SOME ISSUES OF CREATION OF BELARUSIAN LANGUAGE COMPUTER RESOURCES

Natallia Rubashko, Galina Nevmerjitskaia

Faculty of Applied Mathematics and Informatics, Belarus State University, Skorina av., 4, Minsk, 220050, Belarus, e-mail: *roubashko@bsu.by, nevmer@bsu.by*

Abstract. The main reason for creation of computer resources of natural language is the necessity to bring into accord the ways of language normalization with the form of its existence – the computer form of language usage should correspond to the computer form of language standards fixation. This paper discusses various aspects of the creation of Belarusian language computer resources. It also briefly gives an overview of the objectives of the project involved.

Introduction

Computers and computer networks became the main tools for storage, processing and distribution of enormous quantities of information presented in the form of text documents (articles, reports, patents, books, etc.), in other words in the form of natural language (NL) texts.

Because information storage and processing is getting cheaper by a factor of a thousand or so every decade, we know that computer technology will continue to penetrate and reshape society.

These trends create an enormous economic and social opportunity for natural language and speech technology. Where computers are already involved in creating, transmitting, storing, searching, or reproducing speech and text, we have the chance, at little marginal cost, to add new features that improve the quality of the process or that increase the productivity of the human labor involved. Simple examples of this kind include the use of spelling correctors in word processing; the use of speech technology to reduce the workload of telephone attendants, by screening calls with voice recognition or providing information through voice synthesis; and the use of machine-aided translation programs, which make human translators more productive by providing a rough draft to be corrected.

An important paradigm shift is currently taking place in linguistics and language technology. Purely introspective research focussing on a limited number of isolated phenomena is being replaced by a more data-driven view of language. The growing importance of data-driven and stochastic methods opens new avenues for dealing with the wealth of phenomena found in real texts.

In order to build realistic models of human speech and language, researchers need access to masses of linguistic data: speech, text, lexicons and grammars. The process of creating, maintaining, expanding, modifying and distributing databases of the necessary size presents logistical, legal and computational difficulties that tax the resources of large corporations, and may often exclude universities and smaller commercial research groups.

Publication of linguistic resources benefits the entire research community by effectively sharing production costs, avoiding duplication of effort, and lowering start-up barriers. Published resources also provide the basis for replication of published research results, establishing a stable reference point against which different analyses or algorithms can be compared. Finally, published resources can be corrected, improved and further annotated, to the benefit of the entire community.

As an exámple we can consider Linguistic Data Consortium (LDC) which was founded in 1992 to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. Based at the University of Pennsylvania, the LDC is a broadly-based consortium that now includes more than 100 companies, universities, and government agencies. Since its foundation, the LDC has delivered data to 197 member institutions and 458 non-member institutions (excluding those who have received data as a non-member and later joined). The LDC's Catalog contains 190 corpora for different languages [2].

Nowadays Republic of Belarus still remains one of the few states of CIS where the computer resources of the national language only start to form. Partly it is connected that the project of Russian language computer resources which begun to realize already in USSR in 1985-1990 provided as subtasks the creation of computer resources of Ukrainian and Russian languages only [4].

This paper discusses various aspects of the creation of Belarusian language computer resources (BLCR).

Main aspects of the creation of BLCR

The main reason for creation of computer resources of natural language is the necessity to bring into accord the ways of language normalization with the form of its existence – the computer form of language usage should correspond to the computer form of language standards fixation. Therefore it is necessary to create computer resources of Belarusian language which can be considered as a system for integrated automation of linguistic research. The system consists of all kinds of linguistic data: texts, dictionaries, grammars and other linguistic sources published the form of databases available to all researchers, and special software which allows researchers to use these data and to construct new linguistic objects and processes. Such techniques also performing automatic acquisition of linguistic knowledge may help to solve problems of linguistic engineering.

The BLCR's core mission is to provide linguistic resources in support of precompetitive research and development in speech and language technology. It is the model of Belarusian language accessible for observation, analysis, change and application. Its creation needs a productive relationship with linguists, mathematicians, programmers and others interested in the study of language and speech. The widespread availability of new tools for creation and use of language-related data, along with increasingly affordable networked computer power and mass storage, will naturally bring new kinds of researchers into the group of those who prepare, publish and use speech and language databases.

The purpose of the BLCR project is to produce all lexicon resources: general dictionary of Belarusian language, dictionaries of synonyms, homonyms and antonyms, frequency dictionary, proper name dictionary, dictionary with phonetic transcription; bilingual dictionaries (Russian-Belarusian and Belarusian-Russian), terminological dictionaries and others.

This new research paradigm requires very large corpora annotated with different kinds of linguistic information.

In computational linguistics the corpus definition is based on the following principles [3]:

- sampling and representativeness;
- finite size;
- machine-readable form;
- standard reference for the language variety that it represents.

The main problem is the definition of corpus volume because for the purpose of improving the performance of linguistic technologies it is necessary to have tens or even hundreds of millions of words of text to derive useful estimates of the likelihoods of various word sequences, and its performance will continue to improve as its training set grows to include billions of words. To put these numbers in perspective, consider that a typical novel contains about a hundred thousand words, so that we are talking about the equivalent of hundreds or even thousands of novels. The first corpora of English (The Lancaster/Oslo-Bergen Corpus (LOB), the Brown University Corpus) created in 60-th contained 1 million words. But now the British National Corpus (BNC), contains 4124 texts with more than 100 million words [1]. It is necessary to stress that English belongs to inflectional languages whereas the Belarusian language has rich flexion so words depending on a part of speech can contain up to 28 grammar forms.

The next characteristic of a corpus is representativeness .We are interested in creating a corpus which is maximally representative of the variety under examination, that is, which provides us with an as accurate a picture as possible of the tendencies of that variety, as well as their proportions. What we are looking for is a broad range of authors and genres which, when taken together, may be considered to "average out" and provide a reasonably accurate picture of the entire language population in which we are interested.

If corpora is said to be unannotated it appears in its existing raw state of plain text, whereas annotated corpora has been enhanced with various types of linguistic information. Unsurprisingly, the utility of the corpus is increased when it has been annotated, making it no longer a body of text where linguistic information is implicitly present, but one which may be considered a repository of linguistic information. The implicit information has been made explicit through the process of concrete annotation.

For instance, the texts can be given orthographic transcriptions, disfluency annotation, discourse structure annotation, part-of-speech annotation, syntactic structure annotation, word sense disambiguation, phonetic transcription, and intonational annotation. Some of these annotations have introduced new structure, which was then used by others – for instance, the discourse structure annotation used a new phrasal segmentation created by the disfluency annotation. Each of the annotation efforts imposed various informally defined format changes, and most of them also made sporadic to extensive corrections in the underlying orthographic transcription.

The annotated corpus can be used in different applications. For this purpose it is necessary to select an annotation format satisfying the following conditions:

- availability of several levels of the information extracted from annotation separately from each other;
- potential expansibility of information types which are not annotated at the present stage.

The processing of text corpora has already became one of the basic methods of linguistic researches. The corpus is a source not only for systematic updating of dictionaries and grammar data, but a source for obtaining statistical estimations of common use and joint occurrence of lexical and grammar units, statistical analysis of properties of a coherent text, that will allow to improve linguistic researches of the Belarusian language. This corpus can also be of more general use in teaching and learning Belarusian language, especially via the Internet.

For the purpose of bilingual dictionary creation it is necessary to have parallel Belarusian-Russian corpus. Parallel corpora have become an essential resource for work in multilingual natural language processing. They represent resources for automatic lexical acquisition, they provide indispensable training data for statistical translation models and they can provide the connection between vocabularies in cross-language information retrieval. More recently, parallel corpora can be exploited in order to develop monolingual resources and tools, using a process of annotation, projection, and training. For these reasons, parallel corpora can be thought of as a critical resource.

A parallel corpus is not immediately user-friendly. For the corpus to be useful it is necessary to identify which sentences in the sub-corpora are translations of each other, and which words are translations of each other. A corpus which shows these identifications is known as an aligned corpus as it makes an explicit link between the elements which are mutual translations of each other.

Therefore one of the two main objectives of the corpus part of the BLCR project is to produce large monolingual corpora of approx. 20 million running words which are to be morphosyntactically tagged according to predefined tagsets. The other main objective is to construct a parallel corpus Russian and Belarusian languages.

Conclusion

The real usefulness of these results naturally depends on the purpose we intend to use them for. The main purpose of creation of BLCR is to represent as wide a range of modern Belarusian as possible.

The BLCR project will be useful for a very wide variety of research purposes, in fields as distinct as lexicography, artificial intelligence, speech recognition and synthesis, literary studies, and all varieties of linguistics.

Belarusian language currently lacks such computer resources, and we are ready to facilitate the development, publication and distribution of them.

References

[1] *British National Corpus* – Available at http://www.heu.ox.ac.uk/BNC/

[2] *Linguistic Data Consortium* – Available at http://www.ldc.upenn.edu/

[3] Mcenery T., Wilson A. *Corpus Linguistics*. - Edinburgh: Edinburgh University Press, 1996. – 132 p.

[4] *Computer fund of Russian language: concepts and judgements* / Kharaulov Y.N.(ed), 1986. – 240 p. (in Russian)