TraDisc Handbuch

Christoph Malisi

TraDisc Handbuch

Christoph Malisi

Inhaltsverzeichnis

Einleitung

TraDisc ist ein Programm zur Annotation von linguistischen Korpora. Das zu bearbeitende Korpus liegt dabei in einem XML-Format vor. (Möchte man einen Text bearbeiten, der noch nicht in eine XML-Format vorliegt, so kann man das Programm Tokenizer benutzen, um den Text in ein einfaches XML-Format zu bringen).

Ursprünglich wurde TraDisc entwickelt, um Junktoren (Satzkonnektoren) in einem Korpus zu identifizieren und zu annotieren; jedoch kann man mit TraDisc auch beliebige andere Eigenschaften in einem Korpus annotieren.

Teil I. TraDisc

Mit TraDisc lassen sich die Tokens eines Korpus mit vom Benutzer definierten Funktionen oder Eigenschaften annotieren. Diese Eigenschaften, mit denen die Tokens des Korpus annotiert werden sollen, können dabei ein- oder zweidimensional sein. In diesem Handbuch wird zur Illustration ein Beispiel besprochen, in welchem die Personalpronomen im Nominativ in einem Text annotiert werden sollen. Die Eigenschaft, die annotiert werden soll, besteht hier aus den Dimensionen *Person* und *Numerus*. Das Personalpronomen *er* könnte also z.B. mit der Eigenschaft *3. Person-Singular* annotiert werden.

Die Dimensionen der Eigenschaften bilden die Spalten und Zeilen einer Tabelle, des TraDisc-Schemas. Im Beispiel gibt es folglich drei Spalten, nämlich *1. Person*, *2. Person* und *3. Person*, und zwei Zeilen, *Singular* und *Plural*.

Tokens (meist Wörter) können in die Felder dieser Tabelle eingetragen werden. TraDisc bietet Hilfe beim eigentlichen Annotieren der Tokens, und es bietet mehrere Möglichkeiten, sich im Text zielgerichtet zu bewegen, um einfach Markables zu finden. (Tokens, die im Schema aufgelistet sind, heißen Markables. Im Beispiel sind das die potentiellen Personalpronomen *ich, du, er, sie, es, wir, ihr*.) Weiterhin stellt TraDisc diverse Analyse- und Auswertungswerkzeuge zur Verfügung. Sie dienen dazu, den Text anhand der gewählten Annotationskriterien zu bewerten, zum Beispiel die Anzahl und die Art der Markables festzustellen.

Kapitel 1. Das XML-Korpusformat

Anmerkung

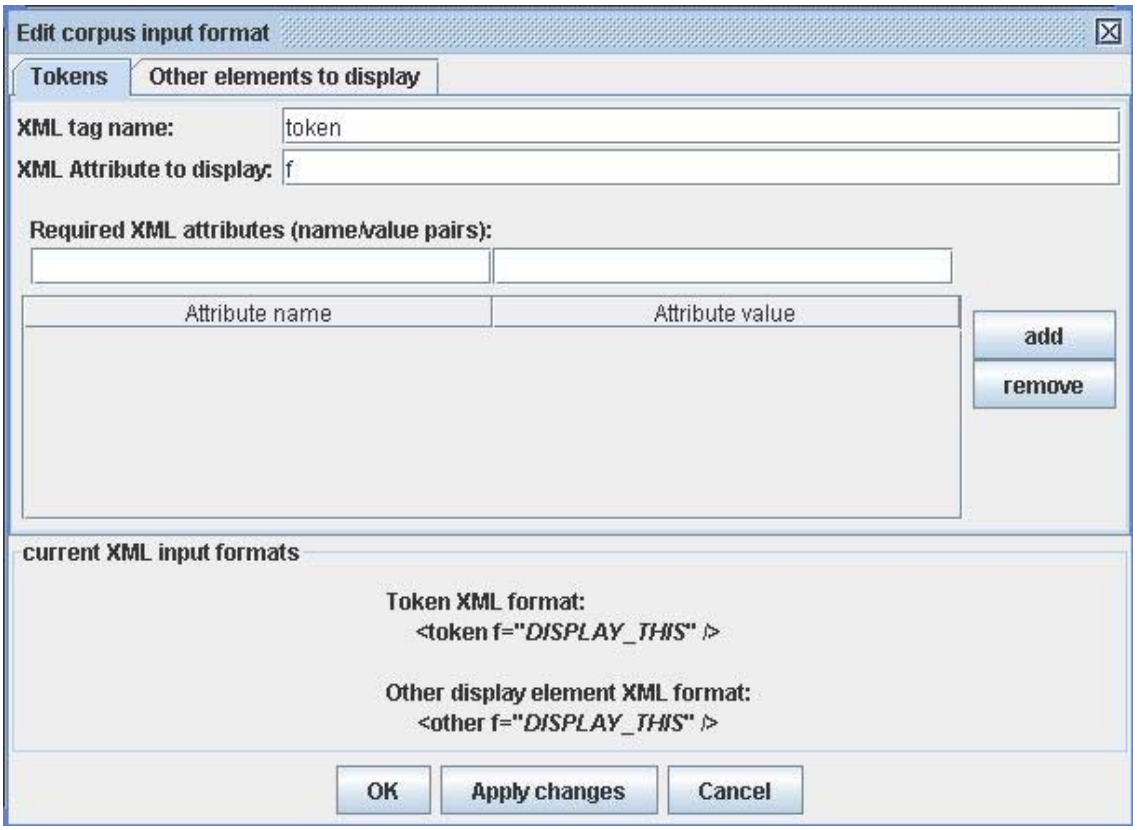
Wenn Sie einen Text mit TraDisc bearbeiten wollen, der im TraDisc Standard-XML-Format vorliegt, also der Text durch den Tokenizer in ein XML-Format gebracht wurde, so ist es nicht unbedingt nötig, dieses Kapitel zu kennen. Die Einstellungen für dieses XML-Format sind standardmäßig voreingestellt.

Spezifizieren des XML-Eingabeformates des Korpus

Um mit einem Korpus in TraDisc arbeiten zu können, muss er in einem XML-Format vorliegen. Dieses Format muss in TraDisc eingegeben werden, damit das Programm weiß, wie der Text vorliegt. Im Menüpunkt "Input Format → Edit corpus input format" erscheint der entsprechende Eingabedialog. Er besteht aus zwei übergeordneten Karteikarten: tokens und other elements to display.

In der Karte Tokens muss das XML-Element spezifiziert werden, das die Tokens des Korpus enthält. Es muss das XML-Tag des Elements angegeben werden (Im Feld XML tag name), außerdem der Name des XML-Attributs, welches das eigentliche Wort enthält, das im TraDisc-Textfeld dargestellt werden soll (im Eingabefeld XML attribute to display). Diese beiden Angaben müssen immer erfolgen. Zusätzlich können noch weitere XML-Attribute samt Attributswerten angegeben werden, die im XML-Format des Korpus im XML-Element eines Tokens immer vorhanden sein müssen. Geben Sie hierzu den Namen des Attributs und des Werts in den linken bzw. rechten Teil des Eingabefelds Required XML attributes ein, und fügen Sie das Attribut mit dem Knopf Add hinzu. Von der Liste der erforderlichen Attribute können Eintrage gelöscht werden, indem der entsprechende Eintrag in der Liste markiert und auf remove geklickt wird. Mit einem Klick auf Apply changes werden die Änderungen am Korpuseingabeformat aktualisiert, das neue Format erscheint im Vorschaufenster Current XML input format.

In der Karteikarte Other elements to display werden XML-Elemente spezifiziert, die keine Tokens enthalten, aber trotzdem im TraDisc Textfenster dargestellt werden sollen. Normalerweise sind das die XML-Elemente, welche die Satzzeichen des zu annotierenden Textes enthalten. Wie bei den XML-Elementen für die Tokens müssen das XML-Tag und das darzustellende Attribut (welches das Satzzeichen enthält) angegeben werden. Es gibt auch hier die Möglichkeit, weitere benötigte Attribute anzugeben, wenn dies erforderlich ist. Die Eingabefelder sind äquivalent zu denen der Karteikarte Tokens.



Der Eingabeformat-Dialog mit den eingetragenen Werten des TraDisc Standardformats.

Das TraDisc Standardformat

Beim Starten von TraDisc ist schon ein XML-Format voreingestellt, das TraDisc Standardformat. Das Tag der XML-Elemente, die Tokens enthalten, ist token, das Attribut, welches das eigentliche Token enthält (das Wort), ist f. Es gibt keine weiteren vorgeschriebenen Attribute des Token-XML-Elements. Satzzeichen werden in XML-Elementen mit dem Tag other gespeichert. Da diese angezeigt werden sollen, ist bei Other elements to display das Tag other eingetragen.

Liegt das zu bearbeitende Korpus im TraDisc-Standardformat vor, muss also nichts im Menü "Input Format" geändert werden.

Beispiel eines Korpus im TraDisc-Standardformat. Der Korpustext lautet: *Herzlich willkommen bei TraDisc!* Das zugehörige XML-Korpus sieht wie folgt aus:

```
<corpus>
   <token f="Herzlich"/>
   <token f="willkommen"/>
   <token f="bei"/>
   <token f="TraDisc"/>
   <other f="!"/>
</corpus>
```

Speichern und Laden eines XML-Eingabeformates

Um das aktuelle Eingabeformat zu speichern, wählt man im Menü "Input Format → Save input format specifications..." aus. Es öffnet sich ein Dateispeicherdialog. Um ein gespeichertes Eingabeformat zu laden, kann man "Input Format → Load input format specifications..." wählen, z.B. wenn mehrere Korpora mit dem gleichen Eingabeformat bearbeitet werden sollen.

Kapitel 2. Das Schema

Mit TraDisc kann man die Tokens in einem Korpus mit zweidimensionalen Funktionen oder Eigenschaften annotieren. Diejenigen Tokens, die für eine Annotation in Frage kommen, die also möglicherweise eine zu annotierende Eigenschaft haben, werden als Markables bezeichnet. Damit das Programm weiß, welche Eigenschaften für ein bestimmtes Markable möglich sind, muss ein TraDisc-Schema erstellt werden. Mit TraDisc kann man zweidimensionale Eigenschaften annotieren, deswegen werden die Markables in eine zweidimensionale Tabelle eingetragen. Die Zeilen- bzw. Spaltentitel dieser Tabelle entsprechen den Eigenschaften, die dieses Markable annehmen kann (siehe auch hier).

Angenommen, man möchte die Personalpronomen im Nominativ eines Textes mit Person und Numerus annotieren, eine entsprechende einfache Schematabelle könnte so aussehen:

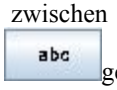
	1. Person	**2. Person**	**3. Person**
Singular	ich	du	er sie es
Plural	wir	ihr	sie

Um diese Schematabelle in TraDisc einzutragen, benutzt man den Schemaeditor.

Der Schemaeditor

Geöffnet wird der Schemaeditor mit dem Menüpunkt "Schema → Edit...". Es wird das gerade geladene Schema angezeigt. Falls noch kein Schema geladen war, wird eine neues, leeres Schema angezeigt, mit zwei Spalten namens "column 1" und "column 2", und zwei Zeilen namens "row 1" und "row 2". Um das Schema seinen Bedürfnissen anzupassen, kann man Spalten und Zeilen mit den folgenden Schaltflächen hinzufügen und löschen:

-  Fügt eine neue Zeile ganz unten an der Tabelle hinzu
-  Fügt eine neue Zeile oberhalb der Zeile hinzu, die das markierte Feld enthält
-  Löscht die Zeile des markierten Felds
-  Fügt eine neue Spalte ganz rechts an der Tabelle hinzu
-  Fügt eine neue Spalte links von der Spalte hinzu, die das markierte Feld enthält
-  Löscht die Spalte des markierten Felds

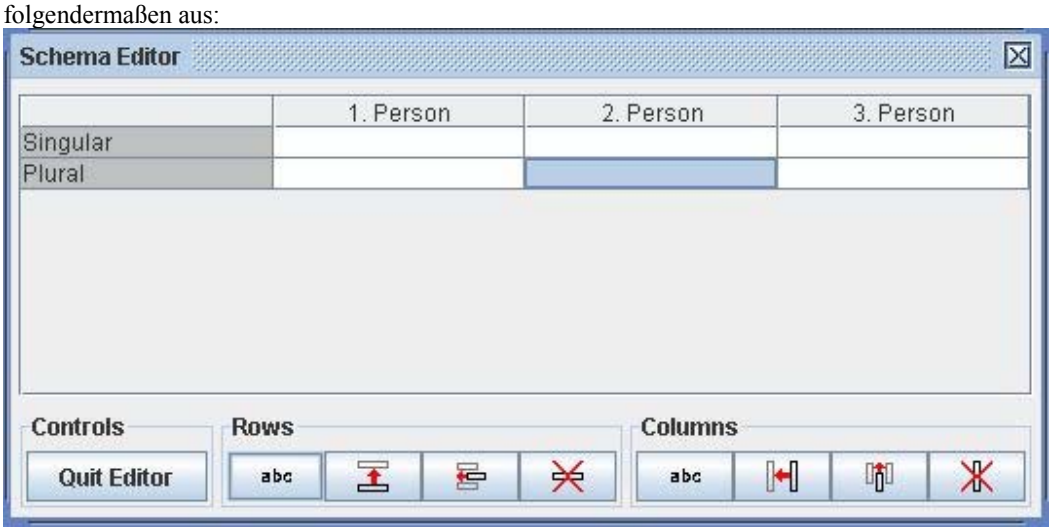
Das markierte Feld ist blau hinterlegt, man kann mit den Pfeiltasten zwischen den Tabellenfeldern wechseln. Die Zeilen- und Spaltentitel können mit den Schaltflächen 

geändert werden. Die Schaltfläche im Rahmen "rows" ändert den Namen der aktuellen Zeile, diejenige im Rahmen "columns" den Namen der aktuellen Spalte. Falls man die Druckfunktion von TraDisc benutzen will, ist es bei der Auswahl der Spaltennamen nützlich, eine Abkürzung oder etwas Ähnliches vor der eigentlichen Bezeichnung zu schreiben, gefolgt von einem Leerzeichen.

Anmerkung

Wenn eine Annotation angefangen hat, können die Dimensionen des Schemas **nicht** geändert werden. Dies bedeutet, dass keine neuen Zeilen und Spalten eingefügt, keine Zeilen und Spalten gelöscht und die Namen der Zeilen und Spalten nicht geändert werden können. Die Einträge in den Tabellenfeldern können jedoch bearbeitet werden wie unten beschrieben.

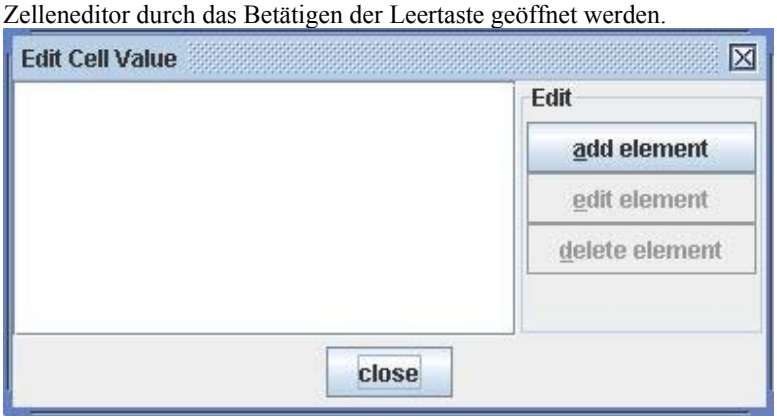
Für unser Beispielschema der Personalpronomen muss also eine neue Spalte in das leere Schema eingefügt werden; und die Zeilen- und Spaltentitel müssen umbenannt und mit den Eigenschaften der Person und des Numerus belegt werden. Die entsprechende Tabelle sieht im Schemaeditor dann folgendermaßen aus:



Um die Felder des Schemas (die Schemazellen) mit Einträgen zu füllen, muss der Zelleneditor geöffnet werden.

Der Zelleneditor

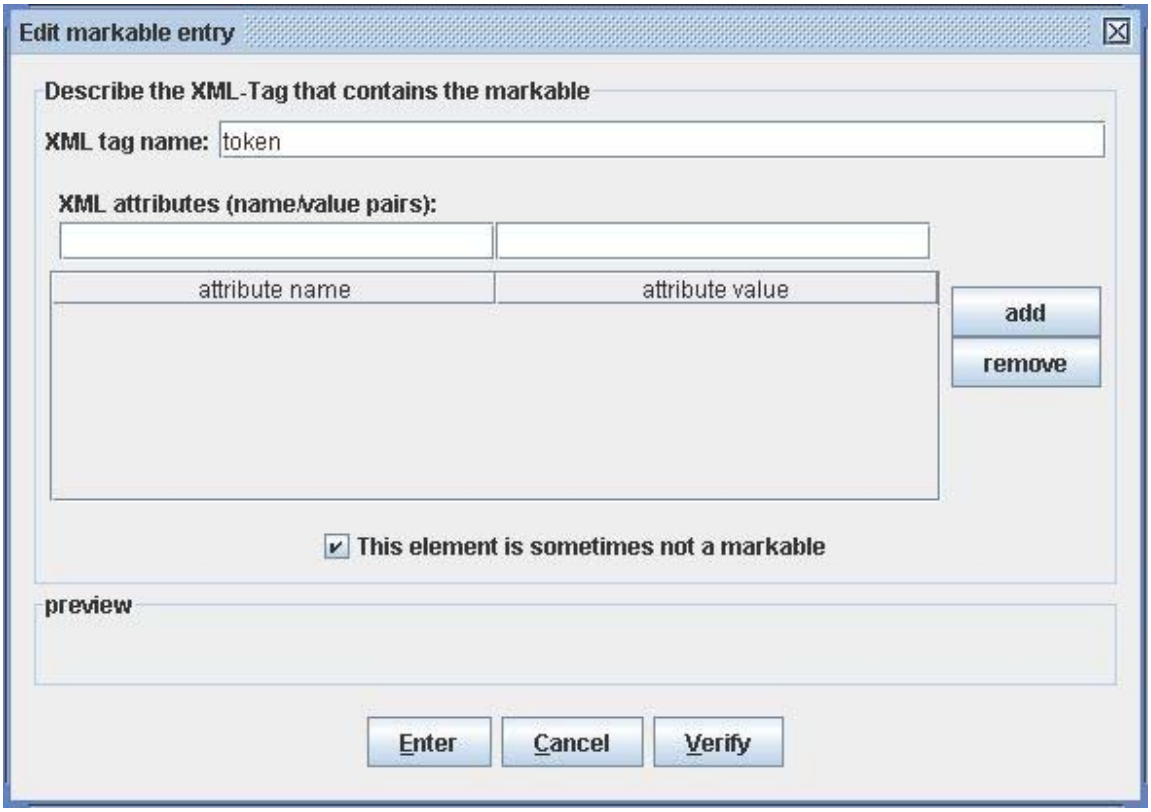
Der Zelleneditor zum Ändern der Einträge eines Felds der Schematabelle wird durch einen Mausklick auf das entsprechende Feld geöffnet; die Zelle kann auch mit Hilfe der Pfeiltasten markiert und der Zelleneditor durch das Betätigen der Leertaste geöffnet werden.



Der Zelleneditor eines Tabellenfelds ohne Einträge.

Neue Markables können durch klicken auf add element hinzugefügt werden. Bereits eingetragene Markables kann man ändern, indem der Eintrag markiert und edit element gewählt wird. Bei beiden Vorgängen öffnet sich der *Markableeditor*. Wenn ein Markable aus der Zelle gelöscht werden soll, so markiert man es und wählt delete element.

Der Markableeditor



Ein leerer Markableeditor bei TraDisc-Standardeingabeformat.

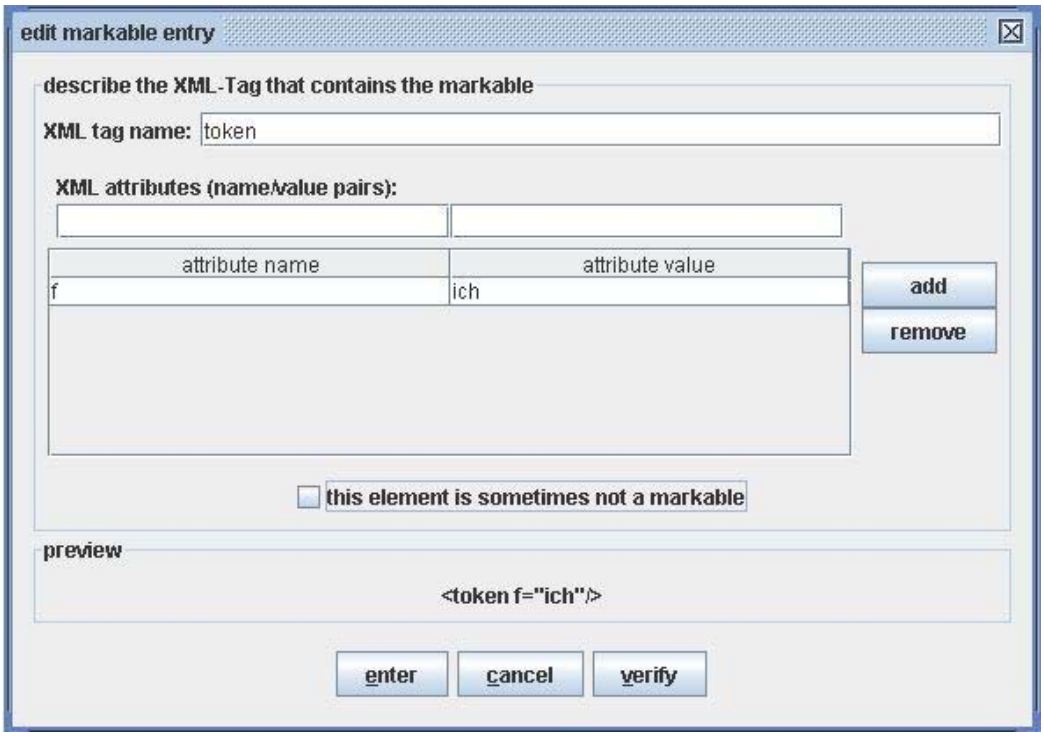
Damit TraDisc Markables im Korpus finden kann, müssen sie zuerst in die Schematabelle im XML-Format des Korpus eingetragen werden.

Im Feld XML tag name trägt man das XML-Tag ein, das die Tokens enthält. Da der Name dieses Tags im XML-Eingabeformat spezifiziert ist, schlägt TraDisc diesen vor. Im TraDisc-Standardformat ist das token. Es müssen zusätzlich noch die XML-Attribute eingetragen werden, die ein Token besitzen muss. Das ist normalerweise das XML-Attribut, welches das eigentliche Wort enthält (das "Display-Attribut"); im TraDisc-Standardformat ist es f. Diese Einträge erfolgen im Textfeld XML attributes (name/value pairs), der Name des Attributs kommt ins Feld attribute name, der Attributswert ins Feld attribute value. Zum Hinzufügen auf add klicken. Die Attributswerte werden **nicht** nach Groß/Kleinschreibung unterschieden, es reicht also aus, eine Version in das Schema einzutragen. Im Attributswert kann das Sternchensymbol * als Platzhalter fungieren. * kann am Anfang und/oder am Ende eines Attributwertes stehen. * ist ein Platzhalter für eine beliebige Buchstabenfolge. Wäre z.B. der Wert eines Attributs *lach**, so würden alle Tokens erkannt, die im entsprechenden Attribut einen Wert hätten, der mit *lach* beginnt, also *lachen, lachst, lache, Lachgeschichten...*.

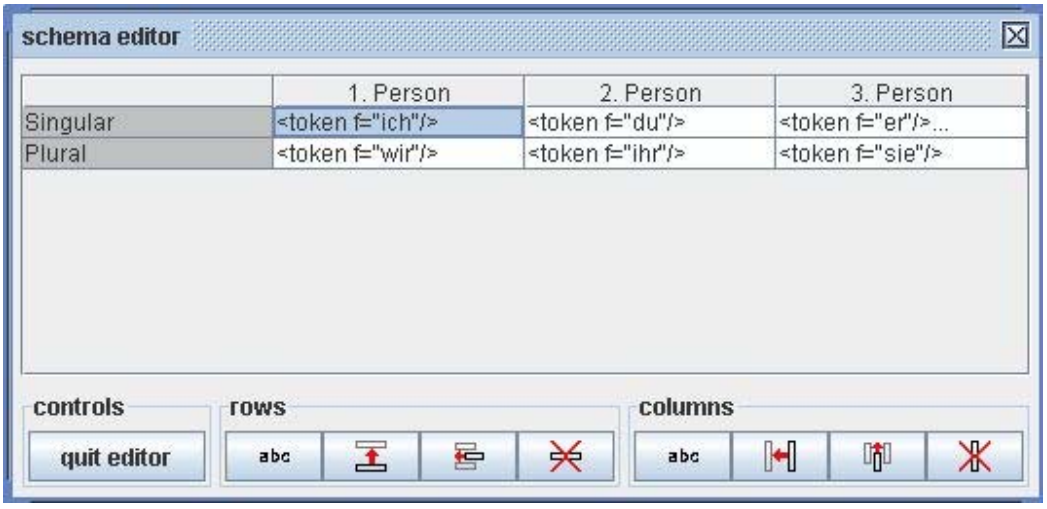
Um ein Attribut zu löschen, das Attribut in der Liste auswählen und remove klicken. Um zu prüfen, ob ein Markable das korrekte Format hat und um eine Vorschau des Markable-XML-Elements zu erhalten, kann man auf verify klicken. Die Vorschau erscheint dann im Feld preview.

Manche Markable haben nicht in jedem Kontext die Funktion oder Eigenschaft, die annotiert werden soll. Das Markable *ihr* aus dem Beispiel kann z.B. auch in einem Kontext vorkommen, in der es keine Funktion als Personalpronomen hat. Also sollen diese Markables nicht immer mit einer der Funktionen der Tabelle annotiert werden. Wenn das so ist, muss das Häkchen this element is sometimes not a markable gesetzt werden.

Im Beispiel mit den Personalpronomen muss man in die linke obere Tabellenzelle das Token *ich* eintragen. Man wählt also die Zelle aus, klickt im Zelleneditor auf add element, und fügt im Markableeditor nun ein Attribut mit Namen f und Wert ich hinzu. Nach einem Klick auf verify sieht man das gewünschte XML-Element <token f="ich" /> im Vorschaufenster. Da *ich* immer als Personalpronomen im Nominativ fungiert, wird das Häkchen in this element is sometimes not a markable entfernt.



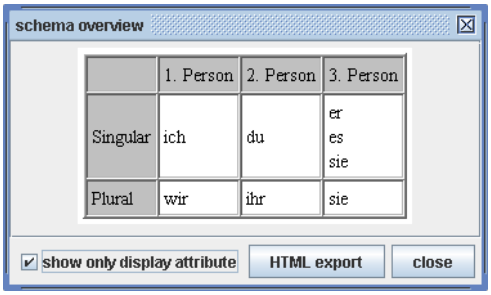
Die restlichen Tabelleneinträge werden entsprechend hinzugefügt. Da *es, sie, ihr* auch in anderen Funktionen (Personalpronomen im Akkusativ bzw. Possessivbegleiter) vorkommen können, wird hier das Häkchen this element is sometimes not a markable gesetzt. Der Schemaeditor sieht mit allen Einträgen dann folgendermaßen aus:

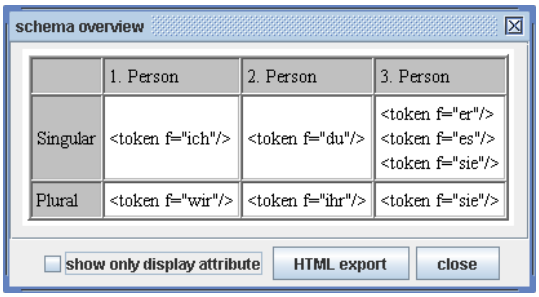


In der Zelle "3. Person - Singular" sind 3 Einträge. Von diesen ist nur der oberste Eintrag sichtbar, dass noch weitere folgen, wird durch die drei Punkte ... angedeutet.

Der Schemaüberblick

Im Menüpunkt "Schema → Overview" befindet sich der Schemaüberblick. Die Schematabelle wird mit allen Einträgen in einem übersichtlichen Layout angezeigt. Dabei kann man zwischen zwei Darstellungsarten wählen, zwischen denen man mit dem Häkchen show only display attribute wechseln kann. Ist das Häkchen gewählt, werden nur die Wert des Display-Attributs des eingetragenen Markables gezeigt. Im TraDisc-Standardformat ist das f, von einem Tabelleneintrag <token f="ich"/> wird also nur das Wort ich dargestellt. Ist das Häkchen nicht gesetzt, wird das gesamte XML-Element gezeigt:





Speichern und Laden eines Schemas

Schematabellen können gespeichert werden, damit ein einmal erstelltes Schema für mehrere Annotationen benutzt werden kann. Zum Speichern muss man den Menüpunkt "Schema → Save" bzw. "Schema → Save as..." auswählen. Ein gespeichertes Schema kann geöffnet werden durch Auswahl von "Schema → Open". (Schemas werden in einem TraDisc-eigenen XML-Format gespeichert.)

Ein neues Schema beginnen

Mit dem Menüpunkt "Schema → New" wird ein neues, leeres Schema in TraDisc und dem Schemaeditor geladen. Sind beim aktuellen Schema noch nicht gespeicherte Veränderungen vorhanden, so weist TraDisc den Benutzer darauf hin und gibt die Möglichkeit, erst zu speichern.

Ein Schema mit einem anderen verschmelzen

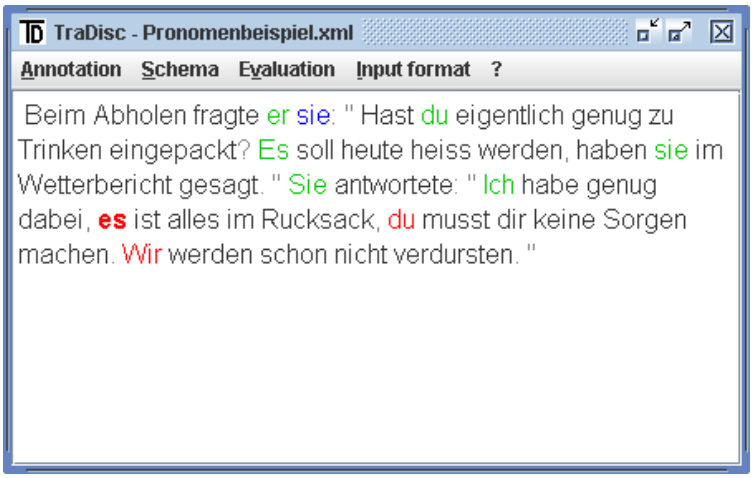
TraDisc bietet die Möglichkeit, ein gespeichertes Schema zu dem gerade aktiven Schema zu addieren, bzw. die beiden Schemata zu verschmelzen. Dann stehen im neuen Schema die Einträge, die in einem oder beiden der Vorgängerschemata stehen. Das zu addierende Schema muss dabei die folgenden Bedingungen erfüllen: Es muß die gleichen Dimensionen haben, also die gleiche Anzahl von Spalten und Zeilen und die gleichen Spalten- und Zeilennamen. Zusätzlich müssen die XML-Eingabeformate der Markables dieselben sein (siehe hier). Alle Einträge des hinzugefügten Schemas, die nicht im aktiven Schema vorhanden sind, werden in dieses übernommen. Im Menüpunkt "Schema → Merge with schema..." kann man ein gespeichertes Schema auswählen, das in der beschriebenen Weise mit dem aktiven Schema verschmolzen wird.

Kapitel 3. Annotationen

Eine neue Annotation beginnen

Um eine neue Annotation zu beginnen, im Menü "Annotation → New..." auswählen. Es öffnet sich ein Dateiauswahldialog. Hier wählt man die zu bearbeitende Korpusdatei. Diese muß dem XML-Korpusformat, das gerade spezifiziert ist, entsprechen. Der Korpustext wird nun geladen und im Annotationstextfenster dargestellt. Außerdem öffnet sich das Annotationskontrollfenster.

Das Annotationstextfenster

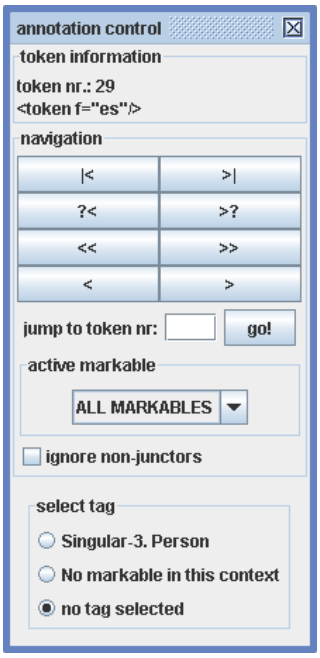


Im Annotationstextfenster ist der zu bearbeitende Korpus dargestellt. Markables, also Tokens, die im Schema stehen, sind farblich gekennzeichnet:

- Rot: Das Markable ist noch nicht annotiert
- Grün: Das Markable ist mit einem Tag annotiert (d.h. einer Funktion oder Eigenschaft, die dem Zeilen- und Spaltennamen einer der Tabellenzellen, in denen das Markable im Schema steht, entspricht. Im Beispiel der Personalpronomen ist ein Tag "Singular-1.Person", ein weiteres "Singular-2.Person", usw.)
- Blau: Das Markable hat in diesem Kontext keine der zu annotierenden Eigenschaften (siehe hier [11])

Tokens, die nicht im Schema vorkommen, also keine Markables sind, sind schwarz dargestellt. Alle anderen dargestellten Elemente des Textes (meist Satzzeichen, siehe auch hier [6]) sind grau dargestellt. Das gerade aktive Element ist fett geschrieben. Um ein Element zu aktivieren, können entweder die Navigationsmöglichkeiten des Annotationskontrollfensters benutzt werden, oder man wählt das gewünschte Element einfach mit einem Doppelklick aus.

Das Annotationskontrollfenster



Das Annotationskontrollfenster ist nur sichtbar, wenn eine Annotation geöffnet ist. Es hat drei Bereiche: Das Feld token information zeigt die Nummer des gerade aktivierten Tokens an, gezählt vom Beginn des Textes. Zusätzlich wird das XML-Element des Tokens angezeigt (d.h. das XML-Element des Tokens in der zugrundeliegenden XML-Korpusdatei, siehe auch hier). Das Feld navigation bietet diverse Möglichkeiten zur Navigation durch den Text. Schließlich kann im Feld select tag ein Tag für das gerade aktive Markable ausgewählt werden (sofern gerade ein Markable aktiv ist).

Navigation

In der Auswahlbox active markable kann man auswählen, über welche Markables man im Text navigieren möchte. Alle Markables, die im Schema eingetragen sind, stehen zur Auswahl; zusätzlich gibt es drei Einträge namens ALL MARKABLES, ALL TOKENS und ALL ELEMENTS.

Nach dem Öffnen einer Annotation ist der Eintrag ALL MARKABLES ausgewählt. Die Navigationsschaltflächen sind dann wie folgt belegt:

- |< >| Springt zum ersten bzw. letzten Markable des Texts.
- ?< >? Springt zum nächsten Markable in der jeweiligen Richtung, das noch nicht annotiert ist (d.h. zum nächsten roten Markable).
- << >> Springt zum nächsten Markable in der jeweiligen Richtung, das mehrere Annotationsmöglichkeiten (Tags) aufweist. Alle noch nicht annotierten Markables, die dazwischen liegen und nur ein mögliches Tag haben, werden automatisch mit diesem annotiert und übersprungen. (Diese eindeutig annotierbaren Markables stehen nur in einer Zelle des Schemas, und das Häkchen this element is sometimes not a markable ist nicht gesetzt. Sie werden dann also mit dem Tag annotiert, das der Tabellenzeile und -spalte des einzigen Tabelleneintrags entspricht.)
- < > Springt zum nächsten Markable in der jeweiligen Richtung.

Wählt man in der Box active markable ein bestimmtes Markable aus (also keinen der Einträge ALL MARKABLES, ALL TOKENS und ALL ELEMENTS), so navigiert man nur über Markables dieses Typs im Text, alle anderen werden übersprungen und ignoriert. Ist ALL TOKENS ausgewählt, so wird über alle Tokens des Textes navigiert, egal ob sie Markables sind oder nicht (also auch über die schwarz dargestellten). Die Wahl ALL ELEMENTS erlaubt die Navigation über alle dargestellten XML-Elemente des Texts, nicht nur über die Tokens. Das bedeutet vor allem, dass auch über Satzzeichen navigiert werden kann.

Mit der Funktion jump to token number kann man direkt zu einem bestimmten Token springen. Ins entsprechende Feld wird die Nummer des Tokens eingegeben, ein Klick auf go! bringt das Annotationsfenster zum gewünschten Token.

Mit der rechten und linken Pfeiltaste kann ebenfalls durch den Text navigiert werden. Sie verhalten sich gleich wie die Knöpfe 

. Ein Doppelklick der Maus auf ein Token oder anderes dargestelltes Element im Textfenster lässt die Annotation direkt zu diesem Element springen.

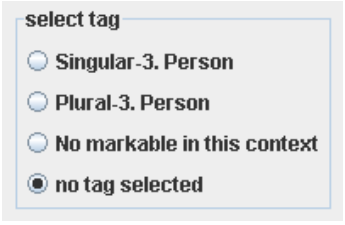
Das eigentliche Annotieren: Auswahl von Tags

Wenn das aktive Element ein Markable ist, so kann es annotiert werden. Alle Einträge im Schema, die auf das Markable passen, werden im Annotationskontrollfenster im Rahmen select tag zur Auswahl präsentiert. Es gibt zwei spezielle Tags:

- "No markable in this context": Es steht zur Auswahl, wenn im Schema die Box this element is sometimes not a markable aktiviert ist. Es wird ausgewählt, wenn das Token in diesem Kontext keine der relevanten Eigenschaften oder Funktionen besitzt, die in dieser Annotation betrachtet werden (vgl. hier [11]).

- "no tag selected": Dieses Tag steht immer zur Auswahl. Es ist ausgewählt, wenn das betreffende Markable noch nicht annotiert wurde. Bei einer neuen Annotation haben alle Markables dieses Tag.

Um ein Markable mit einem Tag zu annotieren, wählt man den Knopf vor dem Tag aus. Wenn danach zu einem anderen Element des Texts navigiert wird, so färbt sich das Markable im Annotationstextfenster entsprechend der Wahl des Tags (siehe hier). Diese Auswahl kann entweder mit einem Mausklick erfolgen oder mit der Tastatur: Mit der Tabulatortaste kann man die verfügbaren Tags nacheinander markieren, die eigentliche Auswahl kann man mit der Leertaste treffen, sobald das gewünschte Tag umrandet ist.



Die Auswahl an Tags für das Markable sie *im Beispiel der Personalpronomen.*

Speichern und Laden von Annotationen

Um eine Annotation zu speichern, wählt man den Menüpunkt "Annotation → Save" oder "Annotation → Save as...". Es müssen zwei Dateien gespeichert werden: Das Korpus und eine Annotationsdatendatei, die im wesentlichen das Schema enthält und Daten darüber, welches Token im Korpus mit welchem Tag versehen wurde; und sie enthält auch die Spezifikation des XML-Formats des Korpus. Man wird also beim Speichervorgang zum Angeben von zwei Dateinamen aufgefordert. Es ist sinnvoll, diese beiden Namen so auszuwählen, dass man sie später beim Laden leicht einander zuordnen kann, z.B. "Annotation1_korpus.xml" und "Annotation1_daten.xml". Bei "Save as..." müssen immer

Dateinamen angegeben werden, bei "Save" nur, wenn die Annotation zum ersten mal gespeichert wird.

Zum Öffnen einer gespeicherten Annotation, im Menü "Annotation → Open..." auswählen. Man wird nacheinander zur Auswahl von zwei Dateien aufgefordert, der Annotationskorpusdatei und der Annotationsdatendatei.

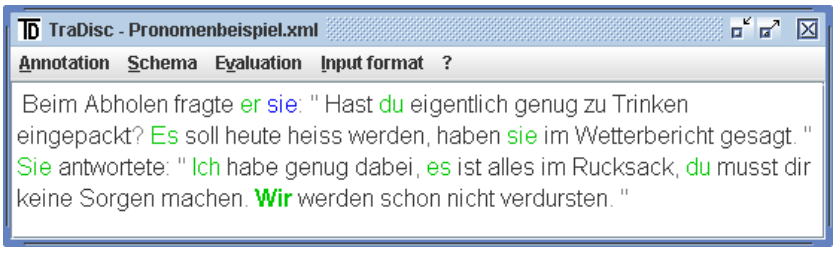
Aufteilen und Zusammenfügen von Annotationen

Aufteilen einer Annotation. TraDisc erlaubt es, die aktive Annotation in zwei Annotationen zu teilen. Dass heißt, dass das Korpus an einer vom Benutzer definierten Stelle in zwei Hälften geteilt wird. Um das zu tun, wählt man durch Doppelklick oder durch die Navigationsschaltflächen das Element des Korpus aus, welches das erste Element des zweiten Teilkorpus nach der Teilung sein soll. Man geht nun im Menü "Annotation → Split". Nun wird man aufgefordert, vier Dateinamen anzugeben: für jede Korpushälfte je einen Namen für die Annotationskorpusdatei und die Annotationsdatendatei (siehe auch hier). Nun hat man zwei gespeicherte Annotationen. Beide Annotationen haben das gleiche Schema und das gleiche XML-Korpusformat.

Zusammenfügen von zwei Annotationen. Um zwei Annotationen zu verschmelzen oder zusammenzufügen, müssen beide dasselbe XML-Korpusformat haben. Außerdem müssen die Dimensionen der jeweiligen Schemata gleich sein, das heißt die Zeilen- und Spaltenanzahl und die Name der Zeilen und Spalten müssen gleich sein. Die Annotation, deren Text am Anfang des neuen, zusammengefügten Texts stehen soll, muss in TraDisc geladen sein. Man wählt den Menüpunkt "Annotation → Merge with..." und wird aufgefordert, die Annotationskorpusdatei und die Annotationsdatendatei der hinzuzufügenden Annotation auszuwählen (siehe auch hier). Der Text des geladenen Korpus erscheint nach dem Zusammenfügen nach dem Text des aktiven Korpus im Annotationstextfenster. Die Annotationsdaten, also die Tags der Elemente des neuen Korpusteils, werden übernommen. Die Schemata der beiden Annotationen werden zu einem neuen Schema verschmolzen, das heißt, dass alle Einträge des hinzugefügten Schemas, die nicht im aktiven Schema vorhanden sind, in dieses übernommen werden (vgl. hier).

Drucken von Annotationen

Durch die Wahl von "Annotation → Print" im Menü kann man den Korpustext samt Annotationen ausdrucken. Jedes annotierte Markable, das im Annotationstextfenster grün erscheint, wird fett gedruckt. Hinter jedem annotierten Markable steht in geschweiften Klammern eine abgekürzte Version des Tags, mit dem das Markable annotiert wurde. Abgekürzt bedeutet, dass nur der Anfang des Zeilen- und Spaltennamens, die das Tag definieren, gedruckt werden. Wenn also wie im Beispiel der Personalpronomen ein Tag *Singular-3. Person* ist, so wird nur *{Singular,3.}* hinter ein mit diesem Tag annotiertes Token gedruckt. Es ist empfehlenswert, dies bei der Benennung der Schemadimensionen zu bedenken.



Beim Abholen fragte **er** *{Singular,3.}* sie: " Hast **du** *{Singular,2.}* eigentlich genug zu Trinken eingepackt? **Es** *{Singular,3.}* soll heute heiss werden, haben **sie** *{Plural,3.}* im Wetterbericht gesagt. " **Sie** *{Singular,3.}* antwortete: " **Ich** *{Singular,1.}* habe genug dabei, **es** *{Singular,3.}* ist alles im Rucksack, **du** *{Singular,2.}* musst dir keine Sorgen machen. **Wir** *{Plural,1.}* werden schon nicht verdursten. "

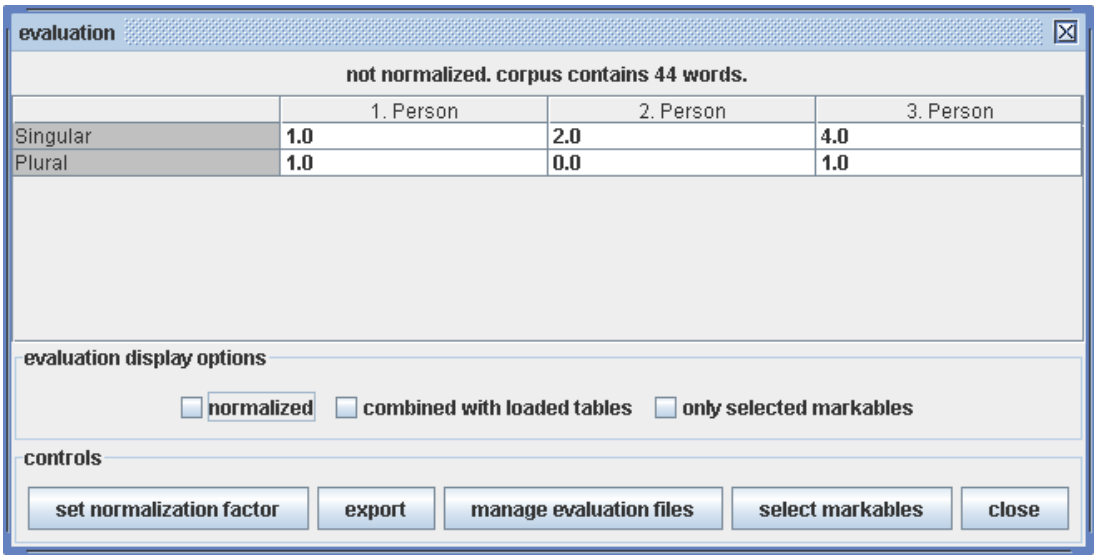
Die Beispielannotation und ihr Ausdruck.

Kapitel 4. Auswertung und Analyse von Annotationen

TraDisc stellt mehrere Möglichkeiten zur Auswertung und Analyse von Annotation zur Verfügung. Sie sind im Menü "Evaluation" zusammengefasst.

Die Evaluationstabelle

Die Evaluationstabelle wird durch Auswahl des Menüpunkts "Evaluation → Show evaluation table" geöffnet. Die Tabelle zeigt, wie oft die verschiedenen Tags (also Zeilen- und Spaltentitel) als Annotation benutzt wurden. In unserem Beispiel zeigt die Tabelle also, wieviele Personalpronomen der 1. Person Singular im Text annotiert wurden, wieviele der 2. Person Singular, der 3. Person Singular usw. Unterhalb der Tabelle befinden sich drei Auswahlboxen, normalized, combined with loaded tables und only selected markables. Beim ersten Öffnen der Evaluationstabelle ist keine der drei Boxen ausgewählt. Die Tabelle zeigt dann die Verteilung der Tags über alle Markables des Korpustextes. Die drei Optionen, die mit den Boxen gesteuert werden, werden in den folgenden Abschnitten beschrieben.



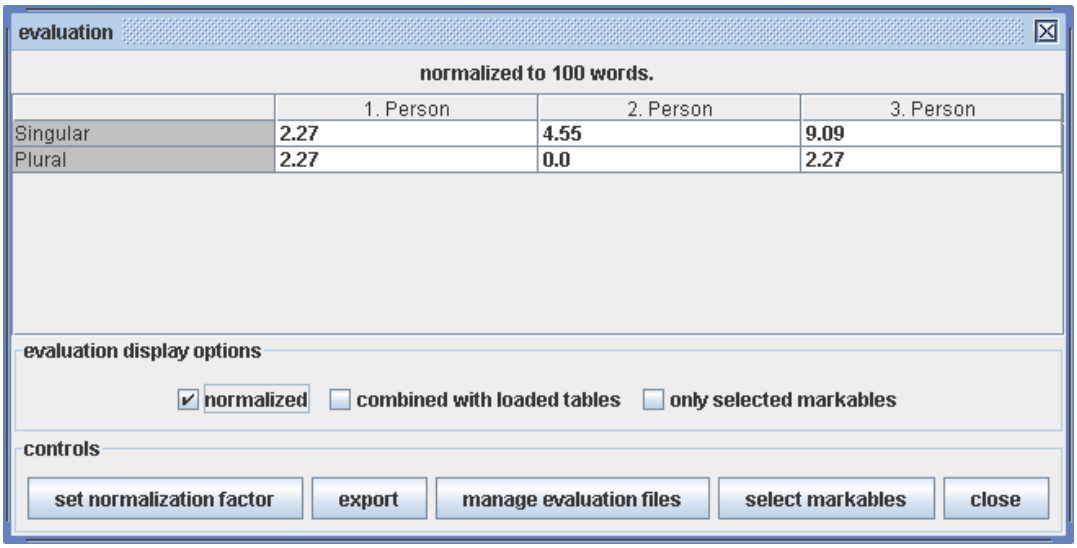
Die Evaluationstabelle des komplett annotierten Pronomenbeispiels.

Normalisieren der Evaluationswerte

In TraDisc bedeutet normalisieren, die Ergebniszahlen der Annotation auf ein Korpus mit einer gewünschten Textlänge umzurechnen, wobei die Textlänge die Anzahl der Tokens im Korpus ist. Diese gewünschte Textlänge wird als der Normalisierungsfaktor bezeichnet. Die Umrechnung der Ergebniswerte in den Tabellenzellen erfolgt nach der folgenden Formel:

$$Normalisierter\ Wert = Normalisierungsfaktor \cdot \frac{Absoluter\ Wert}{Textlänge}$$

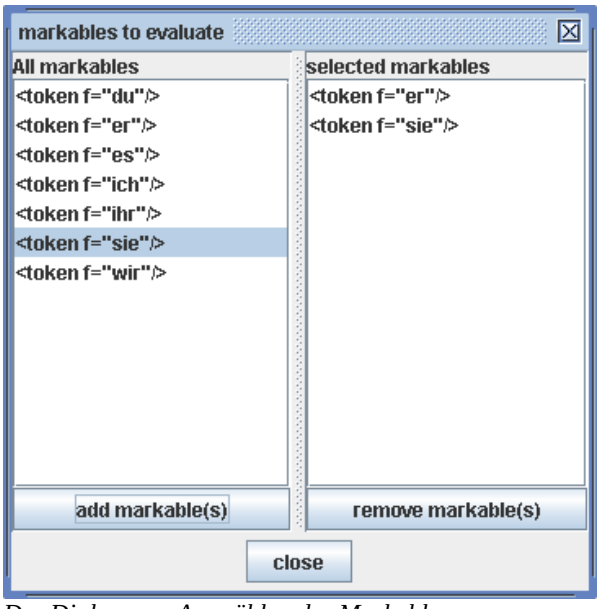
Mit dem Knopf set normalization factor kann man den Normalisierungsfaktor einstellen. Voreingestellt ist ein Normalisierungsfaktor von 1000 Tokens. Wenn das Häkchen normalized ausgewählt wird, werden die Einträge der Tabelle wie oben beschrieben umgerechnet.



Die auf 100 Tokens normalisierte Evaluationstabelle des Pronomenbeispiels.

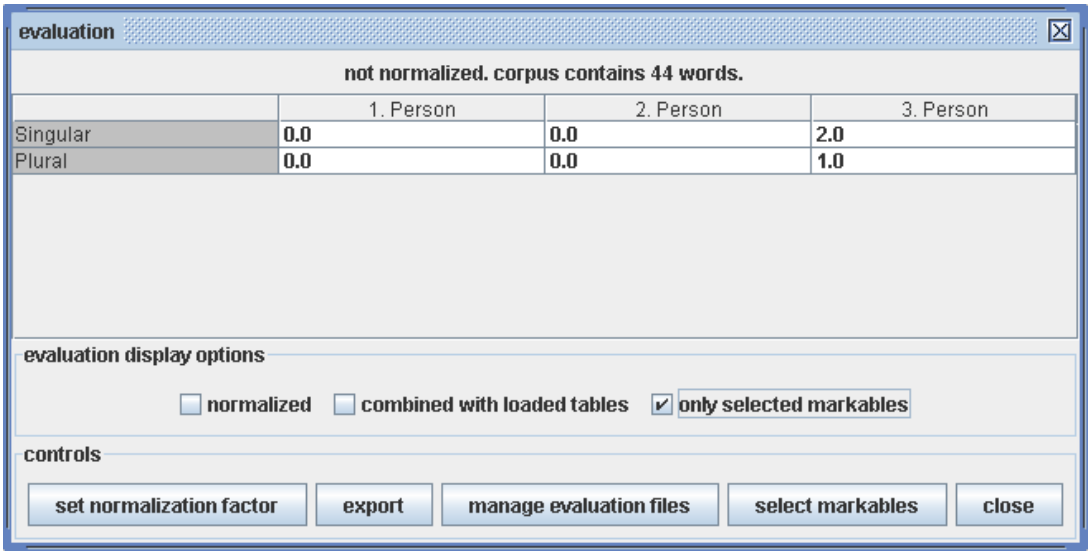
Auswahl der auszuwertenden Markables

Wer nicht die Verteilung von allen Markables auf die Tags der Tabelle erfahren will, sondern nur die Verteilung einiger bestimmter Markables, wählt die Schaltfläche select markables aus. Ein Dialog wird geöffnet, der zwei Listen zeigt: links befindet sich unter der Überschrift all markables eine Liste der Markables, die im Schema eingetragen sind. Hier kann man die Markables auswählen, an deren Auswertung man interessiert ist. Durch einen Klick auf add markable(s) fügt man sie zur rechten Liste selected markables hinzu. Aus dieser Liste können Markables auch wieder gelöscht werden. Dazu müssen sie markiert sein, ein Klick auf remove markable(s) entfernt sie von der Liste der ausgewählten Markables. Durch klicken auf close schließt man den Auswahldialog.



Der Dialog zum Auswählen der Markables.

Wenn die Liste selected markables nicht leer ist, so wird durch setzen des Häkchens only selected markables unter der Evaluationstabelle diese Liste auf die Tabelle angewandt. In den Zellen der Tabelle werden dann nur noch Markables gezählt, die auf der Liste stehen.



Die Evaluationstabelle, die nur die Werte der im obigen Bild ausgewählten Markables zeigt. Es sind also nur die Zahlen der Markables sie *und* er *gezeigt.*

Exportieren von Evaluationsdaten

Mit dem Knopf export kann man die Evaluationstabelle in eine csv-Datei (comma separated values) speichern. Dieses Dateiformat wird von gängigen Tabellenkalkulationsprogrammen wie OpenOffice.org Calc oder Microsoft Excel unterstützt. Es ist also möglich, weitere Berechnungen oder Visualisierung in Diagrammen mit einem dieser Programme vorzunehmen.

Es werden immer die Tabellenwerte exportiert, die gerade auf dem Bildschirm zu sehen sind. Man achte also darauf, wie die Optionshäkchen unter der Evaluationstabelle gesetzt sind, da die angezeigten Werte davon abhängen.

Kombinieren der Evaluationstabelle mit geladenen Evaluationsdateien

Es gibt in TraDisc die Möglichkeit, Evaluationsdaten aus einer Datei zu laden, und die Einträge der Tabelle aus dieser Datei zu den Einträgen der aktuellen Evaluationstabelle zu addieren. Die gewünschte Datei muss dabei eine Exportdatei im .csv-Format einer TraDisc-Annotation sein (vgl. vorherigen Abschnitt), und die Dimensionen (Zeilen- und Spaltenanzahl sowie -namen) der geladenen Tabelle müssen dieselben sein wie die der aktuellen Tabelle.

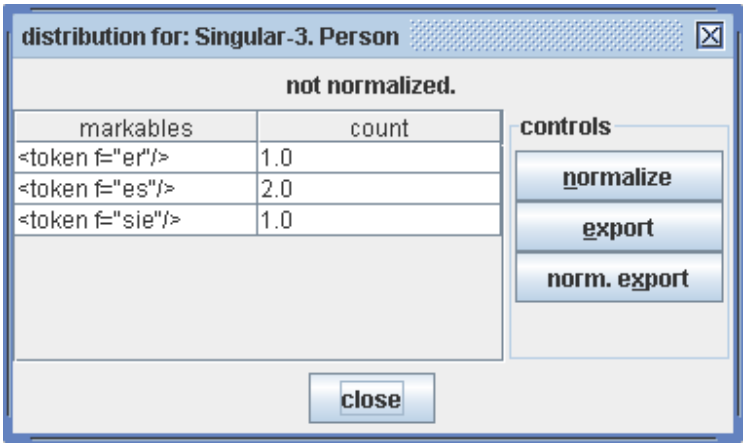
Um eine Datei zu laden, klickt man auf die Schaltfläche manage evaluation files unter der Evaluationstabelle. Ein kleiner Dialog mit dem Titel current evaluation files wird geöffnet, der aus einer Liste der gerade geladenen Evaluationsdateien (diese ist zu Beginn leer) und aus drei Schaltflächen besteht. Durch klicken der Schaltfläche add evaluation file öffnet sich ein Dateiauswahldialog, der es erlaubt, eine Evaluationsdatei zu öffnen. Es können auf diese Weise mehrere Dateien gleichzeitig geladen werden. Um eine Evaluationsdatei von der Liste zu löschen, markiert man sie und betätigt dann die Schaltfläche remove evaluation file. Durch einen Klick auf close wird der Dialog wieder geschlossen.

Um die Werte der geladenen Evaluationsdateien zu der aktuellen Tabelle zu addieren, setzt man das Häkchen combined with loaded tables. Die Werte in den geladenen Tabellen werden dann zu denen der aktuellen Tabelle addiert und die Summen angezeigt. Ist die Tabelle auch normalisiert, so werden einzelnen Werte gewichtet nach der Anzahl der Tokens im jeweiligen Korpus. Werden also zwei Tabellen kombiniert, von denen eine auf einem umfangreicheren Korpus basiert, und die kombinierte

Tabelle normalisiert, so ist das Ergebnis stärker von dieser Tabelle beeinflusst als von der Tabelle, die auf einem kürzeren Korpus basiert.

Auswertung eines Tags

Um festzustellen, für welche Markables ein Tag als Annotation benutzt wurde und wie oft, klickt man auf der dem Tag entsprechenden Zelle der Evaluationstabelle. Ein Fenster wird gezeigt mit einer Liste aller Markables, die im Schema in dieser Zelle eingetragen sind. Hinter den Markables ist aufgetragen, wie oft diese mit dem ausgewählten Tag annotiert wurden. Mit der Schaltfläche normalize kann man auch hier die Werte auf eine bestimmte Korpuslänge normalisieren, es wird der in der Evaluationstabelle eingestellte Normalisierungsfaktor benutzt (vgl. hier). Ein weiterer Klick auf normalize zeigt wieder die gezählten, nicht normalisierten Werte. Auch diese Tabelle kann in eine .csv-Datei exportiert werden, indem export geklickt wird.

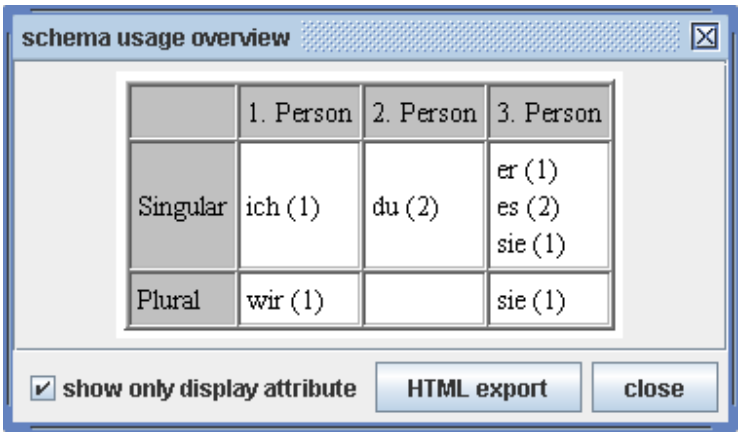


Die Verteilung des Tags "Singular-3. Person", in der rechten oberen Zelle der Evaluationstabelle des Beispiels.

Überblick über die Schemabenutzung

Um einen Überblick zu erhalten, welche Markables des Schemas im Text vorkommen und annotiert sind, kann man den Menüpunkt "Evaluation → Schema usage" wählen. Es öffnet sich ein Dialog, der dem Schemaüberblick ähnlich ist. Hier werden jedoch nur die Markables angezeigt, die mit einem der Tags der Tabelle annotiert sind. Markables, die im Schema eingetragen sind, aber nicht im Text vorkommen, werden in der Tabelle nicht angezeigt. Die Anzahl, wie oft ein Markable mit einem bestimmten Tag annotiert wurde, wird hinter dem Eintrag des Markables in der dem Tag entsprechenden Zelle in Klammern angegeben. Die Werte in einem Feld dieser Tabelle entsprechen den Werten, die im Dialog zur Auswertung eines Tags gelistet sind, der sich beim Klicken auf eine Zelle der Evaluationstabelle öffnet.

Das Häkchen show only display attribute funktioniert wie im Schemaüberblick beschrieben. Die Tabellenansicht kann mit einem Klick auf HTML export als HTML-Datei gespeichert werden.



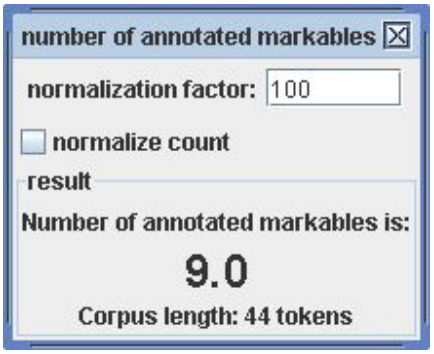
Die Schemabenutzung des Personalpronomenbeispiels.

Die Gesamtzahl der annotierten Markables

Wenn man nicht an die Aufschlüsselung der Zahl der annotierten Markables in der Tabelle interessiert ist, sondern die Anzahl aller annotierten Markables zusammengenommen wissen will, so stellt TraDisc zwei Möglichkeiten zur Verfügung.

Die Zahl aller annotierten Markables im gesamten Korpus

Um zu erfahren, wieviele Markables des Textes mit irgendeinem Tag aus der Tabelle annotiert sind, (das ist also die Anzahl der Markables, die im Annotationstextfenster grün dargestellt sind; rote, noch nicht annotierte Markables und blaue Markables, die in ihrem Kontext keine der zu annotierenden Eigenschaften haben, werden nicht gezählt), so wählt man im Menü "Evaluation → Count annotated markables". Es erscheint ein kleines Fenster, in dem der gewünschte Wert gezeigt wird. Falls die Anzahl der annotierten Markables auf eine bestimmte Textlänge normalisiert werden soll, so trägt man diese Textlänge im Feld normalization factor ein und setzt das Häkchen normalize count.



Der Dialog zum Zählen der Gesamtzahl der annotierten Markables in der Beispielannotation.

Verteilung der annotierten Markables auf Teile des Korpus

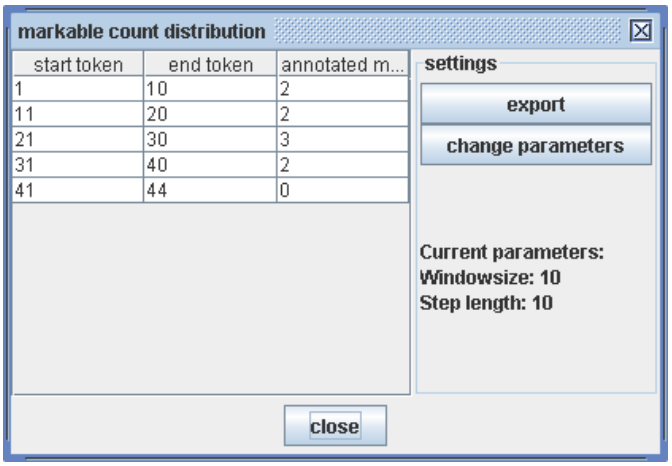
Man kann mit TraDisc das Korpus in Teile gleicher Länge aufteilen und die Anzahl der annotierten Markables in jedem dieser Teile gesondert zählen lassen. Damit erhält man eine Art Verteilung der annotierten Markables im Korpus und kann interessante Teile des Korpus identifizieren, in denen es z.B. überdurchschnittlich viele annotierte Markables gibt. Diese Funktion befindet sich im Menü "Evaluation

→ Distribution of annotated markables". Der Dialog zeigt eine Tabelle, in der die jeweiligen Anzahlen der annotierten Markables in den Teilen des Korpus aufgetragen sind. Jeder Korpusteil beansprucht dabei eine Tabellenzeile. Die Nummer des ersten Tokens eines Teils steht in der Spalte start token, die Nummer des letzten Tokens in der Spalte end token. Man kann sich das so vorstellen, dass ein Fester einer bestimmten Textlänge über den Korpustext geschoben wird. TraDisc zählt für jede Position des Fensters, wieviele annotierte Markables im gerade durchs Fenster sichtbaren Korpusteil vorhanden sind. Wie lang dieses Fenster ist (also wieviele Tokens es jeweils beinhaltet), und um wieviele Tokens es verschoben werden soll (wie groß der "Schritt" von einer Fensterposition zur nächsten sein soll), kann mit den Parametern "Window size" und "Step length" festgelegt werden. Mit einem Klick auf change parameters kommt man auf einen kleinen Dialog, mit dem die Parameterwerte geändert werden können.

Anmerkung

Das letzte Fenster ist möglicherweise kürzer als die eingestellte Fenstergröße, da am Korpusende eventuell nicht mehr genügend Tokens vorhanden sind. Dies kann man anhand der Nummern der Anfangs- und Schlusstokens in der letzten Tabellenzeile feststellen.

Die Verteilungstabelle kann man mit dem Knopf export in einer .csv-Datei speichern (vgl. hier).



Die Verteilung der annotierten Markables in der Beispielannotation, mit einer Fensterlänge und einer Schrittweite von 10 Tokens.

Berechnung des Komplexitätswertes einer Annotation

TraDisc bietet die Möglichkeit, einer Annotation einen sogenannten Komplexitätswert zuzuordnen. Dieser beruht darauf, wie oft die unterschiedlichen Tags des Schemas als Annotationen vergeben wurden. Jedem Tag wird dabei ein Wert entsprechend seiner Komplexität zugeordnet; dies wird mittels der Komplexitätswerttabelle gemacht.

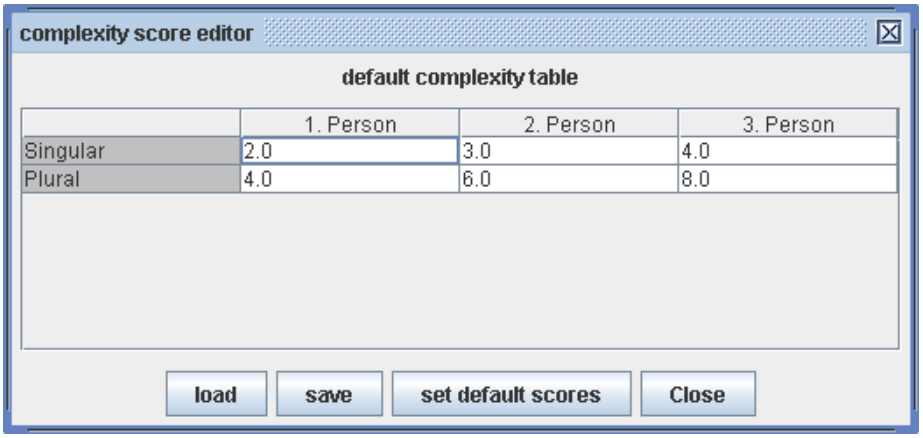
Die Komplexitätswerttabelle

Im Menü "Evaluation → Complexity score table" gelangt man zur Komplexitätswerttabelle. Die Dimensionen sind dieselben wie die der Schematabelle, ebenso die Namen der Zeilen und Spalten. In einer Zelle der Tabelle steht der Komplexitätswert des Tags, das der Zeile und Spalte der Zelle entspricht. Um den Komplexitätswert eines Tags zu ändern, führt man einen Doppelklick auf der entsprechenden Zelle aus. Man kann nun den gewünschten Wert eingeben.

Wenn sie vom Benutzer nicht geändert wurden, sind die Komplexitätswerte der Tags auf Standardwerte

gesetzt. Um alle Werte der Tabelle auf diese Standardwerte zurückzusetzen, kann man die Schaltfläche set default scores betätigen. Der Standardwert der Zellen wird größer, je größer die Zeilen- und Spaltennummern werden. Dies hat seinen Grund darin, dass TraDisc zum Annotieren der Junktoren von Korpora entwickelt wurde. Im dafür benutzten Junktorenschema nimmt die Komplexität der Junktoren zu, je weiter rechts unten in der Schematabelle sich die entsprechende Zelle befindet.

Mit einem Klick auf save kann die aktuelle Komplexitätswerttabelle in eine .csv-Datei gespeichert werden. Solch eine Datei kann auch wieder geladen werden, mit dem Knopf load, Voraussetzung ist, dass die Dimensionen sowie Zeilen- und Spaltennamen der zu ladenden Komplexitätswerttabelle mit denen des aktuellen Schemas übereinstimmen.



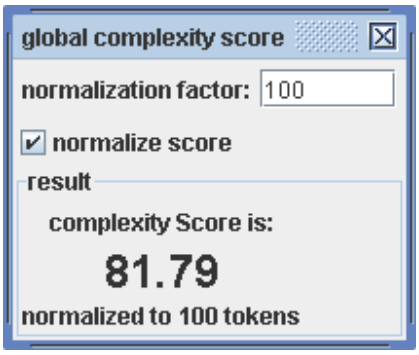
Die Standard-Komplexitätswerttabelle für das Personalpronomenbeispiel.

Berechnung des Komplexitätswertes für eine ganze Annotation

Mit dem Menüpunkt "Evaluation → Complexity score" öffnet man einen Dialog, der den Komplexitätswert der Annotation anzeigt. Die Berechnung des Wertes wird im Folgenden beschrieben. Für jedes Tag wird gezählt, wie oft Markables es als Annotation zugewiesen bekamen. (Das ist dieselbe Zahl, die in der Evaluationstabelle in der Zelle steht, die dem Tag entspricht.) Diese Anzahl wird mit dem Komplexitätswert des Tags aus der Komplexitätswerttabelle multipliziert. Man hat nun also die komplette Komplexität der Annotation bezüglich eines Tags berechnet. Dieser Wert wird für jedes Tag berechnet, und aus allen wird die Summe der Tag-Komplexitätswerte gebildet. Die Gesamtsumme über alle Tag-Komplexitäten ist der endgültige Komplexitätswert der gesamten Annotation, der im Dialog global complexity score dargestellt ist. Etwas formaler ist die Berechnung in der folgenden Formel dargestellt:

$$Gesamtkomplexitätswert = \sum_{Tags} (Anzahl\ der\ Markables\ mit\ dem\ Tag\ als\ Annotation) \cdot (Komplexitätswert\ des\ Tags)$$

Der Komplexitätswert kann auch auf eine beliebige Textlänge normalisiert werden, dazu muss der gewünschte Normalisierungsfaktor in das Feld normalization factor eingetragen werden und das Häkchen normalize score gesetzt sein (vgl. hier).

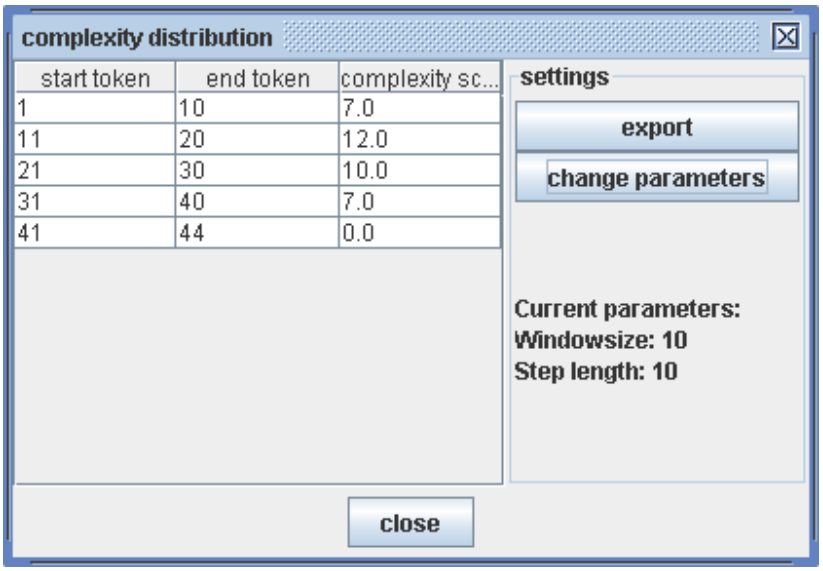


Der auf 100 Tokens normalisierte Komplexitätswert der Beispielannotation.

Verteilung der Komplexität auf Teile des Korpus

Wenn man feststellen möchte, ob die Komplexität der annotierten Markables im ganzen Korpus gleich sind, oder ob es Regionen im Korpus gibt, in denen die Komplexität höher oder niedriger ist, so kann man die Funktion "Evaluation → Complexity Distribution" im Menü wählen. Der Dialog zeigt eine Tabelle, in der die Komplexitätswerte für Teile des Korpus aufgelistet sind, für jeden Korpusteil ist die Nummer des ersten und des letzten Tokens angegeben und der Komplexitätswert. Es werden für dessen Berechnung nur die Komplexitäten derjenigen Markables summiert, die innerhalb der Grenzen liegen.

Die Parameter "Window size" und "Step length" und die Schaltflächen des Dialogs verhalten sich äquivalent zu den gleichnamigen Parametern beziehungsweise Schaltflächen im Dialog für die Verteilung der annotierten Markables.



Die Komplexitätswertverteilung der Beispielannotation mit der Fensterlänge und Schrittweite von je 10 Tokens.

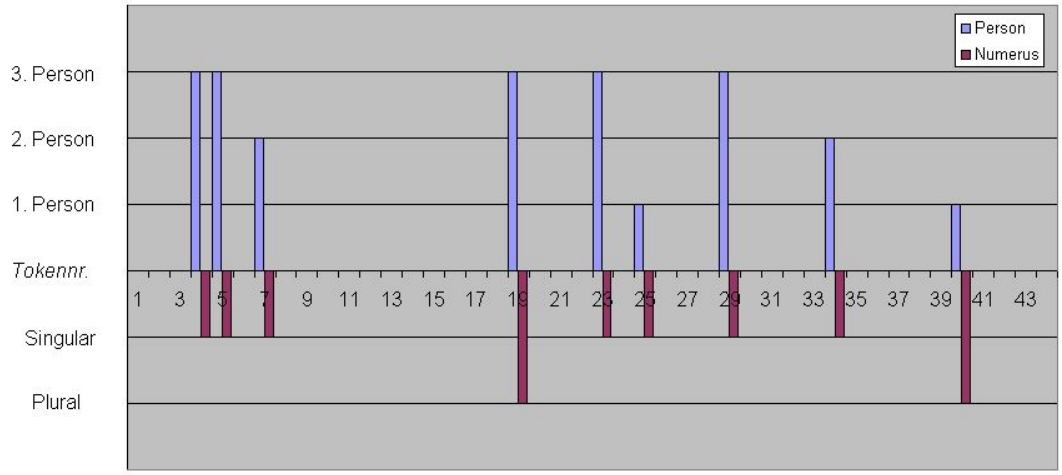
Erstellen von Junktogrammen

Ein Junktogramm ist eine Visualisierung der annotierten Funktionen über den Textverlauf (der Name *Junktogramm* kommt daher, dass TraDisc ursprünglich entwickelt wurde, um Junktoren zu annotieren). Jedem annotierten Markable werden dabei zwei Zahlen zugewiesen; diese sind abhängig davon, mit welchem Tag das Markable annotiert wurde, genauer gesagt, durch welche Schemazelle das Tag repräsentiert wird. Jedem Spaltennamen wird eine positive Zahl entsprechend der Position der Spalte

zugewiesen. Jeder Zeile wird ihre Position als negative Zahl zugewiesen. Im Beispielschema der Personalpronomen sind die Zuordnungen also folgendermaßen: *1. Person: 1*, *2. Person: 2*, *3. Person: 3* für die Spalten und *Singular: -1*, *Plural: -2* für die Zeilen. Den restlichen, unannotierten Markables und den Tokens, die keine Markables sind, wird zweimal Null zugewiesen. Man kann nun ein Schaubild erstellen, das auf der x-Achse die Tokennummer aufträgt, und auf der y-Achse jeder Tokennummer den oben genannten Wert zuordnet. Mit den Visualisierungsmöglichkeiten von Tabellenkalkulationsprogrammen wie OpenOffice.org Calc oder Microsoft Excel kann so eine nützliche Veranschaulichung der Annotation erstellt werden.

Um in einem Tabellenkalkulationsprogramm benutzbar zu sein, wird diese Zuweisung von Werten in eine Tabelle in einer csv-Datei exportiert. Im Menüpunkt "Evaluation → Export junctogram" kommt man zuerst auf einen Dialog, in dem man die Markables spezifizieren kann, die im Junktogramm berücksichtigt werden sollen. Man kann entweder den Knopf all markables wählen (er ist standardmäßig schon vorausgewählt), oder den Knopf markables selected below. Unter den beiden Knöpfen sind zwei Listen gezeigt; in der linken sind alle Markables des Schemas aufgeführt. Hier kann man die gewünschten Markables auswählen und mit einem Klick auf add markable(s) in die rechte Liste hinzufügen. Diese rechte Liste wird dann für das Erstellen der Junktogrammdaten benutzt, falls markables selected below aktiviert ist. Markables, die nicht in der Liste sind, wird in diesem Fall zweimal Null zugewiesen, egal ob sie annotiert sind oder nicht.

Mit einem Klick auf export werden die Daten als Tabelle in einer csv-Datei gespeichert. Diese Tabelle hat vier Spalten: Die Nummer jedes Tokens des Korpus in der Spalte *token number*, das Token in der Spalte *token*, und die beiden Werte, die dem Token wie oben beschrieben zugewiesen wurden, in den Spalten *vertical dimension value* für den Wert der Schemaspalte und *horizontal dimension value* für den Wert der Schemazeile. Die Zuordnung der Zeilen- und Spaltennamen des Schemas zu den Werten ist in der csv-Datei über der eigentlichen Junktogrammtabelle eingetragen. Diese Tabelle kann nun zur Erstellung eines Diagramms benutzt werden.



Ein Junktogramm erstellt aus der Beispielannotation der Personalpronomen.

Teil II. Tokenizer

Mit Tokenizer ist es möglich, ein Korpus in ein für TraDisc lesbares XML-Format zu bringen, das TraDisc Standardformat. Dies ermöglicht es dem Benutzer, jedes Korpus, das in einem einfachen Textformat vorliegt, mit TraDisc zu annotieren.

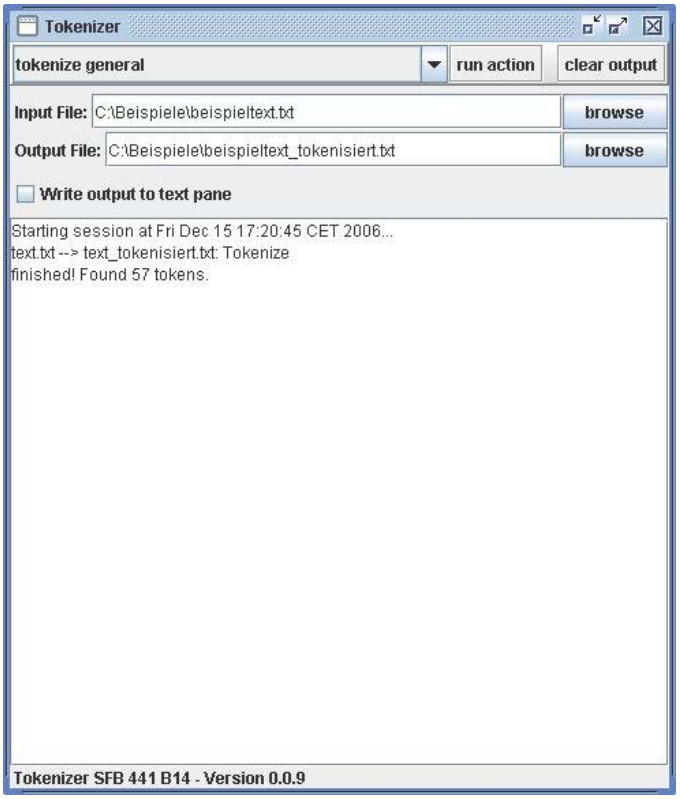
Kapitel 5. Verwendung von Tokenizer

Damit ein Text als Eingabe von Tokenizer verwendet werden kann, muss er in einer *plain text*-Datei vorliegen. Das ist das Standardformat für Textdateien, die Dateiendung ist meist *.txt*.

Tokenisieren des Textes

Das Tokenisieren eines Korpus bedeutet, die einzelnen Tokens des Korpustextes zu identifizieren. Im einfachsten Fall sind das die einzelnen Wörter (und Satzzeichen). Dies wird von Tokenizer automatisch durchgeführt, es wird eine temporäre Datei geschrieben, in der jedes Token in einer eigenen Zeile steht. Um diese Datei herzustellen, muss man in Tokenizer im Feld Input File den Namen und Pfad der Datei eingeben, in der der Korpus gespeichert ist. Mit dem Knopf Browse kommt man auf einen Dateiauswahldialog, mit dem man diese Datei auswählen kann. Im direkt darunter stehenden Textfeld Output File muss der Name der temporären Datei angegeben werden, auch hier kann man den Dateiauswahldialog mit Browse öffnen.

In der Auswahlbox in der linken oberen Ecke kann man auswählen, welche Aktion Tokenizer ausführen soll. Um einen Text zu tokenisieren, wählt man tokenize general aus, das auch schon beim Programmstart vorausgewählt ist. Wenn nun die Ein- und Ausgabedateien festgelegt sind, so startet ein Klick auf run action die gewünschte Aktion. Das große Textfeld in der unteren Hälfte von Tokenizer wird Informationen darüber liefern, ob die Aktion erfolgreich durchgeführt wurde und wieviele Tokens gefunden wurden.



Tokenizer nach dem Tokenisieren des Beispieltextes.

Der Text des Korpus der Beispielannotation lautet folgendermaßen:

Beim Abholen fragte er sie: "Hast du eigentlich genug zu Trinken eingepackt? Es soll heute heiss werden, haben sie im Wetterbericht gesagt." Sie antwortete: "Ich habe genug dabei, es ist alles im Rucksack, du musst dir keine Sorgen machen. Wir werden schon nicht verdursten."

Wenn man nun eine Datei mit diesem Text in obiger Form als Eingabedatei angibt, und den Text tokenisiert wie oben beschrieben, so sieht der Inhalt der temporären Datei folgendermaßen aus:

Beim
Abholen
fragte
er
sie
:
"
Hast
du
eigentlich
genug
zu
Trinken
eingepackt
?
Es
soll
heute
heiss
werden
,
haben
sie
im
Wetterbericht
gesagt
.
"
Sie
antwortete
:
"
Ich
habe
genug
dabei
,
es
ist
alles
im
Rucksack
,
du
musst
dir
keine

Sorgen
machen
.
Wir
werden
schon
nicht
verdursten
.
"

Wenn man vor dem Starten des Tokenisierens das Häkchen Write output to text pane setzt, so werden die zeilengetrennten Tokens auch im Tokenizerfenster im großen Textfeld ausgegeben, in der gleichen Art, wie sie in die temporäre Ausgabedatei geschrieben werden. Um den Text auf der Ausgabefläche zu löschen, kann man die Schaltfläche clear output verwenden.

Anmerkung

Es gibt in der Auswahlbox von Tokenizer noch einige spezielle Optionen zum Tokenisieren von Texten in Altspanisch und Surselvisch, sowie zum Eliminieren von bestimmten Satzzeichen. Diese Funktionen wurden im Hinblick auf die Funktion von TraDisc zum Annotieren von Junktoren in romanischen Texten entwickelt. Sie werden im Rahmen dieses Handbuchs nicht näher erläutert.

Erzeugen der XML-Datei

Um aus der temporären Datei, welche je ein Token in einer Zeile geschrieben hat, nun eine XML-Datei mit dem Text im TraDisc Standard-XML-Format zu erstellen, muss diese temporäre Datei als Eingabedatei im Feld Input file gewählt sein (Man muss also nach dem Tokenisieren die Datei, die als Ausgabedatei gewählt war, nun als Eingabedatei festlegen). Als Ausgabedatei setzt man den Pfad und den Dateinamen, den die XML-Korpusdatei haben soll, normalerweise mit der Endung *.xml*. Man wählt nun in der Aktionsauswahlbox create xml output und klickt auf run action. Tokenizer erstellt eine XML-Datei, in der jede Zeile der Eingabedatei mit einem XML-Element korrespondiert. Tokenizer erkennt automatisch, ob eine Zeile ein Wort enthält oder ein Satzzeichen, und setzt entsprechend die XML-Tag-Namen der Elemente auf token oder other (vgl. hier). Diese XML-Datei kann nun als Korpusdatei in TraDisc benutzt werden.

Wenn man die temporäre Tokendatei aus dem obigen Beispiel in eine XML-Datei wie beschrieben umwandelt, so sieht der Inhalt dieser XML-Datei folgendermaßen aus:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<corpus>
   <token f="Beim"/>
   <token f="Abholen"/>
   <token f="fragte"/>
   <token f="er"/>
   <token f="sie"/>
   <other f=":"/>
   <other f="&amp;quot;"/>
   <token f="Hast"/>
   <token f="du"/>
   <token f="eigentlich"/>
   <token f="genug"/>
   <token f="zu"/>
   <token f="Trinken"/>
```

```
    <token f="eingepackt"/>
    <other f="?"/>
    <token f="Es"/>
    <token f="soll"/>
    <token f="heute"/>
    <token f="heiss"/>
    <token f="werden"/>
    <other f=","/>
    <token f="haben"/>
    <token f="sie"/>
    <token f="im"/>
    <token f="Wetterbericht"/>
    <token f="gesagt"/>
    <other f="."/>
    <other f="&amp;quot;"/>
    <token f="Sie"/>
    <token f="antwortete"/>
    <other f=":"/>
    <other f="&amp;quot;"/>
    <token f="Ich"/>
    <token f="habe"/>
    <token f="genug"/>
    <token f="dabei"/>
    <other f=","/>
    <token f="es"/>
    <token f="ist"/>
    <token f="alles"/>
    <token f="im"/>
    <token f="Rucksack"/>
    <other f=","/>
    <token f="du"/>
    <token f="musst"/>
    <token f="dir"/>
    <token f="keine"/>
    <token f="Sorgen"/>
    <token f="machen"/>
    <other f="."/>
    <token f="Wir"/>
    <token f="werden"/>
    <token f="schon"/>
    <token f="nicht"/>
    <token f="verdursten"/>
    <other f="."/>
    <other f="&amp;quot;"/>
</corpus>
```