Peter Eisenberg / Wolfgang Lezius / George Smith

Die Grammatik des TIGER-Korpus

Das TIGER-Korpus ist ein syntaktisch annotiertes Korpus deutscher Zeitungstexte aus der Frankfurter Rundschau. Die Annotation wird am Institut für Computerlinguistik, Universität des Saarlandes, am Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, und am Institut für Germanistik, Universität Potsdam durchgeführt.[1] Die linguistische Beschreibung basiert auf einem im Projekt NEGRA (*Ne*benläufige *gra*mmatische Verarbeitung) entwickelten Ansatz für die Annotation von Prädikat-Argument-Strukturen (Skut et al., 1997, 1998; Brants et al., 1999). Etiketten (Tagsets) werden auf drei Ebenen verwendet: auf Wortebene, phrasaler Ebene und funktionaler Ebene. Die Darstellung der syntaktischen Struktur erfolgt durch Syntaxgraphen (König und Lezius, 2003). Syntaxgraphen sind Syntaxbäume mit zwei Erweiterungen: Kanten werden mit Etiketten versehen und können sich kreuzen. Kantenlabel werden im TIGER-Korpus dazu verwendet, funktionale Information auszudrücken. Kreuzende Kanten ermöglichen die Repräsentation nichtlokaler Phänomene. Mehrfachabhängigkeiten werden durch sekundäre Kanten dargestellt.

Um die im Korpus kodierte Information der Forschung zugänglich zu machen, wurde die Suchmaschine TIGERSearch (Lezius, 2002; König et al., 2003) entwickelt, mit der ein Benutzer interaktiv Anfragen stellen kann. Dazu wird eine Beschreibungssprache für Syntaxgraphen verwendet, die auf die Bedürfnisse von Linguisten zugeschnitten wurde (König und Lezius, 2003). Eine Anfrage ist eine Beschreibung einer Graphstruktur, die auch unterspezifiziert sein kann. Beschreibungssprache ist also zugleich Anfragesprache. Ergebnis einer Anfrage sind alle Graphen im Korpus, die einen Subgraph enthalten, auf den die Beschreibung passt. Diese Graphen können dann im Fenster der Suchmaschine angesehen werden. Alternativ dazu stehen verschiedene Möglichkeiten des Exports der Daten oder deren weiterer Verarbeitung zur Verfügung.

In einer Anfrage werden zunächst einzelne Knoten anhand von Merkmal-Wert-Paaren beschrieben. Darüber hinaus können zwischen einzelnen Knoten Relationen der Dominanz, der linearen Präzedenz sowie mehrere abgeleitete Relationen durch Operatoren dargestellt werden. Boolsche Operatoren stehen ebenfalls zur Verfügung, um eine logische Verknüpfung einzelner Ausdrücke der Beschreibungssprache zu ermöglichen. Die morphosyntaktische Annotation des Korpus, die ausdrucksstarke Beschreibungssprache sowie die interaktive Verarbeitung von Anfragen durch die Suchmaschine ermöglichen eine gezielte Suche nach grammatischen Strukturen.

---

[1] Die Arbeit wird seit 1999 durch die Deutsche Forschungsgemeinschaft im Rahmen des Projekts TIGER (Linguis*tic* Interpretation of a *Ger*man Corpus) finanziert.

1. Grundlagen der syntaktischen Beschreibung

Abbildung 1 zeigt einen kurzen Beispielsatz, annotiert nach den Richtlinien des TIGER-Korpus, und dargestellt durch die Print-Funktion der Suchmaschine TIGERSearch. Die Struktur besteht aus vier Zeilen sowie einer darüber errichteten Graphstruktur. Ausgangspunkt der Annotation ist die Reihenfolge der Wortformen im Satz, die der Reihe nach in der ersten Zeile stehen. In der zweiten Zeile stehen Tags für Part-of-Speech nach Schiller et al. (1997): ART (Artikel), NN (normales Nomen, also ein Substantiv, das kein Eigenname ist), VAFIN (Verb auxiliar finit) und ADJD (Adjektiv in adverbialer oder prädikativer Verwendung). In der dritten Zeile stehen morphosyntaktische Kategorien, wonach einzelne Wortformen flektiert sind oder die sie regieren. In Abbildung 1 sind das einzelne Kategorien für Genus, Kasus und Numerus beim Substantiv, Person, Numerus, Tempus und Modus beim Verb und Steigerungsgrad beim adjektivischen Prädikatsnomen.

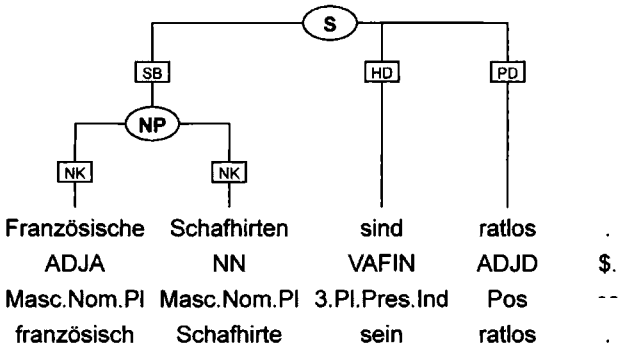

Abbildung 1

Zur Annotation auf Wortebene wird nun eine Graphstruktur hinzugefügt. Die Strukturen sind tendenziell flach, redundante Verzweigungen werden vermieden. So sind sie ohne Informationsverlust besser am Bildschirm zu sehen. Im Graph sind zwei Typen von Etiketten zu finden. Im Knotenlabel steht die syntaktische Kategorie, im Kantenlabel die syntaktische Funktion. So steht in Abbildung 1, dass die Folge von Wortformen *französische Schafhirten* eine Nominalphrase (NP) bildet, die als Subjekt (SB) fungiert.

2. Beschreibung und Anfrage

Vergleichen wir nun den Beispielsatz in Abbildung 1 mit dem Beispielsatz im Abbildung 2, so können wir zugleich die Vorteile sehen, die sich ergeben, wenn eine Beschreibungssprache für Syntaxgraphen als Anfragesprache für eine Suchmaschine verwendet wird.

(1) a. [pos="VAFIN"]
b. [pos="VVFIN"]
c. [pos="ADJD"]
d. [pos="ADJA"]

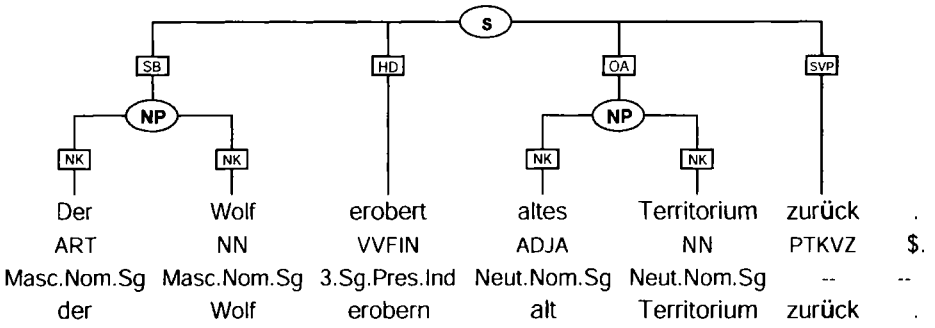

Abbildung 2

Sehen wir uns die Beschreibung einzelner Knoten auf der Wortebene an. Der Satz in Abbildung 1 enthält ein finites Hilfsverb. Der entsprechende Terminalknoten kann (unterspezifiziert) beschrieben werden wie in (1a): ein Knoten mit dem Wert VAFIN für das Merkmal pos (Part-of-Speech). Der Satz in Abbildung 2 dagegen enthält ein finites Vollverb. Die entsprechende Knotenbeschreibung ist in (1b): das Merkmal pos hat hier den Wert VVFIN (Verb voll finit). Wer sich für Strukturen interessiert, in denen Hilfsverben vorkommen, kann eine Knotenbeschreibung wie (1a) als Teil einer Anfrage verwenden. Wer sich dagegen für Strukturen interessiert, in denen Vollverben vorkommen, kann eine Knotenbeschreibung wie (1b) verwenden. Die Adjektive beider Sätze lassen sich auf ähnliche Weise unterscheiden. Das prädikativ verwendete Adjektiv in der Kurzform in Abbildung 1 wird beschrieben wie in (1c), das attributiv verwendete flektierte Adjektiv in Abbildung 2 wie in (1d).

(2) a. [word="Schafhirten" & pos="NN" &
morph="Masc.Nom.Pl" & lemma="Schafhirte"]
b. [word="sind" & pos="VAFIN" &
morph="3.Pl.Pres.Ind" & lemma="sein"]

Eine vollständige Beschreibung des ersten Substantivs in Abbildung 1 ist in (2a) gegeben, die Beschreibung des Verbs desselben Satzes in (2b). Nicht nur Knoten, sondern auch ganze Graphen lassen sich vollständig spezifizieren.

(3) a. [pos="NN" & morph="Masc.Nom.Sg"]
b. [lemma="sein"]

In der Regel aber interessiert sich der Sprachwissenschaftler für bestimmte Aspekte einer gegebenen Struktur. Wer sich für Substantive in Maskulin, Nominativ, Singular interessiert, wird die unterspezifizierte Beschreibung in (3a) als Anfrage verwenden. Wer sich speziell für das Verb *sein* interessiert, verwendet die ebenfalls unterspezifizierte Anfrage in (3b). Es gibt auch verschiedene Wege, auf Mengen von Merkmalswerten zuzugreifen, die in König et al. (2003) näher beschrieben werden.

(4) a. [cat="NP"] > [pos="ADJA"]
b. [cat="NP"] > [pos="ART"]

Wenden wir uns nun der Beschreibung von Dominanzrelationen zu. Die Beschreibung in (4a) trifft auf alle Knoten zu, die den Wert NP für das Merkmal cat (*category*) haben und ein attributiv verwendetes Adjektiv unmittelbar dominieren. Der Operator > drückt die Relation unmittelbarer Dominanz zwischen zwei Knoten aus. Diese Beschreibung, als Anfrage verwendet, würde die Nominalphrase des Satzes in Abbildung 1, sowie auch die zweite Nominalphrase des Satzes in Abbildung 2 finden. Die Beschreibung in (4b) würde dagegen keine dieser Nominalphrasen finden, sondern die erste Nominalphrase in Abbildung 2.

(5) a. [ ] >SB [ ]
b. [ ] >OA [ ]
c. #x:[cat="S"] >SB [ ] &
#x >OA [ ]

Die Relation unmittelbarer Dominanz kann auch näher spezifiziert werden, indem ein Kantenlabel zusammen mit dem Operator verwendet wird. Die Beschreibung in (5a) passt auf alle Knoten, die einen Knoten unmittelbar dominieren, der als Subjekt (SB) fungiert. Die jeweilige Knotenbeschreibungen sind in diesen Beispielen maximal unterspezifiziert. Die Beschreibung in (5b) unterscheidet sich nur durch das verwendete Kantenlabel OA (Objekt Akkusativ). Die Beschreibung in (5c) verwendet eine Variable (#x), die in der ersten Zeile auf Satz-Knoten eingeschränkt wird (Wert S für das Merkmal cat). Diese Beschreibung passt auf alle Satz-Knoten, die sowohl einen Knoten mit der Relation SB als auch einen Knoten mit der Relation OA unmittelbar dominieren, wie z.B. der Satz in Abbildung 2 (siehe König et al. 2003 für weitere Operatoren für Dominanzrelationen und weitere, abgeleitete Relationen, wie z.B. die Geschwisterrelation).

(6) a. #x >SB #sb &
#sb > [pos="ART"] &
#sb > [pos="NN"]
b. #x >OA #oa &
#oa > [pos="ADJA"] &
#oa > [pos="NN"]
c. #x >SB #sb &
#sb > [pos="ART"] &
#sb > [pos="NN"] &
#x >OA #oa &
#oa > [pos="ADJA"] &
#oa > [pos="NN"]

Die Beschreibung in (6a) passt auf alle Knoten, die ein Subjekt unmittelbar dominieren, das seinerseits sowohl ein Substantiv als auch einen Artikel dominiert. Als Anfrage verwendet, würde sie den Satz in Abbildung 2 finden. Die Beschreibung in (6b) würde den Satz ebenfalls finden, aber aufgrund anderer Eigenschaften. Beide Beschreibungen können zu

einer weiteren wie in (6c) zusammengefügt werden, die aber dann noch immer unterspezifiziert ist.

(7) a. [pos="ART"] . [pos="NN"]
b. #x >SB #sb &
#x >HD #fin:[pos="VVFIN"] &
#sb . #fin

Auch Relationen der linearen Präzedenz können zum Ausdruck gebracht werden. In (7a) werden alle Kontexte beschrieben, in denen ein Artikel einem Substantiv unmittelbar vorausgeht, wie z.B. die ersten zwei Wortformen des Satzes in Abbildung 2. Die Beschreibung in (7b) ist etwas komplexer. Hier werden in den ersten zwei Zeilen Subjekt und finites Vollverb durch Dominanzrelationen festgelegt. Dann wird in der dritten Zeile durch den Punkt als Operator für unmittelbare Präzedenz spezifiziert, dass das Subjekt dem finiten Vollverb unmittelbar vorausgeht. Diese Beschreibung, als Anfrage verwendet, findet den Satz in Abbildung 2.

3. Granularität der Analyse

Ein Hauptmerkmal der Grammatik des TIGER-Korpus ist die feine Granularität der Beschreibung auf kategorialer wie relationaler Ebene. Dies wird nun exemplarisch behandelt anhand von Appellativen und Eigennamen sowie einiger Funktionen von Präpositionalphrasen. Ausgangspunkt für die Beschreibung auf relationaler Ebene ist eine Differenzierung der syntaktischen Funktionen Komplement und Adjunkt. Neben Subjekt, Akkusativ- und Dativobjekt werden auch präpositionale Komplemente explizit gekennzeichnet, wodurch sie von präpositionalen Adjunkten zu unterscheiden sind.

(8) a. Der Wolf wartet auf das Schaf
b. Der Wolf wartet auf den Studenten
c. Der Wolf wartet im Wald
d. Der Wolf wartet im Hörsal

In den Sätzen (8a,b) sind die Präpositionalphrasen jeweils Komplemente des Verbs. Im TIGER-Korpus werden solche Präpositionalphrasen als Präpositionalobjekte (OP) kodiert. In den Sätzen (8c,d) sind sie jeweils Adjunkte. Solche Präpositionalphrasen werden im TIGER-Korpus als Modifikatoren kodiert (MO).

(9) a. #s:[cat="S"] >OP [cat "PP"]
b. #s:[cat="S"] >MO [cat "PP"]

Die Sätze (8a,b) lassen sich durch die Anfrage (9a) finden, die Sätze (8c,d) durch die Anfrage (9b).

(10) a. #s:[cat="S"] >SB #np:[cat="NP"] &
#np > [pos="ART"] &
#np > [pos="NN"]
b. #s:[cat="S"] >SB #np:[cat="NP"] &
#np > [pos="ART"] &
#np > [pos="NE"]

Auch auf der Ebene der Wortklassen kann hier differenziert gesucht werden. Der Unterschied zwischen den Tags NN (normales Nomen) und NE (Nomen eigen) ermöglicht die gezielte Suche nach Eigennamen. Die Anfrage (10a) findet die Sätze (8a,c), während die Anfrage (10b) die Sätze (8b,d) findet. Hier kann also gezielt nach Kontexten gesucht werden, in denen Eigennamen mit Artikel verwendet werden.

(11) a. In Frankreich steht die Zukunft des Wolfes zur Debatte
b. Wolfs Vorschlag steht in der Sitzung zur Diskussion

(12) #s:[cat="S"] >CVC #pp:[cat="PP"]

Das Funktionsverbgefüge wird erfasst durch das Kantenlabel CVC (*collocational verb construction*). Die Präpositionalphrasen in (11), die Teil eines Funktionsverbgefüges sind, werden durch die Anfrage in (12) erfasst, die übrigen durch die Anfrage in (9b).

(13) a. #np:[cat="NP"] >AG #ag:[cat="NP"] &
#np >NK #nn:[pos="NN"] &
#nn . #ag
b. #np:[cat="NP"] >AG #ag:[pos="NE"] &
#np >NK #nn:[pos="NN"] &
#ag . #nn

Der Unterschied zwischen (11a) mit einem postnominalen Genitivattribut in der Form einer Nominalphrase und (11b) mit einem pränominalen Eigennamen im Genitiv kann durch die Beschreibungen in (13a) bzw. (13b) dargestellt werden, die dann auch als Anfragen Verwendung finden können, um etwa das Verhältnis von prä- und postnominalem Genitiv zu untersuchen (Eisenberg/Smith 2002; Smith 2003).

4. Zusammenfassung

Das Modell der grammatischen Beschreibung, das bei der Erstellung des TIGER-Korpus entwickelt wurde, nützt die Möglichkeiten aus, die dadurch entstehen, dass eine Korpusbeschreibungssprache für Syntaxgraphen zugleich als Anfragesprache fungiert. Das Inventar an Etiketten sowie die Möglichkeiten der Differenzierung, die eine Graphstruktur bietet, werden in der grammatischen Beschreibung der Sätze des Korpus gezielt so eingesetzt, dass

Daten für möglichst viele für die Germanistische Sprachwissenschaft interessante Fragestellungen erhoben werden können.

Literatur

Brants, Thorsten / Skut, Wojciech / Uszkoreit, Hans (1999): Syntactic Annotation of a German Newspaper Corpus. In: Proceedings of the ATALA Treebank Workshop. Paris, S. 69-76.

Eisenberg, Peter / Smith, George (2002): Der einfache Genitiv. Eigennamen als Attribut. In: Corinna Peschel (Hg.): Grammatik und Grammatikvermittlung. Frankfurt a.M., S. 113-126.

König, Esther / Lezius, Wolfgang (2003): The TIGER language. A Description Language for Syntax Graphs, Formal Definition. Ms. Universität Stuttgart.

König, Esther / Lezius, Wolfgang/Voormann, Holger (2003): TIGERSearch User"s Manual. Ms. Universität Stuttgart.

Lezius, Wolfgang (2002): Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. Stuttgart (= Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS) 8.4).

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. Ms. Universität Stuttgart und Universität Tübingen.

Skut, Wojciech / Brants, Thorsten/Krenn, Brigitte/Uszkoreit, Hans (1998): A Linguistically Interpreted Corpus of German Newspaper Text. In: Proceedings of the Conference on Language Resources and Evaluation LREC-98. Granada, S. 705-711.

Skut, Wojciech / Krenn, Brigitte / Brants, Thorsten / Uszkoreit, Hans (1997): An Annotation Scheme for Free Word Order Languages. In: Proceedings of ANLP-97. Washington, D.C., S. 27-28.

Smith, George (2003): On the Distribution of the Genitive Attribute and its Prepositional Counterpart in Modern Standard German. In: University of Pennsylvania Working Papers in Linguistics 8.1.