

How to:
Importieren einer Texttranskription

Dieses Dokument erklärt, wie Transkriptionen der gesprochener Sprache, die mithilfe eines Texteditors oder Textverarbeitungsprogramms erstellt wurden, in den Partitur-Editor (als "Simple EXMARaLDA" Format) importiert werden. Simple EXMARaLDA ist ein Format für einfache Textdateien, das auch einige grundlegende Annotationen, nicht-verbale Phänomene sowie Überlappungen verarbeiten kann.

Vor dem Lesen dieser Anleitung wird empfohlen, folgendes Dokument zu lesen:

- Understanding the basics of EXMARaLDA

Inhalt

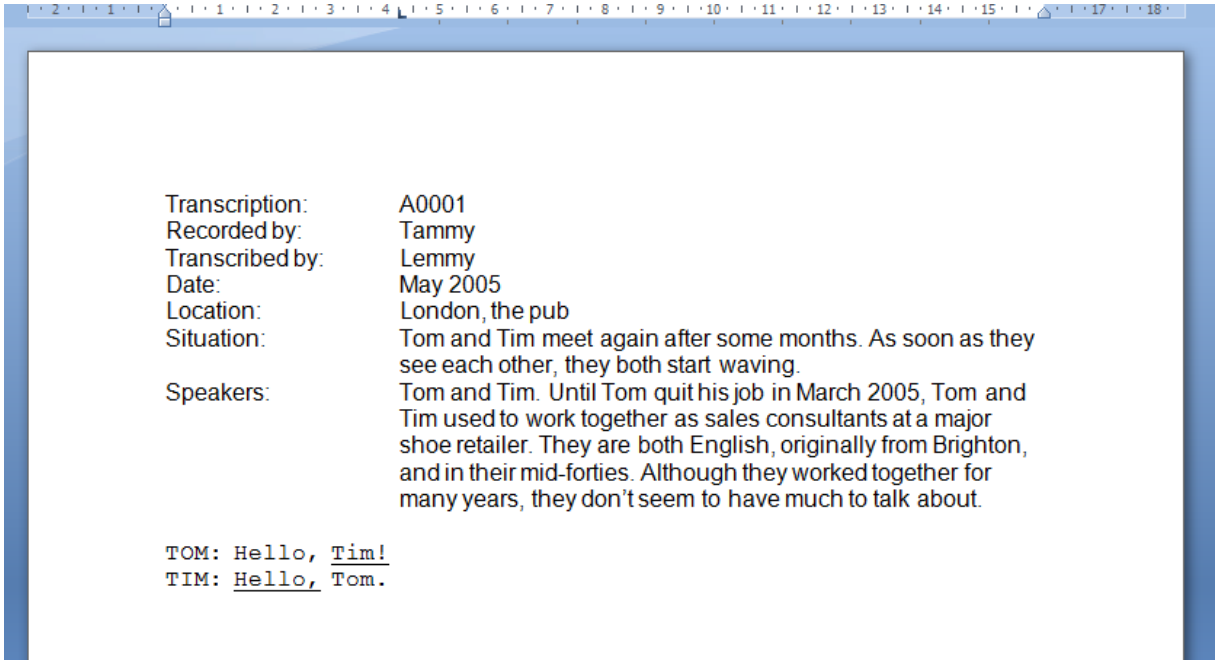
A. Vorbereitung der Datei für den Import

1. Struktur und Informationen der Quelldatei

Der wohl einfachste Weg, eine Simple EXMARaLDA Datei zu erstellen, ist die unten beschriebenen Konventionen von Anfang an zu verwenden. Wenn Sie sich jedoch bereits für EXMARaLDA entschieden haben, wollen Sie sicherlich die Transkripte direkt in den Partitur-Editor eingeben. Das Simple EXMARaLDA Format ist im Normalfall die beste Lösung, um verschiedene Altdaten zu konvertieren. Je nach Gestaltung der Transkription und den darin verwendeten Konventionen, ist die Konvertierung in das Simple EXMARaLDA Format eine mehr oder weniger einfache Aufgabe.

Als erstes muss festgestellt werden welche Art Markup verwendet wurde, um verschiedene Informationen in der transkribierten Kommunikation zu beschreiben. Eine vollautomatische Konvertierung des Transkriptionsformats in das Simple EXMARaLDA Format ist nur dann möglich, wenn das Layout und/oder Markup auf eine konsistente Weise verwendet wurde - sodass unterschiedliche Informationen, die verschieden kodiert sind, ohne menschliche Interpretation entschlüsselt werden können. Gegebenenfalls muss, um festzustellen ob irgendwelche mehrdeutigen Annotationen oder Markups manuell angepasst werden müssen, im Transkriptionsschlüssel nachgesehen werden.

In diesem kurzen Transkript ist das einzige verwendete Markup die Unterstreichung, die die überlappende Rede markiert. Vor dem eigentlichen Beginn der Transkription gibt es im gleichen Dokument Metadaten über das kommunikative Ereignis und die beiden teilnehmenden Sprecher.



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.

Das Simple EXMARaLDA Format kann nur die Transkription verarbeiten. Wenn eine Einleitung mit z.B. Metadaten zur Kommunikation und Sprechern entfernt wird, sollte man daran denken, eine Kopie der Transkription mit den jeweiligen Metadaten zu speichern. Die EXMARaLDA-Transkriptionsformate umfassen auch Strukturen für Metadaten, mit denen z.B. die in der Kommunikation verwendeten Sprachen, Orte oder die L1 und L2 der einzelnen Sprecher verschlüsselt werden können. Wenn diese Information im Partitur-Editor korrekt

eingegeben wird, kann sie auch verwendet werden, um z.B. ein Korpus zu filtern, ein Subkorpus im Corpus Manager zu erstellen oder um Korpusabfragen und Analysen in EXAKT durchzuführen.

2. Das Simple EXMARaLDA Format

Eine Simple EXMARaLDA Datei ist eine Textdatei, die den unten beschriebenen Simple EXMARaLDA Konventionen entspricht.

Jede Zeile beginnt mit einem individuellen *Sprecherkürzel*, gefolgt von einem *Doppelpunkt und einer Leertaste*. Dabei ist zu beachten, dass hierbei die *Groß-/Kleinschreibung* unterschieden wird, z.B.. „Tom“ und „TOM“ werden als zwei unterschiedliche Sprecher behandelt. In dieser Beispiel-Transkription gibt es zwei Sprecher:

```
TOM:
TIM:
```

Da *jede Zeile genau einem separaten Ereignis* in der EXMARaLDA-Datei entspricht, ist es nützlich, jede Äußerung in eine separate Zeile zu setzen. Da jedoch die Basis-Transkription aus der Simple EXMARaLDA Datei erstellt wird, findet *keine* tatsächliche Segmentierung[1] statt. Jede Zeile muss mit einem Zeilenumbruch enden, zusätzliche leere Zeilen, z.B. mit mehr als einem Zeilenumbruch, sind erlaubt.

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

Der Text in *eckigen Klammern vor dem Text* wird als paralleles Ereignis (mit korrespondierenden Start- und Endpunkten) in die *Kommentarspur* eingefügt. Dieses eignet sich für nonverbales Verhalten, wie in diesem Beispiel, wo beide Sprecher winken während sie einander grüßen.

.

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

Der Text in *geschweiften Klammern nach dem Text* wird als paralleles Ereignis (mit korrespondierenden Start- und Endpunkten) in die *Annotationsspur* eingefügt. Dies ist für andere Arten von Informationen, wie z.B. Übersetzung, geeignet. Bitte beachten Sie, dass es *nur* möglich ist den Text *in einer Zeile als Ganzes* zu annotieren, sodass in diesem Beispiel das Winken von Anfang bis Ende der Äußerungen stattfindet und keine Wort-für-Wort Übersetzung vorliegt, obwohl die Wörter in diesem konkreten Fall übereinstimmen.

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

Überlappende Rede ist durch *spitze Klammern* markiert. Die Kennziffer (vorzugsweise eine Zahl) zwischen den beiden schließenden Klammern muss für jede Überlappung einzigartig sein, d.h. sie kann nur in jedem der überlappenden Teile verwendet werden, um anzuzeigen, dass sie einander überlappen.

[1] Wenn Sie eine Segmentation der Transkription benötigen, müssen Sie die EXMARaLDA Segmentationsfunktion anwenden. Näheres dazu gibt es in der Anleitung „How to use segmentation“.

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

Da die eckigen, geschweiften und spitzen Klammern Bedeutung im Simple EXMARaLDA Format haben, können sie in der Transkription ausschließlich mit dieser Bedeutung auftreten.

3. Konvertierung von Dateien in das Simple EXMARaLDA Format

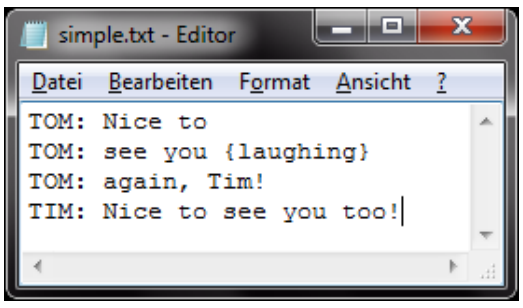
Da die Konvertierung in das Simple EXMARaLDA Format von dem ursprünglichen Dateiformat und den Transkriptionskonventionen abhängt, kann die Transformation von Transkriptionsdateien in das Simple EXMARaLDA Format nicht im Allgemeinen beschrieben werden. Obwohl die Konvertierung vorzugsweise automatisch erfolgen sollte, ist jede solche automatische Konvertierung von Transkriptionen immer ein wenig riskant. Selbst wenn die Konvertierungsschritte als fehlerfrei im Hinblick auf die Transkriptionskonventionen erscheinen, kann es sein, dass die Richtigkeit der mit einem Texteditor oder Textverarbeitungsprogramm erstellten Datei noch nicht geprüft wurde. Dies kann zur Folge haben, dass es Fehler in der Transkription gibt, die wiederum die Inhalte der konvertierten Datei ändern. Nachbearbeitung kann einige von diesen Problemen lösen. Für komplexe Transkriptionsformate von unbekannter Qualität kann der Aufwand der notwendigen Nachbearbeitung, um die konvertierten Dateien in Bezug auf Fehler in der ursprünglichen Datei zu korrigieren, zu hoch sein oder auch angesichts des hohem Zeitaufwands für das Definieren vom Konvertierungsprozess eines komplexen Dateiformates. In diesen Fällen kann es besser sein, sich nur auf einige Teile der Formate zu konzentrieren und z.B. die meisten Annotationen manuell einzutragen.

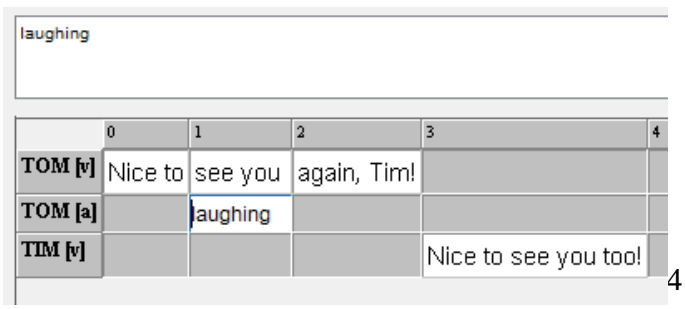
4. „Nur-Text“ (Plain text)

Da der Partitur-Editor Input-Dateien im „Nur-Text“ ("Plain Text", Suffix .txt) Format erfordert und nicht etwa Word, PDF o.Ä., muss diese Datei im .txt Format im Laufe des Konvertierungsprozesses gespeichert werden. Wurden bislang bestimmte Formatierungen (z.B. fett oder kursiv markierte Textstellen) als Markup und/oder für Annotationen oder Sprecher-Kennzeichnung (z.B. Tims Äußerungen in blauer, Toms in gelber Farbe) verwendet, so müssen diese mit dem entsprechenden Simple EXMARaLDA Markup oder zumindest alle Stellen mit einem „Nur-Text“ (Plain Text) Markup, noch vor der Konvertierung ersetzt werden. Microsoft Word und OpenOffice verfügen über eine Option zu regulären Ausdrücken (in der Suchen und Ersetzen Funktion), die Suchen und Ersetzen von Formatierung sowie Verwenden von gefundenen Ausdrücken als Teil des zu ersetzenden Ausdrucks (z.B. hinzufügen von Start-und End-Tags) ermöglicht.

5. Hinzufügen von Spuren

Da sich die Annotation in den geschweiften Klammern immer auf den gesamten Text in der gleichen Zeile bezieht, kann es besser sein, die Transkription anhand von existierenden Annotationen zu splitten um umfangreiche Nachbearbeitung zu vermeiden. Im Beispiel unten wurde Toms Äußerung auf drei Zeilen verteilt, um eine Annotation für die zwei Wörter "See you" zu erstellen.

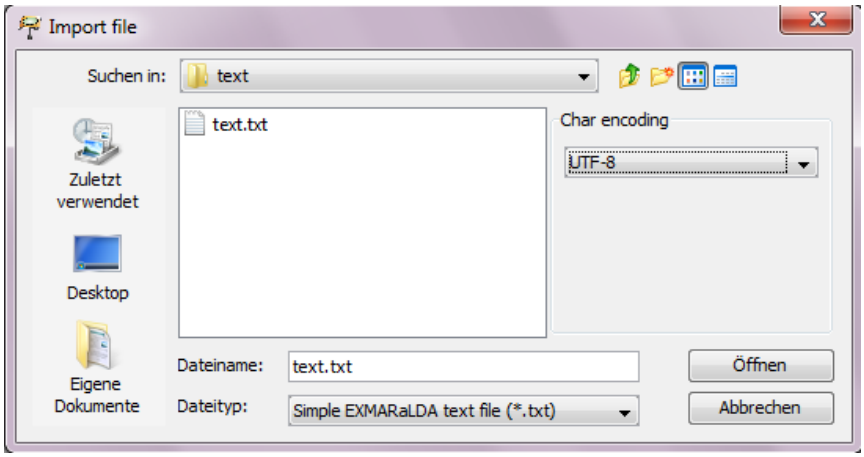




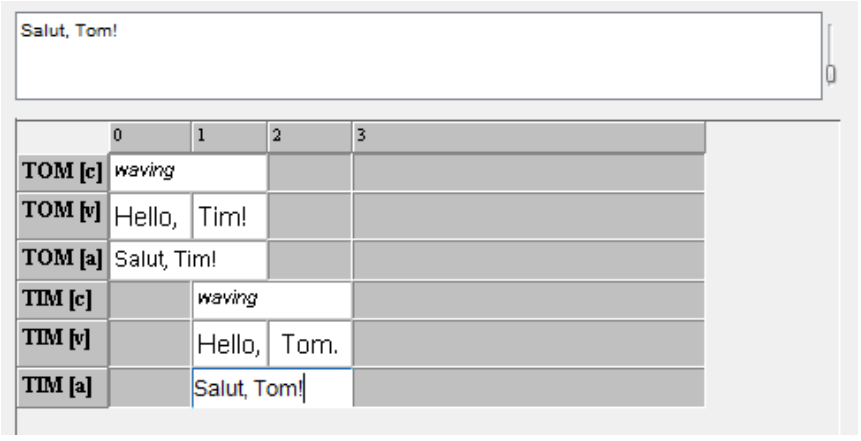
Ein weiteres wichtiges Detail ist, dass nur die Annotation in den geschweiften Klammern am Ende eine Annotationsspur ergibt, wobei der Text in eckigen Klammern in eine Kommentarspur mit Typen-Bezeichnung umgewandelt wird. Das Format kann auch auf andere Weise als beabsichtigt verwendet werden, dabei sollten jedoch mögliche Konsequenzen beachtet werden. Beispielsweise müssen nach dem Import, da die Informationen in den Kommentarspuren anders als Annotationen mit EXMARaLDA - Werkzeugen behandelt werden, zunächst die Spurtypen geändert werden (Punkt **Tier Properties** im Menü **Tier**), um die Werkzeuge wie beabsichtigt benutzen zu können wenn zusätzliche Annotationen auf diese Weise hinzugefügt werden sollen.

B. Importieren der Datei in den Partitur-Editor

Der Import des Textes in den Partitur-Editor erfolgt über den Punkt **Import** aus dem **Datei** Menü. Suchen Sie zuerst die Datei, die Sie importieren möchten. Stellen Sie dann sicher, dass Sie den richtigen Filter ausgewählt haben, z.B. für den Dateityp **Simple EXMARaLDA file (*.txt)** und die entsprechende **Char encoding**, zum Beispiel die gleiche Zeichenkodierung wie in der Textdatei. Wenn Sie nicht wissen, welche Zeichenkodierung verwendet wurde, versuchen
Sie es zuerst mit der Default- Auswahl (**system-default**):

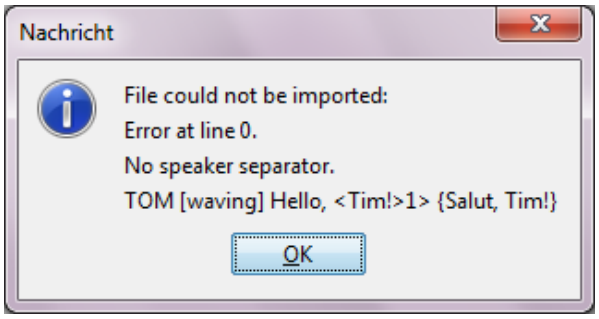


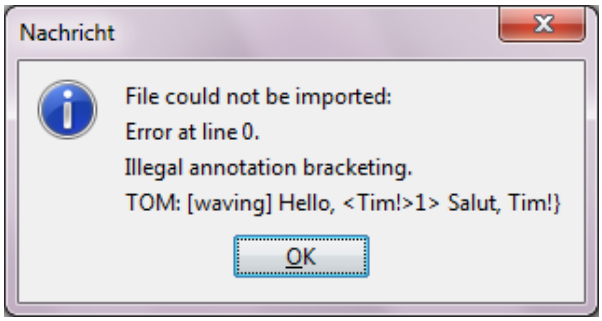
Wenn die ausgewählte Zeichenkodierung nicht mit der Ihrer Datei übereinstimmt, können die Sonderzeichen nach dem Import nicht mehr richtig angezeigt werden. Sollte dies geschehen, versuchen Sie Ihre Textdatei mit einer anderen Zeichenkodierung zu speichern, z.B. UTF-8. Dies kann z.B. durch Auswählen von **Speichern unter...** im Notepad (Betriebssystem Windows) und dann durch Spezifizierung der Zeichenkodierung gemacht werden. Versuchen Sie dann die Datei erneut mit der gewählten Zeichenkodierung zu importieren.



Als nächstes *speichern Sie die Transkription* im .exb Format (EXMARaDLA Basis-Transkription) *bevor* Sie mit der Eingabe von Metadaten oder Editieren der Transkription starten.

Sollte die Datei noch nicht in Ordnung sein, erscheint eine Fehlermeldung mit drei Zeilen, die bei der Korrektur des Fehlers hilft. Die erste Zeile enthält die Zeilennummer, in der der (erste) Fehler aufgetreten ist, die zweite Zeile sagt etwas über den Fehlertyp aus, z.B. "no speaker separator", also der das Sprecherkürzel vom Text trennende Doppelpunkt fehlt; und die dritte ist die fehlerhafte Zeile selbst in der Sie den Fehler sehen können. Achten Sie darauf, dass die Datei dem Simple EXMARaLDA Format und den oben beschriebenen Konventionen entspricht und versuchen Sie dann, die Datei erneut zu importieren.



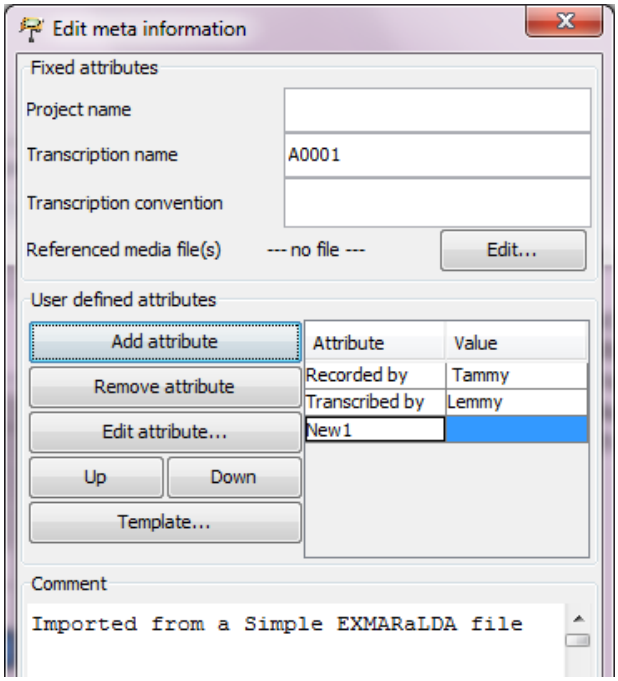


1. Nachbearbeitung (Post-Editing)

Wenn Sie Annotations- und Kommentarspuren für Annotationen verwendet haben, müssen Sie den Spur-Typ von (D)escription in (A)nnotation ändern. Wenn Sie verschiedene Arten von Informationen in eine Annotationsspur gesetzt haben und jetzt einige von ihnen in eine andere Spur bewegen wollen, z.B. um eine Spur für Kommentare zur Aussprache und eine Spur für weitere Kommentare zu haben, können Sie beim Hinzufügen von weiteren Annotationsspuren die Option **Ereignisse kopieren aus** mit dem aktivierten Kontrollkästchen **Copy text** benutzen - dabei werden die Ereignisgrenzen samt Inhalte aus der ersten Spur in die nächste übertragen.

2. Metadata

Bitte vergessen Sie nicht alle Metadaten zu der Kommunikation (**Transkription > Meta-Information...**) und den Sprechern hinzuzufügen (**Transkription > Sprechertabelle...**)!



Metadaten aus der ursprünglichen Transkription werden als Attribut-Wert-Paare zur EXMARaLDA Transkription hinzugefügt.

Metadaten zu den Sprechern werden separat in die Sprechertabelle eingetragen.