University of Brighton

ITRI-03-02

A large-scale inheritance-based morphological lexicon for Russian

Roger Evans, Carole Tiberius, Dunstan Brown and Greville Corbett

May, 2003

Also published in Proceedings of the EACL'3 Workshop on Morphological Processing of Slavic Languages

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

A large-scale inheritance-based morphological lexicon for Russian

R. Evans
ITRI
University of Brighton
roger.evans@itri.bton.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 the mapping of Zaliznjak's (1977) morphological classes into the lexical representation language DATR (Evans and Gazdar 1996). On the basis of the resulting DATR theory a set of fully inflected forms together with their associated morphosyntax can automatically be generated from the electronic version of Zaliznjak's dictionary (Ilola and Mustajoki 1989). From this data we plan to develop a wide-coverage morphosyntactic lemmatizer and tagger for Russian.

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. For this purpose, we will use two corpora, the Uppsala corpus (Lönngren 1993, Maier 1994) and a corpus of Russian newspaper texts from the late 1990's, for which we require detailed morphosyntactic annotation. However, suitably annotated versions of these corpora are not yet freely available and corpus analysis tools for Russian in general are scarce.[1]

We have chosen, therefore, to develop our own lemmatization and tagging technology, based on the electronic version of Zaliznjak's (1977) dictionary (Ilola and Mustajoki 1989), combined with a more detailed and validated hand-crafted analysis of 1500 most frequent noun lexemes (Brown, Corbett, and Fraser 1995; Brown, Hippisley, Corbett and Fraser 1995). In this paper we describe the first step in this process: mapping the basic Zaliznjak data into a hierarchical lexical database implemented in DATR.

1.1 The Zaliznjak dictionary

Zaliznjak (1977) is a reverse dictionary in book form, dealing primarily with Russian inflectional morphology. For each of the almost 100,000 lexical entries, indexes refer the reader to declension types and conjugations, together with stress patterns. Other symbols indicate subregularities and irregularities. As the dictionary uses such indicators, it gives explicit information about every inflectional form and stress. Ilola and Mustajoki (1989: 1-5) describe how the material was adapted for computer use.

Zaliznjak's dictionary has been the starting point for a number of applications. Anciaux (1991) made use of it in the creation of a spell-checker for Russian, and Pavlova, Pavlov, Sproat, Shih and van Santen (1997) used the electronic version to create language-specific tables to fit into the modular architecture of the Bell Laboratories Text-to-Speech system. Brown, Corbett and Fraser (1995) and Brown, Hippisley, Corbett and Fraser (1995) created a DATR lexicon of the 1500 most frequent noun lexemes from Zasorina (1977). The derived forms from this inheritance-based lexicon were all checked manually against Zaliznjak. The forms are represented in a phonological transcription, together with stress information (Brown, Corbett, Fraser, Hippisley and Timberlake 1996). An updated version of this lexicon was used in Brown

[1] For an indication of what is available see: http://talrusse.free.fr. For natural language processing of Slavic languages in general see for example work on the MULTEXT-EAST project by Dimitrova, Erjavec, Ide, Kaalep, Petkevič and Tufis (1998) and work within the INTEX system by Vitas (2001).

(1998) to compare different morphological theories.

1.2 Outline

The paper is structured as follows. Section 2 describes the general principles of the mapping and examples. In Section 3, we discuss the technical framework of our approach and the issues and problems that arise. In Section 4 we discuss the current status of the mapping and principal areas for further development including our approach to lemmatization and tagging. Section 5 concludes the paper.

2 Mapping Zaliznjak into DATR

2.1 The overall approach

In book form, Zaliznjak's dictionary has two parts. The first is a set of tables identifying morphosyntactic classes and defining the realization of morphological features with them. The second is a listing of lexical entries, each followed by an index referring to a table in the first part which gives the paradigm for this particular type. 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.

		1A
M		**завод**
Sg	nom	завóд
	gen	завóда
	dat	завóду
	acc	завóд
	instr	завóдом
	loc	завóде
Pl	nom	завóды
	gen	завóдов
	dat	завóдам
	acc	завóды
	instr	завóдами
	loc	завóдах

Table 1. Zaliznjak's paradigm for masculine inanimate nouns of type 1A.

The electronic form contains just the set of lexical entries (101401 lines, 98729 lexical entries). Thus our mapping process has two distinct components:

1. manual construction of a DATR representation of the morphosyntactic class and realization information from the printed paradigm tables;

2. automatic construction of the individual lexical entries from the electronic dictionary data.

In practice we also introduce a third component, interfacing between the morphosyntactic classes and the automatic entries. As we discuss below, this gives us increased flexibility in the way we interpret the Zaliznjak data.

The information in Zaliznjak's dictionary includes a fair number of subregular and idiosyncratic cases. The target representation, DATR, is specifically designed to support such situations, providing concise representation of hierarchically organized lexicons containing generalizations and exceptions. We have already a formal theoretical model of Russian morphology (Corbett and Fraser 1993; Brown, Corbett, Fraser, Hippisley and Timberlake 1996; Brown 1998) which underlies our approach. In addition, as we have a frequency based resource to check against (Brown, Corbett and Fraser 1995; Brown, Hippisley, Corbett and Fraser 1995), we are in a good position to check the accuracy of our automatic creation of lexical entries for the high frequency, least regular cases. The same framework can also be used to capture generalisations across languages (cf. Cahill and Gazdar 1999; Tiberius 2001), but this is not our current goal.

2.2 The hand-crafted realization component

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 mor-

phological 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 characterize 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.

This information for each lexical entry is used to refer to a table at the beginning of the dictionary which gives an example of the inflectional forms. These tables form the basis of a hand-crafted DATR theory in which each type is represented by a node in the DATR inheritance hierarchy. This results in a hierarchical structure of noun classes, part of which is shown here:

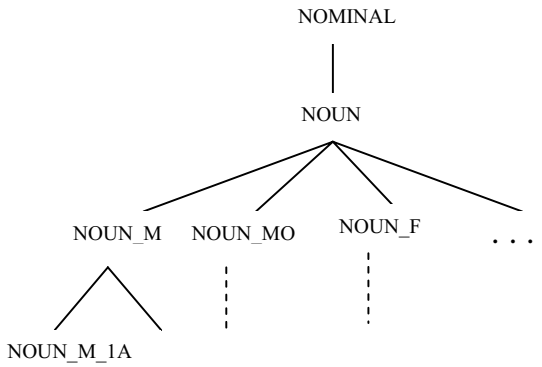

Figure 1. Extract of the DATR hierarchy

In each node, definitions of morphosyntactic realizations specific to that noun class are given. Information that is shared between (or default for) classes is inherited from the parent node. A small fragment of the theory is provided here:[2]

```
NOMINAL:
   <mor> == "<stem>" "<mor suffix>".

NOUN:
    <> == NOMINAL
    <mor suffix pl dat> == ам
    <mor suffix pl instr> == ами
    <mor suffix pl loc> == ах.

NOUN_M:
    <> == NOUN
    <syn gender> == masculine
    <syn animacy> == inanimate.

NOUN_M_1A:
    <> == NOUN_M
    <mor suffix sg nom> == Null
    <mor suffix sg gen> == а
    <mor suffix sg dat> == у
    <mor suffix sg acc> ==
          "<mor suffix sg nom>"
    <mor suffix sg instr> == ом
    <mor suffix sg loc> == е
    <mor suffix pl nom> == ы
    <mor suffix pl gen> == ов
    <mor suffix pl acc> ==
          "<mor suffix pl nom>".
```

Here, `NOMINAL` defines the morphotactics of Russian nominals (nouns and adjectives), comprising a stem followed by a suffix that realizes the morphological features. `NOUN` inherits this definition and defines three plural suffixes that are generally shared between nouns, `NOUN_M` adds specific syntactic features and finally `NOUN_M_1A` fills out the rest of the possible suffixes. Notice that `<stem>` is not defined in this theory – it will be determined on a per-lexical entry basis from the automatically generated entries described below. Notice also the syncretic definitions for suffixes associated with `sg acc` and `pl acc` in terms of their nominative counterparts for inanimate nouns – for a more detailed discussion of the techniques used for representing such a syncretism, see Corbett and Fraser (1993:131) and Brown (1998:154-155).

Classes of feminine and neuter nouns are handled similarly; for feminine nouns, eight types and nine different stress patterns are identified, whereas for neuter nouns eight types and six stress patterns are distinguished. In addition, most of these types are found with both animate and inanimate nouns, and in the DATR theory, two noun classes are distinguished for each type occurring with both animate and inanimate nouns. In total

[2] The DATR code is slightly simplified for expository purposes. Note that the code is written to reflect Zaliznjak's system. 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).

about 100 different noun classes are distinguished per gender in the DATR theory.

In order to make use of this theory, a lexical entry needs to inherit from the node representing its noun class and provide the specific morphotactic elements associated with the class. So for example, a possible definition for *абажур* 'lamp shade' might be:

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

From this definition, plus the preceding example fragment, the standard inference rules of DATR allow all the relevant inflectional forms to be derived:

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

Note that stress is not currently indicated in the derived forms. Our research involves the morphosyntactic analysis of written text which generally does not mark stress. However, as the distinctions related to stress that are made in Zaliznjak (1977) have been kept in the DATR theory, the stress patterns can easily be used in our analysis of syncretism and frequency.

2.3 Automatic generation of lexical entries

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

```
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.

However, inevitably, many of the entries are more complex than this in various ways:

1. Entries can spread over several lines, requiring textual concatenation of just the parts following the line identifier information to build the complete entry.
2. Where inflectional class does not correspond to gender/animacy, it may be specified separately between angle brackets. For example, *дедушка* `МО <ЖО 3*А>` 'grandfather' is a masculine noun which declines as a feminine noun of type 3*A.
3. Alternative values for stress patterns and sometimes classes may be present between square brackets.
4. Additional annotations indicate second locative, second genitive, pluralia tantum, irregular forms, etc.
5. Additional comments may be present enclosed in parentheses.
6. Other punctuation (commas etc.) may or may not be present.

In order to deliver lexical entry information in the form required by the hand-crafted theory, this data needs to be parsed and interpreted into the kind of format we saw above. A standard approach to this task is to use regular expression search and substitute commands to incrementally rewrite the data strings into a more uniform format and ultimately into the required input. However, DATR itself also provides powerful string-rewriting functionality, particularly suited for dealing with awkward exceptional cases, but less efficient for more routine rewriting.

The approach we have taken strikes a balance between these two technologies. Initially we use regular expression rewriting to achieve a basic

parse of the input data: joining multiple lines together, removing duplicate spaces, isolating various bracketed expressions, parsing the remaining fields and finally mapping into a DATR definition for each entry. However this DATR definition is very surface-oriented – little more than a basic segmentation of the input data. This process can be carried out completely automatically with a fairly high accuracy. But in order to link such entries to the core morphological classes, further interpretation of the data fields identified is necessary, and this is achieved dynamically in DATR.

For example, a typical simple lexical entry is the lexeme *аристократка* 'female aristocrat'. Its entry in Zaliznjak is:

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

In the first phase of processing, this is mapped via regular expression search and substitute into a DATR node definition as follows:

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

This node is an instance of the predefined node ZALNODE with index number 30 (meaning simply that it was the 30th node to be processed in this batch). The <src txt> feature (omitted due to lack of space) is the whole original source string, and the other features provide the key components of the entry (cit – citation, str – stressed, gen – gender/animacy, cls – class).

This is the 'surface level' representation of the lexical entry. The DATR node ZALNODE interprets this information to define implicitly a 'deep' representation as required by the morphological classes, roughly equivalent to this:

```
Z-АРИСТОКРАТКА:
    <> == NOUN_FO_3*A
    <root_begin> == аристократ
    <root_end> == к.
```

Here, the gender/animacy and class information have been combined (and transliterated to Latin script) to determine the declension class for this form. The stem forms for this class have been determined from the citation form (the morphotactic specification for NOUN_FO_3*A indicates what components are required – different from the simpler NOUN_M_1A case above, to allow for the possible insertion of a fleeting vowel).

ZALNODE does not actually create a new node definition for the deep representation. Rather, appropriate values for deep features are calculated dynamically when the declension class code requests them, by rewriting and transforming the values provided by the surface form definitions.

The overall effect is that, just as we saw previously, the declension class definitions can use this information to provide the syntax and all the inflected forms for this word:

```
Z-АРИСТОКРАТКА:
    <syn gender> = feminine
    <syn animacy> = animate
    <mor sg nom> = аристократ к а
    <mor sg gen> = аристократ к и
    <mor sg dat> = аристократ к е
    <mor sg acc> = аристократ к у
    <mor sg instr> = аристократ к ой
    <mor sg loc> = аристократ к е
    <mor pl nom> = аристократ к и
    <mor pl gen> = аристократ о к
    <mor pl dat> = аристократ к ам
    <mor pl acc> = аристократ о к
    <mor pl instr> = аристократ к ами
    <mor pl loc> = аристократ к ах.
```

For most of these forms the inflection follows the value of <root_begin> and <root_end>. Notice, however, that in the genitive and accusative plural forms, a fleeting vowel, *о* in this case, is inserted between these two components.

An example of a more complex lexical entry is *армеец* 'soldier' which is a masculine animate noun of type 5*A. This noun has a fleeting vowel which appears in the nominative singular *армеец* (phonologically *armejec*). The writing system also indicates the presence of the phoneme /j/ by the use of *й* in the other case and number combinations. To deal with this allomorphy, the DATR node ZALNODE introduces two values for

<root_begin> in the 'deep' representation of this lexical entry, one which is used in the nominative singular, i.e. *арме*, and one which is used for all other cases, i.e. *армей*. The 'deep' representation for *армеец* looks roughly like this:

```
Z-АРМЕЕЦ:
    <> == NOUN_MO_5*A
    <root_begin 1> == арме
    <root_begin> == армей
    <root_end> == ц.
```

3 The technical framework

The key technical challenge of this exercise was actually rather mundane: we needed to find an environment or set of environments that would allow us to do all the processing required (manual editing, regular expression search and substitute, DATR compilation and dumping) with data that included both Latin and Cyrillic script. In addition, we wanted the resources we created to be maximally reusable in other contexts, so a solution in line with agreed standards was highly desirable.

To achieve these goals, we adopted Unicode as the standard representation for all our data, and identified or adapted tools to work with data in that form. Furthermore we used the simplest encoding of Unicode in data files, the "ucs2" encoding, which stores each 16 bit Unicode character simply as two bytes of data. This is not as compact as other encodings (such as "utf8") but is supported by a wider range of applications, in particular Microsoft Wordpad.

3.1 A Unicode version of Zaliznjak

It is a fairly straightforward task to convert the transliteration used in the electronic form of Zaliznjak to Unicode Cyrillic, using Microsoft Word macros. Disambiguation of the hard and soft signs is required for the first field, (the index word field), as the + character is used for both symbols. However, the third field differentiates the hard and soft sign and, as the number of lexical items written with a hard sign is not great, it is a trivial task to check these. The resulting files are then saved as plain text (ie "ucs2") Unicode files.

3.2 DATR and Unicode

The DATR compiler used for this project was the Sussex/Brighton DATR compiler, which is written in Prolog. The DATR compiler inherits its character-level processing from the underlying Prolog compiler, so in order to process Unicode DATR it was simply necessary to run it in a Prolog system capable of handling Unicode, and modify it slightly to detect when it was given a Unicode file as input. This was achieved using Poplog Prolog,[3] plus a customized version of Sussex/Brighton DATR (soon to be released as version 2.10). This version also includes new support for batch mode processing of DATR theories and a number of compiler enhancements for compiling larger DATR theories.

3.3 Editing and search and substitute in Unicode

Unicode files stored in "ucs2" encoding can be conveniently viewed and edited using Microsoft Word or Wordpad, the latter being more straightforward for the simple text-editing requirements of most of the data files involved here. The automatic rewriting of Zaliznjak entries required a more sophisticated regular expression engine, which we obtained by adapting the Poplog editor's regular expression functionality to work with Unicode. These functions are particularly powerful in allowing multi-line regular expression matching, so that one can match patterns spanning several lines (such as Zaliznjak data continuation lines) and rewrite them to a single line. Limited manual editing of Unicode using the Poplog editor is also possible: it can manipulate arbitrary Unicode data, but its ability to display non-Latin data is platform dependent, and on our platform (Windows 2000) all the Cyrillic characters were displayed as '?'.

4 Current status and future work

The system described in this paper is still very much work-in-progress. The core technologies and

[3] See http://www.cs.bham.ac.uk/research/poplog/freepoplog.html. Unicode support is only available in version 15.53, although currently it is completely undocumented.

structures of the approach have been developed and validated as a viable approach. Population and validation of the data is an on-going process, the current state of which can be summarized as follows:

1. The hand-crafted DATR theory for Zaliznjak's morphological classes has been completed for the noun classes, with adjective classes next to be done. Other classes are lower priority for the present project.
2. Automatic compilation of all 98729 Zanliznjak entries into 'surface' DATR nodes is complete but not validated.
3. Processing of a sample set containing 2062 entries has been undertaken with the following (not fully validated) results:

No. of Zaliznjak entries	2062	100%
No. of DATR nodes	2000	97%
Nodes identified as nouns	1192	60%
Nouns successfully classified	1066	89%

Principal areas for further development include:

1. Completion and validation of noun entries
2. Extension to adjectives (and possibly verbs)
3. Integration of data from the manually validated lexicon of 1500 most frequent nouns, to improve accuracy, particular for irregular forms.
4. Development of a lemmatizer and tagger for Russian using this data.

The last point here deserves further expansion. As we discussed in the introduction, the DATR encoding of Zaliznjak is in part the first step towards lemmatization and tagging technology for Russian. We distinguish **lemmatization,** that is identifying all possible lemmas (plus morphosyntactic features) for a word, which can be carried out on the word in isolation, from **tagging**, that is, identifying the *most likely* lemma (plus features) for a word in context. The primary aim of the project of which this work is a part is to explore ambiguity in lemmatization and its relationship to frequency. For this a high quality lemmatizer is essential.

In principle, once we have a complete set of inflected forms, we could automatically compile it into a lemmatizer. However such a lemmatizer would be extremely cumbersome to produce and use, contain much redundancy and be quite incapable of coping with unknown forms. The approach we intend to take will exploit the hand-crafted components of the framework to the full, using them to construct recognisers for suffixes (and for verbs, prefixes) and identify potential roots, and then using the full lexicon to filter and validate the resulting candidate analyses (we expect the recognition process to overgenerate solutions). This will be more compact, probably faster, and able to cope with unknown root forms.

Beyond such a lemmatizer, we are currently investigating how to combine inheritance-based lexical representation with traditional part-of-speech tagging technology, and hope to apply this work to the Zaliznjak data, to deliver a high quality detailed morphosyntactic tagger for Russian texts.

On the more technological front, current plans include:

1. Consolidating Unicode support in DATR (extending to the Sicstus Prolog version, supporting other file encodings).
2. Packaging key technologies for wider use.
3. Delivering the whole Zaliznjak lexicon as an XML-based DATR database.

5 Conclusions

Zaliznjak's dictionary, both in its book form and electronic version, has proved an invaluable tool. In this paper we have shown how the classes from Zaliznjak can be mapped into a DATR representation. This representation is a structured lexicon from which we can derive all of the associated forms for the entries in Zaliznjak. As well as constituting a valuable computation resource for Russian in its own right, our next step will be to use this lexical database as the foundation for high quality lemmatization and morphosyntactic tagging software for Russian text.

Acknowledgements

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.

Availability

At the time of writing, the Zaliznjak data files are still work in progress, and the tool adaptations (to DATR, Poplog etc.) are still custom extensions. However, it is our intention to make these resources publically available, as far as is consistent with existing licences etc., in the near future.

References

Anciaux, Michele. 1991. Word-form Recognition and Generation: A Computational Approach to Russian Morphology. PhD dissertation, University of Washington.

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, Greville Corbett, Norman Fraser, Andrew Hippisley and Alan Timberlake. 1996. Russian noun stress and network morphology. *Linguistics* 34. 53-107.

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

Cahill, Lynne and Gerald Gazdar. 1999. The POLYLEX architecture: multilingual lexicons for related languages. *Traitement Automatique des Languages*, 40(2):5-23.

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.

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.

Pavlova, E., Y. Pavlov, R. Sproat, C. Shih and J. van Santen. 1997. Bell Laboratories Russian Text-to-Speech System. In G. Kokkinakis, N. Fakotakis, E. Dermatas (eds.) *Eurospeech '97 Proceedings*. Volume 5. 2451 – 2454.

Tiberius, Carole. 2001. Architectures for Multilingual Lexical Representation. PhD Thesis, ITRI, University of Brighton.

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.

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