B.12: Building diachronic corpora for minority languages – facing methodological challenges

Organizers: Ksenia Shagal, Daria Zhornik, Trond Trosterud, Timofey Arkhangelskiy, Alexandre Arkhipov, Jeremy Bradley

Best practices in corpus building conceived and designed for major European languages are of limited applicability when dealing with minority languages lacking a comparable wealth of resources, history of codification, and structural support. The goals of corpus design and development, however, are comparable: irrespective of the size of a (living) language, by default corpora should give speakers and/or scholars the means to view the realities of a language (be it to answer academic questions or just to figure out "how people say this" for miscellaneous reasons), and in the case of diachronic corpora, the history of a language.

Specific challenges encountered in Uralic corpus-building initiatives, especially pertaining to minority languages of Russia, in striving towards this goal include the following:

- The number of speakers imposes a bottleneck on language documentation efforts: 300 speakers cannot produce 300 billion tokens of text; consequently the pressure on the very few speakers, and the value of every last written record, becomes immense for languages with few speakers.
- Documentation of Uralic languages enjoyed a golden age in the late 19th and early 20th century, but then became impossible for foreign scholars for many decades after the revolutions of 1917. This can create the illusion of a "missing middle" in the external view on these languages, and the mirage of a dichotomy between the "old" language of the dialect text collections and the modern heavily Russified languages. Language documentation did however continue in the Soviet period – by scholars from the Soviet Union, both from outside and also within the speaker communities. The materials compiled in this era have oftentimes not been published and remain(ed) in archives in research centers such as Tomsk as well as in personal archives.
- Most Uralic languages were documented first and codified as literary languages later, i.e the aforementioned documentation precedes the codification of literary norms. Field research materials were compiled using diverse and idiosyncratic conventions (esp. in transcription). The inclusion of these materials in diachronic corpora is imperative, but comes with practical, administrative, and methodological challenges – especially when ensuring comparability with literary materials.
- Political upheaval in Russia has often gone hand in hand with orthographic reforms; for many languages of Russia new writing systems were created twice in the first half of the 20th century alone (first in the early Soviet period and again under Stalinism). Consequently even for literary texts, comparability of materials from different eras requires investment of effort.

- The highly diverse nature of writing systems (both orthographies and transcription) used for these languages has in the past made digitization of materials difficult; only recent technological advances in OCR technology have made the semi-automatic digitization of written and printed documents in obscure writing systems time-efficient. Recent additions to the Unicode standard now make it possible to represent almost all Uralic corpora comprehensively.
- There is often a modal mixture of available materials, i.e. written materials (both literary and records of field research) and audio (sometimes even audiovisual) materials pertaining to the same language; they should optimally be accessible to users through connected infrastructures.
- The inclusion of metadata is not trivial, as is the rendering of researchers' comments and additional information in the corpora. The conflict of interest between transparency and researchers' desire for a maximum amount of information (most notably personal background of the speaker) and modern conventions and laws pertaining to data protection can complicate these questions.
- Even (possibly especially) small speaker communities in Russia have very strong standard language cultures, making it necessary to stress the difference between descriptive and prescriptive resources when collecting and processing data and when building infrastructures.
- The accessibility of text collections/corpora to the speaker communities themselves, though historically neglected, must be taken into consideration in modern corpus building initiatives. Consideration in interface design (pertaining to the interface, metalanguage, etc.) is needed to make corpora accessible also to speaker communities.
- Special efforts are needed to make already developed and future corpora comparable and interoperable with respect to annotation conventions, search modalities and the presentation of the search results. While it does not seem realistic to impose common annotation schemes, an intermediary (e.g. ontology-based) layer can be envisaged for making cross-corpora queries possible irrespective of specific corpus design.

Our symposium is built around these methodological challenges: how were they encountered and overcome in individual cases? Which other challenges have been met, and continue to arise? It should be seen as a forum for the exchange of expertise and experience in this domain. We expect contributions dealing primarily with Uralic minority languages, but corpus building initiatives working with other minority languages and facing similar challenges are also welcome.

Contact person: Ksenia Shagal ksenia.shagal@lmu.de