Software Development for Working with Oral Corpora

Elena Frick (IDS) & Thomas Schmidt (LinguisticBits)

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

ZuMult

Open source library improving networking, interoperability, standardization and flexibility of software for spoken language corpora

https://github.com/zumult-org/zumultapi

BACKGROUND

Web applications for access to spoken language corpora: the same functionality, but totally different technical basis strongly tied to the data format and user needs they were designed for (cf. Batinić et al. 2021)

Problem: Reuse of software components for new corpora types and new usage scenarios

Proposed solution: Object-oriented modeling and three-layer client-server architecture for corpus platforms + Use of Standards

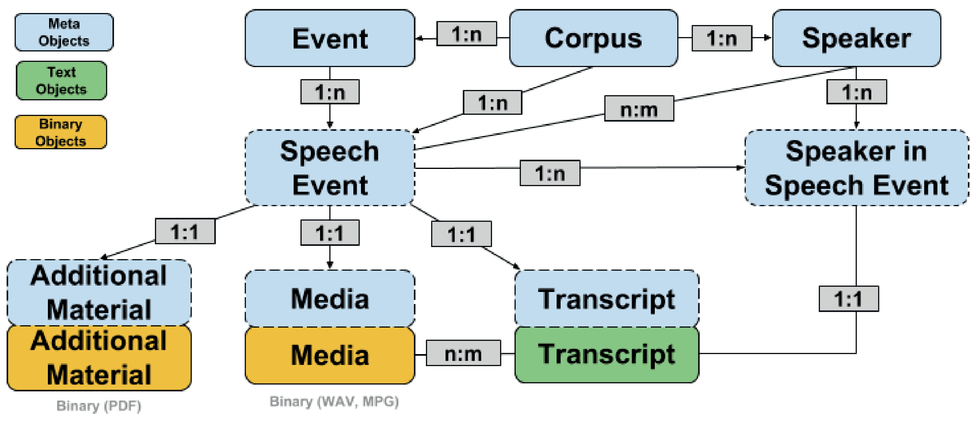
PROJECT

ZuMult stands for German "Zugänge zu multimodalen Korpora gesprochener Sprache" (Access to multimodal corpora of spoken language, Fandrych et al. 2022) – funded by the DFG (LiS-program). **https://zumult.org**

Cooperation partners:

hzsk hamburger zentrum für sprachkorpora | Herder-Institut | UNIVERSITÄT LEIPZIG

OBJECT-ORIENTED CORPUS DATA MODEL



THREE-LAYER SOFTWARE ARCHITECTURE

- **internal layer** – physical storage of data, e.g. in a relational database, a data repository, or a file system on a server
- **conceptual layer** – application logic for data access defining interfaces for corpus objects with their relationships, properties and methods (e.g. for getting a transcript excerpt)
- **external layer** – user interfaces presenting contents to the users in the form of views, e.g. a website in browser displaying a transcript or corpus search results as KWIC.

Software based on such three-layer architecture can be easily adapted to the specifics of corpora located in different repositories and at the same time offers a lot of flexibility for the development of user-group specific applications (e.g. corpus platforms for DaF/DaZ or for conversation analysis research).

IMPLEMENTATION / USE OF STANDARDS

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

	0 [00:00.0]	1 [00:01.1]	2 [00:	3 [00:01.9]
X [v]	You keep (0.2) interrup	ting me.		
X [de]	Du unterbrichst mich immer.			
Y [v]		I am so sor	ry.	
Y [de]		Tut mir leid.		

```
<text xml:lang="de">
    <timeline unit="s">
        <when xml:id="TLI_0" interval="0.0" since="TLI_0"/>
        <when xml:id="TLI_1" interval="1.3" since="TLI_0"/>
        <when xml:id="TLI_2" interval="2.3" since="TLI_0"/>
        <when xml:id="TLI_3" interval="4.3" since="TLI_0"/>
        <when xml:id="TLI_4" interval="6.3" since="TLI_0"/>
    </timeline>
    <annotationBlock xml:id="ab_1" who="X" start="TLI_1" end ="TLI_3">
        <u xml:id="u_1">
            <w xml:id="w1" lemma="you" pos="PRO">you</w>
            <w xml:id="w2" lemma="keep" pos="V">keep</w>
            <pause rend="(0.2)" dur="PT0.2s" />
            <w xml:id="w3" lemma="interrupt" pos="V">inter<anchor
                synch="TLI_2"/>rupting</w>
            <w xml:id="w5" lemma="I" pos="PRO">me </w>
            <pc>.</pc>
        </u>
        <spanGrp type="translation" xml:lang="de">
            <span from="w1" to="w5">Immer unterbrichst Du mich.</span>
        </spanGrp>
    </annotationBlock >
    ...
</text>
```

Media:
PCM-WAV/MP3 (Audio);
MPEG-4 (Video)

Transcriptions and annotations:
ISO 24624:2016
based on XML and Text Encoding Initiative (TEI); interoperable with the formats of the established transcription editors EXMARaLDA, FOLKER, ELAN, Praat und Transcriber

Event und speaker metadata:
XML, CMDI

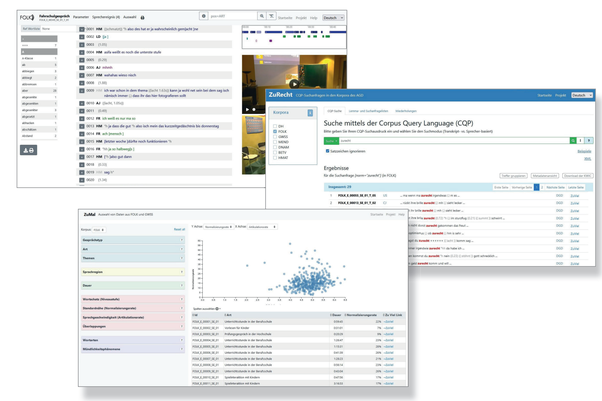
Backend:
Java EE Framework, REST API,
Search functionality realized with MTAS
(open source Lucene-based search engine for querying text with multilevel annotations)
+ CQP Query Language

Client/web application prototypes:
jQuery, XSLT, HTML5, Bootstrap;
Data visualization with XSL, SVG, WebVTT

MTAS-based search index

```
[00001] [0]   [00084]   [w] [so]
[00002] [0]   [00084]   [w.id] [w1]
[00003] [0]   [00084]   [pos] [NGIRR]
[00084] [0-13][00089]   [u.id] [u_d1e17]
[00085] [0-13][00089]   [u.speaker] [LB]
```

Prototypes for user-group differentiated corpus access



Contact:
Elena Frick M.A.
Abteilung Pragmatik
Leibniz-Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim, Germany

Phone: +49 621 1581-492
Fax: +49 621 1581-200
frick@ids-mannheim.de

Dr. Thomas Schmidt
thomas@linguisticbits.de
https://linguisticbits.de/

Street Address:
Leibniz-Institut für Deutsche Sprache
R 5, 6-13
68161 Mannheim, Germany

Phone: +49 621 1581-0
Fax: +49 621 1581-200
info@ids-mannheim.de
www.ids-mannheim.de

Leibniz Association

References:

Batinić, J., Frick, E., and Schmidt, T. (2021). Accessing spoken language corpora: an overview of current approaches. In Corpora, 16 (3): 417–445. Edinburgh University Press.

Fandrych, C., Frick, E., Kaiser, J., Meißner, C., Portmann, A., Schmidt, T., Schwendemann, M., Wallner, F., and Wörner, K. (2022). ZuMult: Neue Zugangswege zu Korpora gesprochener Sprache. In: Kämper, H. et al. (eds.), Sprache in Politik und Gesellschaft: Perspektiven und Zugänge. Jahrbuch des IDS. Berlin etc.: de Gruyter.