decl.)	*mnogomu*	CaMXS3----------
Cv--------------	multiplicative	*triždi*	Cv--------------
V – verb			
VB--*n*---*ertb*----	present (rarely fut.) finite form	*otryvaeš’*	VB--P---2IPI----
VB*g*-*n*----*r*R*b*----	past tense	*čital*	VBM-S----IRI----
Ve-------*r*-*b*----	gerund	*grozja*	Ve-------I-I----
		napisav	Ve-------I-P----
Vf-------*r*-*b*----	infinitive	*spat’*	Vf-------I-I----
Vi--*n*---*er*-*b*----	imperative	*rabotaj*	Vi--S---2I-I----
D – Adverb			
Db--------------	adv. not forming negation/degrees	*tam*	Db--------------
Dg----------*da*--	adv. forming negation/degrees	*sil’nee*	Dg----------2A--
R – Preposition			
RR---*c*----------	nonvocalized prep. with *c* case	*nad*	RR---7----------
RV---*c*----------	vocalized prep. with *c* case	*nado*	RV---7----------
RF--------------	part of a multiword prep.	*nesmotrja*	RF--------------
J – Conjunction			
J^--------------	coordinating conj.	*i*	J^--------------
J,--------------	subordinating conj.	*čto*	J,--------------
T – particle			
TT--------------	particle	*net*	TT--------------
I – Interjection			
II--------------	Interjection		II--------------
Z – punctuation			
Z#--------------	Sentence boundary		Z#--------------
Z:--------------	Punctuation	*!*	Z:--------------
X – special			
X0--------------	part of a multiword foreign phrase		X0--------------
XX--------------	unknown		XX--------------

References

Bermel, N. (1997). *Context and the Lexicon in the Development of Russian Aspect*. Berkeley: University of California Press.

Dahl, O. (1985). *Tense and Aspect Systems*. Basil Blackwell, Oxford.

Feldman, A. (2006). *Portable Language Technology: A Resource-light Approach to Morpho-syntactic Tagging*. Ph. D. thesis, The Ohio State University.

Feldman, A. and J. Hana (2010). *A Resource-light Approach to Morpho-syntactic Tagging*. Language and Computers. Amsterdam–New York: Rodopi Press.

Hajic, J. and B. Hladká (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING-ACL Conference*, Montreal, Canada, pp. 483–490.

Hajič, J. (2004). *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Prague, Czech Republic: Karolinum, Charles University Press.

Hana, J., A. Feldman, L. Amaral, and C. Brew (2006). Tagging Portuguese with a Spanish tagger using cognates. In *Proceedings of the Workshop on Cross-language Knowledge Induction, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), Trento, Italy*.

Hana, J., A. Feldman, and C. Brew (2004). A Resource-light Approach to Russian Morphology: Tagging Russian Using Czech Resources. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, Barcelona, Spain, pp. 222–229.

Isačenko, A. V. (1968). *Die russische Sprache der Gegenwart*. Halle-Saale: Niemeyer.

Karcevski, S. (1927). *Système du Verbe Russe; Essai de Linguistique Synchronique*. Prague: Plamja.

Ruzicka, R. (1952). Der Russische Verbalaspekt [Russian verbal aspect.]. *Der Russischunterricht* (5), 161–69.