How to import text transcription

This document explains how to import transcriptions of spoken language created with a text editor or a word processor into the Partitur-Editor using the "Simple EXMARaLDA" format. The Simple EXMARaLDA format is a format for plain text files that can handle transcriptions with some basic annotations, non-verbal behavior and overlapping speech.

Before you start reading this document, you should read:

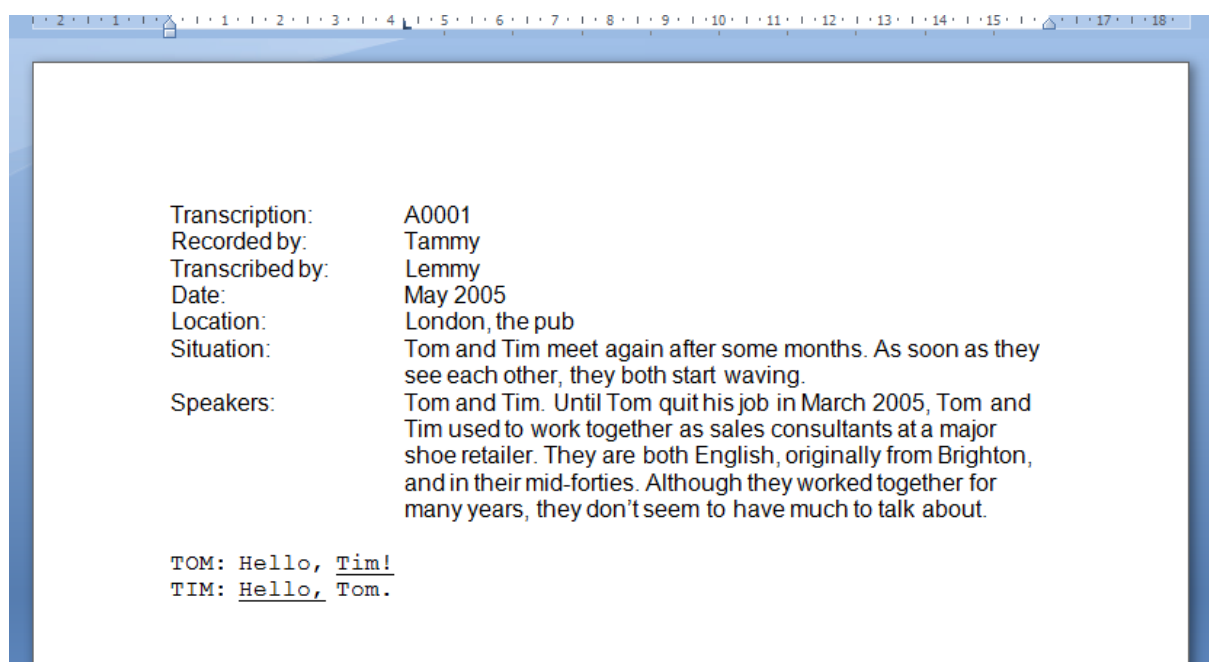- Understanding the basics of EXMARaLDA

Contents

A. Preparing the file for import

1. Source file information and structure

Of course, the easiest way to create a Simple EXMARaLDA file is to use the conventions described below right from the start. But if you've decided on using EXMARaLDA, you probably want to transcribe directly in the Partitur-Editor – therefore, the Simple EXMARaLDA format will most often be useful to convert some kind of legacy data. Depending on your transcriptions' layout and conventions, the conversion to the Simple EXMARaLDA format is a more or less simple task.

First of all, you need to verify which kind markup you've used to describe different kinds of information in the transcribed communication. Fully automatic conversion of your transcription format into the Simple EXMARaLDA format is only possible if you have used layout and/or mark-up in a consistent way, with all different kinds of information encoded differently and in a way that can be recognized without human interpretation. Revisit your transcription key to see if any ambiguous annotations or mark-up need to be adjusted manually.

In this short transcript, the only markup is the underlining used to mark overlapping speech. There is metadata about the communicative event and the two participating speakers written in the same document, but before the actual transcription starts.


```
Transcription:      A0001
Recorded by:        Tammy
Transcribed by:     Lemmy
Date:               May 2005
Location:           London, the pub
Situation:          Tom and Tim meet again after some months. As soon as they
                    see each other, they both start waving.
Speakers:           Tom and Tim. Until Tom quit his job in March 2005, Tom and
                    Tim used to work together as sales consultants at a major
                    shoe retailer. They are both English, originally from Brighton,
                    and in their mid-forties. Although they worked together for
                    many years, they don't seem to have much to talk about.

TOM: Hello, Tim!
TIM: Hello, Tom.
```

The Simple EXMARaLDA format can only handle the transcription itself – if you need to remove some preamble with metadata about the communication and speakers, remember to save a copy of the transcription with the respective metadata. The EXMARaLDA transcription formats include structures for metadata to let you encode e.g. the languages used in the communication or the location, or the L1(s) and L2(s) of each individual speaker. If you add this information properly in the Partitur-Editor, it can be used to filter the corpus to create a sub-corpus in the CorpusManager or for queries and analysis of the corpus in EXAKT.

2. The Simple EXMARaLDA format

A simple EXMARaLDA file is a text file that complies with the simple EXMARaLDA conventions described below.

Each line starts with the unique *speaker abbreviation* followed by a *colon and a space*. Please note speaker abbreviations are *case-sensitive*, i.e. "Tom" and "TOM" will be treated as different speakers. In this example transcription there are two speakers:

```
TOM:
TIM:
```

Since *each line will correspond to a separate event* in the EXMARaLDA file, it might be a good idea to put one utterance on each line. However, since a basic transcription will be created from the simple EXMARaLDA file, this will *not* result in a real segmentation.[1] Each line has to end with carriage return, additional empty lines, i.e. more than one carriage return, are allowed.

```
TOM: Hello, Tim!
TIM: Hello, Tom.
```

Text in *square brackets in front of the text* will end up as parallel events (with corresponding start and end points) in a *description tier*. This is suitable for non-verbal behavior, as in this example, where both speakers are waving while greeting each other.

```
TOM: [waving] Hello, Tim!
TIM: [waving] Hello, Tom.
```

Text in *curly brackets after the text* will end up as parallel events (with corresponding start and end points) in an *annotation tier*. This is suitable for other types of information, e.g. a translation. Please remember it's *only* possible to annotate the text in one line *as a whole*, i.e. the waving is carried out from the start until the end of these utterances, and the translation is not word-by-word, although the words happen to correspond in this particular case.

```
TOM: [waving] Hello, Tim! {Salut, Tim!}
TIM: [waving] Hello, Tom. {Salut, Tom!}
```

Overlapping speech is represented by *angle brackets*. The index (preferably a number) between the two closing brackets should be unique for this overlap, i.e. only used in each overlapping part to indicate they overlap each other.

```
TOM: [waving] Hello, <Tim!>1> {Salut, Tim!}
TIM: [waving] <Hello,>1> Tom. {Salut, Tom!}
```

Since square, curly and angle brackets carry meaning in the Simple EXMARaLDA format, they can only occur in the transcription with this meaning.

3. Converting files to the Simple EXMARaLDA format

Since the conversion to the Simple EXMARaLDA format depends on the source format and transcription conventions, the transformation of transcription files into the Simple EXMARaLDA format can't be described in general. And although the conversion is of course

---

[1] Please remember, if you want a segmentation of the transcription you need to use the EXMARaLDA segmentation function. Please refer to the document How to use segmentation.

preferably done automatically, automatic conversion of transcriptions is always somewhat risky. Even if the conversion steps are correct considering the transcription conventions, the correctness of the files created in a word processor or text editor was perhaps never assessed, and so there might be errors in the transcription, and these might change the contents of the converted file. Post-editing will solve many problems, but for complex transcription formats of unknown quality, the amount of post-editing necessary to correct converted files due to errors in the original transcription files might be too high considering the high amount of time spent on defining the conversion process for the complex file format. In these cases it might be better to focus on some parts of the formats and e.g. add most annotations manually.

4. Plain text

Since the Partitur-Editor requires plain text (extension .txt) as input format, not Word, PDF, etc., you need to save your file in txt format somewhere along the conversion process. If you've been using formatting information (e.g. instances of bold or italic text) as mark-up and/or for annotations or to indicate speakers (e.g. Tim's utterances in blue color, Tom's in yellow), you need to replace the formatting with the corresponding Simple EXMARaLDA markup or at least replace all instances with some plain text mark-up before you carry out this step. Microsoft Word and OpenOffice Writer both have a regular expression option in the Find and Replace function that will let you search for and replace formatting, and use the found expression as part of the replacing expression (e.g. to add start and end tags around it).

5. Adding tiers

As the annotation in the curly brackets always refer to all of the text in the same line, it might be better to be split the transcription according to existing annotations to avoid extensive post-editing. In the example below, Tom's utterance was split into three lines to create an annotation for the two words "see you".
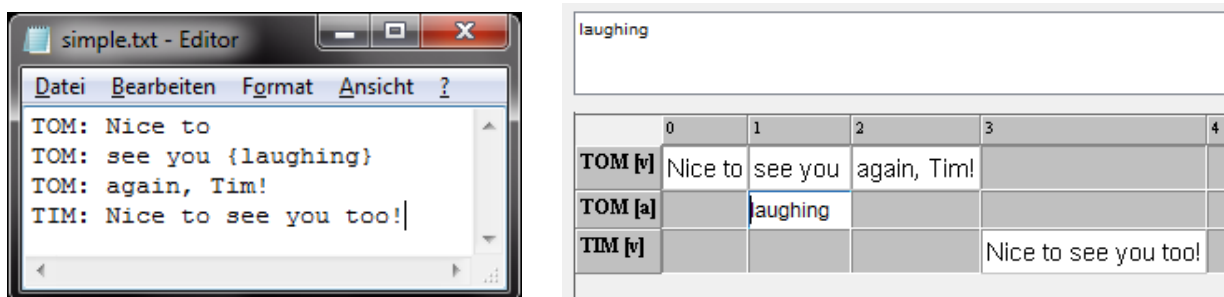

Another important detail is that only the "curly bracket annotations" will end up in an annotation tier, whereas the text in square brackets will go in a tier with the type description. You can use the format in other ways than intended, but please be aware of the consequences. For example, since the information in description tiers are treated differently from annotations by the EXMARaLDA tools, you'll have to change the type of this tier (edit **Tier Properties** in the **Tier** menu) after import to be able to use the tools as intended if you add additional annotations this way.

B. Importing the file into the Partitur-Editor

Import the text into the Partitur-Editor by choosing **Import** in the **File** menu. First locate the text file you want to import. Then make sure you've chosen the right filter, i.e. for the file type **Simple EXMARaLDA file (*.txt)** and the appropriate **Char encoding**, i.e. the same character encoding as in your text file. If you don't know the character encoding, first try the default choice **system-default**:
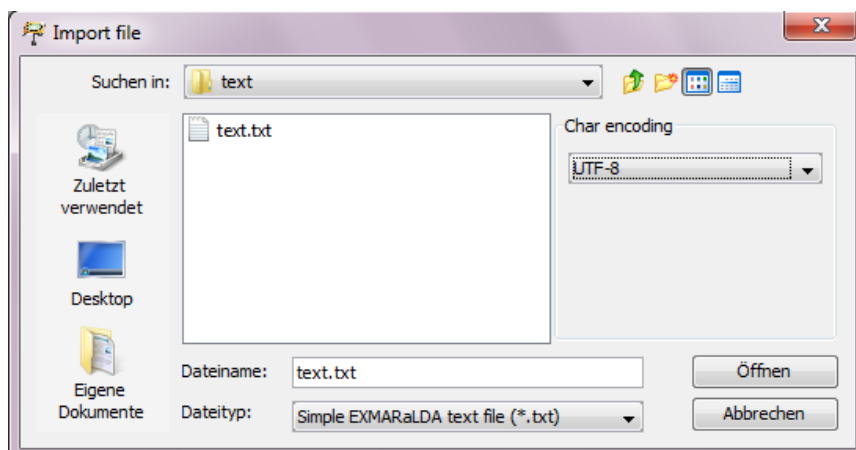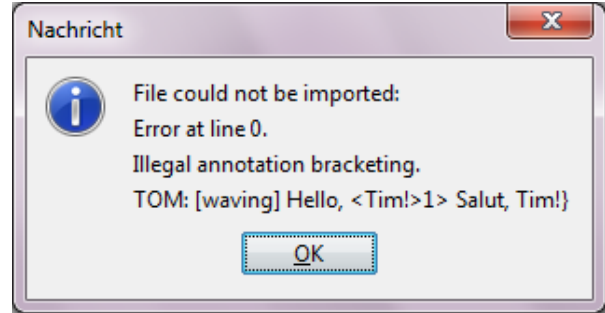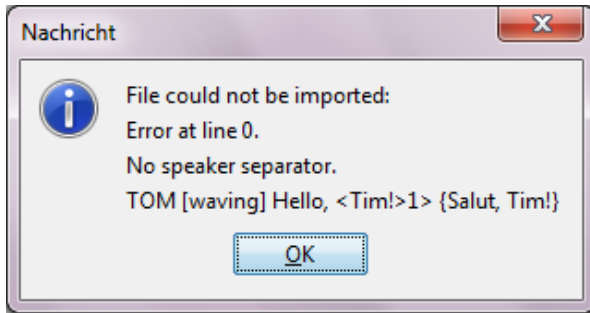

 If the chosen character encoding doesn't match the one of your file, special characters might not display properly after import. Should this happen, try saving your text file with another encoding, e.g. UTF-8. This is done by e.g. choosing **Save as...** in Notepad under Windows and then specifying the encoding. Then try to import the file again with the chosen character encoding.
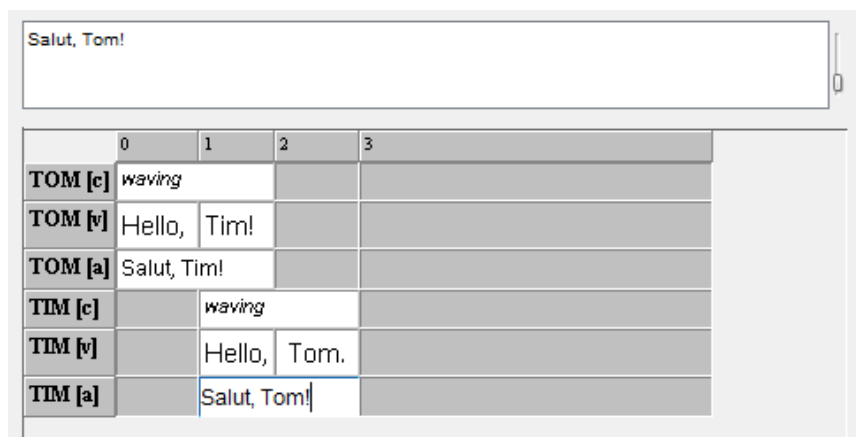
Then *save your transcription* in .exb-format (EXMARaLDA basic transcription) *before* you start adding metadata or editing the transcription.

If there is something wrong with the file, you may get an error message with three lines that will help you correct the mistake. The first line contains the line number where the (first) error was encountered, the second line contains an error type, e.g. "no speaker separator" meaning the colon separating the speaker abbreviation from the text is missing, and the third is the erroneous line itself, where you can see the error. Make sure the file complies with the Simple EXMARaLDA conventions described above and then try to import the file again.

Nachricht — File could not be imported:
Error at line 0.
No speaker separator.
TOM [waving] Hello, <Tim!>1> {Salut, Tim!}
OK

Nachricht — File could not be imported:
Error at line 0.
Illegal annotation bracketing.
TOM: [waving] Hello, <Tim!>1> Salut, Tim!}
OK

1. Post-Editing

If you've used both the annotation and the description tiers for annotations you need to change the tier type from (D)escription to (A)nnotation for the information you put in square brackets in the text transcription. If you've put different kinds of information into one annotation tier and want to move some of them into another annotation tier, i.e. to have one tier for comments on pronunciation and one tier for other comments you can use the feature **Copy events from** with the **Copy text** checkbox checked when adding further annotation tiers, thus copying event boundaries and contents of the first tier into the one you're creating.

2. Metadata

Don't forget to add all metadata on the communication (**Transcription > Meta information...**) and the speakers (**Transcription > Speakertable...**)!
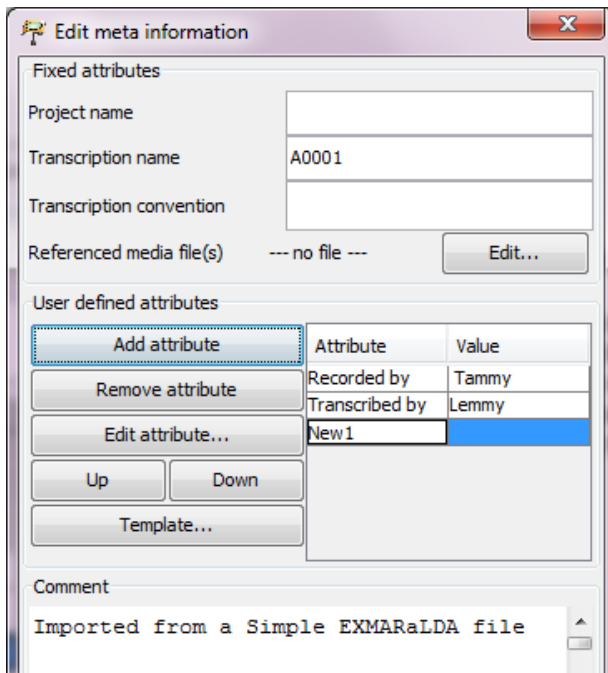
Metadata from the original transcript is added as attribute-value pairs to the EXMARaLDA Transkription.

Metadata on the speaker is added separately, in the speakertable as shown below: