Russian Morphology: Resources and Java Software Applications.

Serge Yablonsky
Petersburg Transport University, Moscow av., 9, St.-Petersburg, 190031, Russia,
Russicon Company, Kazanskaya str., 56, ap.2, 190000, Russia
serge_yablonsky@hotmail.com; root@russicon.spb.su;
http://www.russicon.ru

Abstract

This paper deals with development and application of Russian morphology software and resources. The approach is particularly dependent on advanced morphological analysis. The paper presents the structure, formats and content of Russian dictionaries and corpora. Relevant aspects of the UML data models, XML format and related technologies are surveyed. We introducee the system based on Java and Oracle 9i DBMS.

1 Introduction

Up-to-date language technologies contain efficient morphological analyzers for Romance, Germanic (Karttunen, 1983; Karttunen, Koskenniemi, Kaplan, 1987; 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). This paper attempts to introduce results of 15 years ongoing project on developing of Russian resources and software for building advance Russian language morphological analyzers and their applications that enable a different forms of text indexing and retrieval, and a direct benefit from the Russian morphological analyzers in:

- information-acquisition tools,
- authoring tools,
- language-learning tools,
- translation-tools,
- summarizers,
- semantic web etc.

The objectives of this project are not unique. Several analogous projects have been carried out to different stages. In the late eighties of XX century we developed one of the first Russian morphologic analyzers on PC (Yablonsky S., 1990; Belyaev B.M., Surcis A.S., Yablonsky S.A., 1993; Yablonsky S.A., 1998; Yablonsky S.A., 1999). Now we are developing a set of platform independent Internet/Intranet Russian language processing tools based on Java and Oracle technologies.

2 Russian Resources for Morphology

Russicon company has such main counterparts (Yablonsky S.A., 1998) for Russian morphology software development:

- **Russian lexicon** which is formed from the intersection of the perfect set of Russicon Russian grammatical dictionaries with inflection paradigms (200.000 paradigms that produce more then 6.000.000 inflection word forms). It includes:
    - Russian basic grammatical dictionary.
    - Computer dictionary.
    - Geographical names dictionary.
    - Russian personal names, patronymics and surnames dictionary.
    - Business dictionary.
    - Juridical dictionary.
    - Jargon dictionary.
- **Russicon Russian explanatory dictionary**. The dictionary gives the broad lexical representation of the Russian language of the end of the XX century. More then 100 000 contemporary entries include new words, idioms and their meanings from the language of the Eighties-Nineties. The dictionary is

distinguished by its complete set of entry word characteristics, clear understandable definitions, its guidance on usage. All dictionary information for entries is structured in more then 60 attributes:

- entry word;
- multiple word entries;
- usage notes;
- precise, contemporary definitions;
- derivations;
- example sentences/citations;
- idioms etc.

- **Russicon Russian thesaurus** (set of 11.000 Russian synsets). Synonym list (8 696 synonym rows) plus word list containing approximately 30 000 normalized entry words with inflection paradigms.
- **Russicon Russian Orthographic dictionary** (100 000 normalized entry words plus inflection paradigms with stresses).
- **Russian WordNet** (in development).
- **Russian Corpora**. Today linguistically encoded Russian text corpus includes approximately 2 000 000 words and consist of anthology of Russian prose and poetry of the 20th century, law, business and newspapers. The texts were input from printed resources and Internet (Yablonsky S.A., 1998; 2000).

3 General Set Model of Inflection Morphology

In most language technology applications the encoded linguistic knowledge, i.e. the grammar, is separated from the processing components (Zaene, Uszkoreit, 1996). Linguistic Model (LM) of the language consists of all declarative knowledge about the language, which is concentrated in the set of dictionaries and linguistic tables. The number of word forms P for Slavic inflection languages is rather high (P>>1) for some parts of speech. For example, in Russian language P > 100 for verbs.

For the formal description of inflection morphology model the set theory is used. It is one of the best ways for description of inflection morphology (Bider, Bolshakov, 1976), (Kulagina, 1986). We present the general set model that permits to define mostly all sides of inflection morphology. It is realized in the Russian and Ukrainian morphological analyzers. In this model some concepts for the first time and other have new or more full meaning. We use definitions from (Yablonsky, 1999).

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

[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.

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 $P^2_{k,t} = \{p^2_{k,t,1}, p^2_{k,t,2}, \dots, p^2_{k,t,Nkt2}\}$, then $X^l_{k,t} = \{x_j : x_j \in p^2_{k,t,i}, i = \overline{1,Nkt2}\}$, 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:

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

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, … , $b^*_{NWIS^*}$) and exists WIS_m = (b_1, b_2, … , 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, … , b^*_{No}) = (b_1, b_2, … , b_{No}); (b^*_{No+1}, … , $b^*_{NWIS^*}$) ≠ (b_{No+1}, … , b_{NWIS}). Let us call the ordered sequence $z^s_{I,m}$ = ((b_{No+1}, … , b_{NWIS}), (b^*_{No+1}, … , $b^*_{NWIS^*}$)), 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):

$Z_I = \{z_{I,1}, z_{I,2}, \ldots, z_{I,Nzi}\}$, где $Nzi = |X_{k,t}|$. Each $z_{I,m} = \{z^1_{I,m}, z^2_{I,m}, \ldots, z^{Nzim}_{I,m}\}$, 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 $B^* = \{\ldots, b^*, \ldots\}$ and $B^m = \{\ldots, b^m, \ldots\}$. 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» $\rightarrow$ y_{44} = «ДЕЛАЕМЫЙ». For the verb «СДЕЛАТЬ»: WIS* = «СДЕЛА» $\rightarrow$ 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:

$$W_i = < WIS^*_i, FC, I_i >. \quad (2)$$

Linguistic tables (LT) form the second counterpart of the model FC (1). The structure of each LT depends on its utility type. For example, let us describe LT of ordinary LC (table of inflection classes). Table structure is simple:

TableFC$\rightarrow$FLC = { <FC1,Lc1>, <FC2Lc2>, …,<FCm,Lcm>},

where m – general number of FC; Lc – 8-bit code of general LC.

The interpretation of general LC in 8-bit code (d1,…d8) is like this: 4 low digits (d1-d4) depend on the value of 4 high digits (d5-d8) that distinguish part of speech. For example, the values d1=0, d2=0, d3=0, d4=0 define a noun, d1=0, d2=0, d3=0, d4=1 – an adjective. For nouns other 4 bits could be interpreted as thus: d7=1, d8=0 – ordinary noun, d7=0, d8=1 – substantial noun, d7=1, d8=1 –unchangeable noun. Bits d5, d6 are defined only for ordinary nouns and define gender p2 = {{"masculine", "feminine", "neuter", masculine/ feminine"}: d5=0, d6=0 – masculine, d5=0, d6=1 – femine, d5=1, d6=0 – neutur, d5=1, d6=1 – masculine/feminine.

General set model of inflection morphology we presented (Yablonsky S., 1999) permits to define mostly all sides of inflection morphology for Russian, Ukrainian and other Slavic languages.

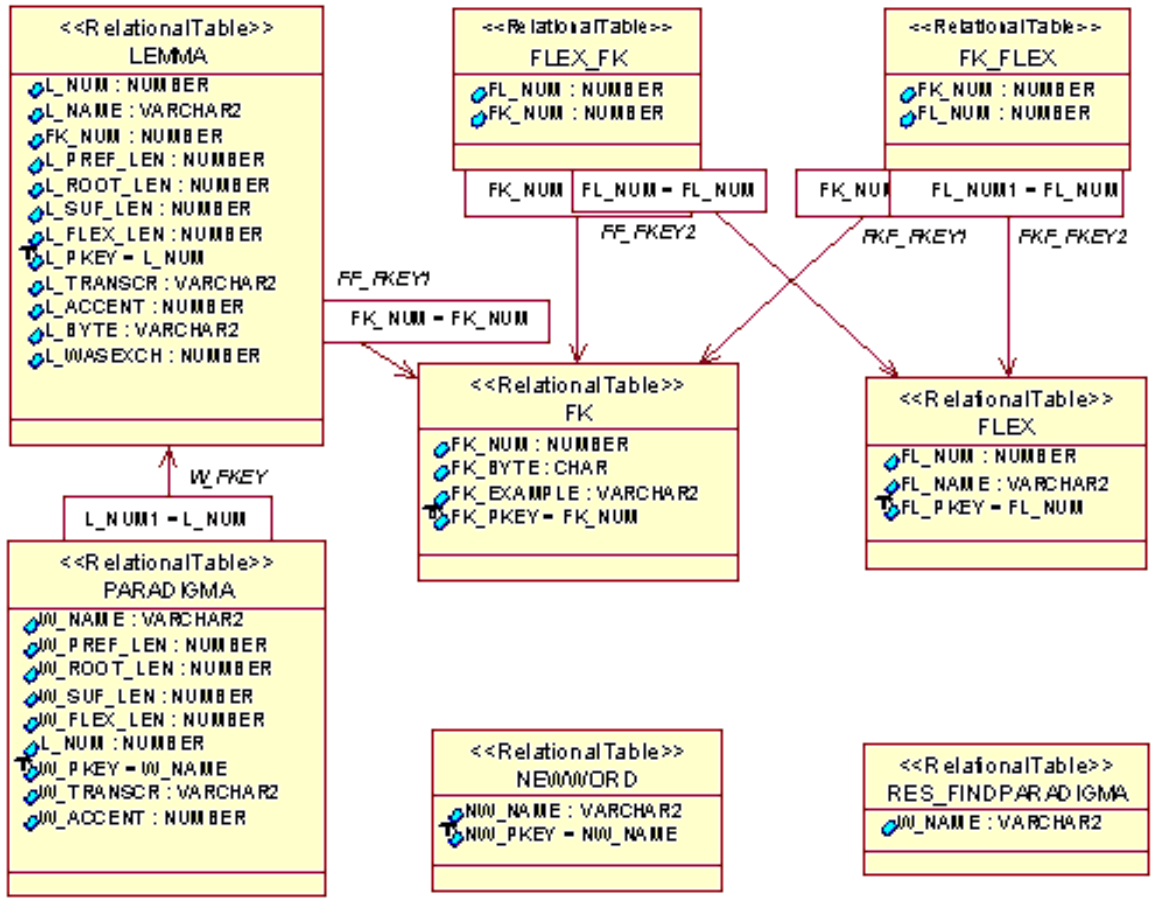

Figure 1. Word dictionary UML notation

4 Russian Lexical Databases for Morphology

The morphological analyzer has two main parts:

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

In general dictionary lexical information is represented in such form:

$$V_i = \{W_i, f_i\}, \quad i = 1 \div N_v, \quad (3)$$

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

4.1 Compressed Database

4.1.1 WFS-dictionary Database

The part of the word including prefix(es) and root is called word formative stem (WFS). The part of the word including prefix(es), root and suffix(es) is called word inflective stem (WIS). In the compressed WFS - dictionary database all WIS are distributed into word forming groups (WFG). Word forming group consists of such set of fours:

$<WFS_i, SUF, FC, I_i>$ (1),

where SUF – suffix (number of the suffix), FC – inflection class number; $WIS_i^* = WFS_i \oplus SUF$. Usually only first 255 maximum frequent suffixes are coded as separate linguistic units in compress WFS-dictionary realization. Other suffixes are included in WFS.

4.1.2 WIS-dictionary Database

For increasing speed of morphological analysis all WIS with stem gradation are generated. In the compressed WIS - dictionary database the ordered sequence of all lexemes is stored. The speed of analysis is increased in 10 times.
Besides, several additional tables are used: 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 (see Yablonsky S., 1999).

4.2 Word-dictionary Database

Today the memory cost is dramatically going down. So words without compression could be stored on HDD/RAM. The simplified UML-notation of the database for storing such resources is shown below.

4.2.1 Word-dictionary database UML Notation

Today Unified Modeling Language (UML) defines a standard notation for object-oriented systems (Booch G., Rumbaugh J., and Jacobson I., 1998). The objective of modeling is to complete a rigorous design with quality checks before we build a word-dictionary database system. The UML is an object-oriented methodology that standardizes modeling language and notation, not a particular method. Using UML enhances communication between linguistic experts, workflow specialists, software designers and other professionals with different backgrounds.
We introduced simplified UML data model (see figure 1) for word-dictionary database and developed a table-based UML mapping according to UML notations.

There are two main tablesin database. *Relational table LEMMA* is destined for storing lemma's linguistic information and has such attributes:
L_NUM – lemma's index;
L_NAME – lemma;
FK_NUM – number of inflection class;
L_PREF_LEN – length of prefixes;
L_ROOT_LEN – length of root;
L_SUF_LEN – length of suffixes;
L_FLEX_LEN – length of inflection;
L_TRANSCR – transcribtion;
L_ACCENT – number of accent letter;
L_BYTE – byte of additional linguistic information.
Relational table PARADIGM is destined for storing paradigm entry-word's linguistic information and has attributes similar to LEMMA table.
Example of Russian entry word мечта 'dream' with its inflection paradigm plus grammatical tags and hyphenation see below.

Lemma: мечта.

Paradigma:

мечта	*noun, feminine, singular, nominative, inanimate*
мечты	*noun, feminine, singular, genitive, inanimate*
мечте	*noun, feminine, singular, dative, inanimate*
мечту	*noun, feminine, singular, accusative, inanimate*
мечтой	*noun, feminine, singular, instrumental, inanimate*
мечтою	*noun, feminine, singular, instrumental, inanimate*
мечте	*noun, feminine, singular, prepositional, inanimate*
мечты	*noun, feminine, plural, nominative, inanimate*
мечтам	*noun, feminine, plural, dative, inanimate*
мечты	*noun, feminine, plural, accusative, inanimate*
мечтами	*noun, feminine, plural, instrumental, inanimate*
мечтах	*noun, feminine, plural, prepositional, inanimate*

Relational database management systems (RDBMS) play a crucial role in the storage of richly structured data. The main advantage of RDBMS is their maturity and almost three decades of experience with them.
In general the main features of Word-dictionary database are implemented using commercial DBMS Oracle9i. Usage of Unicode simplifies Multilanguage Word-dictionary development. PL/SQL-script for database creation is available by e-mail request root@russicon.spb.su.

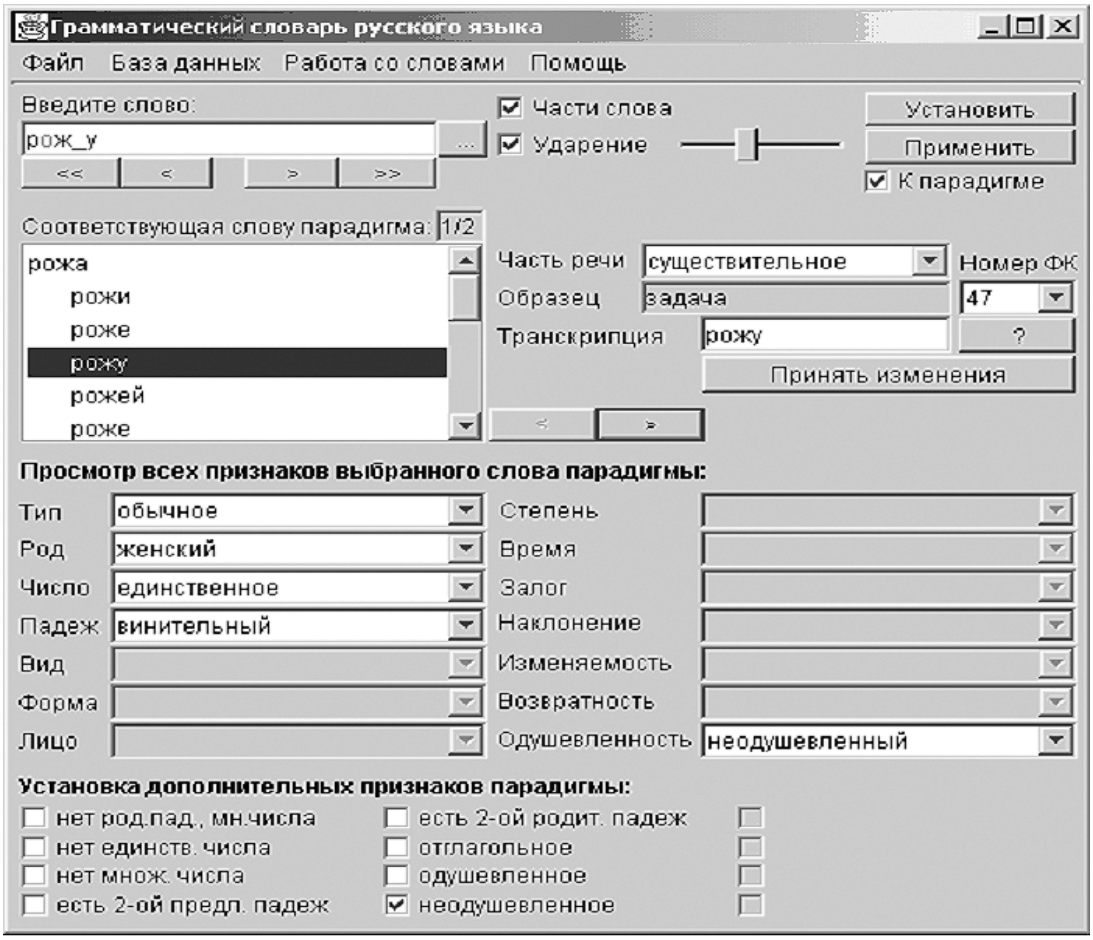

Figure 2. System for construction and support of Russian grammatical dictionaries

5 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, compounds (3000 of most frequent Russian compounds), processing 198 features consisting of morphosyntactic features, derivational features, stylistic features and punctuator features.

6 System for construction and support of Russian Grammatical Dictionaries

System allows to receive morphological information of the word and to build normal form for the word, shows paradigm for the word (see figure 2), constructs new words lexicon, constructs frequency lexicon. It provides such morphological information for new words treatment:

- input of grammatical characteristics of new words, length of word-building and word-changing stems, number of inflexion class etc.;
- generation of different variants of inflexion paradigms for new word containing only one correct variant;
- input of inflection paradigm of new words.

7 Russian Morphological Analyzer

Morphological analyzer and normalyzer allows:

- to define following grammatical characteristic s of a word: part of speech, case, gender, number, tense, person, degree of comparison, voice, aspect, mood, form, type, transitiveness, reflexive, animation;

- to modify a given word to its normal grammatical form/s – lemma/s (normalyzer). A set of applications were build on the base of the processor for three mentioned Slavonic languages.

All applications are designed in Java and Oracle 9i. This makes them platform independent.

References

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., 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 Proceedings of the East-West Artificial Intelligence Conference (EWAIC-93), Moscow.

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.

Booch, G., Rumbaugh, J., and Jacobson, I. (1998) The Unified Modeling Language user guide, Addison-Wesley.

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.

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)

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

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

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.

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

Yablonsky S.A. (1998) Russicon Slavonic Language Resources and Software. In: A. Rubio, N. Gallardo, R. Castro & A. Tejada (eds.) In: *Proceedings First International Conference on Language Resources & Evaluation,* Granada, Spain.

Yablonsky S.A. (1999) Russian Morphological Analyses. In: *Proceedings of the International Conference VEXTAL,* November 22–24 1999, Venezia, Italia.

Yablonsky S.A. (2000) Russian Monitor Corpora: Composition, Linguistic Encoding and Internet Publication. In: *Proceedings Second International Conference on Language Resources & Evaluation*, Athens, Greece.

Yablonsky S. A. (2002) Corpora as Object-Oriented System. From UML-notation to Implementation. In: *Proceedings LREC-2002,* Las Palmas, Spain.

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/).