Minimizing Acquisition and Development Effort in Computational Morphology

Svetlana Sheremetyeva and Sergei Nirenburg

New Mexico State University, Computing Research Laboratory

Abstract

This paper presents a model of morphological analysis which, unlike practically all others, does not rely on a dictionary of stems. Consequently, the effort to acquire the static knowledge sources for systems based on this model is significantly smaller. The model is engineering-oriented: the considerations of economy and efficiency led to the use of non-traditional definitions of morphemes. The model is implemented and tested on the material of Russian and Serbo-Croatian.

1 Introduction

Most current systems of computational morphology rely, in addition to word inflection paradigms, on large lexicons of stems. Morphological modules of larger NLP systems are expected to provide near-perfect coverage. Therefore, the completeness of lexicons is an ongoing concern (especially if one considers that, according to some estimates, on average five new words enter a major language, such as English or Russian, every day). It is often the case that large on-line dictionaries are not available to NLP developers (they either don't exist, e.g. for low density languages, or simply cannot be reached at the moment as was the case with the Corelli project at CRL, NMSU). It would be a clear advantage to be able to carry out morphological processing without the reliance on a large dictionary of stems and, therefore, without a need for a large-scale knowledge acquisition effort. Our system attempts just that.

This paper presents a model of morphological analysis which, unlike practically all other approaches to computational morphology, does not rely on a large dictionary of stems. For the treatment of inflectable open-class items several relatively small-scale lexicons are used in place of the usual large lexicon. The model also covers proper names and hyphenated words. The system is robust and can be used both for lemmatization (base form generation) and for determining the values of morphosyntactic features of word forms. It can be used in text analysis and generation. The model has been implemented in the Rapid Deployment Morphology (RDM) system in the framework of the Corelli project at CRL, NMSU. It was originally developed for Russian and then successfully applied to Serbo-Croatian. We expect it to be extendable to a variety of flective and agglutinating languages. Examples in this paper are all from RDM-Russian.

The central differences between our model and those used in practically all other approaches to computational morphology — two-level "KIMMO" systems

(Koskenniemi, 1984; Antworth, 1990), DECOMP (Allen et al., 1987), MORPHOGEN (Pentheroudakis and Higinbotham, 1991) or the morphological module in the ETAP-2 system (Apresyan et al., 1989), to name just a few — lie in a) the lack of reliance on a comprehensive stem dictionary and b) the choice of the basic units of description. Two-level rules are based on the correspondence between lexical and surface forms; in other models rules are described in terms of abstract *morpheme* categories (MORPHOGEN) or *morphs*, the orthographic representation of morphemes (DECOMP, ETAP-2) and inflection paradigms. In our model, a crucial unit of description, the *quasi-root* (see definition below) is not a regular morpheme, as it is not expected to carry meaning. This decision has been made for reasons of assuring coverage, efficiency and economy of knowledge acquisition. It does not advance theoretical connections between morphology and semantics. This approach is one of those "bags of tricks" which, in the opinion of Yehoshua Bar-Hillel, seconded by Yorick Wilks, are essential components of all practical NLP systems and "will advance computational linguistics in the future" (Wilks, 1996).

One of the requirements for our model was absence of reliance on dictionaries of base forms, roots or stems. In an attempt to fulfil this requirement, we have observed that many Russian words belonging to the same inflectional paradigm have identical strings of characters just before their endings. For example, a Russian character string ***ani*** associated with the following declension paradigm: ***e ja ja j yu jam e ja em jami i jakh*** is common to a large number of nouns such as *zavoev**anie*** (gain), *vyzhiv**anie*** (survival), *obzhalov**anie*** (appeal), *osnov**anie*** (foundation), *sozd**anie*** (creation), *sobr**anie*** (meeting), *pit**anie*** (nourishment), etc. In fact, the Grammatical Dictionary of Russian (Zaliznjak, 1980) lists in the order of 1,500 words of this type. We hypothesized that this shared string (we call it "quasi-root" to distinguish it from the more traditionally defined "root" morpheme) can be used as a determinant of the morphological information for the entire class. This paper reports the results of implementing and testing this hypothesis. In what follows we present the morpheme classes recognized by our model, the lexicons of our model, and the process of morphological analysis in our model.

2 Morphotactics

Our model recognizes the following types of morphemes (some of them are similar to traditionally defined morphemes, while some others are defined in a different manner):

- *reflexive*: a word-final character string whose values are one of {***Ø***, ***sja***, ***s'***};
- *ending*: a word-final or pre-reflexive character string (possibly empty), as found in the lexicon of endings (the endings do not always coincide with those defined in traditional grammar);

- *suffix*: a character string (possibly empty) immediately preceding the ending, as found in the lexicon of suffixes (this lexicon contains only suffixes of full participles and adjectives in the superlative degree);
- *quasi-root*: a three-letter character string immediately preceding the ending or suffix, if present; examples of quasi-roots include: *bu-**mag**-a* (paper), *u-**pak**-ovat'* (pack), *prekr-**asn**-yi* (beautiful), *populj-**arn**-eishij* (most popular*)*, *s-**del**-annyj* (done); if the word does not have enough characters left when the quasi-root is selected, the latter is "padded" on the left with one or two null characters, as in ***Øug**-ol* (angle);
- *prefix*: a word-initial character string (possibly empty), as found in the lexicon of prefixes (this lexicon contains only prefixes of perfective forms of the verbs);
- *body*: a compound character string (never empty) starting after the prefix, if present, and ending before a suffix, if present, or ending;
- *stem*: a word-initial compound character string (never empty) stretching to the ending, as found in the lexicon of stems (see below).

The morphotactics of Russian in our model is then defined by the following very simple grammar:

word ::= stem ending reflexive
stem ::= prefix [string quasi-root | body] suffix

where "string" stands for any character string including the null string.

For example, the correct analysis of the Russian word *sdelala* (did, *feminine, singular*) leads to the following assignments of morph types:

sdelala = stem: ***sdela*** ending: ***la*** reflexive: ***Ø***
stem: ***sdela*** = prefix: ***s*** [string: ***d*** quasi-root: ***ela*** | body: ***dela***] suffix: ***Ø***

In what follows we describe the content and structure of the lexicons.

2.1 Lexicon of endings and reflexives

The lexicon of endings and reflexives (about 900 entries) consists of all the experimentally derived inflections for nouns, verbs, participles and adjectives. Endings recognized in our model do not always coincide with those in traditional grammar. We posit that all the substrings at the end of the word, which change during declension, are endings. Our choice of endings was inspired by the desire to bypass special treatment of alternations. This is why the ending in *ugol* is ***ol*** (because of the existence of *uglom* (singular instrumental), in which ***lom*** is the ending). Theoretically, *ugol* has a null ending, while the ending in *uglom* is ***om***. An entry is of the form:

ending --> $\{\{ f_{11}, ..., f_{1n}\}, ..., \{f_{m1}, ..., f_{mn}\}\}$, where f_{ij} are values of morphosyntactic features associated with this ending, one of which, the part-of-speech feature, is obligatory. It is possible that several feature value sets will be listed for a single part of speech, which reflects the fact that there can be grammatical form homonymy. The entries contain information from inflection paradigms for verbs, adjectives and participles. The entries, in fact, could be considered to be these paradigms, only inversely indexed by the ending and augmented with the information with some additional features, such as aspect for verbal readings or number of declension paradigm for nominal readings. Participial readings are a special case. Their adjectival feature values are listed in this lexicon, while their verbal feature values are obtained from the lexicon of suffixes (see below).

For example, the entry of the ending ***la*** is as follows:

la
noun, 25, masculine, singular, genitive
verb, imperfective, past, feminine, singular
verb, perfective, past, feminine, singular

The number in the nominal reading is that of the declension paradigm. This entry is matched by such words as *ug**la*** (angle, *noun, masculine, singular, genitive*), *sdela**la*** (did, *verb, perfective, past, feminine, singular*), *bezha**la*** (was running, *verb, imperfective, past, feminine, singular*), etc.

The words *peremenn**aja*** (variable, *noun, feminine, singular, nominative*), *beg**aja*** (while running, *adverbial participle, imperfective*), *otrezann**aja*** (cut, *participle, full, perfective, passive, feminine, singular, nominative*), *krasiv**aja*** (beautiful, *adjective, feminine, singular, nominative*) will be analysed using the entry:

aja
noun, 18, feminine, singular, nominative
noun, 30, feminine, singular, nominative
adverbial-participle, imperfective
participle, full, perfective, passive, feminine, singular, nominative
participle, full, perfective, active, feminine, singular, nominative
participle, full, imperfective, active, feminine, singular, nominative
adjective, feminine; singular, nominative

The entries for the two reflexives, ***sja*** and ***s'*** are added to the lexicon of endings because of the possibility of homonymy between reflexives and "true" endings. The difference between reflexives and endings in our system is purely morphotactic, made manifest by differences in continuation classes for the two morpheme classes.

2.2 Lexicon of suffixes

The lexicon of suffixes (58 entries) contains only suffixes of full participles and adjectives in the superlative degree. It consists of entries, a subset of which does not correspond to the traditional participial suffixes. This device, just as in the case of endings, is used for resolving verb form homonymy. For each participle suffix the lexicon lists a unique set of feature values which participles share with verbs. The feature values which participles share with adjectives are recorded in the lexicon of endings. The entry in the lexicon of suffixes is of the form:

suffix --> {f_1, f_2},

where f_1 is aspect and f_2 is voice. A sample entry is as follows:

chenn
participle, full, perfective, passive

This suffix can be identified in such words as *opla**chenn**yj* (paid), *potra**chenn**aja* (spent), *skhva**chenn**oe* (caught).

The entry for adjective suffixes in the superlative degree contains no feature values:

ejsh
adjective, superlative

This suffix is found in such adjectives as *star**ejsh**ij* (the oldest), *krasiv**ejsh**ij* (the most beautiful), *uzhasn**ejsh**ij* (the most terrible).

The lexicon of suffixes is used a) to test participial candidates, distinguish them from candidates belonging to other categories and assign their verbal feature values; and b) to test adjectives for superlative degree forms.

2.3 Lexicon of quasi-roots

The lexicon of quasi-roots (about 2,900 entries) is the main lexicon in the model. Every entry is marked as a nominal, verbal or adjectival quasi-root. The adjectival and verbal entries include endings of their base forms. The nominal entries include a set of declension paradigm numbers with the endings of their base forms. Some sample entries of the quasi-root lexicon are presented below.

ani
*Noun, 3 **ja**; 11 **e***
atsk
*Adjective, **ij***
ata
*Verb, **t'***

eli
Noun, 11 ***e***

This lexicon is used to resolve categorial homonymy among verbal, nominal and adjectival candidates as well as to construct base forms in case of verbs, participles and adjectives.

2.4 Lexicon of prefixes

The lexicon of prefixes (53 entries) contains a list of perfective verb prefixes and is used to resolve homonymy between perfective and imperfective forms of the verb.

2.5 Lexicon of bodies

Each body in the lexicon of bodies (about 1,000 entries) appears with its corresponding base form ending. The lexicon of bodies contains bodies of perfective verbs and is used for homonymy resolution between perfective and imperfective verbs (quasi-roots of such verbs are homonymous) and for finding base forms of perfective participles and perfective verbs.

blokir
ovat'
vra
tit'

The base form endings are also listed in the lexicon of quasi-roots. We repeat them in the lexicon of bodies to help resolve a possible ambiguity among the base form endings of a single quasi-root.

2.6 Lexicon of stems

The lexicon of stems (about 7,000 entries) is used to resolve those cases of quasi-root homonymy for which the quasi-roots do not supply a solution. This lexicon includes:

- Stems of adjectives and participles which can be used as nouns in Russian (used to determine whether an input which declines as an adjective should yield a nominal or an adjectival reading or both); for example, the lexicon of stems contains the entries

peremenn
Adjective ***yj****; Noun, feminine,* ***aja***
upolnomochenn
Participle ***yj****; Noun, masculine,* ***yj***

The above entries show that the word *peremennaja* (variable) can be either a noun or an adjective in Russian, while the word *upolnomochennyj* (plenipotentiary) can be a participle or a noun; RDM outputs both readings.

- Stems of prefixless perfective or bi-aspectual verbs (whose aspect value cannot be determined based on affixation):

  ratifitsirov
  Verb Imperfective, ***at'****; Verb Perfective,* ***at'***

  RDM outputs both readings for *ratifitsirovat'* (ratify).

- Verbal stems which end in ***s'*** in imperative mood but are not reflexives:

  bro
  *Verb, imperative, -****sit'***

  The above entry is used on the input *bros*' (throw, *imperative*), which, in the absence of this lexicon would be understood by RDM as a reflexive.

- Some nominal stems which exhibit homography of quasi-roots; namely, if there are two or more classes of nouns such that (i) they have the same quasi-root and (ii) at least one of the endings is shared among these paradigms, then the paradigm(s) with the smaller number of members are included in the lexicon of stems, while the paradigm with the greatest number of members is processed by RDM in the regular way (incidentally, this is another instance of minimizing the effort of acquisition); for example, the lexicon of stems will include the entry

  shtor
  1 a

  which lists a noun *shtora* (curtain); the quasi-root of this noun and its endings, for instance, in Accusative Singular and Instrumental Plural are homographic with those of a large group of nouns such as *motor* (engine); the paradigm of *motor* has 402 members, while the paradigm of *shtora* has only five, according to Zaliznjak (1980, 536-9, 207); therefore, the members of the latter paradigm are stored in the lexicon of stems.

3 Analysis

The model accepts an inflected Russian word form as input and returns all its legal base forms with all legal morphosyntactic feature value sets. The analysis process starts at the end of the word and proceeds backwards. Segmentation is carried out simply by matching the characters of a word against lexicon entries or by chopping off a certain number of characters to the left of the current segmentation point. We use a bottom-up, depth-first parsing algorithm. The main analysis procedure is *analyse-word*:

```
procedure analyse-word
  find-candidates;
  process-candidates.
```

Procedure *find-candidates* searches the lexicon of endings for all possible matches on the input word, returning all possible pairs of endings and their feature values which are stored with the entries of the lexicon of endings. Each such pair is called a *candidate*. Several candidates may be returned for a single word. For instance, for the word *oplachennymi* (paid, *participle, perfective, passive, plural, instrumental) find-candidates* will return the following candidates (this example, in fact, involves the most ambiguous ending in our model, ***i***; we list only five candidates for this ending overtly[1)].

Table 1: Sample results of find-candidates.

	Ending	**Part of Speech and Feature values**	**Example**
1	ymi	Noun, 17, neuter, plural, instrumental	dannymi
2	ymi	Noun, 18, feminine, plural, instrumental	peremennymi
3	ymi	Noun, 48, masculine, plural, instrumental	uchenymi
4	ymi	Noun 49, masculine, plural, instrumental	postovymi
5	ymi	Participle full, plural, instrumental	oplachennymi
6	ymi	Adjective plural instrumental	krasivymi
7	mi	Verb imperative perfective singular	primi
8	i	Verb imperative perfective singular	posmotri
9	i	Verb imperative imperfective singular	smotri
10	i	Adjective short, plural	khoroshi
11	i	Noun, 2, feminine, plural, nominative	jabloni
...	i	...	...
69	i	Noun, 11, neuter, singular, prepositional	trenii

Further processing will, at least partially, disambiguate the homonymy among the readings. It is carried out by the following procedure:[2]

```
procedure process-candidates
  process-nominal-candidates;
```

[1] We hasten to note that only very few, if any, of these candidates undergo a full treatment in our model; see example analyses in Section 4.

[2] We describe the analysis process conceptually. It can be implemented in many ways.

```
IF no legal nominal readings
THEN  begin
      process-participle-candidates;
      IF no legal participle candidates
      THEN  begin
            process-verbal-candidates;
            IF no legal verbal candidates
            THEN process-adjective-candidates
            end
      end
ELSE  process-verbal-candidates
      UNLESS stems of legal nominal readings are in lexicon of stems
```

The input to every component procedure in *process-candidates* is the input word and a set of corresponding candidates (nominal, participial, verbal and adjectival). The order of the calls to component procedures in the above algorithm is established to minimize the processing time and effort, as it is possible to avoid calling each of the component procedures with each input word. The ordering is based on a set of heuristics such as the following. Noun candidates are processed first because a vast majority of words in a Russian text are nouns and nouns can relatively rarely also be analysed as adjectives, and can very rarely be verbs or adverbs. If a participle candidate has been analysed successfully, there is no need to test any verbal candidates because in Russian participles and finite forms of verbs are practically never homographic. This order is clearly language-dependent (though for Serbo-Croatian, the order developed originally for Russian worked). If the ordering heuristics are not available, the control structure of *process-candidates* can easily be modified to have each subroutine called in any order. We realize that a more fashionable approach is to use a universal rule interpreter and not concern oneself with heuristics-based control issues. However, universal interpreters sometimes make it more difficult to write processing rules, and in any case, in the absence of such an interpreter, the concern for minimization of effort led us to use linguistic heuristics for control.

4 Examples of Processing

We describe the procedures for processing the candidates of different parts of speech through examples. We trace the processing of the participle *oplachennymi* (paid), the noun *peremennymi* (variables), the verb *smotri* (see), and the adjective *krasivymi (*beautiful*)*. These examples are quite complex and we selected them to be able to illustrate as many of the types of processing performed by the system as possible.

4.1 *Oplachennymi*

The candidates for this word are given in Table 1. Procedure *process-candidates* stipulates that the nominal candidates are tested first. Procedure *process-nominal-candidates* first chops off the quasi-root of the input word. The quasi-root in *oplachennymi* with the ending ***ymi*** (for nominal candidates 1-4) is ***enn***; the quasi-root with the ending ***i*** (for nominal candidates 11-69) is ***nym***. Next, these quasi-roots are looked up in the lexicon of quasi-roots. No match with the part-of-speech feature value *noun* is found for ***nym*** in the quasi-root lexicon, so all the candidates 11-69 are immediately discarded. A nominal match is found for ***enn***. The corresponding entry in the lexicon of quasi-roots states that the noun belongs to Declension Paradigms 17, 18, 48 or 49 (which are all paradigms for deadjectival nouns; as an example, Paradigm 18 is as follows: ***aja ye oj yh oj ym uju ye oj ymi oj yh***).

Process-nominal-candidates next checks whether any of these paradigm numbers appear in the candidates. If it is not the case, the candidate is discarded. If it is, the base form of a noun is identified in the quasi-root lexicon. The input word with its base form and the feature value list is then output as a result of the morphological analysis process.

Some declension paradigms (including Paradigms 17, 18, 48 and 49) are applicable to adjectives and full participles. For these paradigms, the procedure finds the stem of the input word and checks whether this stem is listed in the stem lexicon as a nominal stem. The stem ***oplachenn*** is not in this lexicon. Therefore, candidates 1-4 are discarded.

Next, *process-participle-candidate* is called for candidate 5. The procedure first attempts to match a string to the left of the ending to an entry in the lexicon of suffixes. Our input matches the suffix ***chenn,*** as a result of which the feature values *passive* and *perfective* are added to the list of feature values in the candidate. At this point, candidate 5 is as follows:

ymi *participle full, passive, perfective, plural, instrumental*

It remains to determine the base form of the word. For participles we require both the participial and the verbal base form. The former is derived graphotactically — if the participial suffix ends in a sibilant, the base form ending is ***ij***; otherwise, it is ***yj***. The latter is determined using the lexicon of quasi-roots or, in case of prefixed perfective forms, in the lexicon of bodies. In our case the prefix ***o*** is identified and the body ***pla*** (*o-**pla**-chenn-ymi*) is extracted. The body lexicon has the entry ***pla-tit'***. The final output of the analysis is:

oplachennymi *oplachennyj oplatit'*
participle full, passive, perfective, plural, instrumental

4.2 *Peremennymi*

Process-nominal-candidates starts in exactly the same way as in the case of *oplachennymi*: candidates 11-69 are discarded; candidates 1-4 are found to belong to Paradigms 17, 18, 48 and 49, respectively. Next, the quasi-root of the input word is determined (it is ***enn***) and the paradigm number in the corresponding entry of the lexicon of quasi-roots is compared to that of each of the candidates 1-4. The paradigm number in candidate 1 does not appear in the entry for ***enn*** in the lexicon of quasi-roots. Therefore, it is discarded. Since the paradigms of candidates 2-4 are adjectival, the lexicon of stems is consulted. Unlike ***oplachenn***, the stem ***peremenn*** is found there. This means that it is a legal de-adjectival noun. *Process-nominal-candidates* next attempts to unify the feature values for candidates 2-4 with the nominal features of the entry for ***peremenn***. Only one match is found:

> ***peremennymi*** *peremennaja,*
> *noun, feminine, plural, instrumental*

As a side effect of finding this stem in the lexicon of stems, the system immediately outputs the adjectival reading:

> ***peremennymi*** *peremennyj,*
> *adjective, plural, instrumental*

At this point the procedure *process-candidates* terminates.

4.3 *Smotri*

For this word *find-candidates* returns candidates 8-69 from Table 1. *Process-nominal-candidates* for candidates 11-69 chops off the quasi-root ***otr*** for which in the lexicon of quasi-roots there is a match with the part-of-speech feature value *noun* associated with Paradigm 6. There is no candidate belonging to Paradigm 6 in the candidate list; as a result, candidates 11-69 are discarded. The only remaining candidates are verbal and adjectival. Procedure *Process-verbal-candidates* is first applied to candidates 8 and 9. Since the value of the aspect feature of candidate 8 is *perfective,* the input word's body is calculated and looked up in the lexicon of bodies. The first letter of the input word is chopped off as it matches an entry in the lexicon of prefixes. No match is found for the body ***m*** and candidate 8 is discarded. Note that the system gives a correct result because in grammatical terms ***s*** is not a prefix in *smotri.* Indeed, the lexicon of bodies contains only bodies of existing perfective verbs. For candidate 9 the quasi-root ***otr*** is checked and found in the lexicon of quasi-roots with a verbal reading. The candidate, therefore, survives. The only feature value absent at this time is the base form, which RDM finds in the lexicon of quasi-roots. Thus, the output is:

smotri *smotret',*
verb, imperative, imperfective, singular

4.4 *Krasivymi*

The starting point for *process-candidates* is the list of candidates in Table 1. *Process-noun-candidates* fails to find nominal matches for the quasi-roots ***siv*** (candidates 1-4) and ***vym*** (candidates 11-69), and all nominal candidates are discarded. *Process-participle-candidates* cannot find a match for any suffix in the lexicon of suffixes, so that candidate 5 is discarded. Candidates 7 and 8 are also discarded, as *process-verb-candidates* fails to extract a prefix and the stem *krasiv* is not in the lexicon of stems marked as a prefixless perfective verb. Candidate 9 is discarded as well as no verbal match is found for the quasi-root ***vym***. Only candidates 6 (regular adjective) and 10 (short adjective) remain. For candidate 6, its quasi-root ***siv*** is found in the lexicon of quasi-roots with the adjectival reading. For candidate 10 its quasi-root ***vym*** has a match in the lexicon of quasi-roots but the entry has no adjectival reading. Candidate 10 is discarded. The final result is:

krasivymi *krasivyj,*
adjective, plural, instrumental

5 Evaluation

The Russian morphological analyser was tested on an arbitrarily selected raw text from the *Moscow News* newspaper. The corpus comprised about 10,000 usages of 2,562 different lexemes. The results were evaluated manually and are as follows: for 2.6% of the word usages the system overgenerated, i.e. for these words incorrect readings were generated together with the correct ones; for 0.1% of the word usages the system undergenerated, which means that fewer than the full complement of the possible homographic readings was generated; for 1.1% of word usages no fully correct output was generated. In 0.8% of the cases both the base form and the features were incorrect; in 0.2% of the cases the system output an incorrect base form but correct feature sets; and in only 0.08% of the cases an incorrect base form but a correct part of speech feature was returned. In the rest of the cases, the system operated optimally.

We were unable to compare these results directly with those of other Russian morphological analysers by running them on our test text, due to the unavailability of these analysers. In the descriptions of such systems no evaluation results were given. The issue of evaluation is typically mentioned in passing, if at all. For example, the system described in Bolshakov (1993) is said not to provide enough morphosyntactic information about words and could not resolve rampant overgeneration. The morphological analyser of Segalovich (1995) is reported to be optimized for speed at the expense of accuracy and extendability. Mikheev and Liubushkina (1995) mention that their system is implemented for 1,500,000 word-tokens; other works available to us do not touch the matter at all (e.g. Malkov et al., 1983; Ashmanov, 1995).

6 Conclusion and discussion

The objective of this work was to develop a methodology for a quick ramp-up morphology analyser when the resources are scarce. The main distinguishing feature of this approach is (a) its lack of reliance on large traditional lexicons of stems which is crucial in the absence of large on-line resources (which is often the case in practice, especially when dealing with "low-density" languages) and (b) a novel treatment of stem alternations. Both features allow for great economy in the development effort.

The former minimizes acquisition time: while it is necessary in both the traditional and the RDM approaches to construct full inflectional paradigms, the RDM approach requires much less effort to key in the main system lexicons. As an example, consider that in the traditional model, 1,540 different lexicon entries will have to be written for the words of the type of *zdanie* (building) and *kompania* (company), while in RDM all that is needed is one entry in the lexicon of quasi- roots:

ani 11 ***e***; 7 ***a***

where 7 and 11 are paradigm numbers, and the letters refer to endings of base forms. Indeed, the entire set of lexical resources in RDM consists of fewer than 12,000 items, while guaranteeing practically complete coverage of Russian. For comparison, the coverage of the first Russian morphological analysers developed before 1980 was rather limited (up to several thousand words; Malkov et al., 1983). Later systems, in search of better coverage, featured larger lexicons of between 100,000 and 200,000 entries (Belonogov and Zelenkov, 1989; Bolshakov, 1993; Mikheev and Liubushkina, 1995). All the above systems treated *only* the words covered by their lexicons: if the stem of a word was not found in a system lexicon the analysis failed.

The RDM model is robust in that it can process unknown words. For example, a system based on the lexicon of stems of 100,000 words included in the Zaliznjak dictionary (Zaliznjak, 1980) would fail to analyse such regular and frequent newly-coined Russian words as *diskoteka* (disco), *vaucher* (voucher), *evroremont* (European style apartment remodelling), etc. Though in RDM the lexicon of quasi- roots has been also extracted using Zaliznjak (1980), RDM can process these words. The quasi-roots ***tek, cher*** and ***ont*** are found in our lexicon of quasi-roots (extracted, for example, from such words as *biblioteka* (library), *vecher* (evening), *remont* (repair).

It is important to realize that it is impossible to generate a complete lexicon of stems automatically from an on-line dictionary, because the latter does not cover alternations and (often) suppletive forms. The knowledge to cover these phenomena must be acquired manually, and the problem of treating suppletive forms and alternations requires (as was pointed out by many developers; see, e.g. Mikheev and Liubushkina, 1995) the delineation of sets of transformational classes of stems and development of complex control strategies for choosing the right form from the set. Our approach, in addition to minimizing the knowledge acquisition effort, allows us to bypass the need to create such sets of transformational classes and control strategies by simplifying the treatment of alternations:

- In the general case, we posit that the entire word-ending substrings, which change when the word is inflected, are endings. The traditional theoretical approach is followed in RDM when no alternations exist. Thus, the word *stol* (table) which in the traditional Russian morphology belongs to the same paradigm as the word *ugol* is assigned the traditional zero ending and thus belongs to a different RDM paradigm (the RDM ending in ***ugol*** (corner, angle) is ***ol***, because of the existence of the form ***uglom*** (singular instrumental), in which ***lom*** is the RDM ending). Depending upon the paradigm assigned to the word, we extract its quasi-root: ***Øug*** for ***ugol*** (***Ø*** stands for a zero character) and ***tol*** for ***stol***.
- In a small number of special cases in which the above method would lead to undue propagation of the number of inflection paradigms, we deal with alternations directly by introducing two quasi-roots for a single word. For instance, the verb *sberech'* (save) is associated with two quasi-roots ***rech*** and ***reg*** which cover all the alternations (*sberegu*, future, perfective, second person, singular). Thė two quasi-roots, in the general case, appear in more than one lexeme, so that the same kind of alternation can be seen in such word as *uberech'* (save (from)), *priberech'* (keep safe), *sterech'* (guard), *predosterech'* (warn), etc. As the numbers of cases where we use this technique are not numerous, the size of the quasi-root lexicon does not noticeably increase.

The entire selection and compilation work on the RDM lexicons has been accomplished in about ten person-weeks. Of course, we had to include additional lexicons to cover specific problems which the lexicon of quasi-roots could not solve and we expect the number of elements in the lexicons to grow somewhat as we process the larger corpora. However, the sum total of effort for knowledge acquisition in RDM has been negligible compared with that necessary for compiling a standard lexicon for morphological analysis.

We are fully aware that "standard" lexicons will be needed if this model is used as a component of a larger NLP system, such as a syntactic parser or semantic analyser. However, the results of this work can be immediately incorporated in systems for tagging large corpora and other important tasks. In the immediate future we intend to add to this system some context-based rules for selecting from among the multiple outputs of the system.

References

Allen, J., M.S. Hunnicutt and D. Klatt (1987). *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge.

Antworth, E. (1990). PC-KIMMO: A Two-Level Processor for Morphological Analysis. Occasional Publications in Academic Computing 16, Summer Institute of Linguistics, Dallas, TX.

Apresyan, Ju.D., I.M. Boguslavsky, L.L. Iomdin, A.V. Lazursky, N.V. Pertsov, V.Z. Sannikov and L.L. Tsinman (1989). *Lingvisticheskoe obespechenie sistemy ETAP-2* (Linguistic Knowledge Sources in the ETAP-2 System). Nauka, Moscow.

Belonogov, G. and Ju. Zelenkov (1989). An Algorithm for Morphological Analysis of Russian Words, in *Voprosy informatsionnoj teorii i praktiki. Moscow*, 35-42 (in Russian).

Bolshakov, I. (1993). A Multifunctional Thesaurus for Computer Editing of Russian Texts, in *NTI*, s2, no. 7.

Koskenniemi, K. (1984). A general computational model for word form recognition and production, in *Proceedings of COLING-84*. Stanford.

Malkov, M., I. Volkova and T. Gratsianova (1983). Linguistic Processor TULIPS-2: Morphological Component, in *Razrabotka i Primenenie Lingvisticheskih Protsessorov. Novosibirsk*, 34-42 (in Russian).

Mikheev, A. and L. Liubushkina (1995). Russian Morphology: An Engineering Approach, in *Natural Language Engineering* 1(3), 235-260.

Pentheroudakis, J.E. and D.W. Higinbotham (1991). Morphogen: A Morphology Grammar Builder and Dictionary Interface Tool, in *Proceedings of the 1991 Meeting of the Desert Language and Linguistics Society*.

Ritchie, G.D., A.W. Black, S.G. Pulman and G.J. Russell. The Edinburgh/Cambridge Morphological Analyser and Dictionary System. Version 3.0. Unpublished ms.

Segalovich, I. S. (1995). Indexing of large Russian texts with a dictionary built around the sparse hash table, in *Proceedings of Dialog'95 International workshop on Computational Linguistics and its Applications*. Kazan, Russia.

Wilks, Y. (1996). Guest Editor's Introduction, in *Communications of ACM*, 39:1 Special Issue on NLP, 62.

Zaliznjak, A.A. (1980). *Grammaticheskij slovar' russkogo jazyka* (A Grammatical Dictionary of Russian). Russkij Jazyk, Moscow.