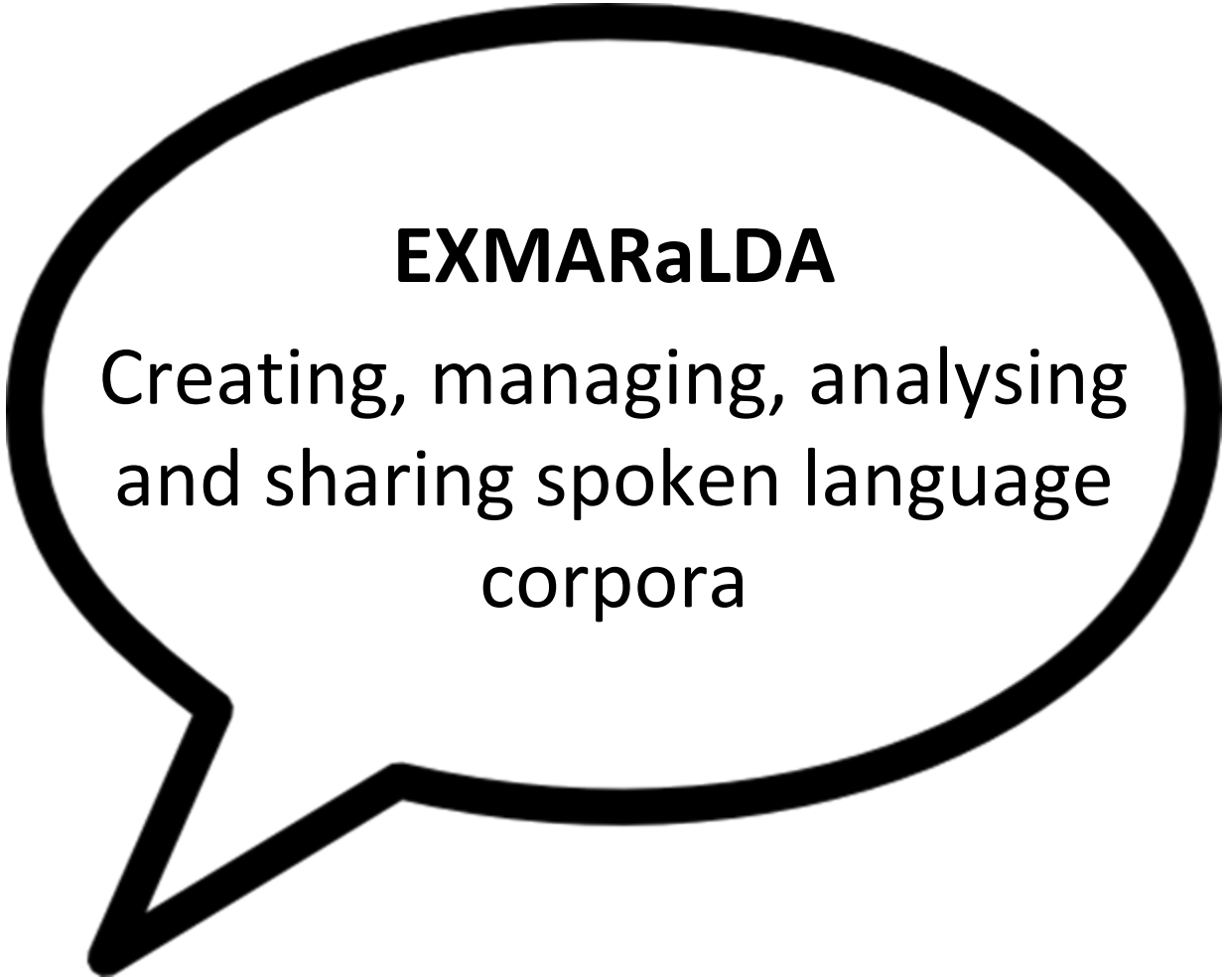

EXMARaLDA
Creating, managing, analysing and sharing spoken language corpora

Outline

- Background
- Aims and principles
- System overview
- Demonstrations
  - Transcription (Partitur-Editor)
  - Analysis (EXAKT)
  - Publication

Background

- Research Centre on Multilingualism, University of Hamburg
- Since 1999, until June 2011
- More than 20 projects in 4 groups
  - Multilingual **individuals**
  - Multilingual **communication**
  - **Historical** multilingualism
  - Transfer group (Practical **applications**)

Background

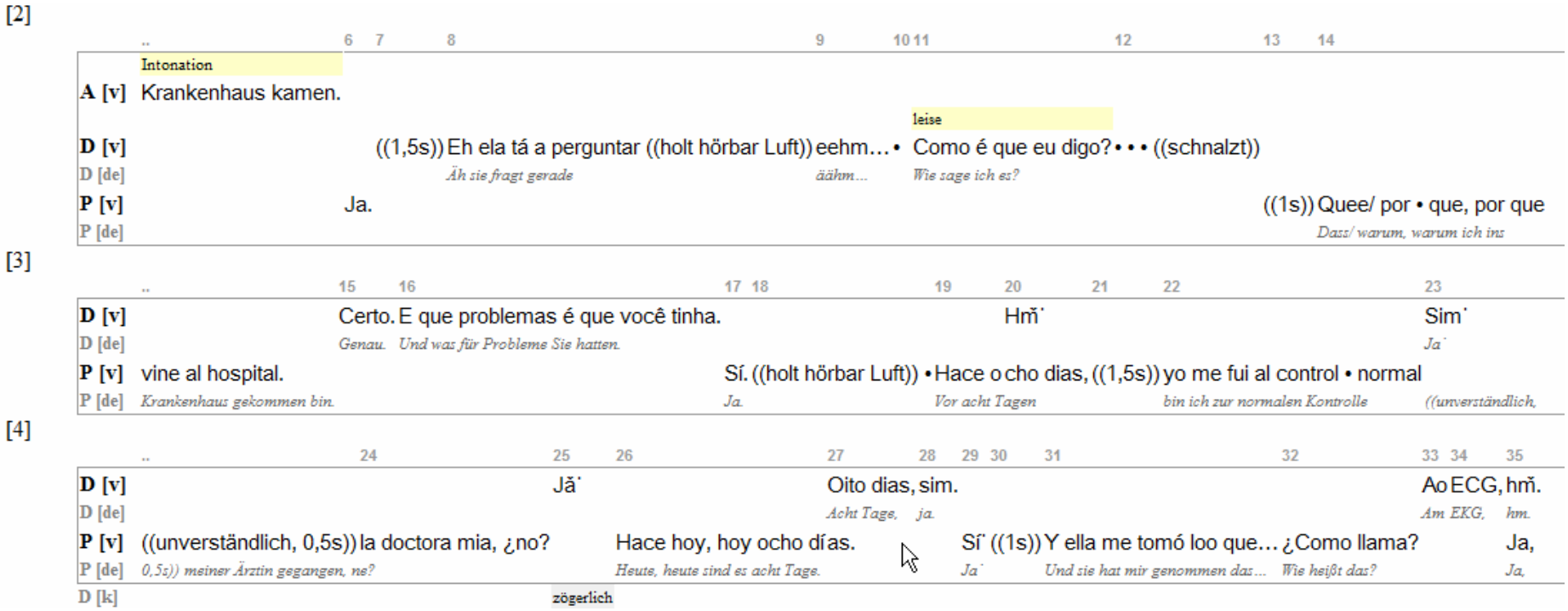
- All work empirical, based on corpus data
- Mostly oral data
  - Language **acquistion** corpora (bilingual children: French, Spanish, Turkish, Portuguese/German Spanish/Basque)
  - Language **attrition** corpora (bilingual adults: Polish, Italian/German)
  - **SLI** data (German / Turkish)
  - **Interpreting** data (Portuguese,Turkish / German)
  - Interscandinavian **communication**
- also: written data (e.g. historical texts, translation corpora)

Example: French/German bilinguals

DUFDE - AN-F069 [Prev] [Next]

	ANNIKA	Interviewer	
00002	de kaka		
		ha ha	
00003	ca pas		An choisit l'un des jeux
		tu veux pas jouer à celui-là?	
00004	non ca c'est un		An mélange les cartes
		tu mélanges bien	
00006	comme ca		An montre à P comment ranger les cartes
		on fait comme ca toutes	
00007	maintenant comme ca		An mélange les cartes
		et moi aussi	
00008	oh tu (x)		
00009	/tu/=toutes les cartes		
		toutes les cartes	
00012	les deux de ca		
00013	de ca		An montre les cartes sur la boîte
		il faut retrouver les deux cartes	
00015	maintenant ici		An se met sur le ventre et continue à ranger les cartes
		attends je finis ici d'abord	
00016	ah maintenant (xx) là		
		et ensuite quel côté	

Example: interpreted doctor/patient communication



Background

- Our project: „Computer-assisted methods for creating and analysing multilingual data“
- Software development → **EXMARaLDA**
- Corpus curation and dissemination → **Data sharing**
- Methodological questions → **Workflows, quality control, transcription conventions**
- Sustainability of data (long-term archiving etc.) → **standardisation, infrastructures**

Aims and principles

- Develop a system which…
  - makes intelligent use of **multimedia** and **hypertext** capabilities of the computer
  - is usable for **different** theoretical **approaches**, different **research interests**
  - is usable in **different technological environments**
  - produces data which is easily **shared**
  - can ingest existing **„legacy data“**
  - produces data which has a chance of surviving the next 30 years

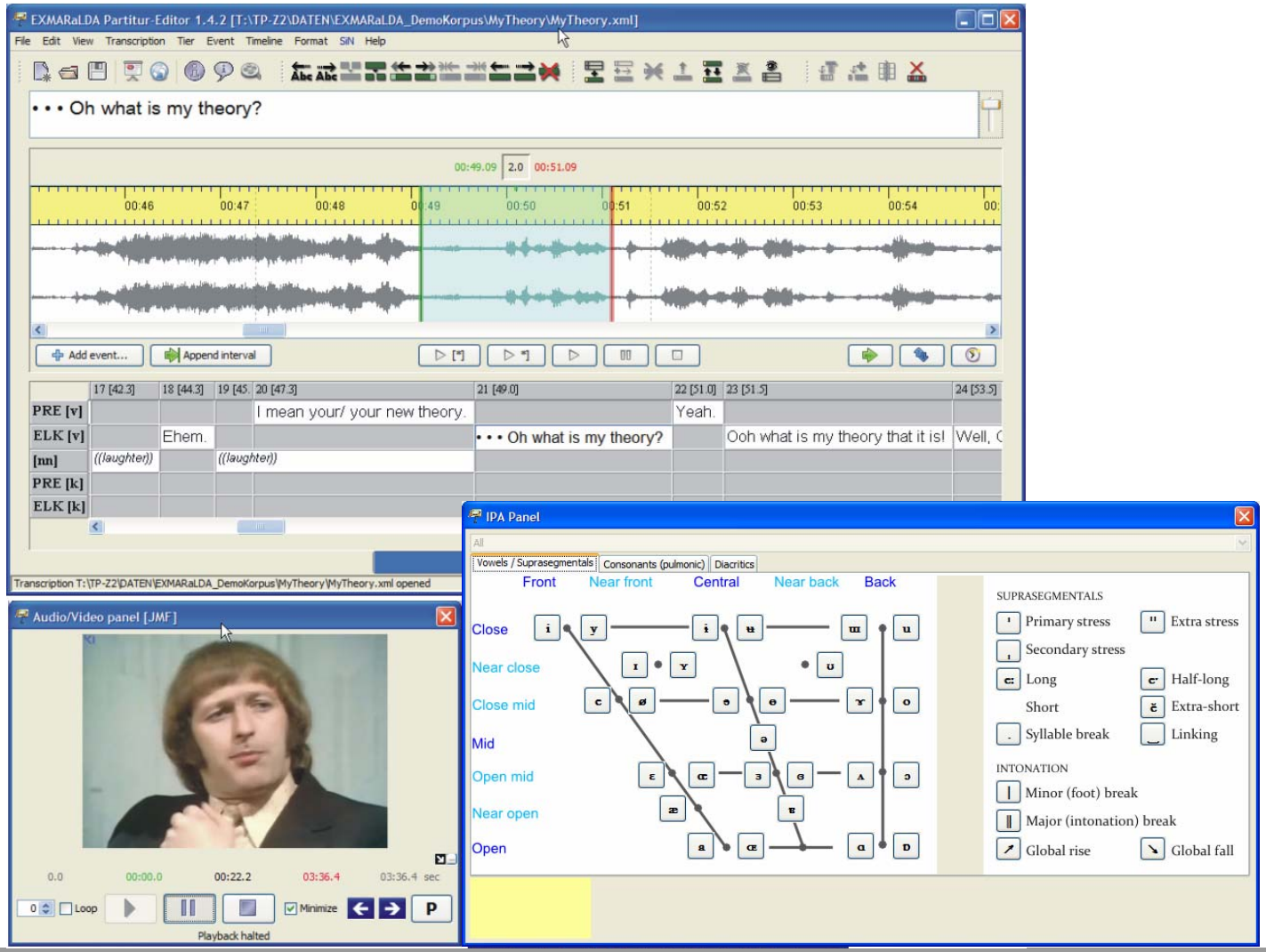
Aims and principles

- Data formats based on open standards (**XML, Unicode**)
- Theory-neutral data model (based on **annotation graphs**)
- Cross-Platform software tools (**JAVA**)
- Interoperability with other widely used tools (**Praat, ELAN, CLAN, Transcriber,** …)
- Interoperability with common desktop formats (**HTML, RTF, PDF**)
- Compatibility with emerging/possible standards (e.g. **TEI**)

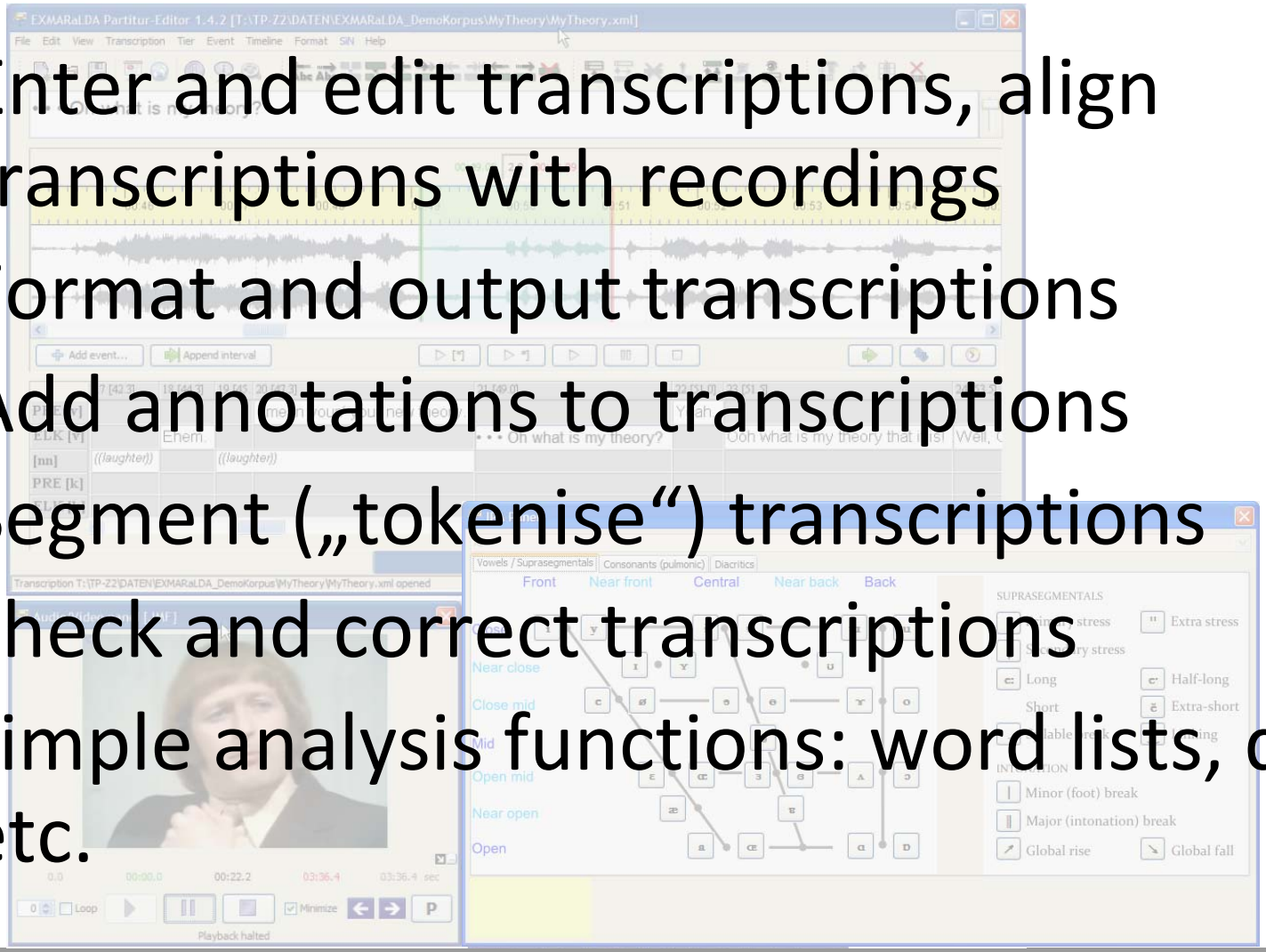
System overview: EXMARaLDA

- Data model, data formats
  - for **transcriptions** (+ annotations)
  - for **corpora** (+ metadata)
- Tools
  - **Partitur-Editor**: Transcription
  - **Corpus Manager**: Corpus Administration, Metadata
  - **EXAKT**: Query, Analysis
  - Utilities for data conversion, corpus publication etc.

System overview: Partitur-Editor

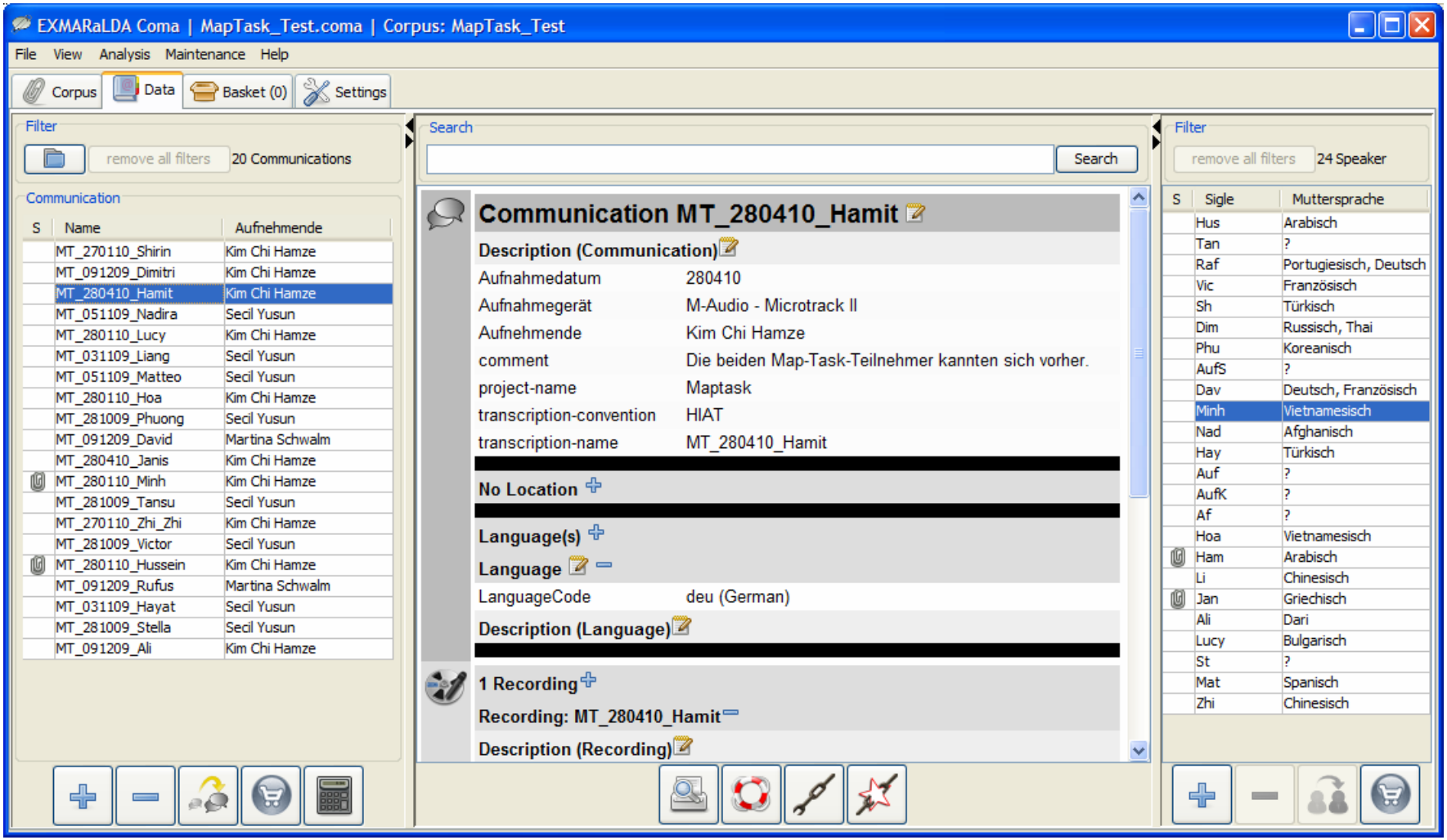


System overview: Partitur-Editor

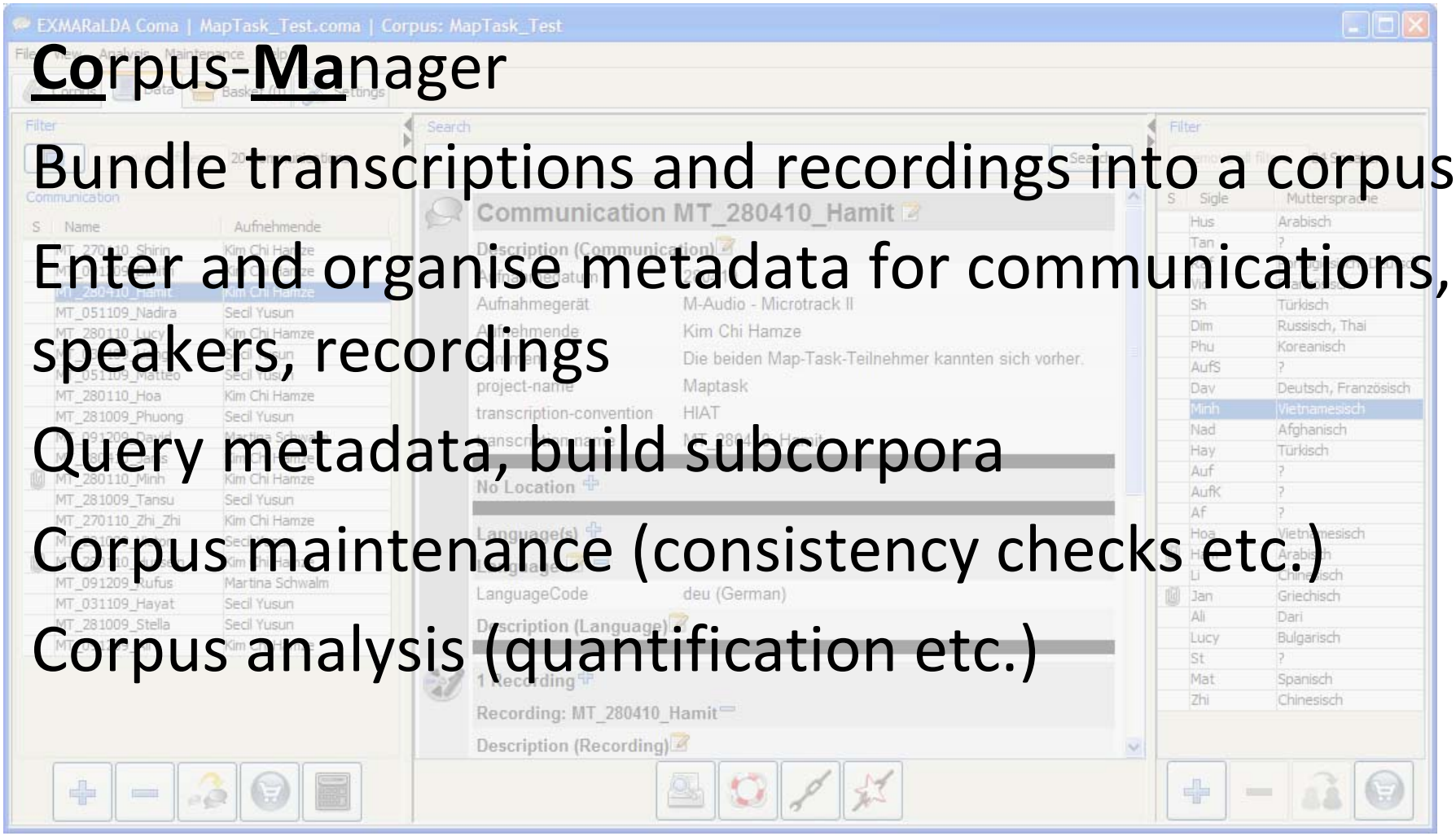


- Enter and edit transcriptions, align transcriptions with recordings
- Format and output transcriptions
- Add annotations to transcriptions
- Segment („tokenise“) transcriptions
- check and correct transcriptions
- simple analysis functions: word lists, query, etc.

System overview: Coma

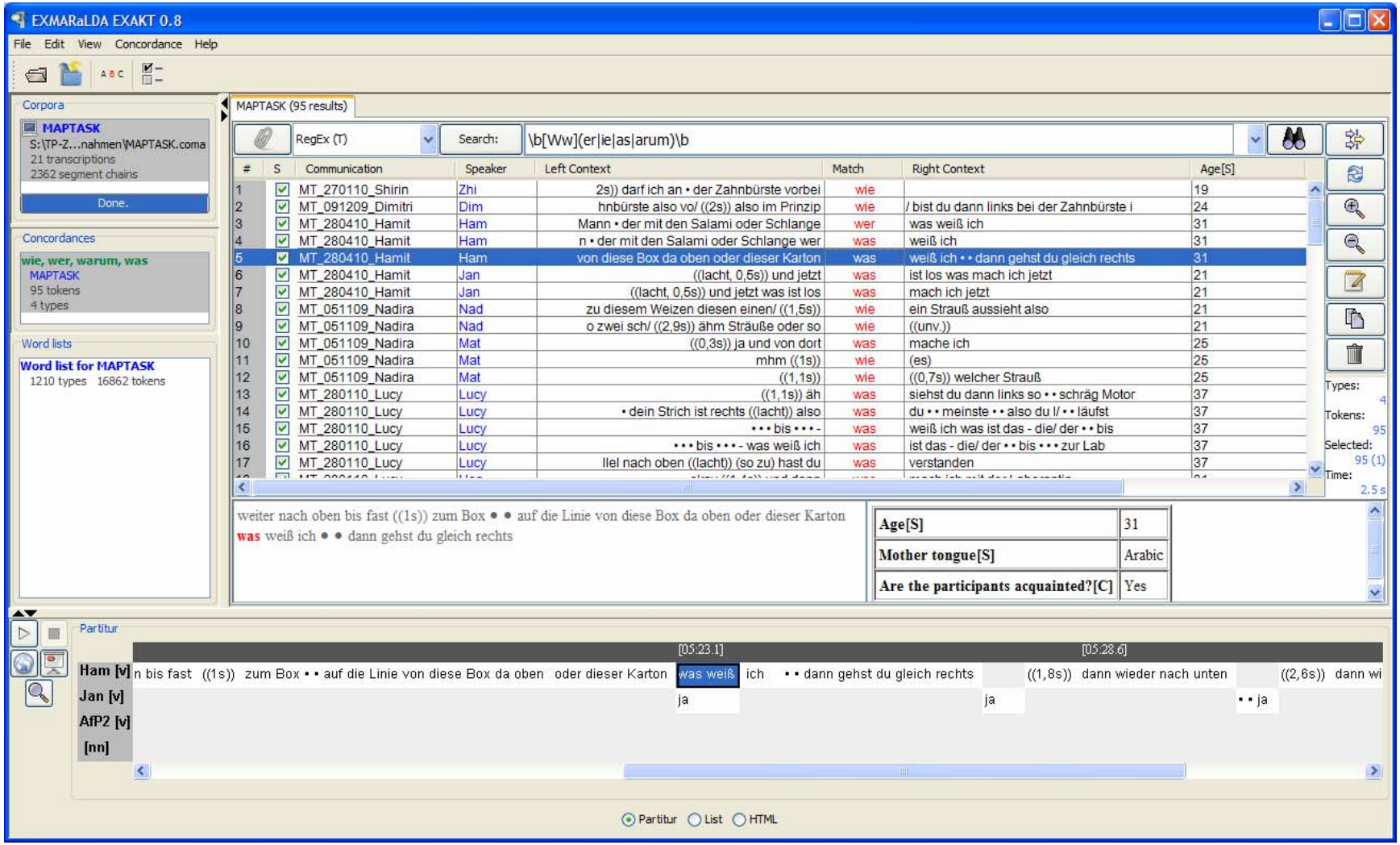


System overview: Coma

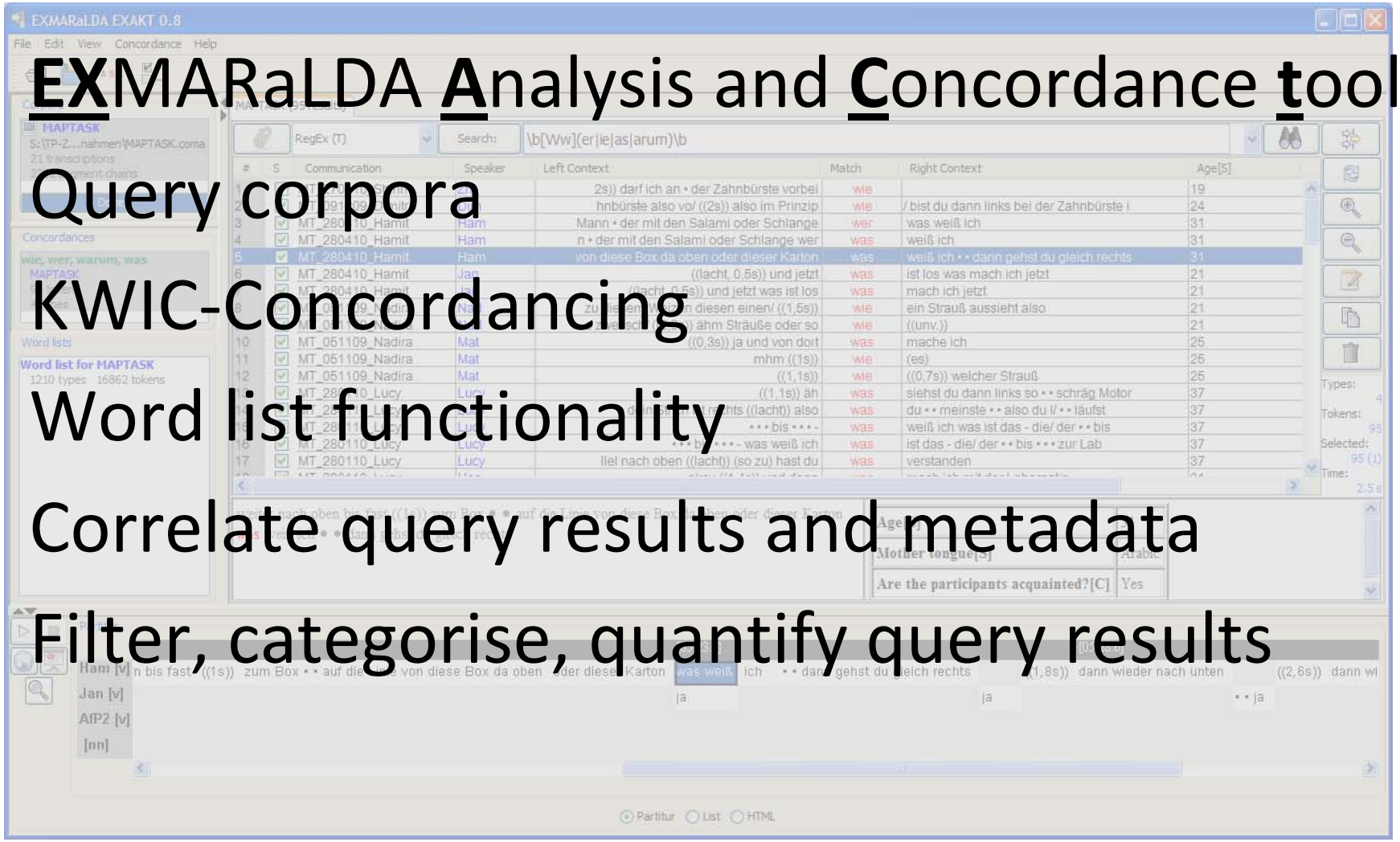


- **Co**rpus-**Ma**nager
- Bundle transcriptions and recordings into a corpus
- Enter and organise metadata for communications, speakers, recordings
- Query metadata, build subcorpora
- Corpus maintenance (consistency checks etc.)
- Corpus analysis (quantification etc.)

System overview: EXAKT

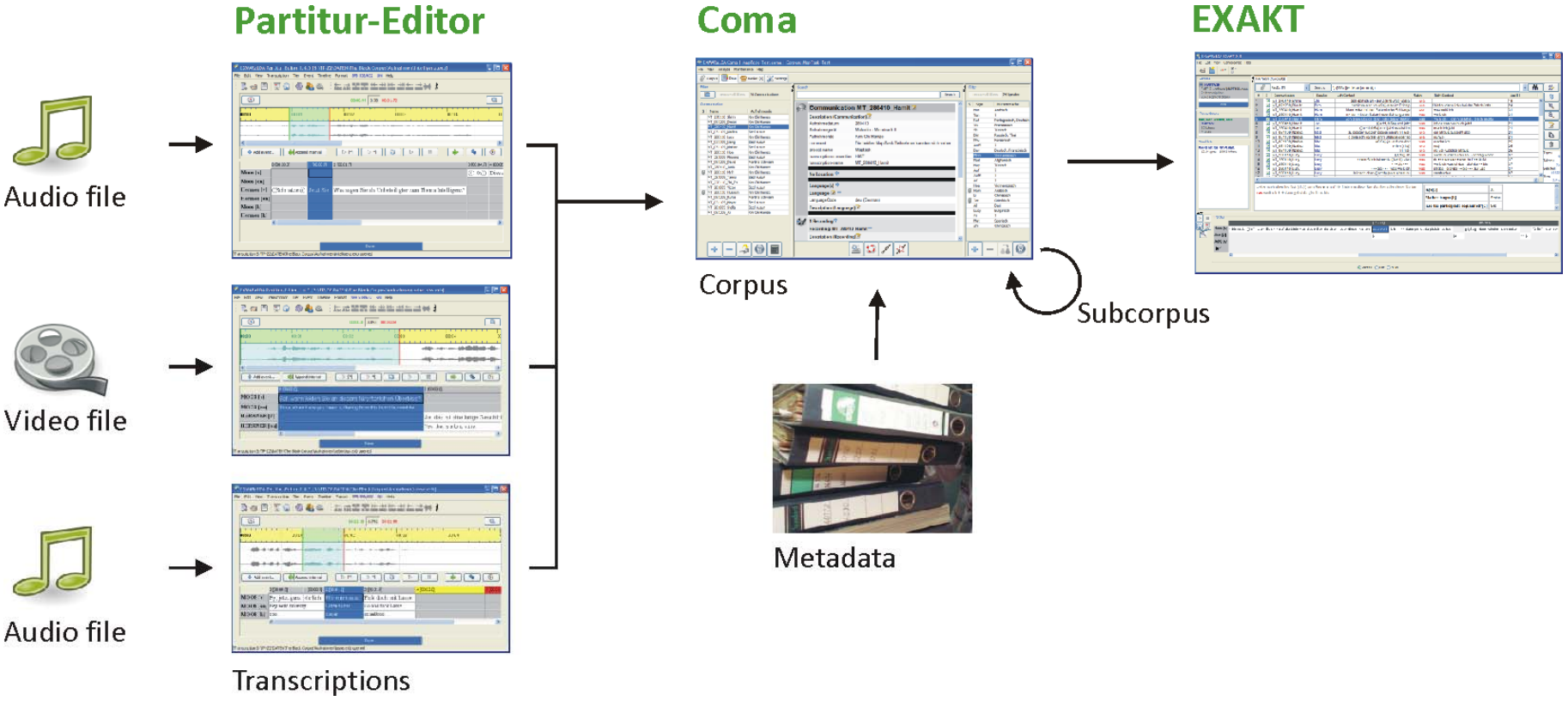


System overview: EXAKT



- **EX**MARaLDA **A**nalysis and **C**oncordance **t**ool
- Query corpora
- KWIC-Concordancing
- Word list functionality
- Correlate query results and metadata
- Filter, categorise, quantify query results

System overview: Corpus workflow



Demonstration: Transcription

- Example: TV debate Royal / Sarkozy
- Video excerpt
- Transcription according to HIAT conventions
- „Verbal“ tier with orthographic transcription, pauses, non-verbal behaviour
- English translation tier
- http://www1.uni-hamburg.de/exmaralda/files/demokorpus/Royal/presentation/royal_partiture.html

Demonstration: Analysis

- Example: Hamburg Map Task Corpus (http://www.exmaralda.org/corpora/en_sfb_z2.html)
- Task: Explaining a route on a map
- Advanced L2 learners of German with (very) different L1s
- Orthographic transcription only
- Metadata about speakers‘ language biographies
  - Learned German when and where?
  - Language use in everyday life?

Demonstration: Analysis

- Methodological issues:
  - Different type of relevant context:
    - Interactional context: other speakers, other modalities
    - Situational context: communication metadata
    - Biographic context: speaker metadata
  - „Corpus-driven“ analysis: development of hypotheses in a „dialogue“ with the data
  - Stepwise or iterative refinement and specification of queries, „explorative“ approach
  - Manual processing of search results
  - Ad-hoc categorisation (categories in the corpus vs. categories in the analysis)

Demonstration: Corpus publication

- Example: Hamburg Map Task Corpus
- Similar for other completed corpora at the Research Centre on Multilingualism:

  http://corpora.exmaralda.org
- Password protected access will be given upon request
- Methods for generating publication format are part of EXMARaLDA
- Similar Corpora elsewhere (e.g. METU Spoken Turkish Corpus in Ankara)
- Online corpus browsing in a web browser
- Online query through EXAKT

Practicalities

- Free download from

  http://www.exmaralda.org
- Documentation, mailing list, demo data, tutorials from the same website
- Published corpora from the same website

Thank you, Merci!