



How to use an EXMARaLDA corpus

This document explains how to use EXMARaLDA corpora from www.exmaralda.org, www.corpora.uni-hamburg.de and other sites. 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

A. Overview	2
B. Online use in a Browser.....	2
1. Corpus overview	2
2. HTML visualizations	4
C. Offline use.....	5
D. Use as remote corpus in EXAKT	6
1. Open a remote corpus	6
2. Choose a corpus	7
3. Login.....	7

A. Overview

There are several ways to use an EXMARaLDA corpus. Which of these is preferable depends on what you want to do with the data (e.g. read it vs. query it), on where you want to use the data (e.g. on your office computer vs. a university course) and, last but not least, on the specifics of the data themselves. For example, the following usage scenarios are common:

- You want to get an overview of the data in a corpus, listen to the recordings, read the transcriptions and inspect the metadata. For this scenario, it is usually recommended to browse the data online. This is explained in more detail in **section B**.
- You want to do a query on the data, e.g. you are looking for uses of a specific lexical item or all items which have been tagged or annotated in a certain way. For this scenario, you will want to use EXAKT. As explained in more detail in **section C**, EXAKT can access a corpus in three different ways – on a local file system (meaning you will have to download the data first), on a remote file system (meaning you can access the data directly on the server) or on a (remote) relational database.
- You want to make your own changes to the data, e.g. add a specific annotation to it or transfer it for use with some other tool. For this scenario, you will have to download the data. This is explained in more detail in **section D**.

B. Online use in a Browser

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

1. Corpus overview

You usually start with a corpus overview. The corpus overview consists of a list of communications on the left and a list of speakers on the right side.

The screenshot displays the EXMARaLDA Demo Korpus interface. At the top, there is a header bar with the text "EXMARaLDA Demo Korpus" and a small logo. Below the header, the interface is divided into two main sections: "List of communications" on the left and "List of speakers" on the right.

List of communications: This section shows a list of 21 communications, grouped by language. The languages listed are English, French, German, German/English, Italian, Polish, and Russian. Each communication entry includes the title, the number of speakers, recordings, and transcriptions. For example, under English, there are entries for "Beckhams", "Monty Python: My Theory", "Paul McCartney: Interview", and "Pear Story".

List of speakers: This section shows a list of 57 speakers, each with a unique identifier and a name. The speakers are listed in alphabetical order by their identifier. For example, the first few are "AC - Arlette Chabot", "AK - Anton Krasovskij", "AL - Alexander Lebedev", "AMT - Ralf Geritzen", "ANR - Hatice Büyükbaz", "AW - Anne Will", "BP - Bernd Peterchen", "BS - Bernd Schwanmeister", "C - Claudio Rossini", "DAV - David Beckham", "EK - Ursula Engelen-Käfer", "ELK - Ann Elk", "ERW - Erwin Schneider", "FB - Sonja Barthel", "FF - Fiona Frick", "FK - Constanze Kastenhuber", "FP - Frank Plasberg", "FS - Fernando Savater", "Fichte - Hubert Johannes Fichte", "GL - Gottfried Ludewig", "GS - Günter Sachs", "GW - Guido Westerwelle", "HB - Daniel Bahr", and "HC - Hans-Gert...

By clicking on an item in one of these lists, information about the chosen item (the communication “Beckhams”) will be displayed:

Beckhams (3 Speakers, 1 Recording, 1 Transcription)

Background information: Victoria und David Beckham sind zu Gast bei Parkinson
 Communication type: television interview
 Project name: EXMARaLDA DemoKorpus
 Source: Parkinson Talkshow auf BBC

Speakers: PAR; VIC; DAV
 Language: eng

Location
 Date: 2007-11-02
 Country: England
 City: London

Recording
 Recording name: Beckhams
 Duration: 4.4 minutes
 mpg file: [Beckhams.mpg]
 wav file: [Beckhams.wav]
 mp3 file: [Beckhams.mp3]
 webm file: [Beckhams.webm]
 ogg file: [Beckhams.ogg]

Transcription
 Transcription name: Beckhams
 Alignment status: fully aligned
 Segmentation algorithm: HIAT
 Transcriber: Fideniz Ercan
 Transcription convention: HIAT (simplified)
 Transcription date: Juli 2011
 Transcription status: fully transcribed

EXMARaLDA: [Transcription] [Segmented]
 Visualisation: [Partiture] [RTF] [PDF] [Utterances] [Words]
 Export: [TEI] [EAF] [PRAAT] [FOLKER] [AG] [Chat] [Plain text]

Callouts:
 1: Points to the title 'Beckhams (3 Speakers, 1 Recording, 1 Transcription)'.
 2: Points to the 'Speakers' section.
 3: Points to the 'Recording' and 'Transcription' sections.
 3a: Points to the 'EXMARaLDA' section.
 3b: Points to the 'Visualisation' section.
 3c: Points to the 'Export' section.

For communications, the top part gives you a list of metadata items (1). This is followed by a list of speakers participating in the communication (2). By clicking on any of the speakers, the corresponding information will be displayed in the speaker list.

The lower part links to all the documents (recordings, transcriptions, visualisations and export formats) which belong to this communication (3). More specifically:

- The section **EXMARaLDA** (3a) links to an EXMARaLDA *basic transcription* (**Transcription**), which can be opened and edited in the EXMARaLDA Partitur-Editor, and an EXMARaLDA *segmented transcription* (**Segmented**), which is the file format used for querying in EXAKT.¹
- The section **Visualisation** (3b) links to *musical score* visualisations in different formats (**Partiture**, **RTF**, **PDF**), to an *utterance list* (**Utterances**) and a *word list* (**Words**) and – in older corpora – to a separate visualisation of the transcription head (HTML).
- The section **Export** (3c) links to several export formats. **TEI** is an XML file corresponding to the guidelines of the Text Encoding Initiative. **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.

¹ For more information about the two formats, consult “Quickstart_Segmentation” and “How to use segmentation”.


FOLKER is a file which can be opened with the FOLKER transcription editor of the IDS in Mannheim. **AG** is an annotation graph file that can be used for data exchange with various annotation tools. **CHAT** is the file format of the CLAN editor of CHILDES. Or, finally, you can also view the transcription in your browser in **Plain text** format.

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.


2. HTML visualizations

If you display the HTML version (**Partiture**) 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 or, in newer corpora, to a HTML5 audio player:

The screenshot shows the 'Demo Korpus - Beckhams' interface. On the left is a video player showing a woman speaking. On the right is a transcription list with four entries, each with a number in square brackets: [1], [2], [3], and [4]. Each entry has a speaker label (PAR or VIC) and a transcription of speech. A box labeled 'Player' points to the video player. Arrows indicate that clicking on a number in the transcription list (e.g., [1]) or a small play button icon in the top row of the transcription will start playback at the corresponding time in the audio recording.

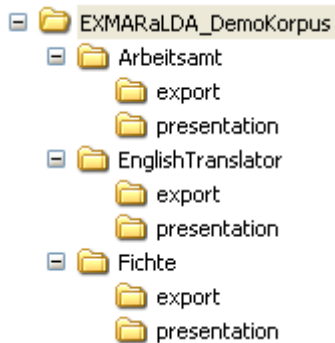
In the musical score, if you click on any one of the little arrows  in the top rows, the player will start playback at the corresponding time in the audio recording. Clicking on any number in the top rows of a musical score will take you to the corresponding place in an utterance list:

The screenshot shows the 'Demo Korpus - Beckhams' interface. On the left is a video player showing a man speaking. On the right is a transcription list with several entries, each with a number in square brackets: [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100]. Each entry has a speaker label (PAR or VIC) and a transcription of speech. A box labeled 'Player' points to the video player. Arrows indicate that clicking on a number in the transcription list (e.g., [1]) or a small play button icon in the top row of the transcription will start playback at the corresponding time in the audio recording.

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

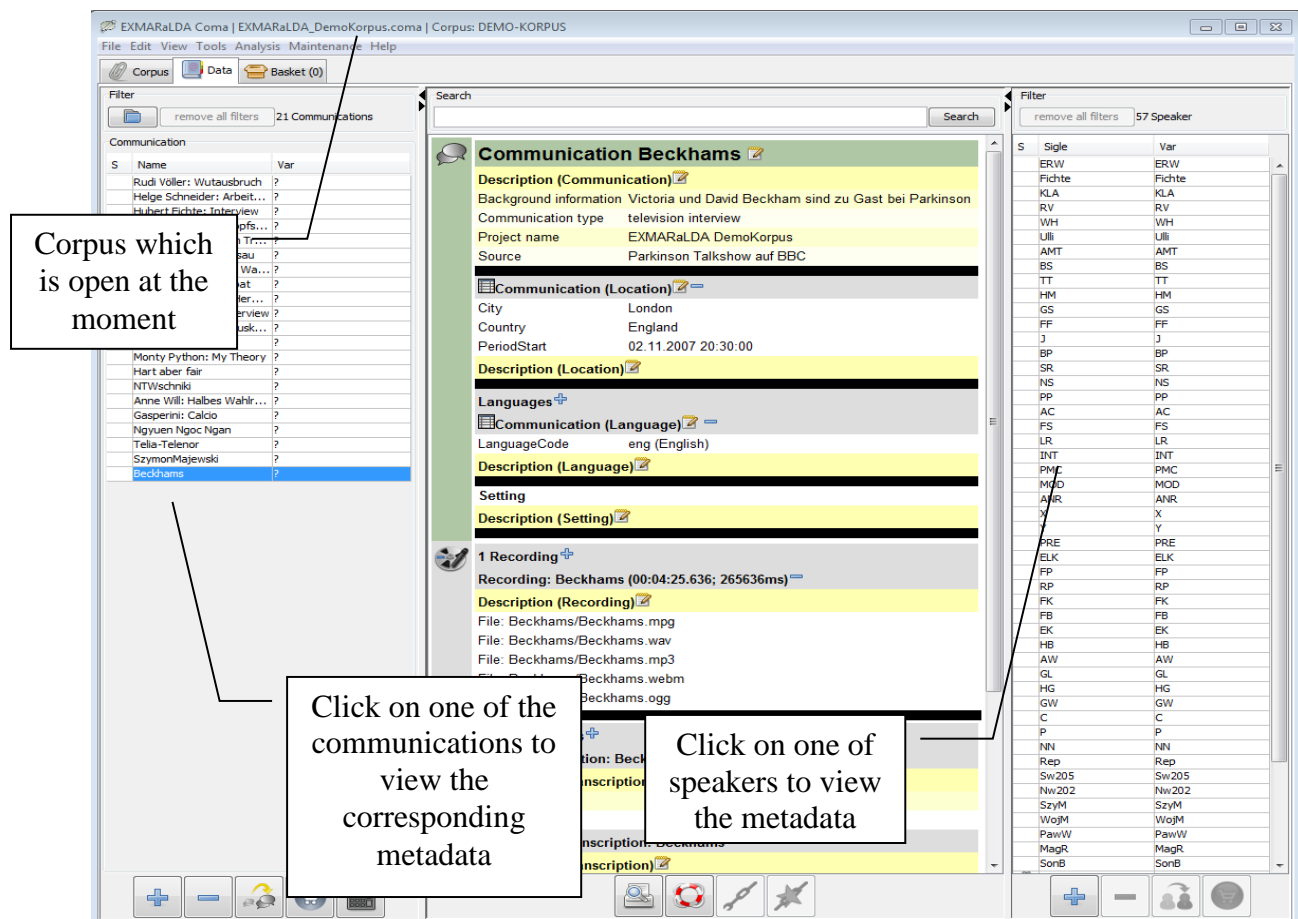
C. Offline use

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 (from: www.exmaralda.org) 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 a corpus file with the suffix ".coma" (for the demo corpus, this file is named EXMARaLDA_DemoKorpus.coma²).

The file can be opened with the Corpus Manager (Coma) to view, edit or query meta-data...



... or you can open it also with EXMARaLDA's query tool EXAKT to carry out corpus queries (from the menu **File > Open corpus...**).

² Older corpora use the suffix ".xml" instead of ".coma"

The screenshot displays the EXMARaLDA EXAKT 1.2 software interface. The main window shows a concordance search for the word "the". The search results are displayed in a table with columns for Speaker, Left Context, Match, and Right Context. The search term "the" is highlighted in red in the Match column. The interface also includes a sidebar with a tree view of corpora, a word list, and a concordance list. A callout box labeled "Query" points to the search bar. Another callout box labeled "Here you can see which corpora are open" points to the corpora tree. A third callout box labeled "Wordlists" points to the word list. A fourth callout box labeled "Concordance(s)" points to the concordance list. A fifth callout box labeled "Partiture view in EXAKT" points to the bottom window showing a transcription file.

Of course, you can also view and edit individual transcription files (you'll find those inside the subfolders, they have the suffix ".exb" or – in older corpora ".xml") with the EXMARaLDA Partitur-Editor.



For further information on how to query an EXMARaLDA corpus, please consult the Coma and EXAKT documentation.

D. Use as remote corpus in EXAKT

EXAKT offers the possibility to do queries on a remote corpus, i.e. on a corpus that is not located on your own computer, but on a remote server. Note that you normally need a broadband connection for this feature to work satisfactorily (other connections are too slow).

1. Open a remote corpus

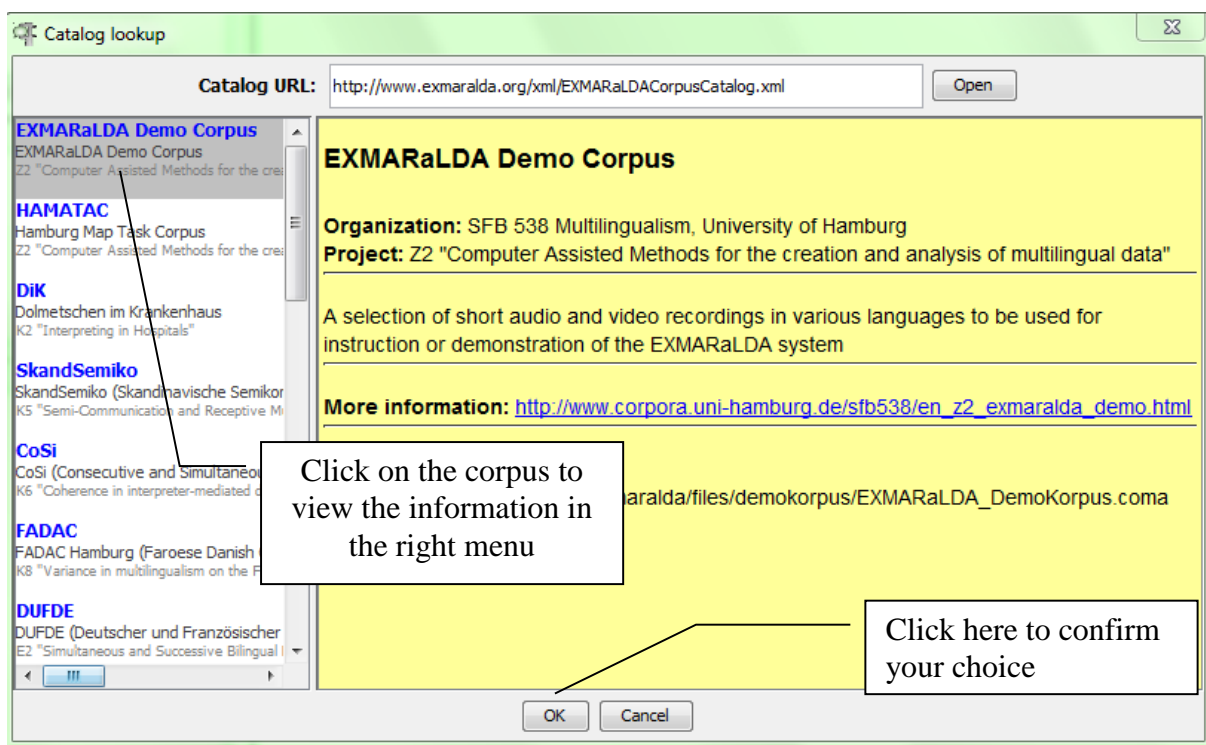
Start EXAKT and choose **File > Open remote corpus...**

The screenshot shows the "Open Remote Corpus" dialog box. It features a central area with a book icon and the text "Catalog lookup...". Below this, there are input fields for "Corpus URL:", "Username:", and "Password:". There is also a checkbox labeled "Anonymous Login". At the bottom, there are "OK" and "Cancel" buttons.

In the following dialog, you have to click on the book symbol to browse through available corpora.

2. Choose a corpus

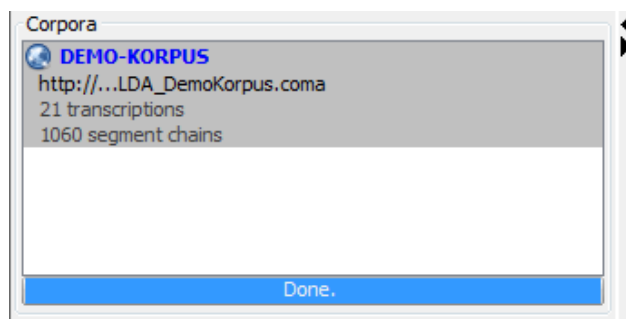
A window with the corpora from the SFB ‘Multilingualism’ will appear automatically. Choose an entry (left menu) and click **OK** – the correct URL will be entered for you:



3. Login

After you have chosen the corpus, a dialog with login request (just like in step 1 “Open a remote corpus”) will appear.

- If the corpus is password protected, you have to enter your username and password in the fields **Username** and **Password**, respectively and then click **OK**. Please note: Check the website www.corpora.uni-hamburg.de³ (Section “Ressourcen und Projekte” > “Korpora”) for access permission and/or contact information.
- If the corpus is not password-protected (as is the case with the EXMARaLDA demo corpus), tick the box **Anonymous login** to disable user authentication. Clicking on **OK** will open the remote corpus and display it in the corpus list in the left upper corner of the EXAKT window (compare also: screenshot in **section C**):



You can then work with this corpus in the usual way, i.e. as if it was on your local computer. Note, however, that *media playback* for remote corpora is currently only supported on Windows. For Macintosh and Linux computers, media playback only works for local corpora.



³ This website is available in German only.