Thomas Schmidt

EXMARaLDA
Analyse & Konkordanz Tool
EXAKT 0.4
Developed by Thomas Schmidt
(C)SFB 538 Mehrsprachigkeit
http://www.exmaralda.org/exakt
This program is licensed under a open source license

A short introduction to EXAKT

Version 0.4., August 2008

INTRODUCTION

EXAKT – the “EXMARaLDA Analysis and Concordancing Tool” – is a tool for searching and analysing corpora of spoken language transcriptions as created by the EXMARaLDA Partitur-Editor and the EXMARaLDA Corpus Manager. EXAKT’s base functionality is that of a concordancer – like WordSmith, MonoConc etc., it lets you enter a search expression and outputs all the instances which match this expression plus a bit of the preceding and the following context. On top of this base functionality, EXAKT enables you to

- display more interactional context as encoded in the transcription (e.g. things which other people said around the same time as the utterance which the search expression matched),
- display situational context in the form of metadata about the communication in question,
- display metadata about speakers,
- listen to the corresponding part of the transcribed (audio or video) recording,
- filter your search results according to various criteria,
- add one or more analyses to your list of search result,
- save, retrieve, combine, output search results and export them to other applications (e.g. Excel, SPSS) for further analysis

This document explains the basic functionality of EXAKT.

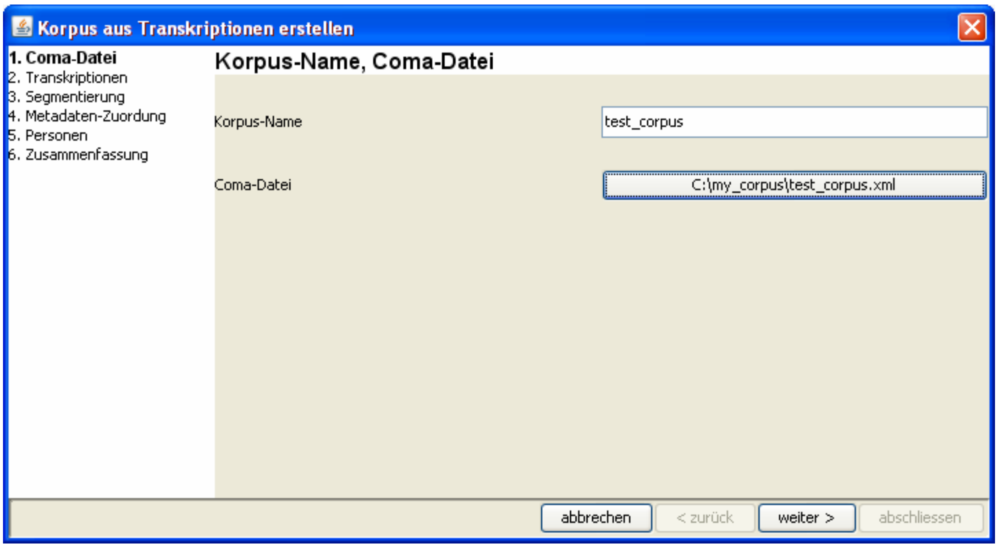
If you are new to EXAKT (and maybe to EXMARaLDA and/or concordancers in general), we recommend that you download the EXMARaLDA demo corpus from http://www.exmaralda.org/en_demokorpus.html and experiment with that before using EXAKT with your own data.

1. OPENING OR GENERATING A CORPUS

If you want to use EXAKT, you need an EXMARaLDA corpus which contains segmented transcriptions (you can create these in the Partitur-Editor with various options in the "Segmentation" menu). Maybe you are already using the EXMARaLDA corpus manager to manage your corpus and know what a segmented transcription is. In that case, all you have to do to get started is choose "File > Open corpus…" from EXAKT's menu and select your corpus file (usually a file whose name ends in ".xml"). (Skip the rest of this chapter if you already have a corpus).

For the EXMARaLDA demo corpus, the corpus file is the file "EXMARaLDA_DemoKorpus.xml" in the top level directory.

If all you have is a list of basic transcriptions created with the Partitur-Editor, EXAKT offers you an easy way to turn those into a corpus. You first have to make sure that all your basic transcriptions are underneath a single folder in your file system. Let's assume this folder is called "c:\my_corpus". You can then choose "File > Generate corpus…" from EXAKT's menu. This will start the corpus creation wizard.

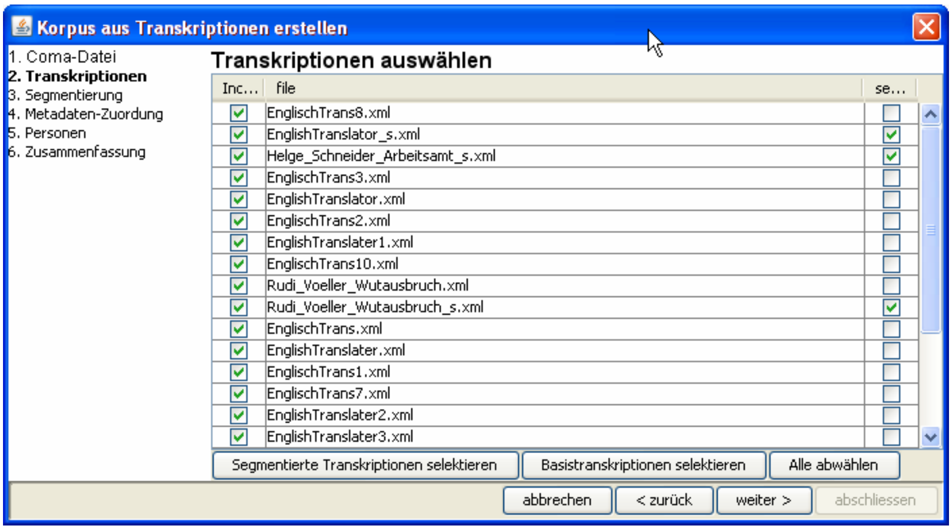


Korpus-Name, Coma-Datei

Korpus-Name: test_corpus

Coma-Datei: C:\my_corpus\test_corpus.xml

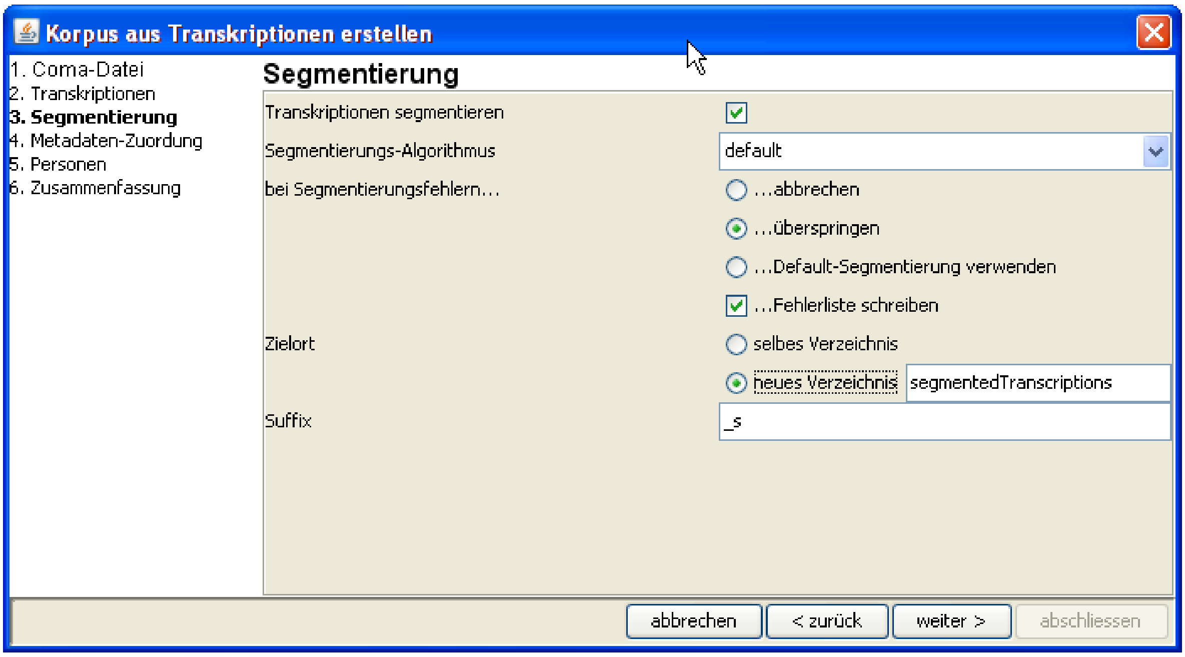
You are asked to enter a name for your corpus ("Korpus-Name") and to choose a path for the corpus manager file ("Coma-Datei"). For the latter, choose the folder underneath which your corpus data can be found (i.e. "c:\my_corpus" in our example). The wizard will then automatically scan this folder for any transcriptions contained in it and its subfolders. When you click on "weiter >" (= "continue"), these transcriptions will be displayed in a table.



Transkriptionen auswählen

Inc...	file	se...
☑	EnglischTrans8.xml	□
☑	EnglishTranslator_s.xml	☑
☑	Helge_Schneider_Arbeitsamt_s.xml	☑
☑	EnglischTrans3.xml	□
☑	EnglishTranslator.xml	□
☑	EnglischTrans2.xml	□
☑	EnglishTranslater1.xml	□
☑	EnglischTrans10.xml	□
☑	Rudi_Voeller_Wutausbruch.xml	□
☑	Rudi_Voeller_Wutausbruch_s.xml	☑
☑	EnglischTrans.xml	□
☑	EnglishTranslater.xml	□
☑	EnglischTrans1.xml	□
☑	EnglischTrans7.xml	□
☑	EnglishTranslater2.xml	□
☑	EnglishTranslater3.xml	□

The second column of this table contains the transcription's name. The first column tells you whether or not it will be included in the corpus, and the last column tells you whether it is a segmented transcription (if it isn't, it is a basic transcription). You can change the selections in the first column according to which transcriptions you want included in your corpus. If you're done, click on "weiter >".



Segmentierung

Transkriptionen segmentieren ☑

Segmentierungs-Algorithmus default

bei Segmentierungsfehlern...
○ ...abbrechen
● ...überspringen
○ ...Default-Segmentierung verwenden
☑ ...Fehlerliste schreiben

Zielort
○ selbes Verzeichnis
● neues Verzeichnis segmentedTranscriptions

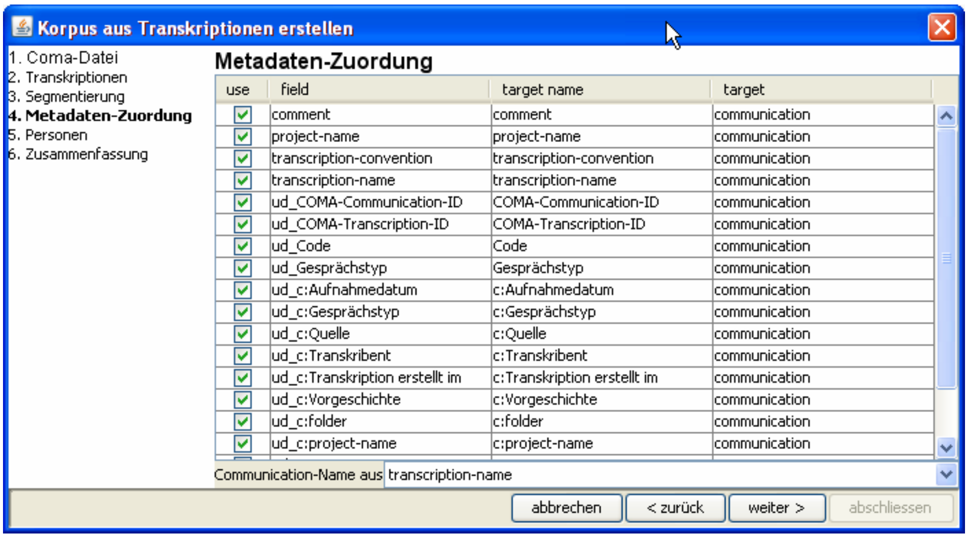
Suffix _s

The next dialog is about creating a segmented version of each of the selected basic transcriptions. If you want to keep things simple, you should choose the following parameters in this dialog:

- Tick the box "Transkriptionen segmentieren" ("segment transcriptions"); this tells the wizard that transcriptions are to be segmented;
- for "Segmentierungs-Algorithmus" ("segmentation algorithm"), choose "default";
- since the default segmentation algorithm never produces any errors, it is not important what you choose under "bei Segmentierungsfehlern" ("in case of segmentation errors");
- for "Zielort" ("target location") choose "neues Verzeichnis" ("new folder"); this tells the wizard to write the resulting segmented transcriptions into a new folder rather than place them side-by-side with the original basic transcriptions;

- finally, choose “_s” as a suffix for the newly created segmented transcriptions; choosing a (i.e. any) suffix will make sure that basic and segmented transcriptions have systematically different names.

When you’re done specifying the segmentation parameters, click “weiter >”.

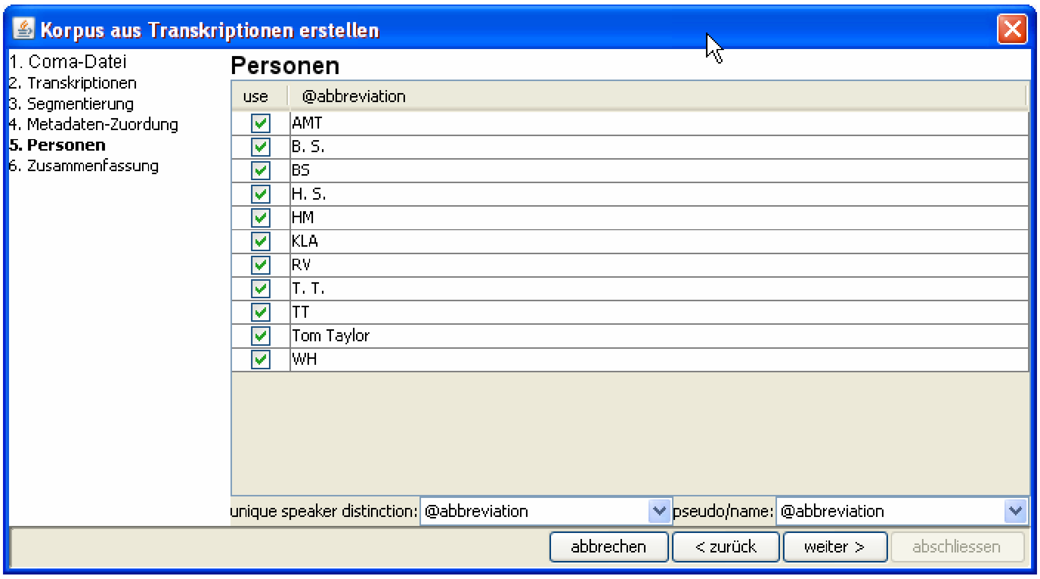


Metadaten-Zuordung

use	field	target name	target
☑	comment	comment	communication
☑	project-name	project-name	communication
☑	transcription-convention	transcription-convention	communication
☑	transcription-name	transcription-name	communication
☑	ud_COMA-Communication-ID	COMA-Communication-ID	communication
☑	ud_COMA-Transcription-ID	COMA-Transcription-ID	communication
☑	ud_Code	Code	communication
☑	ud_Gesprächstyp	Gesprächstyp	communication
☑	ud_c:Aufnahmedatum	c:Aufnahmedatum	communication
☑	ud_c:Gesprächstyp	c:Gesprächstyp	communication
☑	ud_c:Quelle	c:Quelle	communication
☑	ud_c:Transkribent	c:Transkribent	communication
☑	ud_c:Transkription erstellt im	c:Transkription erstellt im	communication
☑	ud_c:Vorgeschichte	c:Vorgeschichte	communication
☑	ud_c:folder	c:folder	communication
☑	ud_c:project-name	c:project-name	communication

Communication-Name aus: transcription-name

The next dialog is about metadata you have entered in your transcriptions (i.e. with the help of “File > Meta information” in the EXMARaLDA Partitur-Editor). For each such metadata field (second column), the wizard asks you whether to include it in your corpus (first column), what to call it in the corpus (third column), and whether to assign it to a communication or to a transcription (fourth column). The most important field is at the bottom of the table: it lets you specify how the wizard determines the name of communications and how it assigns transcriptions to communications. Click on “weiter >” when you have specified everything.

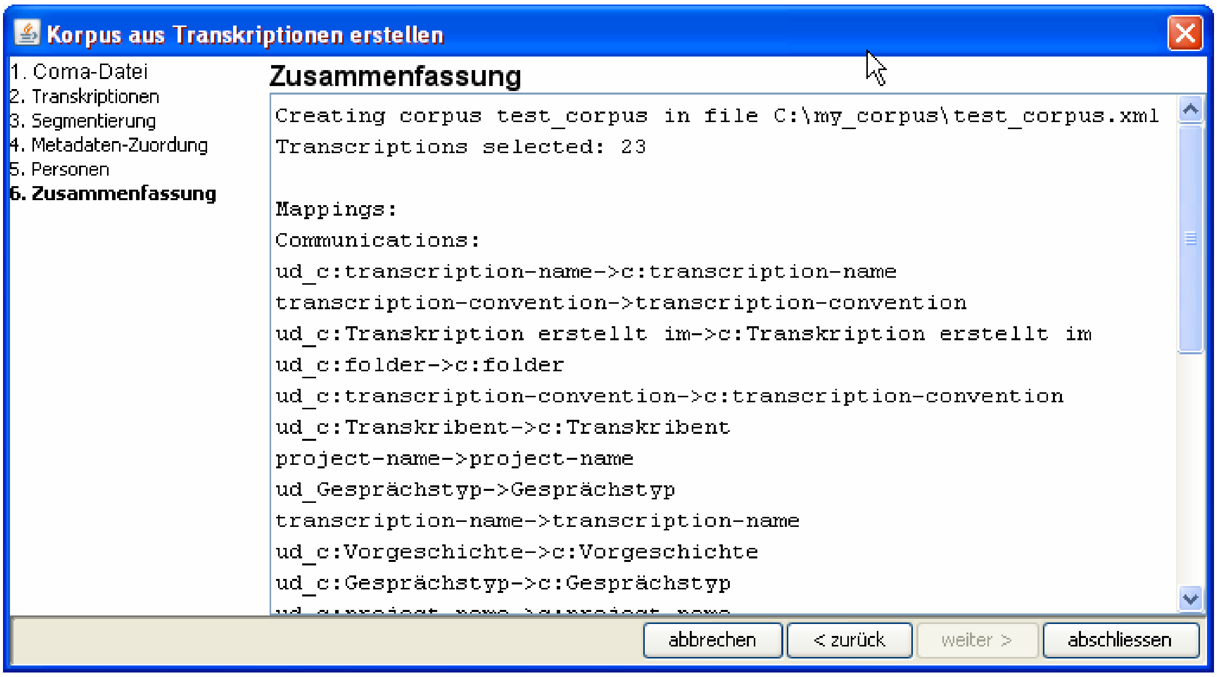


Personen

use	@abbreviation
☑	AMT
☑	B. S.
☑	BS
☑	H. S.
☑	HM
☑	KLA
☑	RV
☑	T. T.
☑	TT
☑	Tom Taylor
☑	WH

unique speaker distinction: @abbreviation pseudo/name: @abbreviation

The last dialog is about the speakers of the corpus. The wizard asks you for a “unique speaker distinction” at the bottom of the dialog. Usually, you will have chosen abbreviations in the EXMARaLDA transcriptions to be unique for each speaker (i.e. no two different speakers will share the same abbreviation). Only if this is not the case, do you need to specify a different unique speaker distinction. Clicking on “weiter >” will get you to a summary of the parameters you have set for the wizard:



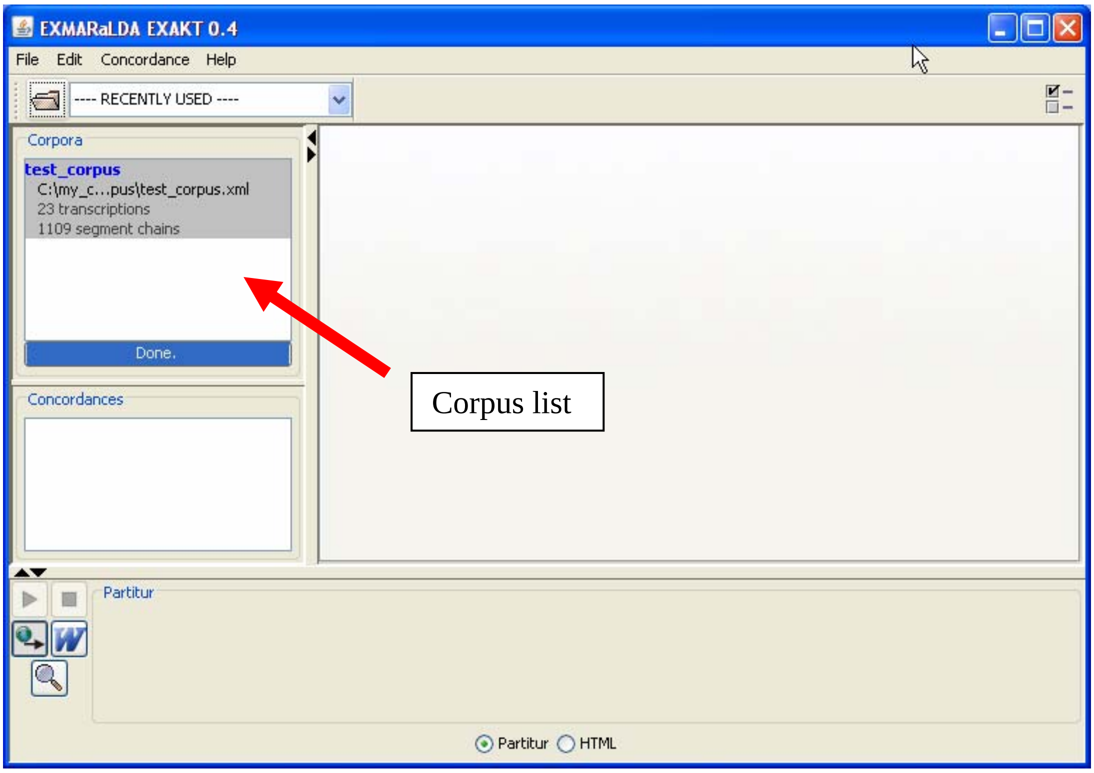
```
Creating corpus test_corpus in file C:\my_corpus\test_corpus.xml
Transcriptions selected: 23

Mappings:
Communications:
ud_c:transcription-name->c:transcription-name
transcription-convention->transcription-convention
ud_c:Transkription erstellt im->c:Transkription erstellt im
ud_c:folder->c:folder
ud_c:transcription-convention->c:transcription-convention
ud_c:Transkribent->c:Transkribent
project-name->project-name
ud_Gesprächstyp->Gesprächstyp
transcription-name->transcription-name
ud_c:Vorgeschichte->c:Vorgeschichte
ud_c:Gesprächstyp->c:Gesprächstyp
```

If you now click on “abschliessen” (“finish”) your corpus will be created and loaded in EXAKT.

2. CREATING A NEW CONCORDANCE

Once you've successfully opened or generated a corpus, EXAKT will display this corpus in the left upper corner of the screen:

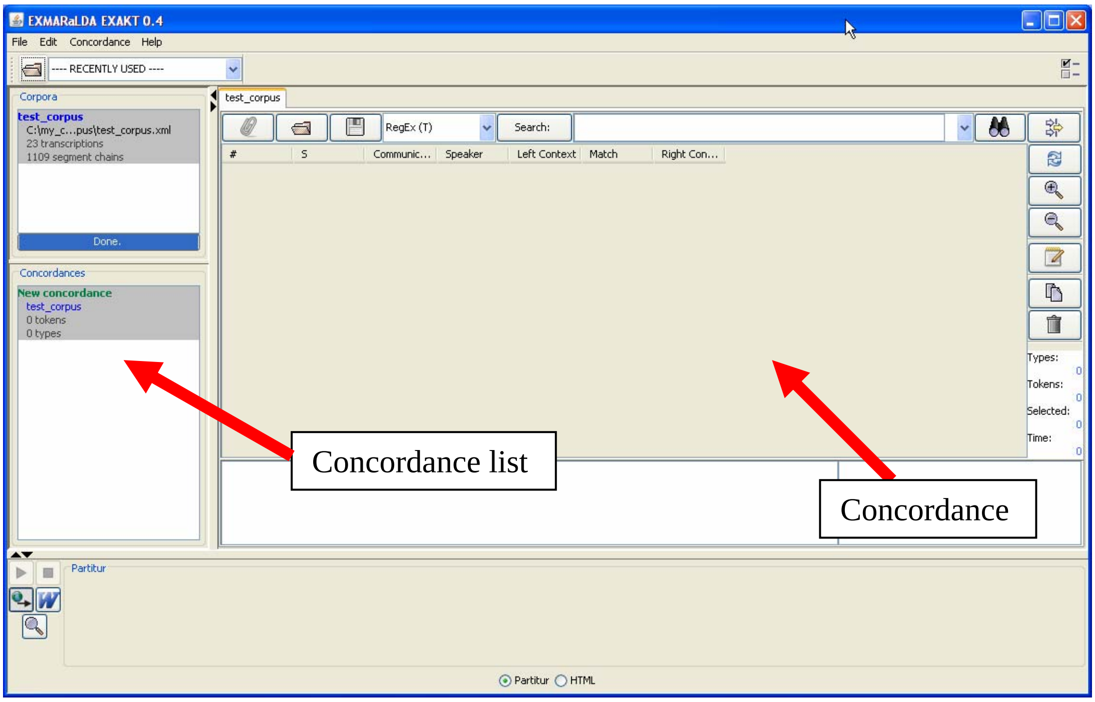


test_corpus
C:\my_c...pus\test_corpus.xml
23 transcriptions
1109 segment chains

Corpus list

It will also tell you how many transcriptions and how many segment chains the corpus contains. Note that you can have several different corpora opened in this list.

To create a new concordance for a given corpus, make sure that this corpus is selected in the corpus list, then click on "Concordance > New concordance".



test_corpus
C:\my_c...pus\test_corpus.xml
23 transcriptions
1109 segment chains

New concordance
test_corpus
0 tokens
0 types

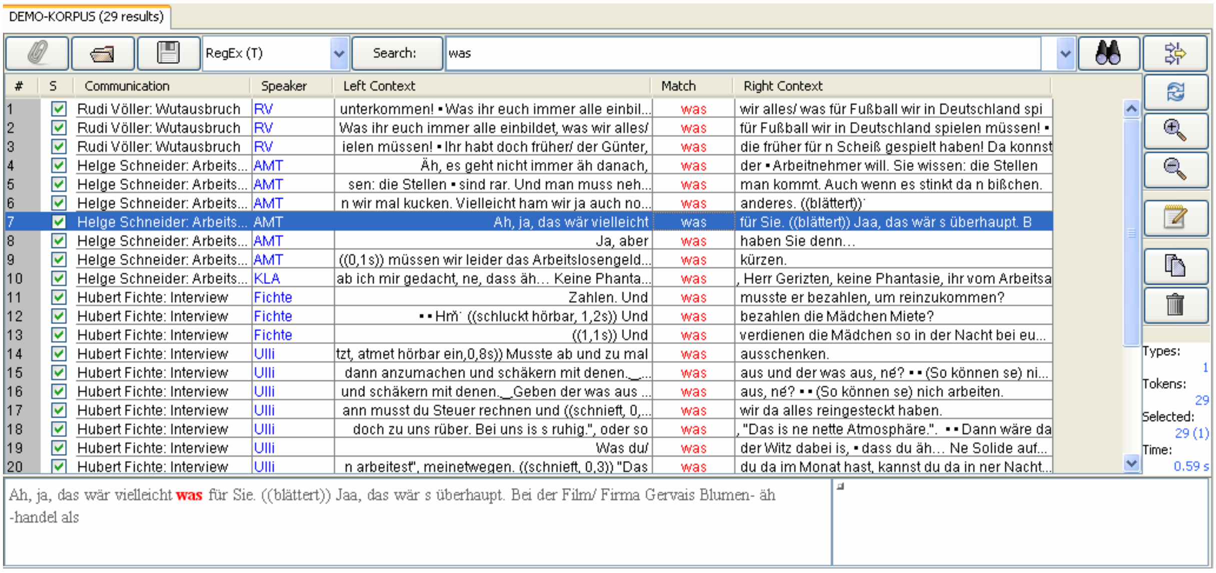
Concordance list

Concordance

You now have a concordance for this corpus which is also shown in the concordance list underneath the corpus list. Note that you can have several concordances for one and the same corpus. A concordance consists of

- a part for entering search expressions (upper part of the concordance window)
- a part for displaying the KWIC (keyword in context concordance) table (centre of the concordance window)
- and a part for displaying additional context (lower part of the concordance window)

To start, enter a simple, frequent word (e.g. “the” for an English corpus, “was” for a German corpus) in the field beside the “Search button” and hit the <Enter> key. You will be given a KWIC concordance displaying all the places in the corpus which match your word.



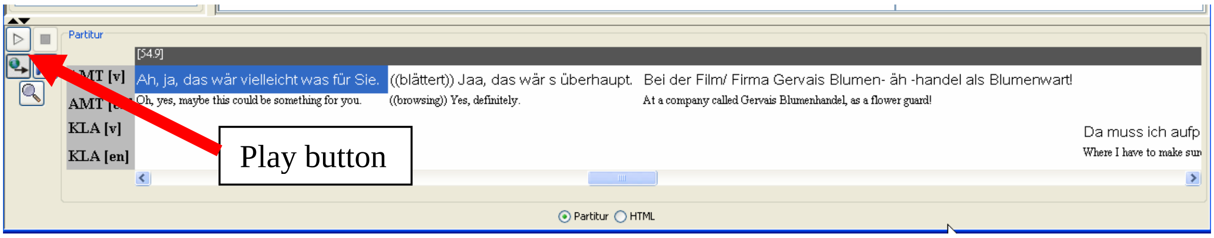
#	S	Communication	Speaker	Left Context	Match	Right Context
1	☑	Rudi Völler: Wutausbruch	RV	unterkommen! • Was ihr euch immer alle einbil...	was	wir alles/ was für Fußball wir in Deutschland spi
2	☑	Rudi Völler: Wutausbruch	RV	Was ihr euch immer alle einbildet, was wir alles/	was	für Fußball wir in Deutschland spielen müssen! •
3	☑	Rudi Völler: Wutausbruch	RV	ielen müssen! • Ihr habt doch früher/ der Günter,	was	die früher für n Scheiß gespielt haben! Da konnst
4	☑	Helge Schneider: Arbeits...	AMT	Äh, es geht nicht immer äh danach,	was	der • Arbeitnehmer will. Sie wissen: die Stellen
5	☑	Helge Schneider: Arbeits...	AMT	sen: die Stellen • sind rar. Und man muss neh...	was	man kommt. Auch wenn es stinkt da n bißchen.
6	☑	Helge Schneider: Arbeits...	AMT	n wir mal kucken. Vielleicht ham wir ja auch no...	was	anderes. ((blättert))˙
7	☑	Helge Schneider: Arbeits...	AMT	Ah, ja, das wär vielleicht	was	für Sie. ((blättert)) Jaa, das wär s überhaupt. B
8	☑	Helge Schneider: Arbeits...	AMT	Ja, aber	was	haben Sie denn...
9	☑	Helge Schneider: Arbeits...	AMT	((0,1s)) müssen wir leider das Arbeitslosengeld...	was	kürzen.
10	☑	Helge Schneider: Arbeits...	KLA	ab ich mir gedacht, ne, dass äh... Keine Phanta...	was	, Herr Gerizten, keine Phantasie, ihr vom Arbeitsa
11	☑	Hubert Fichte: Interview	Fichte	Zahlen. Und	was	musste er bezahlen, um reinzukommen?
12	☑	Hubert Fichte: Interview	Fichte	• • Hm̃˙ ((schluckt hörbar, 1,2s)) Und	was	bezahlen die Mädchen Miete?
13	☑	Hubert Fichte: Interview	Fichte	((1,1s)) Und	was	verdienen die Mädchen so in der Nacht bei eu...
14	☑	Hubert Fichte: Interview	Ulli	tzt, atmet hörbar ein,0,8s)) Musste ab und zu mal	was	ausschenken.
15	☑	Hubert Fichte: Interview	Ulli	dann anzumachen und schäkern mit denen._...	was	aus und der was aus, né? • • (So können se) ni...
16	☑	Hubert Fichte: Interview	Ulli	und schäkern mit denen._Geben der was aus ...	was	aus, né? • • (So können se) nich arbeiten.
17	☑	Hubert Fichte: Interview	Ulli	ann musst du Steuer rechnen und ((schnieft, 0,...	was	wir da alles reingesteckt haben.
18	☑	Hubert Fichte: Interview	Ulli	doch zu uns rüber. Bei uns is s ruhig.", oder so	was	, "Das is ne nette Atmosphäre.". • • Dann wäre da
19	☑	Hubert Fichte: Interview	Ulli	Was du/	was	der Witz dabei is, • dass du äh... Ne Solide auf...
20	☑	Hubert Fichte: Interview	Ulli	n arbeitest", meinetwegen. ((schnieft, 0,3)) "Das	was	du da im Monat hast, kannst du da in ner Nacht...

Ah, ja, das wär vielleicht **was** für Sie. ((blättert)) Jaa, das wär s überhaupt. Bei der Film/ Firma Gervais Blumen- äh -handel als

The KWIC concordance contains the following information:

- column 1 simply counts line numbers for better orientation
- column 2 tells you whether the search result in this row is selected or not
- column 3 gives you the communication in which the search result was found
- column 4 gives you the speaker of the utterance in question
- columns 6 contains the actual search result, i.e. the transcribed text which matched your search expression
- columns 5 and 7 contain the left and right context of that search result

You can sort the table by clicking on any column header. Selecting one search result will also display the corresponding text in the lower left corner of the concordance window. If you double click on a search result, the corresponding transcription will be opened in the lower part of the screen:



Ah, ja, das wär vielleicht was für Sie. ((blättert)) Jaa, das wär s überhaupt. Bei der Film/ Firma Gervais Blumen- äh -handel als Blumenwart!

Oh, yes, maybe this could be something for you. ((browsing)) Yes, definitely. At a company called Gervais Blumenhandel, as a flower guard!

Da muss ich aufp

Where I have to make sur

Play button

If your transcription is aligned with an audio or video file, you can use the play button to playback the corresponding part of the recording.

3. SEARCH EXPRESSIONS

Search expressions can be more than simple strings. In order to find complex patterns in the corpus, you can use regular expressions as search expressions. A regular expression is a text pattern consisting of ordinary characters and meta-characters which is matched against simple strings. Here are some examples:

- The pattern “[Ww]as” will match the strings “was” and “Was”.
- The pattern “komm.{1,2}“ will match “komme”, “kommst”, “kommen”, “komma”, “kommun” etc.
- The pattern „([Ii]ch|[Dd]u)” will match “ich”, “Ich”, “du” and “Du”
- The pattern „\bge[A-Za-z]+?t\b” will match “gemacht”, “gesagt”, “gewusst”, “geht” etc.

By combining regular expressions in various ways, you can formulate rather complex queries with them. The exact syntax and usage of regular expressions is explained at

http://java.sun.com/javase/6/docs/api/java/util/regex/Pattern.html

4. TO BE CONTINUED…

We hope that this little document helps you to get started with EXAKT. We’re working on a more thorough documentation of the tool. Meanwhile, if you have questions regarding the use of EXAKT, please subscribe to the EXMARaLDA mailing list (see http://www.exmaralda.org/en_kontakt.html).