

Adaptation of PolyLex-Programm for Russian

Sebastian Nagel

August 2003

Abstract

The PolyLex programm, designed for analysis of Norwegian and German compounds, has been adapted to Russian. Composition plays a less important role in Russian compared to languages like German. Main motivation for adapting PolyLex has been the fact, that in Russian there are composition-like word formation models which are (1.) located between flexion and derivation and are difficult to treat as either one of them. (2.) Some models produce theoretically infinitely many words. The basic principle when adapting PolyLex has been to separate programm code and linguistic knowledge. So rules for decomposition are stored in a “normal” dictionary in Unitex format. The programm itself checks whether the rules match a given word, and decomposes the word.

‘Word-formation’ (‘словообразование’) includes two models: ‘composition’ (‘слово-сложение’) and ‘derivation’ (‘словопроизводство’).

1 Composition

Composition as a word formation model isn’t that important in Russian as in other languages like German. Most German compound nouns will be expressed in Russian by combination of an adjective and a noun:

- (1) Eisenbahn (‘railway’) → железная дорога

Often there is a distinction between ‘qualitative’ (2a) and ‘relational’ (2b) adjectives supporting this multilexemic word formation model:

- (2) a. дымная комната ‘smoky / smoke-filled room’
b. дымовая труба ‘smokestack’

Anyway, there are compounds in Russian. So the related adjective of *железная дорога* (‘railway’) is *железнодорожный*, a clear case of a compound. But there are other models of composition, too. Composition is here defined as simply putting words or morphemes together with changing them minimally by morphological rules.

1.1 Problematic cases: inflection vs. derivation

There are some problematic cases which are rather systematic and flexion-like but difficult to treat as inflection. Percov (2001: 90) calls them ‘quasi-grammeme’s’:

1. the comparative of adjectives and adverbs
2. adverbial forms of adjectives (*геройски, смело*)
3. the so called ‘attenuative’ comparative (*посмелее*)
4. exhortative (*пойдемте*)
5. negation by prefix *не*: participles (*неустаревший*), but also adjectives (*неэффективный*), nouns (*непедагог*), or pronouns (*ненаш*)

Common to all these forms is that they are written in one word. Two of these cases (exhortative and comparative) are treated as inflection in CISLEX-RU. Adverbial forms of adjectives are no task for PolyLex because they are homonym to short forms.

The rules for decomposition are quite simple. When you have a word beginning with *no* and the remaining part is a comparative then change the tag of the word to <KOMP+atten>. Rules will be written further as:

по <A:k> ⇒ <KOMP+atten>
 <KOMP-atten> ⇒

The rules for negated forms and words are similar:

не <V:E:Q:R:S:T:Z> ⇒ .+neg
 <N> ⇒
 <A> ⇒
 <V:I:G:V:b:H:J> ⇒
 <ADV> ⇒
 <PRFX-neg> ⇒

That is the prefix *не* can be combined with almost all word classes. At present only the feature +neg is added, especially for participles the decomposition in two words might be better. Participles can always be combined with the prefix *не*, there is no semantic restriction. Only the absence of the verbal action is expressed, nothing more. This doesn’t apply for other word classes.

According to Russian orthographic rules *не* is written separately only when an opposition is expressed, as in *не ... а ...* (‘not ...but ...’), *отнюдь не ...* (‘by far not ...’). It is a problem of convention and peculiar to written language. Because the adverbial *не* is clitic there will be no distinction in pronunciation. Because most Tokenizers split the text into tokens by spaces words prefixed by *не* are still problematic.

In other languages like English and German this problem is masked by the fact that prefix and adverb of negation are different. The semantic restrictions stay similar:

- (3) a. unforgotten / unforgettable ↔ not forgetful (*unforgetful)

- b. unvergessen / unvergesslich ↔ nicht vergesslich

Other problematic cases in this category are suffix particles like *-ка*, *-то*, etc. – not treated yet because of the special and ambiguous state of the hyphen sign which can be interpreted as a letter or an interpunctuation character.

- (4) a. пойд^и-ка
b. барыш^н-я-то

1.2 Numeral prefixes

This word formation model is much more frequent in Russian than in German. In English it is rarely used:

- (5) a. двухэтажный – zweistöckig – two storied
b. двухметровый – [?]zweimetrig – two meter long
c. тридцатимиллиграммовый – dreißig Milligramm schwer – of thirty milligrams

Actually about 300 suffixes participating in this model of composition are gathered – in a list of five million unknown words out of a big russian corpus and russian web sides 16 000 words have been recognised. Potentially there are unlimited many words, because any number can be taken as a prefix.¹

First task is to describe the composition of number elements to numeral prefixes. The order is the same as the order of complex numerals. Some auxiliary features are introduced for prefixes:

Num1	= 1-9
Num10	= 10-19
Num20	= 20-90
Num100	= 100-900
Num1000	= 1000
NumMio	= 1000000
NumMia	= 1000000000
NumBig	= 1000 or more
NumUn	= indifferent numbers: <i>много, несколько</i>

Possible combinations are:

¹According to Mel'čuk (1995: 490, 503) compositions with the numerals thousand or higher are ungrammatically, but are accepted by some speakers. In "real" text they are observed frequently.

$\langle \text{Num} \rangle ::= \dots [\langle 1000 \rangle] [\text{Num}100] [[\text{Num}20] [\text{Num}1] | [\text{Num}10]]$
(one element must be given anyway!)
 $\langle 1000 \rangle ::= [\text{Num}100] [[\text{Num}20] [\text{Num}1] | [\text{Num}10]] \text{Num}1000$
(analogous for millions, billions, etc.)

Because word formation rules as described below (chapter 3) affect only the preceding and following element, many rules have to be formulated. Suffixes as word formants are listed separately. They are mainly semantically determined and don't coincide with word classes like parts of speech. And they must not be autonomous words – there is no word **амперный*. Most of the suffixes are adjectives, but there are some nouns (*двухэтажка* ‘two storied building’).

1.3 Common prefixes

This model describes composition of a “normal” word and a modifying prefix like *анти* (‘anti’), *экстра* (‘extra’, *вице-* (‘vice’) and so one. For each prefix it is defined with which word classes it can be combined.

Problems arise for very short prefixes like the alpha privativum *a* in *абиологический* (‘abiological’) which in “real” texts are often spelling errors. Such prefixes are removed from the lexicon. Another frequent error caused by spelling mistakes are prepositions and nouns written in one word (many prepositions are also prefixes). So *междулюдьми* is hardly a form of **междучеловек* (‘middle-man’), but a misspelled *между людьми* (‘among people’). When the resulting decomposition lists are stored as a “lexicon” it seems useful to filter decompositions which are valid combinations of preposition and noun.

1.4 “Real” composition

“Real” composition is defined here as composition of autonomous words. Because not all Russian words or all members of a class (like German nouns) are suitable, there must be lists of composable forms, which can be handled the same way as prefixes (see chapter 1.3).

[Not implemented yet.]

2 Derivation: suffix driven decomposition

Derivation is the most important model in Russian word formation. The difference to the models described above is that not prefixation but suffixation is involved. As substring operations on prefixes and suffixes of words are already part of the syntax of rules, it should be rather simple to integrate derivation to PolyLex. For example if a word has the suffix *ость*, clip it off, append *ый* instead, and make a lexicon look up. If the resulting form is a adjective or a participle in the lexicon, the word will be recognized as a noun. So *балансированность* is o.k., if *балансированный* is already in the lexicon.

The main feature for such a suffix driven decomposition is, that analysis must start from the end of a word. That implies reorganisation of the lexicon. Derivational suffixes must be stored

as a reversed trie, i.e. starting from the end of a word. There are two possibilities: (1.) When having the whole lexicon as reversed trie, no changes to the algorithm are necessary. Only the word to be analysed has to be reversed, too. (2.) When changing the algorithm, that is introducing a new function for derivational decomposition, it should be possible to have only the derivational suffix as reversed forms in the lexicon. Analysis starts on the reversed word, after the suffix has been changed, the resulting word will be looked up in “normal” order.

Proposed example:

ЬСТВО, .RSFX+DR(#<A:neM>=!-0йы; -0ьсто\ .<N\+anim(j)\+gen(F):neF:aeF>)

[Not implemented yet!]

3 Syntax of rules for PolyLex

The syntax of a word-formation rule is defined as:

<rule>	::=	<right context>#<left context>=<then>
<context>	::=	<Unitex-meta>
<then>	::=	[<substring-ops>][\ . <meta-manip>]
<substring-ops>	::=	<substring-op (actual element)>[! <substring-op (following element)>]
<substring-op>	::=	<substr-op>[; <substr-op (undo)>]
<substr-op>	::=	<prefix-op> <suffix-op>
<prefix-op>	::=	- <num><prefix>
<suffix-op>	::=	<num><suffix>
<meta-manip>	::=	<replace> <add>
<replace>	::=	<<Unitex-meta>>
<add>	::=	\ . <codes>

!!! If in the fields <substring-op> or <replace> two operations are given (by preceding and actual element), only the operation given by the actual element is applied. If <replace> is given, <add> is ignored. !!!

Let’s look on a example, say the decomposition of *безытерационного* (‘without iteration’). In the lexicon there are two entries:

безы, .PRFX+DR(#<A-Pron>=1!-0и; -1ы) итерационного, итерационный. A:geM:geN:aeM

After recognizing *безы* as prefix, the rule #<A-Pron>=1!-0и; -1ы is extracted and parsed. <A-Pron> is stored as condition to be fulfilled (matched) by the lexical information of the next segment. Then the substring operations are applied. One letter is removed from the end of the prefix *безы*, the resulting *без* will be pushed on the decomposition list. A lookup for the remaining suffix *итерационного* would be unsuccessful, so first the substring operation for the following element must be applied. The minus sign ‘-’ marks the operation as a prefix one. So *и* is added in front of *итерационного*. The resulting *итерационного* is found in the lexicon, it is of a adjective (<A>) without the grammatical feature +Pron, so the rule matches. But before pushing the decomposed word to the decomposition lexicon, the change *ы/u* has

to be undone for the lemma found in lexicon. This is described by -1ы. The meta-information isn't changed – there is no rule for this –, the codes from *итерационного* are taken. The resulting dictionary entry is:

безытерационного,безытерационный.А:геМ:геN:аоеМ

Let's look at two examples for manipulation of meta-informations, i.e. grammatical and flectional codes:

по,.PRFX+DR(<#<A\ :k>=\.<КОМП\+atten>) + глуже,глубокий.А:k ⇒
 поглуже,поглуже.КОМП+atten
 не,.PRFX+neg+DR(<#<A>=\.\+neg) + нашему,наш.А+Pron:deM:deN ⇒
 ненашему,ненаш.А+Pron+neg:deM:deN

In the first case the codes after ‘.’ stand between <>, so they will replace the codes from the last element of decomposition. Note that also the lemmatization of the word form is cancelled. In the second example the feature +neg isn't included in <>, what means it will be added to the information from the last element.

!!! Because of having a special semantics for Unitex ‘+’ and ‘.’ must be escaped by ‘\’ in all rules. !!!

4 Algorithm of decomposition

“explore” (“word”, “decomposition list”, “rule preceding element”, “dictionary entry preceding element”)

- trie each prefix the word:
 - if the prefix is a valid word in the lexicon
 - get dictionary information
 - extract word-formation rules
 - foreach rule
 - if the prefix is a valid prefix and the rule of the preceding element (calling rule) matches the actual dictionary entry and the actual rule matches the entry of preceding element (empty rules match always)
 - ⇒ push actual element (prefix) to “decomposition list”
 - ⇒ call “explore” (“remaining suffix”, “decomposition list”, “actual rule”, “actual dictionary entry”)
 - if the end of the word is reached and rules and dictionary entries matche (see above)
 - ⇒ word recognised
 - ⇒ generate new lexicon entry by building word, lemma and lexical information out of “decomposition list”, preceding and actual rules, actual dictionary entry

5 Literature

Andrews, Edna

- 1996 *The semantics of suffixation*. LINCOS Studies in Slavic Linguistics 5. München, Newcastle, LINCOS EUROPA.

Efremova, Tat'jana Fedorovna

- 1996 *Tolkovyj slovar' slovoobrazovatel'nyh edinic russkogo jazyka*. Moskva, Russkij Jazyk.

Hippisley, Andrew

- 1996 Productive Russian nominal lexeme formation. Paper

Janko-Trinickaja, Nadija Aleksandrovna

- 2001 *Slovoobrazovanie v sovremennom russkom jazyke*. Moskva, Indrik.

Krylov, S. A.

- 2000 Avtomatičeskij morfoložičeskij analiz russkich slovoform s prefiksali'nyh otricaniem: neskol'ko teoretičeskich problem. *Materialy DIALOGA 2000 2* (prikladnye problemy), 220-225. Protvino, Izdatel'stvo RGGU. http://www.dialog-21.ru/archive_article.asp?param=6510&y=2000&vol=6078 (Stand Juli 2003)

KUZNECOVA&EFREMOVA

- Kuznecova, Araida Ivanovna; Tat'jana Fedorovna Efremova 1986: *Slovar' morfem russkogo jazyka*. Moskva.

Mel'čuk, Igor [=Mel'čuk, Igor' Aleksandrovič]

- 1995 *Russkij jazyk v modeli SMYSL ⇔ TEKST*. Wiener Slavistischer Almanach Sonderband 39. Moskva.

Paumier Sebastián

- 2002 Manuel d'utilisation d'Unitex. <http://www-igm.univ-mlv.fr/unitex/>
2003 *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Thèse de doctorat en informatique, Université de Marne-la-Vallée.

Percova, N.

- 1996 RUSLO: An Automatic System for Derivation in Russian. In: Wanner, Leo (ed.): *Lexical Functions in Lexicography and Natural Language Processing*. Studies in Language Companion Series (SLCS) 31, 307-318. Amsterdam, Philadelphia, John Benjamins Publishing Company.

Rafaeva A.V.

- 2000 Avtomatičeskaja sistema russkogo slovoobrazovanija RUSLO 2. *Materialy DIALOGA 2000 2* (prikladnye problemy). http://www.dialog-21.ru/archive_article.asp?param=6537&y=2000&vol=6078

RDD

- Worth, Dean S.; Andrew S. Kozak; Donald B. Johnson 1970: *Russian derivational dictionary*. New York.

Tichonov, Aleksandr Nikolaevič

- 1985 *Slovoobrazovatel'nyj slovar' russkogo jazyka*. Moskva.
1996 *Morfemno-orfografičeskij slovar' Russkaja Morfemika*. Moskva, Škola-Press.

Townsend, Charles E.

- 1980 [¹1968] *Russian word-formation*. Chelsea (Michigan), Slavica Publishers.