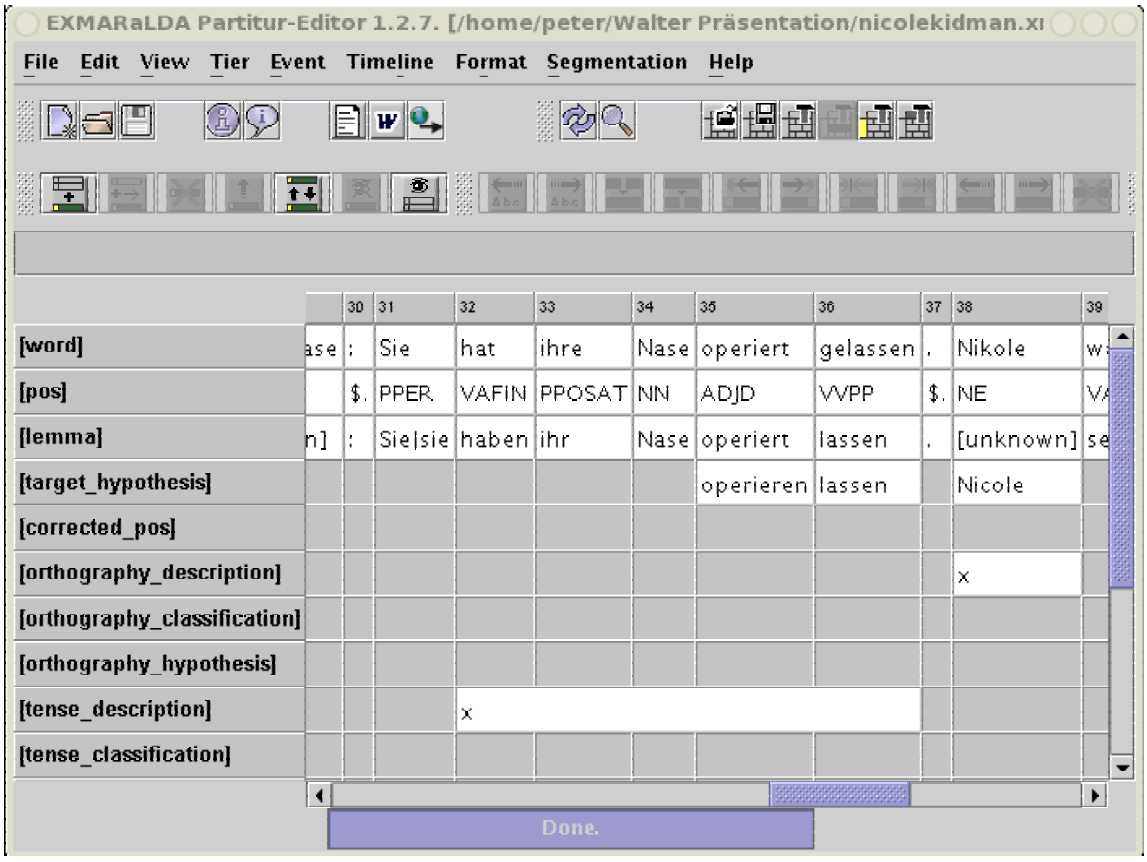
Peter Adolphs
Emil Kroymann

Technik und Arbeitsablauf für FALKO

1 Software

1.1 EXMARaLDA Partitur-Editor

EXMARaLDA ist ein Annotationswerkzeug für linguistische Korpora. Es wurde von der Universität Hamburg entwickelt (http://www.rrz.uni-hamburg.de/exmaralda/). Es ist ein Java-Programm und läuft daher auf Windows, MacOSX, Linux, ...



Installation: siehe http://www.rrz.uni-hamburg.de/exmaralda/

Wie sieht ein Text aus dem Korpus in EXMARaLDA aus?

- eine Annotationsschicht heißt hier *Tier*

- es gibt eine *Timeline* (entspricht in etwa der Korpuspositionen in CQP)
- zwischen verschiedenen Punkten auf der Timeline können *Events* definiert werden (müssen aber nicht)
- Events können sich auch über mehrere Punkte (= mehrere Token) erstrecken
- Events enthalten einen Text (die Annotation bzw. beim word-Tier das Text-Token selbst)
- Texte können Meta-Informationen (siehe File / Meta information)

Idee zur Fehlerannotation:

- Pro Fehlerart (Orthographie, Kongruenz, Tempus, etc.) gibt es eine Annotations*ebene*
- Pro Annotationsebene gibt es 3 *Schichten*: Description, Classification und Hypothesis
- Description: kennzeichnet, *ob* auf der Ebene ein Fehler vorliegt
- Classification: gibt an, um *welchen* Fehler es sich genau handelt (benötigt ein Tagset)
- Hypothesis: gibt eine Hypothese darüber an, *warum* dieser Fehler hier auftritt (benötigt Tagset)

Unsere Annotationskonvention:

- auf der Description-Schicht steht ein "x", falls ein Fehler auf dieser Ebene vorliegt
- für alle anderen Token wird kein Event definiert

1.2 Korpuswalter

Korpuswalter ist das von uns entwickelte Webinterface zur Korpusverwaltung. Korpuswalter übernimmt verschiedene Aufgaben bei der Korpuserstellung:

1. Sammeln von Texten und zugehörigen Annotationsschichten
2. Konsistenz der Annotation absichern
3. Konvertierung der Texte in verschiedene Formate
4. Texte für CQP aufbereiten

Die Texte können von verschiedenen Nutzern gleichzeitig bearbeitet werden. Korpuswalter kann noch nicht verhindern, dass es dabei zu Konflikten kommt. Deshalb müsst ihr solche Konflikte durch eure Absprachen ausschließen.

Korpuswalter ist noch sehr jung ;-) Er hat einige wünschenswerte Dinge noch nicht gelernt, und ist darauf angewiesen, dass Ihr ihm helft!

Die drei Hauptansichten von Korpuswalter sind:

1.2.1 Katalogansicht

In dieser Ansicht werden alle von Korpuswalter verwalteten Korpora aufgeführt, und man gelangt zur *Korpusansicht* eines einzelnen Korpus.

1.2.2 Korpusansicht

Diese Ansicht dient dazu, die Korpusdefinition und den Inhalt eines Korpus einzusehen. Die Korpusdefinition legt fest, welche Headerinformationen und welche Annotationsschichten zu einem Text annotiert werden können oder müssen.

1.2.3 Textansicht

Die Textansicht stellt einen Auszug aus dem Text, den Header und die Annotationsebenen des Texts dar. Von hier erreicht man auch die Aktionen *Text anfordern* und *Text abgeben*.

Mit der Aktion *Text anfordern* wird ein Text im EXMARaLDA XML-Format heruntergeladen. Dabei muss die Annotationsebene ausgewählt werden, die bearbeitet werden soll.

Mit der Aktion *Text abgeben* wird ein bearbeiteter Text im EXMARaLDA XML-Format hochgeladen. Dabei muss die Annotationsebene ausgewählt werden, die bearbeitet wurde. Es muss sich um die selbe Ebene handeln die beim Anfordern des Textes gewählt wurde. Andernfalls schlägt die Aktion fehl. Achtung die Aktion überschreibt eventuell schon vorhandene Informationen auf der gewählten Annotationsebene!

1.3 CQP-Webinterface

FALKO kann über das CQP-Webinterface[1] durchsucht werden. Dafür müssen die Texte in das Corpus-Workbench-Format konvertiert werden. Die automatische Konvertierung erfolgt einmal pro Nacht. Änderungen sind also erst ab dem nächsten Tag im CQP-Webinterface verfügbar.

EXMARaLDA hat eine allgemeinere Sicht auf Korpora als CQP. Folge: Es kann mit CQP nicht nach allem gesucht werden, was auch annotiert wurde! Problematisch sind insbesonders Annotationen, die sich auf mehrere Token gleichzeitig beziehen.

CQP unterscheidet

- positionelle Attribute
  - Annotationen von Token (eigentlich "Korpuspositionen")
  - pro Annotationsschicht ein positionelles Attribut
- strukturelle Attribute
  - erlauben die Zusammenfassung von Bereichen von Token unter einem Namen
  - Anfangs- und Endmarkierer stehen zwischen Token
  - können ineinander geschachtelt werden
  - Aber: keine Überschneidungen möglich!! :-(
  - können selbst Attribut-Wert-Paare besitzen (entspricht XML-Elementen mit Anfangs- und Endtags)

Problem: wie werden Annotationen, die sich in EXMARaLDA über mehrere Token erstrecken, in CQP umgesetzt? (Positionelle Attribute gibt es immer nur pro Token. Strukturelle Attribute dürfen sich nicht überschneiden.)

Antwort: normalerweise werden sie als positionelle Attribute umgesetzt.

Beispiel: mit EXMARaLDA werden in der Schicht "tense_description" Bereiche von Token, in denen ein Tempusfehler vorliegt, zusammengefasst und mit "x" markiert (siehe oben). In CQP gibt es dann *zu jedem Token* das positionelle Attribut `tense_description`. Unter jedem Token, das in EXMARaLDA in einem mit "x" markierten Bereich liegt, steht in CQP ebenfalls ein "x".

Nachteil: wenn zwei solche Bereiche aneinander grenzen, entsteht dadurch in CQP ein einziger großer Bereich. Die Bereiche sind nicht mehr unterscheidbar.

---

[1] CQP (für Corpus Query Processor) ist ein an der Universität Stuttgart im Rahmen der Corpus WorkBench entwickeltes Suchwerkzeug für Korpora, http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/ . Wir haben hier ein Webinterface für CQP entwickelt. Sie können sich dafür unter http://www.linguistik.hu-berlin.de/korpuslinguistik/korpora/index.php anmelden.

[word]	Ich	habe	fertig	.	...
[tense_description]		x			

Abbildung 1: Annotationen über Bereiche von Token in EXMARaLDA

word	Ich	habe	fertig	.	...
tense_description		x	x		

Suche in CQP:

- `[tense_description="x"]`
- `[lemma="haben" & tense_description="x"]`
- `[lemma="haben" & tense_description="x"][tense_description="x"]+`

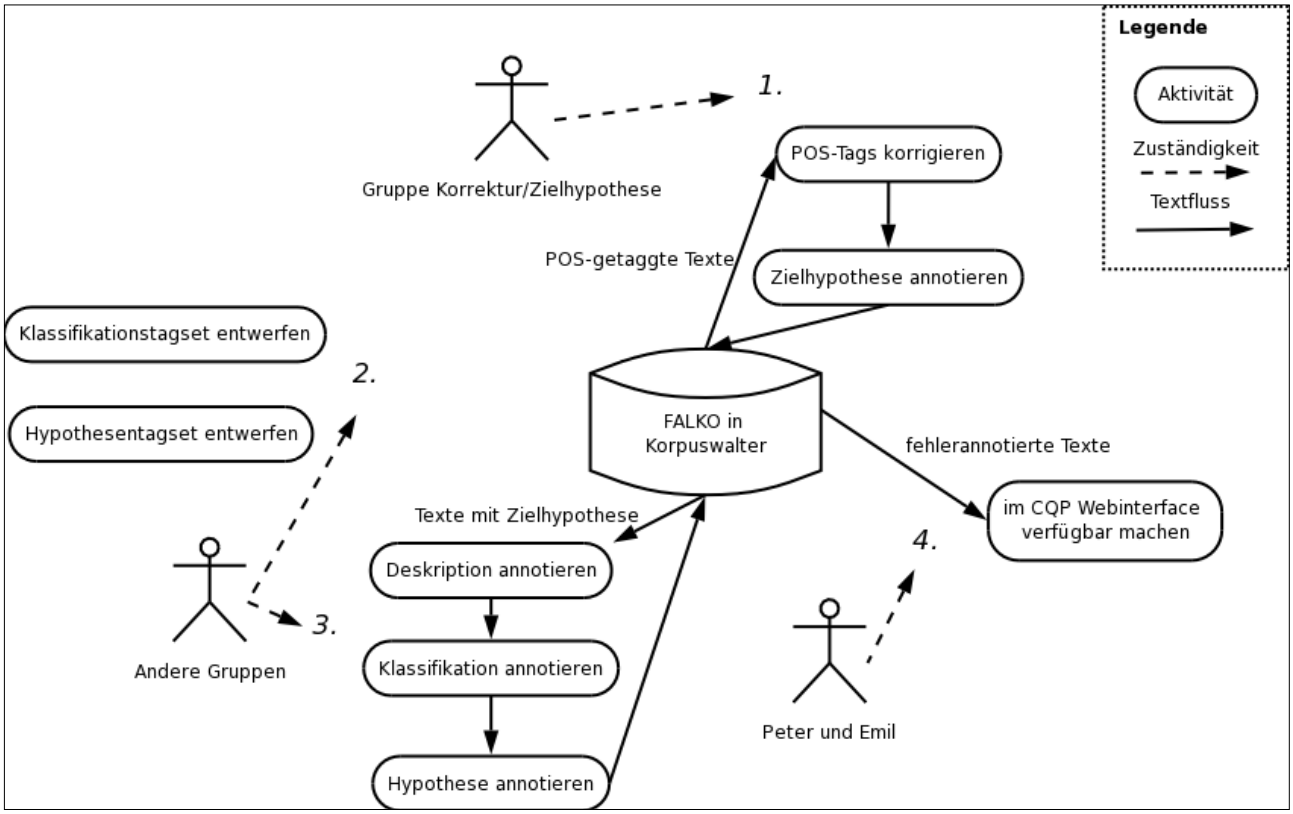
2 Arbeitsablauf

2.1 Wichtige Anmerkungen

Die Aufgaben für jede Gruppe sind genau festgelegt. Dennoch bestehen teilweise Abhängigkeiten zwischen den einzelnen Gruppen. Damit diese Abhängigkeiten nicht, dazu führen dass einige Gruppen auf andere warten müssen, sind viel Kommunikation und verlässliche Absprachen nötig.

2.2 Aktivitäten und Zuständigkeiten

Das Diagramm versucht die einzelnen Aktivitäten und die Zuständigkeiten zu illustrieren.



2.2.1 Vorverarbeitung

Die Vorverarbeitung ist die Aufgabe der Gruppe "Gruppe1". Zur Vorverarbeitung gehören die Schritte:

- Korrektur der Part-Of-Speech Tags auf einer eigenen Ebene
- Annotation der Zielhypothese

Wichtig: Die anderen Gruppen können erst mit der Annotation beginnen, wenn die Vorverarbeitung abgeschlossen ist.

2.2.2 Entwerfen eines Klassifikations- und eines Hypothesentagsets

Für jede Annotationsebene, muss ein Klassifikations- und ein Hypothesentagset entworfen werden,

dass zur Annotation der Klassifikationsschicht bzw. der Hypothesenschicht dieser Ebene verwendet werden soll. Dies ist Aufgabe, der für die jeweiligen Annotationsebenen zuständigen Gruppen.

Wichtig: Das Tagset muss uns mitgeteilt werden, bevor die Annotation beginnt.

2.2.3 Annotation der Texte

Wenn die Vorverarbeitung abgeschlossen ist und die Klassifikations- und Hypothesentagsets entworfen wurden, können die einzelnen Gruppen mit der Annotation der Texte beginnen.

Die Annotation eines Textes verläuft nach folgendem Schema

1. Text von Korpuswalter anfordern, dabei die richtige Annotationsebene auswählen.

2. Annotation mit EXMARaLDA durchführen.

3. Text an Korpuswalter zurückgeben. Auch hier die richtige Annotationsebene auswählen.

Wichtig: Die Gruppenmitglieder müssen sich untereinander einigen, wer welchen Text bearbeitet. Sonst kommt es zu Konflikten. Korpuswalter kann das im Moment noch nicht verhindern.