Josip Batinić,¹ Elena Frick¹ and Thomas Schmidt¹

Accessing spoken language corpora: an overview of current approaches

Abstract

In this paper, we present an overview of freely available web applications providing online access to spoken language corpora. We explore and discuss various solutions with which the corpus providers and corpus platform developers address the needs of researchers who are working with spoken language. The paper aims to contribute to the long-overdue exchange and discussion of methods and best practices in the design of online access to spoken language corpora.

Keywords: corpus access, corpus search platform, spoken language corpora.

1. Introduction

There have been significant developments in the field of spoken language research over the past few decades. New technologies, as well as a more efficient use of the existing ones, have made work with corpora of spoken language more prolific and specialised. Many corpora of spoken language are not in their infancy anymore but have evolved into elaborate systems with dedicated tools and interfaces used to gain access and explore the data they hold. The Database for Spoken German (Datenbank für Gesprochenes Deutsch, DGD),² the tools and corpora of the Hamburg Centre for Language Corpora (Hamburger Zentrum für Sprachkorpora, HZSK)³ and the GeWiss⁴ platform (Gesprochene Wissenschaftssprache, [‘Spoken Academic Language’]) are three significant examples of such systems providing virtual workspace for investigating spoken German language. Each one of these multi-modal platforms is online and in daily use, and

¹ Leibniz-Institut für Deutsche Sprache, R5, 6–13, D-68161 Mannheim, Germany.

Correspondence to: Josip Batinić, e-mail: josip.batin@gmail.com

² See: <https://dgd.ids-mannheim.de>.

³ See: <https://corpora.uni-hamburg.de/hzsk/>.

⁴ See: <https://gewiss.uni-leipzig.de>.

they currently have over 10,000, 1,600 and 1,000 (as of October 2019) registered users, respectively. They are associated with extensive corpus repositories and offer users a full range of practical functionalities for browsing and querying data, as well as evaluating, storing and exporting the search results. Nevertheless, these three web applications focus on different user needs since they originate in separate independent projects with differing target audiences. A recent user survey (Fandrych *et al.*, 2016), however, showed that many research scenarios would benefit from stronger network and compatibility relations between these three corpus platforms.

The ZuMult⁵ project started in 2018 through co-operation between the Leibniz-Institute of German Language (Leibniz-Institut für Deutsche Sprache, IDS) in Mannheim, the Hamburg Centre for Language Corpora (HZSK) in Hamburg, and the Herder-Institute at the University of Leipzig—the research centres of the three spoken corpus platforms mentioned above. The project aims to develop a software architecture that will improve the inter-operability of—and ensure unified access to—spoken language resources located in different repositories (for more information, see Batinic *et al.*, [2019]). The first step toward achieving this goal was to perform an in-depth analysis of DGD, HZSK, GeWiss, as well as several other freely available web applications providing online access to spoken corpora of a wide range of languages (English, French, Russian, Slovenian, Bulgarian, Czech, *etc.*). In that way, we intend to identify what means of retrieving and handling data the platforms use, what user-group specific services are provided for the corpora, and which functionalities the platforms use to achieve that.

Following up on Merkel and Schmidt's (2009) overview of spoken corpora, we present in this paper an updated state-of-the-art review. In contrast to Merkel and Schmidt, we approach here the spoken corpus platforms not only from the user perspective, but also from the functional and technical viewpoint, which is our main focus. We elaborate on this in Section 3 ('Method of analysis').

We start out by briefly presenting the platforms we considered in this research (Section 2: 'Platforms'). This is followed by an elaboration on the adopted methodology (Section 3: 'Method of analysis'). In Section 4 ('Results and discussion'), we systematically examine the outlined categories of classification, point out the major findings, and discuss their relevance for this work, the essence of which is then summarized in the final section (Section 5: 'Conclusion').

⁵ See: <https://zumult.org/>.

2. Platforms

Along with the DGD, GeWiss and HZSK, we examined numerous other online platforms that are available free of charge, and which provide web-browser-based access to spoken corpora. We present some of them in order to demonstrate the diversity of solutions available to spoken language corpora online.

2.1 Datenbank für Gesprochenes Deutsch (DGD)

The DGD (Schmidt, 2017) provides web-based access to spoken language corpora from the Archive for Spoken German (Archiv für Gesprochenes Deutsch, AGD)⁶ in Mannheim. Only a part of the AGD corpora is available through DGD for exploration, but the proportion is increasing annually. Currently it contains thirty-four corpora documenting dialectal and regional varieties of German as well as spontaneous natural interaction in the German language. The data were collected from different regions of Germany and worldwide, and recorded at different time periods, with the oldest dating back to 1955. The corpora cover a wide range of topics, social situations, interaction types and communication formats. The total duration of audio/video files currently contained in the database amount to more than 4,500 hours. The larger part of the corpora contents are transcribed, annotated and time-aligned with the underlying audio recording. The total number of transcribed tokens currently exceeds 18 million.

The platform requires a one-time registration in order to access the database. Once logged in, the user can explore the corpora by browsing metadata, audio and video recordings, transcripts, and additional materials (e.g., map-task documentation). Concordance searches can be performed on user-selected transcripts by specifying the properties of individual tokens (transcribed form, normalised form, lemma and part-of-speech). The token search query can be combined with the metadata, with context or with positional filters in order to refine the results further. Query results can be complemented with corresponding metadata values, exported in various formats or be saved as virtual collections, which can then be shared or stored for later use. It is also possible to create and search virtual corpora (i.e., user-compiled selections of datasets). The DGD platform is built on top of an Oracle database and is available online.⁷

⁶ See: <http://agd.ids-mannheim.de/>.

⁷ See: <https://dgd.ids-mannheim.de/>.

2.2 Hamburger Zentrum für Sprachkorpora (HZSK)

The HZSK (Jettka and Stein, 2014) is a digital corpus archive with a web platform providing access to a wide selection of corpora originating mainly in various projects at the University of Hamburg. Most corpora are spoken language corpora that focus on multilingualism. The corpora are very heterogeneous in their design, size, transcription conventions and available annotations. Some of them can be browsed for metadata, audio/video recordings and time-aligned transcripts, and can be downloaded from the HZSK website without registration. Other corpora are only available for research purposes after a successful single sign-on (SSO) login. Several kinds of data can only be accessed through individual authorisation, which requires detailed information from the researcher about their intended research purposes. The corpora can be downloaded in different formats, thus allowing users to query the data locally using open source tools, such as EXAKT⁸ or Coma.⁹ Some corpora are searchable online on the HZSK website¹⁰ through an installed instance of ANNIS,¹¹ an open-source web-based search and visualisation tool for multi-layer linguistic corpora. Altogether, there are about 2,000 hours of transcribed and non-transcribed audio/video material. The HZSK platform is based on Fedora Commons, an open-source software framework connected through Islandora¹² to Drupal,¹³ and it is available online.¹⁴

2.3 Gesprochene Wissenschaftssprache (GeWiss)

GeWiss (Slavcheva and Meißner, 2014) is a comparable corpus of spoken academic language. The corresponding corpus platform was developed at the University of Leipzig. It provides access to a comparable corpus consisting of recorded and transcribed L1 and L2 speech from German, English, Polish and Italian students and other members of the university. Oral exams and oral presentations by researchers and students are the interaction types recorded for GeWiss.

After a one-time registration and login, users can browse the metadata, audio recordings and the corresponding time-aligned transcripts. Users can also use the concordance view for searching transcribed tokens and single segment-based annotations such as code-switching and discourse comments, available in some GeWiss sub-corpora. Extensive metadata filters

⁸ See: <https://exmaralda.org/de/exakt-de/>.

⁹ See: <https://exmaralda.org/de/corpus-manager-de/>.

¹⁰ See: <https://corpora.uni-hamburg.de/annis/>.

¹¹ See: <http://corpus-tools.org/annis/>.

¹² See: <https://islandora.ca/>.

¹³ See: <https://www.drupal.org/>.

¹⁴ See: <https://corpora.uni-hamburg.de/hzsk/>.

can be applied before or after performing the query, and search results can be exported as a concordance view or in the form of full transcripts.

The GeWiss project uses Typo3 as the technical basis for the website, as well as MySQL/PHP for storing and querying data. It also offers REST-API web services¹⁵ for retrieving corpus data independently of the web interface. The GeWiss corpus contains a total of 1,273,529 tokens, which correspond to audio recordings of a total duration of 126 hours. The platform can be accessed online.¹⁶

2.4 CQPweb

CQPweb (Hardie, 2012) is a web-based corpus analysis system developed by Lancaster University. Like its predecessor, BNCweb,¹⁷ which was designed for searching and querying the original British National Corpus (BNC1994),¹⁸ CQPweb is based on the Corpus Query Processor (CQP) of the IMC Open Corpus Workbench (CWB)¹⁹ in combination with a MySQL relational database, but is corpus-independent and can provide access to multiple corpora at once. It is used to access a large collection of written and spoken corpora hosted at Lancaster University, among which is also the Spoken BNC2014 corpus.²⁰

The Spoken BNC2014 (Spoken BNC2014; Love *et al.*, 2017) was collected by the ESRC Centre for Corpus Approaches to Social Science (CASS)²¹ at Lancaster University between 2012 and 2016. It contains audio recordings from informal conversations collected on various locations across the United Kingdom. The corpus is particular in that the speakers have recorded their interactions themselves using their smartphones. With about 987 hours of speech and more than 11 million transcribed tokens, it is currently one of the largest resources of natural spoken language.

To gain access to the Spoken BNC2014, it is necessary to register an account at the CQPweb server. In addition, a user licence for the Spoken BNC2014 must be signed at the corpus homepage.²² After a successful account creation, the corpus can be explored²³ by querying the audio transcripts. The concordance queries can be formulated in the CQP Query Language (CQP QL)²⁴ or its simplified version—the Common

¹⁵ See: <https://gewiss.uni-leipzig.de/index.php?id=webservices>.

¹⁶ See: <https://gewiss.uni-leipzig.de/>.

¹⁷ See: <http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php>.

¹⁸ See: <http://www.natcorp.ox.ac.uk/>.

¹⁹ See: <http://cwb.sourceforge.net/>.

²⁰ In the rest of this paper, when we discuss CQPweb, we refer to the set of functionalities of the CQPweb application used for accessing Spoken BNC2014.

²¹ See: <http://cass.lancs.ac.uk/>.

²² See: <http://corpora.lancs.ac.uk/bnc2014/signup.php>.

²³ See: <http://corpora.lancs.ac.uk/bnc2014>.

²⁴ See: http://cwb.sourceforge.net/files/CQP_Tutorial.pdf.

Elementary Query Language (CEQL).²⁵ A query builder can be used to add metadata constraints to the current query. A variety of actions can be performed on the search results: users can sort, shuffle or limit the number of hits, view metadata distributions, calculate collocations, and download search results and the query string itself. It is also possible to create user-specific sub-corpora on the basis of metadata, or from transcripts containing at least one query hit. Beside queries, one can also create corpus or sub-corpus frequency lists and compare them with the frequency lists from other corpora within Lancaster's CQPweb instance. The source code of CQPweb is available online.²⁶ Metadata and full transcripts of the Spoken BNC2014 can be downloaded for offline browsing from the corpus homepage.

2.5 KonText

KonText (Machálek, 2017) is a multi-modal corpus analysis platform based on the open source system NoSketch Engine. It provides access to a large collection of written and spoken corpora from the Czech National Corpus (CNC; Křen, 2015), which was developed by the Faculty of Arts at the Charles University in Prague. It contains different types of corpora: written and spoken, synchronic/diachronic, standard/dialect, learner, translated, and parallel corpora with 4.25 billion tokens to date. Out of those, more than 7 million tokens are from spoken language. The CNC spoken corpora contain mainly data of informal spoken Czech from the regions of Bohemia, Moravia and Silesia, but have also data of, for instance, presidential speeches, language used in school settings, and a learner corpus of spontaneous English spoken by advanced speakers with L1 Czech. Some of these corpora are time-aligned and annotated with lemmas and other morphological tags.

KonText allows the user to browse metadata and perform concordance searches in anonymous mode (without registration). Queries can be expressed in SketchEngine Corpus Query Language (CQL),²⁷ a modified version of CQP QL. Alternatively, a query builder can be used for constructing simple queries for individual word forms, word parts, lemmas or sequences of tokens. Building metadata and context restrictions on the current query is also possible. Logging in with an own user account or through SSO unlock further KonText functionalities like sorting, filtering, evaluating and exporting concordance results, as well as storing recent query strings, generating lists of collocations and creating sub-corpora. The

²⁵ See: <http://cwb.sourceforge.net/ceql.php>.

²⁶ See: <http://cwb.sourceforge.net/developers.php#svn>.

²⁷ See: <https://www.sketchengine.eu/documentation/corpus-querying/>.

platform can be accessed online,²⁸ and the latest open-source version can be downloaded.²⁹

2.6 Corpus de LAngue Parlée en Interaction (CLAPI)

More than 600 hours of audio and video of real-life conversations in French were collected in several research projects since 1983 and archived by the Laboratoire ICAR at the Université de Lyon 2. These data are available to the scientific community through the Corpus de LAngue Parlée en Interaction (CLAPI) platform (Baldauf-Quilliatre *et al.*, 2016). Currently, sixty corpora containing more than 130 hours of speech are accessible for research purposes through the CLAPI platform.³⁰

CLAPI can be accessed as a guest (without registration), granting permission to 140 transcriptions, which correspond to about sixty-three hours of media time and about 580,000 transcribed tokens. The corpora contain conversations in various institutional and private contexts (e.g., conversation in the workplace, at home, at the doctor's office, in shops and phone calls) and can be browsed online for metadata and other additional materials (e.g., transcription conventions). For each corpus, an audio or video sample of twenty to forty seconds can be streamed or downloaded. Full-length media and transcript files can be downloaded in different formats for browsing offline, provided the rights allow it.

Beside browsing, CLAPI provides functionalities for collocation analysis and concordance search by means of querying transcribed tokens, stem-based forms, alternatively spelt forms, overlaps, repetitions and pauses. Search by partial matching, as well as metadata, distance, precedence and sensitivity constraints on the query string are also possible. Furthermore, the user can generate statistical information about the usage of individual tokens in relation to their position within a turn, to other oral phenomena, and to the available metadata. The technical base of the CLAPI platform is built with PHP/MySQL. It is also worth mentioning that the CLAPI-FLE site – a platform containing excerpts from CLAPI corpora intended for teaching French – can be found online.³¹

2.7 TalkBank

TalkBank (MacWhinney, 2007, 2017) is the world's largest online repository for spoken language data. It is maintained by Carnegie Mellon University in

²⁸ See: <https://www.korpus.cz/>.

²⁹ See: <https://github.com/czcorpus/kontext>.

³⁰ See: <http://clapi.icar.cnrs.fr/>.

³¹ See: <http://clapi.icar.cnrs.fr/FLE/>.

Pittsburgh and is available online.³² More than 300 corpora in over thirty-four languages included in TalkBank are collected by, and contributed to, by hundreds of researchers based on the principle of open data-sharing. The corpora are divided among fourteen databases associated with different research areas: CHILDES (child language research), PhonBank (phonological development), ClassBank (study of language in the classroom), AphasiaBank (study of communication in aphasia), *etc.* Most corpora are freely accessible and contain transcriptions which are time-aligned with the appropriate audio or video file. The recordings have been transcribed using the CHAT transcription format and can be analysed and queried online using CLAN³³ commands. All corpora can be browsed online up to the level of the individual transcripts and the media file associated with it, or they can be downloaded entirely.

2.8 OpenSoNaR

OpenSoNaR (van de Camp *et al.*, 2017) is a web-based application for the search and exploration of large corpora. It was developed in the CLARIN-NL and CLARIAH projects by a joint team of researchers at the Dutch Language Institute, Tilburg University, and Radboud University. OpenSoNaR provides online access to the Reference Corpus of written Dutch (SoNaR-500) and the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN). The CGN (Oostdijk, 2000) is a collection of about 900 hours of contemporary Dutch speech (adult speakers in the Netherlands and Flanders) recorded within the CGN-project (1998–2004).

After a successful login with a CLARIN account or through a SSO, the corpora can be investigated.³⁴ OpenSoNaR is based on the BlackLab search engine and supports the BlackLab's Corpus Query Language (CQL),³⁵ a modified version of CQP QL. It offers four different search modes with different possibilities to create a query (simple, extended, advanced and expert). There is also an extensive corpus metadata filter that can be used when browsing and searching corpora. The search history function, the possibility to share search results through a URL, and functionality for different kinds of corpus statistics (e.g., metadata distribution, *n*-grams, tokens and POS distributions) are also notable features. The open-source version of OpenSoNaR is available online.³⁶

³² See: <https://talkbank.org/>.

³³ See: <https://talkbank.org/manuals/CLAN.pdf>.

³⁴ See: https://portal.clarin.inl.nl/opensonar_frontend/opensonar/search.

³⁵ See: <https://github.com/INL/BlackLab/blob/master/core/src/site/markdown/corpus-query-language.md>.

³⁶ See: <https://github.com/INL/corpus-frontend>.

2.9 Others

Apart from the corpora and corpus platforms listed above, we also took into consideration Spokes³⁷ (Pęzik, 2015), Eslo³⁸ (Baude and Dugua, 2017), TGDA³⁹ (Boas, 2003), Corpuscle/COLA⁴⁰ (Meurer, 2012), GOS⁴¹ (Verdonik, *et al.*, 2013), MICASE⁴² (Simpson *et al.*, 2002), VOICE⁴³ (Breiteneder *et al.*, 2009), Glossa⁴⁴ (Nygaard *et al.*, 2008), TEITOK/MADISON⁴⁵ (Janssen, 2016), SpoCo⁴⁶ (Waldenfels and Woźniak, 2016), ArchiMob⁴⁷ (Samardžić *et al.*, 2016), Transdanubian Electronic Corpus⁴⁸ (Mladenova, 2006), OCQP⁴⁹ (Barbu Mititelu *et al.*, 2018), Bulgarian Dialectology as Living Tradition⁵⁰

³⁷ Spokes is a web-based corpus query and visualisation tool (see: <http://spokes.clarin-pl.eu/>).

³⁸ Enquêtes Sociolinguistiques à Orléans, Sociolinguistic Investigations in Orleans (see: <http://eslo.huma-num.fr/>).

³⁹ Texas German Dialect Archive (see: <https://tgdp.org/dialect-archive>).

⁴⁰ Corpuscle is a corpus management and analysis system for annotated corpora. It provides access to several types of corpora, among others to Corpus Oral de Lenguaje Adolescente Resource (COLA), a collection of recorded, spontaneous speech among teenagers from different schools and youth clubs in Madrid, Buenos Aires and Santiago de Chile. (see: <http://clarino.uib.no/korpuskel/page?page-id=korpuskel-main-page>).

⁴¹ GOS (GOvorjena Slovenščina, Spoken Slovene), see: <http://www.korpus-gos.net/>.

⁴² Michigan Corpus of Academic Spoken English (MICASE), see: <https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase>.

⁴³ Vienna–Oxford International Corpus of English is available through VOICE Online, at: <https://www.univie.ac.at/voice/page/index.php>.

⁴⁴ Glossa is a web-based multimodal corpus query system (see: <https://tekstlab.uio.no/glossa2/>).

⁴⁵ TEITOK is a web-based system for corpus creation, annotation, and distribution. It is used in different projects, including MADISON – A Sound Map for Portuguese Dialects (see: <http://cards-fly.clul.ul.pt/teitok/madison/>).

⁴⁶ SpoCo is a web-based system for querying spoken corpora with aligned audio files intended for dialect research and used in a network of Slavic dialect projects, (e.g., Ustja River Basin Corpus; see: <http://www.parasolcorpus.org/Pushkino/login.php>), Corpus of Russian spoken in Daghestan (see: <http://www.parasolcorpus.org/dagrus/>), Corpus of Spoken Rusyn (see: <http://parasolcorpus.org/Varchola1/login.php>).

⁴⁷ ArchiMob is a Corpus of Spoken Swiss German, accessible through ANNIS and SketchEngine instances at: <https://www.spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html>.

⁴⁸ Transdanubian Electronic Corpus is a corpus of Bulgarian Transdanubian dialects, accessible through map-based browser and corpus specific concordance, at: <http://www.corpusbdr.info/?goto=home>.

⁴⁹ Oral Corpus Query Platform (OCQP) provides access to the oral part of CoRoLa, a reference corpus of contemporary Romania accessible at: http://89.38.230.23/corola_sound_search/index.php.

⁵⁰ Bulgarian Dialectology as Living Tradition is a searchable and interactive database of oral speech representing a broad range of Bulgarian dialects available at <http://bulgariandialectology.org/>.

(Alexander, 2015), The Pangloss Collection (Michailovsky *et al.*, 2014)⁵¹ and RNC (Grishina, 2006, 2010).⁵²

We are aware that there are many more platforms and tools on the Internet for browsing and/or searching spoken language corpora. We decided to restrict ourselves to this list, because it already reflects the wide spectrum of possible approaches for spoken corpus distribution, and it is sufficient for our project goals.

3. Method of analysis

Our investigation was condensed thematically into the following categories: Access, Search, Presentation of Data and Search Results, and Browsing.

With the ‘Access’ category, we endeavoured to find the ways in which the platforms allow formal access to the data – that is, by requiring (or not) the user to register an account, providing (or not) the user group-specific interfaces (simple *versus* advanced search, teaching *versus* conversation research), and by allowing access to the data through graphical user-interface or WebServices, among others.

The ‘Search’ category deals with the realisation of a search functionality. The level of complexity and the degree of specificity are some of the aspects that we considered here. In this category, we collected observations on the features the platforms offer to the user for constructing and submitting their search queries, especially in terms of support for a query language (QL) and the possibility to search metadata and different types of annotations.

The ‘Browsing’ category encompasses the findings related to the ability to look through the data and meta-information of the corpora manually, traversing its structure step-by-step. We are interested in what corpus elements (corpus metadata, speech event and/or speaker metadata, audio/video files, transcripts, annotations and additional materials) can be browsed and in what form these elements can be viewed. Most significantly, we are interested in whether transcripts in full length are available, and whether the associated media files can be accessed for (continuous) playback.

The ‘Presentation of Data and Search Results’ category groups our findings on the presentation of results and various data statistics (e.g., frequency lists and metadata distributions). The focus is on the number of features available, such as switching between different graphic visualisations, filtering of results, download options, and similar aspects. It should be noted

⁵¹ The Pangloss Collection of LaCiTO (Langues et Civilisations à Tradition Orale, Languages and cultures of oral tradition) is an online archive of linguistic materials in endangered languages of the world. See: http://lacito.vjf.cnrs.fr/pangloss/index_en.html.

⁵² Russian National Corpus (RNC) provides access, among others, to Spoken Russian corpus (<http://ruscorpora.ru/new/en/search-spoken.html>) and Multimodal Russian corpus (<http://ruscorpora.ru/new/en/search-murco.html>).

Single-Sign-On (SSO)	User account required	User account not required
OpenSoNaR KonText HZSK* Glossa* Corpuscle*	DGD GeWiss TGDA KonText CQPweb VOICE	Spokes TalkBank GOS MICASE ArchiMob/Annis Spoken Russian Multimodal Russian Corpus The Pangloss Collection Bulgarian Dialectology as Living Tradition Transdanubian Electronic Corpus OCQP SpoCo TEITOK/MADISON CLAPI** Eslo**

Table 1: Overview of the authentication requirements of the analysed corpora. Some corpora on the platforms marked with an asterisk (*) are accessible without registration. The platforms marked with two asterisks (**) require a login in order to have access to complete data.

that several attempts to provide a fully systematic comparison in the form of a table failed because the heterogeneity of the platforms was too great when considering all criteria at once. For comparisons based on selected specific meaningful criteria, we provide tabular overviews (see Tables 1, 2 and 3).

4. Results and discussion

In this section, we present our observations on the various solutions for accessing corpora of spoken language. The chapter is organised into four sub-chapters, in which we elaborate on the designated aspects of research of spoken language corpora.

4.1 Access

Our analysis has shown that providers of spoken language corpora follow different ways of making their corpus data accessible to the user community. While some of them develop their own tools for searching and/or browsing (e.g., DGD, GeWiss and CLAPI), others use or build on existing open-source systems, like ANNIS (e.g., HZSK), NoSketch Engine (e.g., KonText) or BlackLab (e.g., OpenSoNaR), *etc.*

	SlopeQ	ANNIS QL	CQP QL	Modified versions of CQP syntax			
				TT-CQL ^a	Corpuscle QL ^b	CQL (SketchEngine)	CQL (BlackLab)
KonText						✓	
Spokes	✓						
CQPweb			✓				
OpenSoNaR							✓
Corpuscle					✓		
Glossa			✓				
TEITOK				✓			
SpoCo			✓				
ArchiMob corpus (accessible through ANNIS and SketchEngine)		✓				✓	

Table 2: Overview of corpus QLs (columns) supported by the platforms (rows) analysed in this study. (Key =^a see Janssen (2018), ^b see http://clarino.uib.no/korpuskel/page?page-id=the_query_language.)

In our review, we considered only those platforms which are available online and provide access to corpora of spoken language through a web-based user interface. The investigated user-interfaces are implemented in different languages and frameworks, where the most frequently used is a MySQL/PHP-combination. Some platforms make their API publicly available as web services, allowing users to access the data independently of the user interface (e.g., GeWiss, Spokes and Corpuscle).

A number of platforms provide access to multiple corpora (e.g., DGD and KonText), and some corpora are distributed online through multiple different platforms. For instance, the CLAPI corpus is accessible through its own platform, but some of its parts (metadata and transcripts) are also

Platforms and corpora websites with embedded browsing-functionality	Platforms and corpora websites without embedded browsing-functionality
Bulgarian Dialectology as Living Tradition CLAPI DGD Eslo GeWiss HZSK OpenSoNaR MICASE Spokes SpoCo TalkBank TEITOK/MADISON TGDA Transdanubian Electronic Corpus The Pangloss Collection VOICE	ArchiMob CQPweb GOS KonText OCQP RNC

Table 3: Overview of platforms with and without embedded browsing functionality. (The platforms Glossa and Corpuscle/COLA were excluded from this table, because it was unclear whether they had an embedded browsing functionality or not.)

found on the French Ortolang⁵³ archive platform. The CGN, likewise, can be accessed online through the OpenSoNaR platform, but it is also available as a treebank on the GrETEL⁵⁴ website. Theoretically, such types of corpus distribution are very practical, because corpus platforms with different functionalities provide the user with different views on the same data and allow them to approach their research question from different perspectives. In practice, however, it is often unclear whether the platforms provide access to exactly the same corpus, or whether users are dealing with different versions of the same data.

Some corpora analysed in our study are freely accessible (e.g., Spokes and GOS), while others require a login (see Table 1). Some corpora and corpus platforms allow login only after creating a user account (e.g., DGD), others only through an SSO (e.g., HZSK), and others yet have both options (e.g., KonText). The registration of a user account is handled

⁵³ See: <https://www.ortolang.fr/>.

⁵⁴ Greedy Extraction of Trees for Empirical Linguistics (GrETEL) is a search engine that supports example-based search and is developed by the Centre for Computational Linguistics at the University of Leuven. Available at <http://gretel.ccl.kuleuven.be/gretel3/>.

differently depending on the platform. Some corpus platform providers check the user's authenticity manually (e.g., DGD and GeWiss), while others generate an e-mail with a link, which the user needs to visit in order to confirm registration (e.g., KonText and VOICE Online). In some cases, users are required to register their platform account and also sign the end-user license for the corpus they want to investigate (e.g., BNC2014 accessible through CQPweb, and COLA accessible through Corpuscle). While this diversity in corpus access is understandable because it is explained by different corpus licence restrictions, it is currently one of the main problems the ZuMult-project aims to solve by creating a unified access to spoken language resources located in different repositories. Yet, we observed that more and more platforms are switching to SSO, which seems to herald a trend toward creating a more centralised point of access for spoken corpora. Platforms with free corpus access could also benefit from switching to SSO, as personal user accounts can be customised and thus personalised according to specific user needs.

If a login is required, the platform usually provides some information about the corpus beforehand, so that the user can be informed about the corpus content, size and formats before creating an account (e.g., DGD and GeWiss). There are also platforms that allow for browsing of transcripts and/or media files before making the user sign an agreement. One example of this is KonText, which has a demo version that allows the use of the web application with some limited functionalities, which are fully unlocked on login.

With regards to realisation of user-group differentiated corpus access, there are two approaches that stand out: the one revolving around platforms using a variety of input forms for submitting queries, thereby addressing users with differing competences (beginner *versus* advanced) in using query languages (e.g., Corpuscle); the other is concerned with ancillary applications, which offer specialised access to selected parts of the same corpora, and are designed with specific usages like language teaching and learning, grammatical, lexicographic, variational linguistics or text analysis in mind. These are, for example:

- Lexical Explorer,⁵⁵ for browsing and filtering quantitative data of selected corpora from DGD (Lemmenmeier-Batinić, 2020);
- Kwords,⁵⁶ Morphio⁵⁷ and SyD,⁵⁸ for extracting text keywords, analysing morphological relations and studying variants in the CNC; and,

⁵⁵ See: <https://www.owid.de/lexex/>.

⁵⁶ See: <https://kwords.korpus.cz>.

⁵⁷ See: <https://morfio.korpus.cz>.

⁵⁸ See: <https://syd.korpus.cz>.

- CLAPI-FLE,⁵⁹ a set of resources dedicated to French learning with Clapi Corpora.

4.2 Search

In the domain of corpus linguistics, where one usually works with hundreds of hours of audio/video data and millions of corresponding transcribed words, being able to search quickly and efficiently is one of the key requirements. We analyse this type of functionality of program in terms of the amount and variety of features it presents in its interface.

We found that all corpus platforms, apart from the current version⁶⁰ of TGDA, either have an embedded search function (e.g., DGD and GeWiss) or provide external search software as well as audio transcripts in an appropriate format, which can be downloaded and queried offline (e.g., HZSK). Of the platforms that offer access to more than one corpus, only a few allow simultaneous searches in more than one of those corpora (e.g., DGD and OpenSoNaR). The platforms realise this functionality in different ways. In DGD, for instance, if more than one corpus is selected, only the metadata common to all corpora are available for the search. By contrast, OpenSoNaR makes the metadata of all corpora available to search, even if some of those metadata are not applicable to the selected corpus.

A characteristic common to most platforms with a search engine is the possibility to perform an advanced search, no matter how varying in degree of complexity or usefulness. This generally means being able to apply some filters from the start, rather than doing it successively on the results screen, and/or being able to submit not only simple queries, but also more complex search queries, which can be formulated using specifically designed features. In the platforms we analysed, the advanced options are usually not displayed to the user directly after login. The user is supposed to find them intuitively after concluding that the simple search mode does not fulfil their research purposes. Some platforms explicitly allude to the possibility of gradually increasing the complexity and the depth of the corpus investigation by providing multiple search modes. These allow the user to become familiar with search options and filters step-by-step, and thus come closer to fulfilling their research purposes. This is exemplified by the OpenSoNaR's user interface, which has four different search modes: 'Simple', used for searching for specific words and phrases case-insensitively; 'Extended', used for combining the single-token-search with lemma, POS and other annotations; 'Advanced', which has a query builder for creating search expressions; and 'Expert', used for querying corpora by typing in CQL expressions directly (see Figure 1).

⁵⁹ See: <http://clapi.ish-lyon.cnrs.fr/FLE/accueil.php>.

⁶⁰ The metadata search functionality was offered in the previous version of TGDA (see Merkel and Schmidt, 2009).

The screenshot displays the OpenSoNaR corpus search interface. At the top, a navigation bar includes the OpenSoNaR logo and links for 'OpenSoNaR corpus', 'INT', 'CLARIN', 'CLARIAH', 'Help', and 'About'. Below this, a red bar contains 'Search' and 'Explore' tabs. The main section is titled 'Search for ...' and features four search modes: 'Simple' (selected), 'Extended', 'Advanced', and 'Expert'. Under 'Simple', there are buttons for 'Basics' and 'More annotations'. The search criteria section includes:

- Word:** A text input field with 'Word' as a placeholder and a download icon. A checkbox for 'Case and diacritics sensitive' is below it.
- Lemma:** A text input field with 'Lemma' as a placeholder and a download icon. A checkbox for 'Case and diacritics sensitive' is below it.
- Part of Speech:** A dropdown menu with 'Part of Speech' as the selected option and an edit icon.
- Within:** Radio buttons for 'document', 'paragraph', and 'sentence'.
- A checkbox for 'Split batch queries' at the bottom.

Figure 1: ‘Simple’, ‘Extended’, ‘Advanced’ and ‘Expert’ search modes on OpenSoNaR.

Given that corpora of spoken language are, in general, richly annotated with metadata, being able to use speaker and speech event metadata as filters to restrict the search within transcripts is highly desirable. Most of the corpus platforms we analysed were able to use speaker and speech event metadata as filters (e.g., DGD, GeWiss and KonText), although to varying degrees of complexity, depending on the amount of Boolean operations (e.g., AND, OR or NOT) one can use to combine different metadata restrictions. The amount of metadata also varies and is dependent on the corpus.

Some corpus platforms provide a context filter for specifying words appearing in the context surrounding the searched-for token. The realisation of the context filter varies greatly from platform to platform. In the DGD, one can specify only one token at a time (its transcribed and/or normalised form, lemma and POS) occurring to the right or the left of the desired word. One can also specify the maximum distance between words, and whether they should occur in the same utterance or in a sequence of utterances of the same speaker. CLAPI allows users to specify three words and/or one spoken language phenomenon (pause, overlap) that occur in the vicinity of the target, the maximum distance between the first and the last element in the query, as well as whether they should appear in the same utterance or in the sequence of a certain number of successive utterances. It is also possible to specify the minimum and maximum length of the utterance and the position of the search word in the utterance. GOS’s context filter is special in that it allows

users to specify words (a maximum of three) that should occur or *not* occur in the context of the searched word. It is also possible to specify the maximum distance between words, but also the exact position of the words to the left and/or to the right of the searched word.

We observed that almost all of the platforms with a search functionality can formulate queries that are more sophisticated than just spelling out the single word that is to be searched in the corpora. All of these corpus platforms support a certain query syntax, which, at the very least, allows the use of wildcards in the query (e.g., MICASE and VOICE). Regular expressions (RegEx) are quite common, and automatically contribute a great number of options, even in the case of simple queries. In order to make it easier for the user to work with RegEx, some corpus platforms provide a query builder tool containing a list of descriptive alternatives to RegEx, such as ‘begins with’, ‘ends with’, ‘contains’, *etc.* (e.g., CLAPI). Query builders are generally very popular regardless of whether the platform provides support for the corpus’s query language (QL) or not.

Only a few of the platforms we analysed support a QL that was specifically developed for corpus linguistic queries. Table 2 gives an overview of these platforms and the QLs they support. The QLs, as well as search engines behind the platforms we analysed were developed originally to query written corpora. Most of these search engines use the CQP QL syntax or a modified version of it.

The effectiveness of the use of corpus QLs depends on the richness of the annotations. The better the corpus is annotated, the more powerful and more significant does the query design become. However, learning a whole new syntax (or a set of syntaxes) may seem intimidating at first, and queries formulated in a special corpus QL can become quite lengthy and convoluted, which limits their usefulness.

One of the solutions could be an interactive query builder (such as the one implemented in KonText); users can generate a query by ticking checkboxes or selecting an option from different dropdown menus. As they are making their choices, the corresponding CQL query string is assembled in the search box. This method shows the user the various options they can use, and it passively teaches them the CQL syntax building.

Not all platforms can search within transcripts. The Pangloss Collection of LaCiTO can only search in metadata assigned to the transcript and not the transcribed file itself. This functionality is used as a filter for browsing transcripts. The user can specify values of some metadata (e.g., ‘Title’, ‘Language(s)’, *etc.*), or formulate a character sequence to be found in those metadata fields. After the search is performed, the user can examine the full transcript document, its metadata and the associated recordings.

The key-value-based metadata search is supported by many platforms, either as a filter for browsing transcripts (e.g., CLAPI and TalkBank) or as a filter for displaying a list of speech events/transcripts, which can then be saved as a virtual corpus. The virtual corpus can

then be opened and worked on at a later point in time (e.g., CQPweb and DGD).

DGD stands out when it comes to metadata search, as it is the only platform that provides a full-text search using metadata. The metadata files of DGD corpora are stored in an Oracle Database and can be queried using Oracle QL, whose powerful syntax (including partial matching, Boolean operators, Fuzzy-Search, operators for specifying distance and number of occurrences of the desired items, *etc.*) allows for very complex search queries within metadata.

4.3 Presentation of data and search results

In the majority of cases, we found that the results and data landing pages were the most feature-rich parts of the platforms. This was in part expected, as making sense of the results is generally one of the final and most important steps in the interaction with the platform. Being able to refine the results to one's own needs can help save time and streamline one's research.

Every platform that has a search engine displays excerpts of the transcripts in their results page. This is often realised with the KWIC (Keyword in Context) layout, whereby the queried keyword is highlighted and positioned prominently in the centre of the excerpt that is displayed. This allows the user to quickly gain a sense of context in which the keyword is used. The context surrounding the queried keyword can be increased or decreased in many platforms (e.g., DGD, GeWiss and KonText), although in some cases the context size must be set as a search parameter. The platforms vary in the maximum length to which the context of individual hits can be extended. Only platforms with a browsing functionality have the option to extend the context to the full transcript.

It appears that the KWIC layout has become a sort of a standard, as it is employed in most of our platforms. Another popular way of displaying results is that of the contribution layout (see Figure 2). It is sometimes offered as the more detailed primary view (e.g., in case of CLAPI), but more often as an alternative to the KWIC view.

Some platforms present the search results as an utterance view, where each concordance hit contains the whole utterance in which the queried linguistic phenomenon occurs (e.g., sentence view in KonText). Somewhat unexpectedly, the score view (Figure 3) is not very popular among the platforms we analysed (e.g., DGD, GeWiss and the "Extended Transcript" mode in CLAPI). It might be of some significance to note that the score layout has the advantage of visualising temporal relations, such as speaker overlaps, which are essential in distinguishing spoken corpora from written texts.

Almost all of the platforms that we analysed provide buttons for opening the metadata for the appropriate hit in the concordance line. Some platforms can display values of user-specified metadata keys directly in the

Figure 2: Contribution layout on DGD.

Figure 3: Score view on DGD.

KWIC view (e.g., DGD und GeWiss). This functionality allows for a direct comparison of individual hits.

Display of annotations for individual hits differs greatly from platform to platform (if they are displayed at all). Available annotations are generally presented in the form of a table containing token-based annotations (normalised form, lemma, POS and other morphological characteristics) for every token contained in the hit excerpt. In many cases, token-based annotations are presented in a tooltip, which appears when the user hovers the mouse over a token (e.g., CQPweb) or clicks on it (e.g., RNC). Alternatively, the annotations can be displayed directly in the text. Some platforms allow the user to specify how the annotations should be displayed and whether they should be displayed at all (e.g., Corpuscle and KonText).

Interestingly, not all platforms provide access to the original audio or video files in their results section. BNC2014⁶¹ and MICASE, for instance, only present the transcribed text. This has far-reaching consequences, because being able to analyse the audio/video file along with the transcribed text personally is a rather important aspect for research on spoken language. While transcript-only results may satisfy quantitative linguistic analyses, a combination of transcript and audio (and even more so video) facilitates and encourages more qualitative studies from the areas of phonology and interactional linguistics, for instance. Having access to the original audio and video data means having access also to the paralinguistic features of the recorded speech, as well as the precise context, in which it took place. A transcript aligned with the audio or video file thus becomes a very valuable representation and access point to the primary data itself.

With regard to further processing of search results, the spoken language platforms analysed in this study are very heterogeneous in terms of the features they offer. Only half of the platforms have the ability to randomise results. Some of them can shuffle hits directly in the KWIC view (e.g., CQPweb), while others randomise hits on export (e.g., GOS). Limiting the number of results by applying metadata seems to be a more common practice than limiting result as a percentage of all hits or by specifying the desired number of randomised samples (e.g., OpenSoNaR), or by using a context filter (e.g., DGD and KonText). The latter functionalities are implemented very rarely. The function for inverting the filter (like the one used in DGD and KonText) – that is, for removing the hits that match the condition – is not very widely implemented. None of the platforms allow further searching in search results. That means that they cannot perform further queries within the scope of the texts containing hits found with the initial search request.

Most platforms we looked at were able to sort the results, either by corpus order, by match, or by the specified position of the token to the left or right from the match (to a maximum of five positions). Worthy of mention

⁶¹ BNC2014 plans to make the audios available in the future. See: <http://corpora.lancs.ac.uk/bnc2014/faq.php>.

in this regard are the platforms that could sort according to more than one criterion (e.g., MICASE and OpenSoNaR).

Platforms differ in scope and the means they use to provide information about the quantified distribution of corpus metadata. Such information can be found directly on some online platforms (e.g., GOS) or are sometimes 'hidden' in additional documentation (e.g., GeWiss).

There is a lot of diversity in displaying statistics generated with the results, which generally show figures specific to the outcome of the query. These are generally displayed either during (e.g., CLAPI) or after the searching process is completed (e.g., DGD). The statistical overview is sometimes combined with the filter functionality, which can be used to narrow down the results (e.g., GOS). Some platforms allow the user to specify a metadata dimension or a category of annotations for which they would like to create statistical distribution of results (e.g., KonText and CQPweb). The statistics can also be expressed in charts, graphs and other forms of visualisations, as in the case of Spokes and OpenSoNaR. In addition to the concordance lines, some platforms (e.g., Glossa) use maps for visualising geographical locations associated with the hits. This provides a totally different perspective on the corpus data, which is more visually stimulating and easier to understand than long blocks and tables of raw data.

Almost all platforms make it possible to print or save the KWIC results on the local machine (mostly in .xlsx, .txt, .csv and .xml formats). The platforms that require a login often provide the possibility to save the query string and the KWIC results on a personalised space on the platform's servers (e.g., CQPweb). A few platforms can save individual examples of different search requests in user-defined collections (DGD) or in a bookmarks archive (VOICE). Downloading individual hits is not a very widespread function among the platforms considered for this study, and only DGD allows the user to specify the desired length of the transcript excerpt to be downloaded along with the synchronised audio file and the appropriate metadata.

Platforms developed especially for performing statistical corpus analysis provide more functionalities for working with the results, in contrast to those which focus primarily on the presentation of special language varieties and the cultural and linguistic background of the language community on the Internet (e.g., Transdanubian Electronic Corpus). We also noticed that feature richness of a platform often goes along with reduced usability. For instance, TalkBank uses CLAN commands to filter and perform calculations on the results. This is a very powerful feature, but it requires users to learn the ropes of a specific program before it can be used correctly.

While most platforms provide frequency lists of word forms and lemmas occurring in a given corpus (e.g., DGD), there are platforms that can also generate word frequency lists at the level of individual transcripts (e.g., CLAPI). CQPweb and OpenSoNaR go even further by allowing the user to create frequency lists for individual speakers and sub-corpora. In addition, CQPweb can compare frequency lists of different corpora and sub-corpora

side-by-side. OpenSoNaR, on the other hand, can also create frequency lists and n -grams for POS and phonetic transcriptions.

We noticed that only a few platforms have functionality to perform co-occurrence analyses of spoken language transcripts (e.g., KonText, CQPweb and CLAPI). DGD cannot analyse co-occurrences itself, but the Lexical Explorer (see above) is developed as a tool for this particular purpose.

4.4 Browsing

Browsing could be seen as a non-automated form of searching. Instead of relying on computational power in order to retrieve relevant data, the user is enabled to leaf through the data of the corpora. This allows the user to explore and discover the contents without having to define a specific search pattern. This exploratory aspect encourages users to pay attention to the way the corpora are structured, and it allows them to understand the relationships between the different parts of the corpora, and make sense of them on a more global level.

Most platforms from our analysis provide an embedded browsing functionality (see Table 3), but they implement it in very different ways. CLAPI and DGD stand out especially. CLAPI differs from the majority of platforms with regard to the composition of its browsing functionality. Individual transcript files cannot be viewed directly in the web-browser but have to be downloaded one-by-one in order to be browsed. DGD has a very complex browsing structure. In contrast to the rest of the platforms analysed here, whose corpora consist of transcripts and respective media files (a one-to-one relationship), DGD consists of events that are made up of smaller speech events, which, in turn, consist of one or more transcripts that do not necessarily correspond to the same number of media files.

Most platforms provide a list of corpora (e.g., DGD, CLAPI, OpenSoNaR) or a list of sub-corpora (e.g., GeWiss and VOICE Online) as the starting point for the browsing. In case only one corpus is available, the speech event list is used instead as the starting point for navigating through the corpus (e.g., Spokes and MICASE). It is also worth mentioning that three platforms from our selection (DGD, GeWiss and Spokes) provide lists of speakers for individual corpora/sub-corpora, thus allowing the user to start investigating the corpus by browsing speaker metadata independently of the speech events.

A special mention is reserved for dialect corpora, since they generally offer a browsable geographical map as the entry point into the corpus. Examples of such platforms include the Bulgarian Dialectology as Living Tradition, TGDA, the Transdanubian Electronic Corpus, and The Pangloss Collection.

In platforms that allow browsing, we could see some differences concerning the form of transcripts, the audio playback possibilities, and the

amount of metadata provided for speech events. Firstly, not all transcripts are time-aligned with the corresponding media file. Secondly, time-aligned audio transcripts are generally presented in two ways: in the score view, where a transcript is displayed horizontally and is organised into tiers with time-based annotations (e.g., GeWiss), and in the contribution view, containing a vertical list of speaker contributions (e.g., Spokes and TGDA). The view depends on the pre-processing of the corpus. DGD's FOLK corpus, for example, is annotated in such a way that it allows the user to choose between as many as three different transcript views: the score view, the contribution list view, and the normalised contribution list view.

Thirdly, some platforms are quite frugal in providing speech event metadata, while others offer extensive information, including even metadata for every individual speaker participant in the recording. On the lower side of the metadata availability spectrum is TGDA, which makes available only the ID, the name of the recording, and the location of the recording. On the other side of the spectrum is the DGD, which provides a plethora of detailed information on speakers and speech events. The DGD even distinguishes between a 'speaker' and a 'speaker in the speech event'. The former contains properties related to the speaker in general terms (e.g., sex, mother tongue and birth location), while the latter details the specific properties of the speaker at the time of the recording (e.g., their age, role in the interaction and their current language situation).

Finally, the audio playback function, if provided at all, differs from platform to platform. In the case of a continuous media playback (e.g., Voice-Online), one is presented with an embedded media player that the user can interact with in order to select the time point in the media in which they are interested. Other platforms only play excerpts limited to a few seconds or segmented utterances (e.g., Spokes). Platforms like DGD, GeWiss and HZSK, for example, offer both possibilities (continuous and segment-based playback), allowing for a more flexible user experience. Depending on the purpose of the analysis, one or the other of the options can be preferable. Some platforms stand out in particular in their realisation of this function by marking the places in the transcript that are currently being played (e.g., DGD). Other platforms include a waveform and spectrogram view, which is an essential feature for analysing the corpus phonetically or phonologically (e.g., TEITOK).

The platforms and corpora that do not provide an embedded browsing functionality often allow users to download XML transcripts in full length from the corpus website (BNC2014) or personally request media-files from the corpus provider directly, and thus have the possibility to browse the materials offline (e.g., ArchiMob).

For the most part, it appears that the platforms put more effort into one aspect of the search-browsing dichotomy, and proportionally less into the other. Which option of this 'dichotomy' a platform chooses is often a consequence of whether it was originally developed to query written corpora or whether it was designed from the outset for working with oral data. Most

platforms, which were originally designed for spoken language research (e.g., DGD, CLAPI and TalkBank), offer a more comprehensive and feature-rich browsing interface than others, which were developed not only with spoken language in mind, but also (and perhaps primarily) for querying written corpora.

5. Conclusion

In this paper we presented a state-of-the-art review of some corpus platforms designed to be used for working with spoken language data. Here, we summarise our main conclusions.

Although these platforms were developed in different projects for different purposes and, most importantly, for dealing with different data, they share a lot of common functionalities and similar realisations that could be used as the basis for the development of a cross-platform architecture. There is little evidence to suggest, however, that any two platforms share the same architecture. In fact, the current landscape of spoken language platforms is remarkably heterogeneous, and it is therefore nearly impossible to make a fully systematic comparison between the platforms. This was one of the conclusions of Merkel and Schmidt (2009), and it remains valid ten years after its publication, as this paper confirms.

We expected to be able to browse the data on all the platforms we analysed, as it seems to be a more fundamental and basic functionality when working with spoken language than having a dedicated search engine. However, we observed that not all of our selected platforms offered an option to browse the contents of their corpora freely, and even when they did, it was not always thoroughly implemented.

We noticed that there are a lot of different open-source search engines, and that many platforms use them for querying spoken language, although they were initially developed with written language in mind. CQP syntax seems to have become a *de facto* standard, as most of the search engines use it.

Furthermore, data is presented and visualised in a way that is understandable to the corpus linguist, who is already familiar with the data formats and the terms used. However, users who do not have a solid background in corpus linguistics may find some of the corpus platforms difficult to approach, as they have a steep learning curve. More effort could be directed toward testing the usability of the platform, and making it more open to inter-disciplinary work.

Concerning the realisation of corpus access for individual user groups, we could observe that while simple/expert modes are extensively implemented, the access designed around specific groups of researchers is not common. There are a few attempts to address the needs of language teachers, lexicographers, grammarians and variational linguists,

but the needs of discourse and conversation analysis researchers remain unaddressed, and are therefore still open to future solutions.

Generally speaking, the current review sheds light on the great variety of practical solutions available for working with corpora of spoken language. We presented the various solutions the platforms use to address the needs of the researchers of spoken language, and we believe that this review will be useful to our colleagues who are currently starting to develop a new corpus platform, or who are just looking for the appropriate software for making their spoken language corpora available online.

References

- Alexander, R. 2015. 'Bulgarian dialectology as living tradition: a digital resource of dialectal speech', *Balkanistica* 28, pp. 1–13.
- Barbu Mititelu, V., D. Tufiş and E. Irimia. 2018. 'The reference corpus of the Contemporary Romanian Language (CoRoLa)' in N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga (eds) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1178–85. Paris, France. European Language Resources Association (ELRA).
- Baldauf-Quilliatre, H., I. Colon de Carvajal, C. Etienne, E. Jouin-Chardon, S. Teston-Bonnard and V. Traverso. 2016. 'CLAPI, une base de données multimodale pour la parole en interaction: apports et dilemmes', *Corpus [Online]* 15. Online since 15 January 2017. Accessed 25 October 2019 at: <http://journals.openedition.org/corpus/2991>.
- Batinić, J., E. Frick, J. Gasch and T. Schmidt. 2019. 'Eine Basis-Architektur für den Zugriff auf multimodale Korpora gesprochener Sprache' ['A basic architecture for accessing multimodal corpora of spoken language'] in P. Sahle (ed.) *Book of Abstracts of the Sixth Annual Conference on Digital Humanities in the German-speaking countries (DHd 2019)*, pp. 280–1. Frankfurt.
- Baude, O. and C. Dugua. 2017. 'Les ESLO, du portrait sonore au paysage digital', *Corpus [Online]* 15. Online since 15 January 2017. Accessed 25 October 2019 at: <http://journals.openedition.org/corpus/2924>.
- Boas, H.C. 2003. 'Tracing dialect Death: the Texas German dialect project' in J. Larson and M. Paster (eds) *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, pp. 387–98.
- Breiteneder, A., T. Klimpfinger, S. Majewski and M.-L. Pitzl. 2009. 'The Vienna–Oxford International Corpus of English (VOICE): a linguistic resource for exploring English as a lingua franca', *ÖGAI Journal* 28 (1), pp. 21–6.

- Fandrych, C., E. Frick, H. Hedeland, A. Iliash, D. Jettka, C. Meißner, T. Schmidt, F. Wallner, K. Weigert and S. Westpfahl. 2016. 'User, who art thou? User profiling for oral corpus platforms' in N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis (eds) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 280–7. Paris. European Language Resources Association (ELRA).
- Grishina, E. 2006. 'Spoken Russian in Russian National Corpus (RNC)' in N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk and D. Tapias (eds) *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 121–4. Paris: European Language Resources Association (ELRA).
- Grishina E. 2010. 'Multimodal Russian Corpus (MURCO): first steps' in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 2953–60. Paris. European Language Resources Association (ELRA).
- Hardie, A. 2012. 'CQPweb – combining power, flexibility and usability in a corpus analysis tool', *International Journal of Corpus Linguistics* 17 (3), pp. 380–409.
- Janssen, M. 2016. 'TEITOK: text-faithful annotated corpora' in N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 4037–43. Paris: European Language Resources Association (ELRA).
- Janssen, M. 2018. 'TEITOK as a tool for dependency grammar', *Procesamiento del Lenguaje Natural* 61, pp. 185–8.
- Jettka, D. and D. Stein. 2014. 'The HZSK repository: implementation, features, and use cases of a repository for spoken language corpora', *D-Lib Magazine* 20 (9/10).
- Křen, M. 2015. 'Recent developments in the Czech National Corpus' in P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lungen and A. Witt (eds) *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, pp. 1–4. Mannheim: Institut für Deutsche Sprache.
- Lemmenmeier-Batinić, D. 2020. 'Lexical Explorer: extending access to the database for spoken German for user-specific purposes', *Corpora* 15 (1), pp. 55–76.
- Love, R., C. Dembry, A. Hardie, V. Brezina and T. McEnery. 2017. 'The spoken BNC2014: designing and building a spoken corpus of everyday

- conversations', [International Journal of Corpus Linguistics](#) 22 (3), pp. 319–44.
- Machálek, T. 2017. 'KonText—a modern, customizable corpus query interface' in Abstracts booklet for Corpus Linguistics International Conference, pp. 681–3. University of Birmingham.
- MacWhinney, B. 2007. 'The TalkBank Project' in J.C. Beal, K.P. Corrigan and H.L. Moisl (eds) *Creating and Digitizing Language Corpora: Synchronic Databases*, pp. 163–80. (Volume 1.) Houndmills: Palgrave-Macmillan.
- MacWhinney, B. 2017. 'TalkBank and CLARIN', *Selected papers from the CLARIN Annual Conference 2016. Linköping Electronic Conference Proceedings* 136, pp. 76–89.
- Meurer, P. 2012. 'Corpuscle—a new corpus management platform for annotated corpora' in A. Gisle (ed.) *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*, pp. 29–50. Amsterdam: John Benjamins.
- Michailovsky, B., M. Mazaudon, A. Michaud, S. Guillaume, A. François and E. Adamou. 2014. 'Documenting and researching endangered languages: the Pangloss collection', *Language Documentation and Conservation* 8, pp. 119–35.
- Mladenova, O. 2006. 'Ob elektronnom izdanii korpusa bolgarskich dialektnykh textov iz rumynii' (Об электронном издании корпуса болгарских диалектных текстов из Румынии) ['Towards an Electronic Corpus of the Bulgarian Dialects in Romania'] in A.N. Sobolev (А.Н. Соболев) and A.J. Rusakov (А.Ю. Русаков) (eds) *Jazyki i dialekty malych ètničeskikh grupp na Balkanach* (Языки и диалекты малых этнических групп на Балканах) ['Languages and Dialects of Small Ethnic Groups in the Balkans'], pp. 98–109. Sankt-Petersburg, München: Verlag Biblion.
- Nygaard, L., J. Priestley, A. Nøklestad and J.B. Johannessen. 2008. 'Glossa: a multilingual, multimodal, configurable user interface' in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odiijk, S. Piperidis and D. Tapias (eds) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 617–22. Paris: European Language Resources Association (ELRA).
- Oostdijk, N. 2000. 'The Spoken Dutch Corpus: outline and first evaluation' in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pp. 887–94. Paris: European Language Resources Association (ELRA).
- Pezik, P. 2015. 'Spokes—A Search and Exploration Service for Conversational Corpus Data' in *Selected Papers from the CLARIN 2014 Conference*, pp. 99–109. Linköping University Electronic Press.

- Samardžić, T., Y. Scherrer and E. Glaser. 2016. 'ArchiMob—a corpus of spoken Swiss German' in N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis (eds) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 4061–6. Paris: European Language Resources Association (ELRA).
- Schmidt, T. 2017. 'DGD—Die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim' ['DGD—Database for Spoken German. Spoken corpora at the Institute for German Language in Mannheim'], *Zeitschrift für Germanistische Linguistik* 45 (3), pp. 451–63.
- Schmidt, T. 2018. 'Gesprächskorpora' ['Conversation Corpora'] in M. Kupietz and T. Schmidt (eds) *Korpuslinguistik. Germanistische Sprachwissenschaft um 2020* ['Corpus Linguistics: Germanic Linguistics around 2020'], pp. 209–30. (Volume 5.) Berlin and Boston: de Gruyter.
- Simpson, R.C., S.L. Briggs, J. Ovens and J.M. Swales. 2002. 'The Michigan corpus of academic spoken English', Ann Arbor, Michigan: The Regents of the University of Michigan. Accessed 25 October 2019 at: <https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple>.
- Slavcheva, A. and C. Meißner. 2014. 'Building and maintaining the GeWiss corpus—perspectives on the construction, sustainability and further enrichment of spoken corpora: a showcase' in Š. Ruhi, M. Haugh, T. Schmidt and K. Wörner (eds) *Best Practices for Spoken Language Corpora in Linguistic Research*, pp. 20–35. Newcastle: Cambridge Scholars Publishing.
- Waldenfels, R. von and M. Woźniak. 2016. 'SpoCo—a simple and adaptable web interface for dialect corpora', *Journal for Language Technology and Computational Linguistics* 31, pp. 155–70.
- van de Camp, M., M. Reynaert and N. Oostdijk. 2017. 'WhiteLab 2.0: a web interface for corpus exploitation' in J. Odijk and A. van Hessen (eds) *CLARIN in the Low Countries*, pp. 231–43. London: Ubiquity Press.
- Verdonik, D., I. Kosem, A. Zwitter Vitez, S. Krek and M. Stabej. 2013. 'Compilation, transcription and usage of a reference speech corpus: the case of the Slovene Corpus GOS', *Language Resources and Evaluation* 47 (4), pp. 1031–48.