

# **Expose zur Seminararbeit**

## **Bestimmung semantischer Rollen auf einem deutschen Korpus**

Institut für Informatik & Institut für Linguistik  
Humboldt-Universität zu Berlin

Robert Bärhold (Humboldt-Universität zu Berlin)

Arne Binder (Humboldt-Universität zu Berlin)

Enrique Manjavacas (Freie Universität zu Berlin)

20. Dezember 2013

Seminar: Computergestützte Analyse von Sprache

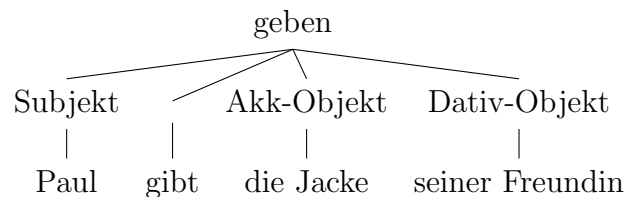
Seminarleiter: Prof. Dr. Ulf Leser

Prof. Dr. Anke Lüdeling

# 1 Einleitung

## 1.1 Semantische Rollen

Prädikate zeichnen sich dadurch aus, dass sie die Komplementationsstruktur einer übergeordneten syntaktischen Einheit (nämlich des Satzes) bestimmen. Zum Beispiel verlangt ein ditransitives Verb wie „geben“ drei weitere Satzkonstituenten mit konkreten syntaktischen Eigenschaften:



Aber das Bedingtheitsverhältnis zwischen dem Verb und seinen Komplementen endet nicht in der Syntax. Diese sind durch das Prädikat auch hinsichtlich ihrer *semantischen* Integrierung in den Satz determiniert. Dem Komplement „Paul“ wird einmal die syntaktische Funktion des Subjekts zugewiesen, es wird aber auch durch die konkrete Semantik der Handlung „geben“ zunächst als „Geber“ charakterisiert. Hierbei spricht man von den Argumenten eines Verbs, und man sagt, dass sie *semantische Rollen* erfüllen.

## 1.2 Frame Semantik

Die traditionelle Semantik war versucht, eine möglichst abstrakte und allgemeine, sprachenübergreifende Formulierung von semantischen Rollen zu formulieren - bei der man z. B. nicht von „Geber“ sondern allgemein von „Agens“ sprechen würde. Demgegenüber steht die von C.J. Fillmore ab den 70er Jahren entwickelte Theorie der Frame-Semantik (FS) Fillmore (1985). Das Fundament der Theorie ist mit dem traditionellen Ansatz in höchstem Maße inkompatibel, für unser Vorhaben reicht es aber, die partikuläre Auffassung von semantischen Rollen zu erwähnen, die der frame-semantische Ansatz mit sich bringt.

Rollen bzw. Frame-Elements (FE) in Termini der FS werden nicht in Bezug auf die Argumentstruktur von Verben formuliert, sondern hinsichtlich der globalen Situation (oder des *Frames*), die durch das Prädikat evoziert wird. Aufgrund des situationellen Charakters der Frames sind nicht nur Verben sondern auch Nomina (Frau, Auge, Montag) oder Adjektive (z.B stolz) potentielle Frame-einführende Elemente, und in diesem Sinne auch Prädikate.

## 1.3 Semantic Role Labeling

Mit Semantic Role Labeling (SRL) ist die automatisierte Erkennung und Annotierung von semantischen Rollen innerhalb eines Satzes gemeint. Die primäre Aufgabe von SRL ist die

genaue Identifizierung der semantischen Beziehung zwischen einem Prädikat und seinen assoziierten Elementen und Eigenschaften, siehe Màrquez u. a. (2008) für eine allgemeine Darstellung. In Bezug auf eine FS lässt sich folgendes feststellen. Da FE Frame-spezifisch sind, stellen sie eine Zwischenstufe in der semantischen Abstraktion zwischen der rein lexikalischen Bedeutung und den traditionellen Verbenübergreifenden semantischen Rollen dar. Die dadurch erreichte „mildere“ Stufe der Abstraktion über die Rollen der verschiedenen Verben, Nomina und Adjektive eignet sich besonders gut für die Lösung unterschiedlicher Aufgabenstellungen in Bereichen des Sprachverstehens wie der Informationsextraktion oder Dialogsystemen Gildea u. Jurafsky (2002).

Die Aufgabenstellung eines FS-orientierten SRL-Systems kann man nun wie folgt untergliedern:

1. zunächst wird der Frame, der durch den Satz realisiert wird, bestimmt.
2. es folgt die Bestimmung der Frame-Elemente bezogen auf den Frame.
3. Zum Schluss erfolgt die Klassifizierung der Frame-Elemente.

## 2 Hypothese

Im Zug unserer Arbeit wollen wir der Frage nachgehen, inwieweit die syntaktischen und lexikalischen Informationen der Satzglieder Aufschluss über die durch die Komplemente erfüllten semantischen Rollen geben. Anhand der erzielten Klassifizierungsergebnisse sowie des entwickelten Klassifikationssystems erwarten wir Aussagen darüber machen zu können, wie die statistische Verteilung von Rollen innerhalb des Satzes allgemein sowie deren Korrelation mit den betrachteten linguistischen Informationen aussieht.

Die Konkretisierung der Hypothese erfolgt sobald alle Einschränkungen bezüglich offener Fragen getroffen wurden.

### 2.1 Korpus

Die Grundlage dieser Arbeit stellt das SALSA-Korpus von Rehbein u. a. (2012) dar, ein Frame-semantisch annotiertes Korpus der deutschen Sprache. Dieses basiert auf dem TIGER-Corpus von Brants u. a. (2004), eine Treebank über deutsche Zeitungsartikel. Das SALSA-Corpus umfasst in seiner zweiten Version 648 Prädikate mit insgesamt 36251 Sätzen. Pro Satz können hierbei mehrere Frames annotiert worden sein. Zur Verfügung gestellt wird das Korpus in einem XML-Format.

Jeder Satz wird dabei getrennt in seine atomaren Konstituenten, die einzelnen Wörter, und die komplexen Konstituenten des über dem Satz erstellten Konstituentenbaums. Jede atomare Konstituente innerhalb eines Satzes ist mit verschiedenen Informationen annotiert, wie beispielsweise dem POS-Tag, dem Kasus, dem Lemma sowie dem Numerus und dem Tempus. Den komplexen Konstituenten ist die Phrasenkategorie sowie den Kanten die grammatikalische Funktion der Tochterkonstituenten zugeordnet.

Das Korpus enthält dabei auch sogenannte „Zweitkanten“ („Sec-Edges“), welche innerhalb des Baums auf entferntere Knoten verweisen, zu denen sie ebenfalls eine grammatikalische Beziehung besitzen. Somit handelt es sich genau genommen eher um einen Graphen, als um einen Baum.

## 3 Realisierung der Anwendung

Im Folgenden soll kurz beschrieben werden, welche Komponenten bei der Realisierung der Anwendung eine Rolle spielen sowie das allgemeine Vorgehen.

### 3.1 Vorgehen

Bei der Herangehensweise wurde der Ansatz von Gildea u. Jurafsky (2002) verfolgt, bei dem versucht wird, den verschiedenen Konstituenten eines Satzes semantische Rollen zuzuweisen. Jurafsky nimmt hierbei den ersten Schritt des allgemeinen Problems als gegeben an (Bestimmung des Frames) und konzentriert sich auf die Klassifizierung der einzelnen Frame-Elemente. Aus Gründen der Komplexitätsreduzierung, haben wir uns dazu entschieden, Frames komplett zu ignorieren und die Frame-Elemente davon unabhängig zu klassifizieren.

Ungeklärt ist bisher die Frage, ob wir bereits im Vorfeld wissen, welches die Frame-Elemente innerhalb eines Satzes sind oder ob jede Konstituente ein potentiell Frame-Element darstellt. Bezogen auf die letztere Variante könnte ein binärer, vorgeschalteter Klassifikator entscheiden, ob es sich bei einer gegebenen Konstituente um ein Frame-Element handelt. Weiterhin unbeantwortet ist die Frage, ob für jedes Prädikat ein eigener Klassifikator erstellt werden sollte oder ob ein globaler Klassifikator über alle Prädikate ebenfalls funktionieren würde.

Da viele Prädikate und Frames des SALSA-Korpus nur mit sehr wenigen Beispielsätzen zur Verfügung stehen, wird für das aktuelle Vorgehen nur ein Auszug des Korpus genutzt. Hierbei ist jedoch noch die Frage offen, ob - in Bezug auf die Anzahl an Beispielsätzen - die Top-20 Prädikate oder die Top-20 Rollen die Korpus-Grundlage bilden.

Die daraus extrahierten Features dienen als Grundlage eines *Naïve Bayes*-Klassifikators. Da hierbei auch verschiedene Feature-Kombinationen gespeichert werden, auch als „Back-Off“ bezeichnet, ist der Klassifikator nicht ausschließlich „naiv“.

### 3.2 Features

Die Auswahl der genutzten Features hat sich ebenfalls an der Arbeit von Gildea u. Jurafsky (2002) orientiert. Für jede atomare sowie komplexe Konstituente werden sowohl lexikalische, als auch syntaktische Features genutzt.

Als lexikalisches Feature nutzen wir das *Kopfelement*, die atomare Konstituente, die den wichtigsten syntaktisch determinierenden Beitrag innerhalb einer komplexen Konstituente leistet. Außerdem leisten Kopfelemente den zur Problemlösung relevantesten semantischen

Beitrag. Hierfür wird die lemmatisierte Form des Wortes genutzt. Für atomare Konstituenten wird vollständigkeitshalber die Konstituente selber als Kopfelement gesetzt.

Anschließend muss das *Target* bestimmt werden. Mit Target ist das Satzglied gemeint, das für die Evozierung des Frames verantwortlich ist. Als Feature dient hierbei das Lemma des Kopfelements der Target-Konstituente.

Als weiteres Feature wird die *syntaktische Kategorie* der Konstituenten ausgewertet. Für atomare Konstituenten entspricht die syntaktische Kategorie dem POS-Tag; bei komplexen Konstituenten wird deren Phrasenkategorie genutzt.

Neben den bisher genannten Features, wird außerdem der *Pfad* zwischen der aktuell betrachteten Konstituente und der Target-Konstituente innerhalb des Konstituentenbaums extrahiert. Er setzt sich aus den verschiedenen Phrasenkategorien der zwischenliegenden komplexen Konstituenten zusammen. Zusätzlich werden die einzelnen Kategorien mit einem Richtungsmarker (absteigend oder aufsteigend im Baum) verknüpft.

Als weiteres, syntaktisches Features wird die *Position* der aktuell betrachteten Konstituente in Bezug zum Target betrachtet. Dabei wird das Kopfelement der Konstituente betrachtet. Dieser kann vor einem atomaren Target (0), vor einem komplexen Target (1), innerhalb eines komplexen Targets (2) und nach einem Target (3) stehen.

Neben den einzelnen Features werden zusätzlich noch Feature-Kombinationen genutzt. Momentan wird exemplarisch die Kombination aus Pfad und syntaktischer Kategorie verwendet. Zukünftig sollten auch die weiteren Features - sofern sinnvoll - miteinander kombiniert werden.

Weiterhin wären folgende Features potentiell interessant:

1. der Kasus der Kopfelementen der zu klassifizierenden Konstituenten sowie der Kasus des Targets,
2. N-Gramme über den Pfad zum Root des Konstituentenbaums,
3. Passiv der übergeordneten Verbalphrase (ja/nein),
4. die Kanteninformationen (z.B. die Kante vom ersten absteigenden Ast innerhalb des Pfades),
5. ist die Konstituente atomar (ja/nein).

Außerdem ist sowohl die Abstraktion der Phrasenkategorien und der POS-Tags möglich sowie die Zusammenfassung verschiedener Pfad-Elemente (zum Beispiel Tilgung von Konjunktionen: NP+CVP-VP-... -> NP+VP-...).

### 3.3 Ablauf der Anwendung

Zuerst wird jeder eingelesene Satz vorverarbeitet und mit zusätzlichen Informationen angereicht. Hierunter fällt bspw. die Bestimmung aller möglichen Pfade zum Wurzelement sowie die Bestimmung des Kopfelements pro Konstituente. Da im SALSA-Korpus nicht alle

komplexen Konstituenten unbedingt ein Kopfelement enthalten, wird regelbasiert Baum-abwärts nach einem Kopfelement gesucht. Dabei werden die verschiedenen grammatikalischen Bezeichner der Kanten sowie die Phrasenkategorien ausgewertet und ein gefundenes Kopfelement bis zum Ursprung der Suche Baum-aufwärts propagiert und dabei für alle Konstituenten auf diesem Pfad gesetzt. Köpfe sind somit immer atomar.

Anschließend werden die Target-Konstituenten des Satzes bestimmt. Im Salsa-Korpus kann dies entweder einer atomaren oder komplexen Konstituente entsprechen, es können aber auch mehrere einzelne Konstituenten als Target annotiert sein („Er *schlug* den Kopf *ab*“). Demzufolge wird beim trainieren die komplexe Konstituente bestimmt, die alle atomaren Konstituenten unmittelbar überdeckt. Beim annotieren wird über alle potentiellen Target-Konstituenten iteriert und jeweils jede Konstituente klassifiziert, da ein Satz mehrere Frames und somit verschiedene Rollen (je nach Target) enthalten kann. Diese potentiellen Target-Konstituenten entsprechen den Konstituenten, deren Kopfelement eines der im Modell gelisteten Target-Lemmata entspricht.

Nach der Bestimmung der Target-Konstituente, werden für alle Konstituenten die Features extrahiert.

Beim Trainieren werden die einzelnen Features und verschiedene Feature-Kombinationen sowohl im Bezug zu den im Satz gegebenen Rollen als auch alleinstehend gezählt. Nach der Verarbeitung aller Sätze werden die absoluten Feature-Häufigkeiten pro Rolle bezogen auf die Gesamt-Häufigkeit des Features normiert. Dieses Modell wird anschließend abgespeichert. Dadurch kann es später sowohl statistisch ausgewertet als auch zum Annotieren wiederverwendet werden.

Beim Annotieren wird ein zuvor erstelltes Modell eingelesen und auf die verschiedenen vorverarbeiteten Sätze angewendet, das heißt, es werden alle Konstituenten in Bezug auf die verschiedenen Target-Konstituenten klassifiziert. Dabei wird jede auftretende Feature(-Kombinationen) im Modell nachgeschlagen. Falls sie nicht existiert, wird für eine Kombination versucht, sie in ihre einzelnen Features zu zerlegen und das Produkt der Einzelwahrscheinlichkeiten genommen. Existiert ein einzelnes Features nicht, so wird eine Glättung („Smoothing“) mit einem vordefinierten Wert (0.000001) angewendet. Das Produkt der relativen Häufigkeiten ergibt dann die Wahrscheinlichkeit, dass die aktuelle Rolle der untersuchten Konstituente zugewiesen werden kann. Alle Wahrscheinlichkeiten werden logarithmiert verarbeitet. Die Rolle mit der höchsten Wahrscheinlichkeit wird anschließend der Konstituente zugewiesen.

Das Ergebnis der Annotation wird ebenfalls in eine Datei geschrieben, um für eine spätere Auswertung zur Verfügung zu stehen.

### 3.4 Auswertung

Die Auswertung der Annotation erfolgt über eine Kreuzvalidierung. Angestrebtes Ziel ist eine Aufteilung 90:10. Verglichen werden sollen hierbei nur die Konstituenten, welche im Ausgangskorpus Rollen besitzen. Hierbei ist zu überlegen, ob die Auswertung einer Teilannotation (im Falle komplexer Konstituenten) positiv gewertet wird oder ob strikt nach dem Prinzip „wahr/falsch“ vorgegangen werden sollte.

## Literatur

- [Brants u. a. 2004] BRANTS, Sabine ; DIPPER, Stefanie ; EISENBERG, Peter ; HANSEN-SCHIRRA, Silvia ; KÖNIG, Esther ; LEZIUS, Wolfgang ; ROHRER, Christian ; SMITH, George ; USZKOREIT, Hans: TIGER: Linguistic Interpretation of a German Corpus. In: *Research on Language and Computation* 2 (2004), Nr. 4, S. 597–620. – ISSN 1570–7075
- [Fillmore 1985] FILLMORE, Charles J.: Frames and the semantics of understanding. In: *Quaderni di semantica* 6 (1985), Nr. 2, S. 222–254
- [Gildea u. Jurafsky 2002] GILDEA, Daniel ; JURAFSKY, Daniel: Automatic Labeling of Semantic Roles. In: *Comput. Linguist.* 28 (2002), September, Nr. 3, S. 245–288. – ISSN 0891–2017
- [Màrquez u. a. 2008] MÀRQUEZ, Lluís ; CARRERAS, Xavier ; LITKOWSKI, Kenneth C. ; STEVENSON, Suzanne: Semantic role labeling: an introduction to the special issue. In: *Computational linguistics* 34 (2008), Nr. 2, S. 145–159
- [Rehbein u. a. 2012] REHBEIN, Ines ; RUPPENHOFER, Josef ; PINKAL, Caroline Sporleder M.: Adding nominal spice to SALSA–frame-semantic annotation of German nouns and verbs. In: *Proceedings of the 11th Conference on Natural Language Processing (KONVENS’12)*, 2012, 89–97