



How to make a corpus from a set of transcriptions

This document explains how to import EXMARaLDA transcriptions and their metadata into the EXMARaLDA Corpus-Manager, creating a corpus that can be used – amongst other things – inside the EXAKT search tool.

Before you start reading this document, you should have read and understood

- Understanding the basics of EXMARaLDA
- Understanding Coma-Metadata.

You should also have a set of EXMARaLDA-Transcriptions, with which you can try out the steps described in this document. This document uses the EXMARaLDA demo corpus.

Create a Coma-Corpus from a Set of Transcriptions

This assistant – which is located in the „File“-Menu, is used to create a Coma file from a set of EXMARaLDA-transcriptions that already have metadata stored in their transcription-heads.

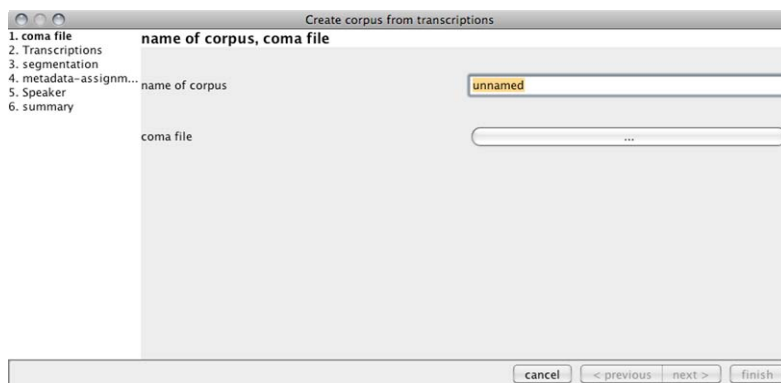
Before using the assistant, please make sure the files on your hard drive are organized as follows:

- Media-files should be located in the same folder or in a folder beneath the transcription they belong to;
- Transcriptions that belong together (thematically or organizationally) should be held in the same directories;
- The Coma-file to be generated should be above these directories.

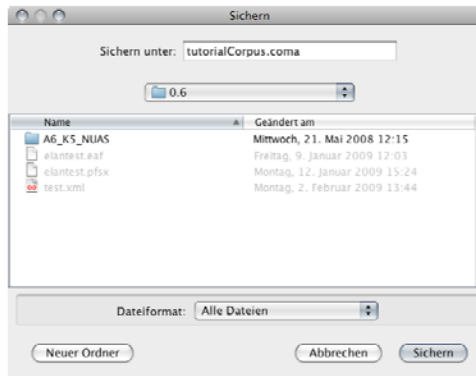
Using the Assistant

Name of corpus, coma file

The assistant to create a Coma-file from transcriptions will lead you through the necessary steps. When invoked, you should see the following screen:



Enter the name for your corpus into the “name of corpus” text field. Then click on the “...”-button and navigate to the directory **above all directories that contain transcriptions to be added to the corpus**. The filename will already be filled with the name you chose for the corpus – if you want to change it, you can do it now.

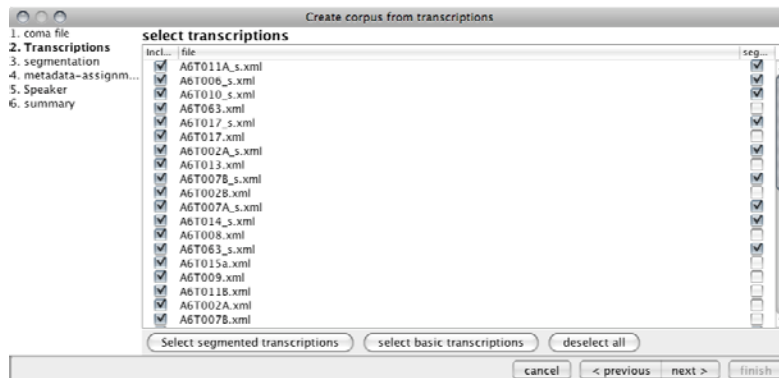


The assistant will then start to search all directories below the chosen one for transcriptions and analyze the metadata in their transcription-heads.



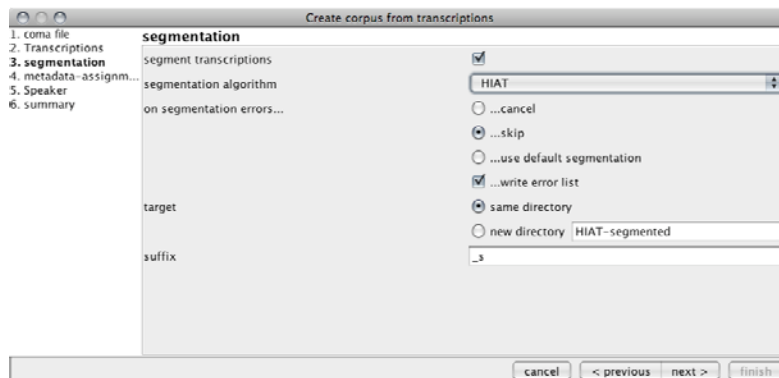
It will then show the number of transcriptions found below the directory-selection button. Click on the “next”-button to get to the selection of the transcriptions.

Select transcriptions



In this step, you can select the transcriptions to be included into the corpus. You can (de-)select single files with the checkboxes in the first column. The checkboxes in the last column are not clickable – they signal whether the transcription is a basic (unchecked) or segmented (checked) transcription. You can choose to include only segmented transcriptions by clicking on the “deselect all” and then on the “select basic transcriptions”-button.

When you are done with your selection, click “next”.



In this step you can decide to create segmented from basic transcriptions while creating the corpus if you have not already done so. Further information about basic- and segmented transcription can be found in “Understanding the basics of EXMARaLDA”.

Selecting the “segment transcriptions” checkbox enables the segmentation mechanism; all other options are only available when this checkbox is checked.

You select the segmentation-algorithm through the following drop-down-box.

The options grouped together under “on segmentation errors...” determine what happens when a segmentation goes sour:

If the “...cancel”-option is selected, the segmentation process stops on an error. The remaining transcriptions after the failed one will not be segmented.

If the “...skip”-option is selected, the erroneous transcription is skipped and the process continues.

If the “...use default segmentation” option is selected, the erroneous transcription will be segmented through an algorithm that is not likely to fail.

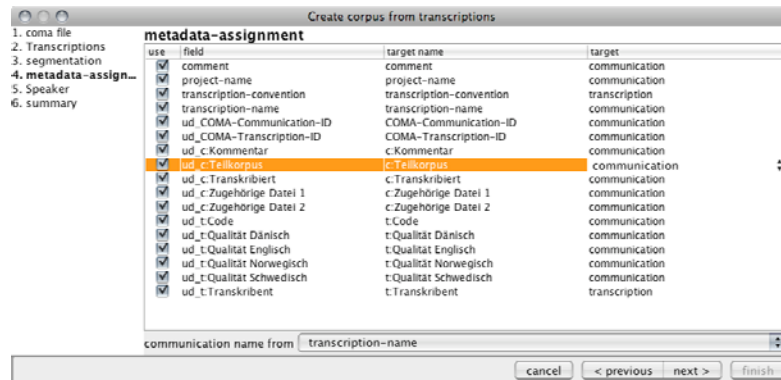
Checking the “...write error list”-box will produce an error-list inside the directory where the Coma-file will be written if there are any segmentation errors.

With the “target”-options, you can select where the segmented transcription files should be written to – either into the same directory as the basic transcriptions or into a subdirectory. The name of the subdirectory can be specified in the textfield.

In the “suffix”-textfield, you can specify a suffix that will be appended to the filenames of the segmented transcriptions.

When you are done with your segmentation settings, click “next”.

Metadata assignment



With “metadata-assignment” you decide, how the metadata found in the transcriptions (all fields found are shown in the second column of the table) will be used in the Coma-metadata-file.

By deselecting a checkbox in the first column, you can prevent fields to be included in the Coma metadata.

The second column displays all fields found in the heads of all transcription to be included in the corpus. User-defined fields have a “ud_” prepended to their name.

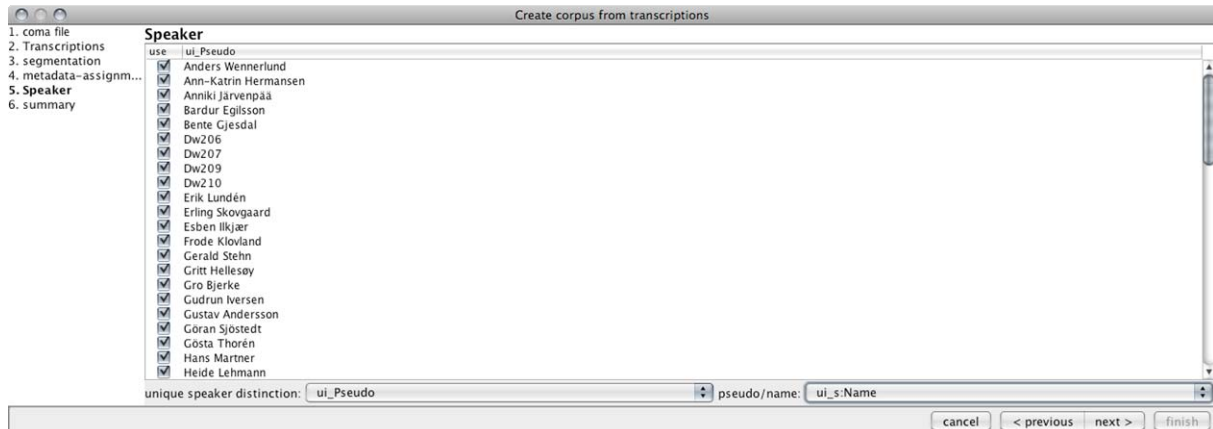
In the third column, you can decide which name these fields should have in the Coma-metadata. Normally, you will just use the given names (note that the assistant has already removed the “ud_” from the field names), but you can use this facility to harmonize field names that have been entered inconsistent in the Partitur-Editor: if you choose the same field name for two different source fields, they will come out as one field in the resulting coma file.

The fourth (“target”)-column is to select the target metadata-container for the source field (see “understanding coma metadata” for details on the different coma metadata containers). The possible targets are “communication” for fields associated to the communicative situation, “recording” for fields associated to the audio- or video-recording of the communication and “transcription” for fields associated to the actual transcription of the communication.

Sometimes, there is more than one transcription that belongs to a communication. With the drop-down-menu at the bottom of the screen, you can decide by which metadata field transcriptions are to be combined to one communication. You have to have a metadata-field that uniquely defines the communication a transcription will later be assigned to if you want to make use of that feature. If each transcription defines one communication, you can use “one file -> one communication” from the drop-down-menu.

If you are done making the metadata-assignments, click “next”.

Speaker-assignment



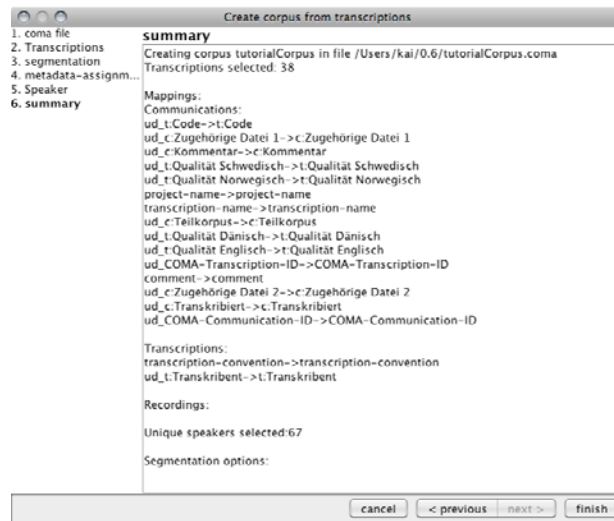
On the “Speaker”-screen you get a list of all speakers found in the speakertables of the found transcriptions. Speakers are by default distinguished by their id, which will in most cases not stand for a unique person. You can select the “unique speaker distinction” with the left drop-down-menu on the bottom of the screen. A change in that menu will be reflected in a different list of speakers right away, so you can try which metadata-field of the speakers in your corpus identifies a speaker uniquely.

If you have found the right metadata-field (you should do that step very thoroughly, since correcting an error afterwards is virtually impossible), you can deselect speakers that you do not want to be included in the Coma metadata by deselecting the checkbox in the first column of the table.

The second drop-down-menu is to select the metadata field that stands for a name or pseudo of the speaker. If you don't have such a metadata field, just select the same field as for the unique speaker distinction.

When you are done with your speaker selection, press “next”.

Summary



The next screen summarizes your choices on one screen. If you realize errors in the setup, just click on the “previous”-button to correct the corresponding setting. When you are comfortable with your settings, click the “finish”-button. If all goes well, the Coma-file is written and opened as the active document inside Coma right away.