Seminar of I. Vekua Institute
of Applied Mathematics
REPORTS, Vol. 41, 2015

IN THE EUROPEAN UNION WITH THE GEORGIAN LANGUAGE - THE AIMS AND BASEMENTS OF THE PROJECT "ONE MORE STEP TOWARDS GEORGIAN TALKING SELF-DEVELOPING INTELLECTUAL CORPUS"

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

Abstract. The paper shortly overviews the aims and fundamentals of the two years project "A One More Step Towards Georgian Talking Self-Developing Intellectual Corpus" and the paper "Strategic Research Agenda for Multilingual Europe 2020" by the META-NET technological board. Also, taking into account the national aim of defending the Georgian language from the danger of digital extinction, as well as, the national aim of joining with the Georgian language the European Union, which according to the strategic research agenda of the Meta-Net is planned to become completely free from language barriers, the current paper underlines that the prioritization of the task of the complete technological foundation of the Georgian language, i.e. the task of creation of the Georgian thinking, speaking and translating system is the question of vital necessity for the Georgian society.

Keywords and phrases: Georgian self-developing intellectual corpus, technological alphabet of the Georgian language, logical grammar of the Georgian language

AMS subject classification (2010): 03B65, 68T50, 68Q55, 91F20.

Introduction

In 2010-2012 with the financial support of the European commission, there was carried out a research "Europe's Languages in the Digital Age" [1]. As a result, in 2012, Meta-Net published a press-release "At Least 21 European Languages in Danger of Digital Extinction - Good News and Bad News on the European Day of Languages" [2], and also Strategic Research Agenda for Multilingual Europe 2020" [3]. These publications, which are very important for us, are overviewed in the paper "Open Letter To The Georgian National Academy Of Sciences Id Est The Fact That European Languages Are At The Danger, Makes It Clear That The Georgian Language Is At Especially High Quality Danger! Id Est, Once Again For Defending The Rights Of The Georgian Language!! Id Est, It's Time To Take Care Of The Georgian Language!!! Short Version" [4]. - Here the main thing is that for today, in the European Union, processes are going on in concordance with the Strategic Research Agenda for Multilingual Europe 2020 with the aims of building such new Europe whose every citizen will be able to have access to any kind of service, knowledge, media, and technologies with their own mother language and, according to this agenda, in this new Europe, there will be no language barriers in communication, and there will be freely accessible high quality translations of domain independent as well as domain specific contents.

The coordinator of Meta-Net, Prof. Hans Uszkoreit, scientific director at German Research Center for Artificial Intelligence (DFKI) says the following: "The results of our study are most alarming. The majority of European languages are severely

under-resourced and some are almost completely neglected. In this sense, many of our languages are not yet future-proof." [2]

This all in sum once again make clear the urgent necessity of declaring as one of the main state priorities of Georgia the researches aimed at defending the Georgian language from the danger of digital extinction. There is also a clear necessity of formation a united Georgian group of researchers, which via collaboration with Meta-Net, will work on the tasks of complete mathematical and technological foundation of the Georgian language, in other words, on the task of creation of the high quality Georgian thinker, talker and translator system. - Without this type of system it will be impossible to join the European Union with the Georgian language, as well as, to defend the Georgian language from the danger of the digital extinction. For us it is clear that if we do not act in this way, and if we again do not manage properly the local processes with the aim of creation Georgian thinker, talker and translator system, i.e. if we continue chaotic, uncoordinated activities, like it is the case today, then the Georgian language will have the future about which Dr. Georg Rehm said in [2]: "There are dramatic differences in language technology support between the various European languages and technology areas. The gap between 'big' and 'small' languages still keeps widening. We have to make sure that we equip all smaller and under-resourced languages with the needed base technologies, otherwise these languages are doomed to digital extinction." - We say the same: We should be certain that we will be capable to defend the Georgian language from the very high danger of digital extinction in the digital age [5-8], and therefore, we should not act chaotically, but in an ordered manner, so that we could minimize today the existing gap instead of making it even bigger.

The aims and basement of the two year project "A one more step towards Georgian talking self-developing intellectual corpus". In 2012, in the Center for Georgian Language Technology at the Georgian Technical University, there was started a long-term project "The Technological Alphabet of the Georgian Language" [9 - 11] with K.Pkhakadze's leadership;[1] in the confines of this project, now center works on the $\mathcal{N}^{\circ}31/70$ project "Foundation of the logical grammar of the Georgian language and its applications in the information technologies" financed by Shota Rustaveli National Science foundation. In addition to it, within this long-term project, the center in March 2014 accomplished a project $\mathcal{N}^{\circ}048$ "Internet Versions of a Number of Developable (Learnable) Systems Necessary for Creating The Technological Alphabet of the Georgian Language "[2] financed by Georgian Technical University. Also, in 2012, there were started the two doctoral theses in the doctoral program "Informatics" at the Georgian Technical University, namely: Giorgi Chichua's doctoral thesis - "Georgian Speech Synthesis and Recognition", and Merab Chikvinidze's doctoral thesis -

[1]This long-term project was elaborated via the further development and completion of a state priority program of the Iv. Javakhishvili Tbilisi State University "Free and Complete Programming Inclusion of a Computer in the Georgian Natural Language System" [12 - 13], which was going on in previous years with K.Pkhakadze's leadership.

[2]The results of this project were successfully presented on the seminar "The Technological Alphabet Of The Georgian Language - One Of The Main Georgian Challenges Of The 21^{st} Century" held on 14 April 2014 that was dedicated to the day of the Georgian language.

"Georgian grammar checker (analyzer)" [14].

In 2014, on the basis of the results achieved within these above mentioned projects and doctoral theses, the center worked out a two year project "One More Step Towards Georgian Talking Self-Developing Intellectual Corpus", which is one more subproject of the long-term project "The Technological Alphabet of the Georgian Language" of the Center for Georgian Language Technology. This project, with which the Center applied for financing to Shota Rustaveli National Science Foundation, aims at building up a complete version of the Georgian self-developing intellectual corpus via further developing the trial version of the Georgian self-developing intellectual corpus, which is already created by us [15-23]. Thus, to build up the Georgian talking self-developing intellectual corpus means to create an automatically developing complete Georgian web-corpus which will be equipped with: the logic of the Georgian natural language systems; with the intellectual procedures constructed on the basis of this logic; and, also, with the Georgian technological alphabet, which is constructed on the basis of this logic and these intellectual procedures, in other words, with the Georgian talking Intellectual System, i.e., with the Georgian written and spoken texts analyzer and generator systems, which are necessary to realize full scale human computer intellectual interaction by means of the Georgian language. Besides it, to build up the Georgian talking self-developing intellectual corpus means to equip it with the two-way translator systems from Georgian to foreign languages, which, in turn, will be constructed on the basis of the above-mentioned Georgian talking intellectual system.

Obviously, it is impossible to build the above-described Georgian Talking Self-Developing Intellectual Corpus in the confines of one two-year project. Therefore, this two year project aims at building above-described Georgian corpus as complete as it is possible, and, also, the project aims to provide the Georgian language with all the necessary resources that are needed in order to be able to participate in those processes that are already going on in concordance with the strategic research agenda for multilingual Europe 2020. - In our opinion, this is the only way to defend the Georgian language from digital extinction in the digital age.

Below, we will very briefly present those results on which the project is based on; they are as follows:

1. A trial version of the Georgian self-developing multilingual and multimodal intellectual web-corpus [15], which despite that it is still only trial one contains already over 144 126 000 words, among which 2 267 700 words are mutually different, and it is already equipped with trial versions of the Georgian intellectual procedures and technological systems, which are listed below and some of which even are unique (see: http://geoanbani.com/Corpus/):

–Taggers, descriptors and generators of the words of the types of V, N and A [16];

–Self-developing syntactic/orthographic spellcheckers and Georgian orthographic corrector [17];

–Georgian-Mathematical/Georgian-English-German translators [18];

–Speech recognizers based on teaching and studying principles [19, 20];

–Georgian e-text and web-page reader [21];

–Georgian multilingual speech assistant and Georgian Spoken Support for Persons with Speech Disorder [22];

–Georgian Multi-lingual Spoken Lexicon and Georgian Extension of Google Translator [23].

2. The foundations of the logical grammar of the Georgian language [24-28], which is elaborated within the project $\mathcal{N}^{o}31/70$ "Foundations of the Logical Grammar of Georgian Language and its Applications in the Information Technologies", and which, on the one hand, is the first logical grammar of the natural Georgian language system. On the other hand, the above-listed intellectual procedures and technological systems are created on the basis of this logical grammar of the Georgian language.

The importance and benefits of the two year project "A one more step towards Georgian talking self-developing intellectual corpus".

For today, the Georgian language in the sense of language resources (resources, data and knowledge basis) and technologies (tools, technologies, applications) is very poorly supported. Even more, the Georgian language is alarmingly lagging compared to almost any of those 21 European languages, which according to the research "Europe's Languages in the Digital Age" [1- 3] done by META-NET, are under the danger of digital extinction in the digital age. All these together clearly indicate the urgent necessity of reducing this lagging as much as it is possible and as soon as it is possible. The aim of two year project "One More Step Towards Georgian Self-Developing Intellectual Corpus" is to reduce this lagging in the shortest possible period, and consequently, to radically change the current state of affairs.

Indeed, in the case of successful completion of the project, which is truly realistic taking into account our existing results that serve as the foundation for the project, in the summer 2017, there will be already built the Georgian self-developing intellectual corpus, i.e. the Self-developing Georgian-net, which will be equipped with the continuously developing Georgian text analyzer *(such as: automatic descriptor of tokens and descriptive databases (that define knowledge and logic of the corpus), automatic extender of intellectual procedures; morphological and syntactic structure generators for words and composed linguistic expressions; the hybrid morphological, syntactic and semantic checker; the Information/knowledge extractor, question-answerer, and logical problem solver-checker), speech processor (such as: the Georgian e-texts semantic reader equipped with possibility to built in it users own voice; the recognizer of synthesized and natural speeches; the various kinds of segmentators of voice and subtitled voice data), automatic translator (such as: the rule based Georgian-English-German and Georgian-Mathematical translators; the hybrid Georgian-English-German translator; the Georgian extension of Google translator; the Georgian spoken lexicon)* and the corpus voice manager systems. In addition, the Georgian-net, i.e. the Georgian self-developing intellectual corpus, from the day of its launch, will extend automatically itself with Georgian and Georgian-foreign texts freely available in the web in a such a way that it will be able to record the source and date of entrance of any newly added Georgian words in it and, accordingly, in the Georgian web space. - It is absolutely obvious that here very shortly but almost completely described the Georgian self-developing intellectual corpus or, shortly, the Georgian-net, from the point of view of technological support, will essentially reduce the existing alarming lagging

with technologically advanced languages.

Besides, if we take into account that within the project it is planned to build Georgian_Thinker&Talker&Translator_1 web-system and mobile apps some of its modules (they are: Georgian multilingual spoken lexicon, Georgian extension of Google translate, Georgian multilingual speech assistant, Georgian e-text and web-page reader), and also to publish monographic work "The Georgian Web-Corpus: Aims, Methods, and Recommendations", it gets even clearer that the project has very high or even groundbreaking importance for the scientific community that is concerned with building Georgian information technology systems.

Acknowledgement. We gratefully acknowledge that the paper was supported with the Shota Rustaveli National Science Foundation grant №31/70 for the project "Foundations of Logical of Georgian Language and Its Application in Information Technology" and with the grant Georgian Technical university grant №048-13 for the project №048 "Internet Versions of a Number of Developable (Learnable) Systems Necessary for Creating The Technological Alphabet of the Georgian Language".

REFERENCES

1. Meta-net white paper series, Europes languages in the Digital Age (32 Different Papers http://www.meta-net.eu/whitepapers/overview), Editors: Georg Rehm, Hans Uszkoreit, Springer, 2012.

2. META-NET At Least 21 European Languages in Danger of Digital Extinction - Good News and Bad News on the European Day of Languages, http://cordis.europa.eu/fp7/ict/language-technologies/docs/metanet-white-paper-press-release-english-international.pdf, September 26, 2012.

3. Presented by the META Technology Council, Strategic Research Agenda For Multilingual Europe 2020, (http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf), *Springer*, 2012, 1-87.

4. Pkhakadze K. Open Letter To The Georgian National Academy Of Sciences Id Est The Fact That European Languages Are At The Danger. Makes it clear that the Georgian Language Is At Especially High Quality Danger! Id Est, Once Again For Defending The Rights Of The Georgian Language!! Id Est, It's Time To Take Care Of The Georgian Language!!! Short Version. *Journal "Georgian Language and Logic"*, **7-8** (2014), 1-20.

5. Pkhakadze K. For Protecting Rights of the Georgian Language. *Journal "The Georgian Language and Logic"* **3-6** (2005), 1-11; 2007, 83-109.

6. Pkhakadze K., Gabunia K., Chichua G., Maskharashvili A., Abzianidze L.,Vakhania N., Pkhakadze N., Chikvinidze B., Gurashvili L., Labadze N., Beriasvhvili M. The Aims of Constructing Georgian Intellectual Computer System and Cultural Perspectives of Georgian Language. *Proceedings of the VI Conference of Arn. Chikobava Institute of Linguistics In "Natural Language Processing"*, 2008, 23-24; 2008, 33-34.

7. Pkhakadze K., Chichua G., Vashalamodze A., Gabunia K., Abzianidze L., Maskharashvili A., Chikvinidze M. The Aims of Elaborations of the Mathematical theory of the Georgian Language and Thinking and the threat in front of the Georgian language. *St. Andrew First Called Georgian University of the Patriarchy of Georgia*, 2009, 1-24.

8. Pkhakadze K., Chichua G., Chikvinidze M., Maskharashvili A. The Georgian Language in the Digital Age and the Technological Alphabet of the Georgian Language. *Book of Abstracts of V International Conference of Georgian Mathematical Union*, 2014, 143-144.

9. K. Pkhakadze. The Technological Alphabet of The Georgian Language The One of The Most Important Georgian Challenge of The XXI Century. *The Works of The Parliament Conference "Georgian Language - The Challenges of The 21st Century"*, 2013, 98-105.

10. Pkhakadze K., Chikvinidze M., Chichua G., Maskharashvili A. The Technological Alphabet of the Georgian Language - Aims, Methods, Results. *Reports of Enlarged Session of the Seminar of I. Vekua Institute of Applied Mathematics*, **27** (2013), 46-49.

11. Pkhakadze K. The Aims and Problems of Creation of the Technological Alphabet of the Georgian Language, Book of Abstracts of III International Conference of Georgian Mathematical Union, pp.59-60, 2012.

12. Pkhakadze K., Chichua G., Abzianidze L., Maskharashvili A. About 1-Stage Voice Managed Georgian Intellectual Computer System, (This work was carried out with the aims of the State Priority Program "Free and Complete Inclusion of a Computer in the Georgian Natural Language System" (2003)). *Seminar of I. Vekua Institute of Applied Mathematics, REPORTS*, **34** (2008), 91-102.

13. Pkhakadze K. Globalization, The Georgian Language and The State Priority Program "Free and Complete Programming Inclusion of a Computer in the Georgian Natural Language System". *Journal "The Georgian Language and Logic"*, **2** (2005), 1-11.

14. Pkhakadze K., Chichua G., Chikvinidze M. The Project "Foundations of Logical Grammar of Georgian Language and Its Application in Information Technology" and Doctoral Themes "Georgian grammar checker (analyzer)" and "Georgian speech synthesis and recognition". *Journal "Georgian Language and Logic"*, **7-8** (2014), 21-36.

15. Pkhakadze K., Chikvinidze M., Chichua G., Maskharashvili A., Beriashvili I. 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. *Reports of Enlarged Session of the Seminar of I. Vekua Institute of Applied Mathematics*, **28** (2014), 67-75.

16. Pkhakadze K., Chichua G., Chikvinidze M., Maskharashvili A., Beriasvili I. A Trial Version of the Georgian Self-Developing Intellectual Corpus and the Aims of Construction of Such Type WebCorpus. *Book of Abstracts of V International Conference of Georgian Mathematical Union*, 2014, 144-146.

17. Pkhakadze K., Chikvinidze M. The Inbuilt Systems of the Georgian Self-Developing Grammatical and Orthographical Checker in the Georgian Self-Developing Intellectual Corpus, *Book of Abstracts of V International Conference of Georgian Mathematical Union*, 2014, 140-141.

18. Pkhakadze K., Chikvinidze M., Chichua G., Maskharashvili A., Pkhakadze N., Beriashvili I. The Inbuilt System of the Georgian-English-German Translator in the Georgian Self-Developing Intellectual Corpus. *Book of Abstracts of V International Conference of Georgian Mathematical Union*, 2014, 142-143.

19. Pkhakadze K., Chichua G., Chikvinidze M. The Inbuilt Trial System of the Georgian Speech Recognition and Voice to Voice Translator in the Georgian Self-Developing Intellectual Corpus. *Book of Abstracts of V International Conference of Georgian Mathematical Union*, 2014, 139-140.

20. Chichua G., Pkhakadze K. The Georgian Speech Alphabet and the Task of the Recognition of the Georgian Speech. *Book of Abstracts of IV International Conference of Georgian Mathematical Union*, 2013, 145.

21. Pkhakadze K., Chichua G., Chikvinidze M. Internet Version of the Georgian None Semantically Text Reader System and First Step Toward Constructing Georgian Semantically Reader-Listener System. *Book of Abstracts of IV International Conference of Georgian Mathematical Union*, 2013, 149-150.

22. Pkhakadze K., Chichua G., Chikvinidze M. Georgian Spoken Support for Persons with Speech Disorder. *Book of Abstracts of V International Conference of Georgian Mathematical Union*, 138-139, 2014.

23. Pkhakadze K., Chichua G., Chikvinidze M., Maskharashvili A., Pkhakadze N., Beriashvili I. The Inbuilt System of the Georgian Extension of the Google Translate in the Georgian Self-developing Intellectual Corpus. *Book of Abstracts of V International Conference of Georgian Mathematical Union*, 146-147, 2014.

24. Pkhakadze K., Chichua G., Chikvinidze M., Maskharashvili A., Pkhakadze N., Beriashvili I., Kurtskhalia D. The Logical Grammar of the Georgian Language as the Theoretical Background

of the Georgian Intellectual Corpus. *Book of Abstracts of V International Conference of Georgian Mathematical Union*, 2014, 147-148.

25. 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 the Georgian Technical Conference University*, 2012, 138-146.

26. Pkhakadze K. The Overview of the Fundamental Questions of the Logical Grammar of the Georgian Language. *Book of Abstracts of IV International Conference of Georgian Mathematical Union*, 2013, 51-52.

27. 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. *Reports of Enlarged Session of the Seminar of VIAM*, **26** (2012), 58-64.

28. Pkhakadze K. About Logical Declination and Lingual Relations in Georgian. (Georgian) *Published in the Journal "Georgian language and logic", N1, "Universali"*, 2005, 19-77.

Received 10.06.2015; revised 25.11.2015; accepted 01.12.2015.

Authors' addresses:

K. Pkhakadze, M. Chikvinidze, G. Chichua, I. Beriashvili,
D. Kurckhalia, N. Pkhakadze
Scientific-Educational Center for Georgian Language Technology
at the Georgian Technical University
Georgian Technical University
77, M. Kostava St., Tbilisi 0175
Georgia
E-mail: gllc.ge@gmail.com

A. Maskharashvili
Loria-Inria Nancy Grand-Est
615, Rue du Jardin botanique, 54600 Vandœuvre-lés-Nancy
France