15

RUSTWOL: A Tool for Automatic Russian Word Form Recognition

LIISA VILKKI

15.1 Introduction

The purpose of this paper is to describe the earlier version of RUSTWOL as a tool for automatic Russian word form recognition. The theoretical foundation of the RUSTWOL program is the two-level model, a language-independent model of morphological analysis and synthesis by Kimmo Koskenniemi (1983). My description is based on a document written by me, when I was working as a linguist at Lingsoft (Vilkki 1997). This earlier version of RUSTWOL was later used at Lingsoft as a basis for a new format. The newer version of RUSTWOL, representing the new format, and its documentation, written by me, are currently available at Lingsoft for customers only.

The main motivation for turning back to history and describing the first quasi-final version of RUSTWOL is that RUSTWOL is one of the most large-scale morphological programs in the tradition of the two-level formalism - yet my document at Lingsoft (Vilkki 1997) is the only presentation of it. The second motivation is that this earlier version has been useful at the Department of Slavonic and Baltic Languages and Literatures at University of Helsinki for the purposes of research and teaching. In 2000, I used this version of RUSTWOL for the morphological analysis of the Russian Corpus of Newspaper Articles, which is available in the University of Helsinki Language Corpus Server (UHLCS) and, in addition, at the University of Tampere. Recently, some changes in the RUSTWOL lexicon and rules have been made by Alexander Paile (2003) for the purposes of the HANCO project (Kopotev, Mustajoki 2003).

Inquiries into Words, Constraints and Contexts.
Antti Arppe et al. (Eds.)

In Koskenniemi's (e.g. 1983,1997) two-level formalism, morphological phenomena are described as relations between lexical and surface levels. One character of the lexical level corresponds to one or a null character of the surface level. Therefore, the two-level formalism is neutral with respect to analysis and generation of word forms.

The RUSTWOL alphabet includes letters, numbers, special characters and diacritic symbols. The surface representations of the word forms are translations of their Cyrillic ortography.

The main components of TWOL are the lexicon and the two-level rules. The RUSTWOL lexicon defines lexical entries and their representations. It also specifies the morphotactic structure of the language and includes part of the morphophonological alternations. The RUSTWOL rule component contains 50 rules, and it deals with fairly natural, transparent alternations. There are also rules for the correct combination of stems and endings. These rules refer to the diacritic symbols of masculinity, animacy, transitivity, reflexivity and imperfective or perfective aspect. In addition, some of the rules control the correct combination of the parts of the compound words. Because the length of this contribution is restricted, it is not possible to describe here the rule component in detail. In addition, only some parts of the lexicon can be focused on.

15.2 An overview of RUSTWOL

The basic lexicon of RUSTWOL is based on a machine-readable version of Zaliznjak (1987). The complete material of this dictionary was not included in the lexicon of RUSTWOL. The dictionaries Jevgen'eva (1981-1984) and Zasorina (1977) were used in order to exclude, e.g. very infrequent words, some words representing special vocabulary and words marked stylistically as colloquial, archaic or local. The word material of the Zaliznjak lexicon was completed with words from Scheitz (1986), Kotelova (1984) and Kahla (1982,1984).

The most productive derivated forms and compounds, listed in Zaliznjak (1987) in their own entries, are treated in RUSTWOL by continuation classes and by the mechanism of compounding. Apart from Zaliznjak, Švedova (1982), Kuznecova and Efremova (1986), Tihonov (1985) and Bukčina and Kalakučkaja (1987) were used on matters relating to the description of derivational morphology and compounding. The evolving RUSTWOL was tested on various text corpora: newspaper and magazine articles and literary texts. These corpora are included in Helsinki Corpus of Russian Texts, which are available in UHLCS.

RUSTWOL described in this paper has a lexicon of approximately 72,000 words. This number is considerably increased by a derivational morphology

and the mechanism for compounding.

RUSTWOL assigns the possible readings of Russian word forms. The readings consist of the base form of the word and the morphological information of the inflected form.

Word form

Reading: "base form" morphological analysis
Reading: "base form" morphological analysis
etc.

RUSTWOL is meant to be used as the basis morphological tool in, for example, text analysis, spelling correction and information retrieval. Here are some central aims and properties of RUSTWOL:

- analyses written standard Russian
- gives the morphological information of Russian word forms
- has a complete inflectional morphology and a fairly extensive derivational and compounding morphology
- contains the basic vocabulary of Russian
- prefers traditional morphological categories

15.3 Lexicon

The lexicon of RUSTWOL consists of sublexicons connected to each other. The sublexicons include mostly root entries but also some full-form entries. Examples of root entries are given in 15.3.1. A full-form entry contains the whole word form and its morphological information. Only the extremely irregular word forms and some compounds, both parts of which are inflected (see 15.3.3), are coded as full-form entries. For example, a pronoun form **c1to-to** 'something' has the following two full-form entries:

c1to-to # "c1to-to INDEF PRON ACC";
c1to-to # "c1to-to INDEF PRON NOM";

15.3.1 Inflection

RUSTWOL incorporates a full description of Russian inflectional morphology. It uses 14 parts of speech. The parts of speech and other morphological properties of word forms are indicated by tags. To each word, a base form and at least one tag is associated.

Verbs

Verbs are labelled with V and they are identified by aspect, mood, tense, person and number, voice and reflexivity. Past tense forms are not identified by person and number but by gender or plurality. Here are some examples of the forms of a verb **delat'** 'to do':

delat’
“delat’” IMPF V INF ACT
delaet
“delat’” IMPF V PRES SG3 ACT
delal
“delat’” IMPF V PAST MA ACT

Participles are labelled with PART and verbal adverbs with V ADV. Long–form participles change according to gender, number and case. Short-form participles have a label SH. The following examples are forms of the verb **c1itat’** ‘to read’:

c1itaemyj
“c1itat’” IMPF V PRES PART MA SG NOM PASS
“c1itat’” IMPF V PRES PART MA SG ACC PASS
c1itaema
“c1itat’” IMPF V PRES PART PASS SH FE
c1itav
“c1itat’” IMPF V PAST V ADV ACT
c1itano
“c1itat’” IMPF V PAST PART PASS SH NE
“c1itat’” IMPF V PART PASS PRED

Verbs are divided into two conjugations (1V and 2V). Some verbs do not clearly belong to any of these conjugations (V). Nearly all inflectional types of verbs have an alternation pattern, which is a sublexicon that lists the lexical representations of suppletion-like alternatives. The stems of verbs are usually formed from roots and from some alternation pattern. Many inflectional types of verbs have variants for impersonal verbs. They are indicated by a tag IMPERS before other tags.

Correct combinations of stems and endings are defined by using proper continuation classes for each alternation entry or ending in the lexicons. In some cases rules are used for forbidding or permitting only certain combinations. For example, rules (33)-(39), concerning diacritics P~, P, V, V~, R and R~, exclude invalid combinations (see below).

Verbal endings are grouped into a few sublexicons. Continuation classes can also consist of a single minilexicon:

1V1: PRES PART ACT
1V2: PRES PART PASS, PRES V ADV
2V2: PRES PART ACT

For example, a verb **pet’** ‘to sing’ has the following entry that refers to the LEXICON oJ-e/1V:

pP~VR~ oJ-e/1V;

LEXICON oJ-e/1V
oJ 1V011 “et’”;
ojP~ 1V1 “et’Q4”;
oJP~ 1V2 “et’”;
e V021 “et’”;
eV V32 “et’Q5”;

The first stem, poJ, has the continuations 1V011, 1V1 and 1V2, and the second stem, pe, has the continuations V021 and V32. All verb entries of this inflectional type refer to this minilexicon. However, only the continuation classes in the first and in the fourth entry of the minilexicon are possible for all verbs of this inflectional type. In the second and third entry, P~ is a symbol of imperfective aspect. For all verb stems that are marked P, a symbol of perfective aspect, the continuation 1V1 and 1V2 are excluded.

There are three types of diacritics in verb inflection. They indicate aspect (P and P~), transitivity (V and V~) and reflexivity (R and R~). Most of the verb stems have either P or P~. In this way, perfective and imperfective stems of the same inflectional type can have the same ending minilexicons. Diacritic V is used in transitive verbs and V~ in intransitive verbs. Only stems marked V can get PAST PART PASS and PART PASS PRED ending. These endings are given in minilexicons V31 and V32. So before continuation classes referring to these minilexicons there is a diacritic V.

In minilexicon 1V2, the entry PRES PART PASS demands both V and P~. These diacritics are also necessary in the case of PASS REFL. Only stems having them can get PASS REFL ending -sa1 (in minilexicon RF0) or -s’ (in minilexicon RF1). These endings can be added after personal (PRES/FUT), past tense or infinitive endings. After PRES PART ACT or PAST PART ACT ending some adjectival ending is added and only after it reflexive ending.

All stems marked R get interpretation ACT REFL. These are called reflexive verbs. Many imperfective verb forms have both ACT REFL and PASS REFL interpretations. For example, a form **c1itau1s2ijsa1** of the verb **c1itat’** ‘to read’ is given the following interpretations:

c1itau1s2ijsa1
“c1itat’” IMPF V PRES PART MA SG NOM ACT REFL
“c1itat’” IMPF V PRES PART MA SG ACC ACT REFL
“c1itat’” IMPF V PRES PART MA SG NOM PASS REFL
“c1itat’” IMPF V PRES PART MA SG ACC PASS REFL

The first and the second interpretations represent the following entry:

c1itaP~V~R J-0/1V-R;

The third and the fourth interpretations are representations of a different kind of entry:

c1itaP~VR~ J-0/1V;

Some verb forms (V ADV, IMPV, PASS PART) cannot get PASS REFL ending, even if their stems have diacritics P~ and V. Therefore, these forms have continuations to minilexicons where PASS REFL interpretation is lacking.

Nouns

Nouns are given a tag N, and they are categorized by gender, number and case. A noun **dom** 'house' has, for example, these inflectional forms:

dom
"dom" N MA SG NOM
"dom" N MA SG ACC
dome
"dom" N MA SG PREP

The main declension types of nouns are determined by gender: masculine (/1SM), feminine (/2SF and /3SF) and neuter (/1SN). All of them have subtypes. These are distinguished on the basis of, for example, **u/u1** ending in MA SG GEN and MA SG PREP, various exceptional plural forms and various alternation patterns.

Some of the subtypes are further divided into two types of minilexicons: words representing the first type cannot be used as the first parts of compound words, whereas words representing the second type can be used (see 15.3.3).

In addition, there are declension types and subtypes for words that are inflected like feminines but are used syntactically as masculines (/2SM) or either as masculines or as feminines (/2SMF) and for words that are inflected like neuters but are used syntactically as masculines (/1SMN). Words that cannot be inflected (/SM-ind, /SF-ind and /SN-ind) and words occurring only in plural have their own types, too.

A nominal declension type usually includes one or more continuation classes of singular forms and plural forms. Some continuation classes consist of only a single minilexicon.

Some endings in nominal ending lexicons have diacritics N (animate), N~ (inanimate) or M (masculine). Therefore, only noun stems having appropriate diacritics can get these endings. The combination of stems and endings is controlled by rules. For example, a noun **divo** 'miracle' has an entry of the following kind:

divM~N~ /1SN;

The stem **div** gets some of its endings in minilexicons 1SM1, SPL1 and SPL3. It has diacritics M~ and N~ and, therefore, SG ACC ending in minilexicon 1SM1 and PL ACC ending in minilexicon SPL3 cannot be added. By contrast, the stem can be combined with PL ACC ending in minilexicon SPL1. A diacritic Q3 in minilexicons is used for compound formation (see 15.3.3).

LEXICON 1SM1
AQ3 TO " SG GEN";
UQ3 TO " SG DAT";
AMNQ3 TO " SG ACC";
OmQ3 TO " SG INSTR";
eQ3 TO " SG PREP";

LEXICON SPL1
AQ3 TO " PL NOM";
AN~Q3 TO " PL ACC";

LEXICON SPL3
Q3 TO " PL GEN";
NQ3 TO " PL ACC";

Other parts of speech

The description of other parts of speech is presented in Vilkki 1997. This version of RUSTWOL does not contain special labels for proper names and abbreviations. However, capital letters in these words have an asterisk (*). Proper nouns can be inflected in various declension types of adjectives, nouns and pronouns. Some of them, like all abbreviations, are not inflected.

15.3.2 Derivation

The version of RUSTWOL described here has only a system of first-degree derivation. Most of the adverbs and predicatives are derived from adjectives. Nouns with various suffixes are derived from adjectives or verbs.

Adverbial or predicative suffixes:

-o/e	otkryt-o
-i	a1nvarsk-i

Nominal suffixes:

-ost'/est'	a1dovit-ost'
-nost'	gotov-nost'
-stvo	grabitel'-stvo
-estvo	imus2-estvo
-instvo	dosto-instvo
(-jstvo)	bespoko-jstvo
-izm	biolog-izm
-'	glub-'
-ina	glub-ina
-nie	avla1-nie
-anie	z1ivopis-anie
-ovanie/evanie	absorbirov-anie
-a1nie	ble-a1nie
-enie	opolz-enie

Besides the kinds of derived words listed above there are, of course, many other kinds of derived words in Russian. The most frequent of these have entries in the lexicon.

15.3.3 Compounding

This first version of RUSTWOL has a mechanism for building compounds, mainly consisting of two parts. The most frequent compounds, consisting of more than two parts, are listed in the lexicon. Only the most productive first parts are chosen in productive compound formation. The bulk of the first parts can occur as independent words, too. The continuation classes of these roots include a continuation to the Stem1 or Stem2 lexicon as one alternative. Many compounds have a hyphen and/or a linking element **O** or **i** between the components. These are usually included in continuation classes. The linking element **O** is realized as **o** or **e**.

Word forms that are permitted as second parts have the following diacritics:

Q1+F1: qualitative adjective (long forms and short forms)
Q2+F1: relative adjective (long forms and short forms)
Q3: noun
Q4: present participle active
Q5: past participle passive

Most of the second parts can be used as independent words. Compounds that are listed in the lexicon have the diacritics mentioned above, too. In this way, the mechanism also permits compounds, consisting of more than two parts. The first parts have a diacritic C1, C2, C3, C4, C5, B1, B2 or B3. They permit the second parts of the following types:

C1: Q2+F1

C2: Q1+F1
C3: Q2+F1, Q3
C4: Q2+F1, Q3, Q4
C5: Q1+F1, Q2+F1
B1: Q3
B2: Q4
B3: Q5

The correct combination of the parts is controlled by rules (40)-(48).

In order to treat compounds, the vocabulary is split up into four main lexicons. Stem1 is the largest one. It contains most nouns, adjectives, verbs, and derivated adverbs and predicatives. In RUSTWOL, the most productive nouns and adjectives can occur as first parts of compounds. The following examples of possible combinations are presented in surface forms, except for #, which is a sign of word boundary:

REL A + REL A	motorno#-parusnyj	(C1 + Q2+F1)
QUAL A + QUAL A	barhatisto#-mohnatyj	(C2 + Q1+F1)
N + N	stroj#bank	(C3 + Q3)
N + PRES PART ACT	gazo#obrazuu1s2ij	(C4 + Q4)
QUAL A + REL A	geroic1eski#-nezemnoj	(C5 + Q2+F1)
N + N	kvartiro#sdatc1ik	(B1 + Q3)
N + PRES PART ACT	luc1e#ispuskau1s2ij	(B2 + Q4)
N + PAST PART PASS	gazo#zas2is2ennyj	(B3 + Q5)

Some of the first parts in Stem1 cannot be used as independent words, for example the following:

ae1ro#s1kola	(C3 + Q3)
gamma#-kvantovyj	(C4 + Q2+F1)

Lexicon Stem2 includes color adjectives. On the one hand, when two color adjectives are combined, there must be a linking element and a hyphen between the parts. On the other hand, only a linking element is needed, when color adjectives are connected to nouns or adjectives in Stem1.

Lexicon Stem3 contains, firstly, pronouns, numerals, proper nouns, abbreviations and non-inflecting parts of speech that are not formed in the declension types of adjectives. In addition, some nouns, adjectives and verbs that are not partaking in productive compound formation are included in this lexicon. Only some pronouns and numerals can occur as first parts of compounds. Lexicon Stem4 contains only numbers 0...9. The continuation classes of numbers account for words like 0, 125, 2.2, 334, 5, 1997-2000, 50-letie. They also include a continuation to Stem1.

15.4 Final Remarks

The most difficult problem that I faced in developing the first version of the RUSTWOL lexicon and rules was the problem of handling compounds, both parts of which are inflected. This problem is not discussed by me in the document Vilkki 1997. Russian has a fairly productive means of forming compounds by inflecting the both parts in the same case and number. Most of these are nouns, but it is also possible to form compound relative adjectives using this kind of compounding. For example, the dictionary of Russian compounds Bukčina and Kalakučkaja 1987 lists 82,000 compounds, and approximately 5,800 these represent compounds, both parts of which are inflected. Here are some examples of these kinds of compounds in the genitive case:

pisatela1-gumanista 'writer-humanist'
funkcii-kriterija 'function-criterion'
z1ens2iny-uc1enogo (sekretara1) 'woman-scientific (secretary)'

Besides genitive, this kind of inflection concerns all the other singular and plural cases. Because it was difficult to find any appropriate way to handle these kinds of compounds adequately, they were totally excluded from the lexicon.

At a more general level, Koskenniemi (1983) understood that his initial two-level model had significant limitations in handling various kinds of non-concatenative morphotactic processes. Several kinds of non-concatenative phenomena are considered in, for example, Beesley and Karttunen (2003:375-420). They rightly state that non-concatenative morphotactis is the cutting edge of computational morphology. I would like to emphasize, however, that the version of RUSTWOL presented here is not the current one. As far as I know, the current RUSTWOL at Lingsoft has, on the whole, a more adequate system of forming compounds. This newer version represents a new kind of format the practical implementation of which is based on suggestions of Koskenniemi.

Appendix

This appendix gives Cyrillic translations of the alphabet used in RUSTWOL. It also lists all the tags contained in RUSTWOL.

Alphabet:

a	a1	b	c	c1	d	e	e1	f	g	h	i	j	k	l
а	я	б	ц	ч	д	е	э	ф	г	х	и	й	к	л

m	n	o	p	r	s	s1	s2	t	u	u1	v	y	z	z1
м	н	о	п	р	с	ш	щ	т	у	ю	в	ы	з	ж

‘	$
ь	ъ

RUSTWOL tag set:

ACC	accusative
ACT	active
ADV	adverb
ADV-CMP	comparative form of adverb
CARD	cardinal number
CMP	comparative
COLL	collective
COMP	compound
CONJ	conjunction
CONST	constituent
DAT	dative
DEF	definite
DEM	demonstrative
FE	feminine
FUT	future
GEN	genitive
IMPERS	impersonal
IMPF	imperfective
IMPV	imperative
INF	infinitive
INDECL	indeclinable
INDEF	indefinite
INSTR	instrumental
INTERJ	interjection
INTERR	interrogative
MA	masculine
N	noun
NOM	nominative
NE	neuter
NEG	negative
NUM	numeral
ORD	ordinal number
PARENTH	parenthetical
PART	participle
PASS	passive
PAST	past tense (preterite)
PCLE	particle
PERF	perfective
PERS	personal
PL	plural
PL1	1st person, plural
PL2	2nd person, plural
PL3	3rd person, plural
POSS	possessive
PRED	predicative
PREP	prepositional
PRES	present tense
PRON	pronoun
REFL	reflexive
SG	singular
SG1	1st person, singular
SG2	2nd person, singular
SG3	3rd person, singular
SH FE	short feminine
SH MA	short masculine
SH NE	short neuter
SH PL	short plural
SUP	superlative
V	verb
V ADV	verbal adverb

Acknowledgements

I wish to thank Kimmo Koskenniemi, Fred Karlsson, Arto Mustajoki and Jouko Lindstedt for the helpful advice on various aspects of the first version of RUSTWOL.

References

Beesley, K., Karttunen, L. 2003. *Finite State Morphology*. Stanford: CSLI Publications.

Bukčina, B.Z., Kalakučkaja, L.P. 1987. *Slitno ili razdel'no*. Moscow: Russkij jazyk.

Jevgen'eva, A.P. (ed.) 1981-1984. *Slovar' russkogo jazyka i-iv*. 2nd ed. Moscow: Russkij jazyk

Kahla, M. (ed.) 1984. *Neuvostoliittolaisten henkilönnimien opas*. Helsinki: Valtion painatuskeskus.

Kahla, M. (ed.) 1982. *Neuvostoliiton paikannimet*. Helsinki: Valtion painatuskeskus.

Kopotev, M, Mustajoki, A. 2003. Principy sozdanija Hel'sinkskogo annotirovannogo korpusa russkih tekstov (HANKO) v seti internet. *Naučno-tehničeskaja informacija*, ser. 2, No. 6: 33-37-

Koskenniemi, K. 1983. *Two-level Morphology: A General Computational Model for Word-form Recognition and Production*. University of Helsinki: Publications of the Department of General Linguistics. No. 11.

Koskenniemi, K.1997. Representations and Finite-State Components in Natural Language. *Finite-State Language Processing*, ed. E. Roche, Y. Schabes, 99-116. Cambridge: The MIT Press.

Kotelova, I.Z. (ed.) 1984. *Novye slova i značenija: slovar'-spravočnik po materialam pressy i literatury 70-h godov*. Moscow: Russkij jazyk.

Kuznecova, A.I., Efremova, T.F. (ed.) 1986. *Slovar' morfem russkogo jazyka*. Moscow: Russkij jazyk.

Paile, Alexander 2003. *Avtomaticeskij analiz russkogo teksta*. Master thesis, University of Helsinki.

Scheitz, E. 1986. *Dictionary of Russian Abbreviations*. Berlin: Veb Verlag Technik.

Švedova, N.J. (ed.) 1982. *Russkaja grammatika i-ii*. Moscow: Nauka.

Tihonov. A.N. 1985. *Slovoobrazovatel'nyj slovar' russkogo jazyka*. Moscow: Russkij jazyk.

Vilkki, Liisa 1997. *RUSTWOL: a System for Automatic Recognition of Russian words*. A technical document, Lingsoft.

Zaliznjak, A.A. 1987. *Grammatičeskij slovar' russkogo jazyka*. 3rd ed. Moscow: Russkij jazyk.

Zasorina, L.N. (ed.) 1977. *Častotnyj slovar' russkogo jazyka*. Moscow: Russkij jazyk.