



How to use an EXMARaLDA corpus

This document explains how to use corpora from www.exmaralda.org. Examples are taken from the EXMARaLDA demo corpus. Other corpora work in a similar way but may differ, for example, with respect to the availability of audio data or the types of export formats.

Contents

1. Online use.....	2
Corpus overview	2
HTML visualizations.....	3
2. Offline use	4

1. Online use

You can use a corpus online (i.e. using your internet browser and without downloading data) to browse meta data, transcriptions and recordings.

Corpus overview

You usually start with a corpus overview:

The screenshot shows the EXMARaLDA Demo Korpus interface. At the top, there is a header bar with the text "EXMARaLDA Demo Korpus". Below this, the interface is divided into two main sections: "11 Communications" on the left and "24 Speakers" on the right. The "11 Communications" section lists several communication entries, including "Helge Schneider: Arbeitsamt (2 Speakers, 1 Transcription)". The "24 Speakers" section lists several speakers, including "AC (Arlotte Chabot)" and "AMT (Ralf Geritzon)".

11 Communications

Rudi Völlers: Wutausbruch (2 Speakers, 1 Transcription)

Helge Schneider: Arbeitsamt (2 Speakers, 1 Transcription)

folder	Arbeitsamt
Gesprächstyp	Telefonisches Arbeitsvermittlungsgespräch
project-name	EXMARaLDA DemoKorpus
Quelle	Helge Schneider: Es rappelt in Karton. 1995 bei Roof Music/EMI Electrola erschienen.
transcription-convention	HIAT (vereinfacht)
transcription-name	Helge Schneider: Arbeitsamt
Vorgeschichte	KLA ist nach einem Auslandsaufenthalt zurück in Deutschland und sucht eine Arbeit.

Speakers: KLA; AMT;

Location: Mülheimer Straße 36, Oberhausen, Deutschland

Institution: Agentur für Arbeit

Start: 1982-03-19T00:00:00

Duration:

Recording (2.363 minutes): Helge_Schneider_Arbeitsamt.mp3

Recording (2.363 minutes): Helge_Schneider_Arbeitsamt.wav

Transcription: Arbeitsamt

EXMARaLDA: [Transcription] [Segmented]

Visualisation: [Partiture] [RTF] [PDF] [XML] [Utterances] [Words] [Head]

Export: [TEI] [AG] [EAF] [Praat]

Hubert Fichte: Interview (2 Speakers, 1 Transcription)

Helge Schneider: Tropfsteinhöhle (2 Speakers, 1 Transcription)

Studio Braun: English Translator (3 Speakers, 1 Transcription)

24 Speakers

AC (Arlotte Chabot)

AMT (Ralf Geritzon)

Sex	male
Ausbildung: beruflich	Fachangestellter für Arbeitsförderung
Ausbildung: schulisch	Realschule
Beruf	Fachangestellter für Arbeitsförderung
Beruf der Lebensgefährtin	Fachangestellter für Arbeitsförderung
Beruf der Mutter	Hausfrau
Beruf des Vaters	Verwaltungsfachangestellte
Familie	Verheiratet
Name	Geritzon
Vorname	Ralf

Language: DEU

Status: LI

In Communications: Helge Schneider: Arbeitsamt,

ANR (Matice ???)

BP (Bernd Peterchen)

BS (Bernd Schwanmeister)

ERW (Erwin Schneider)

FF (Fiona Frick)

Fichte (Hubert Johannes Fichte)

FS (Fernando Savater)

The corpus overview consists of a list of communications (left side) and a list of speakers (right side). If you click on an item in one of these lists, information about the item will be displayed.

The screenshot shows the detailed view for the communication "Helge Schneider: Arbeitsamt (2 Speakers, 1 Transcription)". The view is divided into several sections: a top section with meta data items, a section for speakers, a section for location and institution, a section for recording files, and a section for transcription and export options.

Helge Schneider: Arbeitsamt (2 Speakers, 1 Transcription)

folder	Arbeitsamt
Gesprächstyp	Telefonisches Arbeitsvermittlungsgespräch
project-name	EXMARaLDA DemoKorpus
Quelle	Helge Schneider: Es rappelt in Karton. 1995 bei Roof Music/EMI Electrola erschienen.
transcription-convention	HIAT (vereinfacht)
transcription-name	Helge Schneider: Arbeitsamt
Vorgeschichte	KLA ist nach einem Auslandsaufenthalt zurück in Deutschland und sucht eine Arbeit.

Speakers: KLA; AMT;

Location: Mülheimer Straße 36, Oberhausen, Deutschland

Institution: Agentur für Arbeit

Start: 1982-03-19T00:00:00

Duration:

Recording (2.363 minutes): Helge_Schneider_Arbeitsamt.mp3

Recording (2.363 minutes): Helge_Schneider_Arbeitsamt.wav

Transcription: Arbeitsamt

EXMARaLDA: [Transcription] [Segmented]

Visualisation: [Partiture] [RTF] [PDF] [XML] [Utterances] [Words] [Head]

Export: [TEI] [AG] [EAF] [Praat]

For communications, the top part gives you a list of meta data items. This is followed by a list of speakers participating in the communication. Click on any of the speakers to display the

corresponding information in the speaker list. The lower part links to all the documents (recordings, transcriptions, visualisations and export formats) which belong to this communication. More specifically:

- The section **EXMARaLDA** links to an EXMARaLDA **Basic-Transcription**, which you can open and edit in the EXMARaLDA Partitur-Editor and an EXMARaLDA **Segmented-Transcription**.
- The section **Visualisation** links to **musical score** visualisations in four different formats (HTML, RTF, PDF and XML), to an **utterance list** (HTML) to a **word list** (HTML) and to a separate visualisation of the transcription **head** (HTML).
- The section **Export** links to several export formats. **TEI** is an XML file corresponding to the guidelines of the Text Encoding Initiative. **AG** is an annotation graph file which can be used for data exchange with various annotation tools. **EAF** is an ELAN Annotation File which can be opened and edited with the ELAN tool from the MPI in Nijmegen. **Praat** is a TextGrid which can be opened and edited with the Praat software.

Click on any of these files to display them in your browser. To download them you may have to right-click and choose "Download" from the context menu.

HTML visualizations

If you display the HTML version of a musical score visualisation (and if the corpus makes the audio recordings available), you'll be given a transcription that is linked to a flash audio player (1):

The screenshot shows the EXMARaLDA interface. The top bar has a title and navigation links. A green box labeled '1' highlights the audio player controls. Below the header, the main content area shows a musical score. A green box labeled '2' highlights a small arrow icon in the top row of the score, which is used to jump to a specific time point in the audio recording. The score itself consists of two parts, [1] and [2], each with multiple rows of text and time markers.

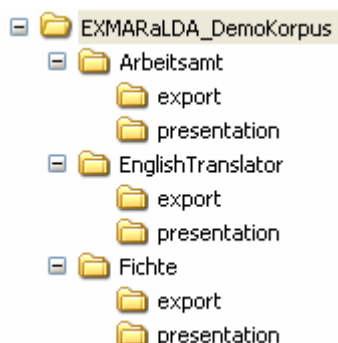
In the musical score, if you click on any one of the little arrows in the top rows (2), the player will start playback at the corresponding time in the audio recording. Clicking on any number in the top rows of a musical score will take you to the corresponding place in an utterance list:

The screenshot shows the EXMARaLDA interface with the utterance list. The top bar is the same as in the previous screenshot. A green box labeled '3' highlights the utterance list on the left side of the interface. The list contains 11 entries, each with a speaker label (AMT or KLA), a number, and a time marker. The text of the utterances is displayed in the main content area. A mouse cursor is pointing at the 10th entry.

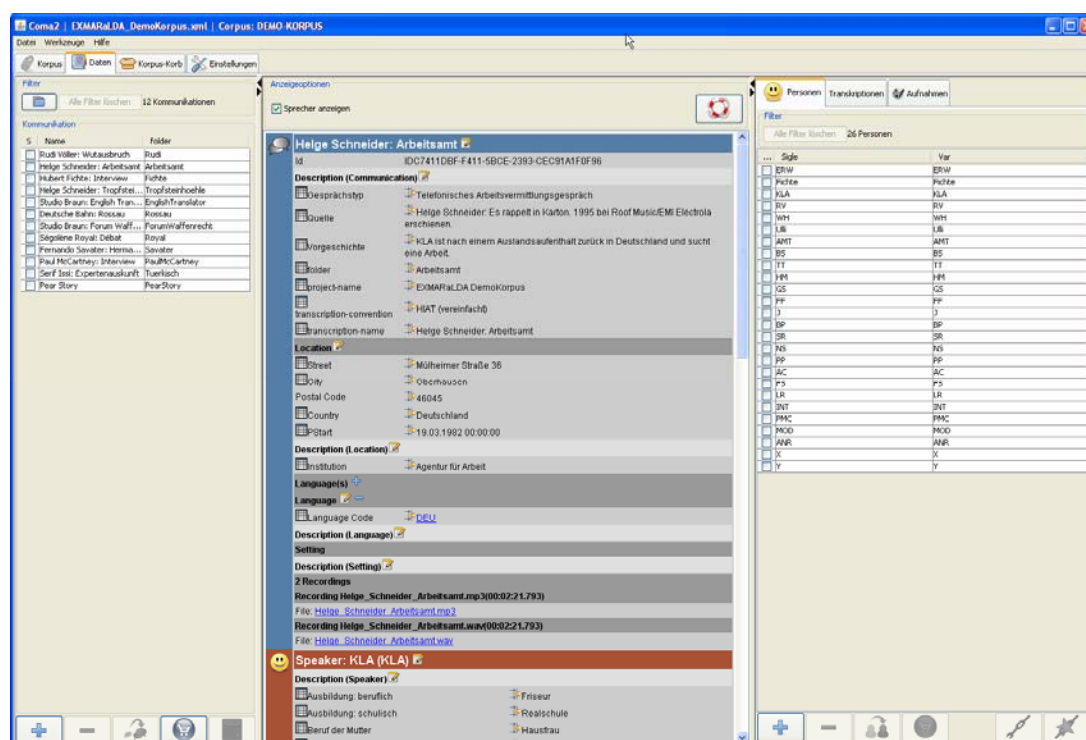
Here too, clicking on an arrow beside an utterance (3) will start the Flash Audio Player. Clicking on a number in square brackets will get you back to the corresponding part of the musical score visualization.

2. Offline use

You can also download an entire corpus for offline use. This is especially useful if you want to edit data yourself or if you want to do corpus queries. To download a corpus, click on the link to the ZIP archive and unpack this archive on your hard disk. This should result in a directory structure like the following:



In the top level directory, there should be an XML corpus file (for the demo corpus, this file has the name EXMARaLDA_DemoKorpus.xml). You can open this file with the EXMARaLDA corpus manager to view, edit or query meta-data...



... or you can open the file with EXMARaLDA's query tool EXAKT to do corpus queries:

EXMARaLDA EXAKT 0.5

File Edit Concordance Help

RECENTLY USED

Corpora

DEMO-KORPUS
T:\IP-2...aLDA_DemoCorpus.xml
12 transcriptions
550 segment chains

Done

Concordances

DEMO-KORPUS
33 tokens
30 types

DEMO-KORPUS (33 results)

#	S	Communication	Speaker	Left Context	Match	Right Context	en
1			Rudi Völter: Wutausbruch	da müssen wir n Gagner auswärts klar beheimen	••Wo in w	••Ihr habt doch früher der Günter, was die früher	what
2			Rudi Völter: Wutausbruch	mal von eurem hohen Ross runterzu runterko	•Was ihr e	•Ihr habt doch früher der Günter, was die früher	what
3			Rudi Völter: Wutausbruch	mal von eurem hohen Ross runterzu runterko	•Was ihr e	•Ihr habt doch früher der Günter, was die früher	what
4			Helge Schneider: Arbeits		Ah, es geh	Sie wissen: die Stellen • sind rar. Und man mus	what
5			Helge Schneider: Arbeits	freier will. Sie wissen: die Stellen • sind rar	Und man	Auch wenn es stinkt da n bißchen.	what
6			Helge Schneider: Arbeits		Ja, aber w		what
7			Helge Schneider: Arbeits		Da muss i		what
8			Helge Schneider: Arbeits		Naj, s'bin i		what
9			Helge Schneider: Arbeits		Och, Sie si		what
10			Helge Schneider: Arbeits		Ohne Luft	Keine Phantasie, was, Herr Gerziten, keine Pha	what
11			Helge Schneider: Tropf	n Mann auf des toten Mannes Kiste. (Ochneidert)	Rach, wad	Oh, da kommt einer. Kommt noch einer. (0,4)	what
12			Helge Schneider: Tropf	Ja, (daß) Stalakken. ((0,2)) Nee, (ja)	Hein, wie	Termen oder so. Die hängen da vonna Decke	what
13			Helge Schneider: Tropf		Ja, sicher	So n Kursus an ner Volkshochschule. Was? Zi	what
14			Helge Schneider: Tropf		Hab ich do		what
15			Studio Braun: English Tra		Hallo Frau		what
16			Studio Braun: English Tra		Da spricht		what
17			Studio Braun: English Tra		Ja ja		what
18			Studio Braun: Forum Wat		Ich hab ab		what
19			Studio Braun: Forum Wat	Da bin ich nun persönlich dran interessiert	Man kann	Stellen Sie sich das mal persönlich vor • bei Ihn	what
20			Ségolène Royal: Débat	fer. ((0,3)) Je ne suis pas sortie de mes gonds.	Je crois qu	((0,6)) Et je pense ((0,7)) qu'il faut que les d	what
21			Fernando Savater: Herm	anos, no tiene sectas. ((0,2)) ((0,1,1,1))	Pues si, es	Pero a lo mejor la gente ya está un poco cansa.	what

Ah, es geht nicht immer uh danach, was der • Arbeitsnehmer will. Sie wissen: die Stellen • sind rar. Und man muss nehmen, was man kommt. Auch wenn es stinkt da n bißchen.

en what

Partitur

AMT [v] Und man muss nehmen, was man kommt. Auch wenn es stinkt da n bißchen.

AMT [en] And you have to take what you get. Even if it's a bit smelly.

KLA [v] Naja, wenn Sie gesundheitlich dadurch gefährdet sind, dann wolln wir ma

KLA [en] Well, if you do have a medical problem, then let's see.

KLA [v] Ja nur dat kann ich auch nich vertragen.

KLA [en] Well yes, but I just can't tolerate that.

Partitur HTML

For further information on how to query an EXMARaLDA corpus, please consult the CoMa and EXAKT documentation. Of course, you can also view and edit individual transcription files with the EXMARaLDA Partitur-Editor.