*SOCIETY. INTEGRATION. EDUCATION*
*Proceedings of the International Scientific Conference. Volume I, May 26th, 2023. 180-189*

# THE CORPORA-ORIENTED PROJECTS AND COURSES – INNOVATION OF THE UNIVERSITY LIFE

**Darejan Tvaltvadze**
Ivane Javakhishvili Tbilisi State University, Georgia

**Irina Gvelesiani**
Ivane Javakhishvili Tbilisi State University, Georgia

***Abstract.*** *The advancement of computer technologies opened up a bunch of opportunities in different spheres of life. Computer science has revolutionized the modes of studying and researching. The educators and policy makers started modernization of educational programs in accordance to the latest achievements in the field. The present paper deals with the presentation of the progressive university courses as well as the projects (winter schools, Akaki Shanidze's Digital Library and Text Corpus, etc.) facilitating the integration of corpora in teaching and researching. The major accept is put on the successful modes of the enhancement of the learners' digital skills (operation, application, problem solving, critical thinking, etc.), corpus-building abilities (annotation, orientation towards the scientific metalanguage, enrichment of metadata, etc.), corpora-based analysis (proper coinage, finding equivalency, plying with concepts and terms, etc.) as well as the corpus-oriented research. Determination of the major challenges and outcomes of the projects and courses, discussion of some gaps, introduction of the convenient opportunities of their "filling" – these are the major issues of the paper. The methodology of research includes observation, analysis and evaluation of the ongoing processes.*
***Keywords:*** *corpora-based analysis, digital library, educational program, language corpus, project.*

## Introduction

The way we live and exist greatly depends on today's world, its challenges, tendencies and perspectives. Nowadays, we may definitely say that COVID-19 drastically changed our lives by shifting the world population from socialization to desocialization and from a face-to-face interaction to an online communication. As a result, we became dependent on computer technologies, electronic data, digital factories, etc.

However, a new "electronic" stage of life started before COVID-19, namely, in the beginning of the 21st century. It involved the digitalization of every field of science, even the humanities. Nowadays, scientists speak a lot about digital humanities, which interprets the cultural and social impact of information


*https://doi.org/10.17770/sie2023vol1.7106*

technologies and creates or/and applies these technologies to interrogate cultural, social, historical and philological questions (Mozhaeva & Mozhaeva Renha, 2016). Moreover, it is usually treated as "a firstborn of the science of the 21st century, whose origin was determined by the ongoing evolution of human consciousness and being, and became the logical and inevitable result of the digital era" (Tandashvili & Kamarauli, 2021).

Although Schreibman et al. (2004) suggest that we consider digital humanities as a discipline in its own right, its purpose and status have continued to be a subject of numerous debates and its very nature is still being negotiated (Luhmann & Burghardt, 2021). We can only admit that digital humanities should be treated as a transdisciplinary scientific field. This fact is definitely evidenced by numerous innovative, application-oriented resources, which were created as a result of a close cooperation of scholars representing different fields of science. These resources are language corpora, digital dictionaries, computer and mobile applications, digital libraries, digital archives, etc. (Tandashvili & Kamarauli, 2021).

A special attention should be paid to a language corpus (pl. corpora), which has become "all the vogue" and "a profound sensation" during the last decades. Scholars believe that a corpus is "a large collection of naturally occurring examples of a language stored electronically" (Bennet, 2010). Its usage during classroom activities has changed a teaching landscape. Moreover, the corpus-based pedagogy has become extremely popular and useful.

The present paper deals with the presentation of the progressive university courses and projects facilitating the integration of corpora in teaching and researching. The major accept is put on the successful modes of enhancement of learners' digital skills, corpus-building abilities, corpora-based analysis, etc. The methodology of research includes observation, analysis and evaluation of the projects and their already-achieved/future outcomes.

## Development of Corpora and Corpora-Oriented Projects

It is generally assumed that corpus linguistics and corpora are the "products" of the previous century. However, some scholars believe that they existed much earlier. According to Kennedy, "there was a tradition of linguistic analysis based on corpora prior to the nineteenth century, long before the arrival of computers, in the context of biblical and literary studies, in lexicography and dialectology, in language education studies and in grammar studies (Assunção & Araújo, 2019). Svartvik and Stubbs present the list of a few examples of 'language corpora BC (before computers)'. The most notable one is the corpus of 5 million citation slips compiled by the volunteers in the second half of the 19th century (Biel, 2009). However, the first computer-based corpus, *Brown Corpus*, was

created in 1961 (Bennet, 2010) and at the end of the 20th century corpus linguistics started flourishing.

It is noteworthy that the development of corpus linguistics and corpora-based researches in Georgia is associated with the name of German linguist and Caucasiologist Jost Gippert from Goethe University. He created the first electronic resource for the Georgian language within the framework of TITUS (*Thesaurus Indogermanischer Text- und Sprachmaterialien*), which was a large-scale electronic platform/thesaurus of the Indo-European languages. Jost Gippert has actively collaborated with the Georgian scholars, who have worked at different higher educational institutions of Georgia (Khalvashi, 2018). The history reveals the products of the cooperation – different bilateral projects oriented towards the creation of corpora. The most prominent product – *The Georgian National Corpus (GNC)* – which covers the complete time range from the earliest attestations of written Georgian in the 5th century up to the present day evolved from several corpus building initiatives that have been realized since the late 1980s, mostly in joint endeavors of the German and Georgian partners (Gippert & Tandashvili, 2015). The major importance of *The Georgian National Corpus* lies in the fact that it unites the work of the scholars and students. Among them are the representatives of Ivane Javakhishvili Tbilisi State University (TSU). They work in two directions – *digitalization of Georgian literary monuments* and *annotation of already-existed texts*. The latter is a form of the enrichment of electronic data with a linguistic meta-information presented on different independent levels. The most complete form, the multi-level annotation, characterizes a lexical unit from a lingual point of view and presents the data of alliteration as well as equivalency (namely, the English counterparts).

In addition to GNC, Georgia's digital reality encompassed different projects oriented towards the development of students' digital skills. They were authorized by different universities, for instance:

- The seasonal schools initiated by Batumi State University – *Digital Humanities and Kartvelology*; *Digital Humanities and Language Documentation,* etc.
- The projects initiated by Ilia University – *The Epigraphic Corpus of Georgian; Georgian Language Corpus; Prosopography of Georgia,* etc.
- The projects initiated by TSU – *Digital Humanities – Kartvelology and the Challenges of the 21st century*; *Development and Introduction of Multilingual Education Programs at Universities of Georgia and Ukraine; Akaki Shanidze's Digital Library and Text-Corpus,* etc.

Let us discuss two winter schools – "*Digital Humanities – Kartvelology and Challenges of the 21st Century*" and *"Digital Kartvelology – Thematic Corpus and Annotation Issues in the Kartvelian Languages"* – organized by Tbilisi State University and Goethe University. Both winter schools gathered the students of different universities of Georgia.

"*Digital Humanities – Kartvelology and Challenges of the 21st Century*" was held in Bakuriani. Within the framework of this project the Georgian and foreign prominent scholars delivered the lectures on the following topics: *Methods of corpus linguistics; Practical effects of corpus-oriented research; Information structuring issues in Georgian; Problems of an interlinear annotation; Corpus research perspectives – modality, e-learning platforms; Georgian Dialect Corpus – interdisciplinary research resource, National Corpus of the Georgian Language – importance and prospects*, etc. Moreover, the accent was put on the importance of the utilization of *OLAT* (*Online Learning and Training*) – one of the most developed e-learning programs, which is oriented on self-control, self-education and self-development. The emphasis was also put on the introduction of *eLecture* – a new format of teaching, which is based on the electronic visualization of the taught material.

The winter school *"Digital Kartvelology – Thematic Corpus and Annotation Issues in the Kartvelian Languages*" was held in Tbilisi. Within the framework of this project the Georgian and foreign scholars delivered the lectures on the following topics: *Basic principles of creating a corpus; Linguistic portrait of Georgia – Georgian thematic corpus; Management of a thematic corpus; Basic annotation schemes and models in corpus linguistics; Basic principles of linguistic annotation; Interline annotation and basic principles of glossing*, etc. Moreover, the principles of utilization of *OLAT* and *eLecture* were discussed.

At the end of both winter schools, almost all students made very interesting presentations created on the basis of the studied topics. They illustrated the acquired knowledge in the field of corpus linguistics, for instance, creation and management of a corpus, principles of information structuring, interlinear annotation and glossing, modality, e-learning platforms, etc. As a result, they demonstrated the practical skills of digital research and structuring a thematic corpus.

After the completion of the winter schools, the best students were involved in the new project *"Akaki Shanidze's Digital Library and Text Corpus"*, which started in 2017. The project aimed at the digitalization of the scientific heritage of famous Georgian scholar Akaki Shanidze, whose scientific works cover different directions: old and new Georgian, dialectology, the history of the Georgian language, the unwritten Kartvelian languages, epigraphic works, lexicology, Rustvelology, etc.

While working on the project, the students performed the following tasks:

- Creation of the electronic versions of the texts in accordance with the international coding standard (UNICODE) (conversion of a text into a digital format);
- Intrastructural processing of the text from the point of view of reference (an electronic version must accurately reflect a structure of a document);

- Processing of the text from the point of view of metadata (entering a text into a special database to make easier for a user to find corpus materials according to relevant signs).

The final products of the project are Akaki Shanidze's text corpus and digital library containing several volumes of his scientific works as well as "Khanmeti Lectionary", "Khanmeti Multichapter", "The Typicon of Petritsoni Monastery" and the prefaces written for the books published under his editorship. In addition, the digital library presents Akaki Shanidze's biography, bibliography, annotated photo archive as well as the books and articles about him.

Akaki Shanidze's text corpus meets the following criteria:

- It is digitized i.e. transferred to the electronic media in order to exist in an electronically processable form (i.e. texts structured with special marks);
- In addition to the primary i.e. linguistic data, it contains secondary information – metadata and linguistic annotation;
- It is equipped with the special corpus management system – the corpus manager (Tvaltvadze, 2019-2020, p. 76).

Akaki Shanidze's text corpus has the multifaceted search program, which facilitates the creation of corpus-based as well as corpus-oriented publications. Moreover, it may be treated as a sub-corpus of the Georgian scientific metalanguage corpus, which enriches the existing database with the new resources.

## Corpus-Oriented Courses

Corpus linguistics is one of the technology-based tools that could be very useful in teaching, but still has not been widely used or tested at the higher educational institutions (Dazdarevic, Zoranic, & Fijuljanin, 2015). Nevertheless, Tbilisi State University, as a driving force of the Georgian educational space, has already implemented several corpus-oriented courses, because a direct application of corpora and corpus tools in a classroom support language teaching theories and concepts related to a learner autonomy, use of *realia* and authentic texts, learner-computer interactions and explicit teaching of language features or patterns (Friginal, Dye & Nolen, 2020). Moreover, corpus-based lessons with an appropriate amount of students' interactions and language use opportunities can stimulate learners' interest and improve a learner autonomy (Ma & Mei, 2021).

Let us discuss the BA courses delivered at two directions of the Faculty of Humanities: English Philology and Georgian Philology. The students of English Philology get acquainted with corpus linguistics and corpora while attending two elective courses: "*Foundations of the Lexicography of the English Language*" and "*Abstracting and Reviewing of the English Text*" (ARET). The former is oriented towards teaching the theoretical issues, namely, corpus linguistics, different types

of corpora, development of corpus linguistics in Georgia, peculiarities of Georgian National Corpus, GEKKO pillar, etc.

"*Abstracting and Reviewing of the English Text"* is oriented to multiethnic groups of students and considers the modern approaches to teaching the vocabulary, grammar and translation. The accent is put on the acquisition of the specialized lexical units via labelling/coinage, plying between terminological units as well as corpus-based analysis. Moreover, ARET deals with the practical aspects, namely, the translation of publications from Georgian into English and vice versa. During translation, the learners are allowed to use different bilingual or multilingual dictionaries as well as online corpora. Specific words are defined by means of a corpus-based analysis i.e. searching for the meaning via determining a proper context and a sentential environment. If the equivalency is not determined, new terms are coined. Accordingly, ARET enables the students to make a practical realization of the knowledge acquired during "*Foundations of the Lexicography of the English Language*".

Within the course "*Abstracting and Reviewing of the English Text"*, students are required to prepare a presentation. One of the necessary conditions is searching for an appropriate empirical material through different types of corpora.

The students' knowledge gained during attending ARET is measured by the midterm and final tests. They are oriented towards checking the theoretical knowledge (topic/topics – score 10/20) and practical skills (multiple choice – score 10; translation – score 10). The following table presents the achievements of three groups of students (total number – 61) attending ARET during fall semester.

*Table 1* ***The results of the midterm and final tests – fall semester, 2022*** *(made by authors)*

| Percentage | The number of students, who wrote the midterm test (the highest score – 30) | The number of students, who wrote the final test (the highest score – 40) |
|---|---|---|
| 91%-100% of the highest score (HS) | 20 | 22 |
| 81%-90% of HS | 18 | 21 |
| 71%-80% of HS | 11 | 11 |
| 61%-70% of HS | 7 | 4 |
| 51%-60% of HS | 5 | 3 |

The above table reveals the students' success throughout the semester that definitely indicates to the usefulness of ARET.

The third BA course oriented towards the usage of corpora is "*The History of the Georgian Literary Language*". It is delivered to the fourth-year students of the direction of Georgian Philology. "*The History of the Georgian Literary Language*" summarizes the knowledge acquired during four years and analyzes

the issues from the synchronic and diachronic viewpoints. It belongs to the group of compulsory courses, whose e-system presents the syllabus, presentations of lectures, teaching materials, tasks and news forum.

Within the course "*The History of the Georgian Literary Language*", students are required to prepare an essay and a presentation. A necessary condition for the preparation is searching for an appropriate empirical material through the electronic databases and corpora presented in the e-system of the course: The Georgian National Corpus and its constituents (*TITUS electronic text base* (University of Frankfurt); *ARMAZI electronic text base* (University of Frankfurt); *GEKKO - Georgian electronic corpus analyzer* (Norway); *Georgian Dialect Corpus* (Georgia), etc.), the corpora created by Ilia University, etc.

During the lectures and seminars, students study how to deal with the mentioned sources. They are instructed by the lecturers, who participated in the seasonal schools and attended the appropriate training courses as well as workshops. Moreover, the flexible search system simplifies research, facilitates the analysis of word forms, makes statistics and draws reliable conclusions. Consequently, a working process becomes easier and more enjoyable. Students learn about modern methods and technologies and use them to conduct their scientific research. The following table presents the achievements of 87 students, who wrote essays while attending "*The History of the Georgian Literary Language*".

*Table 2* ***The results of the assessment of the essays – spring semester, 2022***
*(made by authors)*

| Percentage | The number of students, who wrote an essay (the highest score – 7) |
|---|---|
| 91%-100% of the highest score | 45 |
| 81%-90% of the highest score | 27 |
| 71%-80% of the highest score | 6 |
| 51%-70% of the highest score | 7 |
| The rest | 2 |

## Discussion

The 21st century – the century of technological advancements – gradually shifts the society to the stage of the electronic evolution. All fields of science are revolutionized. The overwhelming technological progress sets new goals before educators and policy makers. Tbilisi State University**,** a leading educational body of the Caucasus region, strives to implement innovative strategies via the cooperation with the prominent western educational institutions. This collaboration is reflected in the bilateral projects oriented towards the creation of different types of corpora and seasonal schools. The latter aim at the development

of students' global skills by introducing the theoretical foundations of corpus-oriented research, basic principles of structuring thematic corpora, glossing, annotation, etc.

The skills acquired during seasonal schools are developed while carrying out the local projects. One of them is *"Akaki Shanidze's Digital Library and Text Corpus"*. It facilitates the presentation of the eminent scholar's name and heritage on Georgia's "digital map". This project has an outstanding importance. On the one hand, scholars and students specialized in the Georgian philology, Kartvelology and related branches may use digitally-presented materials on every stage of studying and researching. On the other hand, student-participants' digital skills (operation, application, etc.) and corpus-building abilities (annotation, orientation towards the scientific metalanguage, enrichment of metadata, etc.) are enhanced.

The raise of digitally-skilled generation is facilitated by delivering corpora-oriented courses at the Faculty of Humanities of TSU. The paper discusses "*Foundations of the Lexicography of the English Language*", "*Abstracting and Reviewing of the English Text*" and "*The History of the Georgian Literary Language*". The former is oriented towards teaching the theoretical issues, while others deal with the enhancement of learners' digital skills (problem solving, critical thinking, etc.), corpora-based analysis (finding equivalency, plying with concepts and terms, etc.) as well as the corpus-oriented research. As a result, the process of learning becomes active and student-centered. The concepts of "traditional teacher-dominated classroom" and marginalized students disappear.

Moreover, Tbilisi State University seems to be the only Georgian university, which offers students two corpora-oriented courses "*Foundations of the Lexicography of the English Language*" and "*Abstracting and Reviewing of the English Text*". The majority of BA programs in English Philology of Georgia's higher educational institutions incorporate "*Foundations of the Lexicography of the English Language*" or "*Foundations of the Lexicography*". ARET is offered only by Samtskhe-Javakheti State University. However, its BA program in English Philology does not incorporate "*Foundations of the Lexicography of the English Language*" or "*Foundations of the Lexicography*". Some non-corpora oriented issues related to the lexicography are presented in the course "*Lexicology of the English Language*", which will be renamed into "*Lexicology-Lexicography of the English Language*" (National Centre for Educational Quality Enhancement at the Ministry of Education and Science of Georgia, 2023). Accordingly, the above-mentioned shows the priority of TSU during dissemination of corpora-related knowledge.

However, taking a closer look at the above-mentioned projects and courses reveals certain gaps, for instance, the courses "*Foundations of the Lexicography of the English Language*" and "*Abstracting and Reviewing of the English Text*" are elective. Accordingly, those BA students of the direction of English philology,

who do not choose them, will not get an appropriate knowledge. It is recommended to make these courses compulsory.

Moreover, "*Abstracting and Reviewing of the English Text"* is a one-semester course. It is recommended to deliver ARET during two or three semesters, because this is the only BA course of the direction of English philology, which focuses simultaneously on translation and intensive corpora-based analysis, coinage of new terms and corpora-oriented research. Moreover, it develops the transferable skills that form a solid foundation for being used at the next level of education or during practical activities.

Finally, it is noteworthy that the seasonal schools are periodically organized by different universities of Georgia, especially, by Tbilisi State University and Batumi State University. It is preferable to organize these schools systematically in order to deepen more students' corpora-oriented skills and make them competitive in the global arena.

## Conclusions

In the 21st century, the world population passes through a new "electronic" stage of life. The advancement of computer technologies makes impact on every branch of science. The advent of corpus linguistics revolutionizes a linguistic research as well as methods of teaching and learning. The challenges of the new era stipulate the creation of projects and courses facilitating the development of learners' digital skills. TSU is one of the leading educational institutions in this respect. The paper presents the projects and courses, which enhance the learners' digital skills, corpus-building abilities, corpora-based analysis, corpus-oriented research, etc. Participation in the projects and attendance of the mentioned courses raise digitally-oriented generation, which becomes competitive throughout the world. Filling the existing gaps will make the courses and projects more progressive. This will be beneficial to the learners, university and country.

Moreover, the highlighted projects or courses may serve as exemplary models for those educational institutions of the developing countries, which strive to implement innovative student-oriented and digitally-enriched strategies of teaching and researching.

## References

Assunção, C., & Araújo, C.S. (2019). Entries on the history of corpus linguistics, *Linha D'Água*, *32*(1), 39-57.

Bennet, G. (2010). *Using corpora in the language learning classroom: corpus linguistics for teachers.* Michigan: Michigan ELT.

Biel, L. (2009). Corpus-based studies of legal language for translation purposes: methodological and practical potential. In C. Heine, & J. Engberg (Eds.), *Reconceptualizing LSP* (1-15). Online proceedings of the XVII European LSP

Symposium 2009. Retrieved from: Biel - 2010 - Corpus-Based Studies of Legal Language for Translation Purposes Methodological and Pra - 1 Łucja Biel* Corpus-Based Studies of Legal | Course Hero

Dazdarevic, S., Zoranic, A., & Fijuljanin, F. (2015). Benefits of corpus-based approach to language teaching. *Balkan Distance Education Network - BADEN Newsletter, 7.*

Friginal, E., Dye, P., & Nolen, M. (2020). Corpus-based approaches in language teaching: outcomes, observations, and teacher perspectives. *Boğaziçi University Journal of Education, 37*(1), 43-68.

Gippert, J., & Tandashvili, M. (2015). Structuring a diachronic corpus. The Georgian National Corpus project. In J. Gippert, & R. Gehrke (Eds.), *Historical corpora challenges and perspectives* (305-322). Tübingen: Narr Francke Attempto Verlag GmbH + Co. KG.

Khalvashi, R. (2018). *Introduction to digital humanities*. Batumi: Batumi Shota Rustaveli State University.

Luhmann, J., & Burghardt, M. (2021). Digital humanities - a discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. *Journal of the Association for Information Science and Technology, 73*(2), 148-171. DOI: https://doi.org/10.1002/asi.24533

Ma, Q., & Mei, F. (2021). Review of corpus tools for vocabulary teaching and learning. *Journal of China Computer-Assisted Language Learning, 1*(1), 177–190.

Mozhaeva, G., & Mozhaeva Renha, P. (2016). Digital humanities: to a question of the directions and prospects of development of interdisciplinarity in humanitarian researches and education. *SHS Web of Conferences, 26*, 1-6. DOI:10.1051/shsconf/20162601017

National Centre for Educational Quality Enhancement at the Ministry of Education and Science of Georgia. (2023). *Record of the meeting of the Accreditation Board of Educational Programs*. Retrieved from: Document (eqe.ge).

Schreibman, S., Siemens, R., & Unsworth, J. (2004). The digital humanities and humanities computing: An introduction. In S. Schreibman, R. Siemens, J. Unsworth (Eds.), *A companion to digital humanities*. Blackwell. DOI: https://doi.org/10.1002/9780470999875

Tandashvili, M., & Kamarauli, M. (2021). *Introduction to digital Kartvelology*. Tbilisi: Iverioni Publishing House.

Tvaltvadze, D. (2019-2020). Akaki Shanidze in the digital epoch (project "Akaki Shanidze's digital library and text corpus"). *Kartvelian Linguistics, 6-7*, 75-84.