IS-Annotation: Beschreibung, Anleitung, Tags

Anleitung zur Annotation

Die annotierten Texte befinden sich im Abschnitt „Annotated Text“ im Korpus.

Parsing

Mithilfe der Information zur Bestimmung der Wortarten in FLEx werden zuvor festgelegte Kombinationen zu Phrasen geparst, z.B. die Abfolge Demonstrativpronomen Substantiv zu Nominalphrase (NP), Adjektiv Substantiv zu NP oder Präfix und Verb zur finiten Verbalphrase (finVP). Der Parser erfasst dabei eine Einheit nach der nächsten und gruppiert eine Phrase, sobald sie einer als Befehl im Skript eingetragenen Kombination entspricht. Auf diese Weise können auch feinere Unterschiede getroffen werden (z.B. NP und postpositionaler Phrase (PostP)). Besteht eine Zeile der Glossierung aus mehreren Sätzen (d.h. mehreren Prädikaten), so können diese mit eckigen Klammern voneinander getrennt werden. Die Bestandteile der einzelnen Sätze erscheinen dann jeweils in einer eigenen Farbe.
Zur besseren Orientierung und Zuordnung der einzelnen Satzglieder wird die fortlaufende Nummerierung mit in der Glossierung angezeigt. Zusammengehörende Satzglieder sind mit eckigen Klammern gekennzeichnet, die zugehörigen Felder stehen unter dem Bestandteil, das den rechten Rand der Phrase bildet, d.h. head-final.

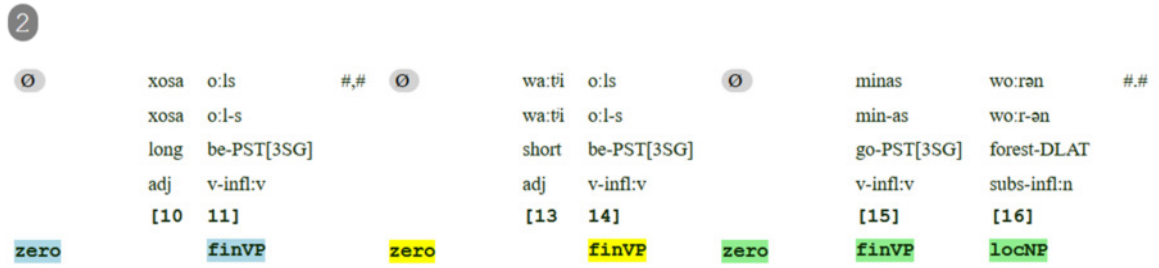

Abbildung 1: Zuordnung in mehrere Sätze (farbig) im Parsing; NM (ID 745, Nr. 2)

Zusätzlich zur Visualisierung unter der Glossierung wird im Feld *phrasal annotation* das Parsing in Form von zuvor festgelegte Parametern schematisch dargestellt. In diesem Feld kann gegebenenfalls das Ergebnis erweitert oder korrigiert werden (s. Abbildung 2).

Nach dem ersten Speichern der vorgeschlagenen Phrasen heißt der entsprechende Knopf dann automatisch „resave phrasal annotation“. Bei Bedarf kann oben auch der Modus „Reset Annotation Mode“ ausgewählt werden (Die Default-Ansicht heißt „Regular Annotation Mode“). Dann erscheint zusätzlich das Feld *original parsing result*, welches nicht bearbeitet werden kann und immer das Ergebnis des automatisierten Parsings enthält, sodass man immer den Originalzustand als Vergleich sieht. Sobald man Änderungen im Parsing vorgenommen hat, werden diese im Feld *original parsing result* gelb hervorgehoben.

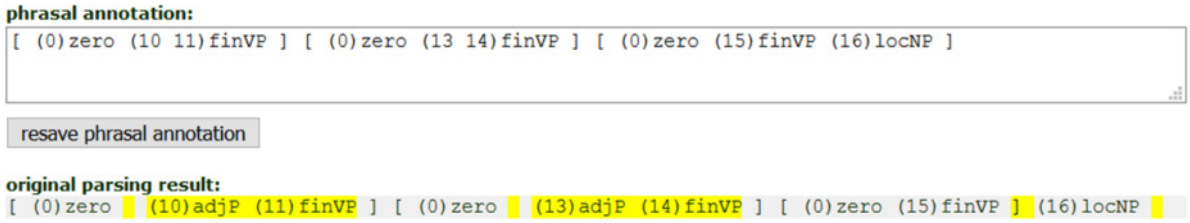

Abbildung 2: Feld *Phrasal Annotation* mit Bearbeitungsmöglichkeit (oben) und Ergebnis automatisches Parsing zum Vergleich (unten, inaktiv); NM (ID 745, Nr. 2)

Die dritte Möglichkeit ist, im Speech Annotation Mode zu arbeiten: dieser funktioniert analog zu den beiden anderen Modi und ist ebenfalls über einen Knopf am Beginn der Seite auswählbar. In diesem Modus erscheint über jedem Satz ein Hinweis, ob es sich hier um direkte Rede handelt. Die Informationen können ebenfalls automatisiert ausgelesen werden (über die jeweilige Information zur Person in der Glossierung) und bei Bedarf manuell korrigiert werden: dazu einfach die Box anklicken, so erscheint ein Häkchen für direkte Rde bzw. erlischt es wieder, wenn nicht gewünscht.

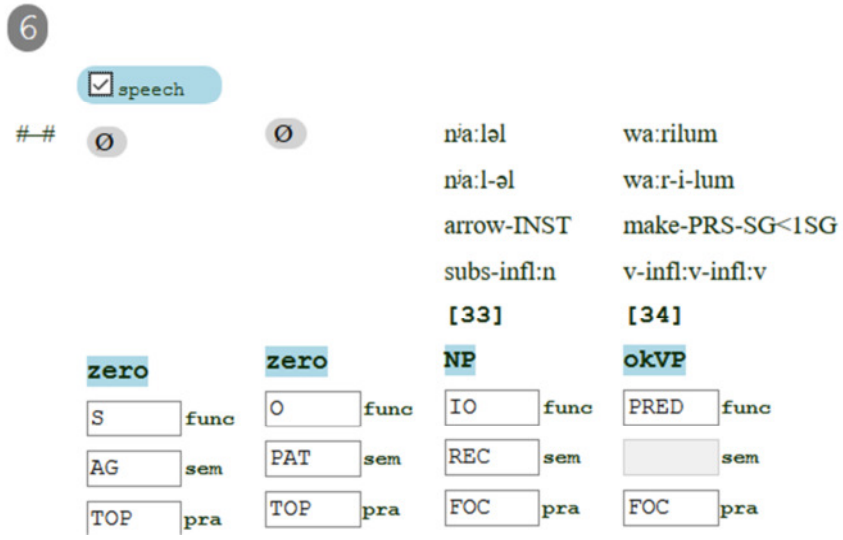

Abbildung 3: Markierung von direkter Rede im Speech Annotation Mode (Häkchen in der Box über dem Satz); NM (ID 1229, Nr. 6)

Parsing-Regeln:

- Nummer der Satzteile müssen in einfachen runden Klammern stehen, die Benennung der Phrase, die sie repräsentieren muss ohne Leerzeichen dahinter stehen: (1)finVP
- Besteht eine Phrase aus mehreren Satzgliedern, müssen diese sich innerhalb derselben Klammer getrennt durch ein Leerzeichen, befinden: (1 2)NP
- Zwischen den einzelnen Phrasen muss ein Leerzeichen stehen: (1 2)NP (3)finVP
- Die Ziffern in der phrasal annotation sollten immer von Klammern umgeben sein. Wenn das Skript eine übersieht, bitte manuell einklammern
- Zeichen für die Nullanapher (Zero): (0)zero
- Besteht eine Zeile aus mehreren Sätzen (mehrere Prädikate), müssen diese mit einer eckigen Klammer um den gesamten Satz voneinander getrennt werden
- Die gesamte Zeile muss von einer eckigen Klammer umgeben sein

Korrekturen können z.B. notwendig werden, wenn:

- die Satzgrenzen falsch gesetzt wurden oder eckige Klammern fehlen;
- nicht alle Bestandteile einer Phrase erkannt wurden;
- Nullanaphern falsch oder in der falschen Anzahl eingefügt wurden;
- Infinitivkonstruktionen (von compC zu infVP und von subC zu ptcpVP oder umgekehrt, siehe Annotationsregeln)

ACHTUNG:

Eine Nullanapher kann jederzeit eingefügt werden, da sie die Ziffer Null (0) hat. Die anderen Satzteile erhalten eine fortlaufende Nummer, die unbedingt in der Reihenfolge beibehalten werden muss, da sich sonst die Wortfolge im Satz verändert! Die Nummerierung schließt auch Satzzeichen mit ein, sodass Lücken in der Nummerierung durchaus vorkommen und ignoriert werden können (endet z.B. ein Satz mit (3) und der Folgesatz beginnt mit (6), so fehlen nicht drei Satzglieder, sondern hier sind drei (Satz)Zeichen (Kommata, Anführungs- oder Bindestriche) mitgezählt worden, die für die Annotation nicht relevant sind). Das Possessivsuffix erhält ebenfalls die Ziffer Null (0).

Stimmt das Ergebnis des Parsings, kann auf den Knopf *save phrasal annotation* geklickt werden. Ein Häkchen, dass das Speichern bestätigt, erscheint unter dem Knopf, gleichzeitig erfolgt ein Reload, der die Satzglieder den Phrasen zuordnet und die Boxen der vier Annotationsebenen diesen hinzufügt.

Abbildung 4: Funktionale, semantische und pragmatische Annotation und Nummerierung der Referenten; NM (ID 745, Nr. 2)

Annotation

Die Boxen der vier Annotationsebenen erscheinen unter der Glossierung. Die Nummern der Satzteile befinden sich über den Boxen zur Orientierung, geparste Phrasen sind mit eckigen Klammern gekennzeichnet. Die Box erscheint dann head-final, d.h. unter dem letzten Bestandteil (in der Regel dem Kopf) der Phrase.

Die meisten Tags erscheinen automatisch in den jeweiligen Boxen. Einige Tags müssen überprüft und ggf. nachkorrigiert werden, einige Boxen müssen von Hand befüllt werden (s.u.). Dazu erscheint eine Auswahl an Tags, wenn man ein Feld anklickt. Sie ist als Vorschlag anzusehen, wobei von oben nach unten das jeweils wahrscheinlichste Tag erscheint, die Auswahl ist aber weder bindend noch auf dieses Set beschränkt. Die Vorschläge sind außerdem thematisch zur besseren Übersicht gruppiert.

--	--	--
S	LOC	1=sister
O	GOAL	2=younger-brother
--	SOURCE	3=village
IO	PATH	4=ahead
--	INST	5=three+man
ATTR	TIME	6=horse
MOD	MANNER	7=village
VOC	CAUSE	8=tea
PAR	CONS	9=river+bank
COLL	QUAL	10=wild-duck
	DEG	12=wild-goose
	--	12a=one+wild-goose
	REC	13=sea
	ADR	14=three+wild-goose+boy
		15=Southland
		16=husband-and-wife
		16a=southland-woman

Abbildung 5: Auswahllisten für funktionale Tags (links), semantische Tags (mitte) und Referenten (rechts)

Der Default Modus ist ‚Regular Annotation', d.h. werden nachträglich Phrasen geändert oder Zeros hinzugefügt, werden die vorhandenen Tags soweit beibehalten wie von den Annoatationsregeln her möglich und die neuen Bestandteile weitestgehend eingegliedert. Sollte dies nicht gewünscht sein, und der Satz komplett neu getaggt werden, ruft man den enstprechenden Modus über den Knopf *‚Reset Annotation' Mode* am oberen linken Rand der Annotationsmaske, unter den Metadaten, auf.

Die Fehlende Tags bzw. Nachkorrekturen können direkt in den Boxen erfolgen. Groß- und Kleinschreibung muss dabei nicht beachtet werden, nach einem Reload der Seite wird automatisch alles in Großbuchstaben konvertiert.

Korrekturen können z.B. notwendig werden bei:

- der semantischen Bestimmung bei mehreren NPn in adverbialer Funktion
- abweichender Reihenfolge von Subjekt und Objekt (der Parser weist der ersten NP/PronP/zero Position die Funktion des Subjekts zu, der zweiten die des direkten Objekts)
- verschiedene syntaktische Funktionen wie Parenthesen
- semantische Rolle des Subjekts im Passivsatz (meist PAT, ab und an aber auch REC); semantische Rolle des direkten Objekts beim Dative-Shift (REC statt PAT)
- Die Annotation der pragmatischen Tags erfolgt soweit es geht automatisch und kann mit dem Hinweis, dass ein Automatisierung hier nur bis zu einem gewissen Grad erfolgen kann, so belassen werden. Bestimmte pragmatische Rollen, die vom Kontext abhängig sind, wie z.B. MFOC oder die Funktion des Possessivsuffixes, können nur von Hand und mit dem entsprechenden theoretischen Hintergrund erfolgen.

Immer nachkorrigiert werden müssen:

- attrP (muss manuell eingefügt werden)
- Die Nummerierung der Referenten. Sobald die ersten Nummern vergeben (und einmal abgespeichert) wurden, erscheinen die Nummern sowie die erste Erwähnung der bisherigen Referenten als Vorschlag, klickt man auf die nächste Box. Dadurch ist auch bei längeren Texten mit mehreren Referenten gewährleistet, dass die Nummerierung der Referenten übersichtlich und fehlerfrei bleibt.

Sind die Boxen zur syntaktischen, semantischen und pragmatischen Annotation befüllt, klickt man auf *save functional annotation*. Die Annotation wird in die Datenbank geschrieben, es erscheint ein weiteres grünes Häkchen.

Hinweise und Tipps

- Man kann mithilfe der Navigationsleiste gezielt zu einem beliebigen Satz navigieren; alternativ kann man der URL am Ende ein # und die Satznummer hinzufügen.
- Grundsätzlich sollten so viele Satzglieder zusammen geparst werden, wie möglich, wenn zu einer Phrase zugehörig. Sind z.B. Satzzeichen dazwischen, empfiehlt es sich, mit mehreren Phrasen zu arbeiten, um keine Verschiebungen in der Annotation zu produzieren.
- Mehrere Satzglieder, die dieselbe syntaktische Funktion ausüben, sollten ebenfalls nach Möglichkeit zusammen geparst werden (z.B. die Abfolge subs subs-DU, bzw. subs subs-COLL)
- Mehrfachnennungen eines Referenten, auch in Abfolge, sollten jedoch einzeln geparst werden, da es einen Grund gibt, weshalb auf ein und denselben Referenten z.B. mit zweierlei Lemmata verweisen wird. Die erste Referenz kann gemäß ihrem Vorkommen im Text getaggt werden (meistens FOC oder RESUME), die nachfolgende dann als REPEAT.
- Elliptische Strukturen, ausgenommen der topikalen Referenten, sollten vermieden werden. D.h Zeros möglichst nur für das Subjekt und ggf. das direkte Objekt, nicht aber für Prädikate o.ä. verwenden. Da letztere Ellipsen oft bei Wiederholungen oder Aufzählungen vorkommen, bietet sich hier an, die Satzteile anhand ihrer syntaktischen und semantischen Funktion zu taggen und das pragmatische Tag MFOC zu benutzen. So werden die Sätze nicht übermäßig zerlegt und mit Leerstellen überfüllt, der Satzzusammenhang bleibt erhalten und die Mehrfachnennung ist gekennzeichnet.
- Texte können, wenn notwendig, neu importiert werden, ohne dass das Parsing und die Annotation verloren gehen oder verrutschen. Voraussetzung hierfür ist a) die Annotationen Satz für Satz nach der Bearbeitung gespeichert zu haben und b) die Anzahl der Elemente in FLEx dürfen sich nicht verändern, d.h. der Sense oder auch die Schreibweise eines Lemmas können verändert werden. Zusammenführen oder Trennen von Morphemen und Wörtern (Präverb und Verb, Possessivsuffix, etc.) resultiert in einer veränderten Anzahl von Elementen im Satz und Parsing und Annotation verrutschen und müssen wiederhergestellt werden, siehe dazu nächster Punkt.
- Der Knopf *Reparse Syntax* (ganz links oben über der Satz-Navigation): ein erneutes Parsing des gesamten Textes (z.B. nach Neu-Import des Textes) erfolgt nur über diesen Knopf und nicht automatisch beim Reload der Seite.
- Bei Unsicherheit, ob ein Tag richtig gewählt ist, dieses mit Fragezeichen in der Box markieren.