RUSLO: An Automatic System for Derivation in Russian

Natalia N. Percova
Moscow state university

The paper describes a linguistic apparatus for a system of Russian derivational analysis and synthesis. Its semantic component uses a list of derivational meanings, which partially coincide with lexical functions from the Meaning<->Text model developed by I. A. Mel'cuk.

Automatic system for Russian derivation

The derivational system RUSLO (_RUsskoe SLOvoobrazovanie_ 'Russian Derivation') can carry out the following tasks:

(1) give derivatives for an input word;
(2) give a source word for an input word;
(3) give a derivational meaning of a derivative (mainly in terms of lexical functions);
(4) generate - for an input word - a derivative with a given meaning.

The system is implemented in the programming language PASCAL and runs on IBM compatible PC's (the programs are written by A. V. Cheremxin). It uses a machine version of A. A. Zaliznjak's "Grammatical Dictionary of Russian" [10] (about 100000 entries) as a sourse of "conventional" Russian words. RUSLO can analyse both conventional and new words (such as neologisms or slang words from [9]); in synthesis only conventional words are allowed as an output.

In the following we give some examples of RUSLO's work (English translations of output words are added by us). "Ancestors" denote words of the formally previous level of derivation with respect to an input word; "brothers" are words of the same level of derivation; and "descendants", finally, are words of the next higher level of derivation.

(1) Task: Give derivatives for _staryj_ 'old'
SYSTEM OUTPUT:
Brothers: _staret'_ 'to become old(er)', _starit'_ 'to make old(er)', _starit'sja_ 'to become old(er)', _star'_ 'old times'
Descendants: _star'jo_ 'old stuff', _starushka_ 'an old woman', _staruxa_ 'an old woman', _starost'_ 'old age', _staroe_ 'the past', _starka_ 'mature vodka', _starica_ 'an old river bed', _starinka_ 'old times', _starina_ 'old times/an old person', _starik_ 'an old man', _starec_ 'an old (and wise) man', _starenie_ 'growing old(er)', _starshij_ 'elder/older', _starcheskij_ 'like in old age', _starovatyj_ 'a little bit old', _staren'kij_ 'diminutive for old'
[Along with these words the system gives two error words: _stara-tel'nyj_ 'careful' and _staranie_ 'a try' because of formal

coincidence]

(2) Task: Find a source word for _rogov'jo_ 'collective for horn'

SYSTEM OUTPUT:

"No word in the dictionary"

Ancestor(s): _rog_ 'a horn', _rogovet'_ 'to become like horn'

(3) Task: Find Sing(_sneg_ 'snow')

SYSTEM OUTPUT:

Sing(_sneg_)=_snezhinka_ 'a snow-flake'

[the system generates two possible variants: _snezhina_ and _snezhinka_, and looks for them in the dictionary [10], where it finds only the second one]

(4) Task: Give derivational meanings for _ximichka_ 'a woman-chemist', _plutovka_ 'a woman-crook', _bezhenka_ 'a woman-refugee', _princessa_ 'a princess'

SYSTEM OUTPUT:

ximichka=Fem(_ximik_ 'a chemist')
plutovka=Fem(_plut_ 'a crook')
bezhenka=Fem(_bezhenec_ 'a refugee')
princessa=Fem(_princ_ 'a prince')

Along with short answers (as in examples given above), the system can describe an input word in more detail, using the following format:

- an input word;
- information from the grammatical dictionary [10] (in case the grammatical dictionary contains this word; otherwise - the message that the word is absent);
- its morph structure;
- its morpheme structure;
- its derivational structure, described with the help of brackets;
- its derivational semantics, described in terms of lexical functions;
- its place within the derivational frame: words of the previous, the same, and the following levels of derivations (its "ancestors", "brothers" and "descendants").

To carry out the described tasks the system uses linguistic information of two basic types: formal and semantic.

Formal information

One of the main problems concerning Russian derivation is the fuziness of morph boundaries. This has already been noticed by the Russian-Polish linguist I. A. Baudoin de Courtanay many years ago. Thus, Baudoin agued [1, pp. 232-233] that morph boundaries can be more or less fuzzy: in many cases different people, or even the same person in different moments, divide the same word into morphs in different ways. This turns out to be true not only for native speakers of Russian, but for researchers as well. For instance, in modern Russian Academic

grammars [2, 3, 8] and in the dictionary of Russian morphemes [5] we find cases of (i) different morph division by different authors as well as (ii) cases of different morph division by the same authors:

(i)

- _kras-av-ec_ 'a handsome man' [5] and _kras-avec_ [2, 3, 8],
- _molch-ank(a)_ 'keeping silence' [5]я_ я.and molcha-nk(a)_ [3, 8],
- _vix-lja(j)_ 'an unbalanced person' [5] and _vixl-ja(j)_ [3, 8],
- _det-v-or(a)_ 'children' [5] and _det-vor(a)_ [8];

(ii)

- _xod-ata(j)_ 'an intercessor' and _xod-at-a(j)_ [5],
- _vid-imost(')_ 'visibility' and _rani-most(')_ 'subtlety' [2],
- _aplodism-ent(y)_ 'applause', _medikam-ent_ 'medicaments' [3] and _aplodis-ment(y)_, _medik-ament_ [8] (the grammers [3] and [8] are written by the same authers).

Thus, as it can be seen from the examples above, the problem is solved by three ways: the questionable intermediate part is treated as a part of a stem (as in _molcha-nk(a)_), as a part of an affix (as in _molch-ank(a)_, _kras-avec_), or as a separate unit sometimes called "interfix" (as in _kras-av-ec_).

In our system, we treat this problem by introducing different types of morhp division markers. It helps us to minimize the number of derivational morphs (the total list of noun derivational suffixes from [2, 3, 5, 8] contains about 1000 units, our suffix list is half as much). We use several types of morph division markers (described in [7]): some of them represent definite boundaries, while others - fuzzy boundaries; in the latter case the intermediate part is considered flexible and can be treated as belonging either to an affix, or to a stem: _vz*n-uzd-a(t')_=_vz-n*uzda(t')_'to bridle'; _kongo*lez-sk(ij)_= _kongo-lez*sk(ij)_ 'kongoleze'; _kruzhk*ov-ec_=_kruzhk-ov*ec_ 'a society member', etc. Formal derivational information includes lists of affixes (prefixes, suffixes and circumfixes), patterns for derivatives and compounds, and some additional rules and restrictions. Among them we shall mention rules of generating those forms of stems which do not coincide with stems in dictionary entries, for example,

- stems used in indirect cases of nouns: _vesn(a)_ 'spring' - _vesenn(ij)_ 'adjecive to spring';
- finite verbal forms: _my(t')_ 'to wash' - _mojk(a)_ 'a wash';
- stems obtained by regular and irregular consonants alternations: _k/ch_ (_ruka_ 'a hand' - _ruchka_ 'a small hand'), etc.

Additional restrictions include constraints on grammatical characteristics of stem words and derivatives; for instance, stem nouns and derivatives with deminutive suffixes usually have the same gender, and therefore the system is allowed to relate _pirog_ 'a pie' -ая_pirozh-ok_ 'a small pie' (which are both masculine), but not _ven(a)_ 'a vein' - _ven-ok_ 'a wreath' (the first noun is feminine, and the second one is masculine), etc.

Different types of derivational information are represented

in a relational data base. (The formal linguistic apparatus for RUSLO is described in more detail in [7].)

Semantic information

To represent the meaning of derivational relations we use a modified version of the system of lexical functions (LF) developed within the Meaning<->Text theory [6, pp. 51-61]. LFs that have been designed for describing syntagmatic and paradigmatic semantic relations among words turn out to be a convenient means for describing regular derivational meanings in Russian.

For this purpose the following LFs can be applied (the following examples are from [6]; merely those in square brackets are added):

LF	Key lexeme	LF value
Syn	_doscheka_ 'a plank'	_dostochka_
Conv:	_varit'_ 'to cook'	_varit'sja_
Anti:	_drugI.1_ 'a friend'	_nedrug_
Dimin:	_brus_ 'a beam'	_brusok1_
Augm:	[_nos_ 'a nose'	_nosische_]
S-0:	_borot'sja1/2_ 'to struggle'	_bor'ba1/2a_
A-0:	_agressija_ 'agression'	_agressivnyj1a_
Adv-0:	_lixoradit'2_ 'to agitate'	_lixoradochno2_
V-0:	_obvinenie1/2_ 'prosecution'	_obvinjat'1/2a_
S-i (=1):	_l'stit'I.1_ 'to flatter'	_l'stec_
S-instr:	[_vykljuchat'_ 'to switch off'	_vykljuchatel'_]
S-med:	[_polirovat'_ 'to polish'	_polirol'_]
S-mod:	[_xodit' 'to walk'	_poxodka_]
S-loc:	_obrezat'I.1a_ 'to cut off'	_obrez1_
S-res:	_vyrezat'1a_ 'to cut out'	_vyrezka2_ / _vyrezka3_
Sing:	_streljat'1.1a 'to fire'_	_vystrelit'1 _ _kirpich1b_ 'bricks' _kirpich1a_ / _kirpichina_
Mult:	_vystrel1a_ 'a shot'	_strel'ba_
A-i (=2):	_vlast'1_ 'power"	-podvlastnyj_
Able-i (=1):	_gnevat'sja_ 'to be angry'	_gnevlivyj_
Magn:	_druzhestvennyjI_ 'amicable'	_druzheskijI_ 'friendly'
Ver:	_zharit'1 _ 'to sting'	_uzharivat'1_
Bon		
+AntiMagn:	_veter_ 'a wind'	_veterok_
Adv-i (=1):	_dosada_ 'annoyance'	_dosadno_
Func-0:	_dozhd'1_ 'a rain'	_dozhdit'1_
Incep:	_znobit'_ 'to feel feverish'	_zaznobit'_
Caus:	_zhenit'sja_ 'to marry'	_zhenit'_
Fin:	_kurit'1_ 'to smoke'	_dokurit'1_
Real-i (=1):	_lechit'1a_ 'to treat'	_vylechivat'_ / _izlechivat'_

Perf: _arestovyvat'_ 'to arrest' _arestovat'_
Imperf: _vr\ezat'1_ 'to cut into' _vrez\at'1_
(the accented vowel is depicted by "\")

The question arises whether this list of LFs is sufficient to describe derivational meanings in Russian. To establish a set of Russian derivatemes (=regular derivational meanings) it seems useful to introduce the following features for their classification.

1. Retainment (-) vs. change (+) of the part of speech (of the source word).
2. Retainement (-) vs. change (+) of semantics (of the source word).
3. Absence (-) vs. presence (+) of a verbal predicate different from the source word within the definition of the derivative word (= "implied predicate").

Then we receive 8 combinations of feature values:

	1	2	3		Examples	
(a)	-	-		-я	_motocikl - motociklet - motocikletka_ 'motor-cycle - motor-cycle - motor-cycle'	
(b)	+	-		-я	_igrat' - igra_ 'to play - a play'	
(c)	-	+		-	_dom - domik_; 'a house - a small house' -seryj - serovatyj_ 'grey - greyish'	
(d)	-	-		+я	IMPOSSIBLE	
(e)	+	+		-я	_prygat'- pryzhok_ 'to jump - a jump' _vykljuchat'- vykljuchatel'_ 'to switch off - a switch'	
(f)	-	+		+я	_kanat - kanatchik_ 'a cable - a worker manufactoring cable'	
(g)	+	-		+я	IMPOSSIBLE	
(h)	+	+		+	_utjug - utjuzhit'_ 'an iron - to iron'	

Two of these combinations - (d) and (g) яare logically impossible, since the presence of the implied predicate entails change of semantics (i.e. if 2 = "+", then 3 = "+"). All the other combinations of features are possible. For further examination of them we shall regard them separately for different possible pairs of parts of speech. In Russian, in "sourse word - derivative" pairs, the source word belongs to one of the following six parts of speech: a noun (N), an adjective (A), an adverb (Adv), a verb (V), a numeral (Num), and an interjection (Int); and the derivative - to four: N, A, Adv, or V.

In the following we list some Russian derivatemes (the

total number of them appears to be several hundreds) for the six possible classes of feature combinations we have introduced above. Within each class derivatemes are ordered in pairs <I, J>, where I is the part of speech of source words, and J - of derivatives.

(a) In this class a source word and a derivative have the same meaning. It is a marginal derivational situation, which occurs mainly in pairs <N, N> (ch. _lisa - lisica_ 'a fox'). Its meaning is described by the single LF - Syn.

(b) If we use Jerzy Kurylovicz's distinction of syntactic vs. lexical derivation (i.e. derivation changing only the part of speech vs. effecting the meaning) [4], this class will be the only case of syntactic derivation. In the system of LFs such derivatives are described by LFs with zero indexes: S-0 for pairs <V, N> (ch. _xodit' - xozhdenie_ 'to walk - a walk') and <A, N> (ch. _sinij - sineva_ 'blue - the blue'); A-0 for <N, A> (ch. _kirpich - kirpichnyj_ 'a brick - brick') and <V, A> (ch. _smjat' - smjatyj_ 'to crumple - crumpled'); Adv-0 for <A, Adv>я(ch. _xolodnyj - xolodno_ 'cold - coldly'); V-0 for <N, V>я(ch. _sneg - snezhit'_ 'snow - to snow') and <Int, V>я(ch. _oj - ojkat'_ 'oh - to produce "oh"').

(c) This class of derivatemes, which can be called modificators, is described by LFs only partially. The subclasses are: <N, N>, <A, A>, <Adv, Adv>, and <V, V>.

For this class the following LFs are relevant: Dimin, Augm, Anti, Sing, Mult, Magn, Plus, Minus, Bon, AntiBon, Ver, Incep, Fin, Real-i, Caus, Perf, Imperf, Plus, Minus (see examples above, in the list of lexical functions).

To these derivatemes described by LFs there must be added some new ones (the cretarion for introducing of a new Lf is repetition of derivational meaning for different derivatives and/or formal derivational means), for instance the following.

- <N, N>: Fem (for 'feminine correlate'): Fem(_uchitel'_ 'a teacher')=_uchitel'nica_; Desc (for 'descendants'): Desc(_lev_ 'a lion')=_l'vjonok_; DescFem(_car'_ 'tchar')=_carevna_; besides, there is a broad range of stylistic and emotional modificators (ch. _parnjaga_ 'a fellow', _starikashka_ and _starikan_ 'an old man', _uchilka_ 'a teacher' - all the words with additional stylistic and emotional meanings);

- <A, A> and <Adv, Adv>: Atten (for 'attenuative'): Atten(_grjaznyj_ 'dirty')=_grjaznovatyj_, Atten(_xolodno_ 'cold-ly')=_xolodnovato_;

- <V, V>: the variety of meanings of Russian verbal prefixes can be reflected by means of some new derivatemes with such meanings as 'for some time' (_po-_, as in _po-begat'_ 'run for some time'), 'too much' (_pere-_, as in _pere-xvalit'_ 'over-praise), etc.

(e) The derivatemes for pairs <V, N> are partially described by the LFs S-i, S-instr, S-med, S-mod, S-loc and S-res; and for pairs <V, A> - by A-i and Able-i. In some cases a derivateme can be represented by a combination of LFs: _pryzhok_= SingS-0(_prygat'_ 'to jump'). This class also

includes some other derivational pairs, such as <A, N>: _staryj - star'jo_ 'old - old stuff (about things or people)', _molodoj - molodjozh'_ 'young - young people', _molodoj - molodnjak_ 'young - young animals'. To describe their semantics, we can use the LF Mult plus an implied object, such as 'Person' for _molodjozh_, 'Animal' for _molodnjak_, etc.

(f) and (h) These two classes are characterized by the presence of an implied predicate in the definition of a derivateme. It seems that the class (f) contains only one pair - <N, N> - with a number of interpretations, such as 'person involved in the sphere N / connected with an object N / having a ("non-normal") part N': _gipertonia - gipertonik_ 'hyper-tension - a hypertension patient' / _kanat - kanatchik_ 'a ca-ble - a worker manufacturing cable'я/ _serdce - serdechnik_ 'a heart - a heart patient'; 'place or an object containing N(s)': _korova - korovnik_ 'a cow - a cow-house',яя_berjoza - bereznjak_ 'a birch - a birch grove', _kofe - kofejnik_ 'coffee - a coffee-pot'; 'place near N': _more - pomor'e_ 'a sea -a beach'. Such types of derivatemes are not described by the LFs from the canonical list. Nor can LFs describe case (h), which includes such pairs, as <A, V> 'look like Aя/ become (more) A...': _belyj - belet'_ 'white - to look like white / become white(r)', <N, V> 'behave like N': _pokrovitel' - pokrovitel'stvovat'_ 'a patron - to patronize, etc.

To obtain more precise derivational meanings, we need information about meanings of source words. For example, verbs formally derived from noun stems with the endingя _-it'_ can generally be defined as 'to do something connected with N'. But if source nouns have certain semantic information assigned to them, definitions can be made more precise thereby:

'N'	'V'	Examples (N - V)
a profession	to work as N	_shofer - shoferit'_ 'a driver - to work as a drive'
a person characterized by the absence of a normal quality	to cause a person to be N	_sirota - sirotit'_ 'an orphan - to make somebody an orphan' _kaleka - kalechit'_ 'a cripple - to make somebody a cripple'
a place	to be in N	_okean - okeanit'_ 'an ocean - to be in an ocean' _berloga - berlozhit'_ 'a den - to be in a den'
an instrument or a means	to do typical action with N (to use N)	_utjug - utjuzhit' 'an iron - to iron' _bomba - bombit'_

'a bomb - to bomb'
vaksa - _vaksit'_
'shoe polish - to polish shoes'

---

an artifact to create N _kopna - kopnit'_
'a haycock - to make a haycock_'

Material for further research: Xlebnikov's neologisms

The model of Russian derivation used in RUSLO currently only describes conventional derivational types. By an extention of derivational types, the range of analyzed words can be broadened. One of the possible sources for such an extension is Solzhenitsyn's dictionary [9], which contains rare and slang words. It turns out that RUSLO analyses correctly most of these words (for instance _rogov'jo_ - see above), i.e. those which do not violate Russian derivational patterns. A richer material for analysis is found in poetic neologisms, such as those invented by the Russian poet Velimir Xlebnikov. Here are some examples of Xlebnikov's neologisms.

1. Conventional suffixes
-OVA(T'):
bozhestvovat' 'to behave like the God'
silarstvovat' 'to use strenth'
-IST(YJ):
mogistyj 'able to act'
letistyj 'able to fly'
_-AR'_
glupostvovar' 'a stupid man'
-gljadar'_ 'a man who looks at something'

2. Unique suffixes
_-M(O)_ (as in _pis'-mo_ 'a letter', derived from _pisat'_ 'to write')
bajmo 'somethig said'
_Dostoevskij-mo 'something reminding of Dostoevskiy's manner of writing'
-OSH(Ъ) (as in _jun-osha_ 'a young man', derived from _junyj_ 'young')
veselosha 'a gay man'
grexosha 'an evil man'
-EBEN (as in _mol-eben_ 'public prayers', derived from _molit'sja_ 'to pray')
rjazheben 'a fancy-dressed ball'
deeben 'a public action'

3. Xlebnikovian derivational suffixes
_-ES(A)_ (as in _neb-esa_ 'the skies', interpreted as 'place')
inesa 'different place'

rovesa 'place of people's equality'
_-JAZ(') (as in _kn-jaz'_ 'a prince', interpreted as 'a creature / a governor'
nebjaz' 'an inhabitant of the skies'
ljubjaz' 'a person in love'

The vocabulary of morphologically interpreped neologisms in published Xlebnikov is about 4200 items. We have been composed this vocabulary, and now we plan to change our derivational model with regard to these neologisms in order to be able to cope with the analysis of neologisms, as well as to have more precise restrictions concerning conventional words.

References

1. _I. A. Baudoin de Courtanay. Ob otnoshenii russkogo pis'ma k russkomu jazyku. - I. A. Baudoin de Courtanay. Izbrannye trudy po russkomu jazyku. T. 2. M.: Izd-vo AN SSSR, 1963, 209-236._
2. _Grammatika russkogo jazyka. T. 1. Fonetika i morfologija. M.: Izd-vo AN SSSR, izd. 2-oe, 1960._
3. _Grammatika sovremennogo russkogo literaturnogo jazyka. M.: Nauka, 1970._
4. J. Kurylovicz. Derivation lexicale et derivation syntaxique (Contribution a la theorie des parties du discours). BSL, 37, 1936, 79-92.
5. _A. I. Kuznecova, T. F. Efremova. Slovar' morfem russkogo jazy- ka. M.: Russkij jazyk, 1986._
6. I. A. Mel'cuk, A. K. Zholkovsky. Explanatory combinatorial dictionary of modern Russian. Wiener Slawistischer Almanach, Sonderband 14, 1984.
7. _N. N. Percova, A. V. Cheremxin. Aeksperimenty po postroeniju formal'noj sistemy russkogo slovoobrazovanija. - Trudy mashinnogo fonda russkogo jazyka. T. II. M.: IRJA RAN, 1992, 86-103._
8. _Russkaja grammatika. T. 1. Fonetika. Fonologija. Udarenie. Intonacija. Slovoobrazovanie. Morfologija. M.: Nauka, 1980._
9. _A. I. Solzhenitsyn. Russkij slovar' jazykovogo rasshirenija. M.: Nauka, 1990._
10. _A. A. Zaliznjak. Grammaticheskij slovar' russkogo jazyka. M.: Russkij jazyk, 1977._