RUSSIAN MORPHOLOGICAL ANALYSIS

Serge A. Yablonsky

Department of Computing, St.–Petersburg University of Transport

Russicon Company

Kazanskaya 56, Ap. 2

190000 St.–Petersburg, Russia

e-mail: root@russicon.spb.su; tel/fax: +7-812-312-72-13

Abstract

In this paper the approach to the organization of Russian inflexion morphologic model and its application for the Russian language morphological analysis and disambiguation are described. We are concerned with the pos tagging of 150-million-word Russian corpora. The approach is particularly dependent on the language processor Russicon, and on wide usage of Russicon's electronic dictionaries.

1. Introduction

Up-to-date language technologies contain efficient morphological analyzers for Romance, Germanic (Karttunen, 1983; Karttunen, Koskenniemi, Kaplan, 1987; Varile, Zampolli, 1996; Zaenen, Uszkoreit, 1996) and some Slavic (Chanod, 1997) languages. In the last 15 years Russian computational morphology has advanced at a great rate from first quite restricted systems towards large-scale practical morphological analyzers (Ashmanov I., 1995; Belonogov, Zelenkov, 1989; Belyaev, Surcis, Yablonsky, 1993; Bolshakov, 1990; Mikheev, Liubushkina, 1995; Segalovich, 1995). Now Russian word-form morphological analyzers are able to:

- classify an input string as a valid word of the supported lexicon and categorize it morpho-syntactically: part-of-speech (pos) category, number, gender, etc.
- generate a form of a word in accordance with certain morpho-syntactic features.

Although all these systems have an impressive lexicon (more then 1,500,000 word-tokens and 80,000 – 200,000 word paradigms) and are computationally efficient, few of them have user friendly interface to extend their lexicon with new words (Mikheev, Liubushkina, 1995; Yablonsky, 1990, 1998) and robust analysis of unrestricted Russian texts.

Russian part-of-speech (pos) tagging provides developing of disambiguation algorithms on the top of morphological text analysis. Alternate morphological analyses occur because of high categorial homonymy of inflective language. Only the sentential context used by disambiguation normally decides which analysis is appropriate (Zaenen, Uszkoreit, 1996).

In this paper the approach to the organization of Russian inflexion morphologic model and its application for the Russian language morphological analysis and disambiguation are described. Our purpose is to mark each word of 150-million-word Russian corpora (Yablonsky, 1998, a; 1999) with a pos and lexical categories tags.

The approach is particularly dependent on the language processor Russicon, and on wide usage of Russicon's electronic dictionaries (Belyaev, Surcis, Yablonsky, 1993; Yablonsky, 1990, 1997, 1998).

2. General model of inflection morphology

For the formal description of *inflection morphology model* the set theory is used. It is one of the best ways for description of Russian inflection morphology (Bider, Bolshakov, 1976; Kulagina, 1986).

Here we present the general set model that permits to define mostly all sides of inflection morphology. It is realized in the morphological analyzer of the Russicon language processor (Yablonsky, 1990; Belyaev, Surcis, Yablonsky, 1993; Yablonsky, 1998). In this model some concepts are introduced for the first time and other have new or more full meaning.

2.1. Definitions

Lexicon of the inflective language is named Q = $\{q_1, q_2, \dots, q_{Nq}\}$, where q_i, i = $\overline{1,Nq}$ — legal sequence of alphabet L characters named as a *word*. Set Q is an infinite denumarable set, because lexicon is permanently enlarging. Denurable system W of in general intersecting finite subsets W_i named as lexemes is given in Q, $\bigcup_{Wi \in W} Wi$ = Q. In the unstrict, lexeme W_I = $\{w_{I,1}, w_{I,2}, \dots, w_{I,Nwi}\}$, where $w_{I,j} \in Q$, j = $\overline{1,Nwi}$, is a set of words that has the same interpretation and etymology. Words $w_{I,j} \in W_I$ are named as *word forms or inflective forms of lexeme W_i*.

Let us introduce for each inflection form w of lexeme W_I number N_{WIS} from the interval 1÷|w|. The combination of a surface form and its analysis as a canonical form and inflection is called a *lemma*. *Paradigm* of the word is composed of all inflective forms of one lemma.

For every word one or several paradigms could be mapped:

$$W_i \rightarrow \overline{\beta}_m (\beta_{m,1}, \beta_{m,2}, \dots, \beta_{m,P}).$$

The number of word forms P for inflective language is rather high (P >> 1) for some parts of speech. For example, in Russian language P > 100 for verbs.

The part of the word including prefix(es), root and suffix(es) is called *word inflective stem (WIS).* The length L_{WIS} of WIS is calculated L_{WIS}= |WIS|=L_{WORD}-L_{END}, where L_{WORD} is the length of the word and L_{END}– the length of the ending.

For the character sequences WIS = (b_1, … ,b_{Nwis}), and INFL = (b_{Nwis+1}, … ,b_{Nb}) – *ending of inflection form (inflection),* condition w = WIS ⊕ INFL is fulfilled. Operation of two lists concatenation is denoted by '⊕'; |WIS| = L_{WIS} = 1÷|w|; |INFL| = 0÷(|w|-1). Null length inflection is called *empty* and is denoted by '+', and corresponding flexion is denoted by '–'. WIS of the lemma is denoted by WIS*.

The part of the word including prefix(es) and root is called *word formative stem (WFS).* The length L_{WFS} of WFS is calculated L_{WFS} = |WFS|=L_{WORD} - L_{SUF} - L_{END} = L_{WIS} - L_{SUF} = L_{PREF} - L_{ROOT}, where L_{SUF} is the length of the suffix(es), L_{PREF} – the length of the prefix(es), L_{ROOT} – the length of the root(s). The range of the introduced parameters are: L_{ROOT} = 1÷ L_{WIS}, L_{PREF} = 0÷(L_{WIS} – 1), L_{SUF} = 0 ÷ (L_{WIS} –1), L_{WFS} = 1÷ L_{WIS}.

The morphological analyzer has two main parts:

- dictionary with linguistic knowledge of the language;
- program realization of morphologic model's algorithms.

In general dictionary is represented in such form:

$$V_i = \{W_i, f_i\}, \quad i = 1 \mid N_v,$$

where W_i = (a_1, a_2, … , a_{Li}) — lexical part of dictionary's article: the word or phrase, composed from the alphabet characters A = {a_S : s = 1, … , N_a}; tag part f_i = (f_1, f_2, … , f_k) — subset of tags from the set F = { f_r : r = 1, … , N_f}, N_v — number of the words (word-tokens) in the dictionary, for large-scale dictionaries of inflective languages usually $N_v > 1500\,000$.

2.2. Morphological analysis model

Let $H = \{ h_1, h_2, \dots , h_{Nh} \}$ be the set of part-of-speech (pos) categories and $P = \{ p_1, p_2, \dots , p_{Np} \}$ — lexical categories (LC) of gender, number etc. Each element $p_i \in P$, where $i = \overline{1, Np}$, represents the set of concrete realizations of lexical category $p_i = \{p_{i,1}, p_{i,2}, \dots ,p_{i,Ni}\}$. Let us chose one element in P (for definiteness p_1) named *type* and denoted by T ($T = p_1, T \in P$), $T = \{t_1, t_2,\dots, t_{Nt}\}$.

For example, Russian language model includes: H[1] = {h1 = "noun", h2 = "adjective", h3[2] = "verb", h4 = "particle", h5 = "parenthetic word", h6 = "modal word", h7 = "adverb", h8 = "conjunction", h9 = "interjection", h10 = "preposition", h11 = "abbreviation", h12 = "unit of measure", h13 = "pronoun", h14 = "numeral", h15 = "adverbial participle“, h16 = "composition or special prefix"}. P = {p1 = "case", p2 = "gender", p3 = "number", p4 = "time", p5 = "person", p6 = "degree", p7 = "voice", p8 = "aspect", p9 = "mood", p10 = "form", p11 = "transitivity", p12 = "reflexive", p13 = "animate"}, where p1 = {"nominative", "genitive", "dative", "accusative", "instrumental", "prepositional"}, p2 = { "masculine", "feminine", "neuter", "masculine/feminine"}, p3 = {"singular", "plural "}. p4 = {"present", "past", "future", " present / future "}, p5 = {"1st person", "2nd person", "3rd person "}, p6 = {"superlative", "comparative"}, p7 = {"active ", "passive"}, p8 = {"imperfective", "perfective", "perfective and imperfective"}, p9 = {"indicative", "imperative"}, p10 = {"full", "short (predicative)", "infinitive"}, p11 = {"transitive ", "intransitive"}, p12 = {"reflexive", "irrevocable"}, p13 = {"animate", "inanimate"}.

We take that $(\forall h_k : h_k\in H)\ (\exists T_k : T_k \subset T, T_k \neq \varnothing)$, i.e. at least one type exists for each part of speech and is named *ordinary*, and also $T_k = \{t_{k,1}, t_{k,2}, \dots , t_{k,Nk}\}$, where $k = \overline{1, Nh}$.

For example, in Russian for h_1 = "noun", T_1 = {"ordinary", "invariable", "substantival"}. For (hk, tk,j) = ("noun", "ordinary"): P1k,t = {"gender"}, P2k,t = {"case","number"}, P3k,t = {"animate"}, Xk,t = (X*k,t = {"nominative case", "singular number"}, {"genitive case", "singular number"}, {"dative case", "singular number"}, {"accusative case", "singular number"}, {"instrumental case", "singular number"}, {"prepositional case", "singular number"}, {"nominative case", "plural number"}, {"genitive case", "plural number"}, {"dative case", "plural number"}, {"accusative case", "plural number"}, {"instrumental case", "plural number"}, {"prepositional case", "plural number"}, {"2-nd genitive case", "singular number"}, {"2-nd instrumental case", "singular number"}, {"2-nd prepositional case", "singular number"}, {"2-nd accusative case", "singular number"}, {"2-nd accusative case", "plural number"}).

Then

$$(\forall h_k \forall t_{k,j} : h_k\in H, t_{k,j}\in T_k, k=\overline{1, Nh}, j=\overline{1, Nk})$$
$$(\exists\ P_{k,t} : P_{k,t}\subset P, \bigcup_{k=1}^{Nh} \bigcup_{t=1}^{Nk} P_{k,t} = P),$$

i.e. for each part of speech exists it's own, may be empty, set of LC. Elements of $P_{k,t}$ are named LC of t-type h_k. For all $P_{k,t}$ exists partition on three nonoverlapping and may be empty subsets, named $P^1_{k,t}$, $P^2_{k,t}$, $P^3_{k,t}$:

$$(\forall P_{k,t} : P_{k,t}\subset P)\ (\exists P^1_{k,t}\exists P^2_{k,t}\exists P^3_{k,t} :$$
$$P^1_{k,t} \cup P^2_{k,t} \cup P^3_{k,t}=P_{k,t}, P^1_{k,}t \cap P2k_{,t} \cap P^3_{k,t} = \varnothing).$$

Elements of $P^1_{k,t}$ are named as *ordinary LC*, elements of $P^2_{k,t}$ — *special LC*, elements of $P^3_{k,t}$ —*individual LC* of t-type $h_{k.}$ For each $P_{k,t}$ set, if $P_{k,t} \neq \varnothing$, there exists ordered sequence of sets $X_{k,t} = (X^1_{k,t}, X^2_{k,t}, \dots , X^{Nxkt}_{k,t})$. That is, if

[1] Our model for Russian slightly differs from the classic: we include in the sets H and P some additional elements.
[2] In the paradigm of the verb we include participle and adverbial participle.

$P^2_{k,t} = \{p^2_{k,t,1}, p^2_{k,t,2}, \dots, p^2_{k,t,Nkt2}\}$, then $X^1_{k,t} = \{x_{,j} : x_{,j} \in p^2_{k,t,i}, i = \overline{1,Nkt2}\}$, where $l = \overline{1,Nxkt}$, and if $P^2_{k,t} = \varnothing$, then it is considered that $X_{k,t} = (\varnothing)$.

We shall call sequence $X_{k,t}$ s the *sequence of lexical categories of word inflective paradigm of t-type h_k*. One of lexical categories, usually the first, is named $X^*_{t,k}$ and called *normalized.*

There exists a single pair $(h_k,t_{k,j})$, ($h_k \in H$, $t_{k,j} \in T_k$, $k = \overline{1,Nh}$, $j = \overline{1,Nk}$) for every lexeme and, therefore, ordered list of LC $X_{k,t}$.

Let us define function $f_{l \to W}(l)$, $l = \overline{1,Nx}$,with range of values $W'_I = W_I \cup \{\varepsilon\}$, where ε – *dummy*, nonexistent word form. Thereby, for every lexeme W_I an ordered sequence of word-forms $Y_{Wi} = (y_1, y_2, \dots, y_{Nx})$ ($y_j \in W_I$ for $j = \overline{1,Nx}$) could be formed. Such sequence is called *word changing paradigm of lexeme W_i (WCP)*. If for lexeme W_I exists such l, that $f_{l \to W}(l) = \varepsilon$, it is said that lexeme W_I has a *dummy word changing paradigm* (Apresyan, 1989).

If the pair $(h_k, t_{k,j})$ corresponds to lexeme W_I and for some $l = l^*$ from $\overline{1,Nx}$ conditions: $f_{l \to X}(l^*) = X^*_{k,t}$ and $y^* = f_{l \to W}(l^*)$ ($y^* \in Y_{Wi}$, $y^* \neq \varepsilon$), are fulfilled, then we shall call the word form y^* as *normalized form or lemma of lexeme W_i*. Usually $y^* = y_1$. As a rule, infinitive is a lemma for the verb etc.

Let Y_{Wi} — WCP of lexeme W_i. Then word form's inflections of paradigm Y_{Wi} form ordered sequence denoted by Y_{FLCi}. Inflection class (FC) number I denoted by FC_I is the five:

$$FC_I = < h_k, t_{k,j}, P^1_{k,t}, X_{k,t}, Y_{FLCi} >,$$

where h_k – some part of speech; $t_{k,j}$ – some realization of LC *type* for corresponding part of speech; $P^1_{k,t}$ – ordinary LC, corresponding to $t_{k,j}$; $X_{k,t}$ – sequence of special LC of WCP, corresponding to $t_{k,j}$; Y_{FLCi} – some I-th sequence of inflections, also called WCP of FC, where $|X_{k,t}| = |Y_{FLCi}|$. Inflection class concept was first used by (Belonogov, Zelenkov, 1985), although inflection class was understood only as ordered sequence of inflections.

Let for lexeme W $WIS^* = (b^*_1, b^*_2, \dots, b^*_{NWIS^*})$ and exists $WIS_m = (b_1, b_2, \dots, b_{NWIS})$, where $m = 1 \div |X_{k,t}|$, such, that $WIS_m \neq WIS^*$. Consequently, exists natural number N_0, $N_0 = 0 \div \min(N_{WIS}, N_{WIS}^*)$, such, that $(b^*_1, b^*_2, \dots, b^*_{No}) = (b_1, b_2, \dots, b_{No})$; $(b^*_{No+1}, \dots, b^*_{NWIS^*}) \neq (b_{No+1}, \dots, b_{NWIS})$. Let us call the ordered sequence $z^s_{I,m} = ((b_{No+1}, \dots, b_{NWIS}), (b^*_{No+1}, \dots, b^*_{NWIS^*}))$, allowing to obtain lemma WIS from some word form WIS, *direct substitution*. Here I is a FC number, m – position number in WIS FC, s – exact pair number among other pairs in the m-th position. For each I-th FC is defined ordered set Z_I (may be empty): $Z_I = \{z_{I,1}, z_{I,2}, \dots, z_{I,Nzi}\}$, где Nzi = $|X_{k,t}|$. Each $z_{I,m} = \{z^1_{I,m}, z^2_{I,m}, \dots, z^{Nzim}_{I,m}\}$, where $m = 1 \div |X_{k,t}|$, also is a set of pairs of direct substitutions (may be empty). If the pair $z^s_{I,m} = (b^m, b^*)$ is a direct substitution, then the pair (b^*, b^m) is called *reverse substitution.* Reverse substitution allows obtaining some m-th word form WIS from lemma WIS. There is one-to one correspondence between the sets $B^* = \{\dots, b^*, \dots\}$ and $B^m = \{\dots, b^m, \dots\}$. Thus, $|B^*| = |B^m|$; if (b^*, b^m_1) and (b^*, b^m_2), then $b^m_1 = b^m_2$; if (b^*_1, b^m) and (b^m, b^*_2) , then $b^*_1 = b^*_2$. The letters from the constant part of WIS could be added to the beginnings of such character sequences for achievement of this term.

For example, the genitive of the plural noun *КОПЕЙКА* (copeck) with lexeme WIS* = (КОПЕЙК) is WIS_7 = (КОПЕЕК). Direct substitution should be (ЕК, ЙК), but for the lexeme of the same inflexion class (FC = 154) «ПУЛЬКА» (kitty or pellet or pool) direct substitution in the same position must be (ЕК, ЬК). This generates ambiguity. Therefore, two pairs of direct substitutions: (ЕЕК, ЕЙК) и (ЛЕК, ЛЬК) are formed in the morphology model for inflexion class 154 and m = 7. Thus, for some inflection classes the set of direct substitutions should be formed.

So, for obtaining word form of WCP with given LC it is enough to define WIS of the lemma, number of the inflexion class and the number of word form in WCP, thus the three $< WIS^*, FC, l >$. If Y_l = '–', then for given FC and, accordingly, for given lexeme the word form with such LC does not exist. However, even if $Y_l \neq$ '–', paradigm of the given lexeme could be dummy. Such situation is described with the help of the set $P^3_{k,t}$ of individual LC of given lexeme.

For example, lexemes «ДЕЛАТЬ» (do) and «СДЕЛАТЬ» have the same inflection class 175 and, hence, the same realization of ordinary and special LC, but they have different value of aspect: verb «ДЕЛАТЬ» – imperfective aspect, verb «СДЕЛАТЬ» – perfective aspect. So LC aspect should be the individual LC for this pair. Additionally, the individual LC could impose restriction on the existence of some inlections of the word. In the above example for FC = 175 FLC_{44} = «ЫЙ» и $Z_{175,44}$ = («ЕМ», ~), where sign '~' designates empty sequence. For the verb «ДЕЛАТЬ» WIS* = «ДЕЛА» → y_{44} = «ДЕЛАЕМЫЙ». For the verb «СДЕЛАТЬ»: WIS* = «СДЕЛА» → y_{44} = «СДЕЛАЕМЫЙ». This contradicts with Russian language standard.

So in the morphologic model should be the rules "rejecting" some inflection forms according their individual LC information. Such exclusion for given lexeme could be set explicitly by indicating the number of concrete inflection.

For example, for lexeme «МЕЧТА» (dream) there is no y_8 – plural genitive inflection. The set of individual LC realizations of lexeme inflections and numbers of forbidden inflections of WIP are considered to be individual feature of lexeme and are marked by I.

Thus, LC of every lexeme W_i could be given by three:

$$W_i = < WIS^*_i, FC, I_i > \quad (1).$$

2.3. Derivational morphology and compounding

Derivational morphology is based on detection of fixed expressions (more then 2000 of Russian idioms, proverbs, sayings), multiword prepositions, prefixes/suffixes with strong derivation functions and productive central derived forms (Kuznetcova, Efremova, 1986; Efremova, 1996), compounds (3000 of most frequent Russian compounds), processing 198 features consisting of morpho-syntactic features, derivational features, stylistic features and punctuator features.

3. Russian morphological dictionary

General lexicon of the dictionary is formed from the intersection of such lists (Yablonsky, 1998):

- Russian basic grammatical dictionary (80 000 word paradigms);
- Russian thesaurus (8 696 synonym rows, word list containing approximately 30 000 word paradigms);
- Large Russian explanatory dictionary[3] (more then 130 000 word paradigms from the language of the Eighties and the beginning of the Nineties);
- Orthographic dictionary[4] (60 000 word paradigms);
- Computer dictionary (1 500 word paradigms);
- Geographical names dictionary (1 500 word paradigms);

---

[3] Russicon company has the copyright on electronic version of the paper dictionary: Kuznetcov S.A. ________ _________ ________ _________ ______ (Large Explanatory Dictionary). St.-Petersburg. 1998, 1536 p. (in Russian)

[4] Russicon company licensed electronic version of the paper dictionary: Solov'ev N. V. Orthographic dictionary plus orthographic and punctuation reference guide, S.-Petersburg, 1996.

- Russian personal names, patronymics and surnames dictionary (10 000 word paradigms of Russian personal names, diminutives and patronymics, surnames of the world famous and Russian famous people);
- Business dictionary (1 500 word paradigms);
- Juridical dictionary (1 500 word paradigms);
- Jargon dictionary (5 000 entry word paradigms).

Joint number of all different Russian word paradigms is approximately **160 000**. Comprehensive description of all mentioned dictionaries is done in (Yablonsky, 1998).

3.2. Compressed database of Russian morphological dictionary

3.2.1. WFS - dictionary

In the compressed WFS - dictionary database all WIS are distributed into word forming groups (WFG). Word forming group consists of such set of fours:

$$< WFS_i, SUF, FC, I_i > (2),$$

where SUF – suffix (number of the suffix), FC – inflection class number; $WIS_i{}^* = WFS_i \oplus SUF$. Only first 255 maximum frequent suffixes are coded as separate linguistic units in compress WFS-dictionary realization. Other suffixes are included in WFS. Thus WIS are distributed into **42 874** WFS. Capacity of the compressed dictionary is **990 K.**

3.2.2. WIS -realization

For increasing speed of morphological analysis all WIS with stem gradation were generated. This formed **179 289** word inflection stems for database. In the compressed WIS - dictionary database the ordered sequence of all lexemes represented by (1) is stored. The speed of analysis is increased in 10 times.

Besides, we use several additional tables: table of inflection classes, inflection class — inflections, inflection — inflection classes, inflection class — right direct substitutions, joint right direct and right inverse substitutions, direct and inverse tables of suffixes, prefixes and substitutions in prefixes, and some other.

4. Evaluation of the morphological analyzer

Morphologic analyzer has been tested on various texts including more then 50 million words of literary and newspaper texts, Russian laws (1990–1995 years) available from Russicon text corpora (Yablonsky S.A., 1999, a, b). The results demonstrated right recognition of 95 – 98 % of text words. For Russian language the morphological analyzer leaves 1-5 % of all words in running text without the correct analysis when 10-15 % of words still have two or more analyses because of high categorial homonymy of inflective language. Only the sentential context used by disambiguation normally decides which analysis is appropriate.

C-realizations of morphologic analyzer and disambiguater are currently available for MS DOS and Windows 9X/NT.

Detailed description of Russicon Slavonic language processor could be find in forthcoming Belyaev B.M., Surcis A.S., Yablonsky S.A. / Yablonsky S.A. (ed.), (1999). Slavonic Language Processor RUSSICON, St.-Petersburg, Russia.

5. Concluding remarks

Now we are developing a rule-based disambiguater that uses Russian constraint grammar that consists of 1000 disambiguation rules, syntactic markers of lexemes.

Syntactic markers consist of government, concord and parataxis models for verbs, nouns, adjectives, prepositions and other parts of speech (Apresyan, 1985, 1989; Crockett, 1975; Iomdin, 1990; Russian grammar, 1980). To further improve of the disambiguating power of the tagger, the grammar and the number of syntactically marked lexemes is to be extended.

Russian disambiguater is now tested against several kinds of Russian texts from Russicon collection (Yablonsky, 1999, a, b). The first results (500 000 words) showed that tagger has fully disambiguated (no homonymy) from 76% (contemporary Russian artistic avant-garde texts) up to 90 – 95% “neutral“ Russian texts.

5. Acknowledgements

My thanks go to university and Russicon company colleagues Boris Belyaev, Anatol Surcis and Michael Kazakov. We worked together on a common design.

6. References

Apresyan Y.D., (1985) Синтаксические признаки лексем (Syntactic markers of lexeme). Russian Linguist. Vol. 9, №2/3, p. 289–318. (in Russian)

Apresyan Y.D., ed. (1989) Лингвистическое обеспечение системы ЭТАП-2 (Linguistic framework of the system ЭТАП-2). Moscow. (in Russian)

Ashmanov I. (1995) Grammar and Style Checker for Russian Texts. In Proceedings of Dialog'95 International Workshop on Computational Linguistics and its Applications. Kazan, Russia.

Belonogov G.G., Kuznezov B.A. (1983) Языковые средства информационных систем (Language tools for information systems). Moscow, 288 p. (in Russian)

Belonogov G.G., Zelenkov Y.G. (1985) Алгоритм морфологического анализа русских слов (An Algorithm for Morphological Analysis of Russian words). In journal “Issues of information theory and practice", №53. Moscow. (in Russian)

Belyaev B.M., Surcis A.S., Yablonsky S.A. (1993) Russian Language Processor RUSSICON: Design and Applications. In the *Proceedings of the East-West Artificial Intelligence Conference EWAIC-93*, Moscow, pp.175-180.

Bider I.G., Bolshakov I.A. (1976) Формализация морфологического компонента модели "Смысл<=>Текст". I. Постановка проблемы и основные понятия (Formalization of morphological component within the Meaning↔Text framework). Reports of USSR Acadimy of Science on Technical Cybernetics.№6, pp.42–57. (in Russian)

Bolshakov I.A. (1990) A Large Russian Morphological Vocabulary for IBM Compatibles and Methods of its Compression. In the *Proceedings of the 13th International Conference on Computational Linguistics COLING-90.* Helsinki, Finland.

Belyaev B.M., Surcis A.S., Yablonsky S.A. / Yablonsky S.A. (eds). (1998). Forthcoming *Slavonic Language Processor RUSSICON.* St.- Petersburg.

Chanod J. (1997) Current development for Cenral and Eastern European Languages. In *Proceedings of the Second European Seminar “Language Applications for a Multilingual Europe”*, Mannheim/Kaunas.

Crockett D.B. (1975) Agreement in contemporary standart Russian. Cambridge (Mass.). 458 p.

Denisov P.N.,ed. (1978) Учебный словарь сочетаемости слов русского языка (Russian word combinability dictionary). Moscow. (in Russian)

Efremova T.F.(1996) Толковый словарь слово-образовательных единиц русского языка (Explanatory dictionary of Russian language word forming units). Moscow.636 p. (in Russian)

Iomdin L.L. (1990) Автоматическая обработка текста на естественном языке: модель согласования (Automatic Natural text Processing: a Model of Grammatical Agreement). Moscow. 168 p. (In Russian)

Kuznetcova A.I., Efremova T.F.(1986) Словарь морфем русского языка (Russian morpheme dictionary). Moscow. (in Russian)

Karttunen L. (1983) KIMMO: a general morphological processor. In Dalrymple et al (Eds.). Texas Linguistic Forum, 22, Department of Linguistics, University of Texas at Austin, pp166-186

Karttunen L., Koskenniemi K., Kaplan R. (1987) A Compiler for Two-Level Phonological Rules. Techical Report. Center for the Study of Language and Information. Stanford University.

Kulagina O.S. (1986) Морфологический анализ русских именных словоформ (Morphologic analysys of Russian word forms). Internal Publication of the IPM. Moscow, Academy of Sciences, №10, 26p. (in Russian)

Kuznetcov S.A. (1998) Большой толковый словарь русского языка (Large Explanatory Dictionary). St.-Petersburg. 1536 p.(in Russian)

Mikheev A.S., Liubushkina L.A. (1995) Russian Morphology: An Engineering Approach Natural Language Engineering 1 (3), Cambridge University Press, pp. 235–263.

Varile, G. B., Zampolli, A. (1996) *Survey of the State of Art in Human Language Technology*. Cambridge: Cambridge University Press.

Popov E. V. (1986) Talking with Computers in Natural Language. Springer-Verlag, 305p.

Russian grammar (1980). Moscow. Vol. 1,2. (in Russian)

Rybakov F.I., Rudnev E.A., Petukhov V.A. (1980) Автоматическое индексирование на естественном языке. (Automatic indexing using natural language). Moscow. (in Russian)

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.

Tapanainen P., Jarvinen T. (1994) Syntactic Analysis of Natural Language Using Linguistic Rules and Corpus-based Patterns. In *Proceedings of Fifteenth International Conference on Computational Linguistics (COLING '94).* Kyoto, Japan.

Tikhonov A.N. (1985) Русский словообразовательный словарь (Russian word forming dictionary). Moscow. Vol.1,2. (in Russian)

Yablonsky S.A. (1990) Russian Language Processor RUSSICON. In *Actual problems of computer linguistics,* Tartu, Estonia. (in Russian)

Yablonsky S.A. (1997) New Capabilities for Russian and Ukrainian Language Learning Based on the Language Processor Russicon. In Sake Jager, John Nerbonne and Arthur van Essen (eds). Forthcoming "*Language Teaching and Language Technology".* Lisse: Swets and Zeitlinger.

Yablonsky S.A., (1998). Russicon Slavonic Language Resources and Software. In: W. Teubert, E. Tognini Bonelli and N. Volz (eds.) *Proceedings of the Third European Seminar Translation Equivalence,* (pp. 217-227), Montecatini Terme, Italy.

Yablonsky S.A. (1999, a). Russian Written Language Corpora Development. In: Proceedings of the International Seminar Dialog99, May 30-June 8, Tarussa, Russia.

Yablonsky S.A. (1999, b). Russian 20th Century Literature Digital Library for Language Teaching. In: Proceedings of the International Digital Libraries for Humanities Scholarship and Teaching JUNE 9-13, 1999, (pp. 252 - 253), University of Virginia Charlottesville, Virginia, USA.

Zaenen A., Uszkoreit H. (1996) Language Analysis and Understanding. In *Survey of the State of the Art in Human Language Technology* (http://www.cse.ogi.edu/CSLU/HLTsurvey/).