Proceedings of he 8th Workshop on Challenges in the Management of Large Corpora, pages 40–46
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Using Full Text Indices for Querying Spoken Language Data

Elena Frick, Thomas Schmidt
Leibniz-Institute for the German Language
R5, 6-13D-68161 Mannheim, Germany
{frick, thomas.schmidt}@ids-mannheim.de

Abstract

As a part of the ZuMult-project, we are currently modelling a backend architecture that should provide query access to corpora from the Archive of Spoken German (AGD) at the Leibniz-Institute for the German Language (IDS). We are exploring how to reuse existing search engine frameworks providing full text indices and allowing to query corpora by one of the corpus query languages (QLs) established and actively used in the corpus research community. For this purpose, we tested MTAS - an open source Lucene-based search engine for querying on text with multilevel annotations. We applied MTAS on three oral corpora stored in the TEI-based ISO standard for transcriptions of spoken language (ISO 24624:2016). These corpora differ from the corpus data that MTAS was developed for, because they include interactions with two and more speakers and are enriched, inter alia, with timeline-based annotations. In this contribution, we report our test results and address issues that arise when search frameworks originally developed for querying written corpora are being transferred into the field of spoken language.

Keywords: MTAS, spoken language data, oral corpora, TEI, query

1. Introduction

When talking about large corpora, one would think automatically of text corpora in the size of billions of tokens. In the context of spoken language, however, corpora with only over one million tokens already qualify for this group. The reasons why written and spoken corpora are looked upon from different perspectives regarding the size are foremost the costs of transcribing the audio/visual material. Additionally, there are difficulties in terms of field access and data protection for collecting authentic and spontaneous interaction data – even more so when various interaction types required for representative language research need to be covered (see Kupietz and Schmidt (2015)).

Even if today the need for search engine optimization (to retrieve huge amounts of big data within a reasonable time) is not a paramount concern in the development of spoken language platforms, there are good reasons to address the issue: The question is whether and how the efficient solutions developed to handle large written corpora can be applied for indexing and querying spoken language transcripts in order to provide uniform ways for accessing written and spoken language data. Could high-performance frameworks be adopted to spoken language without complex modifications? Or would it be necessary to rethink the basic concepts and reimplement the whole software from scratch to suit the special features of spoken language?

Our review of the state of the art of corpus platforms shows that some search engines (e.g. ANNIS[1], Sketch Engine[2], CQPWeb[3], BlackLab[4]), developed for querying written corpora, are already actively applied as search environments on multimodal spoken language corpora (see e.g. Spoken BNC2014[5], Spoken Dutch Corpus[6] and ArchiMob corpus[7]). Unfortunately, no publications could be found that discuss the difficulties that arise when search frameworks originally developed for querying written corpora are being transferred into the field of spoken language.

MTAS[8] (Multi-Tier Annotation Search) developed by the KNAW Meertens Institute[9] in Amsterdam is another open source search engine for querying on text with multilevel annotations. As a part of the ZuMult-project[10], we are currently testing this technology for indexing and querying corpora from the Archive of Spoken German[11] (Archiv für Gesprochenes Deutsch, AGD, Stift and Schmidt, 2014) at the Leibniz-Institute for the German Language[12] (IDS). In this contribution, we are sharing our experience in applying MTAS on three corpora stored in the TEI-based ISO standard for transcriptions of spoken language (ISO 24624:2016) and enriched with different kinds of annotations, especially timeline-based annotations.

In what follows, we first give a short description of our project (Section 2) and then present MTAS - the search engine framework that is in the focus of the present study (Section 3). In the remaining sections, we describe our test data (Section 4), evaluation method (Section 5) and results (Section 6), and discuss some challenging aspects involved in creating and searching indexes of *spoken* language corpora. Section 7 includes the conclusions of our research and provides an outlook on possible future developments.

2. Background

ZuMult (Zugänge zu multimodalen Korpora gesprochener Sprache, Access to Multimodal Spoken Language Corpora) is a cooperation project between three research institutes: the AGD in Mannheim, the Hamburg Centre for Language Corpora (Hamburger Zentrum für Sprachkorpora, HZSK) and the Herder-Institute at the University of Leipzig. This

[1] https://corpus-tools.org/annis
[2] https://www.sketchengine.eu
[3] https://corpora.linguistik.uni-erlangen.de/cqpweb
[4] https://inl.github.io/BlackLab
[5] http://corpora.lancs.ac.uk/bnc2014/
[6] https://www.clariah.nl/en/new/news/search-written-and-spoken-dutch-with-opensonar
[7] https://www.spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html
[8] https://textexploration.github.io/mtas/
[9] https://www.meertens.knaw.nl/cms/en/
[10] https://zumult.org/
[11] http://agd.ids-mannheim.de/index.shtml
[12] https://www1.ids-mannheim.de/

project started in 2018 with a twofold purpose: On the one hand, a software architecture for a unified access to spoken language resources located in different repositories should be developed. On the other hand, user-group specific web-based services (e.g. for language teaching research or for discourse and conversation analysis) should be designed and implemented based on this architecture. The concept involves two parallel modules: 1) Object-oriented modeling of spoken language corpus components (audio- and video data, speech event and speaker metadata, transcripts, annotations and additional materials) and their relationships; 2) Providing the search functionality that is fully compatible with typical characteristics of spoken language. While the first module is primarily intended for explorative browsing on the data, the second query module should enable a quick and targeted access to specified parts of transcripts and thus a systematic research in a corpus linguistic approach. Both components are going to be available through a REST API. In this contribution, we focus only on the developments in the second (search) module and describe our work in progress towards selecting a suitable framework for querying spoken language data.

3. MTAS

MTAS (Brouwer et al. 2016) is an approach for creating and searching indexes of language corpora with multi-tier annotations. It was developed to be primarily used in the Nederlab project[13] for querying large collections of digitized texts.

MTAS builds on the existing Apache Lucene approach[14] and extends this by including complex linguistic annotations in the Lucene search index: During tokenization of a document, MTAS handles linguistic structures and span annotations as the same type as textual tokens and stores them on their first token position as Lucene would do this with n-grams. In the Lucene approach, text files to be indexed are stored as *Documents* comprising one or more *Fields*. Each *Document Field* represents the key-value relationship where a key is "content" or one of the metadata categories (e.g. author, title) and the value is the term to be indexed (e.g. in case of the category "title", it can be a token or a token sequence from the title of the text). MTAS indexes linguistic annotations and text in the same Lucene *Document Field*. The combination of prefix and postfix is used as a value of every token to distinguish between text and various annotation layers (cf. Table 1). In addition to the Lucene inverted index, MTAS provides forward indices to retrieve linguistic information based on positions and hierarchical relations.

We chose MTAS because it supports parsing of annotated texts in multiple XML-based formats, among others the TEI-based ISO standard for transcriptions of spoken language, which is used for transcripts in the AGD. To map data with custom annotations to the MTAS index structure only requires adjusting the parser configuration file. Many

IDs	Position	Parent	Token (Prefix [] Postfix)
[00001]	[0-43]	[ ]	[annotationBlock][ ]
[00002]	[0-43]	[00001]	[u][ ]
[00003]	[0-16]	[00002]	[seg][ ]
[00004]	[0-16]	[00002]	[seg.speaker][RH_0233]
[00005]	[0-16]	[00002]	[seg.speaker.sex][female]
[00006]	[0-16]	[00002]	[seg.type][contribution]
[00007]	[0]	[00003]	[word][ja]
[00008]	[0]	[00003]	[id][w122]
[00009]	[0]	[00003]	[norm][ja]
[00010]	[0]	[00003]	[lemma][ja]
[00011]	[0]	[00003]	[pos][NGIRR]
[00012]	[1]	[00003]	[pause][ ]
[00013]	[1]	[00003]	[id][p26]
[00014]	[1]	[00003]	[pause.type][micro]
[00015]	[2]	[00003]	[word][ähm]
[00016]	[2]	[00003]	[id][w123]
[00017]	[2]	[00003]	[norm][äh]
[00018]	[2]	[00003]	[lemma][äh]
[00019]	[2]	[00003]	[pos][NGHES]
[00020]	[3]	[00003]	[word][vielen]
[00021]	[3]	[00003]	[id][w124]
[00022]	[3]	[00003]	[norm][vielen]
[00023]	[3]	[00003]	[lemma][viele]
[00024]	[3]	[00003]	[pos][PIAT]

Table 1: List of tokens extracted from the transcript excerpt presented in Figure 1.

different types of annotations (incl. stand-off annotations, hierarchical relations and overlaps) are supported in MTAS and can be queried using the MTAS Corpus Query Language[15] (MTAS CQL) - a modified version of CQP Query Language[16] (CQP QL) introduced by the IMS Open Corpus Workbench[17] (CWB). Moreover, MTAS is a Lucene-base framework, which speaks in favor of scalability. MTAS is implemented in Java and is freely available as open source code[18].

4. Data

For testing MTAS, we selected three spoken language corpora from our archive (cf. Table 2). These are the Research and Teaching Corpus of Spoken German (Forschungs- und Lehrkorpus Gesprochenes Deutsch, FOLK, Schmidt, 2017), the German part of the Comparative Corpus for Spoken Academic Language (Gesprochene Wissenschaftssprache, GeWiss, Fandrych et al. 2017) and the Corpus of Mennonite Low German from North and South America (Mennonitenplautdietsch in Nord- und Südamerika, MEND, Kaufmann, in print). These corpora with a total size of almost 3.5 million transcribed tokens were collected between 1999 and 2019. While FOLK and GeWiss comprise authentic spontaneous interactions in German language with two and more native as well as non-native speakers recorded in various communication situations in Germany and abroad, the MEND corpus contains Plautdietsch translations of English, Spanish and Portuguese sentences recorded in the USA and South America. Extensive metadata for speakers and speech events are provided.

[13] https://www.nederlab.nl/onderzoeksportaal/
[14] https://lucene.apache.org
[15] https://textexploration.github.io/mtas/search_cql.html
[16] http://cwb.sourceforge.net/files/CQP_Tutorial/
[17] http://cwb.sourceforge.net/
[18] https://textexploration.github.io/mtas/download.html

Corpus	Data Type	Recording Time	Size (h)	Transcribed Tokens	Speech Events	Documented Speakers	Annotations
FOLK	interactions, audio, video	2003-2019	250	2429489	306	876	normalized forms, part-of-speech tags, lemmas, phonetic annotations, speech-rate
GeWiss	interactions, audio	2009-2012	92	743402	257	480	normalized forms, part-of-speech tags, lemmas, code-switching incl. translations, discourse comments
MEND	dialect corpus, audio	1999-2002	40	296867	321	322	normalized forms, part-of-speech tags, lemmas, prompt/translations, number of target prompt sentence

Table 2: AGD corpora selected for testing MTAS.

The audio- and video recordings are transcribed in modified orthography ("literarische Umschrift") according to the guidelines for the cGAT minimal transcript (Schmidt et al., 2015). Time-aligned speech segments are tokenized, orthographically normalized and enriched with different kinds of timeline- or transcription-based annotations. The annotations were either performed manually or generated automatically. They include e.g. part-of-speech tags, lemmatization, phonological annotations, speech-rate information, code-switching and discourse comments. The corpora differ according to the annotations they include, but taken together, the selected three corpora cover all types of annotations occurring in the entire corpus archive.

The audio transcripts and annotations are stored in the ZuMult format based on the ISO-TEI standard for transcriptions of spoken language. The ZuMult specification requires the mandatory use of <annotationBlock> elements for grouping utterances[19] of the same speaker and the stand-off annotations referring to them (see Figure 1). <annotationBlock> elements consist of exactly one <u> element containing the basic orthographic transcriptions and may contain an arbitrary number of <spanGrp> elements used to represent annotations of different types. Speaker utterances are fully tokenized and represented as a sequence of word tokens (<w> elements), pauses (<pause>), vocalized but non-lexical phenomena (<vocal>) and non-verbal events (<incident>). All these elements are embedded in <seg> elements directly beneath the <u> element. In our corpora, the <seg> elements correspond to speaker contributions – units of segmentation which are linked in time with the audio signal and which are terminated either by a silence of more than 0.2 seconds or by a change of speaker.

The temporal structure is represented by @start and @end attributes pointing to the @xml:id of <when> elements defined in the timeline. Additional <anchor> elements can be provided inside the <seg> element to specify further time points of interest, e.g. for a detailed representation of speaker overlaps. All elements within <annotationBlock>, except for <anchor> elements, require a unique @xml:id to be addressable for search. All token-based annotations like normalized forms, part-of-speech tags, lemmas etc. are encoded as attributes on the respective <w> element. Alternatively, these token-based annotations as well as all other types of annotations can be presented as spans within a <spanGrp> element. Figure 1 illustrates how transcription-based discourse comments (<spanGrp type ="DK">)[20] and timeline-based speech-rate information (<spanGrp type ="speech-rate">) are represented in our corpora.

5. Method

Before testing MTAS, we conducted an overview analysis of 20 existing search platforms providing access to spoken language corpora (a.o. DGD[21], KonText[22], Spokes[23], CQPweb, OpenSoNaR[24], Corpuscle[25], Glossa[26] and TEITOK[27]). Based on this overview analysis, we collected a set of search use cases and features supported by these platforms, regardless of the use of a query builder or one of the corpus query languages (CQP QL, ANNIS QL etc.) in order to submit queries on spoken language corpora. After that we incorporated the MTAS library into the search component of our corpus access architecture (Batinić et al. 2019) and implemented a simple frontend, in which a corpus can be selected and queries in MTAS CQL can be submitted. Our interest was focused on the following two aspects: 1) whether MTAS can be configured for mapping all types of annotations existing in our spoken language corpora 2) whether we can use MTAS CQL to formulate use cases that we are interested in.

[19] The utterance element (<u>) "is the fundamental unit of organization for a transcription, roughly comparable to a paragraph (<p> element) in a written document. It corresponds to a contiguous stretch of speech of a single speaker." (ISO 24624:2016, p. 6)

[20] "DK" stands for German "Diskurskommentierungen" (=discourse comments)

[21] https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome

[22] https://kontext.korpus.cz/

[23] http://spokes.clarin-pl.eu/

[24] https://portal.clarin.nl/node/4195

[25] http://clarino.uib.no/korpuskel/page?page-id=korpuskel-main-page

[26] https://tekstlab.uio.no/glossa2/

[27] http://www.teitok.org/

```
<annotationBlock xml:id="c26" who="RH_0233" start="TLI_67" end="TLI_82">
        <u xml:id="u_d43e2475">
                <seg type="contribution" xml:id="seg_d43e2475">
                        <anchor synch="TLI_67"/>
                        <w xml:id="w122" norm="ja" lemma="ja" pos="NGIRR" >ja</w>
                        <pause xml:id="p26" rend="(.)" type="micro"/>
                        <w xml:id="w123" norm="äh" lemma="äh" pos="NGHES">ähm</w>
                        <anchor synch="TLI_68"/>
                        <w xml:id="w124" norm="vielen" lemma="viele" pos="PIAT">vielen</w>
                        <w xml:id="w125" norm="Dank" lemma="Dank" pos="NN">dank</w>
                        <w xml:id="w126" norm="für" lemma="für" pos="APPR">für</w>
                        <w xml:id="w127" norm="die" lemma="d" pos="ART">die</w>
                        <w xml:id="w128" norm="freundliche" lemma="freundlich" pos="ADJA">freundliche</w>
                        <w xml:id="w129" norm="Einführung" lemma="Einführung" pos="NN">einführung</w>
                        <anchor synch="TLI_69"/>
                        <vocal xml:id="b6"><desc rend="°h">short breathe in</desc></vocal>
                        <anchor synch="TLI_70"/>
                        <w xml:id="w130" norm="äh" lemma="äh" pos="NGHES">ähm</w>
                        <anchor synch="TLI_71"/>
                        <incident xml:id="n14"><desc>schmatzt</desc></incident>
                        <w xml:id="w131" norm="äh" lemma="äh" pos="NGHES">äh</w>
                        …
                </seg>
        </u>
        <spanGrp type="DK">
                <span from="w124" to="w129">D2_Anfang</span>
                <span from="w132" to="w141">D1_Thema</span>
                <span from="w142" to="w146">D2_Vorstellung</span>
        </spanGrp>
        <spanGrp type="speech-rate">
                <span from="TLI_68" to="TLI_69">3.44</span>
        </spanGrp>
</annotationBlock>
```

Figure 1: An excerpt of the GeWiss corpus presented in ZuMult format.

6. Results and Discussion

6.1 Indexing

The MTAS configuration file provides a large repertoire of settings allowing us to consistently map our audio transcripts including all types of linguistic annotations to the MTAS search index. This requires no major modifications to the MTAS source code. Still, some difficulties arose because of essential structural differences between written and spoken language corpora.

The main challenge we faced in mapping spoken language data to the MTAS search index was to decide what elements of a transcript (word tokens, pauses, non-verbal sounds, time anchors etc.) can be considered as an equivalent to a text token.

From the point of view of calculating token distances, it would be more appropriate not to consider pauses and other audible and visible non-speech events in the same way as genuine word tokens. But querying these phenomena is very important for many use cases from discourse analysis. Therefore, they should be stored in the search index. Because MTAS does not provide an extra type to parse and index such kinds of annotations, we coded them at the word token level. We did this for <pause>, <vocal> and <incident> elements that are placed between word tokens (<w>) within a <seg> element (see Figure 1).

Furthermore, when talking about word token distances in spoken language, we should consider fillers like “äh” that could occur at any place in a word sequence. Therefore, users have to explicitly specify in their queries if the token sequence may or may not contain such fillers between desired word tokens. In the same way, optional pauses and other non-verbal events may be specified in queries as in (A). Users can be supported by query builders when formulating such complex queries.

(A) [word="herr"]([word="äh"]|<pause/>|<vocal/>|<incident/>)?[pos="NE"]
This query looks for word token “herr” followed by a proper name, where one filler, a pause or another non-verbal phenomenon can occur between “herr” and the proper name

A further general difficulty in querying spoken language corpora stems from the fact that individual tokens are often not synchronized with the audio sound because the audio alignment is usually made in contributions and other units above the word level (mainly due to reasons of efficiency in transcribing). Therefore, the temporal order of any two individual tokens is not always fully determined, and the document order of tokens does not always reflect their temporal order in the recording. This applies when speakers’ contributions overlap. It can be exemplified by the transcript excerpt in Figure 2. In the transcription document, the word token “hm” of speaker “HA” in line 0003 is directly preceded by the word token “ne” of speaker “PS” in line 0002. According to the timeline alignment, however, “hm” is preceded by and overlaps with the word token “okay”.

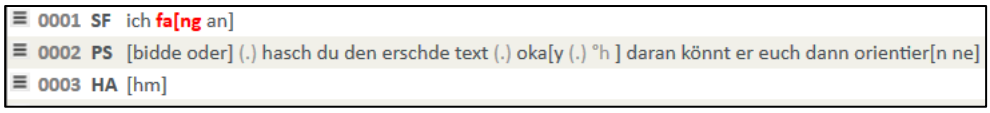

Figure 2: An audio transcript excerpt with speaker overlaps.

The same problem arises when dealing with token distances. Although the tokens "okay" and "hm" from the example in Figure 2 overlap, the token distance between these words according to the transcript would be 10, because 9 tokens occur between "okay" and "hm" in the transcript file.

The given problems with the token distance and precedence in spoken language corpora pose a lot of questions, that still remain unanswered and should be discussed beyond individual projects. The main question is whether the word token level is the right one to be a base tokenization/position level for indexing spoken language transcripts. Another question is whether individual speakers should perhaps be indexed separately (in a multiple tokenization model). MTAS for its part as search framework provides a flexible and transparent indexing approach that could serve as a starting point for further experiments with different tokenization models.

With regard to linguistic annotations, our experiments revealed that the MTAS indexing approach is suitable for dealing with

- token-based annotations (e.g. normalized form, lemma, POS)
- transcription-based span annotations that refer to a sequence of tokens coming from one speaker
- timeline-based span annotations that fully overlap with the structures (segments, utterances) placed within the same <annotationBlock>
- annotations coming from different annotation sources like different projects or tools for automatic annotation (e.g. Tree Tagger[28], MATE-Parser[29], OpenNLP[30])

Our intervention was needed for coding timeline-based annotations referring to a part of a segment. In MTAS, the end and the start of such annotations are automatically synchronized with the end and the start of the annotation block they are located in, because – according to the time references – the position of particular annotations cannot be encoded. We reimplemented the MTAS parser to replace time references with IDs of tokens located nearest to the respective time anchor. In that way, we achieved a more precise output, especially when annotations refer to a small part of a very large segment.

Finally, we would like to mention the difference between text and audio transcript with regard to metadata. While speech event information (i.e. information pertaining to the interaction or recording as a whole, such as date of recording, interaction type) is technically comparable with text metadata, speaker metadata (such as sex, age, education, etc.) have to be handled in a special way, because they can refer to discontinuous parts of a transcript rather than to the transcript itself. This applies for corpora consisting of interactions of two and more speakers. By using MTAS, we could easily index speaker information in the same way as structures and span annotations at the first token position of every segment originated from the respective speaker. For a query example, see Example (E).

6.2 Query

Once the MTAS index is created, it can be searched by using MTAS CQL. A closer look at this query language (QL) shows that MTAS CQL differs from all known QLs coming from the CQP family (e.g. Poliqarp[31], Sketch Engine's CQL[32], BlackLab's CQL[33]) and therefore represents yet another CQP dialect. It supports different types of search queries including positional constraints (A, B), containment (C, D) and intersecting relations (E, F). It allows to specify the distance and the precedence relation between query elements (G, H) as well as to use RegEx and Boolean operators for specifying token conditions (D, I).

(A) <seg>([word="vielen"][word="dank"])
This query looks for segments starting with "vielen dank"

(B) [incident="lacht"]</seg>
This query looks for a laughter at the end of a segment

(C) <seg.speaker="SF"/> !containing [lemma="äh"]
This query looks for segments of speaker "SF" not containing any forms of the filler "äh"

(D) [pos=".V.*"] within <DK="D1_Zeit"/>
This query looks for verbs in passages annotated with the tag "D1_Zeit" [34]

(E) <seg.speaker="PS"/> intersecting (<seg.speaker.sex="female"/> containing [lemma="hm"])
This query looks for segments of speaker "PF" intersecting with segments coming from female speakers and containing any forms of "hm"

(F) <seg/> fullyalignedwith ([word="so"]{2})
This query looks for segments consisting of two word tokens "so"

(G) [word="ich"][]{1,3}[word="du"]
This query looks for "ich" and "du" with a minimum of one and maximum of 3 tokens in between

[28] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger
[29] https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools/
[30] http://opennlp.apache.org/
[31] http://www.nkjp.pl/poliqarp/help/ense3.html
[32] https://www.sketchengine.eu/documentation/cql-basics/
[33] https://github.com/INL/BlackLab/blob/master/core/src/site/markdown/corpus-query-language.md
[34] "D1_Zeit" is a discourse comment used in GeWiss corpus to annotate passages where speakers mention the time limitation of their reports.

(H) [norm="Untersuchung"] precededby [w="die"]
This query looks for all transcribed forms of "Untersuchung" if they are preceded by token "die"

(I) [norm="wir|mir" & !word.type="assimilated"]
This query looks for all transcribed but not assimilated forms of "wir" and "mir"

Our tests revealed certain limitations of MTAS CQL, namely, the absence of some operators that are important for querying use cases typical for the spoken language research, e.g.

- comparison operators "<=" and ">=" that could be used for querying numerical values, e.g. searching pauses or speech-rates shorter or longer than N
- RegEx "*" (0 or more) and "+" (1 or more) that can be used in a token sequence to find e.g. two certain word tokens also if some fillers, pauses and other transcribed phenomena occur in between
- variables that can be used to refer to query elements as implemented in Poliqarp (J) or SketchEngine (K). Such references are important to search for repetitions and speaker overlaps (L).

(J) [case=$1 & pos=adj] [case=$1 & pos=subst]
This query is formulated in Poliqarp and looks for an adjective followed by a noun in case agreement with the preceded adjective

(K) 1:[] 2:[] & 1.word = 2.word
This query is formulated in Sketch Engine's CQL and looks for a word token repetition

(L) (<seg.speaker="$1"/>) intersecting (<seg.speaker="$2"/>) & $1 != $2
This is a fictional query looking for speaker overlaps (= segment A intersecting with segment B whereby both segments contain contributions coming from different speakers)

Our findings were reported to the MTAS developer, and meanwhile, some operators that we missed in MTAS CQL during our tests are already implemented in the current MTAS version (v. 8.4.1.1.).

What should be particularly emphasized is the flexibility of MTAS QL regarding different types of annotations: new annotation levels can be added to transcripts without the need to adapt the QL or to change other settings in the MTAS configuration. Just adding a new <spanGrp> element to the transcript, specifying its @type attribute and reindexing the corpus is sufficient to be able to search for these new annotations. For example, if disfluency annotations are added as shown in (M), queries <disfluency/>, <disfluency="TROUBLE"/> or <disfluency="REPAIR"/> can be used to find the spans corresponding to these annotations.

(M)
```
<spanGrp type="disfluency">
    <span from="w874" to="w875">
        TROUBLE</span>
    <span from="w876" to="w880">
        REPAIR</span>
</spanGrp>
```

6.3 Search Output

Every hit retrieved from the MTAS index contains all tokens occurring at the matched positions. For example, searching for [lemma="äh"] in the index excerpt from Table 1 would return the following list of MTAS tokens:

[annotationBlock][ ], [u][ ], [seg][ ],
[seg.speaker][RH_0233], [seg.speaker.sex][female],
[seg.type][contribution], [word][ähm], [id][w123],
[norm][äh], [lemma][äh], [pos][NGHES]

From this output, token IDs can be extracted and used to find the corresponding place in the appropriate transcript. All structures and linguistic annotations for the match are also available for different representations in the user interface.

The difficulty arises when determining the context of the match, e.g. for the presentation in a KWIC view. Here, we come across the problem that was already mentioned in Section 6.1. The context around words in a transcript document (consisting of a list of speaker contributions) is not necessarily identical to the immediately preceding and following context in the audio. The real context can be determined only if all individual tokens are aligned with the original recording. It is against this background that further questions arise, e.g. what exactly is the context of one word occurring within speaker overlaps? Is KWIC maybe not the optimal output/visualization form for all types of search results in case of spoken language? Even if these issues do not primarily concern MTAS, we find it important to mention them in this paper, because sooner or later, any developer of search platforms for spoken language corpora will be faced with these questions.

7. Conclusion and Future Work

Applying MTAS for indexing and querying corpora described in Section 4 revealed that this framework is suitable to be used as a search environment for AGD corpora in their present state. With MTAS, we achieve a good first approximation to a query mechanism for spoken language corpora which is both sufficiently similar to established query mechanisms for written language, and which can at the same time handle a substantial proportion of the structures and annotations specific to spoken language.

As a next step, we plan to enrich our data by discontinuous annotations, relations and annotations that do not refer to the concrete speaker but to parts of the interaction itself like annotations of sequences of social actions as they are used in the research field of Conversation Analysis (cf. ten Have, 2007). It would be interesting to see how such annotations can be indexed and searched with MTAS. We suspect there will be challenges of two kinds: 1) to find the right form for the presentation of such annotations and this form should suit both the ISO-TEI and the MTAS input format 2) to specify the search output if annotations refer to passages with speaker overlaps.

The clear and structured code of MTAS offers opportunities for further development. We see potential for merging the MTAS indexing component with one of the more advanced Lucene-based search modules, e.g.

Korap[35]. Korap supports Koral QL[36] – a serialization of Corpus Query Language Franca (CQLF, ISO 24623-1:2018) – and therefore provides an extensive set of search possibilities.

The MTAS indexing approach itself has convinced us. It stands out with its extensive parser configuration options. From our point of view, it can be used and is **w**orth a recommendation for indexing spoken language corpora.

8. Acknowledgements

We would like to thank Matthijs Brouwer, the developer of MTAS, for friendly support to better understand the framework. Furthermore, we are very grateful to the anonymous reviewers whose insightful comments helped to improve and clarify this paper.

9. Bibliographical References

Batinic, J., Frick, E., Gasch, J. and Schmidt, T. (2019). Eine Basis-Architektur für den Zugriff auf multimodale Korpora gesprochener Sprache. Digital Humanities im deutschsprachigen Raum, DHd 2019 28.3.2019, Frankfurt.

Brouwer, M., Brugman, H. and Kemps-Snijders, M. (2016). MTAS: A Solr/Lucene based Multi-Tier Annotation Search solution, Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence.

ISO 24624:2016. Language resource management — Transcription of spoken language.

ISO 24623-1:2018. Language resource management — Corpus query lingua franca (CQLF) — Part 1: Metamodel.

Kupietz, M. and Schmidt, T. (2015). Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In Eichinger, L. M. (Ed.), *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven*, pp. 297–322 - Berlin/Boston: de Gruyter, 2015. (Jahrbuch des Instituts für Deutsche Sprache 2014).

Schmidt, T., Schütte, W. and Winterscheid, J. (2015). cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2). Working paper available at https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/4616/file/Schmidt_Schuette_Winterscheid_cGAT_2015.pdf

Stift, U.-M. and Schmidt, T. (2014). Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch. In Institut für Deutsche Sprache (Eds.), *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache.* Redaktion: Melanie Steinle, Franz Josef Berens, pp. 360–375 - Mannheim: Institut für Deutsche Sprache, 2014.

ten Have, P. (2007). *Doing Conversation Analysis : A Practical Guide.* London: Sage Publications.

10. Language Resource References

Fandrych, C., Meißner, C. and Wallner, F. (2017). Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora. Tübingen. Stauffenburg.

Kaufmann, G. (in print). The World Beyond Verb Clusters: Aspects of the Syntax of Mennonite Low German. In Auer, P., Hinskens, Frans L. und Kerswill, P. (Eds.), *Reihe Studies in Language Variation*. Amsterdam/Philadelphia: John Benjamins.

Schmidt, T. (2017). Construction and Dissemination of a Corpus of Spoken Interaction – Tools and Workflows in the FOLK project. In Kupietz, M. and Geyken, A. (Eds.), *Corpus Linguistic Software Tools, Journal for Language Technology and Computational Linguistics* (JLCL 31/1), pp. 127–154.

[35] https://korap.ids-mannheim.de, source code: https://github.com/KorAP

[36] https://korap.github.io/Koral