DOCUMENTATION OF ENDANGERED LANGUAGES AS CONTEMPORARY PRACTICE OF PRESERVING INTANGIBLE CULTURAL HERITAGE IN EURASIA

Andrey Y. Filchenko, PhD
Chair, Department of Languages, Linguistics and Literatures
School of Sciences and Humanities
Nazarbayev University
Nur-Sultan, Kazakhstan
andrey.filchenko@nu.edu.kz

ABSTRACT

Based on contemporary estimates, between 50% and 80% of the world's linguistic diversity will disappear during the 21st century. This means that out of estimated 6,500 languages currently spoken in the world, over 3,500 will seize to be used.

While every language manifests a unique wealth of human knowledge accumulated over millennia, sadly multiple languages in Central and Northern Eurasia are at risk. Often, these languages and language varieties are little known outside their immediate communities and are less studied by the academic community, which exacerbates the problem even further.

Contemporary methodologically and technologically advanced language documentation serves to establish the connection between the intangible and tangible. Modern electronic libraries and archives that result from such documentation projects, including those developed at NU SSH department of LLL, aim to address the problem of preserving linguistic and cultural diversity of Eurasia. The information and data that these projects produce offer an important empirical contribution to debates pertaining to the history, evolution, variation, and change in the languages and cultures of the region. The experiences collected from these projects are also useful in providing evidence to the discussion of the role of digital technologies in minority language/culture maintenance and revival, as well as in developing best practices for mitigating language endangerment.

While these projects naturally focus on collecting primary language and culture data in relevant communities, their important component has also been digital archiving of legacy materials associated with a range of technical, methodological, and ethical issues and their impacts on language/culture endangerment and vitality.

Based on various estimates, between 50% and 80% of the world's linguistic diversity will disappear during the 21st century. This means that that out of estimated 6,500 languages currently spoken, over 3,500 will cease to be used in various parts of the world. Most of these languages have no standardized or even any at all written form, and as such exist and are transferred across generations in their most natural form – as oral spoken idioms.

Sadly, multiple languages in Central and Northern Eurasia are at risk of disappearing, in some cases, without any significant trace. More often than not, these are languages and

language varieties that are understudied and are not well described, which exacerbates the problem even more. Every language is a unique wealth of human knowledge accumulated over millennia, offering insights into universal and unique features as well as possible limits to human cognition and communication.

Therefore, it is extremely important to recognize that documentation and study of linguistic diversity and change of the region should be among the essential components of any cultural heritage preservation program. Methodologically and technologically rigorous modern documentation of linguistic diversity is essentially an exercise in transforming the intangible – the illusive and the ever changing language and culture into tangible – the robust and accessible linguistic databases and repositories with wide applied potential.

Zooming into the Central Asian region and Kazakhstan in particular, one can see this area as an exciting arena of multidirectional interaction, variation, and change of various languages (genetically affiliated and not), including most Eurasian language families: Turkic, Mongolic, Uralic, Paleoasiatic, Indo-European, Sino-Tibetan, and Semitic. However, in Kazakhstan, it appears that most of the existing local research traditions focus primarily either on written Turkic / Altaic artefacts of Central Asia and South Siberia, or on the established literary tradition and folklore, utilizing often conservative theoretical frameworks and research methodologies. Sociolinguistics, though one of the dynamically emerging research domains in Kazakhstan, still sees most projects being of the survey nature, focusing on identification and description of multilingualism situations and aspects of language policies towards previously identified speech communities.

Very few recent studies have been addressing the language situation in Kazakhstan, which has been identified as exoglossic and unbalanced, with diverse genetically affiliated and unaffiliated languages of various structural types co-existing in the situation of extended contact. It should be noted, however, that in the survey and census data, often used in these studies, linguistic affiliation is often confused by the respondents with ethnic ancestry, while actual level of language proficiency and functional spheres of language use remain by and large unclear and unverified, particularly for the languages which are not characterized by high-density localized diaspora situations (as in the case of Uighur and Korean in Kazakhstan, for example).

At NU, there have been recent and ongoing projects focusing on documentation and analysis of the languages of Central and Northern Asia, including the endangered languages of indigenous ethnic minority groups. One project focused on comprehensive documentation of three endangered Turkic minority languages of Southern Siberia: Teleut, Eushta-Chat and Melets Chulym. As a result of this five year project, a large multimedia digital database has been created, containing audio-visual recordings and rich metadata, organized into accessible annotated multimedia corpora and digital lexica of these languages, maintained and curated under the auspices of the international Endangered Languages Archive at SOAS in London.

Another recent international collaborative project undertakes to develop the Multimedia corpus of modern spoken Kazakh language. This project, launched in January 2021, aims to fill the empirical gap in modern research infrastructure for the study of Kazakh language

in its contemporary state by implementing the state-of-the-art richly annotated multimedia corpus of natural Kazakh speech in its social and regional variation. Designed with an intense educational component in mind, this project explores aspects of Kazakh language diversity, contact, variation, and change in the wider regional context of Central Asia, uncovering linguistic details of the present and past conditions of cultural composition of the region. Apart from its considerable applied potential, this project also provides crucial missing components to the existing resources on Kazakh language (such as existing corpora projects focusing mainly on standardized written language variety), building a complex data system for further use in research, education, policy-making, and industry.

The research programs utilizing these data repositories and methodology of their application focus on such large research areas as:

- studying local linguistic diversity, variation, and change in wider areal context;
- aspects of multilingualism in natural speech behavior (code-switching, code-mixing, language shifts);
- natural linguistic diversity and variation vis-à-vis language pedagogy;
- historical, typological, social and cognitive aspects of multilingualism;
- traditional and contemporary cultural productions in multilingual contexts, and
- digital humanities and cultural heritage preservation in the region.

Contemporary cross-disciplinary research frameworks (for example the LAG research ideology that productively combines in a "triangulation-style" analysis of linguistic, cultural, and genetic data), allows for a more rigorous cross-disciplinary verification of disciplinary-specific theories increasingly. It also provides for more objective, discipline-external interpretations of data by a range of new synergetic methods, building universal models of local ancient and recent human histories. Integration of data from various disciplines (linguistics, anthropology, and archeology, and ancient and modern DNA analysis), has already proven to be an extremely productive approach to making inferences about processes underlying the distribution of modern or ancient genetic, cultural, and linguistic variation. Examples of such cross-disciplinary LAG-inspired programs include the "GENES, LANGUAGE, CULTURE" research program by the Max Planck Institute for the Science of Human History in Jena.

Cross-disciplinary approaches to the analysis of human history have been shaping throughout the 20th century, if not before. In linguistics, for example, early recognition of the utility of linguistics <-> archeology interface dates back to early proto-home reconstructions for the Indo-European, Uralic, and Altaic language families, in the 1960-1980s.

The late 1990s and early 2000s saw first applications of mathematical modeling and statistical methods adapted from population genetics to the analysis of linguistic similarity and variation (Archibald et al., 2003; McMahon & McMahon, 2005). Bayesian phylogenetic methods originally used to study the evolutionary relationships and divergence times of biological species have been applied to comparisons of basic vocabulary within established language families to quantitatively estimate the heterogeneity within the families and the robustness of their subgroupings. Such studies address in a more rigorous empirical

manner the long-standing issues plaguing historical-comparative linguistics, such as the effects of contact interaction versus genetic inheritance.

In other studies, quantitative phylogenetic methods applied to linguistic data help elucidate the evolutionary history of linguistic communities. Following the biological models, the macro-evolutionary processes in language families were seen as driven by either cultural interactions or environmental changes that affect linguistic diversification. These studies allow to verify or challenge traditionally produced chronologies for various groups and subgroups within linguistic families and suggest co-occurrence with environmental changes and cultural interactions of populations (Honkola et al., 2013; Trudgill, 2011).

As noted by the participants of the **GENES, LANGUAGE, CULTURE program at MPIRS SHH** in analyzing genetic data alongside cultural and linguistic data, a set of broad research questions is addressed:

- How or where do genetic histories match those from archaeology and linguistics?
- What kind of cultural processes create / maintain genetic diversity?
- How can genetic inferences be best integrated with those from historical sciences?

(https://www.shh.mpg.de/DLCE-research-overview)

Again, as stated by the MPIRS SHH programs, the s*cientific relevance of the new interdisciplinary (LAG) research ideology is in that it allows to:*

- Provide reference tools for geneticists, linguists, and archeologists
- Extract information on genealogical relatedness and demography for non-geneticists
- Frame questions on human history and diversity in a multidisciplinary perspective
- Develop more realistic understanding of the complex mechanisms behind cultural transmission
- Understand that change of cultural features through time not only impacts our ability of tracing back human prehistory, but also influences the definition of "population" as the unit of research.

Max Planck Institute for the Sciences of Human History: https://www.shh.mpg.de/553680/gelato-genes-and-languages-together

The models that NU future projects could pursue in the development of cross-disciplinary approaches to the study of Human History of Central Eurasia may follow the best practices exemplified by some of programs at MPIRS SHH, for example, and include:

- projects in documentation and analysis of the region's linguistic and cultural diversity, and developing methods for making inferences about human prehistory, relationships between languages and processes of language change (example: Glottobank (MPIRS SHH)).
- projects developing or contributing to the databases for exploring how languages within family (or families) relate to each other in their lexicon serving both qualitative and quantitative research purposes (example: Cognacy in Basic Lexicon (MPIRS SHH)).
- projects on interactions between human genetic and cultural evolutionary systems, integrating methods and data from linguistics-anthropology-archaeology-population genetics and quantitative history in a comprehensive approach to the study of human evolutionary history, (example: Gene-culture coevolution (MPIRS SHH)).

- events fostering understanding and interaction between specialists in archaeology, genetics and linguistics converging on complementary and holistic studies of human prehistory (Cross-Disciplinary Prehistory of the region (MPIRS SHH)).
- studies of local regional variety of cultural traditions, factors affecting the evolution of diversity, understanding the historical and contemporary processes generating diversity and convergence in the languages and cultures of local communities of the region, understanding how features of religion co-evolved with the structure of human social systems and the physical environments (example: Regional Languages & Lifeways (MPIRS SHH)).
- studies bringing together cultural, linguistic, environmental and geographic information for the regional societies, linking societies to their geographic locations and their shared linguistic ancestry, employing computational methods to investigate the roles of environment, spatial proximity and cultural ancestry in observable patterns of cross-cultural diversity across the region (example: Places, Languages, Culture and Environment (MPIRS SHH)).

The cornerstone of all such research programs is modern quality empirical data, increasingly in the form of a digital database.

In this area of digital humanities, exemplified by cultural heritage preservation programs, libraries play an important role, for example, that of providing long-term hosting and sustained maintenance of linguistic databases and repositories.

Similar to re-thinking the role of modern museums, modern libraries re-invent their place in contemporary and future research and education as institutions that go beyond passively archiving artefacts of cultural productions. They evolve into research and education facilities actively curating and managing diverse research and education resources, including linguistic digital databases and repositories.

In the context of linguistic databases, examples of the past are ample that demonstrate that beyond creation of a corpus project, in all its complexity, a possibly even more challenging and resource-intensive task is in its sustained long-term maintenance and further development, adding further research and education value to it, by keeping it updated, accessible, supported by expert consultations, in other words, a live resource with a wide applied value.

In this light, university libraries are uniquely positioned to not only ensure sustained maintenance of valuable linguistic/cultural resources in digital databases and repositories, but also to guide users in state-of-the-art, contemporary theory-informed applications of these unique empirical resources to diverse projects in research and education, policy making, preservation, and revival of linguistic and cultural diversity in the region.