INTERNET DATA ACCUMULATION AND PROCESSING COMPLEX

O.Makhambetov*, A.Makazhanov, Z.Yessenbayev, B. Matkarimov, I.Sabyrgaliyev, A.Sharafudinov

NURIS, Nazarbayev University, Astana, Kazakhstan, *omakhambetov@nu.edu.kz

INTRODUCTION.

A language corpus is a collection of texts written in that language and classified by genres. Corpora are actively used by researchers from different fields (most notably linguists and computer scientists) and by industry (Google, Yandex, etc.)

METHODOLOGY AND RESULTS.

Computer Scientists from NURIS were first to assemble an open, annotated corpus for Kazakh language, the KLC (Kazakh Language Corpus) [1]. KLC is designed to be a large scale corpus containing over 135 million words and conveying five stylistic genres: literary, publicistic, official, scientific and informal. In addition to that KLC has a grammatically annotated sub-corpus and a sub-corpus with the annotated speech (audio) data.

KLC is currently up and running as a Web service providing visitors with abundance of data and advanced search opportunities. This resource is of great value to linguistics and computer science research communities not only in Kazakhstan, but all around the world. Indeed, from the moment KLC was presented at EMNLP 2013 conference in Seattle in October 20, 2013 the corpus received 128 visits from 13 countries. Amongst them our colleagues from USA, Russia and China, who are interested in collaboration in the field of computational linguistics. The work on the corpus resulted in developing prototypes of the language analysis tools that can be integrated into software packages used world-wide, such as OpenOffice, Lucene, Mozilla products, and Hunspell. In addition to that, one research paper has been published in the proceedings of a top-ranked conference, and two more were submitted. The potential of the corpus development is far from being exhausted!

CONCLUSIONS.

The future work includes the following:

- Development of machine translation systems;
- Improvement of the existing speech recognition/generation system;
- Development of Kazakh sentiment analysis systems, i.e. detecting emotional flavor of a text (sad, happy, angry, etc.)

There is a lot of other potential applications; we have just listed the ones which are of primary interest. It is important to understand that the corpus is a resource that lasts in time. Let us not forget that one of the first corpora was developed in Brown University in 1964, and it is still used by many researchers and referred to by the university name as Brown corpus [2].

ACKNOWLEDGMENTS.

We would like to thank the Ministry of Education and Science of the Republic of Kazakhstan for supporting this work through a grant under the 055 research program.

REFERENCES.

1. O. Makhambetov, A. Makazhanov, Z. Yessenbayev, B. Matkarimov, I. Sabyrgaliyev, A. Sharafudinov. (2013). Assembling the Kazakh language corpus. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1022–1031, Seattle, Washington, USA, October 2013, Association for Computational Linguistics.
2. W.N. Francis, H. Kucera. (1979). Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers. Department of Lingustics, Brown University, USA.