|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
| **How to:**  **Importieren einer Texttranskription** | | |
|  | | |

Dieses Dokument erklärt, wie Sie Transkriptionen gesprochener Sprache, die mithilfe eines Texteditors oder Textverarbeitungsprogramms erstellt wurden, in den Partitur-Editor (als Simple EXMARaLDA" Format) importieren können. Simple EXMARaLDA ist ein Format für einfache Textdateien, das auch einige grundlegende Annotationen, nicht-verbale Phänomene und Überlappungen verarbeiten kann. Bevor Sie dieses Dokument lesen, sollten Sie sich mit „Understanding the basics of EXMARaLDA“ auseinander setzen.

Inhalt

A. Vorbereitung der Datei für den Import 2

1. Struktur und Informationen der Quelldatei 2

2. Das Simple EXMARaLDA Format 3

3. Konvertierung von Dateien in das Simple EXMARaLDA Format 4

4. „Nur-Text“ (Plain text) 4

5. Hinzufügen von Spuren 4

B.Importieren der Datei in den Partitur-Editor 5

1. Nachbearbeitung (Post-Editing) 6

2. Metadaten 7

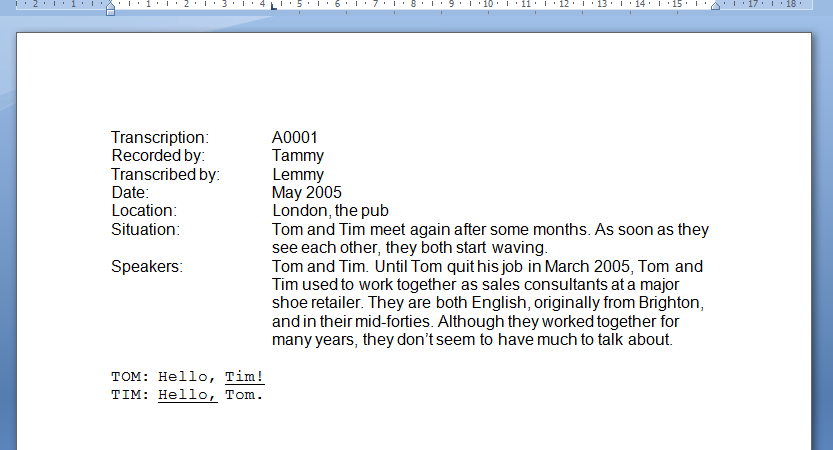
# A. Vorbereitung der Datei für den Import

## 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, wünschen Sie sicherlich die Transkripte direkt in den Partitur-Editor einzugeben. 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.

Zunächst muss festgestellt werden, welche Art Markup verwendet wurde, um verschiedene Informationen in der transkribierten Kommunikation beschreiben zu können. 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, verschieden kodierte Informationen ohne menschliche Interpretation entschlüsselt werden können. Um festzustellen, ob einige mehrdeutige Annotationen oder Markups manuell angepasst werden müssen, muss gegebenenfalls im Transkriptionsschlüssel nachgesehen werden.

In dem folgenden kurzen Transkript ist das einzige verwendete Markup die Unterstreichung, die zur Markierung überlappender Rede genutzt wird. Vor dem eigentlichen Beginn der Transkription gibt es im selben Dokument Metadaten über das kommunikative Ereignis und die teilnehmenden Sprecher.



Das Simple EXMARaLDA Format kann nur die Transkription verarbeiten. Deshalb ist es ratsam eine Kopie der Transkription mit den entsprechenden Metadaten zu speichern, wenn die Transkription durch Metadaten eingeleitet wird. 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 beispielsweise ein Korpus zu filtern, ein Subkorpus im Corpus Manager zu erstellen oder um Korpusabfragen und Analysen in EXAKT durchzuführen.

## 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 zwischen *Groß-/Kleinschreibung* unterschieden wird: „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]](#footnote-1)statt. Jede Zeile muss mit einem Zeilenumbruch enden, zusätzliche leere Zeilen, mit zum Beispiel 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 Kommentarspureingefügt. Dieses Vorgehen eignet sich für nonverbales Verhalten. In unserem Beispiel zur Beschreibung, dass beide Sprecher winken, während sie einander grüßen.

TOM: [waving] Hello, Tim!

TIM: [waving] Hello, Tom.

Der Text in geschweiften Klammern nach dem Textwird als paralleles Ereignis (mit korrespondierenden Start- und Endpunkten) in die Annotationsspureingefügt. Dies ist für andere Arten von Informationen, wie z.B. Übersetzungen, 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 Überlappungen anzuzeigen.

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.

## Konvertierung von Dateien in das Simple EXMARaLDA Format

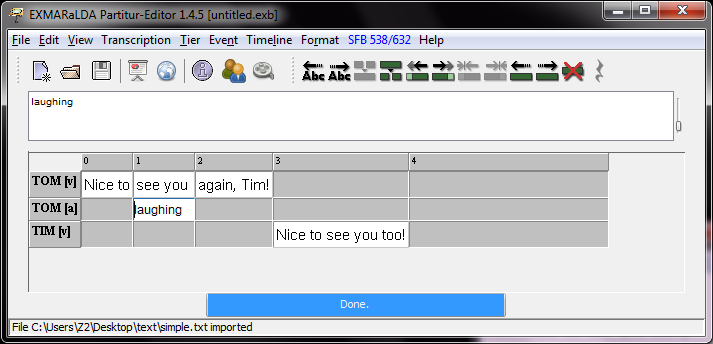
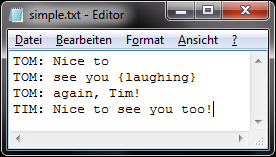
Da die Konvertierung in das Simple EXMARaLDA Format von dem ursprünglichen Dateiformat und den Transkriptionskonventionen abhängt, kann die Transformation von Transkriptionsdateien in das Simple EXMARaLDA Format nicht allgemeingültig beschrieben werden. Obwohl die Konvertierung vorzugsweise automatisch 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 und für das Definieren vom Konvertierungsprozess eines komplexen Dateiformates zu hoch sein. In diesen Fällen ist es vorteilhafter, sich nur auf bestimmte Teile der Formate zu konzentrieren und z.B. die meisten Annotationen manuell einzutragen.

## „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.

## Hinzufügen von Spuren

Da sich die Annotation in den geschweiften Klammern immer auf den gesamten Text in der gleichen Zeile bezieht, sollten Sie in Erwägung ziehen die Transkription anhand von existierenden Annotationen zu splitten, um umfangreiche Nachbearbeitung zu vermeiden. In unserem Beispiel wurde Toms Äußerung auf drei Zeilen verteilt, um eine Annotation für die zwei Wörter "See you" zu erstellen.

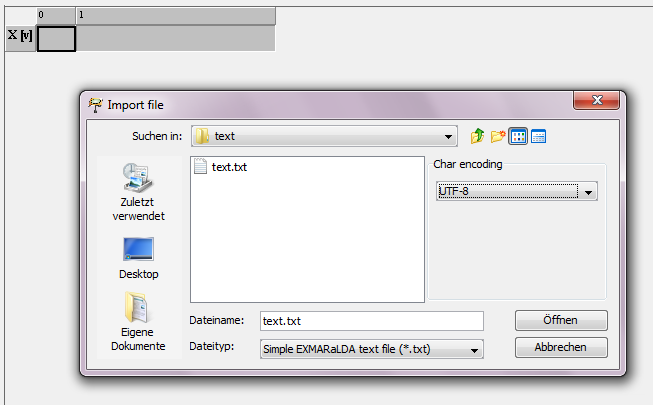
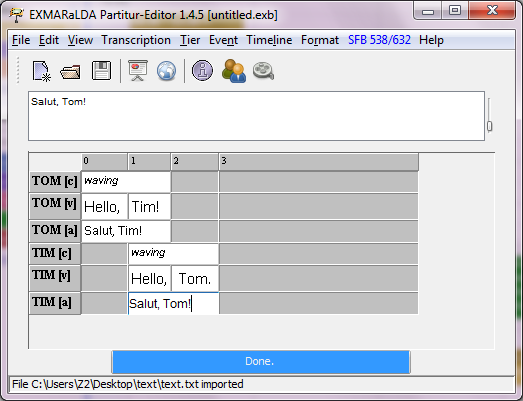


Ein weiteres wichtiges Detail ist, dass nur die Annotation in den geschweiften Klammern am Ende eine Annotationsspur ergibt, wobei der Text in eckigen Klammern in eine Kommentarspur mit Typen-Bezeichnung umgewandelt wird. Das Format kann auch auf andere Weise als beabsichtigt verwendet werden, dabei sollten jedoch mögliche Konsequenzen beachtet werden. Die Informationen in den Kommentarspuren werden anders behandelt als Annotationen mit EXMARaLDA Werkzeugen. Daher müssen beispielsweise nach dem Import zunächst die Spurtypen (via Tier Properties im Menü Tier) geändert werden. Erst dann ist es möglich die Werkzeuge wie gewohnt zu nutzen.

# B.Importieren der Datei in den Partitur-Editor

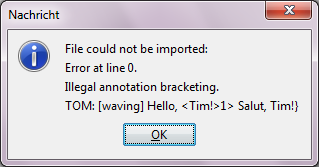
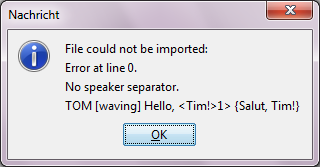
Der Import des Textes in den Partitur-Editor erfolgt über den Punkt Import aus dem Datei Menü. Suchen Sie zunächst die zu importierende Datei und stellen Sie sicher, dass Sie den richtigen Filter ausgewählt haben, also beispielsweise für den Dateityp Simple EXMARaLDA file (\*.txt) und die entsprechende Char encoding. Letzteres sollte die gleiche Kodierung sein, die auch in der Textdatei verwendet wurde. Wenn Sie sich dessen nicht sicher sind, versuchen Sie es zunächst 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 vollzogen werden. Versuchen Sie dann erneut Ihre Datei mit der gewählten Zeichenkodierung zu importieren.



Im nächsten Schritt 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 fehlerhaft sein, erscheint eine Fehlermeldung. Die erste Zeile gibt an, in welcher Zeilennummer der erste Fehler auftritt. Die zweite gibt Informationen über den Fehlertyp, beispielsweise „no speaker separator“, was bedeutet, dass ein Doppelpunkt zwischen dem Sprecherkürzel und dem Text fehlt. Die dritte Zeile ist die fehlerhafte Zeile selbst. Achten Sie darauf, dass die Datei dem Simple EXMARaLDA Format und den oben beschriebenen Konventionen entspricht und versuchen Sie dann erneut die Datei zu importieren.

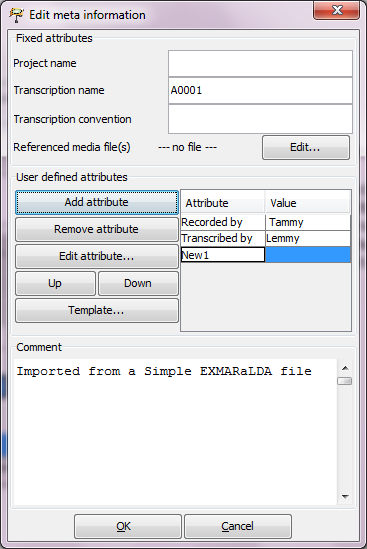


## Nachbearbeitung (Post-Editing)

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 einige davon in eine andere Spur bewegen möchten, weil Sie zum Beispiel eine Spur für Kommentare bezüglich der Aussprache wünschen, können Sie die Option Ereignisse kopieren aus (das Kontrollkästchen Copy text sollte angewählt sein) zum Hinzufügen weiterer Annotationspuren nutzen. Durch dieses Vorgehen werden die Ereignisgrenzen mit ihren Inhalten aus der ersten Spur in die nächste übertragen.

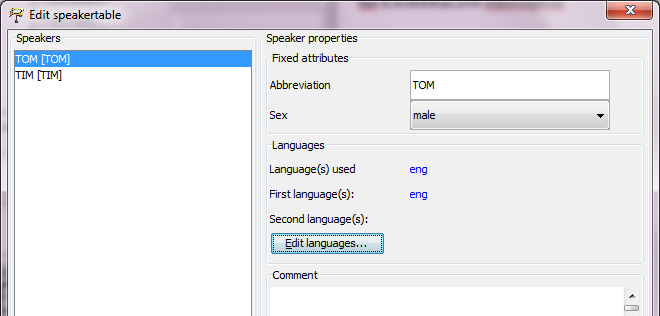
## Metadaten

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.



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“. [↑](#footnote-ref-1)