Reports of Enlarged Sessions of the
Seminar of I. Vekua Institute
of Applied Mathematics
Volume 32, 2018

GEORGIAN UNIVERSAL SMART CORPUS AND IN IT INBUILT ABKHAZIAN, CHECHEN, KABARDIAN, LEZGIAN, AND MINGRELIAN SELF-DEVELOPING CORPUSES – RESULTS, PERSPEQTIVES

Konstantine Pkhakadze Merab Chikvinidze Giorgi Chichua
David Kurtskhalia Shalva Malidze

Abstract. In the paper is shortly overviewed Georgian Universal Smart Corpus, which is elaborated on the base of Pkhakadze's Logical Grammar of Georgian Language and, also, the first trial versions of Abkhazian, Chechen, Kabardian, Lezgian and Mingrelian corpuses, which are functioned as inbuilt systems of the Georgian Universal Smart Coprus.

Keywords and phrases: Georgian Universal Smart Corpus, Logical Grammar of Georgian Language, Abkhazian, Chechen, Kabardian, Lezgian, and Mingrelian Self-Developing Corpuses.

AMS subject classification (2010): 68T50.

1 Introduction. In the paper is overviewed the Self-Developing Georgian Intellectual Web-Corpus, in other words, the first trial version of the Georgian Universal Smart Corpus [2], which is elaborated on the basis of Pkhakadze's Logical Grammar of Georgian Language. The Corpus was constructed in confine of the AR/122/4-105/14 project "One More Step Towards Georgian Talking Self-Developing Intellectual Corpus". This project, which is a subproject of the long-term project "Technological Alphabet of the Georgian Language" of the Center for the Georgian Language Technology at the Georgian Technical University, was financed by Shota Rustaveli National Science Foundation (in 2017 year, Shota Rustaveli National Science Foundation announced AR/122/4-105/14 project as a successful project because of its very important results (see materials located at the address `http://www.rustaveli.org.ge/en/Success-Stories/page/1988`).

Thus, it can be said, that the results obtained within the successfully completed AR/122/4-105/14 project can be estimated as a groundbreaking step for scientific field engaged by the complete technology processing of the Georgian state languages. In particular, the project researches provide the results of fundamental importance in sense of protection from danger of digital extinction of Georgian and Abkhazian languages [1], which, in turn, are Georgian state languages. Therefore, these researches should be estimated as: 1. One more very important step towards defending the Georgian language from the real danger of digital extinction; 2. First very important step towards defending the Abkhazian language from the real danger of the digital extinction. Moreover, we underline that, in rapidly forthcoming digital age thanks to the results of the AR/122/4-105/14 project, Georgian and Abkhazian languages become more protected from threats of digital extinction than they were before. In addition, the first trial version of the

Georgian Universal Smart Corpus, which, as was already said, was constructed in confine AR/122/4-105/14 project [2], was expanded with the Megrelian, Kabardian, Lezgian and Chechen languages and, already, in trial mode, it is the only one self-developing (it contains its own self-developing tools, which are functions automatically, as for Georgian, Abkhazian and Megrelian, as well as for Chechen, Kabardian and Lezgian languages) multimodal (it contains the trial tools for constructing corpuses of titred data for Georgian, Abkhazian, Megrelian, Chechen, Kabardian and Lezgian spoken languages) and multilingual (together with Georgian it contains Russian, English, Abkhazian, Megrelian, Chechen, Kabardian and Lezgian Self-Developing Corpuses and tools for matching i.e. parallelizing them with Georgian one) Georgian corpus, which, at the same time, is the largest (for today (25.03.2018) it contains 278 614 032 words, where 4 322 571 are different) and technologically most of all supported Georgian self-developing corpus. This makes clear, that together with protection Georgian language from danger of digital extinction our researches should be estimated also as a step towards the protecting the above Ibero-Caucasian languages from the threat of digital extinction.

2 The Abkhazian, Chechen, Kabardian, Lezgian and Mingrelian Self-Developing Corpuses – Results and Perspeqtives. Below, as our new results, we present the short descriptions of the Abkhazian, Chechen, Kabardian, Lezgian and Mingrelian Self-Developing Corpuses, which are constructed on the basis of tools elaborated within the project AR/122/4-105/14. This tools, in turn, are constructed on the base of the Pkhakadze's Logical Grammar of Georgian Language elaborated within the FR/362/4-105/12 project "Foundations of Logical Grammar of Georgian Language and Its Application in Information Technology", which was financed by the Shota Rustaveli National Science Foundation. Thus, for today (25.03.2018):

1. Our Abkhazian corpus is most volumetric (it contains 2050446 words, where 173223 are different) and only one self-developing Abkhazian corpus, which is already equipped with trial Abkhazian text reader and with trial corpus of titrated Abkhazian speech data;

2. Our Chechen corpus is most volumetric (it contains 122159 words, where 17798 are different) and only one self-developing Chechen corpus;

3. Our Kabardian corpus is most volumetric (it contains 102539 words, where 55237 are different) and the only one self-developing Kabardian corpus;

4. Our Lezgian corpus is most volumetric (it contains 402018 words, where 130483 are different) and only one self-developing Lezgian corpus;

5. Our Mingrelian corpus is most volumetric (it contains 967521 words, where 182840 are different) and only one self-developing Mingrelian corpus, which is already equipped with trial Mingrelian text reader and with trial corpus of titrated Mingrelian speech data.

In addition to this, Ibero-Caucasian corpuses shortly described above are already realized searching by words and sentences in such a way, that:

1. It is possible to indicate the length and/or the type of search sentence;

2. Together with searched sentence the search provides us with links to sources, by the help of which we can have access to these sources (see Figure 1);

3. In search with a word the statistical counters of the corpus provide informations about this search word. They are:

3.1. The frequency and relative frequency of the searched word in the corpus;

3.2. Date of entry of the searched word in the corpus (for example, the last picture of the figure shows the search result launched by the Abkhaz word „Аəгьы“; from which became known that the date of entry into corpus of this word is 2017-04-15, the frequency is 68; Relative frequency - 0,0000242175).

Besides, the corpuses, consider above as it was mentioned, are already equipped with trial tools for constructing corpuses of titrated speech data. Moreover, Abkhazian and Mingrelian corpuses are equipped with text readers (these readers, according to our information, are the first and only readers for Abkhazian and Mingrelian languages). In particular, by activating "Read= წაკითხვა" button, which is located below the search phrase (see Figure 1), the corpus allows you to listen to the searched phrase.

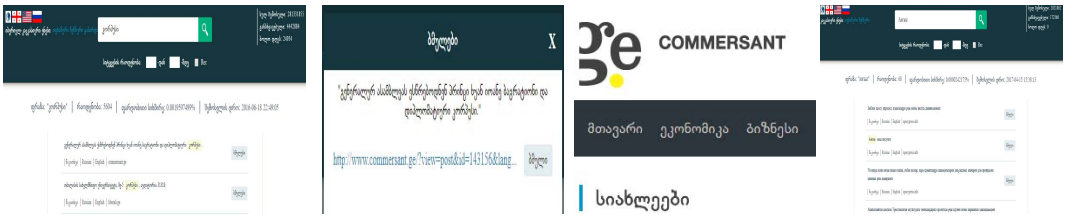

Figure 1: Georgian (first 3 pictures) and Abkhazian (last picture) corpuses in action

Now about perspectives: We plan to equip above shortly described Ibero-Caucasian corpuses with all such systems with which our self-developing Georgian intellectual corpus [2] are already equipped they are:

1. Into text analysis direction – the systems for processing Georgian text; the taggers, descriptors and generators for the Georgian words of V, N and A type; the Georgian self-developing syntactic/orthographic checkers; the Georgian texts analyzer, the Georgian question-answerer and the Georgian logical tasks/analogies generating and testing automatic systems.

2. Into automatic translation direction - the Georgian-Mathematical and hybrid Georgian-English-German translators; the internet and mobile applications of the Georgian multilingual voice lexicon, Georgian multilingual spoken assistant, and Georgian extension of Google translator; internet and mobile applications of the multilingual textual and voice messages between Georgian smart papers.

3. Into speech interaction direction - the Georgian spoken browser and the Georgian voice managed reader; the internet and mobil applications of Georgian spoken assistant for Georgian speech disorder persons; the Georgian adapted Internet and Computer; the analyzer, segmentator and generator for Georgian titrated speech data and the voice inbuilt tools in Georgian reader-listener systems.

3 Conclusion. It is clear from the above-said, that the complete construction of Georgian, Abkhazian, Chechen, Kabardian, Lezgian and Mingrelian Universal Self-Developing Smart corpuses have a vital importance for defending these languages from danger of digital extinction.

REFERENCES

1. Pkhakadze K., Chikvinidze M., Chichua G., Kurtskhalia D., Malidze Sh. The open letter to the Georgian parliament, government, national academy of sciences and Georgian and Abkhazian society i.e. Key principles of the unified program of complete technological supporting of the official language of Georgia (Georgian, Abkhazian) i.e. In the future cultural world with the completely supported Georgian and Abkhazian languages. *The Georgian Language and Logic*, **11** (2017-2018), 21-164.
2. Pkhakadze K., Chikvinidze M., Chichua G., Kurtskhalia D., Beriashvili I., Malidze Sh. The Georgian Intellectual Web – Corpus: Aims, Methods, Recommendation. *Additional Publication of the Journal The Georgian Language and Logic*, (2017), 4-320.

Received 11.05.2018; revised 12.10.2018; accepted 21.12.2018.

Author(s) address(es):

Konstantine Pkhakadze
Center for Georgian Language Technology
Georgian Technical University
77, Kostava Str., 0160, Tbilisi, Georgia
E-mail: gllc.ge@gmail.com

Merab Chikvinidze
Center for Georgian Language Technology
Georgian Technical University
77, Kostava Str., 0160, Tbilisi, Georgia
E-mail: gllc.ge@gmail.com

Giorgi Chichua
Center for Georgian Language Technology
Georgian Technical University
77, Kostava Str., 0160, Tbilisi, Georgia
E-mail: gllc.ge@gmail.com

David Kurtskhalia
Center for Georgian Language Technology
Georgian Technical University
77, Kostava Str., 0160, Tbilisi, Georgia
E-mail: gllc.ge@gmail.com

Shalva Malidze
Center for Georgian Language Technology
Georgian Technical University
77, Kostava Str., 0160, Tbilisi, Georgia
E-mail: gllc.ge@gmail.com