

# EXMARaLDA EXAKT

## Manual

Version 1.0

Last update: 26 November 2010

Thomas Schmidt

INTRODUCTION .....	3
1. OPENING OR GENERATING A CORPUS .....	4
1.1. Opening an existing corpus .....	4
1.2. Opening a remote corpus.....	4
1.3. Generating a word list .....	5
1.4. Generating a corpus from EXMARaLDA transcriptions .....	6
1.5. Generating a corpus from FOLKER transcriptions .....	9
1.6. Generating a corpus from CHAT transcriptions .....	10
1.7. Generating a corpus from ELAN annotation files.....	10
1.8. Generating a corpus from Transcriber files.....	10
2. WORKING WITH CONCORDANCES .....	11
2.1. Creating a new concordance .....	11
2.2. Understanding concordances.....	12
2.3. Going from a search result to the transcription.....	13
2.4. Outputting and saving search results.....	15
3. SEARCH EXPRESSIONS .....	17
3.1. Regular Expressions.....	17
3.2. XPath Expressions.....	20
4. DISPLAYING METADATA .....	21
5. ADDING ANALYSIS COLUMNS.....	23
6. FILTERING SEARCH RESULTS.....	25
7. USING WORD LISTS.....	27
8. DIFFERENT TYPES OF SEARCHES.....	29
8.1. Regular Expression Search over Transcription tiers [RegEx (T)].....	29
8.2. Regular Expression Search over Annotation tiers [RegEx (A)] .....	29
8.3. Regular Expression Search over Description tiers [RegEx (D)] .....	29
8.4. XPath Search over Transcription tiers [XPath (D)].....	29
8.5. Multilevel searches.....	29

## 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 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](http://www.exmaralda.org/en_demokorpus.html) and experiment with that before using EXAKT with your own data.

## 1. OPENING OR GENERATING A CORPUS

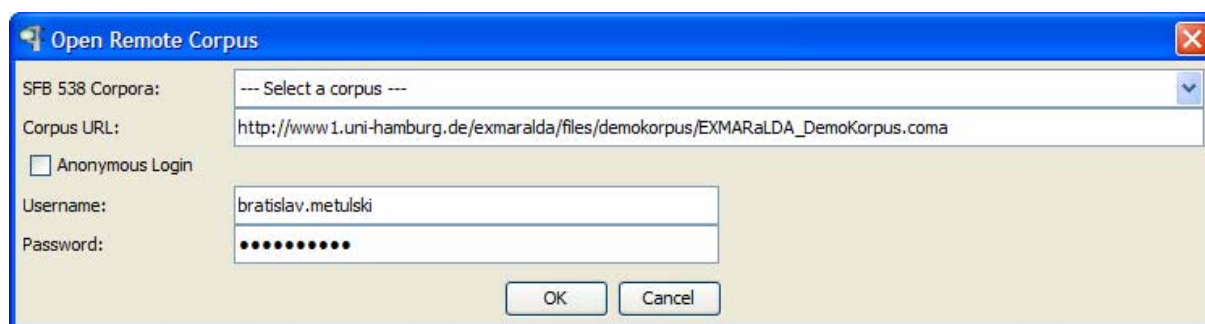
### 1.1. Opening an existing corpus

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 ".coma").

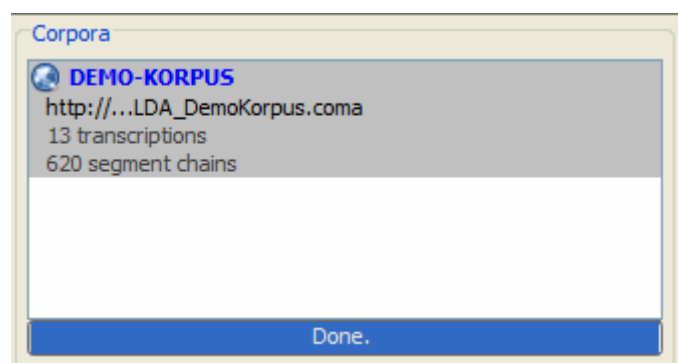
For the EXMARaLDA demo corpus, the corpus file is the file "EXMARaLDA\_DemoKorpus.coma" in the top level directory.

### 1.2. Opening a remote corpus

Whereas choosing **File > Open corpus...** will open a corpus whose COMA and transcription files are on your local computer, you can use **File > Open remote corpus...** to access a corpus and its transcriptions on a server. This saves you from having to transfer large amounts of data onto your computer. Note, however, that you need a fast internet connection to access a remote corpus, and that media playback for remote corpora currently only works on Windows.



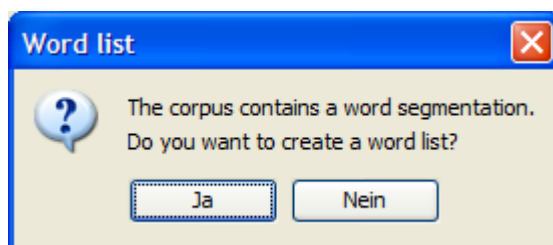
In the following dialog, you have to enter the URL address of the corpus' COMA file in the field **Corpus URL**. For corpora from the SFB 'Multilingualism', you can simply choose an entry from the field **SFB 538 Corpora** and the correct URL will be entered for you. If the corpus is password protected, you have to enter your username and password in the fields **Username** and **Password**, respectively. 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.



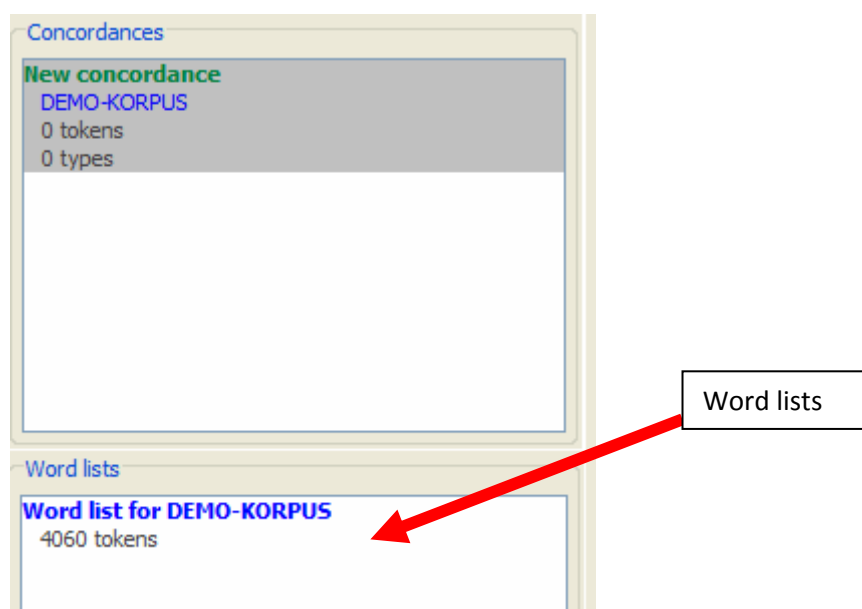
You can then work with this corpus in the usual way, i.e. as if it was on your local computer.

### 1.3. Generating a word list

After a corpus has been read and indexed, EXAKT checks whether the transcriptions in it have been segmented for words. This is the case if an appropriate segmentation algorithm has been used for the generation of segmented transcriptions. For instance, all HIAT corpora available from the SFB 538, including the DEMO corpus, have this kind of word segmentation. If a corpus has been segmented for words, EXAKT asks you whether you want to create a word list for it.



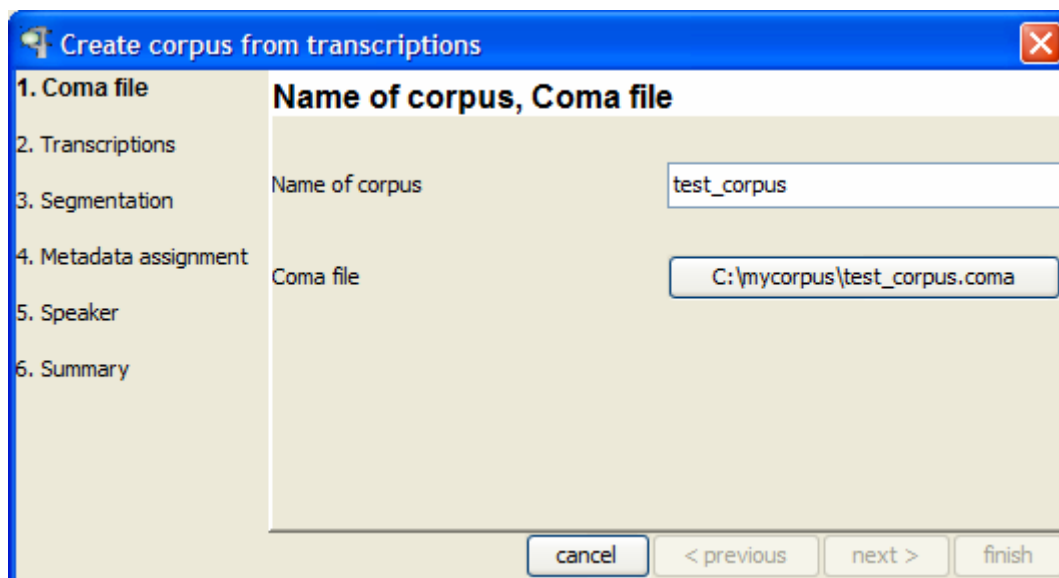
If you choose **Yes**, the word list will be created and displayed in the word list section of EXAKT.



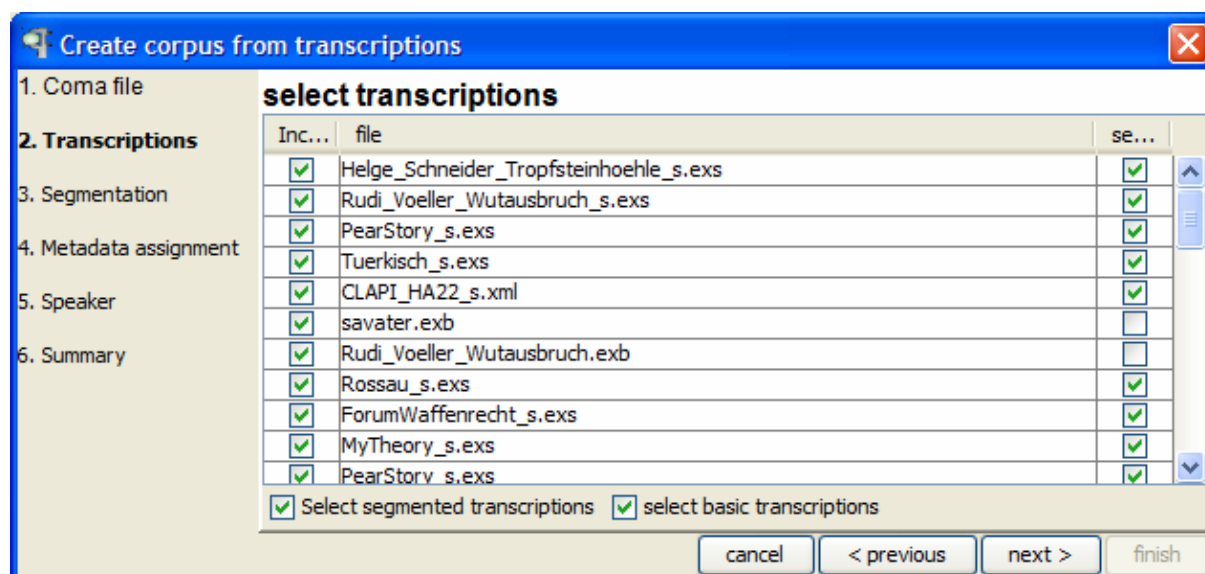
Double click on the word list to display it. See below for an explanation of how to use word lists.

## 1.4. Generating a corpus from EXMARaLDA transcriptions

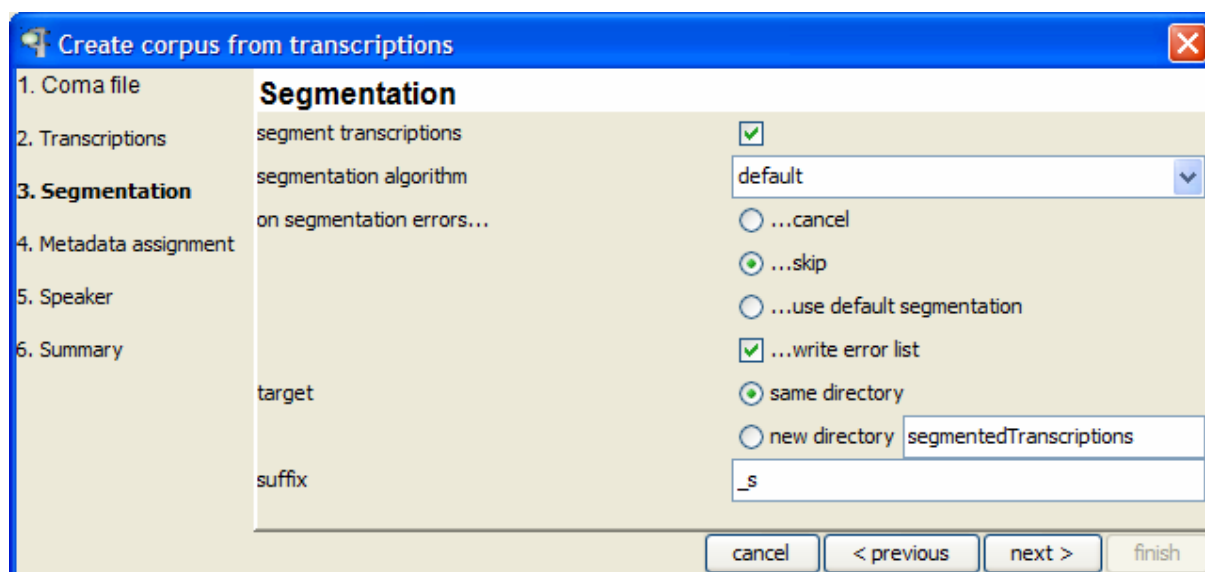
If all you have is a set of basic transcriptions created with the Partitur-Editor, EXAKT offers you an easy way to turn those into a corpus. You first have to make sure that all your basic transcriptions are underneath a single folder in your file system. Let's assume this folder is called "c:\my\_corpus".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



You are asked to enter a name for your corpus (**Name of corpus**) and to choose a path for the corpus manager file (**Coma file**). 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 **next >**, these transcriptions will be displayed in a table.



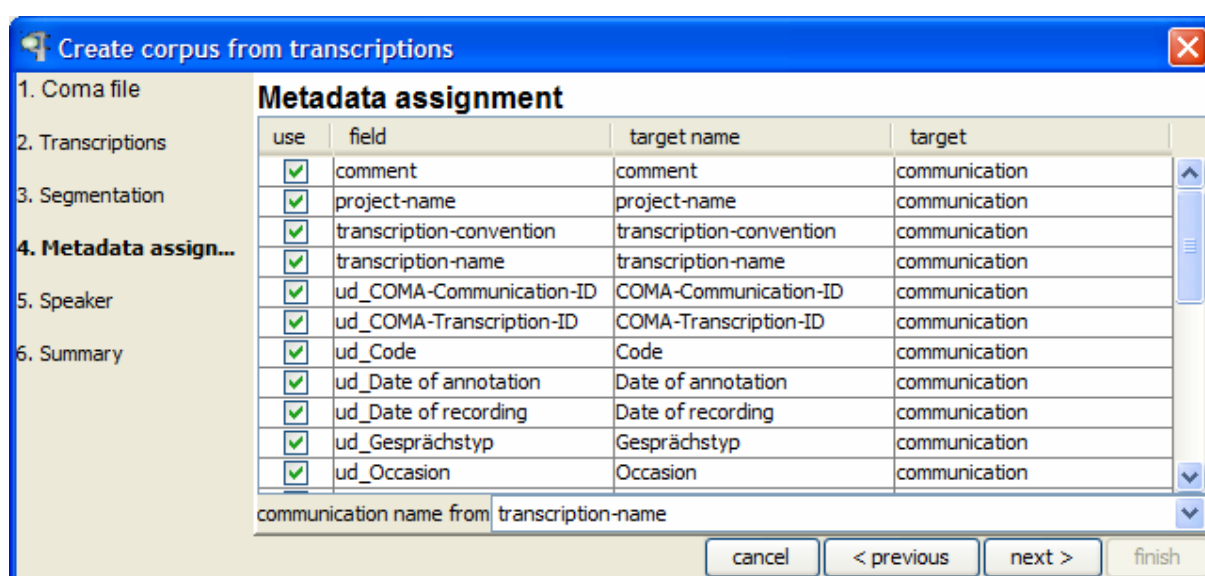
The second column of this table contains the transcription's name. The first column tells you whether or not it will be included in the corpus, and the last column tells you whether it is a segmented transcription (if it isn't, it is a basic transcription). You can change the selections in the first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



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:

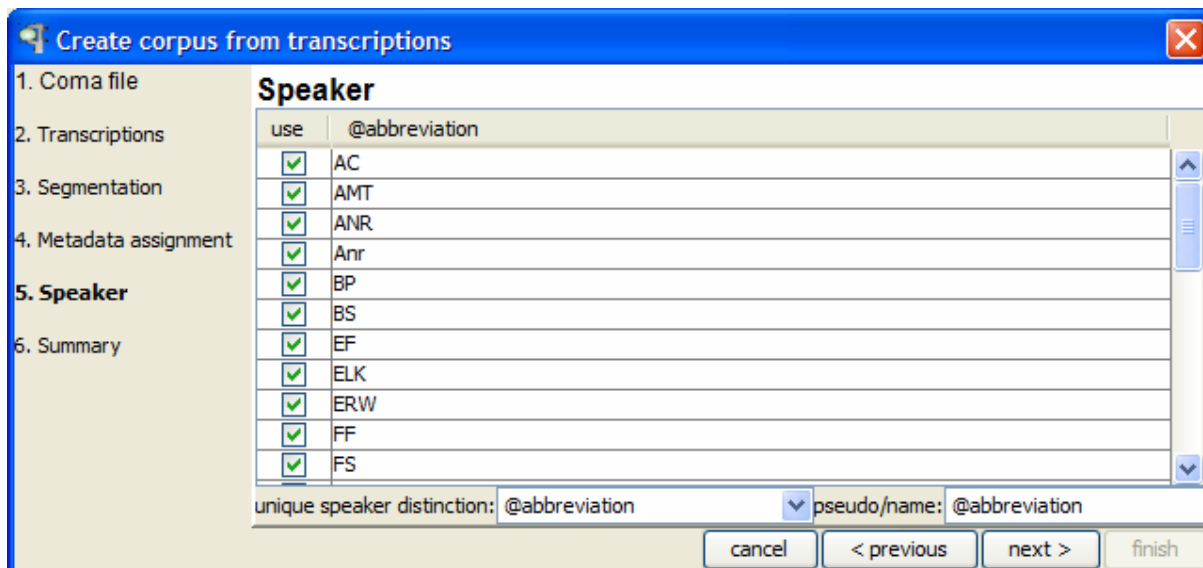
- Tick the box **segment transcriptions**; this tells the wizard that transcriptions are to be segmented;
- for **segmentation algorithm**, choose **default**;
- since the default segmentation algorithm never produces any errors, it is not important what you choose under **on segmentation errors**;
- for **target** choose **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 **next >**.

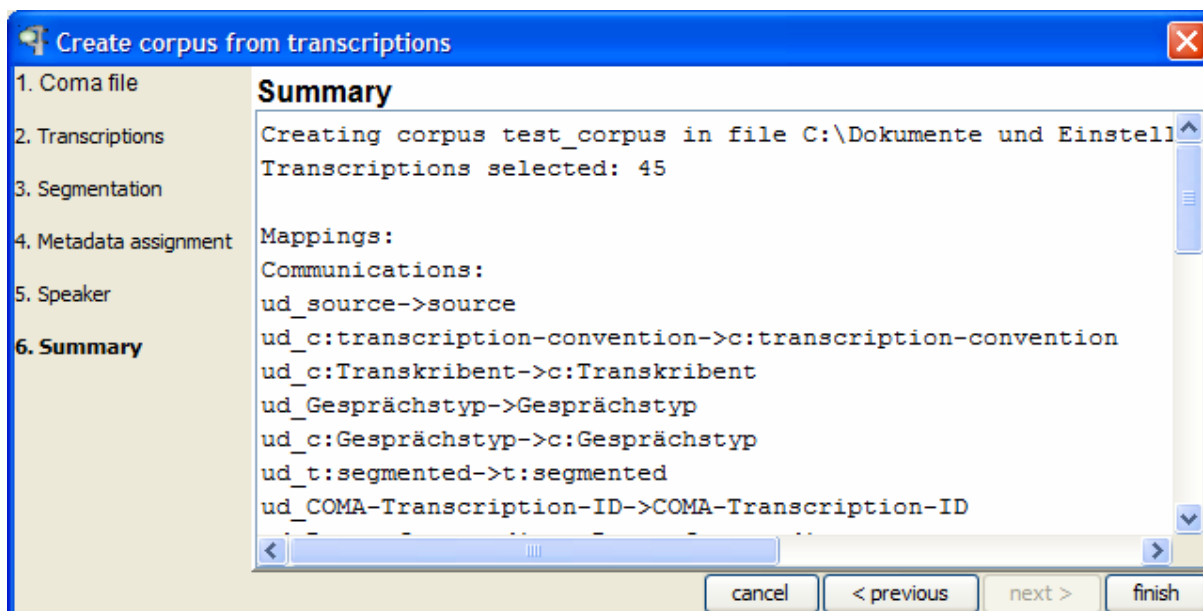


The next dialog is about metadata you have entered in your transcriptions (i.e. with the help of **File > Meta information** in the EXMARaLDA Partitur-Editor). For each such metadata field (second column),

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 **next >** once you have specified everything.



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 **next >** will get you to a summary of the parameters you have set for the wizard:

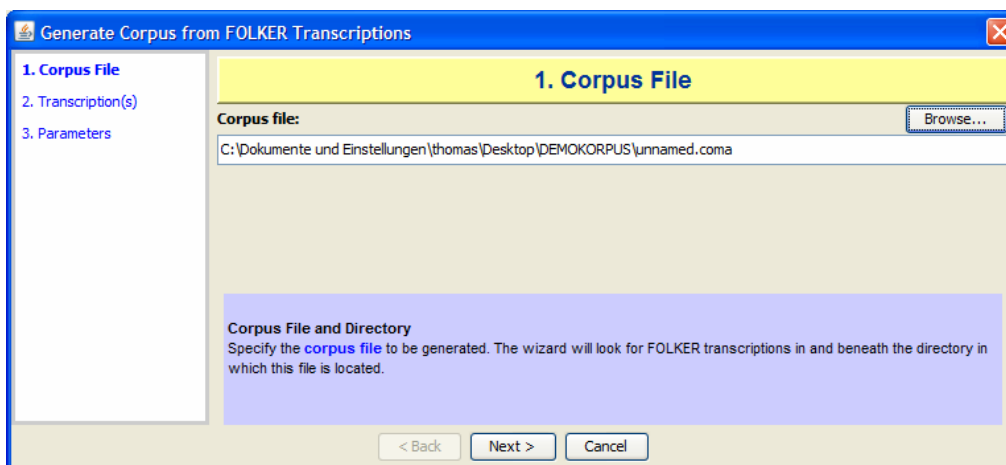


If you now click on **finish** your corpus will be created, saved under the name you specified in step 1 and loaded in EXAKT.

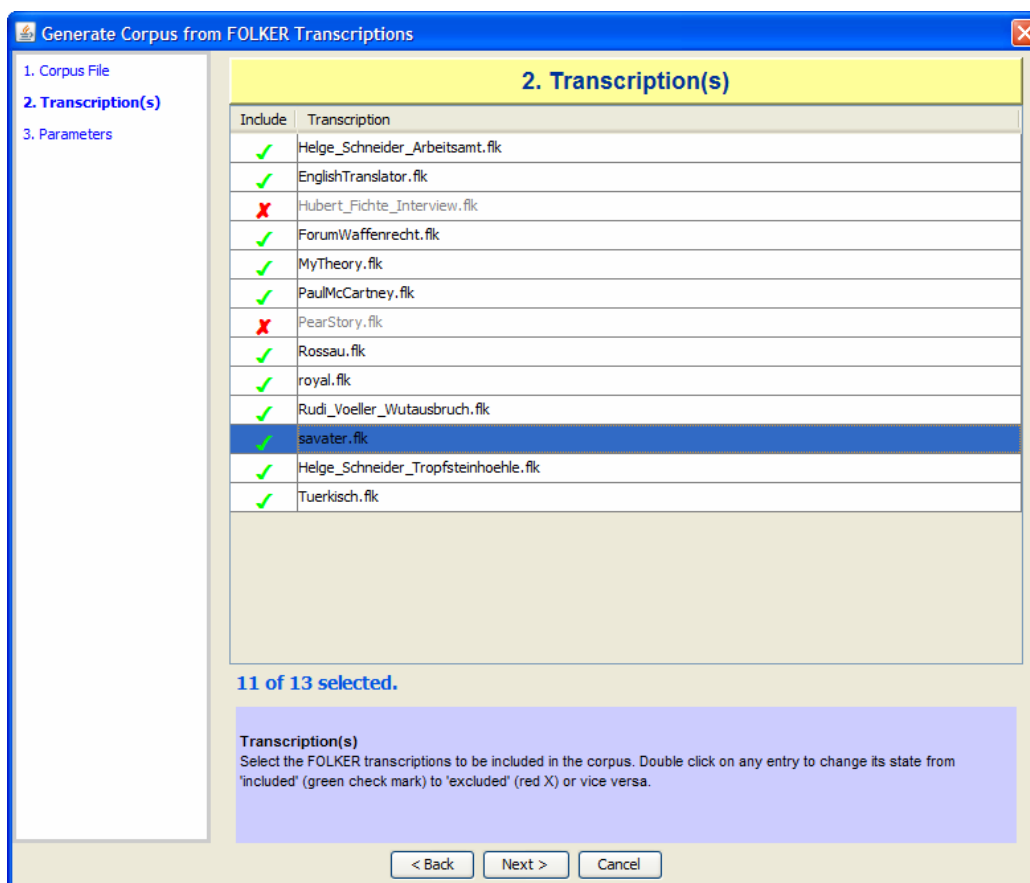
## 1.5. Generating a corpus from FOLKER transcriptions

Via **File > FOLKER corpus...** you can start a wizard for creating an EXMARaLDA corpus from a set of FOLKER transcriptions (FOLKER is a transcription editor published by the IDS at Mannheim, see <http://agd.ids-mannheim.de/html/folker.shtml>). The wizard is similar to the one described in the previous section. Since FOLKER does not store any transcription or speaker metadata, steps 4 and 5 are skipped. The wizard itself contains some additional instructions for each step in the corpus generation.

### Step 1: Specify a Coma file to which the resulting corpus will be written



### Step 2: Select FOLKER transcriptions to be included in the corpus



## 1.6. Generating a corpus from CHAT transcriptions

Via **File > CHAT corpus...** you can start a wizard for creating an EXMARaLDA corpus from a set of CHAT transcriptions (CHAT is the file format written by the CLAN editor from the CHILDES system, see <http://childes.psy.cmu.edu/>). The wizard works in a similar fashion to the one described in the previous section.

## 1.7. Generating a corpus from ELAN annotation files

Via **File > ELAN corpus...** you can start a wizard for creating an EXMARaLDA corpus from a set of ELAN annotation files (see <http://www.lat-mpi.eu/tools/elan/>). The wizard works in a similar fashion to the one described above.

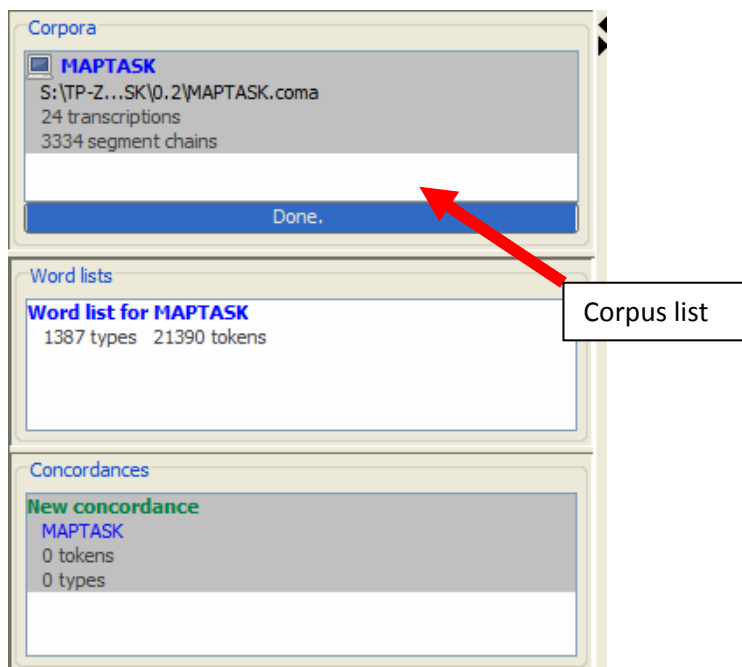
## 1.8. Generating a corpus from Transcriber files

Via **File > Transcriber corpus...** you can start a wizard for creating an EXMARaLDA corpus from a set of Transcriber files (<http://trans.sourceforge.net/>). The wizard works in a similar fashion to the one described above.

## 2. WORKING WITH CONCORDANCES

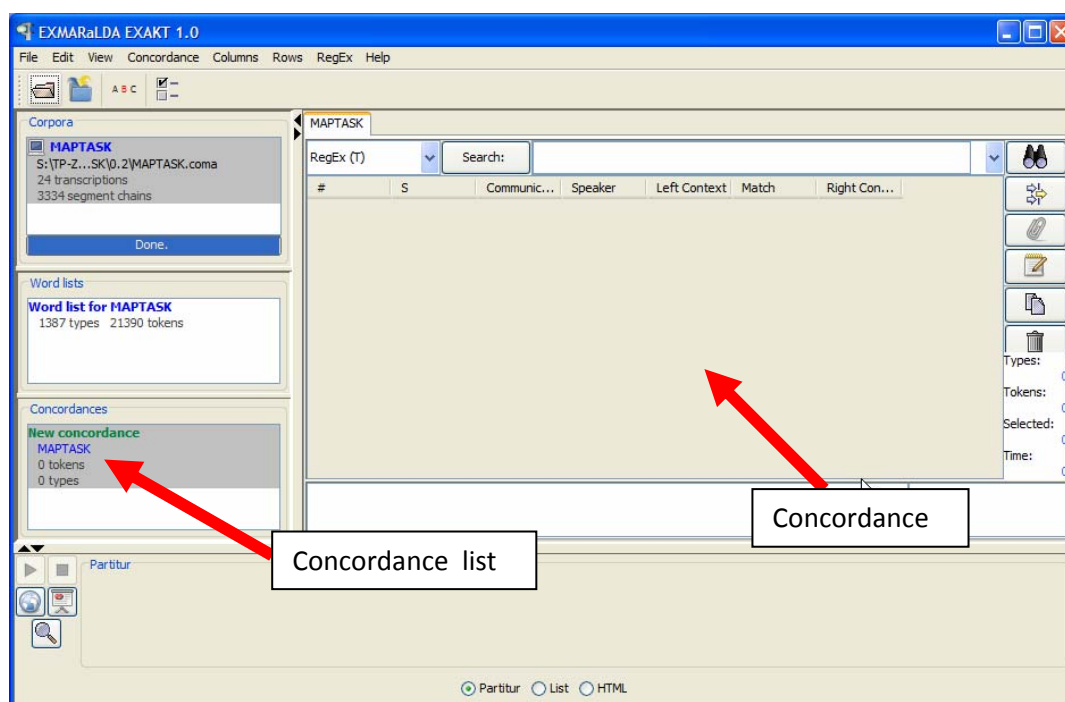
### 2.1. Creating a new concordance

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



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** (One such concordance is created automatically after you've opened a corpus).



## 2.2. Understanding concordances

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.

DEMO-KORPUS (66 results)

RegEx (T) Search: was

#	S	Communication	Speaker	Left Context	Match	Right Context
1	<input checked="" type="checkbox"/>	Rudi Völler: Wutausbruch	RV	Was ihr euch immer alle einbildet,	was	wir alles/ was für Fußball wir in
2	<input checked="" type="checkbox"/>	Rudi Völler: Wutausbruch	RV	mer alle einbildet, was wir alles/	was	für Fußball wir in Deutschland spi
3	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	AMT	Ah, es geht nicht immer äh danach,	was	der • Arbeitnehmer will. Sie wisse
4	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	AMT	n • sind rar. Und man muss nehmen,	was	man kommt. Auch wenn es stinkt da
5	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	AMT	n. Vielleicht ham wir ja auch noch	was	anderes. ((blättert))
6	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	AMT	Ah, ja, das wär vielleicht	was	für Sie. ((blättert)) Jaa, das wär
7	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	AMT	Ja, aber	was	haben Sie denn...
8	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	AMT	wir leider das Arbeitslosengeld et	was	kürzen.
9	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	KLA	cht, ne, dass äh... Keine Phantasie,	was	,Herr Gerizten, keine Phantasie, i
10	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Fichte	Zahlen. Und	was	musste er bezahlen, um reinzukomme
11	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Fichte	Hm: ((schluckt hörbar, 1,2s)) Und	was	bezahlen die Mädchen Miete?
12	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Fichte	Hm: ((1,2s)) Und	was	verdienen die Mädchen so in der Na
13	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	ar ein,0,8s)) Musste ab und zu mal	was	ausschenken.
14	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	und schäkern mit denen. Geben der	was	aus und der was aus, né? •• (So
15	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	t denen. Geben der was aus und der	was	aus, né? •• (So können se) nich
16	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	du Steuer rechnen und ((schnieft))	was	wir da alles rümgesteckt haben.
17	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	ber. Bei uns is s ruhig., oder so	was	„Das is ne nette Atmosphäre.“ ••

er äh danach, was der • Arbeitnehmer will. Sie wissen: die Stellen • sind rar. Und man muss nehmen, was man kommt. Auch wenn es stinkt da n bißchen.

Types: 1  
Tokens: 66  
Selected: 66 (1)  
Time: 0.5 s



#	S	Communication	Speaker	Left Context	Match	Right Context
1	<input checked="" type="checkbox"/>	Rudi Völler: Wutausbruch	RV	Was ihr euch immer alle einbildet,	was	wir alles/ was für Fußball wir in
2	<input checked="" type="checkbox"/>	Rudi Völler: Wutausbruch	RV	mer alle einbildet, was wir alles/	was	für Fußball wir in Deutschland spi
3	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	AMT	Ah, es geht nicht immer äh danach,	was	der • Arbeitnehmer will. Sie wisse
4	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	AMT	n • sind rar. Und man muss nehmen,	was	man kommt. Auch wenn es stinkt da
5	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	AMT	n. Vielleicht ham wir ja auch noch	was	anderes. ((blättert))
6	<input checked="" type="checkbox"/>	Helge Schneider: Arbeitsamt	AMT	Ah, ja, das wär vielleicht	was	für Sie. ((blättert)) Jaa, das wär

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
- column 5 contains the actual search result, i.e. the transcribed text which matched your search expression
- columns 6 and 7 contain the left and right context of that search result

You can sort the table by clicking on any column header. Text in the left context column will be sorted reversely so that words closer to the matched text get the priority. This makes it easier to discover similarities or patterns in the left context.

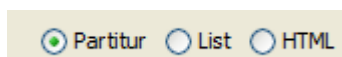
X	th: A rooster crows. ((1,4s)) ((breathes in)) And	the	en you see ehm a maan in maybe his fiftie
X	ars into • em a basket. • • • ((breathes in)) And	the	n you hear a goat. ((1,4s)) ((breathes in)) A
X	guy is just walking (ago). And just walks by. And	the	y don't say anything. • • And then he goes
X	st walks by. And they don't say anything. • • And	the	n he goes/ the man goes up the tree again
ELK	d • • • much much thicker in the middle • • • and	the	n thin again at the far end.

You can reduce or increase the amount of text in the left and right context columns by clicking on the buttons  and , respectively.

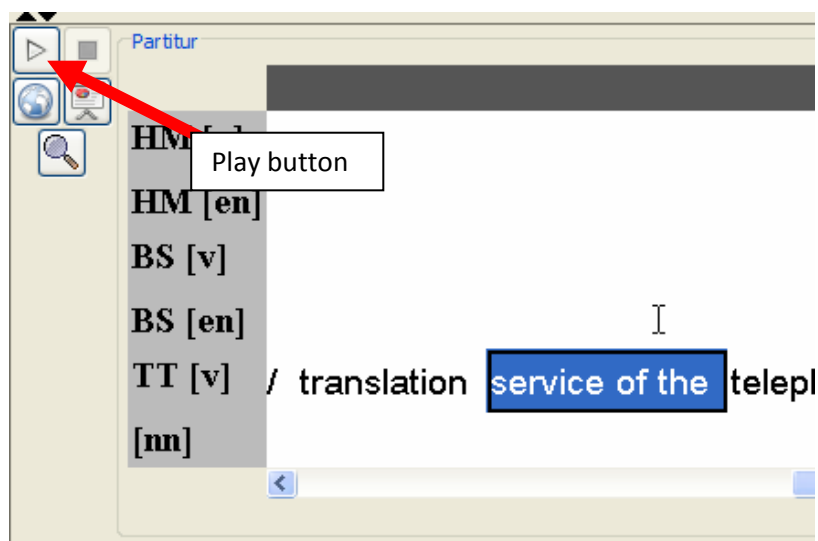
Selecting one search result will also display the corresponding text in the lower left corner of the concordance window.

### 2.3. Going from a search result to the transcription

If you double click on a search result, the corresponding transcription will be opened in the lower part of the screen. You can choose to display the transcription as a partitur, as a list, or as a HTML document generated through a stylesheet transformation. You make this choice by selecting the appropriate radio button from the list in the lower part of the screen:



If you've chosen the **Partitur** option, the transcription will be displayed as a musical score, as in the Partitur-Editor. The event containing the search result is highlighted:



You can freely navigate in this transcription to explore the context of the search result. 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.

If you've chosen the **List** option, the transcription will be displayed as a list of segment chains. The segment chain containing the search result is highlighted:

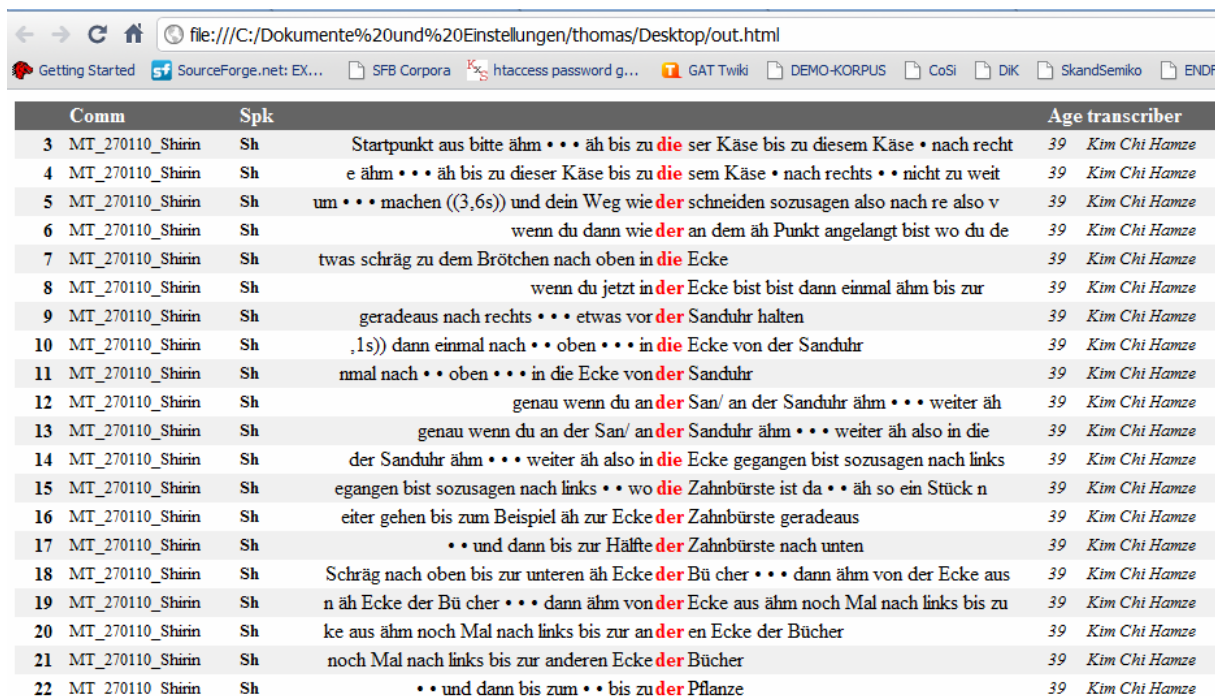
List

	Speaker	Text	
10	HM	Yes.	
11	TT	((Einatmen)) Aand he is calling because of the eh erasure gums.	▲
12	HM	Ja, aber er hat doch eben/ mmh he ha/ he has spoken in German. Why...	■
13	TT	Yes he is talking in German to me and I am a translation (eh) tr/ translation service of the telephone company.	
14	HM	I see.	
15	BS	((0,2s)) Und die R/ Radiergummis sind aus den Sechzigerjahren?	
16	HM	Ganz recht. Ja, aber/ why why is this (this) this gentleman speaking English to me?	▼

Double clicking on any line will playback media aligned with this line. A subsequent single click stops the media player.

## 2.4. Outputting and saving search results

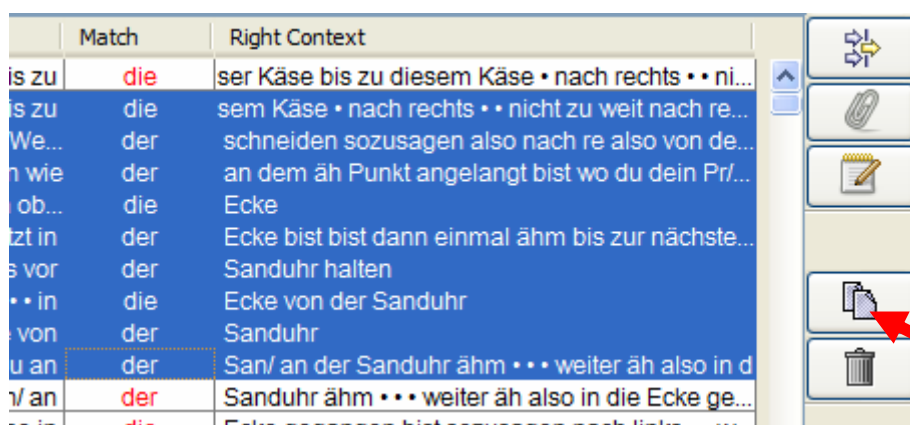
In order to print your search results, use them inside a word processor (e.g. MS Word) or further process them in some other application (e.g. MX Excel), you have several options. Choose **Concordance > Save Concordance as...** in order to save the whole KWIC concordance to a file. The file dialog gives you a choice between HTML and XML. Saving as HTML enables you to open the resulting file in a browser, in a word processor or in any other application which reads HTML (e.g. MS Excel). By default, a built-in style-sheet is used to generate the HTML. In a browser, an exported HTML will look something like this:



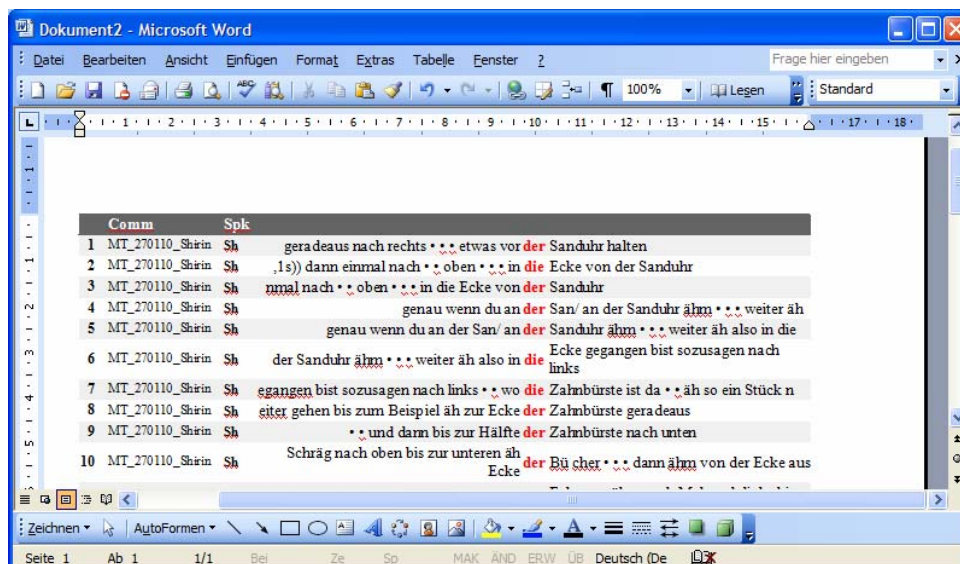
Comm	Spk		Age	transcriber
3	MT_270110_Shirin	Sh Startpunkt aus bitte ähm ••• äh bis zu <b>die</b> ser Käse bis zu diesem Käse • nach recht	39	Kim Chi Hamze
4	MT_270110_Shirin	Sh e ähm ••• äh bis zu dieser Käse bis zu <b>die</b> sem Käse • nach rechts •• nicht zu weit	39	Kim Chi Hamze
5	MT_270110_Shirin	Sh um ••• machen ((3,6s)) und dein Weg wie <b>der</b> schneiden sozusagen also nach re also v	39	Kim Chi Hamze
6	MT_270110_Shirin	Sh wenn du dann wie <b>der</b> an dem äh Punkt angelangt bist wo du de	39	Kim Chi Hamze
7	MT_270110_Shirin	Sh twas schräg zu dem Brötchen nach oben in <b>die</b> Ecke	39	Kim Chi Hamze
8	MT_270110_Shirin	Sh wenn du jetzt in <b>der</b> Ecke bist bist dann einmal ähm bis zur	39	Kim Chi Hamze
9	MT_270110_Shirin	Sh geradeaus nach rechts ••• etwas vor <b>der</b> Sanduhr halten	39	Kim Chi Hamze
10	MT_270110_Shirin	Sh ,1s)) dann einmal nach •• oben ••• in <b>die</b> Ecke von der Sanduhr	39	Kim Chi Hamze
11	MT_270110_Shirin	Sh nmal nach •• oben ••• in die Ecke von <b>der</b> Sanduhr	39	Kim Chi Hamze
12	MT_270110_Shirin	Sh genau wenn du an <b>der</b> San/ an der Sanduhr ähm ••• weiter äh	39	Kim Chi Hamze
13	MT_270110_Shirin	Sh genau wenn du an <b>der</b> San/ an <b>der</b> Sanduhr ähm ••• weiter äh also in die	39	Kim Chi Hamze
14	MT_270110_Shirin	Sh der Sanduhr ähm ••• weiter äh also in <b>die</b> Ecke gegangen bist sozusagen nach links	39	Kim Chi Hamze
15	MT_270110_Shirin	Sh gegangen bist sozusagen nach links •• wo <b>die</b> Zahnbürste ist da •• äh so ein Stück n	39	Kim Chi Hamze
16	MT_270110_Shirin	Sh eiter gehen bis zum Beispiel äh zur Ecke <b>der</b> Zahnbürste geradeaus	39	Kim Chi Hamze
17	MT_270110_Shirin	Sh •• und dann bis zur Hälfte <b>der</b> Zahnbürste nach unten	39	Kim Chi Hamze
18	MT_270110_Shirin	Sh Schräg nach oben bis zur unteren äh Ecke <b>der</b> Bü cher ••• dann ähm von der Ecke aus	39	Kim Chi Hamze
19	MT_270110_Shirin	Sh n äh Ecke der Bü cher ••• dann ähm von <b>der</b> Ecke aus ähm noch Mal nach links bis zu	39	Kim Chi Hamze
20	MT_270110_Shirin	Sh ke aus ähm noch Mal nach links bis zur an <b>der</b> en Ecke der Bücher	39	Kim Chi Hamze
21	MT_270110_Shirin	Sh noch Mal nach links bis zur anderen Ecke <b>der</b> Bücher	39	Kim Chi Hamze
22	MT_270110_Shirin	Sh •• und dann bis zum •• bis zu <b>der</b> Pflanze	39	Kim Chi Hamze

If you want to change the appearance, you can write a custom XSL stylesheet and tell EXAKT to use it via **Edit > EXAKT Preferences... > Stylesheets > Concordance Output**.

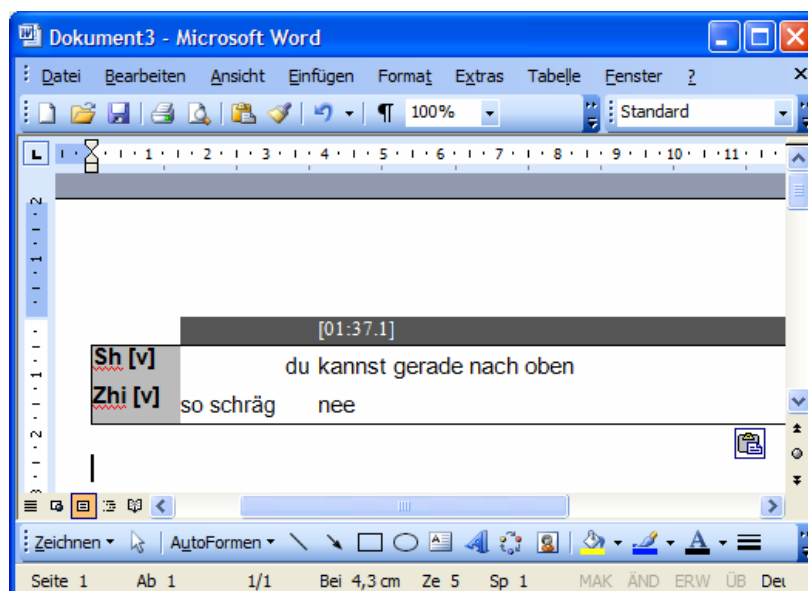
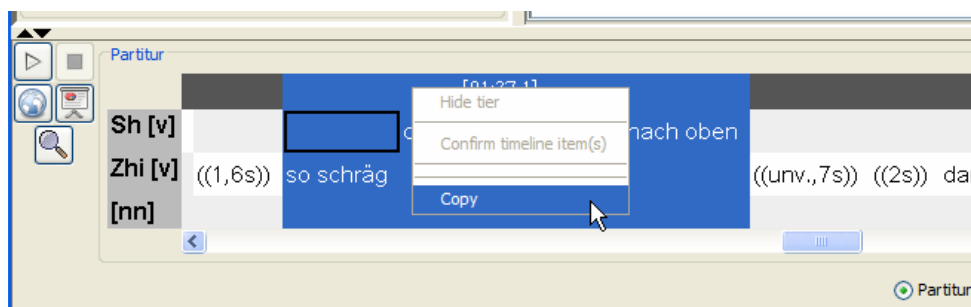
If you want to copy and paste a selection of search results into your word processor, simply make the selection inside the concordance then choose **Edit > Copy Selection** or press the **Copy button** beside the concordance. The selected search results will be copied to the clipboard and you can paste them from there.



Copy button



You can also use **Ctrl + C** or **Copy** from the Partitur's context menu to copy a part of the transcription to the clipboard.



Sometimes (especially if you have added manual annotations to a KWIC concordance), you want to save a search result not to display it via some other application, but to reopen it in EXAKT itself. In order to do that, proceed as follows:

- 1) Choose **Concordance > Save Concordance as...**
- 2) Select the **XML** option and specify a filename

- 3) To reopen the search result: first open the corpus from which it was derived
- 4) Then select this corpus in the corpus list and choose **Concordance > Open concordance...**

Of course, you can also use the exported XML file to do your own further processing of the search results (e.g. by transforming them via an XSL stylesheet).

### 3. SEARCH EXPRESSIONS

#### 3.1. Regular 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:

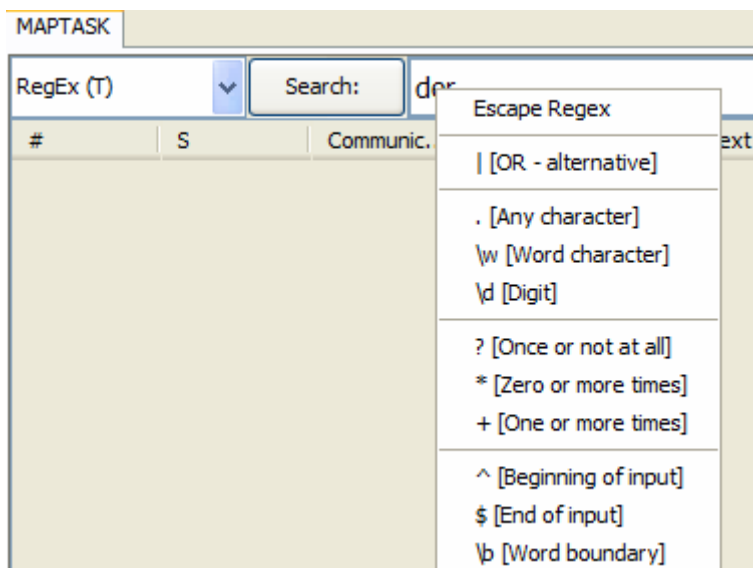
- The pattern `[Ww]as` will match the strings “was” and “Was”.
- The pattern `komm.{1,2}` will match “komme”, “kommst”, “kommen”, “komma”, “kommun” etc.
- The pattern `([Ii]ch|[Dd]u)` will match “ich”, “Ich”, “du” and “Du”
- The pattern `\bge[A-Za-z]+?t\b` will match “gemacht”, “gesagt”, “gewusst”, “geht” etc.
- `[Tt]h(is|at|ose|ese)` will match the words *this*, *that*, *those* and *these* and their capitalized variants.
- `\bin[a-z]+abl[ey]\b` will match words starting with *in* and ending in *able* or *ably* like *indisputable*, *indescribably*, *ineffable*, *indistinguishable* etc.
- `(\b[A-Za-z]+\b){3,3}\?` will match all sequences of three words followed by a question mark, i.e. the last three words of questions
- `\btou(s|t|te|tes)\b` will match the French quantifier words *tous*, *tout*, *toute* et *toutes*
- `\b([MmTtSs](a|on|es)|[Ll]eur(s)?|[VvNn](os|ôtre))\b` will match all French possessive pronouns (*mon*, *ma*, *mes*, *ton*, *ta*, *tes* etc.)

By combining regular expressions in various ways, you can formulate rather complex queries with them. Useful as they are, regular expressions are not very easy to learn. There are very many books and websites explaining regular expressions. We recommend that you consult at least one of those and use it as a reference when working with EXAKT.

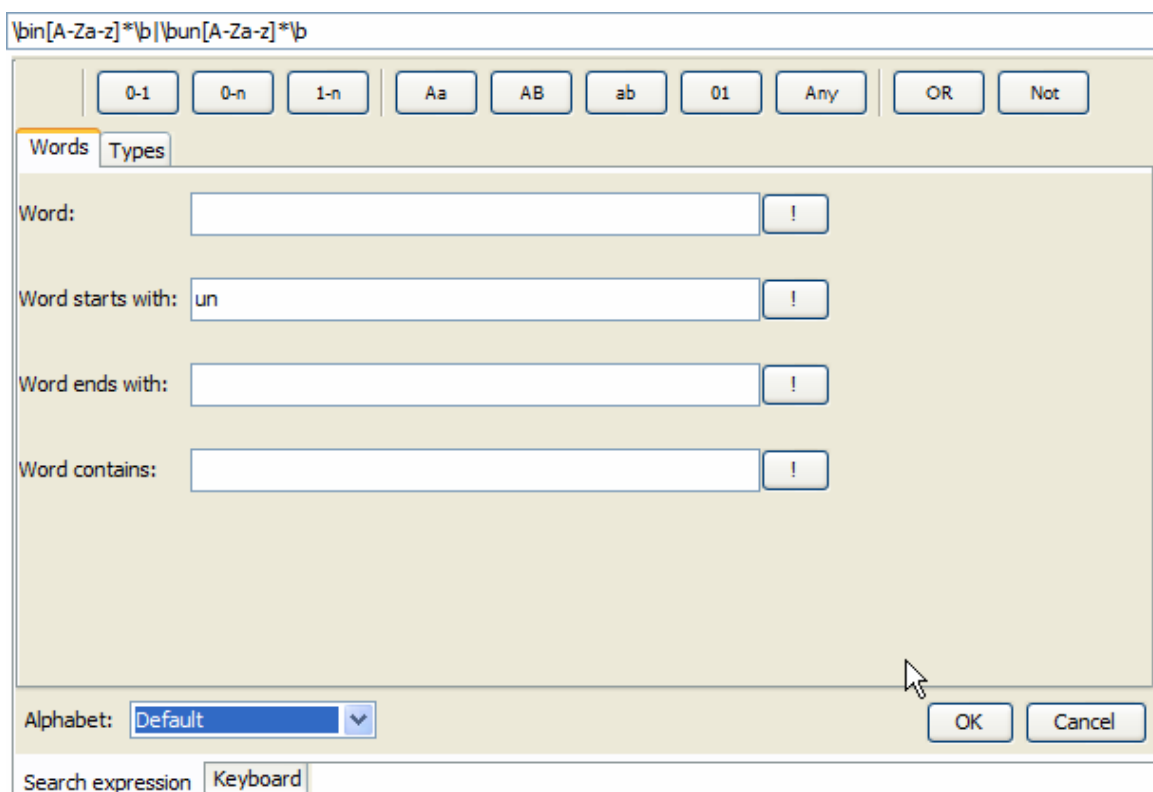
For those not afraid of formal specifications, the exact syntax and usage of regular expressions is explained at:

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

EXAKT provides help for working with regular expression in several places. The simplest is a context menu in the search expression text field which lists some commonly used meta-characters. Choose any entry in that context menu to paste the respective character into the search expression text field.



If you click on the **Search** button to the left of the search expression text field, a dialog will come up which helps you to formulate some commonly used types of regular expression.



For example, if you want to look for all words starting with “in” or “on”, you could proceed as follows:

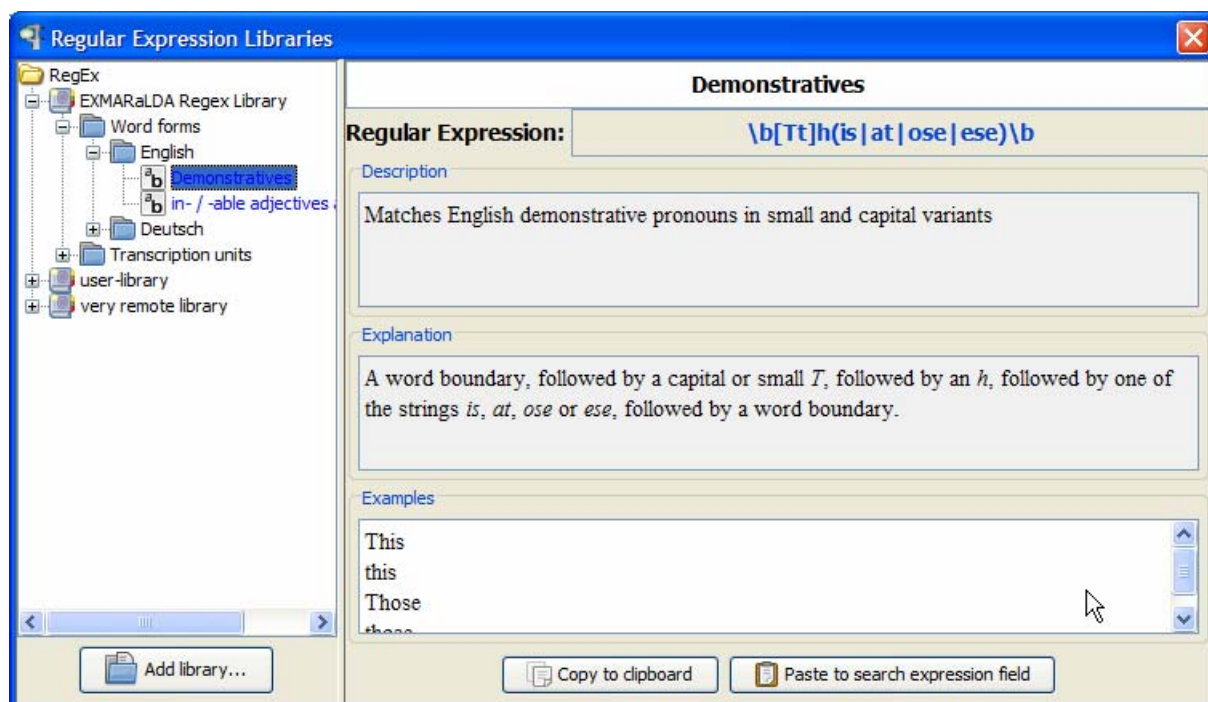
- Choose the **Alphabet** you want to work with. The default is the English alphabet with the letters a-z and their capitalized variants. Other alphabets have additional characters, e.g. German has the Umlauts ä, ö and ü and the “sharp s” ß.

- enter the string “in” in the field **Word starts with:** and press the button with the exclamation mark !
- press the button **OR**
- enter the string “un” in the field **Word starts with:** and press the button with the exclamation mark !
- press **ENTER** to paste the whole search expression into the search expression text field of the concordance and close the helper dialog

The expression constructed should then look something like this:

$$\backslash\text{bin}[A-Za-z]^*\backslash\text{b}|\backslash\text{bun}[A-Za-z]^*\backslash\text{b}$$

A third type of help is to be found under the menu item **RegEx > Regex Library Dialog**. This will bring up a dialog with different regular expression libraries. One of them – the **EXMARaLDA Regex Library** – is built into EXAKT. It contains some commonly used search patterns for different languages and different transcription systems. A second one – the **user-library** – can be used to store and describe your own regular expressions for reuse. Finally, one or more **remote libraries** which are stored under some URL in the WWW can be loaded by clicking on the **Add library...** button.



Libraries are organised into a tree of folders. Each entry in such a folder consists of the regular expression itself, a description saying what the expression will match, an explanation of how it will match it, and a few examples of matching strings. If you want to use the displayed entry, simply click on **Paste to search expression text field**. In order to add your own entry to the user-library, enter an expression in the concordance, apply it to the corpus and then choose **RegEx > Add to library...** The following dialog will come up in which you can enter a name, a description and an explanation for the new entry. Some examples taken from the current search result will be automatically provided.

**New Regex Library Entry**

**Words starting with 'in' or 'un'**

**Regular Expression:** `\\bin[A-Za-z]*\\b | \\bun[A-Za-z]*\\b`

Description

Explanation

Examples

- unter
- uns
- unverst
- un...

OK Cancel

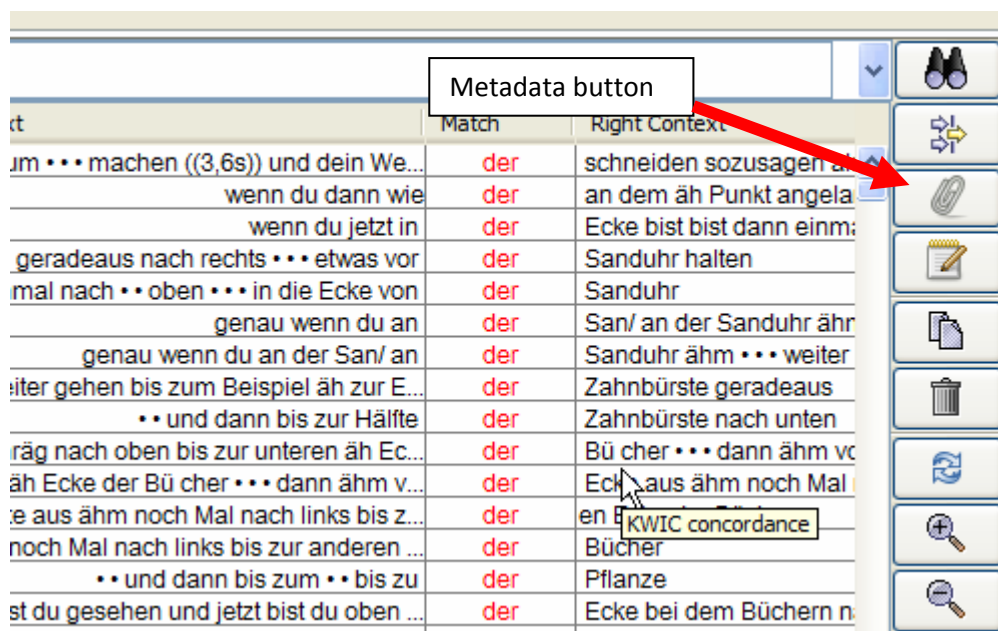
Clicking on OK will add the entry to the user-library. The user-library is saved when you exit EXAKT so that all entries you have added will still be available at the next start.

## 3.2. XPath Expressions

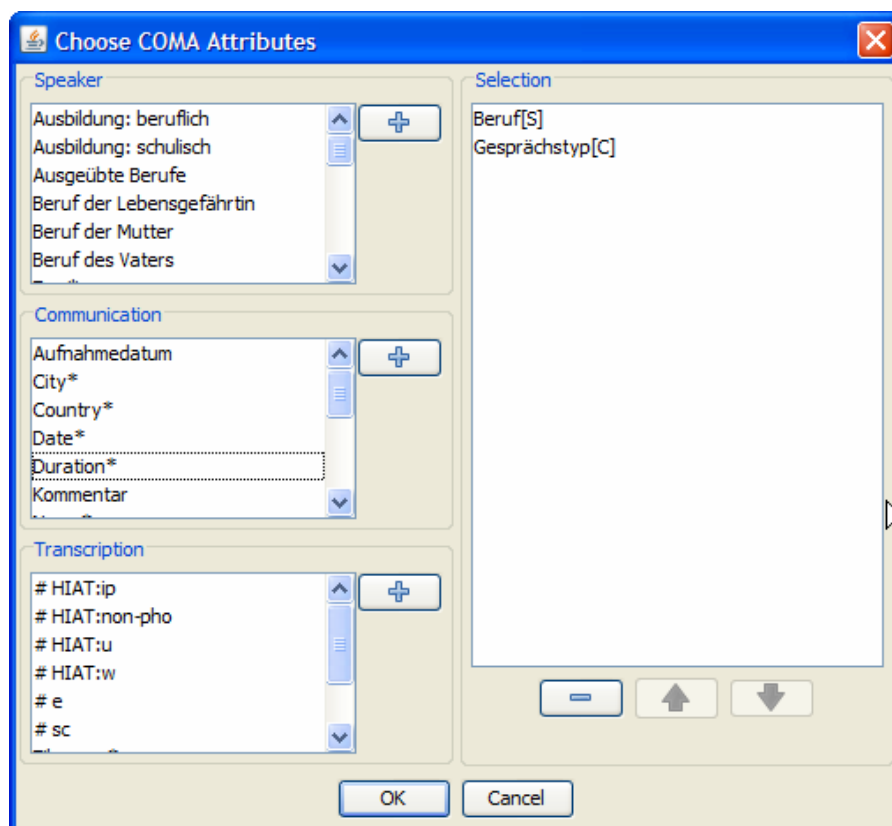
\*\*\* TODO \*\*\*

#### 4. DISPLAYING METADATA


Once you have carried out a search, you can display metadata, i.e. data about speakers, communications or transcriptions in additional columns of your KWIC concordance.



Click on the metadata button or choose **Columns > Metadata...** The following dialog is displayed:



It lists metadata attributes for different entities of the corpus – the speakers, the communications and the transcriptions. You can select any number of these attributes by double-clicking on them or

by selecting them and clicking on the corresponding  button. When you close the dialog by clicking on OK, each selected metadata attributes will be given an additional column in the KWIC concordance, and the corresponding values will be displayed in that column.

	Match	Right Context	Gesprächstyp[C]	Beruf[S]
lle einbildet,	was	wir alles/ was für Fußball w	Fernsehinterview	Teamchef der Fußballnationalm
as wir alles/	was	für Fußball wir in Deutschla	Fernsehinterview	Teamchef der Fußballnationalm
äh danach,	was	der • Arbeitnehmer will. Sie	Telefonisches Arbeitsvermittlungsgespräch	Fachangestellter für Arbeitsförde
ss nehmen,	was	man kommt. Auch wenn es stin	Telefonisches Arbeitsvermittlungsgespräch	Fachangestellter für Arbeitsförde
a auch noch	was	anderes. ((blättert))	Telefonisches Arbeitsvermittlungsgespräch	Fachangestellter für Arbeitsförde
är vielleicht	was	für Sie. ((blättert)) Jaa, d	Telefonisches Arbeitsvermittlungsgespräch	Fachangestellter für Arbeitsförde
Ja, aber	was	haben Sie denn...	Telefonisches Arbeitsvermittlungsgespräch	Fachangestellter für Arbeitsförde
losengeld et	was	kürzen.	Telefonisches Arbeitsvermittlungsgespräch	Fachangestellter für Arbeitsförde
Phantasie,	was	, Herr Gerizten, keine Phanta	Telefonisches Arbeitsvermittlungsgespräch	Arbeitslos

Metadata columns

These metadata columns can be sorted and filtered just like the other columns of the KWIC concordance. If you now select a search result, its metadata will also be displayed in the right text field below the KWIC concordance.

11	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Fichte	(schluckt hörbar, 1,2s) Und	was	bezahlen die Mädchen
12	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Fichte	Hm ((1,2s)) Und	was	verdienen die Mädchen
13	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	,0,8s) Musste ab und zu mal	was	ausschenken.
14	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	chäkern mit denen. Geben der	was	aus und der was aus,
15	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	n. Geben der was aus und der	was	aus, né? • • (So könn
16	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	uer rechnen und ((schnieft))	was	wir da alles reingeste
17	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	ei uns is s ruhig.", oder so	was	, "Das is ne nette Atr
18	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	Was du/	was	der Witz dabei is, • d
19	<input checked="" type="checkbox"/>	Hubert Fichte: Interview	Ulli	istungen ((schreit)) "Das	was	du da im Moment best

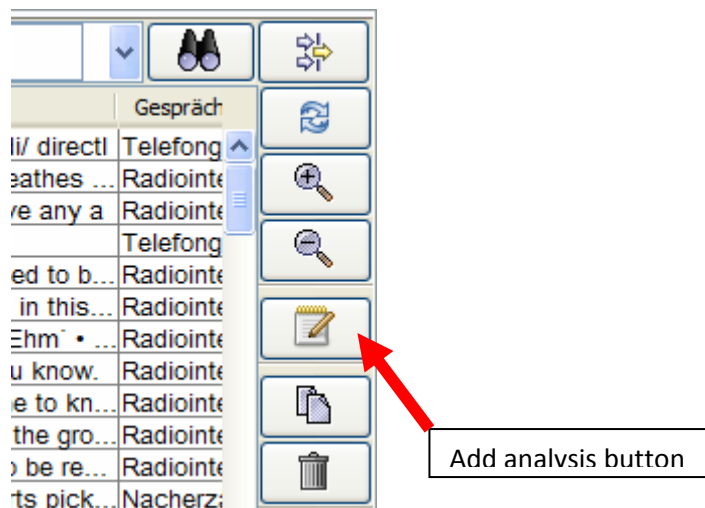
Ja. ((holt Luft,0,3s)) Ja, dat hab ich mir gedacht, ne, dass äh... Keine Phantasie, was, Herr Gerizten, keine Phantasie, ihr vom Arbeitsamt, ne? ((holt Luft,0,5s)) Also euch, äh n/ äh n... N

Gesprächstyp[C]	Telefonisches Arbeitsvermittlungsgespräch
Beruf[S]	Arbeitslos

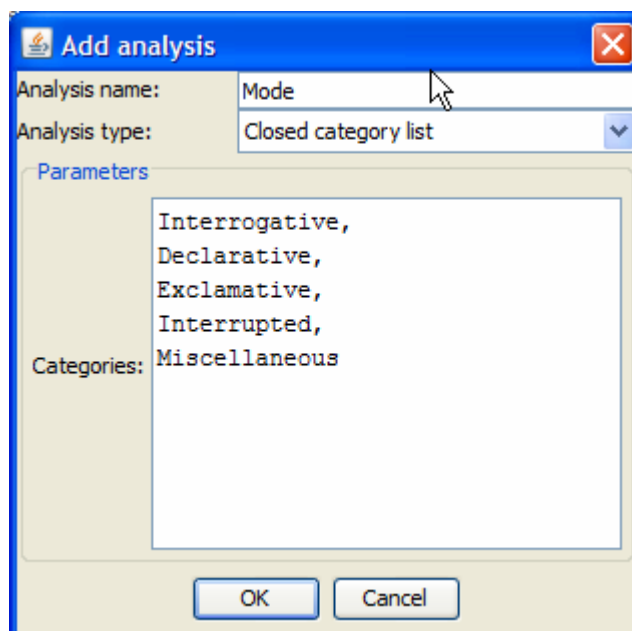
Metadata display for the selected search result

## 5. ADDING ANALYSIS COLUMNS

In order to classify or categorize your search results, you can add one or several analysis columns to the KWIC Concordance. Click on the **Add analysis...** button on the right side of the application or choose **Columns > Add analysis...**



This will display the following dialog:



Enter a name for your analysis column in the **Analysis name** field. Under **Analysis type**, you can choose between three types of analyses:

- A **Free** analysis lets you enter an arbitrary text. This can be useful, for instance, for making free comments on individual search results.
- A **Closed category list** analysis lets you choose from a predefined set of categories. Enter the list of categories into the provided text field, separating individual categories with a comma.
- A **Binary** analysis lets you tick or untick a check box for each search result. This can be useful, for instance, to manually distinguish relevant from irrelevant search results.

When you close the dialog by clicking on OK, an additional column in the KWIC concordance will be created in which you can carry out your analysis:

	Match	Right Context	Mode
(0,3s) I'd/ I'd I want to speak with	the	German/ with the German int/ eh peopl	
nt to speak with the German/ with	the	German int/ eh people who is interest	
int/ eh people who is interesting in	the	gums.	Interrogative
to)... We speak German to each o	the	r!	Exclamative
I am	the	translator of Mr Bernd Schwanmeister.	Interrogative
en)) Aand he is calling because of	the	eh erasure gums.	Interrupted
ation (eh) tr/ translation service of	the	telephone company.	Declarative
(Hello), I am in	the	line all the time. Now eh it's a serv	
(Hello), I am in the line all	the	time. Now eh it's a service from (the	Interrogative
≥ time. Now eh it's a service from (	the	) eeh English telephone company.	Declarative
(So) he is asking what's about	the	eraser gums?	Exclamative
Yeah, y/ you can. He is in	the	line.	Interrupted
y to speak German with eh each o	the	r and I try to translate in between, O	Miscellaneous

Analysis column

Analysis columns can be sorted and filtered just like the other columns of the KWIC concordance.

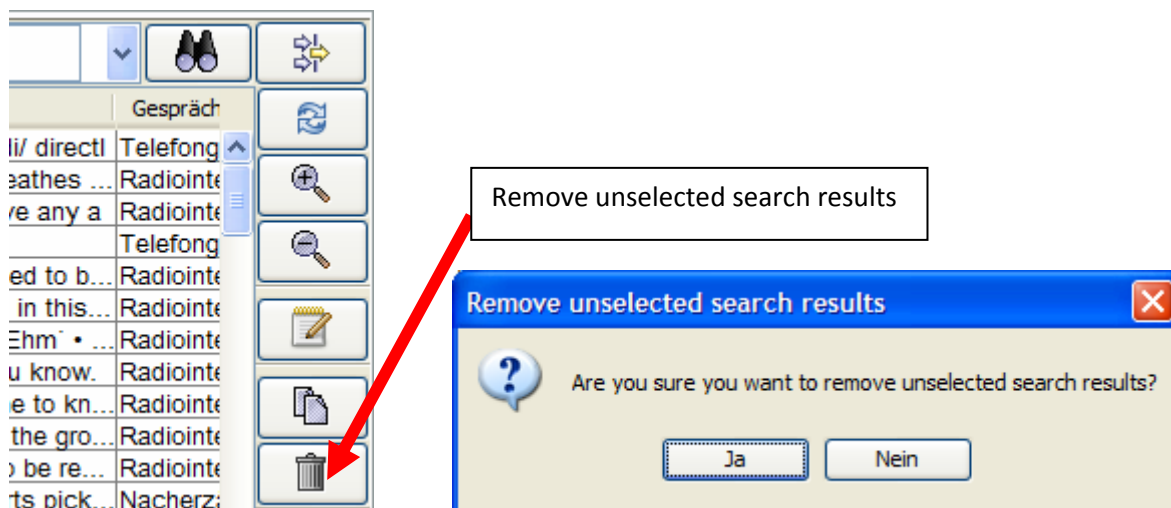
## 6. FILTERING SEARCH RESULTS

Often, your list of search results contains some unwanted or irrelevant instances. In order to get rid of those you can have two options.

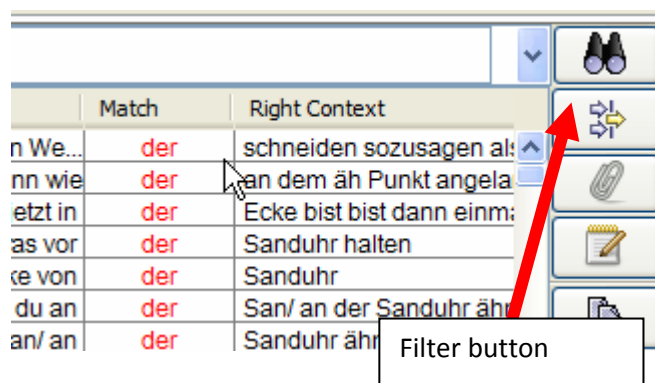
Option 1 is to manually go through the KWIC concordance and unselect the unwanted instances in the column with the check boxes.

#	S	Communication	Speaker	Left Context	Match	Right Context
1	<input type="checkbox"/>	Hart aber fair	FB	n. • Das war ein ewiges Affen	the	ater. • • ((atmet ein)) Das h
2	<input checked="" type="checkbox"/>	Hart aber fair	HB	die meisten Sendungen zu Rand	the	men gemacht. Sie ham über Mana
3	<input type="checkbox"/>	Monty Python: My Theory	PRE	• • You • • have a new	the	ory about the brontosaurus.
4	<input checked="" type="checkbox"/>	Monty Python: My Theory	PRE	• • have a new theory about	the	brontosaurus.
5	<input checked="" type="checkbox"/>	Monty Python: My Theory	PRE	I mean your/ your new	the	ory.
6	<input type="checkbox"/>	Monty Python: My Theory	PRE	Well eh this	the	ory of yours appears to have
7	<input checked="" type="checkbox"/>	Monty Python: My Theory	PRE	of yours appears to have hit	the	nail on the head.
8	<input type="checkbox"/>	Monty Python: My Theory	PRE	ears to have hit the nail on	the	head.
9	<input type="checkbox"/>	Monty Python: My Theory	PRE	ery much for coming along to	the	studio.
10	<input checked="" type="checkbox"/>	Monty Python: My Theory	ELK	one moment that I have a new	the	ory about the brontosaurus.
11	<input checked="" type="checkbox"/>	Monty Python: My Theory	ELK	at I have a new theory about	the	brontosaurus.
12	<input checked="" type="checkbox"/>	Monty Python: My Theory	ELK	• • • Oh what is my	the	ory?
13	<input checked="" type="checkbox"/>	Monty Python: My Theory	ELK	Ooh what is my	the	ory that it is! • • Well, Chr
14	<input checked="" type="checkbox"/>	Monty Python: My Theory	ELK	u may well ask me what is my	the	ory.

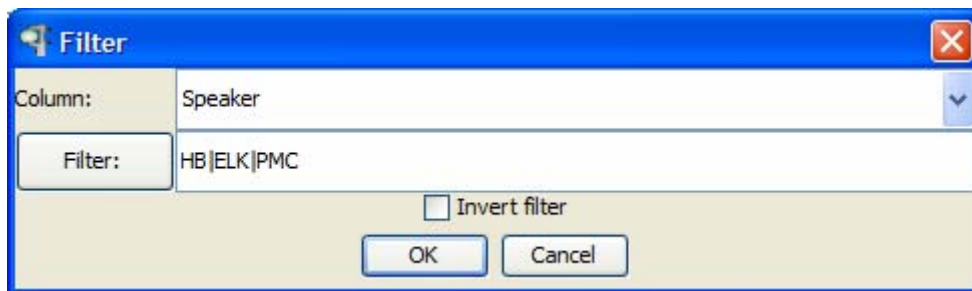
If you then click on the trash bin button on the right side of the window, all unselected search results will be removed from the concordance.



The other option is to automatically filter your concordance according to certain criteria. Click on the **Filter** button to the right of the search button.



This will open the following dialog:



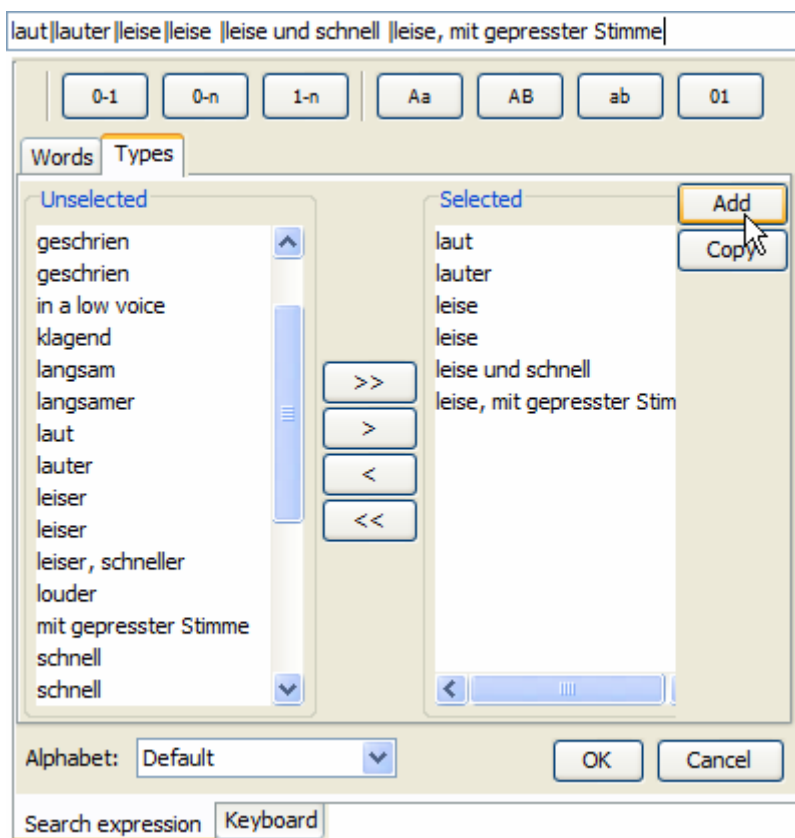
Under **Column**, you can specify the column in the concordance according to which you want to filter. Under **Filter**, you can specify a filter in the form of a regular expression (see above). For example, the screen shot shows a filter that will go through the **Speaker** column of the concordance and select all search results in which the value in that column matches the regular expression...

HB | ELK | PMC

...and unselect all search results which do not match the expression. In order to exchange the roles of selected and unselected search results, tick the box **Invert filter**.

Clicking on **OK** will give you a KWIC concordance in which the selection check boxes are ticked according to your filter. You can then throw away unselected search results as described above.

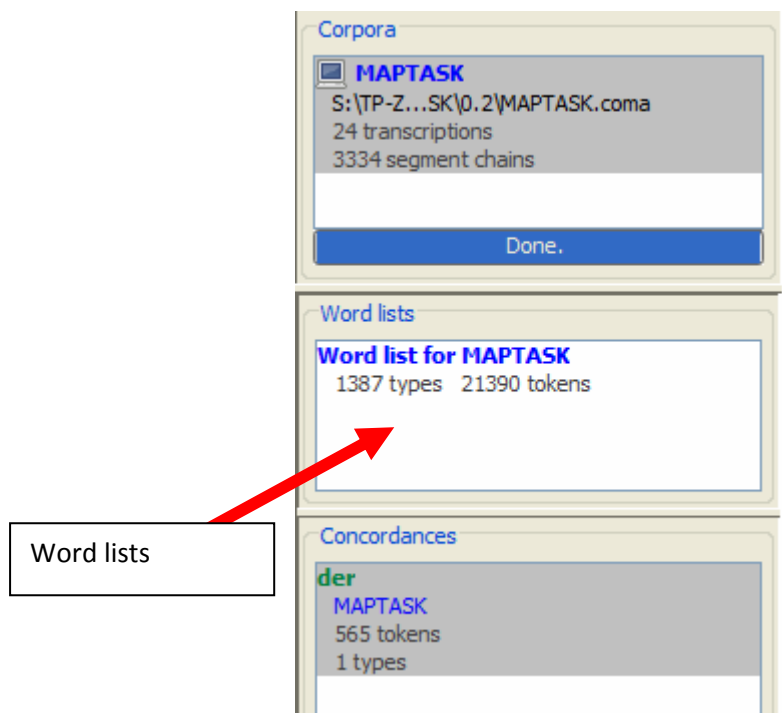
A useful help for many filtering tasks is a token list, i.e. a list containing all distinct forms from a column. If you click on the **Filter:** button such a list will be displayed for the currently selected columns (this will not work for the left and right context columns, though).



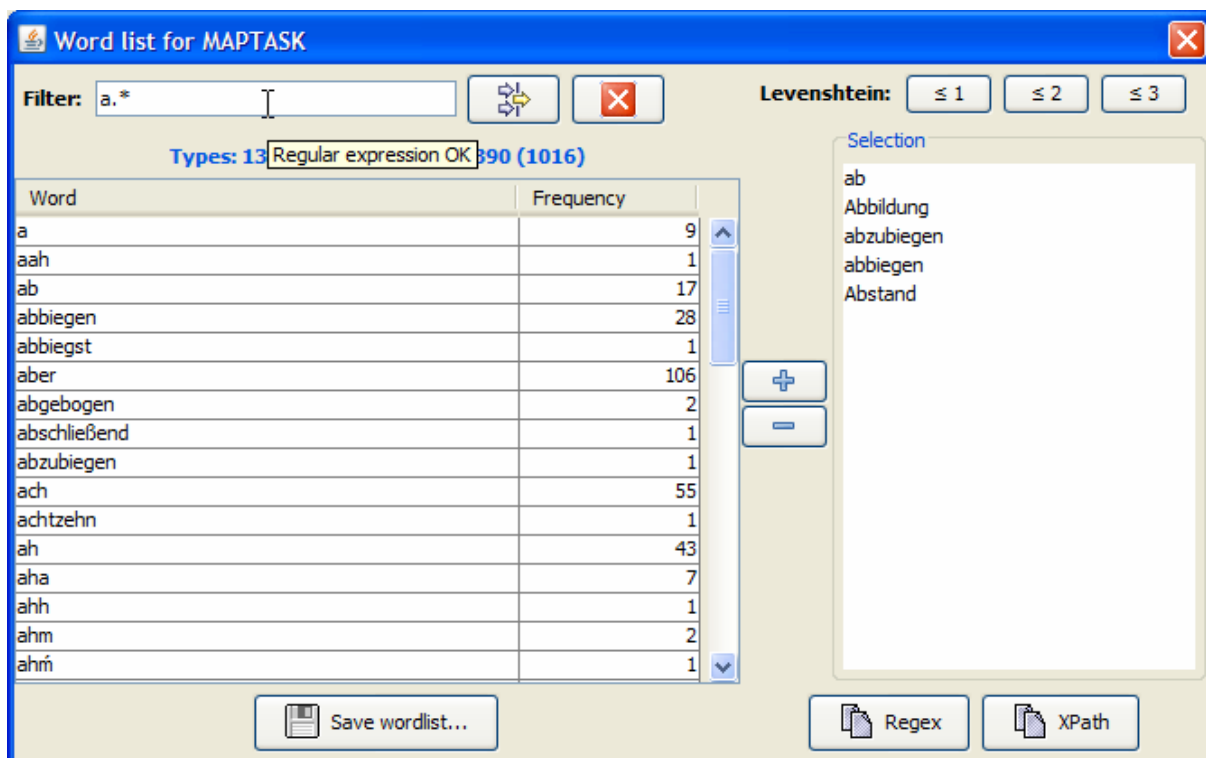
The **Unselected** list displays all tokens found in the column. You can add tokens to the **Selected** list by clicking on **>**. Clicking on **Add** will produce a regular expression corresponding to the selected tokens. By pressing ENTER, you can paste this expression into the Filter dialog.

## 7. USING WORD LISTS

If you open a corpus which has been segmented for words (this depends on the segmentation algorithm used), EXAKT offers you the possibility to generate and use a word list. All word lists are displayed on the left hand side under **Word lists**.

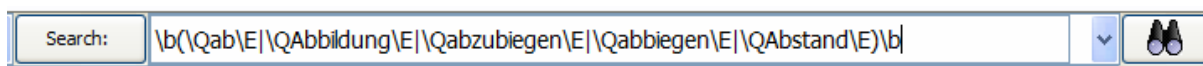
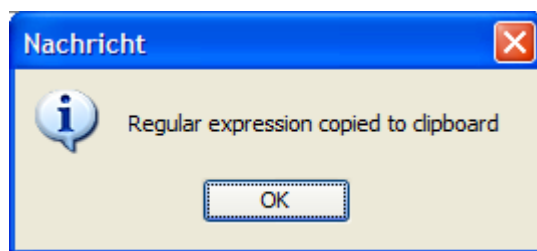


If you double click on an entry of that list, a dialog of the following type will be displayed:



This lists all the word types occurring in the corpus together with their frequency. Click on the table header to sort words alphabetically or by their frequency. Click on **Save wordlist...** to generate a HTML version of the wordlist. You can filter the list using a regular expression. On the right side of the dialog, you have a list with selected words. Double click on any entry in the word list or click on the button with the plus sign to add a word to the selection.

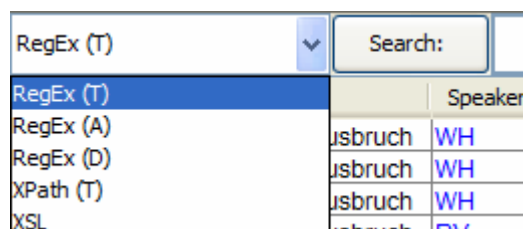
Clicking on the button **Regex** will copy a regular expression to the clipboard with which EXAKT can search exactly those word forms contained in the selection list. Click on the search expression text field and press **Ctrl + V** to paste that regular expression into the field.



Above the selection list, you have three buttons for automatically extending the selection via a Levenshtein function. The Levenshtein Distance between two strings A and B is defined as the minimal number of insertion, deletion or substitution operations necessary to get from A to B. For example, the Levenshtein distance between *car* and *war* or between *run* and *runs* is 1 because it requires one substitution or one insertion, respectively, to get from one string to the other. If you click the button **≤1**, all words from the word list will be added to the selection whose Levenshtein distance to any one of the entries in the selection list is less than or equal to 1. The buttons **≤2** and **≤3** work in an analogous manner.

## 8. DIFFERENT TYPES OF SEARCHES

So far, we have demonstrated all functionality with the example of a regular expression search over transcription text. Actually, however, this is only one of several types of searches you can do with EXAKT. Which type of search is carried out is specified via the combo box beside the **Search:** button.



### 8.1. Regular Expression Search over Transcription tiers [RegEx (T)]

\*\*\* TODO \*\*\*

### 8.2. Regular Expression Search over Annotation tiers [RegEx (A)]

\*\*\* TODO \*\*\*

### 8.3. Regular Expression Search over Description tiers [RegEx (D)]

\*\*\* TODO \*\*\*

### 8.4. XPath Search over Transcription tiers [XPath (D)]

\*\*\* TODO \*\*\*

### 8.5. Multilevel searches

\*\*\* TODO \*\*\*