|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
| **Howto:**  **Verwendung des Partitur-Editors mit geschriebenen Daten** | | |
|  | | |

Dieses Dokument erläutert die Verwendung vom EXMARaLDA Partitur-Editor bei der Arbeit mit geschriebenen Daten.

Diese Anweisungen gelten für „gewöhnliche" (geschriebene) Texte, d.h. nicht für Transkriptionen gesprochener Sprache. Für weitere Informationen über Transkriptionen, die mit einem Texteditor oder einem Textverarbeitungsprogramm erstellt wurden, lesen Sie bitte die Sektion „Simple EXMARaLDA Format" im Dokument „How to import texttranscriptions“.

Bei einigen Abschnitten handelt es sich um Anweisungen, wie einige Import-Funktionen für einen „individuellen" Text-Import verwendet werden können, d.h. auf eine Weise, die ursprünglich nicht vorgesehen war. Diese Informationen befinden sich in grau unterlegten Feldern. Sollten Sie keine zusätzlichen Informationen benötigen, können Sie diese überspringen.

Bevor Sie beginnen, dieses Dokument zu lesen, konsultieren Sie:

* „Understanding the basics of EXMARaLDA”

Inhalte

[A. Optionen für den Text-Import in EXMARaLDA 2](#_Toc372193181)

[1. Importieren von Nur-Text (‘Plain Text’) 2](#_Toc372193182)

[2. Importieren von TreeTagger Output 4](#_Toc372193183)

[3. Importieren des Simple EXMARaLDA Format 7](#_Toc372193184)

[B. Der SFB 632 EXMARaLDA-importer 7](#_Toc372193185)

[C. Annotieren des Textes 8](#_Toc372193186)

[1. Der AUT Sprecher 8](#_Toc372193187)

[2. Annotationsspuren 8](#_Toc372193188)

[3. Annotieren im Partitur-Editor 9](#_Toc372193189)

[D. Segmentierung 10](#_Toc372193190)

# A. Optionen für den Text-Import in EXMARaLDA

Der Text-Import kann im EXMARaLDA Partitur-Editor auf drei unterschiedlichen Wegen erfolgen: neben dem Import von Nur Text Dateien können auch möglich (Text-)Dateien, die mithilfe vom TreeTagger erstellt wurden und Textdateien im Simple EXMARaLDA Format (spezielles Format für Transkriptionen, die mit einem Texteditor oder Textverarbeitungsprogramm erstellt wurden) importiert werden. Die folgenden Anweisungen beziehen sich auf den Import geschriebener Daten.

## 1. Importieren von Nur Text (‘Plain Text’)

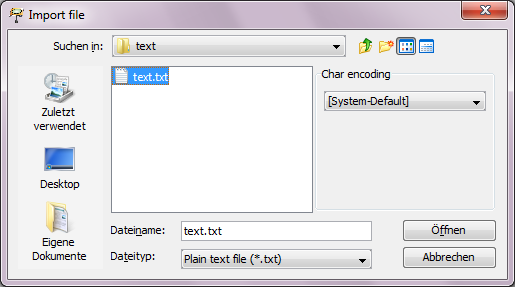
Sollte der Text nur manuell annotiert und für die Analyse im EXAKT verwendet werden, handelt es sich hierbei um die schnellste Option für den Einstieg.

**a.Vorbereiten der Dateien für den Import**

Beim Speichern des Dokuments im Nur Text (\*.txt) gehen die Formatierungen (z. B. fett oder kursiv markierte Stellen) verloren. Wenn das für Sie kein Problem darstellt,[[1]](#footnote-1) brauchen Sie Ihr Dokument lediglich im Nur Text-Format speichern, indem Sie Speichern unter und im Dropdown-Menü Nur Text(\*.txt) auswählen. Die Textdatei kann dann z.B. mit dem Editor (unter Windows) geöffnet und bearbeitet werden.

**b. Importieren der Datei in den Partitur-Editor**

Der Import des Textes erfolgt über den Punkt Import aus dem Datei Menü. Zuerst lokalisieren Sie die Datei, die Sie importieren möchten. Stellen Sie im nächsten Schritt sicher, dass der richtige Filter, d.h. Nur Text (\*.txt), und die richtige Zeichenkodierung (entsprechend der Textdatei) ausgewählt wurden. Wenn Sie nicht wissen, welche Zeichenkodierung verwendet wurde, versuchen Sie zunächst die System-Default (Standard-Auswahl).



Wenn die gewählte Zeichenkodierung nicht mit der Ihrer Datei übereinstimmt, ist es möglich, dass die Sonderzeichen nach dem Import (s. unten) nicht korrekt angezeigt werden. Sollte dies geschehen, versuchen Sie Ihre Textdatei mit einer anderen Zeichenkodierung zu speichern, z. B. UTF-8. Zu diesem Zweck wählen Sie im Editor (unter Windows) Speichern unter und legen die Zeichenkodierung fest. Danach versuchen Sie, die Datei erneut mit der gewählten Zeichenkodierung zu importieren. Im nächsten Schritt des Importprozesses wählen Sie die Aufteilung des Textes in Ereignisse aus.

Es ist wichtig sich darüber bewusst zu sein, dass die Ereignisse einer EXMARaLDA- Basistranskription, dem Format, das sie nach dem Import von Nur Text erhalten, ausschließlich zeitbasiert, im Fall geschriebener Daten wohl eher „Leerzeichen-basiert“ ist. Auch wenn die Ereignisse erstellt wurden, um den Text z.B. in Wörter zu splitten, ist nicht garantiert, dass der Text auf eine Weise segmentiert oder tokenisiert wurde, die von den EXMARaLDA-Werkzeugen umgesetzt werden kann. Die Segmentierung des Textes in Wörter ist ein anderer Prozess, in dem Start- und Endpunkte der Wörter automatisch erkannt werden. Diese können den Start- und Endpunkten entsprechen, die für „Leerzeichen-basierte" Ereignisse beim Import erstellt wurden. Man betrachtet die „Leerzeichen-basierten" Ereignisse und Zeitpunkte eher als eine Art Raster oder Orientierungshilfe beim Alignieren von Textabschnitten, möglicherweise Wörter, mit Annotationen, z.B. den jeweiligen POS-Tags.

**c. Auswählen des Text-Splitters**

Im Partitur-Editor können Sie aus verschiedenen Text-Splittern wählen:

#### Split at paragraphs

Einfache Zeilenumbrüche werden durch diese Funktion als „Absätze“ erkannt.Zusätzliche Leerzeilen zwischen Absätzen, wie sie auch in diesem Dokument verwendet werden, werden daher zusätzliche leere Ereignisse hervorrufen.Solche Ereignisse können mit dem regulären Ausdruck ^13^13 in der MS Word-Funktion **Suchen und Ersetzen** korrigiert werden. Bitte denken Sie daran Ihre Datei nach dem Import zu speichern.

#### Split at non-word character

Diese Option ist kein Tokenisierer, sondern eine Funktion, die Ereignisse erzeugt, die grob mit Wörtern korrespondieren. In jedem Ereignis gibt es ein „Wort“, gefolgt von einer Reihe nicht-alphabetischer Zeichen. Hier einige Konsequenzen für nicht alphabetische Zeichen, die Sie berücksichtigen sollten:

* ElementareInterpunktionszeichen (. ! ? : ; ,)

|existerat. |

* ÖffnendeKlammern

|fönsterrutor (|

* SchließendeKlammern

|kaféerna) |

* ÖffnendeAnführungszeichen

|där“|

* SchließendeAnführungszeichen

|gubbarna” |

* Bindestrich

|på1980-|

* Schrägstrich

|Torsgatan/|

* Zahlen

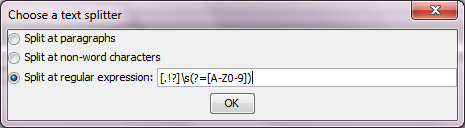
|pilsner 2,8 |

Bitte denken Sie daran Ihre Datei nach dem Import zu speichern.

#### Split at regular expression

Mit dieser Option können die Ereignisgrenzen mit regulärem Ausdruck bestimmt werden. Was die Erkennung von Wörtern angeht, ist das Splitten von Texten mit einigen regulären Ausdrücken nicht so erfolgreich wie mit einem sprachspezifischen Tokenisierer. Wenn dies für Ihre Arbeit relevant ist, sollten Sie die Arbeit mit einem Tokenisierer in Erwägung ziehen. Um Ihren Text mit dem Partitur-Editor zu tokenisieren, können Sie entweder den EXMARaLDA-Importer (entwickelt am SFB 632 in Potsdam) oder den Tokenisierer des TreeTaggers verwenden. Beide Optionen werden im Folgenden beschrieben. Wenn Sie eher ungewöhnliche „Textstücke” annotieren möchten, die in dieser Weise beschrieben werden können, oder mit Sprachen arbeiten, für die es keine zuverlässigen Tokenisierer gibt, ist die Option Split atr egular expression die bessere Wahl.

Der Ausdruck in diesem Beispiel würde den Text nach jedem, von einem Leerzeichen gefolgte, Punkt, Ausrufezeichen oder Fragezeichen, splitten, solange die darauf folgenden Zeichen entweder Großbuchstaben von A-Z oder Zahlen sind. Die Syntax für reguläre Ausdrücke finden Sie auf [hier](http://java.sun.com/javase/6/docs/api/java/util/regex/Pattern.html#sum)[[2]](#footnote-2). Unabhängig von dem verwendeten Ausdruck wird der Text an den Absatzgrenzen, z. B. Zeilenumbrüchen aufgeteilt. Bitte denken Sie daran Ihre Datei nach dem Import zu speichern.

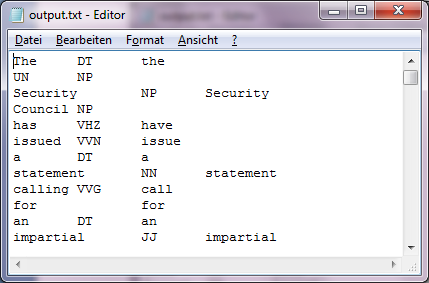


## 2. Importieren von TreeTagger Output

Der Output für den weit verbreiteten TreeTagger (Schmidt 1994, [Webseite](http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/))[[3]](#footnote-3) kann in den Partitur-Editor importiert werden. In der EXMARaLDA-Datei wird für den Text eine Transkriptionsspur mit jeweils einem Token je Ereignis erstellt. Des Weiteren werden für jedes Token ein oder zwei Annotationsspuren erzeugt, die die jeweiligen POS-Tags und Lemma (wenn die TreeTagger Option benutzt wurde) enthalten.Diese Option ist praktisch, wenn Sie vor der manuellen Annotation eine automatische Lemmatisierung und/oder POS-Tags für Ihren Text verwenden.Es gibt TreeTagger Parameter-Dateien für mehrere Sprachen und Tagsets, allerdings kann der Tagger auch „trainiert“ werden. Er kann lernen, die Daten mithilfe von eines beliebigen Tagset zu taggen, sofern es für das Tagset manuell annotierte Daten für Übungszwecke bereitstehen.

**a. Vorbereiten der Dateien für den Import**

Anweisungen für die Verwendung von TreeTaggern finden Sie auf [dieser Internetseite](http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/).[[4]](#footnote-4) Für diejenigen, die die graphischen Benutzeroberflächen bei der Eingabeaufforderung bevorzugen, gibt es eine separate [Windows Benutzeroberfläche](http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winttinterface.htm)[[5]](#footnote-5)für den TreeTagger. Die Installation dieser Benutzeroberfläche und des TreeTaggers wird auf den jeweiligen Internetseiten ausführlich beschrieben. Je nach Sprache und Tagset, sollte die Output-Datei wie folgt aussehen:



Die TreeTagger (und Simple EXMARaLDA) Import-Optionen im Partitur-Editor sind nicht nur bei der Verwendung vom TreeTagger oder Transkriptionen, die in Word erstellt wurden, interessant.Sie können auch zur Anpassung von Text-Importo-Optionen eingesetzt werden. Grundsätzlich wird beim TreeTagger-Import eine Transkription aus einer Textdatei, die durch Tabulatoren getrennt wurde, erstellt.

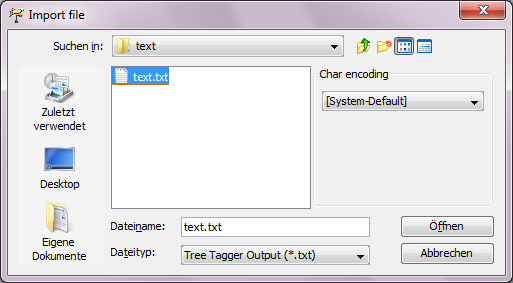
Für jede neue Zeile wird ein Ereignis erzeugt. Für die zweite und dritte Spalte werden Annotationsspuren erzeugt, deren Eigenschaften sich ganz einfach bearbeiten lassen. Das heißt, dass Sie diese Import-Option für eine beliebige Textdatei mit TreeTagger als Input- (ein Token pro Zeile) oder Output-Format (eine Spalte für den Text, zwei oder drei Spalten für die Annotationen, ein Ereignis pro Zeile) verwenden können.

Da der TreeTagger

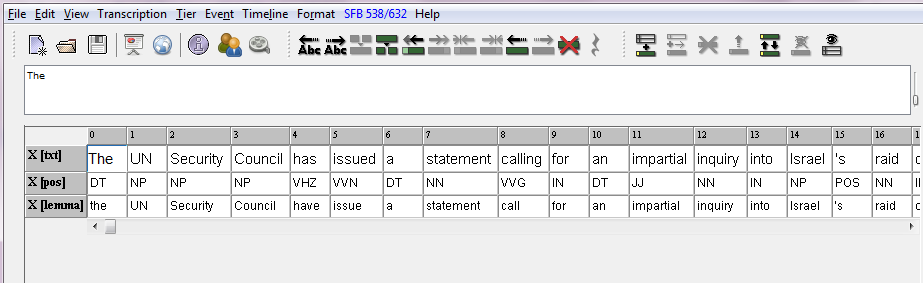
Tokenisierer die Leerzeichen der ursprünglichen Textdatei auflöst, wird während des Imports nach jedem Token/Ereignis in der Text/Transkriptionsspur ein zusätzliches Leerzeichen eingefügt.

**b. Importieren der Datei in den Partitur-Editor**

Importieren Sie die Datei über Datei>Importieren…und wählen Sie TreeTagger Output (\*.txt), um eine Transkriptionsdatei zu erzeugen und speichern Sie Ihre Datei.



Wenn Sie die bereits vorhandene EXMARaLDA-Segmentierung und Visualisierung für den Text verwenden möchten, müssen Sie während der manuellen Kontrolle des Tagger-Ergebnisses die eingefügten Leerzeichen nach Wörtern,denen Interpunktionszeichen folgen, entfernen.Wenn Sie beispielsweise möchten, dass nicht wiedererkannte, zusammengesetzte Substantive, die mit einem Tag versehen sind, als solche identifiziert werden, verwenden Sie die Funktion Verbinden aus dem Ereignis Menü. Die Graphik illustriert diese Problematik anhand von „UN Security Council“:



## 3. Importieren des Simple EXMARaLDA Formats

Diese Option eignet sich für Transkriptionen im .txt Format, die in einem gewöhnlichen Texteditor oder Textverarbeitungsprogramm erstellt wurden. Mit dem Simple EXMARaLDA Format ist es möglich, zusätzliche Annotationsspuren für Formatierungen, Markup und/oder Annotationen während des Imports anzulegen und Ereignisgrenzen anhand der Information aus der ursprünglichen Datei zu definieren. Das Dokument „How to import texttranscriptions" enthält weitere Informationen über das Simple EXMARaLDA Format und mögliche Verwendungsszenarien. Wenn Sie dieses Format verwenden möchten, um den Import des Textes anzupassen, sollten Sie sich zunächst damit befassen.

**a. Vorbereiten der Dateien für den Import**

Um Annotationen (beschreibende Ereignisse) zu erstellen, können Zeilenumbrüche und geschweifte oder eckige Klammern gesetzt werden. Der annotierte Text sollte in eine separate Zeile, die mit einem Sprecherkürzel beginnt, eingesetzt werden. Der Annotationstext hingegen sollte in geschweiften Klammern am Ende der Zeile stehen. Microsoft Word und OpenOffice verfügen über eine Option für reguläre Ausdrücke (Suchen und Ersetzen), die das Suchen und Ersetzen von Formatierungen ermöglicht. Die gefundenen Ausdrücke werden als Teil des zu ersetzenden Ausdrucks verwendet. Um die Formatierungsinformationen zu behalten, könnte der Satz „This is **very** weird.” wie folgt umgewandelt werden:

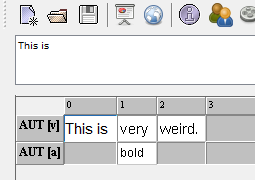
AUT: This is

AUT: very {bold}

AUT: weird.

**b. Importieren der Datei in den Partitur-Editor**

Der Import erfolgt über Datei >Importieren.... Wählen Sie Simple EXMARaLDAtextfile (\*.txt)als Dateityp. Sie können die Spurkategorien und die angezeigten Namen ändern. Hierfür klicken Sie auf das Sprecherkürzel der Spur, um diese hervorzuheben. Wählen Sie anschließend den Sprecher über Spur > Spureigenschaften… aus.



# B. DerSFB 632 EXMARaLDA-Importer

Der [SFB 632](http://www.sfb632.uni-potsdam.de)[[6]](#footnote-6) hat einige Datei-[Importer](https://141.89.100.100/homes/d1/services/paula_webservice/for_KorpTA/index_en.php)[[7]](#footnote-7) für die automatische Generierung von EXMARaLDA-Dateien (auch für Dateien in anderen Formaten, die von Annotationswerkzeugen verwendet werden) aus Nur Text-Dateien entwickelt. Der Importer erzeugt eine EXMARaLDA-Datei mit dem Text aus der Nur Text-Datei, indem der Text in die Transkriptionsspur der Kategorie „word" und mit Ereignisgrenzen nach jedem Wort eingesetzt wird.

In den beiden Annotationsspuren befinden sich Annotationen, die mit den Satzgrenzen (Kategorie „sent”, Annotation „S”) und Absatzgrenzen (Kategorie „para”, Annotation „P”) korrespondieren. Laden Sie zunächst eine Textdatei hoch und Sie erhalten einen Link zu der korrespondierenden Datei im EXMARaLDA-Basistranskriptionformat. Laden Sie diese herunter. Standardmäßig wird für die Erkennung von Wort-, Satz- und Absatzgrenzen während dieses Prozesses eine Tokenisierung durchgeführt. Sollte das Resultat der Tokenisierung nicht zufriedenstellend sein, kann der Text auch manuell korrigiert (Tokens getrennt durch Leerzeichen, ein Satz pro Zeile, Absätze getrennt durch leere Zeile) und mithilfe der Option für den Import von tokenisierten Text importiert werden.

Der Potsdamer „Dialekt“ und die restlichen EXMARaLDA-Werkzeuge teilen die Interpretation der EXMARaLDA- Grundkonzepte, wie Ereignis und Segmentierung, nicht. Solange diese Art von Basistranskription nicht in eine segmentierte Transkription konvertiert wird, ist es auch nicht möglich, EXAXT für Konkordanzen und Analysen heranzuziehen. Mit den inhärenten Textimport-Optionen würden die Leerzeichen erhalten bleiben, die in der ursprünglichen Datei als Worttrenner verwendet wurden, beim TreeTagger-Import werden sogar einige hinzugefügt. Im Potsdamer Dialekt hingegen werden die Leerzeichen während der Konvertierung aufgelöst. Der Partitur-Editor benötigt Leerzeichen für die automatische Segmentierung, d.h. um eine segmentierte Transkription zu erzeugen.Auch EXAKT arbeitet ausschließlich mit segmentierten Transkriptionen.

# C. Annotieren des Textes

## 1. Der AUT Sprecher

Da EXMARaLDA ursprünglich für die sprachliche Transkription mehrerer Sprecher entwickelt wurde, basiert es auf dem Prinzip, dass es für jeden Sprecher eine Transkriptionsspur gibt. Auch wenn beim Arbeiten mit geschriebenen Daten, der „Sprecher" unbekannt oder irrelevant ist, muss einer angegeben werden. Annotationsspuren, und auch Annotationen, stehen in Verbindung mit der Transkriptionsspur und somit auch mit dem annotierten Text. Daher müssen auch alle hinzugefügten Spuren über das Sprecherkürzel der Transkriptionsspur zugewiesen werden. Aus diesem Grund wird vom Partitur-Editor seit Version 1.4.5. automatisch ein Dummy-Sprecher (AUT) samt der Transkriptionsspur für den importierten Text erzeugt.

Wenn Sie eine ältere Version von EXMARaLDA verwenden, ist eine Aktualisierung auf die neueste Version empfehlenswert. Sie können auch manuell einen Sprecher für den importierten Text anlegen (Transkription > Sprechertabelle… > Sprecher hinzufügen) und ihn einer (Transkriptions)Spur zuordnen: Hierfür klicken Sie auf das Sprecherkürzel der Spur, um diese hervorzuheben. Dann wählen Sie den Sprecher über Spur > Spureigenschaften… aus.

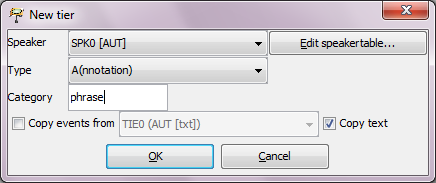
## 2. Annotationsspuren

Um Ihren Text zu annotieren, müssen sie zuerst zusätzliche Annotationspuren des Typs „A(nnotation)" erstellen. Die Anzahl der Annotationsspuren und deren Kategorien hängt von Ihrem Annotationsschema ab. Da der Partitur-Editor für die Transkription von gesprochener Sprache entwickelt wurde, kann er sehr gut mir parallelen, unabhängigen Ereignissen oder Annotationen arbeiten. Allerdings verfügt das EXMARaLDA-Datenmodell über keine integrierte Möglichkeit Hierarchien oder andere Relationen zwischen Annotationen und Annotationsspuren auszudrücken.

Um eine Annotationsspur hinzuzufügen, klicken Sie auf Spur anfügen…(links) oder Spur einfügen…(rechts) oder wählen den entsprechenden Punkt aus dem Menü Spur.

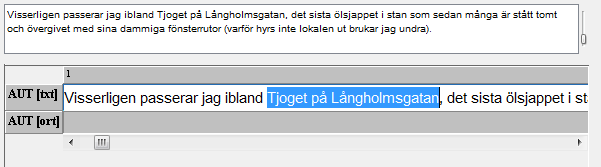
Anfügen / Einfügen

Wenn Sie Spur einfügen… wählen, können Sie die Platzierung der Spur bestimmen. Die Option Spur hinzufügen… wird die neue Spur als letzte anfügen. Wählen Sie den Typ A(Annotation), wählen Sie den Sprecher aus und definieren die Kategorie für die neue Spur.

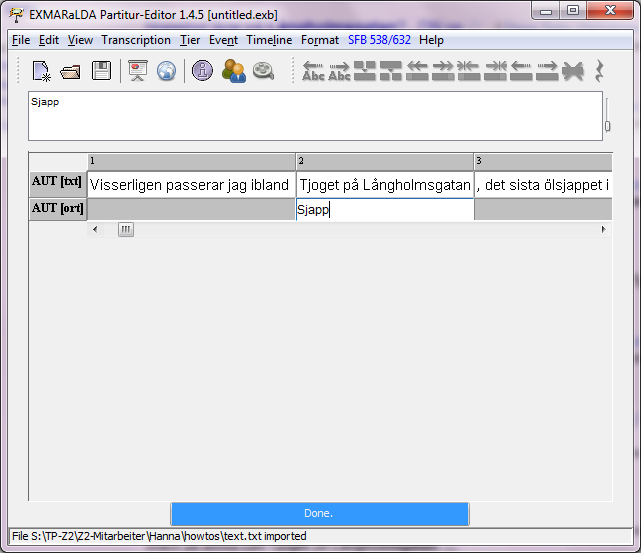


## 3. Annotieren im Partitur-Editor

Wenn es sich um einfache Annotationen handelt, d.h. eine Art Kommentar, können Sie sofort mit dem Annotieren beginnen. Um Ereignisse für die Annotationen zu erzeugen, bestimmen Sie zusätzliche Ereignisgrenzen: Für eine einfache Grenze platzieren Sie den Cursor im Text dort, wo die Grenze liegen soll. Dann drücken Sie entweder Strg+2 oder klicken auf das Teilen-Symbol oder wählen Teilen aus dem Ereignis-Menü.Um ein Ereignis für einen Textabschnitt zu erzeugen, drücken Sie entweder Strg+3 oder wählen Sie Zweifach teilen aus dem Menüpunkt Ereignis.



Das Bild oben zeigt den markierten Text, im unteren Bild sehen Sie das Resultat von Zweifach teilen.



Das Annotationswerkzeug (unter Ansicht> Annotationswerkzeug) erleichtert Ihnen den konsistenten Annotationsprozess, zum Beispiel durch Hinzufügen von Annotationen zu mehr als einem Ereignis, welche dann automatisch verbunden werden. Ihr Annotationsschema mit Annotationsrichtlinien wird für die Person, für die eine Annotation vorgenommen werden soll, sichtbar sein. Das Werkzeug schlägt Annotationen vor, wird jedoch keine Beschränkungen auferlegen.In dem Dokument „How to use the Annotation Panel“ wird seine Anwendung detailliert behandelt.

# D. Segmentierung

Die anderen EXMARaLDA-Werkzeuge - CoMa und EXAKT - erfordern segmentierte Transkriptionen. Die Segmentierung in EXMARaLDA basiert gewöhnlich auf einigen Transkriptionskonventionen, in denen jedes Symbol oder Symbolpaar eine eindeutige Bedeutung trägt.

Da dies bei der Standard-Schriftsprache nicht der Fall ist[[8]](#footnote-8), kann der Text nur in Segmentketten und Wörter segmentiert werden. Dies wiederum ist nicht das Resultat der Tokenisierung, sondern hängt nur von der Verwendung von Leerzeichen als Trennzeichen ab. Für weiterführende Informationen über Segmentierung und Verwendungsmöglichkeiten anderer Segmentierungsalgorithmen und deren Anpassung, konsultieren Sie „How to use segmentation“.

1. Wenn die Formatierung für Ihr Dokument entscheidend ist, können Sie als Zwischenlösung entweder die etwas komplexeren Import-Optionen (Beschreibung s.u.) verwenden oder Sie konvertieren das Dokument direkt in das Basistranskriptionsformat XML. [↑](#footnote-ref-1)
2. http://java.sun.com/javase/6/docs/api/java/util/regex/Pattern.html#sum [↑](#footnote-ref-2)
3. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ [↑](#footnote-ref-3)
4. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ [↑](#footnote-ref-4)
5. http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winttinterface.htm [↑](#footnote-ref-5)
6. http://www.sfb632.uni-potsdam.de/ [↑](#footnote-ref-6)
7. https://141.89.100.100/homes/d1/services/paula\_webservice/for\_KorpTA/index\_en.php [↑](#footnote-ref-7)
8. Zum Beispiel wird ein Punkt sowohl als ein Äußerungsendzeichen, als auch als Teil einer Abkürzung verwendet. [↑](#footnote-ref-8)