University of Brighton

ITRI-03-23

Russian Lemmatisation with DATR

Roger Evans and Carole Tiberius and
Dunstan Brown and Greville Corbett

October, 2003

Supported by ESRC grant no. RES-000-23-0082 to Surrey University

Information Technology Research Institute Technical Report Series

ITRI, Univ. of Brighton, Lewes Road, Brighton BN2 4GJ, UK
TEL: +44 1273 642900 EMAIL: firstname.lastname@itri.brighton.ac.uk
FAX: +44 1273 642908 NET: http://www.itri.brighton.ac.uk

Russian Lemmatisation with DATR

R. Evans
ITRI
University of Brighton
roger.evans
@itri.brighton.ac.uk

C. Tiberius, D. Brown, G.G. Corbett
Surrey Morphology Group
University of Surrey
{c.tiberius,d.brown,g.corbett}
@surrey.ac.uk

Abstract

In this paper, we describe an approach to lemmatisation for Russian nouns, which makes use of a large-scale inheritance lexicon implemented in the lexical representation language DATR (Evans and Gazdar 1996). The lexicon was compiled semi-automatically from Zaliznjak's morphological dictionary (Zaliznjak 1977, Ilola and Mustajoki 1989) and automatically generates fully inflected forms together with their associated morphosyntax for around 40,000 Russian nouns. From this resource, we have automatically extracted wordform recognition rules and compiled them into a lemmatiser which hypothesises possible citation form and morphosyntactic features for nominal wordforms. We describe the construction of the lemmatiser and the results of our initial evaluation of its accuracy.

1 Introduction

Our goal is to undertake a detailed corpus analysis of Russian texts, focusing on the relationship between morphological ambiguity (syncretism) in nouns and adjectives and the comparative frequency of the relevant grammatical categories[1]. For this purpose, we have chosen to develop our own lemmatisation and tagging technology, based on the electronic version of Zaliznjak's dictionary (Zaliznjak 1977, Ilola and Mustajoki 1989). We have previously reported (Evans, Tiberius, Brown and Corbett 2003) the construction of an inheritance-based morphological lexicon from this resource. In this paper we briefly review this lexicon and describe the next phase in our programme, the development of a wide-coverage morphosyntactic lemmatiser.

The morphological lexicon and lemmatiser are represented using the lexical representation language DATR (Evans and Gazdar 1996). This allows information to be shared via an inheritance hierarchy, while supporting subregularities and exceptions through the use of overriding of defaults. The hierarchical structure is exploited in several stages of processing: compilation of the lexicon, derivation of morphological forms and extraction of lemmatiser rules. For our resources to be maximally reusable in other contexts, we have adopted Unicode (UTF-8) as the standard representation for all our data.

The paper is structured as follows. Section 2 reviews the large-scale inheritance-based morphological lexicon for Russian derived from Zaliznjak's dictionary; Section 3 describes the construction and evaluation of the lemmatiser; Section 4, discusses principal areas of further development and Section 5 concludes the paper.

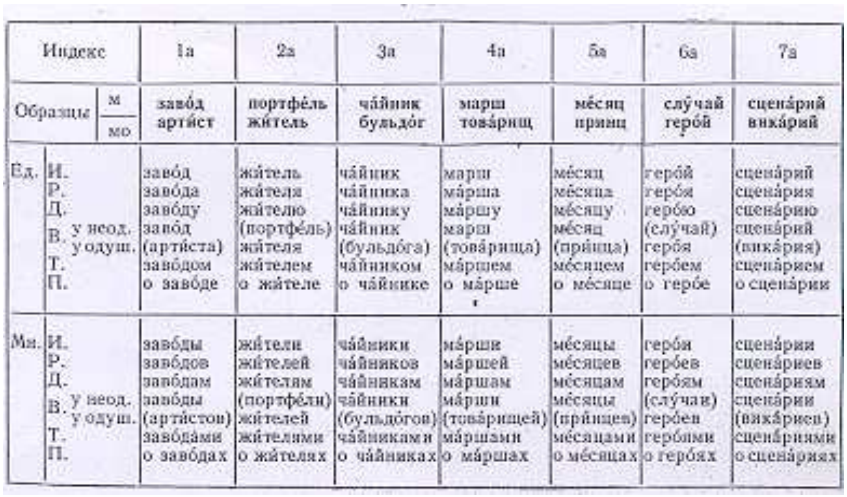
2 An inheritance-based morphological lexicon for Russian

2.1 The Zaliznjak dictionary

Zaliznjak's dictionary deals primarily with Russian inflectional morphology giving explicit information about inflectional forms and stress. There are two versions of the dictionary, a printed version which was first published in 1977 and an electronic version which was created

[1] The research reported here is supported by the Economic and Social Research Council (UK) under grant RES-000-23-0082 'Paradigms in Use'. Their support is gratefully acknowledged.

by Ilola and Mustajoki (1989). The 1977 printed version is a reverse dictionary consisting of two parts. The first is a set of tables identifying morphosyntactic classes and defining the realisation of morphological features with them. The second is a listing of almost 100,000 lexical entries each including an index into a realisation table in the first part. For example, the word *абажур* 'lamp shade' is a masculine noun of type 1A and as such follows the inflectional pattern of *завод* 'factory' which is given as the example paradigm for masculine nouns of type 1A (see Figure 1). The electronic form contains just the set of lexical entries (101401 lines, 98729 lexical entries).



Индекс		1а	2а	3а	4а	5а	6а	7а
Образцы	м / мо	заво́д артист	портфе́ль жи́тель	ча́йник бульдо́г	марш това́рищ	ме́сяц принц	слу́чай геро́й	сцена́рий вика́рий
Ед.	И.	заво́д	жи́тель	ча́йник	марш	ме́сяц	геро́й	сцена́рий
	Р.	заво́да	жи́теля	ча́йника	ма́рша	ме́сяца	геро́я	сцена́рия
	Д.	заво́ду	жи́телю	ча́йнику	ма́ршу	ме́сяцу	геро́ю	сцена́рию
	В. у неод.	заво́д	(портфе́ль)	ча́йник	марш	ме́сяц	(слу́чай)	сцена́рий
	В. у одуш.	(арти́ста)	жи́теля	(бульдо́га)	(това́рища)	(при́нца)	геро́я	(вика́рия)
	Т.	заво́дом	жи́телем	ча́йником	ма́ршем	ме́сяцем	геро́ем	сцена́рием
	П.	о заво́де	о жи́теле	о ча́йнике	о ма́рше	о ме́сяце	о геро́е	о сцена́рии
Мн.	И.	заво́ды	жи́тели	ча́йники	ма́рши	ме́сяцы	геро́и	сцена́рии
	Р.	заво́дов	жи́телей	ча́йников	ма́ршей	ме́сяцев	геро́ев	сцена́риев
	Д.	заво́дам	жи́телям	ча́йникам	ма́ршам	ме́сяцам	геро́ям	сцена́риям
	В. у неод.	заво́ды	(портфе́ли)	ча́йники	ма́рши	ме́сяцы	(слу́чаи)	сцена́рии
	В. у одуш.	(арти́стов)	жи́телей	(бульдо́гов)	(това́рищей)	(при́нцев)	геро́ев	(вика́риев)
	Т.	заво́дами	жи́телями	ча́йниками	ма́ршами	ме́сяцами	геро́ями	сцена́риями
	П.	о заво́дах	о жи́телях	о ча́йниках	о ма́ршах	о ме́сяцах	о геро́ях	о сцена́риях

Figure 1. Extract of Zaliznjak's paradigm tables for masculine nouns (Zaliznjak 1977: 39)

2.2 Mapping Zaliznjak into DATR

The mapping process from Zaliznjak into DATR has been described in detail in Evans, Tiberius, Brown and Corbett (2003). Here we briefly summarise the slightly refined process we currently use. The mapping involves two distinct components: a) the manual construction of a DATR representation of the morphosyntactic class and realisation information from the printed paradigm tables (see Section 2.2.1), and b) the automatic construction of the individual lexical entries from the electronic dictionary data (see Section 2.2.2), which is now achieved in a more efficient way than previously described.

2.2.1 The hand-crafted DATR theory

Zaliznjak does not use the traditional division of words into declension types in his dictionary, but divides nouns into types according to the last grapheme of the stem. (Ilola and Mustajoki 1989:9) For example, he distinguishes eight types for masculine nouns numbered `1` to `8`. These morphological types are then further divided according to stress. The masculine noun types can occur with six different stress patterns indicated by subcategories `A` to `F`. Thus the most basic masculine noun classes might be named `M 1A`, `M 3C`, etc.

Special characters are used to further characterise the different morphological types. For instance, types with an `*` indicate the presence of a fleeting vowel such as in *свекор* 'father-in-law (husband's father)' which has the instrumental *свекром*. Animacy is indicated in combination with gender, so that a class such as `MO 1*A` is masculine, animate, type 1, stress pattern A with a fleeting vowel.

For each resulting type, Zaliznjak's realisation tables specify how each morphosyntactic form is constructed, and we have used those tables as the basis of our hand-crafted DATR theory. Each type is represented by a node in the DATR inheritance hierarchy (see Figure 2)[2].

[2] In the node names, M stands for masculine, F for feminine and A for Animate.

In total, there are about a 100 noun classes per gender[3]. Each node contains definitions of morphosyntactic realisations specific to that noun class. Information that is shared between (or default for) classes is inherited from the parent node.

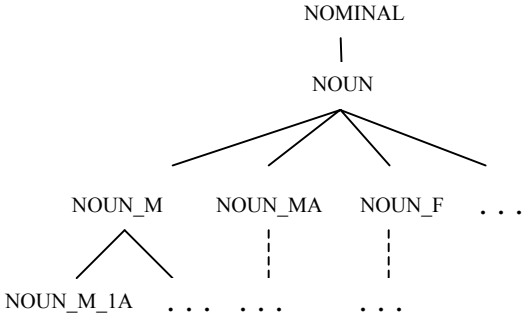


Figure 2. Extract of the DATR hierarchy

In order to make use of this lexicon, a lexical entry needs to inherit from the node representing its noun class and provide the specific morphotactic elements required for the class. So for example, if we know that *абажур* 'lamp shade' is a masculine noun of type 1A, a class which just requires the morphotactic element `stem`, then a possible DATR lexical entry for *абажур* would be:

```
АБАЖУР:
    <> == NOUN_M_1A
    <stem> == абажур.
```

This defines a DATR node called 'АБАЖУР', which specifies a `stem` feature 'абажур' and inherits everything else from class 'NOUN_M_1A'. From this definition, plus the hand-crafted theory, the standard inference rules of DATR allow all the relevant inflected forms to be derived[4]:

```
АБАЖУР:
    <mor sg nom> = абажур            <mor pl nom> = абажур ы
    <mor sg gen> = абажур а          <mor pl gen> = абажур ов
    <mor sg dat> = абажур у          <mor pl dat> = абажур ам
    <mor sg acc> = абажур            <mor pl acc> = абажур ы
    <mor sg instr> = абажур ом       <mor pl instr> = абажур ами
    <mor sg loc> = абажур е          <mor pl loc> = абажур ах
```

2.2.2 Automatic generation of lexical entries

In its electronic form, Zaliznjak represents each lexical entry as a text string of the sort given here for *абажур*:

```
АБАЖУР    0101 АБАЖ<УР М 1А
```

Here, the first item is the (uppercase) citation form of the word, the second is a line identifier (line 01 of 01 lines), the third is the word annotated with stress information, the fourth is gender/animacy information and the fifth morphological type[5]. Compilation of these entries into DATR nodes has two phases. Initially, we use regular expression search and substitute operations to rewrite the Zaliznjak source line into a DATR node definition containing essentially the same 'source' information. We then use DATR to compile this into a form

[3] We would like to emphasise that the DATR theory is written to reflect Zaliznjak's system and that the main goal has not been elegance and economy of representation. For theoretically-driven inheritance representations of Russian morphology using DATR see Corbett and Fraser (1993), Fraser and Corbett (1995), and Brown (1998).
[4] Note that stress is not currently indicated in the derived forms.
[5] Many of the entries are more complex than this in various ways – spread over several lines, with additional inflectional class information, alternative forms and comments etc..

suitable for accessing the morphological theory, exploiting the hierarchical structure of the theory to control how different noun classes are compiled.

For example, the entry for *аристократка* 'female aristocrat' has the following form:

```
АРИСТОКРАТКА 0101 АРИСТОКР<АТКА ЖО 3*А
```

This is first mapped into the DATR definition 'Z-АРИСТОКРАТКА', and from there compiled into 'CZ-АРИСТОКРАТКА':

```
Z-АРИСТОКРАТКА:
    <> == ZALNODE
    <index> == 30
    <src cit> == 'АРИСТОКРАТКА'
    <src str> == 'АРИСТОКР<АТКА'
    <src gen> == 'ЖО'
    <src cls> == '3*А'.
```

```
CZ-АРИСТОКРАТКА:
    <> == NOUN_FA_3*A
    <root_begin> == аристократ
    <root_end> == к.
```

In 'CZ-АРИСТОКРАТКА', the gender/animacy and class information have been combined (and transliterated to Latin script) to determine the noun class for this form. This class then also provides information about both the morphotactic components required and how to determine them from the citation form. In this case two components are specified (to allow for the insertion of a fleeting vowel in some forms) and notice also that the final vowel of the citation form is dropped completely (the final *a* of the nominative form is re-generated by the DATR theory). In total twelve different morphotactic configurations are defined, and their association with individual noun classes is controlled by the main inheritance hierarchy of the theory.

2.2.3 Combining theory and lexical entries

Finally we combine the compiled form of the lexical entry with the hand-coded theory and use this information to provide the morphosyntactic specifications and all the inflected forms for the automatically generated lexical entries as is illustrated here for the lexeme *аристократка*:

```
CZ-АРИСТОКРАТКА:
   <mor sg nom> = аристократ к а      <mor pl nom> = аристократ к и
   <mor sg gen> = аристократ к и      <mor pl gen> = аристократ о к
   <mor sg dat> = аристократ к е      <mor pl dat> = аристократ к ам
   <mor sg acc> = аристократ к у      <mor pl acc> = аристократ о к
   <mor sg instr> = аристократ к ой   <mor pl instr> = аристократ к ами
   <mor sg loc> = аристократ к е      <mor pl loc> = аристократ к ах
```

Notice that for most of these forms the inflectional suffix follows the concatenation of `<root_begin>` and `<root_end>`. However, in the genitive and accusative plural forms, a fleeting vowel, *o* in this case, is inserted between these two components.

2.3 Evaluation of the lexicon

To date, the hand-crafted DATR theory for Zaliznjak's morphological classes has been completed for the noun and adjective classes. Other classes are lower priority for our present project. Initial evaluation of the Zaliznjak DATR theory for nouns was undertaken by selecting a random sample of 500 lexemes of a wordlist of some 40,000 nominal headwords taken from Zalizjnak. The results are as follows:

Number of Zaliznjak entries	500
Number of DATR entries	499
Number of correct wordforms	5988
Number of incorrect wordforms	0

Of the 500 lexemes, one did not generate a compiled DATR entry, because the corresponding noun class was missing from the DATR theory (so it was not possible to compile from the DATR source form to the compiled form). All 5988 inflected forms for the 499 lexemes were generated and manually checked. No errors were found. Once the missing noun class was added all inflected forms for this lexeme were also generated correctly. We conclude from this that the automatic generation of lexical entries from Zaliznjak's source data works satisfactorily.

3 Lemmatisation

We take *lemmatisation* to mean the task of determining from a wordform the set of all possible analyses it has as a lemma plus morphosyntactic features. Lemmatisation is often the first step in *part of speech tagging*, where the second step is disambiguating the lemmatisation results by using the surrounding context of the wordform within a text. Such disambiguation is our long-term goal, but for the present paper we are not concerned with it, and consider only the analysis of wordforms in isolation.

Traditionally there are essentially three approaches (plus combinations of them) to the lemmatisation problem[6]:

1. **Wordform lookup** maintains a simple list of all possible forms, and their associated morphosyntactic analyses. Lookup is very fast, but the size of the wordform list can be very large, especially for languages with complex morphology. In addition, this approach is not robust – it has no strategy to deal with unknown forms or new usages.
2. **Hand-crafted rule-based lemmatisers** use a set of manually designed pattern matching rules to analyse the wordform. These are much smaller than wordform lists, typically quite fast in execution, and in principle robust, since they are generic in nature. But they are reliant on coding of rules to cover all cases which can be difficult to achieve by hand, especially in more complex languages and for irregular forms.
3. **Automatically-generated rule-based lemmatisers** use a morphological theory or corpus to derive rules for lemmatisation automatically. These will typically have more rules than a hand-crafted solution, but are likely to more systematically cover all cases, although they too are only as good as the morphological data they are derived from. In general, they lose a little on space (but not necessarily time, as their rules may be simpler to apply), but gain on robustness.

In the present context, we have the morphological resources of the Zaliznjak dictionary at our disposal, comprising a detailed morphological theory and a wordlist of some 40,000 (noun) headwords. Option 1 is a possibility since we can in principle generate a wordform list of around half a million forms from this resource, but we feel it is unlikely that this list alone would have sufficient coverage for practical corpus analysis. Option 3 offers a more interesting alternative, as the theory has far greater coverage of morphological types than the wordlist has of wordform tokens. If we can use the theory to generate lemmatisation rules we should be able to produce a very robust lemmatiser. In addition, as the morphological theory is extended or modified, it will be possible to derive a new lemmatiser automatically which takes advantage of the improvements.

3.1 Automatic generation of lemmatiser rules

Our approach to generating lemmatiser rules can be outlined as follows:

1. Extract a sample from Zaliznjak's dictionary containing one word for each distinct morphosyntactic type represented in the hand-crafted morphological theory.

[6] See Minnen, Carroll and Pearce (2001) for a good discussion of approaches to lemmatisation for English. For a morphological analyser for Russian, see Dialing: http://www.aot.ru/download.htm.

2. Use the sample to create dummy lexical entries whose ‘citation form’ is the string ‘987654321’.
3. Use the morphological theory to derive wordforms from these lexical entries: these ‘wordforms’ will contain a mixture of inflectional endings and digits derived from the dummy citation form.
4. Use these wordform patterns to derive wordform recogniser rules.
5. Compact the set of recogniser rules into a lemmatiser.

We will now illustrate these steps in more detail.

3.1.1 Extracting a morphologically representative sample

The morphology tables in Zaliznjak’s dictionary distinguish 290 basic noun classes, as described in Section 2.2.1 above. In addition there are 24 hybrid noun types, for which the base class and the declension type are different (for example, masculine nouns declining according to an adjectival pattern). Thus in total we have 314 noun classes with distinct morphosyntactic behaviour. One example of each class was manually extracted from the Zaliznjak source, and this set was used as our sample of lemma types.

3.1.2 Creating ‘dummy’ lexical entries

The framework described in Section 2 allows us to process the lemma sample right from source through source-node and compiled-node to listing of inflectional forms. During the process, the citation form is analysed into morphotactic components according to noun class (and to a limited extent, the particular properties of the citation form), and these components are used to construct the inflected forms. We can use this architecture to generate lemmatiser patterns by separating these two activities: the citation form of the lemma is still used to determine how the morphotactic components should be constructed, but the form used to construct them is replaced by the fixed string ‘987654321’. Using such a string has a number of useful properties:

1. In the inflected forms, it is easy to distinguish components derived from the ‘citation’ form from inflectional affixes, because the former are made up of digits, not letters.
2. Apart from the left-most component, (that is, the ‘main’ stem, for suffixational languages), the morphotactic components encode their expected length in characters within a wordform.
3. The actual digits occurring in the morphotactic components encode character position in the citation form. This information can be inverted to allow the lemmatiser to hypothesise citation forms from matched wordforms.

Using this technique the ‘compiled node’ form of ‘Z-АРИСТОКРАТКА’ would look like this:

```
CZ-АРИСТОКРАТКА:
    <> == NOUN_FA_3*A
    <root_begin> == 9876543
    <root_end> == 2.
```

From this definition, we see that in general, for nouns in this class, the `root_end` component consists of the second (from the right) character of the citation form, the `root_begin` component is the third from the right, plus everything to its right, and the right-most character of the citation form is not part of the real ‘stem’ at all (that is, it is discarded).[7]

[7] The use of ‘987654321’ is not intended to imply that all citation forms have exactly nine letters: the left-hand segment is always a placeholder for ‘all the rest’. The important thing is that there are enough numbers to cover all the possible characters that might be removed from the citation form. In the case of Zaliznjak, this is 7, for citation forms such as МЫШЬЮНОЧЕК, with singular instrumental form analysed as мышь + оночком.

3.1.3 Deriving wordform patterns

Deriving the patterns for each inflected wordform is a straightforward application of the DATR inference rules over the morphological theory, in exactly the same way as it would be for the original entry. But the resulting forms have numeric 'morphotactic' components instead of the original citation components:

```
CZ-АРИСТОКРАТКА:
   <mor sg nom> = 9876543 2 а        <mor pl nom> = 9876543 2 и
   <mor sg gen> = 9876543 2 и        <mor pl gen> = 9876543 о 2
   <mor sg dat> = 9876543 2 е        <mor pl dat> = 9876543 2 ам
   <mor sg acc> = 9876543 2 у        <mor pl acc> = 9876543 о 2
   <mor sg instr> = 9876543 2 ой     <mor pl instr> = 9876543 2 ами
   <mor sg loc> = 9876543 2 е        <mor pl loc> = 9876543 2 ах
```

Listings like this for all the lemma nodes contain all the information required to build the lemmatiser: the right hand sides of the '`mor`' equalities can be used as patterns to match against a word form (and to construct a possible citation form), the left-hand sides provide the morphological features, and the surrounding context provides the lemma type (via the node-name itself).

3.1.4 Creating wordform recognition rules

The wordform recognition rules are constructed by simple reorganisation of each wordform line. A line such as:

```
<mor sg nom> = 9876543 2 а
```

in the above listing becomes a rule such as:

```
[а  * **] → АРИСТОКРАТКА <mor sg nom>  98765432*
```

Here the square-bracketed expression is a pattern which matches a wordform from right to left. * matches any single character. ** matches the rest of the form. So this pattern matches any wordform that ends in 'а' and has at least one additional character. To the right of the arrow is the answer: this wordform could be an instance of the lemma type represented by 'АРИСТОКРАТКА', in the singular nominative form. The final component is a pattern to hypothesise the citation form of the word. Here the numbers refer to digit position *in the wordform* which should be used to construct the citation form. In this example all the letters up to letter 2 are used in their original places, and another unknown letter (represented by *) is on the end (recall that the morphological analysis threw away the last letter of the original citation form, which is why we cannot now hypothesise what it should be). Here is a slightly more complex example:

```
<mor pl acc> = 9876543 о 2

[* о **] → АРИСТОКРАТКА <mor pl acc>  98765431*
```

Here we are matching a form with a fleeting vowel 'о'. The citation form template again places letters '9876543' of the wordform in the same place in the citation form, but this time the second from the right is the *rightmost* letter of the wordform (index 1), and again the rightmost letter of the citation form is unknown.

From this example we see how citation templates are constructed. The numbers in the original lemma output represent the relationship between citation form and wordform by placing numbers representing position in the citation form into the appropriate position in the wordform. To construct the template, we invert this mapping, so placing numbers

representing position in the *wordform* into the corresponding position in the *citation form* template.

Using this technique we generate a set of 3451 wordform recognition rules directly and automatically from the morphological output of the set of lemma type nodes.

3.1.5 Compacting the lemmatiser

The set of rules created in the preceding step can in principle be used directly to create a lemmatiser, which operates simply by applying each rule in turn and collecting up the results for rules which match. However the ruleset is highly redundant – many of the rules have identical patterns, or patterns which are prefixes of other patterns. By combining such rules a much more efficient lemmatiser can be constructed. Two such compaction operations have been implemented[8]. Firstly, rules with identical patterns are combined into a single rule which returns all the results of the original rules as a list. So for example:

```
[* о **] → АРИСТОКРАТКА <mor pl acc>  98765431*
[* о **] → АРИСТОКРАТКА <mor pl gen>  98765431*
```

become:

```
[* о **] → АРИСТОКРАТКА <mor pl acc>  98765431* /
           АРИСТОКРАТКА <mor pl gen>  98765431*
```

Using this technique, the original 3451 rules reduce to 105 rules. The lemmatiser operates in the same way – it tests all the rules and collects up the results of all that match.

The second compaction is that any rule whose pattern extends another rule can be extended to return all the values the shorter pattern would have returned. For example:

```
[и  * **] → АРИСТОКРАТКА <mor sg gen>  98765432*
[и  м а * **] → АРИСТОКРАТКА <mor pl instr>  987654*
```

become:

```
[и  * **] → АРИСТОКРАТКА <mor sg gen>  98765432*
[и м а * **] → АРИСТОКРАТКА <mor pl instr>  987654* /
               АРИСТОКРАТКА <mor sg gen>  987654ам *
```

This does not reduce the number of rules, but if the rules are searched longest-pattern-first, then the lemmatiser can stop at the first successful match, since its result will subsume all shorter matches.

3.2 Reducing the number of analyses

The lemmatisation framework just described will effectively analyse the wordforms covered by its underlying morphological theory. However, because of the very fine-grained approach adopted by Zaliznjak's morphological analysis, it does return a rather large number of analyses for every form. As an indication of this, among the 3451 base rules generated, 186 match *any* wordform, that is, they analyse the wordform as an unaffixed stem. Thus the lemmatiser returns *at least* 186 answers for every wordform processed. This is perhaps a slightly surprising result, and certainly is not ideal for use with statistical part of speech taggers, where data sparseness would surely preclude training across the whole set of possible analyses. There are a number of possible explanations for this situation:

[8] An alternative approach to achieving these efficiency gains would be to view the search patterns as a finite state automaton, and minimise it using standard algorithms. In particular, if the automaton is a 'Moore machine' (that is, capable of emitting results in every state, not just final states), the effect of both compactions could be achieved, with a maximally efficient search of all patterns simultaneously. We have not yet attempted to implement this, although it would be straightforward to do so.

1. Zaliznjak's analysis may be too fine-grained for our purpose, making distinctions we are not really interested in (for example, between stress patterns not represented in textual documents).
2. Our encoding of Zaliznjak's analysis may be inadequate, for example by not capturing constraints on stem forms within classes.
3. The analysis is correct, but the idea of doing lemmatisation in isolation, rather than in context (that is, as part of the tagging process) is not practical for languages with complex morphology.

Our view is that there may be some truth in all these possibilities. Regarding the second point we would like in future work to encode additional constraints (such as whether a stem ends in a vowel or a consonant), and make use of wordlist data to restrict the possible analyses for known words, and the third point seems to have considerable potential for future research. But for the present, we restrict ourselves to consideration of the first option.

If Zaliznjak's analysis is too fine-grained, we need to identify something less fine-grained to replace it. Possible approaches include abstracting in the inheritance hierarchy (that is, ignoring the bottom layer of the inheritance tree, so that fewer distinctions are made), or simply omitting some of the information returned as a result (most notably the lemma type information). However the success of such approaches depends on the validity of the internal structure of the morphological theory, which may have not been designed with this purpose in mind. An alternative, more reliable, route is to leave the morphological theory intact, but to add to it a second lemma type classification which is less fine-grained. The inheritance structure allows us to specify such lemma types high up in the hierarchy, but also gives us the control we need to distribute types correctly across all the Zaliznjak classes. A good candidate less fine-grained lemma typology is the one developed within Network Morphology (Corbett and Fraser 1993, Fraser and Corbett 1995, Brown 1998), which distinguishes 6 lemma types relevant for the data considered here (`N_I, N_II, N_III, N_IV, N_ON_STEM, A_I`). We can distribute these correctly across the 314 Zaliznjak classes with just a handful of high-level specifications. If we use this classification instead of Zaliznjak's to distinguish our results, then the number of distinct analyses of an unaffixed stem drops from 186 to 8, and all other results are reduced correspondingly.

3.3 Lemmatiser evaluation

To evaluate the output of the lemmatiser we used the lemma typology developed in Network Morphology. A random sample of 100 wordforms was created from a randomly selected subset of 1000 lexical entries from Zaliznjak for which all the inflected forms were automatically generated. We then calculated the precision, recall and F-measure[9] ($\alpha = 0.5$) of the lemmatiser by determining the number of correct analyses (984 affixed and unaffixed; 672 affixed analyses only), wrong analyses (557 affixed and unaffixed; 69 affixed analyses only) and missing analyses (7 affixed and unaffixed; 7 affixed only). The results are as follows:

	Affixed and Unaffixed Analyses	**Affixed Analyses**
Precision	0.64	0.91
Recall	0.99	0.99
F-measure	0.78	0.95

At the moment, our lemmatiser overgenerates as basically no constraints on stems have been built in. For example, the fact that an unaffixed form of a Russian inflectable noun will generally end in a consonant has not been taken into account when generating the lemmatiser

[9] (Manning and Schütze 1999: 268-269).

rules. Thus, each wordform gets at least eight analyses as it can be an instance of one of the eight unaffixed forms provided by the Zaliznjak DATR theory. We have counted those unaffixed analyses for wordforms ending in a vowel as incorrect, which explains the lower value for precision when we take both affixed and unaffixed analyses into account. Similarly, a form such as *баталере*, the locative singular of *баталер* 'petty officer responsible for provisions', gets a possible analysis as a plural nominative of the Network Morphology Class I. Theoretically, this is a possible analysis, as Class I nouns such as *армянин* 'Armenian' get a nominative plural ending in *е*, but this is only the case for those Class I nouns ending in *янин* or *анин*. Again we have counted this as an incorrect analysis, even though the lemmatiser was not given this information. To improve the performance of the lemmatiser we might want to consider building in some of these constraints.

4 Future Work

This paper describes intermediate results in a programme of work that is in progress. In this section we briefly note the next steps for us, as part of our corpus analysis work for Russian, and also possible areas for more general future exploration.

We have already created a DATR fragment based on the adjectival classes in Zaliznjak (1977) and the next step will involve the automatic generation of lemmatisation rules for adjectives. Once this is completed we will combine the wordform recognition rules with a wordform list which can be generated from the combined theory and lexical entries as described in section 2. In other words we will combine the wordform lookup solution with the ruled-based lemmatiser solution in order to create a lemmatiser which is fast and robust, and does not overgenerate needlessly. We will also combine the Zaliznjak morphological lexicon with data from a manually validated lexicon of 1500 most frequent nouns (Brown, Corbett and Fraser 1995, Brown, Hippisley, Corbett and Fraser 1995), to improve accuracy, particularly for irregular forms. We are also interested in looking at *co-lemmatisation*, that is, combining information about several unknown inflected forms in a document to make the best hypothesis for a single lexeme that subsumes them all.

Our main aim with the lemmatisation work is to derive figures on the distribution of inflectional syncretism (grammatical ambiguity) in texts. Lemmatisation give us a measure of inflectional ambiguity for isolated words. The next step is to attempt disambiguation using the textual context, in a process not dissimilar to raditional part of speech tagging. We are currently exploring traditional and DATR-based methods for doing this.

In the longer term we are keen extend this work to languages which exhibit radically different kinds of morphology from Russian. The underlying approach to lemmatisation developed here is independent of the actual morphological theory, but at present only copes with suffixational morphology. A more sophisticated approach to wordform matching would be required to cope with other inflectional processes.

5 Conclusions

In this paper we have described an approach to lemmatisation of Russian nouns based on a large-scale inheritance-based morphological theory. The approach combines a hand-crafted implementation of Zaliznjak's basic theory with automatic derivation of lexical and lemma definitions, inflected forms, lemmatiser rules and finally the compacted lemmatiser itself. Although this work is in a preliminary stage, the results of evaluation are encouraging.

The benefits of basing this framework on an inheritance-based model of morphology are evident in every step of the process: the hierarchy is used to determine the morphotactic components of lexical entires, to determine inflected forms and hence lemmatiser rules, and to distribute Network Morphology class information across the Zaliznjak classes. Furthermore the use of the DATR representation language makes it easy to extend the framework to other word classes, and to add irregular and subregular cases to improve and fine-tune the coverage. After any such modification, a new lemmatiser can simply be compiled out, reflecting the content of the theory, without being burdened in any way by the overhead of default

inheritance. Finally, the lemmatisation framework is entirely modular, and could be used equally well for any other suffixational language for which a suitable morphological theory exists.

References

Brown, Dunstan, Greville Corbett and Norman Fraser. 1995. rusnoms.dtr – a fragment for the nominal system of Russian. Available from the DATR archive http://www.datr.org

Brown, Dunstan, Andrew Hippisley, Greville Corbett and Norman Fraser. 1995. rusnlex.dtr - lexicon of frequent Russian noun. Available from the DATR archive http://www.datr.org

Brown, Dunstan. 1998. From the General to the Exceptional: A Network Morphology Account of Russian Nominal Inflection. PhD thesis, University of Surrey.

Corbett, Greville G. and Norman M. Fraser. 1993. Network morphology: A DATR account of Russian nominal inflection. *Journal of Linguistics* 29. 113-42.

Dimitrova, Ludmila, Tomaž Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevič, Dan Tufis. 1998. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of COLING-ACL '98.* 315-319.

Evans, Roger and Gerald Gazdar. 1996. DATR: A Language for Lexical Knowledge Representation. *Computational Linguistics* 22. 167-216.

Evans, Roger, Carole Tiberius, Dunstan Brown and Greville Corbett. 2003. A large-scale inheritance-based morphological lexicon for Russian. In Tomaž Erjavec and Duško Vitas (eds.) *Proceedings of the Workshop on Morphological Processing of Slavic Languages*, 10th Conference of the European Chapter of the Association for Computational Linguistics, April 12-17. EACL: Budapest. 9-16. Also available as ITRI technical report ftp://ftp.itri.bton.ac.uk/reports/ITRI-03-02.pdf

Fraser, Norman M. and Greville G. Corbett. 1995. Gender, animacy and declensional class assignment: a unified account for Russian. In G. Booij and J. van Marle (eds.) *Yearbook of Morphology 1994*. Dordrecht: Kluwer. 123-150.

Ilola, Eeva & Mustajoki, Arto. 1989. Report on Russian Morphology as it appears in Zaliznyak's Grammatical Dictionary. Helsinki: Helsinki University Press.

Lönngren, Lennart (ed.) 1993. *Častotnyj slovar´ sovremennogo russkogo jazyka*. Uppsala: Uppsala University. (=Studia Slavica Upsaliensia 32).

Maier, I. 1994. Review of Lönngren (ed.) Častotnyj slovar´ sovremennogo russkogo jazyka. *Rusistika Segodnja* 1. 130-136.

Manning, Christopher. D. and Hinrich Schütze 1999. *Foundations of statistical natural language processing.* MIT Press. Cambridge: MA.

Minnen, Guido, John Carroll and Darren Pearce 2001 `Applied morphological processing of English', *Natural Language Engineering,* 7(3). 207-223.

Russian Newspaper Corpus. 1996. V.V. Vinogradov Institute of Russian Language of the Russian Academy of Sciences. Available at: http://irlras-cfrl.rema.ru:8100/newspap.htm

Vitas, Dusko. 2001. Intex and Slavonic Morphology. In *Proceedings of the 4th Intex workshop*.Bordeaux. Available online at: http://grelis.univ-fcomte.fr/intex/downloads/Dusko_Vitas.pdf

Zaliznjak, A. A. 1977. *Grammatičeskij slovar' russkogo jazyka*. Moscow: Russkij jazyk.