VolkswagenStiftung FUNDED BY THE VOLKSWAGEN FOUNDATION, GERMANY

International Symposium

Digital Caucasiology

A Change of Paradigm?

OCT 4-8, 2023

OPENING:

October 4, 10:00
Eisenhower-Saal, Campus Westend,
I.G.-Farben-Haus, Goethe University,
Frankfurt am Main

SESSIONS:

Digital Manuscript Studies, Digital Lexicography, Digital Processing of Manuscripts and Inscriptions, Session for PhD Students and Post Docs, Digital Methods in Caucasian Studies, Perspectives of Digital Methods

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Institute of Empirical Linguistics, Goethe University Frankfurt

International Symposium

Digital Caucasiology – a Change of Paradigm?

(October 4–8, 2023)

Book of Abstracts

Frankfurt a/M

2023

Caucasiology diachronically

Caucasiology, i.e. the study of the peoples, languages and cultures in the Caucasus and their history, has always been an area that aroused the interest of researchers worldwide. In the 18th and 19th centuries, it was especially German (or German-born) scholars such as Güldenstädt, Klaproth, Uslar, Schiefner, Dirr, and Schuchardt who paved the way for a systematical development of Caucasiology. During the Soviet period, research concentrated in Georgia (at the Institute for Linguistics under the direction of Arn. Chikobava) as well as in Russia (Moscow, St. Petersburg). Numerous studies on different Caucasian languages originate from this time, illustrating the enormous linguistic diversity of the region.

Caucasiology today

In comparison with other disciplines of the Humanities, Caucasiology switched at a relatively early stage from traditional forms of research to new, digitally based methods of investigation. Meanwhile, corpora of both written and spoken languages (including documentations of endangered Caucasian languages) have been created in Europe, America, and the Caucasus itself, and the Georgian National Corpus (http://gnc.gov.ge) is one of the largest diachronic corpora world-wide (more than 220 million tokens). On this basis, we are today witnessing the development of “Digital Caucasiology” as a new interdisciplinary discipline in the Humanities, based upon corpora of different Caucasian languages, electronic dictionaries, and digital archives including audio-video-materials.

Goals of the Symposium

The symposium is meant to provide, for the first time ever, an international forum for the discussion of topical questions, with a view to developing common structures and to clarifying the relation between digital and “traditional” approaches. It aims to provide an overview of the existing digital resources that have been created by different scholars or groups at different research institutions, to enable an exchange of current research projects and approaches, to discuss the corpora and technologies of digital language processing and to identify possible cooperation plans. In addition, it is meant to be a platform for the exchange of views concerning the accessibility and usability of digital resources and their present and future impact on Caucasiology. For these purposes, the symposium will try to determine the present state of the art in the application of digital methods in Caucasiology on the basis of presentations concerning running projects.

Sessions of symposium:

Session I – Digital Manuscript Studies
Session II - Digital Lexicography
Session III - Digital processing of manuscripts and inscriptions
Session IV – Session for PhD students and Post Docs
Session V – Digital methods in Caucasian Studies
Session VI – Perspectives of digital Methods

Symposium is funded by Volkswagen Foundation

Jost Gippert (Hamburg)

A New Look on Georgian *Suffixaufnahme*: Finding out the True Rules

The German term *Suffixaufnahme*, first introduced by Franz Nikolaus Finck in his famous “Haupttypen des Sprachbaus” (1910: 141), denotes a peculiar characteristic of Old Georgian, namely, the extension of adnominal genitives by the case ending of the head noun they determine. Although the phenomenon has been dealt with in many grammatical treatises of the Georgian language, the rules of its application and their exceptions have not yet been clarified entirely, and some important questions are still open. Among them, we may mention, e.g., the questions whether or not the extension could be iterated more than two times, to what extent dissimilations of genitive endings could take place, or whether and to what extent *Suffixaufnahme* was also possible in a preposed position. Having the huge textual tradition of Georgian since the beginning of its literacy at hand in digital form, in the TITUS and GNC corpora, and having more and more online access to original manuscripts and to rare literature, these questions can now be addressed with much more chance of reliability, including statistical analyses that were never possible before. On this basis, the presentation will provide some new insights into the regularities under concern.

Bernard Outtier (Saint Martin de la Mer-Lavau)

Colligite fragmenta: Fragments of Dispersed Caucasian Manuscripts Virtually Reunited

Unfortunately, we do not have so many binders’ colophons in Caucasian manuscripts. Nevertheless, we may sometimes virtually re-unit fragments bound as fly-leaves in new manuscripts. One question arises: was it possible in every scriptorium to bind manuscripts? This will be illustrated with Armenian and Georgian fragments, either till now kept bound as fly-leaves or left today as separate, unbound, fragments.

Sergey Kim (Lausanne)

Transcribing the Old Georgian Minuscule Script: Some Perspectives

The paper discusses problems and perspectives of digital approaches toward Old Georgian manuscripts and Old Georgian calligraphy. Results of model training for Transkribus are presented, as applied to Old Georgian scripts of different periods.

Theresa Zammit Lupi (Graz)

A Codicological Description of the Georgian Lectionary in Graz

This presentation focuses on the codicological features of this rare seventh century Georgian manuscript. 2058/1 is a lectionary belonging to a set of five Georgian manuscripts coming from Mount Sinai that date between the seventh and eleventh centuries. The five items are housed in the Special Collections Department of Graz University Library.

From the existing literature it appears that the lectionary has already been studied for its content and philology, but not for the importance of its material nature. Elements such as the page layout, sewing structure, sewing routes, quire formation, spine preparation and cover are features that have not been investigated and therefore merit further research. The manuscript has also undergone various interventions over the course of its history, and these treatments too require analysis. Furthermore, there are loose fragments kept with the manuscript which seem to have been used as sewing guards. These recycled pieces of parchment also await study.

Codicology is not only about structure and function but also involves the study of materials and their identification. In this research paper I will discuss the different kinds of materials that were used to produce this manuscript such as parchment, inks and threads. It is hoped that this will shed new light on 2058/1 which will in turn contribute to the enormous puzzle in the understanding of the Georgian book production in the first millennium. Like palaeography and art historical analysis, codicology plays a likewise significant role in the study of any historic book. Up until recently the focus has always been on the text and/or decoration, but it is clear that the physical make-up of a book provides valuable evidence in equal measure.

Christa Müller-Kessler (Jena)

Georgian Manuscripts as a Gateway to the Early Christian Palestinian Aramaic Tradition

Christian Palestinian Aramaic, a branch of Western Aramaic, can be studied today due to the fact that it was preserved for the early period (5th to 7th century) in the form of palimpsests. These early manuscripts were re-used in later centuries by scribes writing Arabic, Georgian, Hebrew, or Syriac. In this context one particular Georgian scribe plays an important role. The tenth-century scribe Ioanne Zosime started to make use of these parchments written in Christian Palestinian Aramaic in the Laura of Mar Saba near Jerusalem in order to write new Georgian texts. He later moved these parchments to the Monastery of St Catherine on Mount Sinai, where he fulfilled the

role of a librarian. There he continued to dismember more of these Palestinian Aramaic manuscripts for his own purposes. Despite the dismantlement of this early text material, he helped to preserve unique texts relevant for the study of early church history by reusing them as palimpsests. Among them are the earliest witnesses of the Catecheses of Cyril of Jerusalem, the Dormition of Mary, the Old Jerusalem Lectionary, and unknown or very rare martyrdoms of which the Greek sources have been lost.

Donald Rayfield (London)

The Indigestible Impact of English on the Modern Georgian Lexicon

From 1990 a century-old trickle of anglicisms (American and British) into Georgian became a major torrent with several subsidiary streams. One stream is the introduction of new products (from cell-phones to pickup trucks), new customs (even morals), new commercial transactions etc., for which no existing words (as distinct from descriptive paraphrases) existed: more often than not, an Anglo-Saxon word was adopted. A second stream was a determination to rid certain spheres of Russianisms, notably the terms for automobile parts (a field in which Russians and Armenians, rather than Georgians did the dirty work) by devising (or reviving) words of Georgian origin, or importing Anglo-Saxon terms (a similar process has occurred in fields such as rugby football, played from 1959 to 1989 only against Slavonic teams, but now on a world basis). A third stream, after internet use became widespread for personal communication and commerce, was the devising of a terminology parallel to that of western internet users and commerce. A prime example is the word იშოპინგეთ!, ‘Continue shopping’. Native Georgian words and phrases flourished best in the area of slang and obscenity, since from 1990 words and phrases appeared in print that had never been read since the days of Sulkhan-Saba Orbeliani: here the absence of foreign influence (if we discount the Hebrew of Georgian Jews) is remarkable.

This study is based on internet chat forums, commercial material, plays (notably by dramatists who presented their work for translation and development at the Royal Court Theatre in London in 2013), novels by Lasha Bughadze etc. The provisional conclusion is that an alien flood, an indigestible input, has since 1990 struck the Georgian language, comparable to what happened with Iranian in the 10th and Russian in the early 19th century, but that, like earlier floods, a process of absorption and rejection will take place, leaving only a percentage of digestible material to be integrated with the body of the language. Much depends on the equilibrium of norms and innovations reached between a strong and conservative school-teaching establishment and an active and anarchic internet community.

Arguably, another important factor is to be found in the actual phonological structure of innovations: words such as რითრითი (‘retreat’) and დაიგუგლება (‘can be googled’) have a choice and distribution of consonants and vowels which integrate easily with Georgian, whereas აბეკაპებს (‘backs up’) still has to compete with the periphrastic Franco-Russian სარეზერვოდ ადუბლირებს.

The grammatical implications are also complicated: იშოპინგეთ is marked as a transitive verb with an ergative agent – a process that is also noted in other ‘ergative-inclined languages’, where an imported intransitive verb acquires ergative marking, e.g. Basque *nere kochek ez du funtzionatzen* ‘My car isn’t working.’

Ketevan Datukishvili, Nana Loladze, Merab Zakalashvili (Tbilisi)

“Lexicographer” – a Software Tool for Lexicographical Work

The program “Lexicographer“ is a software tool which supports lexicographic activities. It enables the creation of dictionaries in printed and electronic formats. “Lexicographer” is based on a conceptual model which reflects the structural elements (components) of dictionary units and governs their sequence.

The model has yielded a database in which the above-mentioned components are placed in separate fields. Thus, the material makes it possible to work efficiently using the dictionary. It is also possible to classify the data based on any parameter of the database and focus on each parameter separately.

An expert can process the data based on alphabetical order or selected by other parameters, for instance, semantic fields (plants, animals, technology), parts of speech (adverbs, interjections, particles, etc.), and so on. This kind of approach enables uniformity of information and definitions of lexical units belonging to the same type. It also reveals the conceptual attitude of the authors regarding concrete issues. Besides, such databases serve as tools for scientific research, in particular, statistical analyses.

The program “Lexicographer” has a design editor which enables various formats for different types of dictionaries. Each field in the database can be made in a desired style: font type and font size, colour, and so on. For example, the following components of dictionary units are given in different styles: title, definition, illustrative examples, etc. The sequence of components of dictionary units can also be altered.

The program tool enables the automatic preparation of the dictionary information. On the one hand, this saves time and human resources, and, on the other hand, it excludes “human errors”.

“Lexicographer” also enables the efficient administration of activities. The information in the database is available for every lexicographer, whereas the functions of editing and making entries are distributed: some experts can edit only a certain field of the database, others can only see the material, yet others can process the entire database and make changes in it.

“Lexicographer” was created by a team dealing with linguistic technologies who compiled and published two dictionaries: the printed “Georgian Dictionary” of 2014 (Bakur Sulakauri Publishing House), embracing more than 25 thousand dictionary units; and an electronic dictionary (https://www.ganmarteba.ge) of 2019. Currently, the electronic dictionary embraces 41455 units.

“Lexicographer” was created especially for explanatory dictionaries. However, it may also serve as a tool for compiling different other types of dictionaries (orthographic, terminological, and so on).

Tinatin Margalitadze (Tbilisi)

Digital Lexicography in Georgia

The history of digital lexicography in Georgia begins in the twenty-first century. Its development was hampered by the historical events connected with the break-up of the Soviet Union, political turmoil, difficult years following the declaration of independence by the country, etc. It should be noted though that the digitization of Georgian texts and the compilation of a Georgian National Corpus (http://gnc.gov.ge) started as early as the 1980s in collaboration of Georgian and German scholars under the scientific supervision of Prof. Jost Gippert from Goethe University in Frankfurt am Main (Gippert, Tandashvili 2015). The development of corpus linguistics in Georgia contributed a lot to the digitization of Georgian lexicography.

Digital lexicography in Georgia, like in the entire world, followed two main trends: (1) digitizing existing printed dictionaries and placing them on the Internet; and (2) the generation of online dictionaries, equipped with control panels allowing revision, updating and expanding the dictionary material.

The next stage in the development of Georgian digital lexicography was the composition of electronic and online dictionaries in special Dictionary Writing Systems (DWSs), rather than on the basis of .doc files (Margalitadze, Keretchashvili 2013). Within such a system, the dictionary material is distributed to separate fields of the database, which enables the generation of online and electronic dictionaries of much higher quality. Georgian Lexicographers started using in-house DWSs, although, at a later stage, they applied professional tools for the compilation of dictionaries (Margalitadze, 2018).

The development of corpus linguistics in Georgia created several monolingual, parallel, and dialect corpora, which paved the way for corpus-based lexicography. Entirely corpus-based Georgian dictionaries opened a new page in modern Georgian digital lexicography. These are dictionaries of some Georgian dialects, composed on the basis of wordlists by means of a special corpus editor, which processed corpus data from the Georgian Dialect Corpus (Beridze, Lortkipanidze, Nadaraia 2015).

Modern Georgian lexicography is marked by the development of new genres of lexicography: frequency dictionaries, learner’s dictionaries, Georgian Sign Language Dictionary, etc. Georgian lexicographers also pay great attention to the compilation of online terminological dictionaries based on ISO standards.

One more important direction of modern Georgian digital lexicography is the application of lexicographic data in different translational and educational applications. This new tendency demands new approaches to the processing and structuring of dictionary data that will increase the effectiveness of the application of lexicographic data not only for stand-alone dictionaries but also for Natural Language Processing purposes. This tendency created the need for new methods in lexicography which is becoming an independent, interdisciplinary branch of knowledge. This goal can be achieved by theoretical studies, as well as by educating a new generation of Georgian lexicographers. For this purpose, new MA and Ph.D. programs were developed at Ilia State University which aim to teach lexicography in close relation to information technologies, computer linguistics, general linguistics, and lexicographic practice.

The new trend in modern Georgian lexicography is dictionary didactics which aims to improve the dictionary culture at schools and universities of Georgia.

Hélène Gérardin (Paris)

Building a Dialectological Lexical Database of Georgian Cognates for Digital Analysis

The paper presents my work on the collection of dialectological data in Georgian that I carried out as part of my postdoc position in the LaDyCa project (Language Dynamics in the Caucasus) at the University of Sorbonne (Paris), with a focus on the application of Levenshtein distances.

Lia Karosanidze (Tbilisi)

The Georgian Term Bank (GTB) and Issues of the Standardization and Normalization of Terms in Georgia

The dysfunction of terminology management policy in Georgia has often made us face salient problems. Uncoordinated work on terminology has caused the emergence of a number of undesired processes in Georgia, specifically:

1. simultaneous and chaotic work
2. unreasonable expenditure of resources
3. suspension of the development of field terminologies
4. decrease in scientific studies on terminology
5. terminological diversity occurring in all kinds of documentations that question the application of bilingual corpora as sources for glossaries of terms.

Years-long analyses of these and other hardships and the pertaining conclusions have made us choose a difficult but, in our opinion, the only correct way, namely, to amalgamate glossaries of terms into a single database. The work on the Georgian terminology database (GTB) was initiated in 2014 by the Department of Bilingual Dictionaries and Scientific Terminology at the Arnold Chikobava Institute of Linguistics. Owing to the lack of terminology policy in Georgia, the GTB has been designed on principles that are different from those of European term banks. European term banks combine specialized glossaries of normalized terms while the GTB will become a scientific database for glossaries of normalized terms. The main goal and function of the database was formulated since the very beginning: to combine and study all Georgian glossaries of specialized terms compiled at the Institute of Linguistics and, based on them, to identify a normalized term from a multiplicity of synonyms while establishing its English equivalent.

Alongside glossaries of other institutions, all editions of the *Glossary of Georgian Technical Terms* were entered into the GTB, specifically the glossaries published in 1920, 1921, 1935, 1960, 1977, 1982, and an unpublished glossary of 1993. Besides, the GTB also includes minutes from sessions of the Terminological Commission and comments by various experts, frequently considered by us when working on a new version of a glossary of technical terms. Based on the data of the GTB, thematic explorations have been incepted, specifically inquiries into terms pertaining to the thematic groups of ironware, kinds and parts of machines, tanks and trains, building materials, construction, computation, medical technologies, etc., Experts from various industries have been involved in these activities.

Obviously enough, the GTB needs support on the part of the authorities: we should primarily encourage and recruit young researchers. The digital era has spotlighted a need for the enhancement of a research institute for language studies,

given that the present day requires the development of technologies and linguistic inquiry. Alongside old, traditional approaches, in order to enhance language studies, we have to establish up-to-date methodologies, to develop terminology scientifically founded on scholarly research involving theoretical and practical work on terminology. In the current digital world, alongside European languages, the full-fledged use of Georgian can only be provided for by means of a language policy that is founded on consistent international standards. The application of a term bank for restoring terminology policy and coordinated terminology work has been a common approach for European terminology standards.

Bastien Kindt (Louvain-la-Neuve)

The GREgORI Project: General Overview, Goals, and Concrete Achievements Relating to Armenian Language

The GRE*g*ORI Project aims to produce corpora and make them available online. These corpora bring together sources written in classical languages or in languages of the Christian East (Greek, Armenian, Georgian, Syriac, Arabic, etc.) and are analyzed (lemmatization, part-of-speech tagging, inflexional tagging, etc.).

The presentation will focus on some of the concrete achievements of the project: 1) indexes published in critical editions; 2) word-forms or lemmatized concordances; 3) on-line interfaces enabling researchers to consult texts.

Gabriel Képéklian (Louvain-la-Neuve)

Presentation of Desamb, a Tool for Disambiguating the Results Produced by Automatic Lemmatisations

The main digital resource useful for the automatic processing of a natural language is its dictionary of forms. The latter can be more or less rich, but never complete. It is also more or less reliable, that is to say that we have been able to verify and validate its content. It can be built up in three ways. The first, the classic one, consists in manually lemmatizing texts from which one draws the lists of the forms present to increase text after text a dictionary of forms. The second, which is algorithmic, is based on generating possible forms according to what the grammar allows. And the last is the combination of the first two methods.

Knowledge of the language makes it possible to define its morphosyntactic framework and its inflectional rules in algorithmic form. Technically, we need to

represent this knowledge formally and populate the data structures created. The generation of forms can be partially validated using the form dictionary resulting from the first method. As for the dictionary, augmented with the generated forms, it can be used to lemmatise texts, either programmatically or using AI tools such as neural networks. It should be noted here that the dictionary of forms enables the programmed lemmatisation of texts which are then used to train the network.

For the work carried out as part of the GREgORI project, we appreciate the concurrent use of several methods. This consolidates the confidence we can have in the results. To this end, we have developed a tool for disambiguating the results produced by dictionary and neural lemmatisations. We will present this tool.

Emmanuel van Elverdinghe (Louvain-la-Neuve)

Benefits of a Combined RNN and Dictionary-based Lemmatization: the Three Armenian Editions of the Apocalypse of John as a Test Case

The methodology of lemmatization in the GREgORI Project combines two procedures: a traditional dictionary-based approach, relying on manually verified lemmatized and POS-tagged forms from previous corpora, and an analysis by a recurrent neural network model, which produces predictions based on the same dataset. This paper illustrates the differences between these two approaches and demonstrates the complementarity of their outcomes, using as a test case the Armenian text of the Apocalypse of John, accessible under three recensions in three (more or less unsatisfactory) editions. Once lemmatized, these texts can be compared with each other and with the Greek originals more easily and accurately, paving the way for a study of their textual history and translation technique.

Ramaz Khalvashi, Khatuna Beridze (Batumi)

Batumi Linguoculturological Digital Archive: New Perspectives of the Documentation of the Adjara Region

The paper introduces the "Batumi Linguocultural Digital Archive," an initiative that was implemented through financial support from the Shota Rustaveli National Science Foundation during the period of 2016 to 2019. This project was specifically designed to document the linguistic and cultural heritage of the Adjara region, Georgia. The archival holdings encompass an array of digital resources, including audio and video materials, totaling 100 hours of content, formatted in MP3 and AVI.

Among these holdings, a subset of 30 hours consists of audio and video materials that have undergone transcription. Additionally, 10 hours of video materials have received comprehensive multimedia annotation, featuring transcriptions, glosses, and interlinearization in the form of an EAF-type resource. Furthermore, the archive includes 5 hours of video content with multimedia annotations that extend to English translations, also categorized as EAF-type resources.

The development of the Linguistic and Cultural Digital Archive marked a pivotal moment in advancing digital documentation and archiving practices in Georgia. This initiative not only introduced innovative research methodologies but also leveraged digital resources, thereby significantly enhancing the efficiency of endeavors focusing on the research in humanities in Georgia. The primary achievement of the "Batumi Linguocultural Digital Archive" project lies in the creation of a modern, standardized framework for the documentation and archiving of cultural and linguistic materials in Georgia, a milestone that is poised to foster the development of interdisciplinary studies within the country.

Furthermore, it is worth highlighting that the resources acquired and cataloged under the project's auspices have transcended national boundaries, serving as valuable assets with international applicability. These materials have been translated into English, rendering them accessible to a broader, non-Georgian speaking audience. This effort toward internationalization underscores the project's commitment to facilitating cross-cultural exchange and knowledge dissemination.

Maia Lomia, Ketevan Margiani (Tbilisi)

The Theoretical Grounds and Practical Value of Distinguishing the Marker of Evidentiality in the Verb Paradigms of Kartvelian Languages

Evidentiality is a universal category. It is found in every language although the linguistic means of its expression may vary even within related languages. The group of Kartvelian (South Caucasian) languages embraces Georgian, Megrelian, Laz and Svan. Out of these, Georgian is the literary language whereas Megrelian, Svan and Laz are non-literary. In some of these languages, evidentiality is a grammatical category, whereas in others it is a lexical-semantic feature. In cases where evidentiality is a grammatical category, the chief means of its expression is the verb. Evidential verb-forms are of two types: perfect and imperfect. Perfect verb forms are more widespread, whereas imperfect forms are rare and more peculiar of the non-literary languages. It should be noted that evidential verb forms express both non-modal and modal semantics.

The present paper focuses on the following issues: a) the theoretical grounds of non-modal evidential **perfect** and **imperfect** verb formation and segmentation, and

b) the practical value of distinguishing the markers of evidentiality for the creation of electronic resources of Megrelian-Laz and Svan texts.

Perfect evidential verb forms are a common Kartvelian phenomenon. The expression of evidentiality in this case is a secondary function on the synchronic plane. Perfect evidential verb forms do not have a special morphological marker. They use lexical means to express that something **has not been seen.** In this way, they oppose the aorist verb forms expressing the action that was **seen**. As we have mentioned, in this regard, all Kartvelian languages are similar.

Imperfect evidential verb forms appear only in the non-literary Kartvelian languages (Megrelian, Laz and Svan). In Laz, evidentiality is expressed lexically (in a descriptive way), whereas in Megrelian and Svan, evidentiality has a special morphological marker.

In Megrelian, the present and imperfect verb forms express a seen action. They are opposed to evidential imperfect verbs, which denote an unseen action. In Svan, the neutral imperfect is realized in two forms: the evidential Imperfect I in the superessive version and the evidential Imperfect II. Initially, the Megrelian Evidential Imperfect I and Evidential Imperfect II, and the Svan Evidential Imperfect II were descriptive. As a result of transformation, these forms turned into organic formations and developed special markers of evidentiality.The Svan Evidential Imperfect I in superessive version seems to have been organic from the very start. Thus, in Svan and Megrelian, evidentiality has developed as a morphological category.

The segmentation and glossing of evidential verb-forms in the non-literary Kartvelian languages and the distinction of the marker of evidentiality serve certain practical aims, namely, a) to reveal the rich capacity of the language with regard to the expression of a universal category, b) to describe the genesis of the category of evidentiality and prove its authenticity, and c), to serve the purpose of a morphological annotation of Megrelian and Svan texts.

Chahan Vidal-Gorène (Paris)

Automatic Information Retrieval from Medieval Handwritten Documents. Examples on Armenian and Eastern Scripts

The rapid advancements in computational methods and artificial intelligence have brought significant disruptions to traditional disciplines such as historical sciences and palaeography, particularly in the study of texts. Currently, digital humanities heavily rely on automatic text recognition and information analysis, which have become well-established processes in handling collections written in the Latin script.

However, despite the success of certain projects in dealing with non-Latin scripts like Syriac, Hebrew, and certain Arabic scripts, Eastern scripts or specific

corpora still suffer from significant under-resourcing and limited coverage by intelligent systems. Consequently, our presentation aims to present an up-to-date overview of the available technologies and models in text recognition and information retrieval, encompassing elements like date identification, name entity recognition, and morphological analysis.

Our focus will be on the advancements made by Calfa specifically for Eastern scripts, including Armenian, Georgian, and various Arabic scripts. By presenting the achievements and capabilities of the Calfa technology, on different use cases, our intention is to highlight the progress made in computational methods and artificial intelligence for Eastern scripts, and the good practices to overcome issues of a lack of data. We will delve into the potential applications of these systems in effectively exploring vast corpora, unlocking new possibilities for research and analysis in these scripts' traditions.

Irina Nevskaya (Frankfurt a/M)

Paradigm Change in Old Turkic Runic Studies

Old Turkic runiform inscriptions are the earliest Turkic sources providing valuable information on the history, culture and language of ancient Turkic peoples. The most famous and studied ones are the “classical” Orkhon inscriptions in Mongolia, followed by runic epitaphs in the reaches of the river Yenisei in the Republics of Tyva and Khakassia (Russian Federation). Written in an autochthonous script on rocks, steles and everyday objects, they present a great challenge for researchers trying to decipher them.

Old Turkic runic inscriptions in the Republic of Altai have become an object of special research only in the course of the recent decades, first by traditional, contact, methods of fixation. However, the contact methods of documentation have proved to be not sufficient because of the peculiarities of the Altai runiform signs: they are written with very thin cuts, 0.1–1.0 mm wide, less than 0.1 mm deep, often on the surfaces with earlier and later graffiti and petroglyphs. Thus, one of the main challenges in their deciphering is to discern the lines of the inscriptions from numerous further lines on the stone surface, both of natural and artificial origin.

In 2017–2020, a three-dimensional documentation of these inscriptions has been done using the method of digital photogrammetry. It has allowed us to correct earlier readings of well-known inscriptions and decipher those discovered recently. These readings showed that runiform inscription in the Altai Mountains are an important source of information on the language, traditional culture, religion and beliefs of the ancient Turkic speaking population of the Altai Mountains.

Maia Machavariani (Tbilisi)

Digital Humanities and Korneli Kekelidze Georgian National Center of Manuscripts

The research of written sources using modern technologies has created completely new perspectives for humanitarian fields. Digital Humanities has become a foundation for interdisciplinary studies. As a result of the joint work of specialists of codicology and textual studies, history and source studies, and art-historians on the one hand and of the professionals of the Digitization Laboratory on the other hand, it has become possible to study and present the research issues in a much more complete way.

The first Digital Humanities project at the Korneli Kekelidze Institute of Manuscripts was implemented in 1997-1999 when the first database of historical documents kept in the institute’s repositories was created. Since then, the National Center of Manuscripts has implemented a number of projects in the direction of Digital Humanities, such as:

1. digitization – making digital copies of the manuscript heritage (books, documents, archives) preserved in the Center
2. cataloguing – the creation of thematic, sectoral, chronological and other types of electronic catalogues of the stocked materials
3. various databases
4. various websites – on Georgian historical figures, Georgian, Persian and Ottoman painted documents, and others.

In the present report, we will also talk about future projects, the most important of which is the creation of digital descriptions of manuscripts.

David Maisuradze (Tbilisi)

Georgian Paleographic Fonts and Their Role in Scientific Studies

Paleographic fonts allow graphemes with paleographic features to be printed. They contain vector representations of graphemes. Modern fonts are computer programs that are equipped with artificial intelligence and can perform various functions.

Depending on what the graphematic shapes of the font are based on, paleographic fonts can be categorized as date-based, manuscript-based, calligrapher-based, or calligraphic school-based.

Paleographic fonts are multi-functional. On the one hand, they are used by the community for various purposes and they contribute to the popularization of science. However, the primary function of paleographic fonts is to aid in scientific studies.

Paleographic fonts can be viewed primarily as a digital alternative to printed paleographic albums and tables. Fonts have several important advantages over such printed publications:

1. The number of graphemes in a font is unlimited. Multiple alternate outlines of the same grapheme can be incorporated into one font.
2. The font has the ability to represent an unlimited number of ligatures and their alternatives, which are often ignored in printed publications.
3. The historical punctuation, often ignored in printed editions, which varies depending on the century and the calligrapher, is fully integrated into the fonts.
4. A font enables the dynamic representation of writing. It allows displaying not only individual symbols but also a text loaded with paleographic features.
5. The frequency of use of certain stylistic alternates and ligatures in the typed text can be determined programmatically. This will make the typed text more authentic in comparison with the original manuscript.
6. A font can also display spacing between words or lines specific to a particular manuscript, calligrapher, or era.
7. All graphic images in the font are vectors, which is why they can be infinitely enlarged without loss of quality.

A paleographic font creates visual aids for scholarly literature. It is common for a verbal description of a symbol to be unaccompanied by an illustration because there is no way to print the symbol. This deficiency is compensated by paleographic fonts.

Paleographic fonts facilitate the rapid study of paleography. The student has the opportunity to type the text himself using the dynamic font and observe how a word, a sentence is expressed in the handwriting of a specific era or calligrapher.

The following paleographic fonts have been created by me since 2010:

a) fonts based on the handwriting of individual Georgian figures of the 18th–20th centuries
b) *asomtavruli*, *nuskhuri* and *mkhedruli* fonts according to three manuscripts preserved in the National Archives of Georgia
c) *nuskhuri* fonts according to Georgian manuscripts from Sinai, 10th century
d) fonts according to a *nuskhuri* manuscript preserved in the National Center of Manuscripts, 11th–13th centuries
e) *mkhedruli* fonts according to royal acts of the 11th–18th centuries.

In the future, when a large database of Georgian paleographic fonts will have been created, it will be possible to compare manuscripts with the fonts in the database using artificial intelligence. By comparison, the artificial intelligence will then be able to more or less accurately date the manuscript, assume its origin, copyist and calligraphic school.

Mariam Kamarauli (Hamburg)

Digital Methods of Comparative Manuscript Research: Passion of St. Febronia

Over the last 20 years, considerable progress has been made in the analysis of palimpsests as the oldest written specimens of the three Caucasian literary languages, Georgian, Armenian and Caucasian Albanian, and the results have provided substantial new insights into their historical development. These insights, which have hitherto been confined to the individual languages, are now for the first time ever to be brought into a cross-language synthesis, which will yield a completely new view on the emergence and spread of writing in the region, taking into account the interrelations between the three languages and the Christian cultures represented by them as well as the influence of external religious and linguistic factors. The project “DeLiCaTe – The Development of Literacy in the Caucasian Territories” aims to assail the task of developing a first cross-language synthesis of the common conditions and circumstances of the development of literacy in the Caucasus. For the successful execution of this project, it is important to decipher and edit the palimpsests, which has proved to be a difficult task. Several methods for the decipherment of these palimpsests need to be developed and applied, such as working with multispectral images, counting of letters and lines, comparing different redactions, identifying the exact text passages and analyzing and explaining the differences given during the comparison. The present paper will deal with these topics on the basis of several examples that illustrate the methods of analysis applied and the progress made so far.

Sarah Dopierala (Frankfurt a/M), **Max Ionov** (Köln)

Towards an Automatic Approach to Extracting Abkhaz Language Examples from Written Text

Combing through sources and grammars looking for language examples for a particular language phenomenon is a process familiar to many linguists. This is especially common in descriptive or comparative research on languages without large or well-established corpora. One particular type of material that linguists may be looking for when working with these texts is (interlinerarly glossed) language examples. When working with machine readable PDFs (i.e. those not coming from scanned books), it might be possible to search the file for things like keywords or certain sequences of letters (i.e., certain morphemes in language examples) to bring

the linguist quickly to a particular example. However, even if this function is available, the linguist may have to sift through many matches, including words containing said sequence of letters, and potential examples containing the query item, and they will likely have to transcribe manually, in some manner, each example they want to use from the text into another format they will use in their research. In this talk, we present a workflow to help make this process more efficient and reliable by automating some of the manual labour that is inherent in this type of work. The proposed workflow consists of three parts: First, we identify and extract the examples from machinereadable PDFs. Second, we standardise the transliteration and the orthography as they vary across sources while preserving the original source form. Finally, we provide a user-friendly interface that allows these examples to be queried for specific morphemes, either by their form or their grammatical glosses. The search results can then be exported to established formats such as Toolbox and FLEx.

Mariam Rukhadze (Tbilisi)

Homonymy Caused by Grammaticalized Elements in a Corpus Linguistics View

Language is a system of a signs and represents a complex phenomenon in itself. The primary function of language – its use in communication – is carried out through the linguistic inventory of **lexical** units and **functional** elements. The inventory of the language is divided into different paradigmatic classes. Due to changes in natural languages, a sign can be transposed from one paradigmatic class to another. This type of change in a current language system is called **grammaticalization**.

Grammaticalization is a well-known phenomenon in typology. During the process of **grammaticalization** lexical units lose their historically developed semantics and after desemanticization acquire some specific grammatical function; consequently, we obtain a functional element. In typological terms, the transition of verb forms into functional elements is particularly interesting.

Functional elements create special difficulties in processing language by computational methods, particularly in computer linguistics. Today no one argues about the establishment of digital bases of languages and the importance of their processing with digital methods. The creation of automatic analyzers is one of the main tasks today. This requires a thorough analysis of the linguistic system on phonological, morphological, syntactic, semantic and pragmatic levels. The Georgian language material is not yet functionally properly processable, given that the Georgian National Corpus (GNC) does not have an algorithm to subtract functional elements from elements belonging to other paradigmatic classes, thus causing homonymy and leaving the problem of **disambiguation** unsettled.

The present paper refers to the process of **grammaticalization**, the functional and semantic analysis of **grammaticalized** items, and it demonstrates the problem of **homonymy** arising in computer linguistics due to **grammaticalized** elements of verbal origin. The research is based on both corpus-based and corpus-oriented methods and is carried out on the basis of the Georgian National corpus (GNC) (www.gnc.gov.ge). Empirical data is collected from the Old, Middle and Modern Georgian corpora, as well as the corpora of juridical and political texts.

The aim of the paper is to investigate the **functional and semantic** characteristics of **grammaticalized** items. For this purpose we use methods of **lexical substitution** and **elimination**; to describe the functions of the item we use a **language competency test**. The results of the test are illustrated by charts.

The goal of the paper is to describe the problem of **homonymy** caused by **grammaticalized** items in corpus linguistics and to find ways of removing the ambiguities. It suggests rules that are created by the analysis of the collected data.

Mariam Gobianidze (Tbilisi)

The Issue of Equivalence of Aphorisms in the English Translations of Shota Rustaveli's "The Knight in the Panther's Skin"

The 12th-century Georgian poet Shota Rustaveli is the author of an outstanding medieval literary monument: "The Knight in the Panther's Skin". The epic, which is a most vivid expression of Georgian spiritual culture, has reached us in the form of more than 160 manuscripts.

The topicality and importance of the epic is recognized far beyond the borders of Georgia. With translations into 56 languages, it has for long aroused the interest of foreign readers. In some languages, there are several versions made by different translators. The aim of the present paper is to analyze the aphorisms in the English translations of the epic.

"The Knight in the Panther's Skin" was translated into English by five different translators working at different times. The first English translation of the epic, a prose version, was published by Marjory Wardrop in 1912. A second English translation, in verse form, was made by Venera Urushadze in 1968. The author of the third translation was provided by Katherine Vivien in 1977, with the main text being in prose but the prologue and the epilogue in verse form. The fourth translation was made by Robert Stevenson in 1977, in the form of rhythmical prose. The most recent translation, a poetic version again, is by Lyn Coffin, based on a word-by-word translation by Dodona Kiziria; it appeared in 2015.

The present research analyzes the translations by Marjory Wardrop, Venera Urushadze and Lyn Coffin. The aim is to evaluate the quality of the translation of aphorisms from "The Knight in the Panther's Skin" in the poetic translations. For this purpose, it is based on the poetic translations and the literal translation by Dodona Kiziria. However, as Marjory Wardrop's translation was the earliest and has possibly served as a source for the later translations, this prosaic translation is also embraced.

Rustaveli's epic "The Knight in the Panther's Skin" abounds in aphorisms which are especially used to express wise ideas. The selected aphorisms are discussed in the present paper on the basis of up-to-date Rustvelological research by analyzing the opinions of Rustvelologists regarding the aphoristic speech in the English versions of "The Knight in the Panther's Skin". The evaluation of the translations is based upon research on the equivalence of the aphoristic formulae, on the example of two aphorisms (ch. 36, str. 878 and ch. 30, strophe 764. Based on an analysis of these two aphorisms and their translations, we attempt to assess the semantic precision and functional adequacy of their translations in the different versions, focusing on the meaning of the aphorisms in the original text of "The Knight in the Panther's Skin" and the semantics of their English correlates in each of the selected English translations. The aim is to find out whether the translators have managed to preserve the literary meaning of the original text. The research also dwells on the strategies selected by each of the translators in translating the aphorisms.

Giorgi Jgharkava (Tbilisi)

Digital Processing of Proverbs in Kartvelian Languages – Theoretical and Technological Framework

Language is a universal form of spiritual heritage, realized in texts of diverse genres and content. Among these, special attention should be paid to proverbs because they are marked with peculiar national features. Proverbs represent specific aspects of the life, culture and history of a given nation and are distinguished by the ability to preserve most ancient information. Therefore, research of proverbs is important for the study of the national spiritual culture and for the further development and promotion of the given field. Human wisdom and the system of perspectives concentrated in proverbs determine the core nature of a certain culture. Thus, the structural-semantic analysis of proverbs demarcates the general picture built upon the common beliefs, viewpoints and global perspectives of a given society.

The creation of a database of proverbs of Kartvelian (Georgian, Megrelian, Laz and Svan) languages and the implementation of linguocultural research using this database is especially interesting and topical because it proves the genetic unity of the Kartvelian languages and the existence of a common cultural-mental and linguistic space of the Georgian people. The Kartvelian languages, as well as the communities

speaking them, are closely related. Therefore, essential differences between the proverbs of Kartvelian languages are rare. Differences are due to the geographical location, local rites and customs, specifically toponyms and anthroponyms, as well as other local features. All this has been reflected in the proverbs. The comparison and harmonization of Kartvelian proverbs based on their semantic equivalence yields a significant resource for interdisciplinary research since proverbs are interesting not only linguistically but also from the viewpoint of folklore, ethnography, ethnopsychology, sociolinguistics, psycholinguistics, translation and cultural studies, etc. Proverbs represent important empirical material for general typological research as well because they are characterized by isomorphism and allomorphism, variance and invariance.

In the modern digital age, the acquisition and processing of relevant empirical material is strongly tied to the use of the latest technologies and methods, which is why interdisciplinary studies based on extensive empirical material are moving to a new stage. Considering the recent scientific literature created in the field of paremiology (Cignoni & Coffey 2000; Laporte 2012; Steyer 2014, 2017; Solano & Rondineli 2021 etc.), it is clear that the existence of digital resources of proverbs and their investigation by corpus-linguistic methods open up new research perspectives. Against the background of all this, the creation of an electronic database of Kartvelian proverbs is important not only for the development of thematic corpora but also for the implementation of corpus-based and/or corpus-driven studies. The present paper discusses the theoretical and technological framework for a digital processing of proverbs in the Kartvelian languages.

Jesse Wichers Schreur (Leiden)**, Max Ionov** (Köln)

Automatic Conversion of Interlinear Data to LaTeX – Caucasian Languages and Beyond

The applications Toolbox and its successor FLEx (Fieldworks Language Explorer) are pieces of software developed by SIL in order to facilitate corpus creation, lexicography and interlinear glossing of small and medium-sized datasets. They are widely used by linguists of many kinds (fieldworkers, typologists, historical linguists) and by students in many linguistics programmes. Interlinear glossed text can be exported in a number of ways. Toolbox provides its own so-called Standard Format Text structure, while FLEx provides multiple output options such as XML, Word and OpenOffice. However, judging from our collaborations, our courses where we teach FLEx, and many comments in the linguistics community, these output options do not meet the standard for most users. All options are, to varying degrees, judged to be lacking in (1) customisability and (2) aesthetics.

Our solution has been to create a web service and a standalone application that can be used for transforming Standard Format Texts and FLEx XML into LaTeX representation in a flexible way.

The decision to provide both a web service and a standalone application comes from limitations and preferences of potential users that might want to use the application in places with unstable internet connection or not want to upload their data before publication. Additionally, this solves a common problem with academic software which often stops working after the end of the project.

The final result is an application that gives the user the ability to (1) fully customise the placement of different lines of interlinear glossing from the source text, (2) make use of additional options to transform characters and scripts in original or additional lines, (3) use the powerful LaTeX environment to fully customise the font, typesetting and general aesthetics of the glossed examples. This talk aims to present the application and its present and future functions, but also to engage in a discussion to collaboratively answer questions regarding the demands of modern linguists of our software in general.

Felix Anker (Jena)

Digitizing the Caucasus – Where we are and Where to go

As digitization is picking up pace in all areas, it is worth taking a look at where we are right now and where to go. The digitization of research data and outcomes allows us to share knowledge without the hitherto analog boundaries. In this talk, I will present two of our institute’s past and ongoing projects and how they can be used by both scientific and non-scientific audiences:

(i) The electronic dictionary of Sanzhi Dargwa

The Sanzhi Dargwa dictionary has been published in *Dictionaria* which is an open-access journal publishing electronic dictionaries (Forker 2019). Sanzhi Dargwa belongs to the East Caucasian language family and the research has been conducted by Diana Forker since 2012. Together we compiled the dictionary which also contains example sentences and audio recordings. It is now publicly available to both researchers and non-researchers interested in the language. All the necessary tools for creating an electronic dictionary are available for free and I will provide a short overview on how to get started.

(ii) The comprehensive lexical database *LexCauc*

The goal of the *LexCauc* project (https://lexcauc.github.io) is to build a lexical database to investigate the linguistic diversity in the Caucasus with the help of quantitative statistical and phylogenetic tools. The database will contain transcriptions, translations, and audio recordings of more than 1000 concepts. The website will furthermore give information on the languages, cultures, and peoples of

the region and is therefore not only valuable for scientific research but also for the general public.

These projects are just two examples of how scientific data on the Caucasus can be digitized and processed in a way that is not only useful for science but also educational for a wider audience. They also paved the way for new projects that have already been launched and will contribute to the digitization of the Caucasus.

Tatia Tsetskhladze (Batumi), **Anastasia Kamarauli** (Wels)

Contemporary Georgian Political Speech from a Gender Perspective (Contrastive Analysis)

Political speech and, correspondingly, political discourse is a multidimensional phenomenon. Due to its nature, it represents an object of interdisciplinary research. In the beginning of the past century, the works of W. Lippmann (1922), G. Lasswell (1948), and P. Lazarsfeld (1944) formed grounds for the development of a new interdisciplinary field: **political linguistics**, which was further developed in Armin Burkhardt’s work on “Politolinguistik” published in 1996.

Both the lexical and structural means used in oral political speech are aimed at obtaining or preserving power. Political speech is a powerful tool affecting mass mentality. It is effectively used by politicians and people discussing political issues. Thus, research of political speech is important from the viewpoint of the analysis of language behavior and its impact on the society.

The scientific study of political communication started in Georgia in the 21st century. Currently, numerous works have been dedicated to political discourse in the form of monographs, scientific papers and doctoral theses. Due to the topicality of the issue, numerous blogs have appeared in the Georgian media, focusing on the analysis of political speech. Despite this, political linguistics is still in the process of development and requires large-scale theoretical research as well as the introduction of new methods and approaches.

Scientists are now focusing on the gender aspect of political speech, studying the dynamics of activity of female politicians and identifying the language differences between the speeches of male and female politicians. Gender-oriented research of oral Georgian political speech is topical because in the past two decades, the number of female members of parliament (MPs) has increased radically. The political activation of women is proved by a legislative act issued in 2021: the statutes of the Parliament of the 10th convocation underline that 1 out of every 4 candidates in the election lists of parties was to be a woman. The activation of female politicians has changed the political culture and affected gender stereotypes in the country. These changes have

aroused our interest towards the manipulation techniques and strategies in Georgian oral political speeches from a gender perspective.

The present paper introduces the analysis of oral political speeches of four Georgian politicians and shows similarities and differences in the speech of them.

In selecting our empirical resources, we have considered the principle of balancing:

1. **balance of the political spectrum**: for the analysis, we selected representatives of both ruling and opposition parties
2. **thematic balance**: for the analysis, we selected thematically similar political speeches, in particular, with the topic of Georgia's European integration;
3. **gender balance**: the number of female politicians' speeches is equal to the number of male politicians' speeches.

Our research of manipulation techniques and strategies in Georgian oral political speeches from a gender perspective will support interdisciplinary research into the Georgian political culture and increase public awareness of it.

Nino Sabadze (Tbilisi)

The Problem of the Face of the Dragon in Caucasian and European Mythic-Epic Texts

The face of the dragon occupies an important place in folklore, mythology and literature. This ancient mythological character has never lost its relevance, because over the centuries it has acquired various symbolic meanings and undergone many transformations. The dragon sometimes brings benefit to people, sometimes it causes problems and trouble.

The dragon, as a mythological archetype, appears in texts of various genres: cosmogonic myths, fairy tales, narratives, epics, legends and pseudo-historical tales related to the foundation of a village. It can be said that it is a comprehensive form of world folklore.

The study of the dragon's face is extremely interesting from the point of view of the comparative method of research. Until now, the dragon has not been considered as a mythological character in a comparative sense contrasting European and Caucasian legends. The present paper draws the image of the dragon according to the Anglo-Saxon Beowulf epic and several Caucasian legends, namely, Georgian, Ossetian, Abkhazian, Chechen, and Ingush materials. In order to determine the similarities and differences and the functional characteristics between them, a special place is devoted to the typological analysis of the material collected from the Georgian *Amiraniani*.

The paper discusses both universal and specific national characteristics of dragons, their types, action space, ways and means of defeat. The purpose of the paper is to clearly show the face of the dragon in epic and mythological works, its functions and role in the life of the protagonist.

The dragon is connected to many areas of the world and includes good and evil aspects. At the same time, the dragon can be considered as a challenge in a hero’s life, without which he could not prove his heroism. Every mythic-epic hero needs his dragon on the path of life to prove that he is really a hero; in this case the dragon is needed as a necessary power for the hero. Therefore, it can be safely said that the dragon is a challenge for the hero, rather than an outright personification of evil. Even when the dragon devours the hero, it is a way of rebirth and initiation for the hero, after which he returns to society with new knowledge and vision. In the bosom of the dragon, the hero shares sacred knowledge that he could not have received otherwise. Often, a relationship with a dragon is followed by a series of prohibitions in the hero's life, which he is unable to follow, and that is why he is punished.

It is interesting that the appearance of the dragon in Caucasian mythology is different from the dragon prevailing in European literature. In the Georgian language, dragons can be addressed by several synonymous words, and the usual word for dragon itself is a compound. In addition, the characteristics differ in how they are presented in Caucasian mythology and how they appear in European literature. In the Caucasian material, the dragon does not have the function of spewing fire and flying, whereas in the epics of the Caucasian peoples, as well as in the ancient epics, the dragon is rather connected to the elements of water and also appears as a treasure guard. In the epics of the peoples of the North Caucasus, the motif of the hero’s fight with the dragon can indeed be found, but unlike Georgian and European legends the episode of absorption-and-exit is not confirmed in them. All the existing differences are particularly interesting because this is where cross-cultural differences come into play.

Elene Kadagishvili (Tbilisi)

Basic Models in the Research of Sentiment Analysis in the Georgian language

Sentiment analysis, a branch of natural language processing, provides insights into the emotional tone of textual content, distinguishing between positive, negative, and neutral sentiments. This presentation focuses on sentiment analysis conducted on the Political Corpus of the Georgian language.

Our research involves defining and categorizing various models employed in sentiment analysis. Specifically, when neutral nouns are paired with positive or negative adjectives, the overall sentiment is altered accordingly.

- A (Pos) + N (Ntr) >>> Sentiment (Positive)
- A (Neg) + N (Ntr) >>> Sentiment (Negative)
- A (Ntr) + N (Ntr) >>> Sentiment (Neutral)

Conversely, if the noun itself conveys a positive or negative sentiment, the resultant sentiment follows suit. For instance, the sentence *Friends, this one year after the putsch has led us to a national* ***<neg>disaster</neg>*** demonstrates a negative sentiment due to the negative sentiment of the noun phrase:

- *A (Ntr) + N (Pos) >>> Sentiment (Positive)*
- *A (Ntr) + N (Neg) >>> Sentiment (Negative)*

Notably, noun sentiment plays a vital role not only when adjectives convey neutrality but also when adjectives possess their own sentiments. Several scenarios are explored:

- *A (Pos) + N (Pos) >>> Positive*
- *A (Pos) + N (Neg) >>> Negative*
- *A (Neg) + N (Neg) >>> Negative*
- *A (Neg) + N (Pos) >>> Negative*

By analyzing combinations and contextual cues, the present paper makes it evident that sentiment determination relies heavily on phrases rather than isolated words. Furthermore, it shows that verb phrases (VP) are usually more influential than noun phrases (NP) in sentiment determination. It also addresses sentiment intensifiers and enhancement mechanisms and associated challenges encountered in sentiment analysis, such as ambiguity and semantic compatibility issues.

Nina Dobrushina (Lyon)

Nakh-Daghestanian Languages: Digital Resources and Some Examples of Their Use

The Nakh-Daghestanian language family is the biggest in the Caucasus in terms of different languages and language branches. In the last decade, there have been several collaborative projects aimed at documenting and systematically studying these languages. As a result, a number of new resources on Daghestanian population, multilingualism, vocabulary, and grammar emerged. In this talk, I will present the resources that are already available online and those that are in the process of being developed.

The demographic database of Dagestanian villages is a result of digitalization of several sources: rural registers from 1886 and 1895 and national censuses of 1926 and 2010. Creation of this database made it possible to study the dynamics of the population of Daghestanian languages across 150 years with great accuracy. Since Daghestanian villages are ethnically and linguistically homogeneous, the population of villages provide very good estimates for the number of language speakers starting from the end of the 19th century.

High linguistic density in a small territory was the reason why, before the advent of Russian, many inhabitants of Daghestan spoke several languages. The results of a large-scale field study of Daghestanian multilingualism are presented at **multidagestan.com**. The user can see the dynamics of multilingualism in various villages in Daghestan by filtering the data by language, gender and year of birth of the speakers.

It is known that multilingualism is often the cause of language change – convergence of languages in the domains of vocabulary, phonetics and grammar. The **DagLoans** project emerged as an attempt to test the correlation between the number of borrowings from different languages and the intensity of multilingualism among the speakers of the same languages. A database of borrowed words collected in a number of villages in Dagestan is available online.

In contrast, the **DagSwadesh** project aims at collecting basic vocabulary (100 Swadesh lists) at the level of individual villages. This allows us to clarify the genealogical classification of not only languages but also dialects, and to raise questions related to the correlation of linguistic and geographical distance.

The systematic comparison of phonetic, lexical, and grammatical features of the languages of Daghestan is reflected in the form of **DagAtlas**, a typological atlas of the languages of Daghestan. The languages of Daghestan (including Turkic languages, Armenian and Georgian) are compared on the basis of a number of features which are displayed on maps in a uniform way.

The project of documentation of the dialectal variation of the Rutul language is conceived in a somewhat similar way, though with a much higher level of granularity and on a much smaller geographical scale. Based on a field survey of twelve villages, **a database of Rutul dialects** was created, and is in the process of being put online in the form of an atlas where each feature is reflected on a map. The project will make it possible to carry out a quantitative comparison of divergence between Rutul dialects.

Finally, I will talk about **the database of *wishes*** in Daghestan. The idea of this project is to run a cross-linguistic comparison of traditional expressions of blessings and curses in Daghestanian languages in order to reveal areal distributions in this domain. The data was collected mainly from the dictionaries of Nakh-Daghestanian languages, but also from the published collections of local folklore.

Paul Meurer (Bergen)

Towards a Treebank of Abkhaz – The AbNC, Analyzing Abkhaz, and the Importance of Good Tools

In this talk, I will present my effort to build an Abkhaz treebank in the Universal Dependencies (UD) framework, based on texts from the Abkhaz National Corpus (AbNC) and a morphological analyzer developed in the AbNC project.

UD is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across languages. Currently, there are UD treebanks of widely varying sizes available for 141 languages, but Abkhaz is not among them. This project is intended to fill that gap.

The AbNC

The AbNC was developed in the years 2016–2018 in a project financed by USAID, with participants from Sukhumi, Tbilisi, Frankfurt and Bergen (Meurer 2018). It comprises more than 10 million tokens of texts from a variety of genres and is morphologically annotated. The corpus is hosted in the Corpuscle corpus management tool, which has advanced possibilities for searching and viewing the corpus texts (https://clarino.uib.no/abnc).

In addition to being a corpus-linguistic resource and tool, the corpus also serves as a digital library and as a pedagogical tool for language learning. The texts can be read in a neatly typeset, paginated presentation, and pages can be bookmarked for later reference. Most importantly, when the user clicks on a word in the text, grammatical information about that word is displayed, and in addition, the corresponding articles in the integrated Abkhaz-Russian Dictionary (Kaslandzia, 2005) are shown.

Morphosyntactic analysis of Abkhaz

The lexicon lookup module of the analyzer is implemented as a finite state transducer (Meurer 2011), and Constraint Grammar (CG) rules are used for disambiguation. The main challenge in building the analyzer was coping with the extraordinarily high degree of homonymy of many word forms, owing to the polysynthetic nature of the Abkhaz verb, and the absence of stress in written text. CG rules, which take syntactic context into account, can be used to a certain degree to disambiguate a given form, but often, semantic information is needed to fully disambiguate a word. Semantic information can be (and is being) partially integrated into the CG parser, but a principled approach would have to be based on a word net, or on statistical information (e.g., word vectors) derived from a gold standard corpus.

Building a treebank

Dependency treebanks can be built in a variety of ways: they can be manually constructed, they can be built using either a statistical or a rule-based parser, or by a combination of those methods. In my project, I am using a rule-based parser, followed by manual correction of the output. The dependency rules are written in the CG formalism and constitute an add-on to the disambiguation module. The rules rely heavily on the morphosyntactic features from the lexicon lookup. In order to obtain best results, it is therefore crucial that the input to the syntax rules be fully (and correctly) disambiguated. After parsing, the resulting dependency analysis must often be manually corrected, or rules must be refined or added in order to handle difficult cases. I have implemented several tools that help streamline this workflow: The analysis of each word (either wrongly analyzed, or ambiguous) can easily be changed, and the resulting trees can be reorganized in a graphical web tool. In my talk, I will also provide a demo of those tools.

Karina Vamling (Malmö)

An Exploration of the Urban Linguistic Landscape of Batumi: The Case of Luka Asatiani Street

The linguistic landscape is formed by the multitude of linguistic signs present in the public space, ranging from official and commercial signs to private notes, signs and graffiti. Thus, it reflects both official and professional as well as private (written) language use and language choices in the community.

Georgian has a long history of competition with Russian in different spheres of society. In the last post-Soviet decades, the use of Russian in Georgia decreased. However, with the tourist boom in recent years and the high increase in the number of speakers of Russian in Georgia following February 2022, it is to be expected that these developments would have an impact on the Georgian linguistic landscape. The present study sets out to explore this multilinguality, and aims at detecting tendencies in the domains of language use in the public space.

The Batumi street chosen for this case study is Luka Asatiani street, named after the city’s first mayor. It is one of the older streets of Batumi. It runs 2,2 kilometres, starting in the most prestigious part of the city by the seaside boulevard, passing the 19th century City Hall, crossing several major streets, and ends in a predominantly residential area closer to the hills.

In this field study, conducted in the spring of 2023, all signs along the entire Luka Asatiani street were photographed, capturing over 400 textual units. A database of the digital photos and textual units was set up in Excel, allowing for a study of relative frequency of a number of categories and combinations of these, such as

linguality (mono-, bi-, trilingual), language choice (Georgian, English, Russian, Turkish or other languages), scripts (Georgian, Latin, Cyrillic), domains of language use, physical appearance of signs and other categories, being further outlined and discussed in the paper.

The general picture emerging from this study is that the Georgian language dominates in monolingual textual units, followed by a substantial use of English. Monolingual texts in Russian are encountered more rarely. Bilingual texts are to a large extent written in Georgian and English, thus adhering to the official language policy prescribing the use of English in parallel to Georgian.

Zhaolin Li, **Jan Niehues** (Karlsruhe), **Monika Rind-Pawlowski** (Frankfurt a/M)

Automatic Speech Recognition for Khinalug

Khinalug is a Nakh-Dagestanian (aka East Caucasian) language spoken endemically by appr. 2,300 people in only one village, Khinalug, in Northern Azerbaijan, and, with a decreasing level of fluency, by a diaspora of at least 10,000 community members in Azerbaijan and Russia. The position of Khinalug within the Nakh-Dagestanian family is debated.

As of 2011, an annotated corpus of natural speech has been developed and consistently extended and revised within the frame of the projects “Documentation of Khinalug” (2011–2016, Volkswagen Foundation), “Linked Open Dictionaries” (2016–2020, BMBF), and “Computational Language Documentation by 2025” (CLD2025, as of 2020, DFG/ANR), now comprising around 300,000 tokens of both transcribed recordings and literature produced by the Khinalug community.

The project CLD2025 aims at the development of Automatic Speech Recognition (ASR) methodologies particularly for languages with small corpora, including Khinalug. ASR refers to the process of converting spoken language into written text. So far, the ASR models developed for low-resource (and mostly endangered) languages tend to perform inferior compared to models for high-resource languages (such as English, Chinese, German etc.). The reason is that developing ASR models requires the training data of speech and corresponding transcription, which is less available for low-resource languages

To address the problem, our project explores the effectiveness of cross-lingual representation learning in ASR for low-resource languages. Cross-lingual representation learning enhances the overall speech recognition capability by leveraging knowledge from multiple languages. The motivation is that many languages share underlying acoustic and linguistic patterns, even though they may have different vocabularies and phonetic variations. Previous research shows that by learning from high-resource languages, the ASR model can identify commonalities

and transfer knowledge across languages, so that a lower amount of training data from target low-resource languages is required.

Training a successful ASR model for multiple languages is expensive and not always practical because of the requirements on computational resources and datasets availability. Fortunately, there are public models that are pre-trained with multilingual datasets. To leverage the cross-lingual representation learning, we build the mode based on the pre-trained ones and then fine-tune the model with the annotated data of Khinalug. Specifically, we utilize the public wav2vec2.0 models that are pre-trained with different numbers of languages ranging from 1 (English) to 1406, corresponding to the datasets of up to 55,000 hours.

With pre-trained models, we investigate the relationship between speech recognition performance and the number of pre-training languages. Results show that cross-lingual representation learning benefits performance, but increasing the number of pre-trained languages from 53 to 1404 brings no apparent improvement. In addition, as the prediction of wav2vec2.0 is on the character level, we explore the effectiveness of involving contextual information on the word level. We develop a word-level 5-gram language model that considers the contextual information from previous 5 words for the next word, and incorporate the language model with the ASR model. Results show adding the language model brings clear improvement.

Khatuna Beridze (Batumi)

Postcolonial Translation of Georgian Texts

Since the 1990s, research in literary translation has expanded beyond linguistic theories and delved into the realm of culture. The focus of study shifted towards viewing translation as a metaphorical representation of political events. Philosophical perspectives on power and ideology, such as deconstructionism and cultural studies, became influential in understanding translation and its activities. The metaphorical nature of translation, resembling imperial politics, led to its association with concepts like language policy, territorial expansion, and cultural transformation. As a result, translation became a symbol of power and academic research sought to explain the sociological and ideological influences of translation within the realms of literature and politics. In each narrative, the notion of “language politics” serves as a concept, reflecting the subjective choices of plot and characters made by the storyteller. The author presents a complete picture of the narrative, incorporating various elements like genre, syntax, and lexical choices, which are motivated by their own ideology. This language policy is a subjective expression of the author’s attitude towards the story and its characters. Furthermore, linguistic information is supplemented by extralinguistic elements. During the translation process, the clash between the

author’s ideology and language policy and the translator’s ideology and language policy can either result in harmony or in dissonance. The metaphorical nature of translation, influenced by deconstructionist ideas, disrupts the homogeneous composition of the source text. Through the subjective interpretation of textual signs, translation introduces a duality of meanings and purposes. Linguistic and stylistic considerations alone no longer suffice, as common and distinct artistic-political conditions between the author and translator narrow down the possibilities. Within the postcolonial paradigm, translators attempt to critique the discord between the author’s voice and the translator’s voice at the intersection of culture, politics, and language. The present study characterizes translation as a political metaphor, exploring the ideological and political goals embedded in the textual structures of the Source Language (SL) and the Target Language (TL).

Manana Tandashvili (Frankfurt a/M), **Mariam Kamarauli** (Hamburg)

Translation of the Aphoristic Style of “The Knight in the Panther’s Skin”

Shota Rustaveli’s epic “The Knight in the Panther’s Skin” is a masterpiece of Georgian literature. It plays a special role not only in the cultural memory of the Georgian people but also in the history of world literature. The epic has been translated into 56 languages and is included in the registry of the world cultural heritage of UNESCO.

Shota Rustaveli’s epic “The Knight in the Panther’s Skin“ is at the same time an exceptional monument of aphoristic style. Alongside with other stylistic devices (paroemia, maxims), the author makes abundant use of aphorisms in order to express his wisdom. Some of the aphorisms in the epic paraphrase quotations from the Bible and from the works of philosophers. Along with their origin, Rustaveli’s aphorisms can be distinguished according to their thematics (love, friendship, destiny, etc.) as well as their purpose (didactic, evaluative, verifying, and so on): love-related aphorisms usually derive from the biblical books of the Apostles as in the case of “A wise man cannot abandon his beloved friend“ (cf. Rom. 13:10 and I Cor. 13:7: “Love does no harm to a neighbour”). In the epic, the concept of love is elevated to the concept of divine love (“A friend should spare himself no trouble for his friend’s sake; he should give heart for heart, love as a road and a bridge“), thus agreeing with the concept of love in the works of the Apostles and holy fathers (cf. I John 4:21: “Anyone who loves God must also love their brother and sister”). Here we may also compare the didactic aphorisms that are based on parables from the Gospels: “A pearl falls to the lot of none without buying and bargaining” (cf. Mt. 13:45–46: “The kingdom of heaven is like a merchant looking for fine pearls”).

The origin of the aphorisms is sometimes complex because several parallel sources can be envisaged. For instance, the concept of **fate** as a term implying **God's will** or **destiny** is found in both Christian theological literature and ancient Greek philosophy. The concept of **fate/destiny** as an idea of omnipotence is also peculiar of the Georgian linguocultural space as it is found in folklore, especially in proverbs. Therefore the analysis of the origin of Rustaveli's aphorisms concerning fate (e.g., "The deed which is inevitably decreed above cannot be avoided"; "No creature of flesh hath power to thwart Providence") requires a multifaceted approach.

Although numerous works have been dedicated to the aphorisms contained in "The Knight in the Panther's Skin", opinions of scholars still vary regarding their origin, their content, their structure and even their number in the epic. Currently, there are about 30 monolingual or bilingual editions of Rustaveli's aphorisms, in which the number of examples varies from 52 (1943) to 234 (1991). The main reason for this diversity is that there are no strict criteria for distinguishing between aphorisms and similar expressions (didactic phrases, parables, proverbs etc.).

In our paper we will present a **theoretical framework** created for the classification and qualification of aphorisms in general (origin, speaker, theme, structure, semantics) which will be used to create a **technological basis** for the search of aphorisms. The multilingual parallel corpus of translations of "The Knight in the Panther's Skin" created at the Institute of Empirical Linguistics of the University of Frankfurt will also be presented.

Michael Job (Göttingen / Baden-Baden)

Caucasian Studies amidst the Digital Revolution: Observations by a Born Analogist

When I started to study in 1966, my "word-processing" was characterised by handwriting and a portable typewriter for the term papers. Compared to this quite restricted starting position, we are, these days, in a situation we would not have been able to dream of in the sixties.

What I have witnessed in the past fifty years, can be divided into different stages with regard to digitisation:

- **(i)** the slow departure from the analogous world in the everyday practice of the humanities; and parallel to this,
- **(ii)** the development of technical progress in word-processing;
- **(iii)** the use of mainframe computers in universities also for dealing with text-related projects;
- **(iv)** the emergence of PCs with the possibility for institutes and individuals to work on even demanding questions with affordable hardware;

(v) the triumph of the Internet with an unimagined networking of researchers worldwide;

(vi) at the same time, the availability of huge amounts of digitised data, including Jost's TITUS project, which has made large amounts of texts and, in part text analyses, available for students and researchers of Indo-European and Caucasian languages. In other parts of the world, masses of texts have been digitised in the same way for other language areas, e.g., the Classical languages;

(vii) digitisation of books and journals beyond what was previously imaginable. Copyright has (of course) not always been respected in the process;

(viii) the most recent phase of this development is determined by the application of artificial intelligence, which on the one hand provides powerful translation programmes (such as *DeepL*), but above all: a couple of months ago, text generation tools, such as *34chatGPT* and similar programmes have been launched, the effects of which on the academic world cannot yet be assessed.

My paper will address selected items of the development outlined above with a focus on what many of us – I mean: of the older generation – have experienced in the past decades and what may be expected in the foreseeable future.

George Hewitt (SOAS, retired)

Studying Kartvelian/Caucasian Languages in the Soviet Period and Today

The general conditions existing in the USSR during the 1970s are summed up in Rosemary Sullivan's 2015-book *Stalin's Daughter: The Extraordinary and Tumultuous Life of Svetlana Alliluyeva* (NY: HarperCollins), namely: 'In the mid-1970s, virtually no citizens of the Soviet Union were permitted to travel, unsupervised, outside the Eastern bloc, and any and all contacts with foreigners inside the USSR were still considered treasonous' (pp. 421-22).

Against this background I describe the experiences I 'enjoyed' or 'endured' as a post-graduate student in Tbilisi during the academic year 1975–76, and parallels were similarly experienced by my two American predecessors Dee Ann Holisky and Alice Harris who were in Tbilisi the year before my own arrival. I was able to spend a further academic year there in 1979–80 still as a post-graduate, whilst my sojourn during the last 5 or so months of 1987 was at the level of university-lecturer there to spend a term's sabbatical leave (from Hull University). And so, I had the twin-perspectives by virtue of living/studying in Tbilisi as both a (post-graduate) student and at the more senior level of lecturer.

Thus, I propose to describe for the Soviet period procedures (with attendant difficulties) under the following sub-headings:

1. Initial access allowing one to spend an extended period in Georgia during the late Soviet period (specifically the 1970/80s) for under-grad/post-grad students and lecturers
2. Making arrangements for one’s study and for receiving tuition on this or that language
3. Working in circumstances where essential/desired books, journals and/or recordings might not be easy to find, let alone acquire
4. Arranging travel within the country, whether for research-purposes or just simple pleasure
5. How to ‘play’ the system by utilising knowledge acquired on the ground!

I shall then close with some words on the post-Soviet period, including the odd account of my ‘brushes’ with aspects of now-existing technologies.