Ruprecht von Waldenfels, Michał Woźniak

SpoCo - a simple and adaptable web interface for dialect corpora

We present SpoCo, a simple, yet effective system for the web-based query of dialect corpora encoded in ELAN that provides users with advanced concordancing functions, as well as the the possibility to edit and correct transcriptions if needed. SpoCo is easy to use and maintain, and can be adapted to different spoken corpora in a straightforward way. Simplicity is emphasized to facilitate use by a wide range of users and research groups, including those with limited technical and financial resources, and encourage collaboration and data exchange across such groups. Relying on existing technology and pursuing a modular architecture, SpoCo is developed bottom-up: it was initially devised for a specific dialect project and is being continually adapted for use in other projects in a network of Slavic dialect projects that cooperate in tool development and data sharing. SpoCo thus takes a middle position between systems that are developed for the purposes of a specific dialect corpus, on the one hand, and general-use systems designed for a wide range of data and usage cases, on the other.

1 Overview

While the last years have seen the development of a number of corpus query systems that support spoken data, we observe a a lack of powerful, yet simple and effective corpus tools for dialect corpora with aligned audio that are accessible and manageable for linguists with limited computational expertise. Consequently, many dialect projects still do not realize the potential that modern corpus methods provide for their work.

We present SpoCo, a system that provides a workable, stable, and adaptable environment for the presentation of audio-aligned **spo**ken **co**rpora (the acronym alludes to Polish *SpoCo*, 'it's all right, don't worry'). It offers concordancing, statistical functions and user-provided correction of spoken data using standard corpus and web technologies and relying on the de facto standard ELAN format for its input files. SpoCo relies on the possibilities that are provided by the well-established corpus manager OpenCWB (Evert and Hardie, 2011) and adds only a single function - transcription correction by user feedback - to the existing set of functions.

A main feature of SpoCo is its *simplicity* which we see as key in an effort to provide a tool that is easily accessible for researchers that are not particularly versed with computational tools. Despite being user-friendly in its simplicity and intuitiveness, SpoCo does not forgo the possibilities of a modern corpus system, and in fact is one of a handful of systems available that deal with audio primary data.

SpoCo was first developed as a tool for dialectologists in the Ustja River Basin Corpus Project (von Waldenfels et al., 2014) on Russian and has been subsequently adapted for two other projects working on Slavic dialects, namely the Corpus of Spoken Rusyn (Rabus and Šymon, 2015) and the Corpus of the Spisz Dialect (Grochola-Szczepanek, ta). These and other projects work together on tool development and data sharing in the research network *SlaSpoCo*[1]. SpoCo is a system that is developed bottom-up, meeting the needs of specific projects. At the same time, care is taken to develop an adaptable system so that development work can benefit the whole ecosystem.

The present paper is organized as follows. In the following Section 2, we list the requirements and aims of SpoCo. In Section 3, we describe the architecture in some detail. Section 4 covers procedures of data import and automatic annotation; Section 5 describes the user interface in more detail. In Section 6 we illustrate the use of the interface with a typical usage case. We conclude with a perspective on future developments.

2 Requirements and aims of SpoCo

SpoCo was developed to meet four key requirements.

A first requirement was to create an interface that is simple and intuitive without restricting the complex possibilities that a modern corpus manager offers. Simplicity and intuitive accessibility were crucial requirements in the design of SpoCo because a large part of the intended audience of our corpora consists of dialectologists who work in a traditional, rather than variationist or corpus-based, paradigm, and are easily dissuaded from using corpus tools if they present a learning curve that is too steep. This issue is exacerbated by the fact that dialectologists working on Russian who do have corpus experience are typically used to the Russian National Corpus (RNC, www.ruscorpora.ru), which has had an exceptionally simple interface from its very beginnings[2].

In general, we find simplicity to be an undervalued, but key issue in spreading corpus use in and beyond the research community; one of the few cases where this issue was explicitly raised and evaluated was during the construction of the GigaFida corpus of Slovenian, the user base of which was considerably broadened by an effective redesign of its query interface (Arhar Holdt et al., 2012, 19). This issue is similarly relevant to our corpora, which we make accessible to interested lay people and scholars from other fields such as anthropology or history. Overall, we think that simplicity is key in the enhanced relevance of such projects such as ours in the context of the *digital humanities.*

[1] For the Ustya River Basin Corpus, see `http://parasolcorpus.org/Pushkino`; for the Corpus of Spoken Rusyn, see `www.russinisch.uni-freiburg.de/corpus`; for the Corpus of the Spisz Dialect, see `https://spisz.ijp-pan.krakow.pl`. The projects collaborate as part of the network *Corpus-based Research into Sociolinguistic and Dialectal Variation in Slavic Languages* (with the Acronym *SlaSpoCo*, which stands for *Slavic spoken corpora*; see `parasolcorpus.org/Spoken_Slavic`), as well as in other ways.

[2] This is due to the fact that the RNC was initially developed by non-computational linguists in collaboration with the search engine company Yandex and modeled on other interfaces aimed at a general audience. Most other corpora, in contrast, were first developed by computational linguists and more directed at a computer-savvy audience.

A second requirement of SpoCo was to enable researchers to make the actual audio recordings available for listening and download, so that detailed analyses can then be made in specialized tools such as PRAAT. As opposed to corpora of written language, the primary data of a dialect corpus is actual speech in its audible form; any representation of this data in written form constitutes an interpretation that to some extent reflects the primary research question. Making this data directly available is thus crucial. Conversely, ready access to the audio data alleviates the demands on the written representation, as it can be viewed to merely represent an access point to the primary data – therefore, rather pragmatic solutions such as transcription in a standard orthography can be pursued. Standard orthography additionally has the desirable effect that it makes the use of standard tools for annotation, such as taggers and lemmatizers, much more straightforward.

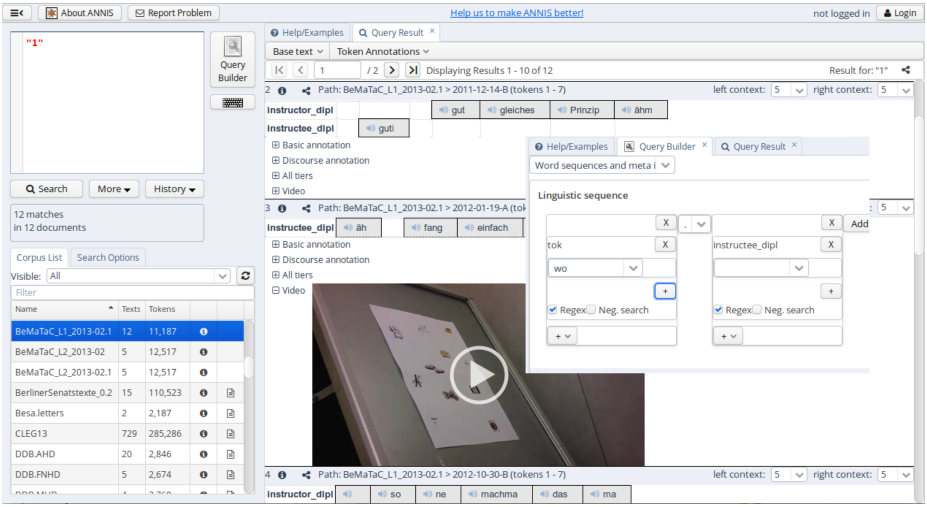
The third requirement has to do with flexibility: as the interface is in continuous development and used for multiple corpus projects, the interface must be easily adaptable. To achieve this, we use AngularJS as a programming tool. In the current version of SpoCo new search fields representing different tiers of annotation and transcription, as well as metadata categories, can easily be specified and semi-automatically integrated into the interface. We feel this is crucial in addressing the inherent contradiction between avoiding a cluttered interface and ensuring simplicity of interaction with the GUI (graphical user interface) on the one hand, and using the interface for a wide range of dialect corpora, on the other. The general workflow used to adopt SpoCo is described in section 5.2.

A fourth, basic requirement is the adherence to best practices in data formats and handling. Most importantly, this means using standard formats wherever possible. While adherence to a standard has obvious advantages such as making it easier to use already existing tools, this requirement also has to do with our view of the status of our tool, which we see as principally provisional. We assume that SpoCo will be superseded by more advanced tools in the coming years, and that the data will be migrated to a new system. Since the data have a much longer life expectancy, potentially being archived for decades or longer, it is imperative that we work with formats that are as standard as possible and will be not be problematic from a middle or long-term perspective. For this reason, we store transcriptions in the XML-encoded ELAN[3] file format, WAV-encoded files for the audio data, and transparently encoded XML files for speaker metadata. We choose ELAN since it has become a de facto standard for spoken corpora with a wealth of available corpora and the capacity to represent complex data in a stand-alone format time-aligned to media files themselves[4].

This specialization and these requirements distinguish SpoCo from other corpus tools that also make time-aligned audio data available, but cater to a wider range of tasks.

[3] ELAN is developed at the Max Planck Institute for Psycholinguistics in Nijmegen and available at `http://tla.mpi.nl/tools/tla-tools/elan/`; see Sloetjes and Wittenburg (2008)

[4] ELAN as a tool is used in some, but not all the projects using SPOCO; in the URB project, most transcribers prefer PRAAT, which is more stable and arguably affords a quicker work flow. Here ELAN is used only to convert the PRAAT files to ELAN format before inclusion.



Name	Texts	Tokens
BeMaTaC_L1_2013-02.1	12	11,187
BeMaTaC_L2_2013-02	5	12,517
BeMaTaC_L2_2013-02.1	5	12,517
BerlinerSenatstexte_0.2	15	110,523
Besa.letters	2	2,187
CLEG13	729	285,286
DDB.AHD	20	2,846
DDB.FNHD	5	2,674

Figure 1: An example query result in ANNIS3, with query builder pasted into the image. ANNIS is very powerful, but also rather complex in use.

Two such systems seem to be particularly relevant for the kind of task that we are faced with. First, ANNIS3 (Krause and Zeldes, 2016) is a corpus system that is geared towards handling annotations of great complexity with multiple corpora of many types. However, ANNIS3 does not fulfill at least two of our requirements since it has a rather complex interface (see figure 1) and does not allow users to download chunks of the aligned primary audio data for further analysis. A second such tool is GLOSSA (Kosek et al., 2015), developed at the University of Oslo for the inclusion of a great variety of data, including the Nordic dialect corpus; GLOSSA has a number of functions that directly cater to dialect corpora building. However, in our experience, it proved difficult to install and it is unclear how to maintain and adapt it to our specific needs without offering the GUI simplicity that we are looking for. Moreover, it does not address the archiving problem since the corpus is essentially kept in CWB vertical format, rather than a standard XML format of some sorts.

Other, more specialized tools are likewise too complex to handle for a small dialectological group; these include the tools developed by the Czech national corpus project for the DIALEKT and ORTOFON corpora (Kopřivová et al., 2014), or the Edisyn interface (Barbiers, 2015).

In the design of SpoCo, we aim to cover the ground between complex, one-size-fits-all interfaces, such as GLOSSA, ANNIS, or CQPWeb (Hardie, 2012) on the one hand, and specialized corpus interfaces on the other hand, such as Edisyn Barbiers (2015) and many other in-house solutions that are never released to the general public as software.

Our approach is to straddle these two worlds by developing a system that is constructed bottom-up, driven by concrete tasks, but at the same time stays flexible and adaptable to new projects, all of which share development costs in a network of related projects.

3 SpoCo architecture and set-up

3.1 Overview of the architecture

SpoCo consists of three main components: 1) the actual linguistic data, 2) the corpus management back end, which supplies concordancing and statistical functions, and 3) the web interface. Each part is largely independent from the rest, which makes changing or replacing individual parts straightforward. SpoCo is designed to be deployed on a standard machine running Ubuntu Linux with Apache (LAMP server) and CWB; no further components are required. Currently, a number of different technologies are involved in the corpus preparation and management process, most notably XML, XSLT, PHP and Apache Server, perl, python and AngularJS. Below we describe the three components in more detail.

3.2 Linguistic data: types and import procedures

The linguistic resources SpoCo uses consist of three types: audio recordings, transcriptions of these, and speaker metadata (i.e., gender, age, place of residence and recording, mobility, etc.). Speaker metadata are technically optional, but they play a potentially crucial role in analysis and are thus useful when searching and presenting results. During corpus encoding, the transcription is split into text segments that were delimited as utterances during transcription in ELAN (or a different transcription tool which the data is converted from). Sound files (supplied in lossless wav format) are split into the corresponding audio segments.

Audio and transcription data are kept in two separate directories and are implicitly linked by identical files names. Adding new files is as simple as copying them into the appropriate directory and issuing a command to re-encode the corpus. Note that while new sound files are typically only added after field work trips, ELAN files are added and updated continuously as transcriptions become completed.

Metadata concerning speakers and recordings are managed separately in a dedicated database (for that purpose we use a DJANGO-based system not described here[5]). For archiving and inclusion purposes into the query system, they are exported and saved as an XML-file; specific metadata fields which should be available in the corpus can be specified during installation.

3.3 Corpus management back end

For storing and querying corpus data, SpoCo uses Corpus Workbench (CWB), a stable and powerful corpus management software which provides sophisticated query and

[5] Metadata management was implemented as part of the TriMCo project by Ilya Khait, Leipzig.

statistical functions. CWB is widely supported, so that, e.g., integration into R is easily accomplished, and it is actively under development (Hardie and Evert 2014). CWB is not resource intensive, in our experience very stable and easy to install, and thus ideal for our purposes. We anticipate that the current version of CWB will eventually be replaced, quite possibly by its successor CWB4 which promises better handling of XML files as well as a new, improved data structure (Evert and Hardie, 2015)

3.4 Web interface

The transcribed data are available for advanced querying through a corpus interface that prepares and sends queries to the corpus manager; this interface is based on previous interfaces for parallel and diachronic data (von Waldenfels, 2011; von Waldenfels and Rabus, 2015). CWB is configured to return the results in XML, which are then displayed using XSLT sheets. This approach affords the advantage of clear separation of corpus manager and output display, as well as simple adaptation of the result page to different needs. Thus, using a different corpus manager is greatly simplified and adding new export formats (e.g., csv) is as simple as specifying a different XSLT sheet in the output. Altogether, this makes the inclusion of new data types straightforward.

Currently, this interface exists in two versions, both of which are geared towards maximal simplicity to make it accessible for a wide range of users. Both versions share most of their functionalities: user management, corpus querying, a correction module, full-text browsing. The main difference between them is the technology they are built on: the initial version (developed for the URB) uses mostly simple HTML and some JavaScript and PHP, while the second version is built with the modern JavaScript framework AngularJS (version 2.1). We chose this framework because it is interface oriented, flexible and scalable; web-page content is easily updated without the need to refresh, and therefore features such as the construction of the CQL query on-the-fly or switching interface languages are easily accessible for both the user and the developer. Both interfaces produce identical output: CQP queries that are channeled through the back end. A more detailed comparison is provided in the next section. Both interfaces are completely interchangeable, which is a good example of the flexibility that the SpoCo modular architecture allows.

4 Using SpoCo I: the back end

4.1 SpoCo integration

For the installation of SpoCo for use with a specific corpus, the interface is copied into a directory that Apache can access, and the settings files are set to contain paths to CWB and data directories, as well as (in the second version) information about the metadata fields in use and languages available in the interface. Depending on specifics of the corpus data (e.g. the number of transcription levels and automatic annotation procedures), the web interface and the inclusion script require adaptation.

4.2 Corpus preprocessing and conversion

After the ELAN-encoded transcriptions are added to the corpus (i.e., copied into the appropriate directories), these are enhanced using automatic tools and converted to the corpus manager CWB for easier querying and simple html files for reading. This involves the following steps triggered by a shell script calling a heterogeneous set of utilities:

- the ELAN files and the XML file containing the metadata are converted into a single, CWB-compatible file in *vertical format*: one token or xml tag on each line
- further annotation such as lemma and morphological tags are then added to this file using standard tools. This is, in general, corpus-specific: in the case of the URB, the Treetagger (Schmid, 1999) is employed using a model trained on the Russian National Corpus. In the case of Rusyn corpus, a custom-made approach to tagging with Levenshtein distances is being developed to take into account variation due to diverse transcription standards (ongoing work by Achim Rabus, Freiburg, and Yves Scherrer, Geneva.); this approach is expected to be relevant for other corpora in the network, as well.
- based on the segmentation in the ELAN file, the audio files are cut into small chunks for downloading and broadcasting
- html versions of each transcribed text are prepared for full in-context reading

This script is invoked each time the corpus data is changed and can be triggered by users via the web interface.

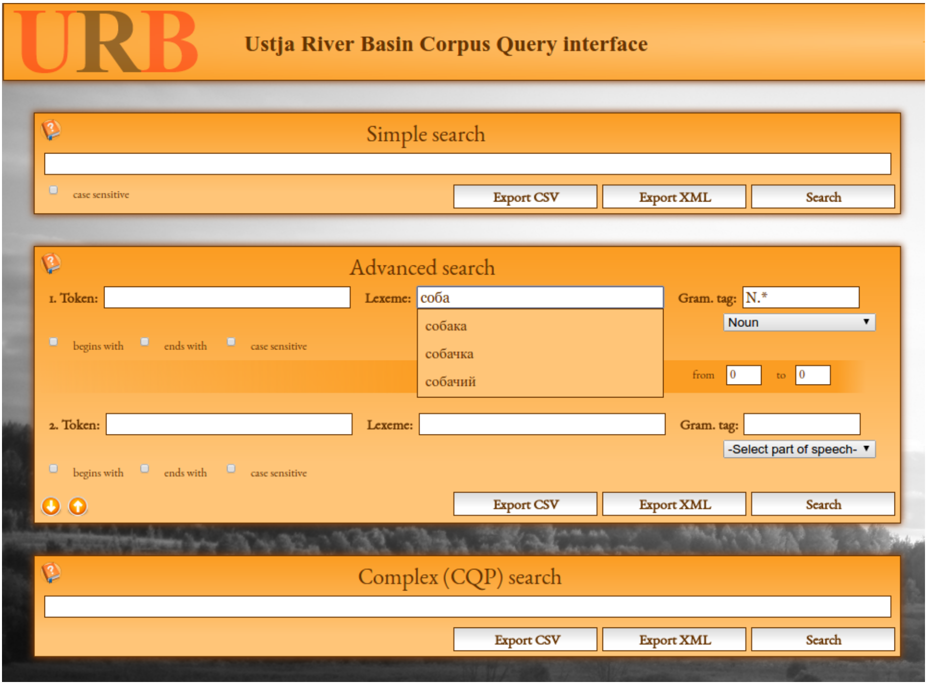
5 Using SpoCo II: interface features

5.1 Corpus query

In the following, we focus on describing the initial version that was developed for the URB but also highlight differences in the newer version used in the two other projects.

Initially, users are asked to log in to reach the query page. There are three user categories: *guests* can only search the corpus; *registered users* can also correct transcriptions; *administrators* can validate corrections and re-encode the corpus.

Figure 2 shows the main query page (a second pane with help, sample queries and corpus statistics is not shown here for reasons of space). Queries are available on three levels of complexity: *simple search* allows users to search for contiguous phrases just like in the search field of a word processor; *advanced search* allows users to specify search words in terms of wordform, lemma and morphological tag with variable distance between them. This option is very similar to the RNC's interface and therefore familiar to scholars working on Russian. Finally, the *complex (CQP) search* allows users to enter any valid CQP query. This is used for more complex queries, sorting and metadata



Lexeme: соба

Gram. tag: N.*

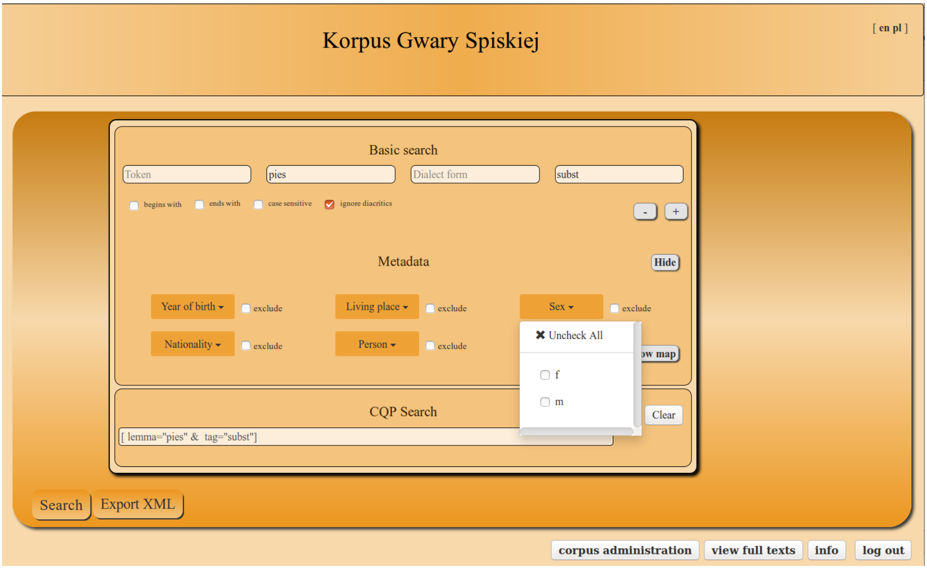
Noun

from 0 to 0

Figure 2: Search interface for the URB, with simple, advanced and expert search options, offering drop-down lists of word forms and lemmata and an interactive panel for morphological tag construction.

filtering; for example, negative conditions or restrictions on specific categories of speakers can be formulated here.

For the *Corpus of Spoken Rusyn* (Rabus and Šymon, 2015) and subsequently for the *Spisz Dialect Corpus*, the query page was taken to the next step. While the rest of the system remains essentially identical, the query page was reprogrammed in AngularJS, allowing for multiple interface languages to be included and for search fields to be more easily adapted, and thus for customization for different corpora. It introduces three new features (see Figure 3): First, metadata can now be searched for in the GUI; this is customized in a settings file where these fields, their display names and default values are specified during installation. Second, the interface now integrates a Google maps application for geographic visualization and filtering. The interface is described in detail in Rabus and Šymon (2015). Third, the basic and CQP searches are now linked dynamically: filling in fields in the *basic search* automatically constructs the *CQP search*, which can then be manually adapted as needed for more advanced queries.



Korpus Gwary Spiskiej

Basic search

Token | pies | Dialect form | subst

□ begins with □ ends with □ case sensitive ☑ ignore diacritics

Metadata

Year of birth □ exclude Living place □ exclude Sex □ exclude

Nationality □ exclude Person □ exclude

✖ Uncheck All

□ f

□ m

CQP Search

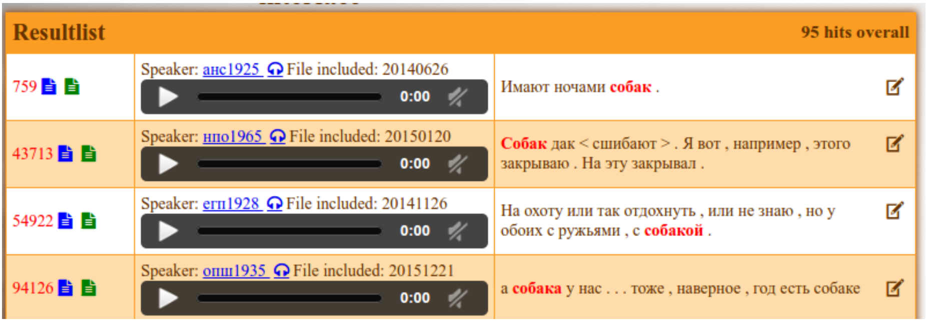
[lemma="pies" & tag="subst"]

Figure 3: The search interface for the *Spisz Dialect Corpus*, with additional metadata search options, interactive maps for places of recording (not shown), and a dynamic link between the *basic search* and *CQP search*.

This last feature not only has the effect that complex CQP queries can be constructed more quickly, it also makes learning CQP much easier for novice users who can now observe the CQP queries as they are being constructed in response to their filling in the *basic search* fields. Then, users can learn to use CQP by changing these queries, rather than having to construct them from scratch. In our experience, this greatly lowers the threshold for beginning to use complex queries in CQP.

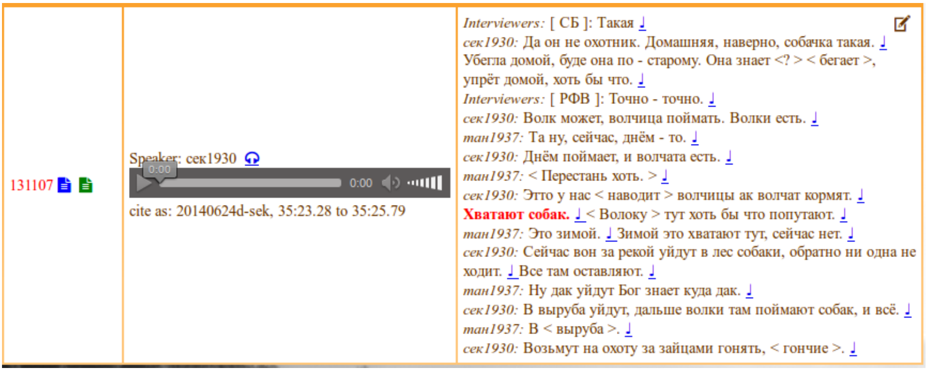
5.1.1 Corpus query results

Figure 4 displays the beginning of the concordance for the lexeme *sobaka* ‘dog’; here, each example includes the respective audio segment. Each corpus hit is provided with links to a tab-delimited csv view, with directly downloadable audio segments for analysis in PRAAT or other speech analysis software, and with a file containing basic metadata for each speaker. Users can also examine each example with more context in a separate window (see Figure 5); the URLs for these separate windows serve as a unique identifiers based on the location in the audio timeline. To the very right, registered users can click the paper-and-pen icon in order to be able to edit the transcription.



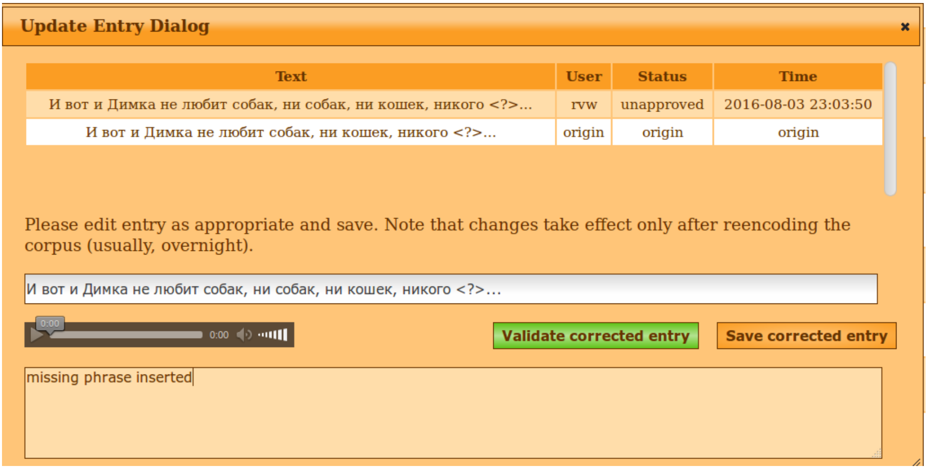
Resultlist		95 hits overall
759	Speaker: анс1925 File included: 20140626 0:00	Имают ночами **собак** .
43713	Speaker: нпо1965 File included: 20150120 0:00	**Собак** дак < сшибают > . Я вот , например , этого закрываю . На эту закрывал .
54922	Speaker: егп1928 File included: 20141126 0:00	На охоту или так отдохнуть , или не знаю , но у обоих с ружьями , с **собакой** .
94126	Speaker: опш1935 File included: 20151221 0:00	а **собака** у нас . . . тоже , наверное , год есть собаке

Figure 4: Query results for *sobaka* 'dog'. Each row in the result list provides access to (from left to right): context view with persistent URL (blue icon), csv-export view (green icon), link to *speaker metadata* found in a google spreadsheet, audio fragment in wav format (headphones icon), the date the transcription was added, the text itself with search item highlighted in red, a link for registered users to provide corrections (paper-and-pen icon).



131107	Speaker: сек1930 0:00 0:00 cite as: 20140624d-sek, 35:23.28 to 35:25.79	*Interviewers:* [СБ]: Такая ♩ *сек1930:* Да он не охотник. Домашняя, наверно, собачка такая. ♩ Убегла домой, буде она по - старому. Она знает <? > < бегает >, упрёт домой, хоть бы что. ♩ *Interviewers:* [РФВ]: Точно - точно. ♩ *сек1930:* Волк может, волчица поймать. Волки есть. ♩ *ман1937:* Та ну, сейчас, днём - то. ♩ *сек1930:* Днём поймает, и волчата есть. ♩ *ман1937:* < Перестань хоть. > ♩ *сек1930:* Этто у нас < наводит > волчицы ак волчат кормят. ♩ **Хватают собак.** ♩ < Волоку > тут хоть бы что попутают. ♩ *ман1937:* Это зимой. ♩ Зимой это хватают тут, сейчас нет. ♩ *сек1930:* Сейчас вон за рекой уйдут в лес собаки, обратно ни одна не ходит. ♩ Все там оставляют. ♩ *ман1937:* Ну дак уйдут Бог знает куда дак. ♩ *сек1930:* В выруба уйдут, дальше волки там поймают собак, и всё. ♩ *ман1937:* В < выруба >. ♩ *сек1930:* Возьмут на охоту за зайцами гонять, < гончие >. ♩

Figure 5: Context view, with citation instructions and multiple speakers.



Text	User	Status	Time
И вот и Димка не любит собак, ни собак, ни кошек, никого <?>...	rvw	unapproved	2016-08-03 23:03:50
И вот и Димка не любит собак, ни кошек, никого <?>...	origin	origin	origin

Please edit entry as appropriate and save. Note that changes take effect only after reencoding the corpus (usually, overnight).

И вот и Димка не любит собак, ни собак, ни кошек, никого <?>...

missing phrase inserted

Figure 6: Window for transcription correction.

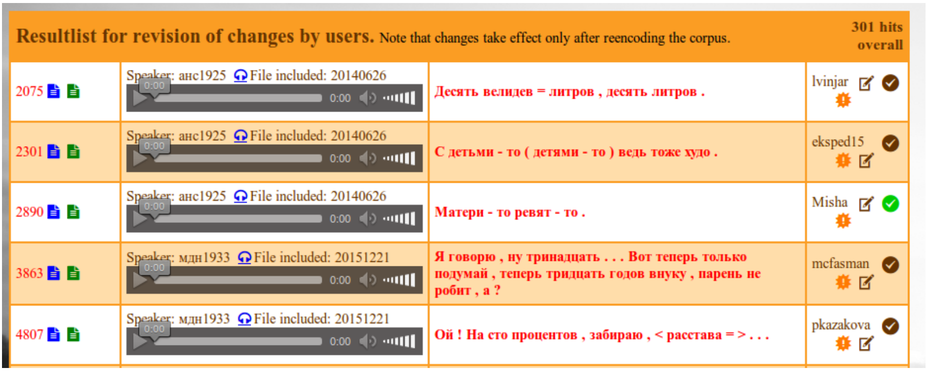
5.1.2 Transcription edit function (error correction)

The interface for editing transcriptions is implemented using XSLT and JavaScript and shown in figure 6. It is only available to registered users. The user is provided with an option to enter a new transcription and leave a comment; the segmentation itself cannot be changed online. This new transcription is written directly to the ELAN files as an additional transcription together with the user name, time stamp, and a status field stating that this is an unconfirmed correction. This new and all previous versions are kept in the ELAN file in parallel; unconfirmed changes need to be reviewed and validated by an administrator before they are marked as accepted (or reverted if necessary); see Figure 7. In this way, each correction is double-checked.

Changes appear in corpus results only after the corpus has been re-encoded in CWB; this is done every night automatically or on demand after the validation process. Corrections are flagged in the query results, together with their status as confirmed or unconfirmed, and the edit function gives access to all previous versions of the transcription segment. This function is used quite a lot; in the URB corpus today, which has roughly 750 000 tokens (excluding interviewers), around 3000 lines have been corrected.

5.1.3 Full text views

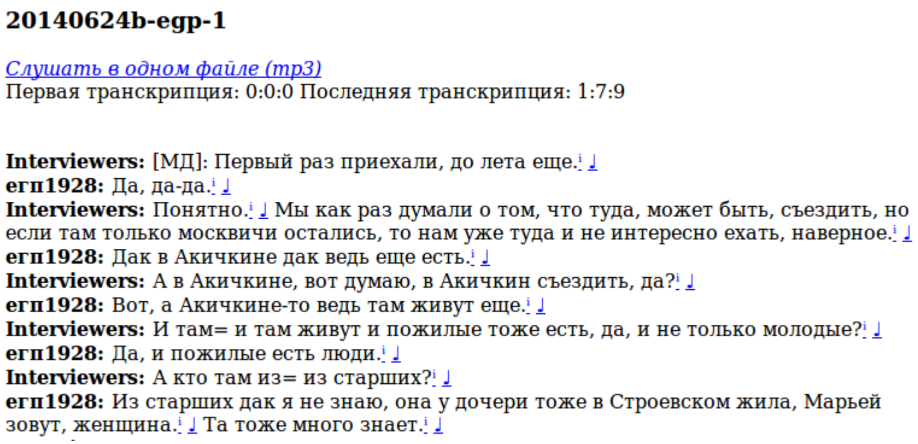
In many cases, users want to read interview transcripts in their entirety as full text. In addition to using the query page, registered users may listen to and read complete transcriptions, as shown in figure 8. These full texts are derived from the ELAN files



Resultlist for revision of changes by users. Note that changes take effect only after reencoding the corpus. **301 hits overall**

2075	Speaker: анс1925 File included: 20140626 0:00	Десять велидев = литров , десять литров .	lvinjar
2301	Speaker: анс1925 File included: 20140626 0:00	С детьми - то (детями - то) ведь тоже худо .	eksped15
2890	Speaker: анс1925 File included: 20140626 0:00	Матери - то ревят - то .	Misha
3863	Speaker: мдн1933 File included: 20151221 0:00	Я говорю , ну тринадцать . . . Вот теперь только подумай , теперь тридцать годов внуку , парень не робит , а ?	mcfasman
4807	Speaker: мдн1933 File included: 20151221 0:00	Ой ! На сто процентов , забираю , < расстава = > . . .	pkazakova

Figure 7: List of unconfirmed edits for administrators, with basic information on who made the change, its status, and the possibility of one-click confirmations.

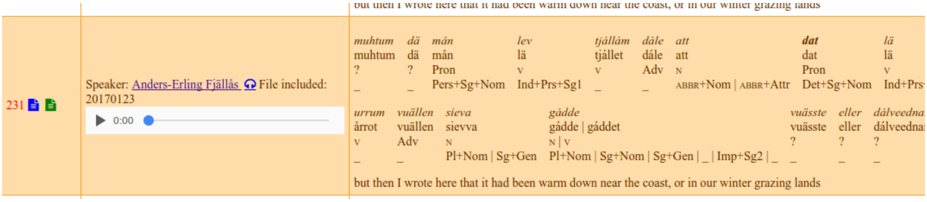


20140624b-egp-1

Слушать в одном файле (mp3)
Первая транскрипция: 0:0:0 Последняя транскрипция: 1:7:9

Interviewers: [МД]: Первый раз приехали, до лета еще.ⁱ ♩
егп1928: Да, да-да.ⁱ ♩
Interviewers: Понятно.ⁱ ♩ Мы как раз думали о том, что туда, может быть, съездить, но если там только москвичи остались, то нам уже туда и не интересно ехать, наверное.ⁱ ♩
егп1928: Дак в Акичкине дак ведь еще есть.ⁱ ♩
Interviewers: А в Акичкине, вот думаю, в Акичкин съездить, да?ⁱ ♩
егп1928: Вот, а Акичкине-то ведь там живут еще.ⁱ ♩
Interviewers: И там= и там живут и пожилые тоже есть, да, и не только молодые?ⁱ ♩
егп1928: Да, и пожилые есть люди.ⁱ ♩
Interviewers: А кто там из= из старших?ⁱ ♩
егп1928: Из старших дак я не знаю, она у дочери тоже в Строевском жила, Марьей зовут, женщина.ⁱ ♩ Та тоже много знает.ⁱ ♩

Figure 8: Full text view. The note symbols after each sentence enable the researcher to listen to audio segments and jump to the corpus view for each utterance.



but then I wrote here that it had been warm down near the coast, or in our winter grazing lands

231

Speaker: Anders-Erling Fjällås File included: 20170123

▶ 0:00

<table>
<tr><td><i>muhtum</i></td><td><i>dä</i></td><td><i>mån</i></td><td><i>lev</i></td><td><i>tjållåm</i></td><td><i>dåle</i></td><td><i>att</i></td><td><b><i>dat</i></b></td><td><i>lä</i></td></tr>
<tr><td>muhtum</td><td>dä</td><td>mån</td><td>lä</td><td>tjållet</td><td>dåle</td><td>att</td><td>dat</td><td>lä</td></tr>
<tr><td>?</td><td>?</td><td>Pron</td><td>V</td><td>V</td><td>Adv</td><td>N</td><td>Pron</td><td>V</td></tr>
<tr><td>_</td><td>_</td><td>Pers+Sg+Nom</td><td>Ind+Prs+Sg1</td><td>_</td><td>_</td><td>ABBR+Nom | ABBR+Attr</td><td>Det+Sg+Nom</td><td>Ind+Prs</td></tr>
</table>

<table>
<tr><td><i>urrum</i></td><td><i>vuällen</i></td><td><i>sieva</i></td><td><i>gådde</i></td><td><i>vuässte</i></td><td><i>eller</i></td><td><i>dålveedna</i></td></tr>
<tr><td>årrot</td><td>vuällen</td><td>sievva</td><td>gådde | gåddet</td><td>vuässte</td><td>eller</td><td>dålveedna</td></tr>
<tr><td>V</td><td>Adv</td><td>N</td><td>N | V</td><td>?</td><td>?</td><td>?</td></tr>
<tr><td>_</td><td>_</td><td>Pl+Nom | Sg+Gen</td><td>Pl+Nom | Sg+Nom | Sg+Gen | _ | Imp+Sg2 | _</td><td>_</td><td>_</td><td>_</td></tr>
</table>

but then I wrote here that it had been warm down near the coast, or in our winter grazing lands

Figure 9: A result page from the glossed Pite Saami corpus.

using an XSLT transformation, and provide links to the full audio file, audio segments as well as to the corpus view for each utterance.

5.2 Adapting SpoCO: a sample case case

In late 2016, SpoCo was adopted for two non-Slavic corpora; first, a corpus of Pite Saami (Wilbur, 2017), and second, a corpus of dialectal Lithuanian (as part of the planned TrimCo[6] corpus). As many projects aimed at documenting endangered languages, these corpora supply morphological glosses; support for this feature was added to SpoCo on this occastion. The adaptation of SpoCo consisted in three steps:

- adaptation of the scripts converting ELAN to CWB; this entailed decoding the hierarchical relationship in the ELAN-file to obtain token-based glosses and encoding free translations as xml attributes at utterance level
- adaptation of settings in SpoCo to query and return glosses and free translations
- adaptation of the XSLT sheet that displays the resulting XML using an open source library that displays glossed text

Figure 9 shows the results of a query in the Pite Saami corpus.

6 A sample workflow and desirable features

To exemplify the role of SpoCo, below we describe the workflow of a typical investigation of a dialect variable in the URB project. Specifically, our example concerns the dialectal shift of /a/ to /e/ between palatalized consonants (a-raising, see Požarickaja 2005, 42f.) that is characteristic of the speech of older speakers of the Ustja dialect. In the URB, as well as in most of the projects using SpoCo, transcriptions are written in standard orthography with only very limited representation of dialectal features; see von Waldenfels et al. (2014); Gerstenberger et al. (2017) for discussions of the advantages of such an approach.

[6] Triangulation Approach for Modelling Convergence with a High Zoom-In Factor; see `http://www.trimco.uni-mainz.de/`. PI: Björn Wiemer, Mainz; Lithuanian Subcorpus: Kirill Kozhanov, Moscow.

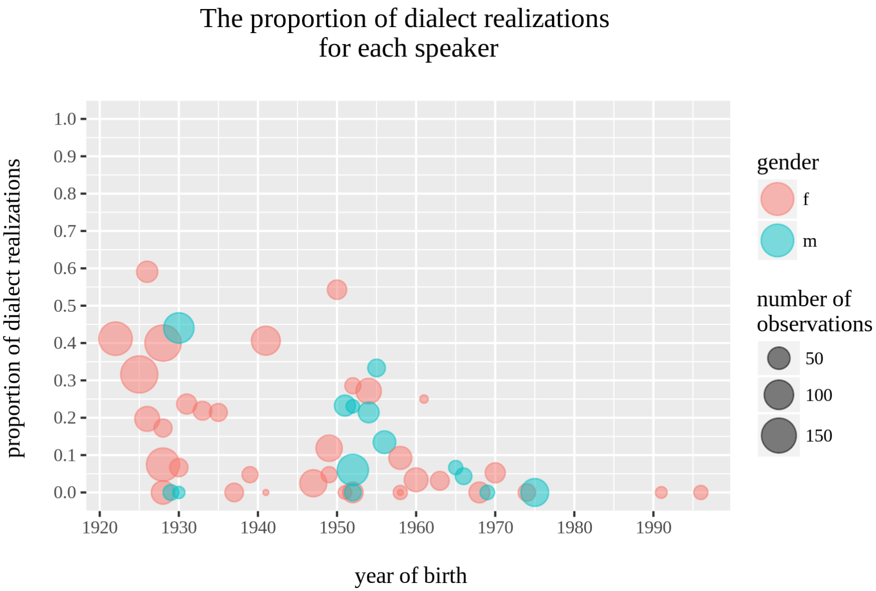


Figure 10: An example variable (Kazakova, 2016). The plot gives the relative instances of historical /a/raising to [e] between soft consonants for speakers born between 1922 and 1995; red circles represent females, green circles males. The size of the circles represent the number of instances in the study.

In a first step, the envelope of variation is defined and searched for in terms of the standard language – in this case, all word forms that contain /a/ between soft consonants in the standard orthography are queried, downloaded as csv, copied, and pasted in OpenOffice Calc or some other offline tool. The CSV contains links to the audio segment files, so that each example is categorized with respect to the actual audio data (rather than a transcription). Since the download also includes basic speaker metadata, the resulting categorization thus affords a simple plot of speakers, ordered by date of birth, and with respect to the relative proportion of dialectal as opposed to standard pronunciation. Figure 10 gives an example of such a plot, which nicely shows the dialectal feature's tendency to recede in an apparent time perspective.

Transcription into standard orthography as opposed to a phonetic alphabet (IPA or similar) effectively allows the phonetic analysis to be postponed until it is needed for specific research question, thereby streamlining and focusing it and limiting the overall work load involved. In the future, it would be highly desirable to allow users to upload the result of such annotation tasks so that they can be viewed and used by future users.

7 Summary and future developments

We have presented SpoCo, a system to query and analyze spoken corpora with aligned audio data. The system is *pragmatic* in that it aims to provide facilities that are needed in a number of concrete Slavic dialect projects. At the same time, it follows an overarching agenda to enable collaborative tool development across different projects; with this in mind, care is taken that the system is modular and expandable for use with other, related projects.

An important aim of SpoCo is to create a system with a low threshold of use for a wide range of projects, including those with limited computational expertise and resources. Specifically, we aim for a stable, hassle-free, easy-to-use system that is simple yet effective. We see this as as an important methodological contribution to the field since using such a system and its collaborative development goes hand in hand with data sharing and the adoption of innovative research methods. An important aim for the future is thus to make the deployment of SpoCo easier for new projects and to work on making the customization of SpoCo simpler yet more flexible as well as to work on further integration of the workflow of corpus file management (including distributed archiving), metadata acquisition, and subsequent data annotation.

References

Arhar Holdt, Š., Kosem, I., and Logar Berginc, N. (2012). Izdelava korpusa Gigafida in njegovega spletnega vmesnika. In Erjavec, T. and Žganec Gros, J., editors, *Zbornik Osme konference Jezikovne tehnologije*, pages 12–17. Institut Jožef Stefan, Ljubljana.

Barbiers, S. (2015). European dialect syntax: Towards an infrastructure for documentation and research of endangered dialects. In Jones, M., editor, *Endangered Languages and New Technologies. Cambridge: CUP*. CUP, Cambridge.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference, Birmingham, UK*. University of Birmingham.

Evert, S. and Hardie, A. (2015). Ziggurat: A new data model and indexing format for large annotated text corpora. In *Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora (CMLC-3)*, page 21–27, Lancaster, UK.

Gerstenberger, C., Partanen, N., Rießler, M., and Wilbur, J. (2017). Utilizing language technology in the documentation of endangered Uralic languages. In Pirinen, T. A., Trosterud, T., and Tyers, F. M., editors, *Northern European Journal of Language Technology: Special Issue on Uralic Language Technology.*

Grochola-Szczepanek, H. (t.a.). Korpusowe badania języka mieszkańców spisza w polsce – cele i zadania. *Jezikoslovni zapiski*, 22(2).

Hardie, A. (2012). CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Kazakova, P. (2016). Alternation of [a]/[e] between palatalized consonants under stress in the dialect of the village mikhalevskaya. Unp. manuscript.

Kopřivová, M., Klimešová, P., Goláňová, H., and Lukeš, D. (2014). Mapping diatopic and diachronic variation in spoken Czech: the ORTOFON and DIALEKT corpora. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 376–382, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kosek, M., Nøklestad, A., Priestley, J., Hagen, K., and Johannessen, J. B. (2015). Visualisation in speech corpora: maps and waves in the Glossa system. In Grigonytė, G., Clematide, S., Utka, A., and Volk, M., editors, *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, pages 23–31. Linköping University Electronic Press.

Krause, T. and Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Požarickaja, S. K. (2005). *Russkaja dialektologija.* Gaudeamus, Moscow.

Rabus, A. and Šymon, A. (2015). Na novŷx putjax isslidovanja rusyns'kŷx dialektu: korpus rozhovornoho rusyns'koho jazŷka. In Koporova, K., editor, *Rusyn'skŷj literaturnŷj jazŷk na Slovakiji. 20 rokiv kodifikaciji / The Rusyn literary language in Slovakia. 20th anniversary of its codification. IV. International Congress of the Rusyn Language. Prjašiv, 23. - 25. 09. 2015*, pages 40–54.

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to German. In Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., and Yarowsky, D., editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Processing*, pages 13–26. Kluwer Academic Publishers, Dordrecht.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).*

von Waldenfels, R. (2011). Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In Majchráková, D. and Garabík, R., editors, *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011, Modra, Slovakia, 20–21 October 2011*, pages 156–162, Bratislava. Tribun EU.

von Waldenfels, R., Daniel, M., and Dobrushina, N. (2014). Why standard orthography? Building the Ustya River Basic Corpus, an online corpus of a Russian dialect. In *Komp'juternaja lingvistika i intellektual'nye technologii: Po materialam ežegodnoj Meždunarodnoj konferencii «Dialog» (Bekasovo, 4 — 8 ijunja 2014 g.) Vyp. 13 (20)*, Moskva. Izd-vo RGGU.

von Waldenfels, R. and Rabus, A. (2015). Recycling the metropolitan: building an electronic corpus on the basis of the edition of the *Velikie Minei Čet'i. Scripta & e-Scripta*, 14/15:27–38.

Wilbur, J. (2008–2017). Pite saami. In *Endangered Languages Archive (ELAR)*. SOAS, University of London.