IS-Annotation: Beschreibung, Anleitung, Tags

Tags

Im Folgenden werden die Tags der vier Annotationsebenen beschrieben. Alle zu verwendeten Tags werden aufgelistet und beschrieben. Darüber hinaus werden Regeln zum Tagging bestimmter Vorkommen zur Verfügung gestellt.

Phrasen Tags

Die Phrasal Tags werden beim Parsing der Sätze anhand eines Skriptes zugewiesen. Das Skript enthält sämtliche Abfolgen, die zu Phrasen zusammengeführt werden sollen. Mitunter können diese Automatismen vom tatsächlichen Beispielsatz abweichen und müssen dann nachkorrigiert werden. Erkennt der Parser eine Phrase einmal nicht, kann sie manuell anhand der Liste ergänzt werden:

phrase tag	phrase type	
NP	noun phrase	
pronP	pronominal phrase	
zero	zero anaphora	
locNP	noun phrase with local case	
postP	postpositional phrase	
QP	interrogative pronominal phrase	
advP	adverbial phrase	
attrP	attributive phrase	(nur manuell zu taggen!)
adjP	adjective phrase	
NUM	numeral	
finVP	finite verbal phrase	
passVP	finite verbal phrase in passive voice	
okVP	finite verbal phrase with objective conjugation	
ptcpVP	participial (verbal) phrase	teilweise manuell statt subC taggen
infVP	infinite (verbal) phrase	teilweise manuell statt compC taggen
subC	subordinated clause	
compC	complement clause	
CVB	converb	
CONJ	conjunction	
PTCL	particle	
NEG	negation	
px	possessive suffix	

Functional Tags

Anhand der Wortabfolge im Satz werden den einzelnen Phrasen ihre syntaktischen Funktionen zugewiesen; auch hier kann daher eine Nachkorrektur erforderlich sein. Nicht alle Komponenten des Satzes, die geparst werden, müssen jedoch auch annotiert werden, Orientierung bietet die folgende Liste. Es wurde darauf geachtet, das Set so klein und übersichtlich wie möglich zu halten. Die verwendeten Tags sind so gewählt, dass sie sich möglichst selbst erklären und orientieren sich an Dik 1997.

functional tag	description
S	Subject of a sentence
O	Direct Object of a sentence
IO	Indirect Object of a sentence
ADV	Adverbial (Place, Time, etc.)
VOC	Vocative, constituent of a sentence (i.e. a referent) with no other syntactic role other than being the addressee of the utterance
ATTR	Attribute preceding the head of a phrase
PRED	Predicate of a sentence
SUBPRED	Predicate of a subordinate clause
MOD	Modifier of a phrase (cf. possessive suffixes)
COLL	Collocation
AGR	Agreement
PAR	Parenthesis, insertion with no syntactic function
CON	Connector

Semantische Tags

Auch die semantischen Rollen der Satzglieder werden getaggt. Hier wurde das Set ebenfalls so übersichtlich und selbsterklärend gehalten wie möglich. Wir unterteilen daher nicht in Experiencer und Agens oder Patiens und Thema, sondern orientieren uns ebenfalls an Dik 1997. Besondere Relevanz beim Tagging der semantischen Rollen haben dabei die Animatheit des Referenten (sodass wir sowohl COM als auch INST verwenden, je nachdem, ob ein Referent belebt oder unbelebt ist) oder die ausgedrückte Lokalangabe. Diese wird in den ob-ugrischen Sprachen mit einem differenzierten System von Kasussuffixen und Postpositionen ausgedrückt, welches in den semantischen Tags Berücksichtigung findet. Die Besonderheit, im Ob-Ugrischen den Referenzpunkt mithilfe des Possessivsuffixes auszudrücken, wir ebenfalls mit dem Tag REF (nach Mackenzie 1983) eigens markiert.

Tag	Description
AG	Agent
PAT	Patient
REC	Recipient
ADR	Addressee (animate)
COM	Comitative (animate)
LOC	Location
GOAL	Goal
SOURCE	Source
PATH	Path
INST	Instrument (inanimate)
TIME	Time
MANNER	Manner
CAUSE	Cause
CONS	Consecutive
QUAL	Quality
DEG	Degree
REF	Reference-Point
PWH	Part-Whole Relation (for possessive suffixes as marker of collocations)
ADVERS	Adversative
COMP	Comparative
ADD	Additive
DISJ	Disjunctive

Pragmatische Tags

Die Annotation der pragmatischen Rollen kann nur zu einem bestimmten Grad automatisiert erfolgen, benötigt die meiste Nachkorrektur und kann nicht ohne bestimmte Vorkenntnisse erfolgen. Sie wurde daher so vereinfacht wie möglich konzipiert, ohne dabei ihre Berechtigung, überhaupt annotiert zu werden, zu verlieren. Zum theoretischen Hintergrund siehe (gesondertes Dokument IS in OU). Die Tags wurden so transparent wie möglich konzipiert, ihre Verwendung so eindeutig wie möglich festgelegt. Zur Vereinfachung der pragmatischen Rollen zählt u.a. dass wir lediglich mit Fokus und Topik arbeiten, Background und New, die eigentlichen Gegenstücke werden nicht als Tags vergeben. Die neue Information (eigtl. New) wird daher immer als fokussiert und daher als Fokus des Satzes angesehen. In der Konsequenz wird das Prädikat des Satzes automatisch mit FOC getaggt. Analog dazu ist der Background gleichgesetzt mit dem Topic. Das Topic wird als pragmatische Rolle und nicht als global topic und dergleichen verstanden, da für uns die Realisierung und Fortführung der topikalen Referenten im Vordergrund steht. Zusätzlich wird der kontrastierende Fokus berücksichtigt, d.h. die Wahl eines Referenten aus einer Gruppe von mehreren Möglichkeiten. Kontrast wird daher als zwischen Topik und Focus stehend verstanden – die hervorgehobene alternative, jedoch vorerwähnte pragmatische Rolle. Dies trifft generell z.B. auf Personalpronomina zu. Wir verzichten auf Verweise auf den Status einer topikalen Rolle, wie z.B. Diskurs-Topik oder Satz-Topik, auch werden die verschiedenen topikalen Rollen nicht nummeriert. Dies erfolgt alles über das referentielle Tagging, mit dem Vorteil, dass weder eine Wertung der Topikalität einzelner Referenten vorab impliziert wird und auch nicht nur topikale Rollen, sondern das Vorkommen des Referenten insgesamt erfasst werden kann.

Zusätzlich wurden bestimmte Formen der Wiederaufnahme berücksichtigt sowie Wiederholungen bestimmter Satzstrukturen. Vervollständigt wird die Liste mit einigen Tags, die eindeutig bestimmbare Besonderheiten der ob-ugrischen Textstruktur markieren.

TOP	Topic	Satzteil, der das Topik des Satzes (hier den topikalen Referenten) kodiert.	nominal (zero)
FRAME	Frame-Setting	Für gewöhnlich ein einleitender Satz zu Beginn einer Erzählung oder einem Abschnitt mit neuer Handlung; meist Informationen zu Zeit, Ort, beteiligten Personen. Das Frame-Setting muss nicht zwangsläufig im ersten Satz erfolgen, ist jedoch dort typisch. Beginnt ein Text stattdessen gleich mit dem tatsächlichen Handlungsverlauf, kann auch mit FOC getaggt werden. Umgekehrt kann auch zu Beginn eines neuen Abschnitts ein Frame-Setting erfolgen	nominal + verbal
CTR	Contrast	Satzteil, der eine hervorgehobene Alternative darstellt, vorerwähnt, z.B. Personalpronomina; Diese Alternative kann auch das Gegenüber in einem Dialog sein	nominal (pronominal)
MFOC	Mirror Focus	„gespiegelter Fokus“, zwei aufeinanderfolgende Sätze oder Satzteile mit identischem Aufbau, jedoch wird auf einen anderen Referent verwiesen; ZWEI REFERENTEN, GLEICHE HANDLUNG	nominal + verbal
KG	Known Group	Herausgreifen eines Referenten aus einer Gruppe, d.h. der Referent war allein bislang nicht Teil der Handlung, jedoch als Teil einer Gruppe und ist daher nicht vollkommen unerwähnt	nominal
FOC	Focus	Satzteil, das den Fokus (hier die neue Information) beinhaltet	verbal + nominal
THL	Tail Head Linkage	Repetition of an action; either the same sentence is uttered several times or the action is repeated in a participal clause while the main clause introduces a new action and/or participant	verbal (subpred) Konjuntionen
DISC	Discourse Marker	Tag für Adverbial / Partikel, die auf verschiedene Arten dein Diskursverlauf steuern	Konjunktionen
ANCH	Anchoring	Tag für das Possessivsuffix, wenn dieses einem zuvor nicht erwähnten Referenten angefügt ist	PX

Referentielle Tags

Die vierte Annotationsebene beinhaltet das Reference-Tracking. Hierzu wird zunächst jedem Referenten im Text eine Nummer zugewiesen. Jedem Referenten bedeutet, sobald ein Referent mehr als einmal im Text vorkommt, bzw. immer in einer Possessivkonstruktion. Letztere (sog. Possessor UND Possessum!) werden immer gezählt. Eine spätere Auswertung der Vorkommen jedes nummerierten Referenten gibt dann Aufschluss über die Häufigkeit des Auftretens eines jeden Referenten, sowie darüber, welche Referenten topikal sind (vorwiegend animat, in die Handlung über mehrere Sätze involviert) und deren Stellung in der topikalen Hierarchie (Diskurs-Topik, Paragraphen-Topik oder z.B. Satz-Topik).

Ref-Nummern können auf einzelne Referenten oder Gruppen verweisen. Wird ein Referent, der zuvor nur in der Gruppe erwähnt wurde, aus dieser herausgegriffen, so erhält er keine neue Nummer, sondern wird mit der Nummer der Gruppe und einem Kleinbuchstaben versehen. Dasselbe gilt für zwei Referenten aus einer Gruppe. Für den Fall, dass mehrere Referenten aus einer größeren Gruppe herausgegriffen werden (mehr als zwei aus einer Gruppe von mehr als dreien), so wird lediglich die Nummer der Gruppe verwendet.

1	Referent 1
1+2	Referent 1 und Referent 2, Dual
1+2+3	Referent 1 und Referent 2 und Referent 3, Plural
2a	Referent 2 bezieht sich auf eine Gruppe, Referent 2a ist ein in der Gruppe vorerwähnter, jetzt alleine an der Handlung beteiligter Referent
2a+2b	Referent 2 bezieht sich auf eine Gruppe, Referent 2a und 2b sind zwei in der Gruppe vorerwähnte, jetzt zu zweit an der Handlung beteiligte Referenten
1+group	Formen mehrere Referenten im Textverlauf neue Gruppen, kann der Übersichtlichkeit halber der Referent, der den Kern der Gruppe bildet, als Nummer gewählt werden zusammen mit dem Hinweis +group

Achtung:

Unter jeder NP befindet sich eine Box für die Nummerierung des Referenten. Diese wird jedoc hnciht immer befüllt, d.h. nicht jede NP wird als Referent interpretiert und nummeriert (insb. aus Gründen der Übersichtlichkeit). Zunächst werden ausschließlich belebte Referenten getaggt. Sind Gegenstände / Objekte zwar unbelebt, aber in gewisser Hinsicht personalisiert, können sie ebenfalls referentiellen Status haben und werden mit getaggt. Auch solche Gegenstände, die (über einen bestimmten Zeitraum / Abschnitt im Text) mehrfach wiederholt werden. Genrell sollte eher einmal zuviel getaggt werden als zu wenig.

Weitere Annotationsregeln

Attribute

Attribute sind „Beifügungen“ zum Kopf, sie gehen dem Kopf voran und fügen Informationen zu Eigenschaften über diesen hinzu. Attribute sind Adjektive, Demonstrativpronomina, Numeralia, etc. Diese werden automatisch mit dem Substantiv zu einer Nominalphrase geparst. Attribute können auch Substantive sein, diese sind allgemeinhin als Possessoren bekannt. Im Unterschied zu Attributen fügen diese Substantive jedoch keine qualitative Zusatzinformation zum Kopf hinzu, sondern stehen in Relation zu diesem (z.B. in einem Teil-Ganzes Verhältnis). Sie haben außerdem im Unterschied zu anderen Attributen eine eigene Referentialität, die ggf. die des Kopfes beeinflusst, in der Kognitiven Linguistik spricht man daher von Referenzpunkt (dem sog. Possessor) und vom Target, wenn man sich auf den Kopf der Phrase bezieht. Die syntaktische Bezeichnung für den Referenzpunkt ist Modifizierer, um den Unterschied zum nicht-referentiellen Attribut auszudrücken. Modifizierer können bis auf seltene Ausnahmen, in denen der Modifizierer eine Angabe zur Referenz bedarf (siehe dazu PX), mit dem Kopf zu einer Phrase geparst werden. Dasselbe gilt für wenige Fälle in denen die Attribute selbst aus so vielen Bestandteilen bestehen, dass es Sinn macht, diese separat zu parsen. Beides muss von Hand nachkorrigiert werden. Das Phrase Tag attrP muss manuell geändert werden, es bewirkt, dass diese Bestandteile bei der Annotation nicht irrtümlich als kernsyntaktische Rollen getaggt werden. Das syntaktische Tag für das Attribut ist dann ATTR und MOD für den Modifizierer. Semantisch erhält das Attribut das Tag QUAL und der Modifizierer das Tag REF (in diesem Fall beziehend auf den Referenzpunkt, siehe hierzu auch PX). Das pragmatische Tag muss anhand der Information und Vorerwähntheit gewählt werden.

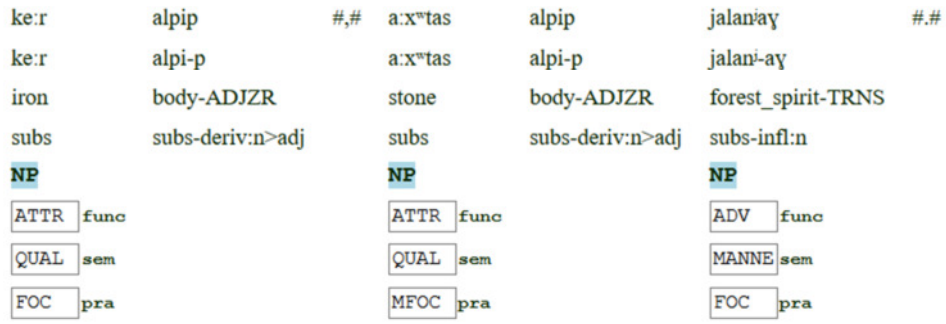

Abbildung 1: Separat geparstes mehrteiliges Attribut; NM (ID 750, Nr. 82)

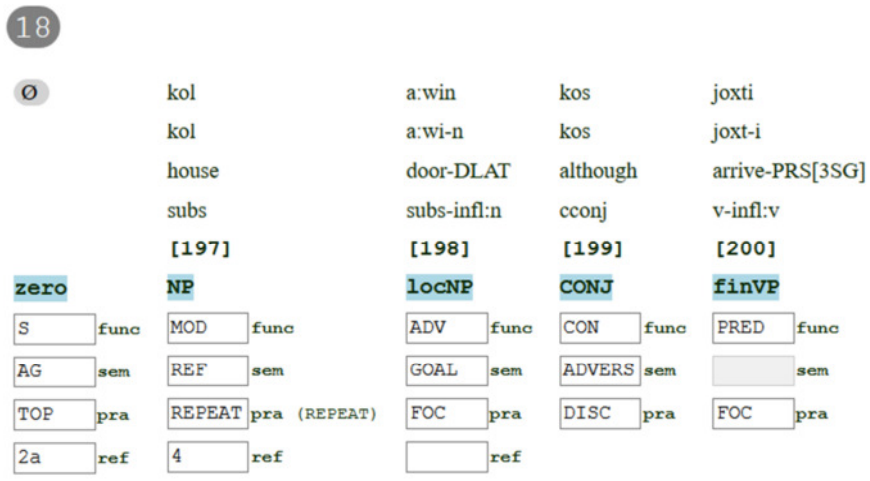

Abbildung 2: Separat geparster Modifizierer, da der Referent (Haus) vorerwähnt ist und über ein referentielles Tag verfügt; NM (ID 750, Nr. 18)

Pronomina

Pronomina sind Pro-Formen, d.h. sie setzen Vorerwähntheit voraus. Hier jedoch gibt es drei Dinge zu beachten:

a) Personalpronomen setzen Vorerwähntheit voraus, werden aber nur in bestimmten pragmatischen Kontexten verwendet, die reguläre Kodierung von vorerwähnten Referenten ist die Leerstelle, d.h. das Agreement am Verb. Die Verwendung des Personalpronomens impliziert eine gewisse Auswahl, bzw. Kontrastierung d.h. „ich von uns beiden bin es und nicht du“. Personalpronomen werden daher automatisch mit CTR getaggt.

Bei emphatischen Personalpronomen muss von Hand das Tag FOC gesetzt werden, sie dienen tatsächlich in hohem Maße der Fokussierung, d.h. Betonung.

b) Demonstrativpronomina haben hinweisende Funktionen und können mit FOC getaggt werden.

c) Indefinitpronomina setzen insofern Vorerwähntheit voraus, als dass ein Referent impliziert ist, es wird jedoch nicht auf einen spezifischen, zugänglichen Referenten verwiesen. Daher muss hier von Hand das Tag FOC gesetzt werden.

Possessivsuffixe (PX)

Possessivkonstruktionen

Auch Modifizierer können mit dem Kopf zu einer Phrase geparst werden, da fast alle Referenten mit topikalem Status, die als Modifizierer fungieren, zusätzlich als Possessivsuffix kodiert werden. Somit kann die Referenz am PX selbst getaggt werden. Bei sog. Possessivkonstruktionen mit Possessivsuffix (d.h. NPn bestehend aus Personalpronomen und Substantiv, Substantiv und Substantiv sowie Substantiv). Aufgrund der linearen Abfolge des Parsings kann es vorkommen, dass das PX im Parsing dann erst zum Schluss einer Phrase und nicht unmittelbar hinter dem Kopf erscheint, z.B. bei Phrasen mit Postpositionen.

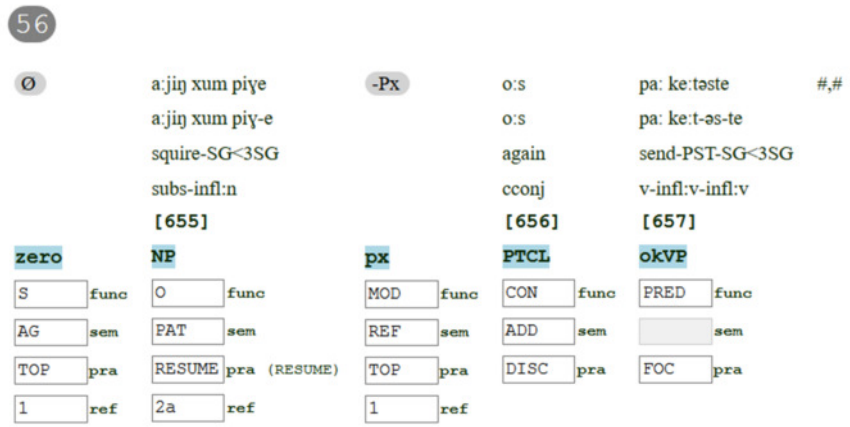

Abbildung 3: Parsing des Possessivsuffixes; NM (ID 750, Nr. 56)

Das syntaktische Tag beim PX ist MOD (analog zur Funktion des darin kodierten Referenten). Die semantische Funktion ist REF, da das PX auf den Modifizierer referiert. Vom kognitiv-linguistischen Standpunkt aus wird der Referenzpunkt kodiert, sodass hier wie beim Modifzierer das Tag auch für Referenzpunkt stehen könnte. REF als semantische Funktion wird hier nach (Mackenzie 1983) benutzt und synonym zum kognitiven Referenzpunkt verstanden. Der pragmatische Status richtet sich neben dem des kodierten Referenten auch nach der Funktion des Possessivsuffixes in der jeweiligen Verwendung der Konstruktion und hängt u.a. von der Vorerwähntheit und dem Kontext ab und bedarf weiterer Analyse für das Ob-Ugrische und wird deshalb in einer vereinfachten Form getaggt. Nur auf ganz spezifische, erkennbare Funktionen wird hingewiesen.

Ist der Referent, der als Kopf der Konstruktion dient, neu und zum ersten Mal im Text erwähnt, so ist die Funktion des Possessivsuffix das sog. Anchoring, d.h. ein neuer Referent wird durch Relation zum Referenzpunkt zugänglicher. Das Tag ist ANCH. Ist das Target vorerwähnt, d.h. findet kein Anchoring statt, wird vorerst ebenfalls das Tag TOP für das Possessivsuffix verwendet.

Kollokationen

Einige Possessivkonstruktionen mit PX ähneln den Izafet-Konstruktionen des Ungarischen, in welchen das Possessivsuffix eher konstruktive als pragmatische Funktionen auszuüben scheint, d.h. das Possessivsuffix referiert nicht auf einen Referenzpunkt, sondern zeigt die Zugehörigkeit des Kopfes zur Konstruktion an. In diesen Konstruktionen wird meist ein Teil-Ganzes Verhältnis ausgedrückt und der Referenzpunkt wird als Substantiv realisiert, was sich im Ob-Ugrischen aufgrund der im Possessivsuffix mit ausgedrückten Topikalität eigentlich ausschließt. Kollokationen sind daher Konstruktionen vom Typ Substantiv Substantiv-PX mit Ausdruck eines Teil-Ganzes Verhältnisses. Das syntaktische Tag ist COLL. Das semantische Tag ist PWH (Part-Whole).

Konstruktionen mit Postpositionen und PX

Ähnlich wie bei der Kollokation scheint das Possessivsuffix in Konstruktionen mit Postpositionen eher eine konstruktive Funktion auszuüben, da das Personalpronomen in der Konstruktion obligatorisch scheint. Somit referiert nicht das Possessivsuffix auf den Referenzpunkt, sondern das Personalpronomen selbst. Das Possessivsuffix fungiert als Agreement-Marker zwischen Kopf und Modifizierer, ähnlich der Izafet-Konstruktion und wird daher nicht abgetrennt. Das funktionale Tag der Postposition ist ADV, das semantische richtet sich nach der in der Postpostion ausgedrückten adverbialen Bestimmung. Postpositionen erhalten ebenfalls eine Box für die Nummerierung der Referenten, hier wird der als Substantiv oder Pronomen zugehörige und ggf. als Possessivsuffix angefügte Referent eingetragen.

Konstruktionen mit Verbalnomina und Possessivsuffix

Je nach topikalem Status und Funktion des Referenten in einer Partizipial- oder Konverbkonstruktion kann dieser mit einem Possessivsuffix kodiert werden. In diesem Fall wird nur ein Referent kodiert, d.h. keine Relation zwischen Kopf und Modifizierer erstellt. Das Possessivsuffix fungiert ähnlich einer Personalendung am finiten Verb und wir daher als AGR getaggt. Das semantische Tag richtet sich nach der semantischen Rolle des Subjekts der im Verb ausgedrückten Handlung, meist AG. Das pragmatische Tag ist TOP, da die Topikalität eine der Voraussetzungen für die Kodierung im Possessivsuffix ist.

Zur Funktion der Partizipial- und Konverbkonstruktion, siehe unten.

In Konstruktionen mit Partizip, Possessivsuffix und Verben der Wahrnehmung hingegen wird das Possessivsuffix nicht abgetrennt. Es handelt sich bei dieser Konstruktion wahrscheinlich um eine Grammatikalisierung, sodass das Possessivsuffix hier obligatorisch erscheint und über keine referentielle Verweiskraft verfügt bzw. das Vorkommen nicht pragmatisch bedingt ist.