Reports of Enlarged Session of the
Seminar of I. Vekua Institute
of Applied Mathematics
Volume 28, 2014

# AN OVERVIEW OF THE TRIAL VERSION OF THE GEORGIAN SELF-DEVELOPING INTELLECTUAL CORPUS NECESSARY FOR CREATING GEORGIAN TEXT ANALYZER, SPEECH PROCESSING, AND AUTOMATIC TRANSLATION SYSTEMS

Pkhakadze K., Chikvinidze M., Chichua G., Maskharashvili A., Beriashvili I.

**Abstract**. In this paper, we will shortly overview a Trial Version of the Georgian Self-Developing Intellectual Corpus, which among the other trial systems created in our group[1] is located on the link `http://geoanbani.com/Corpus/`. The corpus is created in the Scientific-Educational Center for Georgian Language Technology (GLTC) at the Georgian Technical University (GTU) by the leadership of K.Pkhakadze within the researches pursued in the confines of the projects "Foundations of Logical Grammar of Georgian Language and Its Application in Information Technology" and "Internet Versions of a Number of Developable (Learnable) Systems Necessary for Creating The Technological Alphabet of the Georgian Language" [2] [3], which are the sub-projects of the long-term project "Technological Alphabet of the Georgian Language"[1].

**Keywords and phrases**: The Georgian self-developing Intellectual Corpus, the logical grammar of the Georgian language, the technological Alphabet of the Georgian language.

**AMS subject classification**: 03B65, 68T50, 68Q55, 91F20.

**1. An Overview of the Trial Version of the Georgian Self-Developing Intellectual Corpus.** Below overviewed Trial Version of the Georgian Self-Developing Intellectual Corpus, shortly, Georgian Intellectual Corpus (GI_Corpus) is created in GTU-GLTC under K.Pkhakadze's leadership. The idea of creating such web-corpus [8], and, also, the methods and algorithms of the intellectual procedures that are inbuilt in the GI_Corpus belongs to him. The software part of the trial version of the corpus is directed by M.Chikvinidze. By now he is doing research to improve the quality of the Georgian Self-Developing Grammar Checker (Analyzer) and Georgian-English-German (Geo-Eng-Ger) Rule based translator systems. In the theoretical researches with the aim of improving quality of these systems A.Maskharashvili and I.Beriashvili also play an important part. G.Chichua is doing researche to improve quality of the Georgian Speech Recognizer and Georgian E-Texts Reader Systems.

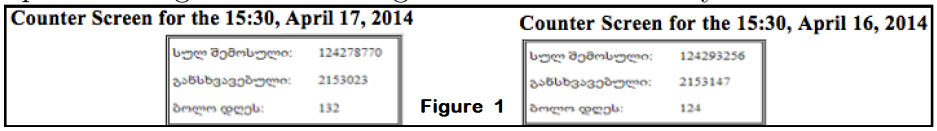
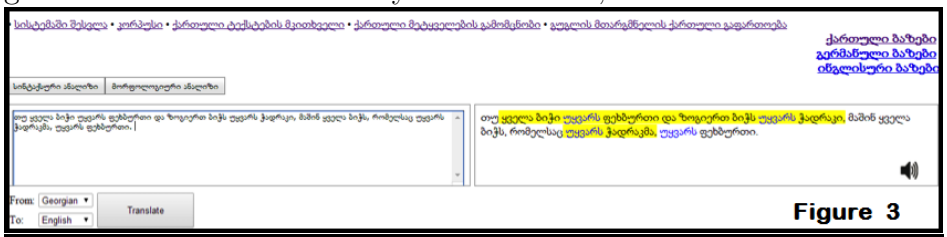
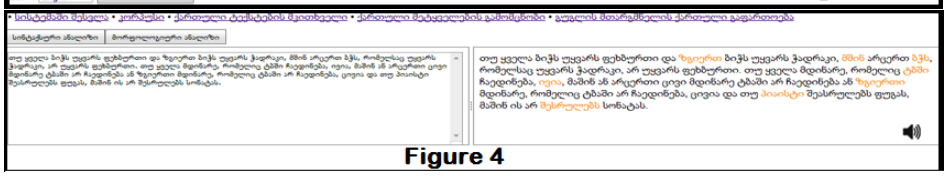


Figure 1

[1]These systems are: Georgian Self-Developing Grammatical Checker; Georgian Self-Developing Spelling Checker; Georgian Speech Recognizer; Georgian E-text Reader; Reader System for Georgian Internet Sites; Rule based Georgian-English-German Translator system; Georgian Extension of the Google Translate; Georgian speech to speech translator; Georgian Multilingual Spoken Lexicon; Multilingual Spoken Support for (Georgian) Persons with Speech Disorder [1], [2], [3].


The main page of GI_Corpus is available on the address http://geoanbani.com/Corpus/. On this page, one can find the counter. In Figure 1, the counter shows that during 24 hours (from 15:30 of 16.04.2014 to 15:30 of 17.04.2014) the total number of the words in the corpus has increased by 14486 words (from 124278770 to 124293256); the number of newly added words has increased by 124 (from 2153023 to 2153147). In Figure 2, the information is given that the corpus gives for a word "patrioti" (patriot), according to which the word has the relative frequency equal to 0.0008%, and the frequency equal to 950. Also, the corpus gives sentences and the left, right, and both-side contexts of the word. Besides, using Google translator, the Corpus gives English and German written and spoken translations of the given word; moreover, by clicking icon of Speakers, the user can hear the spoken form of this word which is synthesized by the Georgian E-Text Reader created by K.Pkhakadze and G.Chichua (see `http://geoanbani.com`). Using the Georgian E-Text Reader system, together with written forms of the word GI_Corpus saves their synthesized forms as well. We need to have the synthesized forms of the words in order to improve Georgian Speech Recognizer created by K.Pkhakadze and G.Chichua (see `http://geoanbani.com`) and, also, to improve Reader System for Georgian Internet Sites created by K.Pkhakadze, G.Chichua and M.Chikvinidze.




In Figure 3, we see the page of our website that is not freely accessible (some of the links of the page are not freely accessible). In Figure 3, the result of the functionality of the Georgian grammar checker is shown, i.e. analyzer system, which is the first and only grammar checker for the complex and simple sentences of the Georgian language.[2] The

[2]Here it is noteworthy, that L.Abzianidze [9] implemented HSPG style grammar checker for

system is designed by K.Pkhakadze and M.Chikvinidze on the basis of K.Pkhakadze's logical grammar of the Georgian language [7]. It is especially important to note that we have already implemented intellectual procedures that automatically can recognize nouns, adjectives and verbs with quite wide coverage and sufficiently high accuracy. Thus, we can say that on the basis of GI_Corpus our grammar analyzer can extend its knowledge automatically. Because of this the analyzer system is called the Georgian Self-Developing Grammar Analyzer.

In Figure 4, the Georgian Self-Developing Spelling Checker is shown. On the basis of its current knowledge it is analyzing an input text and on web site it is coloring in the red the detected incorrect words in the text.

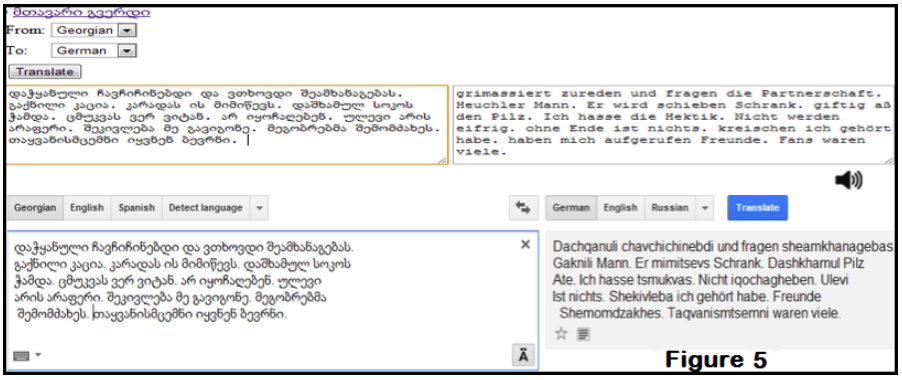



**Figure 5**

The upper part of Figure 5 shows a Georgian–German translation of a text using Google Translator system, and below the result of a translation done by the system is shown that is our extension of the Google Translator system. Finally, it should be underlined that we already began to connect a rule based Georgian-English-German translator system elaborated by K.Pkhakadze, M.Chikvinidze, I.Beriashvili, A.Maskharashvili to the corpus. In particular, the methods are elaborated that will allow the knowledge base of this rule based system to be extended automatically.

**Acknowledgement.** The paper has been prepared in the GTU-GLTC with the financial support from the grant №31/70 by Shota Rustaveli National Science Foundation.

## REFERENCES

1. Pkhakadze K. The Technological Alphabet of The Georgian - The One of The Most Important Georgian Challenge of The XXI Century. (Georgian) *The Works of The Parliament Conference "The Georgian Language - The Challenge of The 21st Century"* (2013), 98-1005.

the simple sentences of the core part of Georgian language basing on K.Pkhakadze's Logical Grammar of the Georgian Language [7]. Here, also, it is noteworthy P. Meurer's [10] corpus (from this corpus in GI_Corpus we have taken the part of its words) and his restricted parser for Georgian (see `http://iness.uib.no/gekko/corpus-list?session-id=236474260345447 http://clarino.uib.no/iness/xle`)

2. Pkhakadze K., Chichua G., Chikvinidze M., Maskharashvili A. The short overview of the aims. methods, and results of the logical grammar of the Georgian language, *Rep. Enlarged Sess. Semin. I.Vekua. Appl. Math.*, **26**, (2012), 58-64.

3. Pkhakadze k., Chichua G., Chikvinidze M., Maskharashvili A. The project "foundations of logical grammar of Georgian language and its application in information technology" - grounding results and planed aims. *Proceeding of The International Scientific Conference Dedicated to the 90 anniversary of Georgian Technical University*, (2012), 138-146.

4. Pkhakadze K., Abzianidze L., Maskharashvili A. Georgian language's theses. *Seminar of I.Vekua Institute of Applied Mathematics REPORTS*, **34**, (2008), 108-121.

5. Pkhakadze K., Chichua G., Abzianidze L., Maskharashvili A. About 1-stage voice managed Georgian intellectual computer system. *I. Vekua Institute of Applied Mathematics*, **34**, (2008), 96-107.

6. Pkhakadze K. About achieved results and the future aims of the state program "free and complete programming inclusion of a computer in the Georgian natural language system", i.e. about necessity extension of the Georgian mathematical school with mathematical linguistics. *Reports of III republic seminar week in "Logic, Language and Computer", TSU I.Vekua Institute of applied mathematics*, (2007), 22-52.

7. Pkhakadze K. About Logical declination and lingual relations in Georgian. (Georgian) *Journal "Georgian language and logic", N1, "Universali",* (2005), 19-77.

8. Schäfer R. Bildhauer F. Web Corpus Construction. *Morgan Claypool Publishers*, 2013.

9. Abziandize L. An HPSG-based Formal Grammar of a Core Fragment of Georgian Implemented in TRALE. *Master's thesis, Charles University in Prague*, 2011.

10. Meurer P. A computational grammar for Georgian. *Logic, Language, and Computation. Springer-Verlag*, (2007), 1-16.

Received 09.05.2014; revised 21.10.2014; accepted 29.12.2014.

Authors' addresses:

K. Pkhakadze,[1] M. Chikvinidze,[1] G. Chichua,[1] A. Maskharashvili,[1,2] I. Beriashvili[1]
1: Scientific-Educational Center for Georgian Language Technology Georgian Technical University
Web-site: `http://gtu.ge/gltc/index.htm ; http://geoanbani.com/`
77, Kostava St., Tbilisi 0175
Georgia
E-mail: `gllc.ge@gmail.com`

2: Universite de Lorraine, Loria-Inria Nancy Grand-Est
615, rue du Jardin Botanique, 54500 Villers-les-Nancy
France