Development of Russian Lexical Databases, Corpora and Supporting Tools for Speech Products

Serge A. Yablonsky

Department of Computing
St.-Petersburg State Transport University, St.-Petersburg, Russia
Russicon Company, St.-Petersburg, Russiaoot@russicon.spb.su

Abstract

The situation with regard to Russian language resources is fragmented and disorganized. For this reason, it is important to promote for Russian the development of its basic resources in one package that could be used for development of speech products.

The paper presents a design of the Russian lexical databases, corpora and supporting tools (system for construction and support of lexical databases, system for transcription, morphological analyzer and normalyzer) developed for wide usage in speech engineering.

1. Introduction

Up-to-date speech technologies contain language processing for Romance and Germanic languages [1-3]. Progress in Slavonic natural language processing [4] and particular in Russian language processing affords an opportunity for applying its results for creating different language resources and software [6, 8].

The situation with regard to Russian language resources is fragmented and disorganized. For this reason, it is important to promote for Russian the development of its basic resources in one package that could be used for development of speech products. The paper presents a design of the Russian lexical databases, corpora and supporting tools developed for wide usage in speech engineering.

Russian lexical databases – structured electronic lexicons containing formally represented grammatical and phonetic properties of lexical items – are used for training and testing essential components of the various speech technologies, as an ingredient for acoustic modeling, and for use as exception dictionary in the resulting systems. Several specific lexical databases are needed containing abbreviations and acronyms and their properties, names (person names, geographic names, organization names, etc.) with their properties. Finally, a large database with general lexicon with less frequent fully inflected word forms and their properties, is needed.

The lexicons that are required for the processing of the training set and include only the orthography of words are called ‘orthographic lexicons’. The processing of the training set is not the only purpose of the orthographic lexicons. They can also be used to reduce the manual check of the most frequent words of the training set.

The lexicons used to build the product lexicons usually include orthographies linked with their phonetic transcriptions. Lexicons with phonetic transcriptions are called ‘transcription lexicons’. The phonetic transcriptions are used in the product lexicons and in the acoustic training. Providing phonetic transcriptions is a time-consuming task for the lexicon supplier.

We developed different Russian language lexicons [6]:
- Person names;
- Geographic names;
- Company & brand names;
- Organization names;
Russian general lexicon consists of approximately 77,717 normalized entry words (lemmas) with hyphenation plus inflexion paradigms) and procreates approximately 3,462,558 words. Lexicon has been tested on various texts including more than 50 million words of literary and newspaper texts, Russian laws (1990–1995 years) available from Russicon text corpora, various CD’s and Internet Russian recourses. The results demonstrated right recognition of 95 – 98 % of text words.

Russian proper names lexicon consists of geographic names, person names, diminutives, patronymics and last names. Each entry of the lexicons describes the properties of a word. If a word has multiple grammatical properties (e.g. if it can be both a noun and a verb), multiple entries are created for this word. The properties described in each entry can be grouped into a number of major classes:

- Orthographic properties.
- Grammatical Properties.
- Abbreviations and Acronyms.

We use such grammatical properties:

- Part of speech: noun, adjective, particle, adverb, conjunction, parenthetic word, modal word, interjection, participle, preposition, abbreviation, verb unit of measure, verbal adverb, pronoun, numeral.
- Case: nominative, genitive, dative, accusative, instrumental, prepositional.
- Gender: masculine, feminine, neuter, masculine/feminine.
- Number: singular, plural.
- Tense: present, past, future.
- Person: 1st, 2nd, 3rd person.
- Degree of comparison: superlative, comparative.
- Voice: active, passive.
- Aspect: perfective, imperfective.
- Mood: indicative, imperative.
- Form: full, short (predicative), infinitive.
- Type: quantitative, verbal, ordinal, collective, invariable, possessive, pronominal, fractional.
- Transitivity: transitive, intransitive.
- Reflexive: reflexive, irrevocable.
- Animation: animate, inanimate.

This corresponds to the model of Russian morphology [7].

3. Electronic Text Corpora

Examples of electronic text corpora are electronic versions of recent newspapers, magazines, journals, covering the general domain and also special domains such information technology, finance, business, economy. Other examples are electronic versions of business correspondence, e-mails, electronic versions of fiction (literature) and essays. The text corpora is used as training and testing material for the development of the technologies to create language models, and serve as input to create lexical databases.

The Russian text corpora is collected

- from the Web;
- from newsgroups (usenet text including newsgroups from the IT domain);
- from publishers, press agencies, linguistic organizations, consumer products (CD ROMs with texts), etc.;
- newspaper text (including business/finance related articles);
- magazine text (including IT related and business/finance related articles);
- correspondence (personal and business);
- chat texts;
- business documents like memo’s, meeting minutes, reports, notes, etc.;
- scientific texts;
- fiction texts.

Today Russian corpora includes approximately 500,000,000 words. This collection will continue to grow as resources are created and encoded. The corpora files are prepared in text, HTML and XML formats. Every word of the corpora simultaneously is the entry word of Russian reference lexicon.

4. Language processor Russicon

In general, language processor Russicon is language independent, that is, it permits to build multilingual applications [6-7]. Now processor includes such main blocks for Russian language:

- System for construction and support of lexical databases. System allows to receive morphological information of the word and to build normal form for the word, shows paradigm for the word, constructs new words lexicon, constructs frequency lexicon, provides
morphological information treatment of new words not in the base dictionary.

- System for transcription of lexical databases.
- Morphological analyzer and normalizer. It gives possibility to define following grammatical characteristics of a word: part of speech, changeability, animation, case, number, gender, person, aspect, tense, transition, mood, form, reflexive (verb). Additionally algorithm of morphological analysis uses derivational morphology and compounding. This considerably increases coverage of word recognition. The evolving program realization has been tested on various texts including more than 10 Gb of literary and new spaper texts, Russian laws (1990–1995 years) available from Russicon text corpora, various CD’s and Internet Web-pages. The results demonstrated recognition of 95–97 % of text words. Russian normalizer (lemmatizer) modifies a given word to its normal grammatical form/s (lemma/s), and it can form the hole inflection paradigm for a given word.

5. Conclusions

In this paper we present a design of the Russian lexical databases, corpora and supporting tools (system for construction and support of lexical databases, system for transcription, morphological analyzer and normalizer) developed for wide usage in speech engineering. Detailed description of described Russian lexical databases, corpora and supporting tools could be find on www.russicon.ru.

6. Acknowledgements

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

7. References

[1] Dan Jurafsky, et all, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition".