

Disambiguierung mit Wortklassenagenten¹

Hermann Helbig, Andreas Mertens, Marion Schulz

FernUniversität Hagen
Praktische Informatik VII/Artificial Intelligence
D-58084 Hagen, Germany

E-mail:

hermann.helbig@fernuni-hagen.de
andreas.mertens@fernuni-hagen.de
marion.schulz@fernuni-hagen.de

Phone/Fax:

+49 2331 987 374/5

Zusammenfassung

Natürlichsprachliche Ausdrücke können aus verschiedenen Gründen mehrdeutig sein. Deshalb müssen bei der Verarbeitung natürlicher Sprache entsprechende Lexikoninformationen und geeignete Disambiguierungsmethoden verfügbar sein. Die in unserem Beitrag behandelten Disambiguierungsprinzipien werden in der wortagentenorientierten Analyse des Literaturrecherchesystems LINAS an der FernUniversität Hagen eingesetzt. Von den Mehrdeutigkeiten, die typischerweise in praktisch angewandten Systemen auftreten, konzentrieren wir uns in diesem Beitrag auf solche, die mit Hilfe des Lexikons aufgelöst werden können. Im Mittelpunkt steht insbesondere die Verwendung von lexikalischen Informationen bei der Disambiguierung struktureller Mehrdeutigkeiten.

¹Eine gekürzte Version dieses Berichts ist erschienen in: Harald Trost (Hrsg.): KONVENS '94 – Verarbeitung natürlicher Sprache, Wien, 28.–30. September 1994. Berlin: Springer-Verlag 1994, S. 151–160

Disambiguierung mit Wortklassenagenten

Hermann Helbig, Andreas Mertens, Marion Schulz

Abstract

Natural language expressions are ambiguous for many reasons. Therefore, natural language processing must be based on adequate lexical information and suitable strategies in order to support the disambiguation processes. The disambiguation principles underlying our contribution are implemented in the word-agent oriented analysis of the bibliographic information system LINAS at the Open University of Hagen. From the range of the disambiguation problems that typically arise in practical systems, we concentrate here on the lexically resolvable ones. More specifically, we emphasize the role of lexical information in resolving structural ambiguities.

Zusammenfassung

Natürlichsprachliche Ausdrücke können aus verschiedenen Gründen mehrdeutig sein. Deshalb müssen bei der Verarbeitung natürlicher Sprache entsprechende Lexikoninformationen und geeignete Disambiguierungsmethoden verfügbar sein. Die in unserem Beitrag behandelten Disambiguierungsprinzipien werden in der wortagentenorientierten Analyse des Literaturrechersystems LINAS an der FernUniversität Hagen eingesetzt. Von den Mehrdeutigkeiten, die typischerweise in praktisch angewandten Systemen auftreten, konzentrieren wir uns in diesem Beitrag auf solche, die mit Hilfe des Lexikons aufgelöst werden können. Im Mittelpunkt steht insbesondere die Verwendung von lexikalischen Informationen bei der Disambiguierung struktureller Mehrdeutigkeiten.

Inhaltsverzeichnis

1	Einleitung	4
2	Analyseprinzipien und Disambiguierungsprobleme	5
2.1	Analyseprinzipien	5
2.2	Disambiguierungsprobleme	7
3	Lexikoninformation	8
3.1	Verben	8
3.2	Präpositionen	10
3.3	Nomen	11
4	Disambiguierung struktureller Mehrdeutigkeiten	13
4.1	Die Verweislisten OList und CList	13
4.2	Unterordnungsmöglichkeiten	14
5	Praktische Anwendung und Ausblick	18
6	Literaturverzeichnis	18
7	Abbildungsverzeichnis	21
8	Tabellenverzeichnis	21

1 Einleitung

Im Bereich der sprachorientierten KI-Forschung gibt es bereits seit Anfang der achtziger Jahre Ansätze zur automatischen Verarbeitung natürlicher Sprache, die eine Beschreibung des grammatikalischen Wissens in Form von Wortexperten bzw. Wortklassenfunktionen vorschlagen. Dazu gehören insbesondere das Wortexperten-Parsing (WEP) [Small 81], die wortklassengesteuerte funktionelle Analyse (WCFA) [Helbig 86], das wortorientierte Parsen [Eimermacher 88] und die Analyse mit Wortaktoren [Bröker, Hahn, Schacht 93]. Allen genannten Sprachbeschreibungsmodellen ist die grundlegende These gemeinsam, daß dem Wort die zentrale Rolle beim Sprachverstehen zukommt.

Die hier vorgestellten Disambiguierungsmethoden basieren auf dem Wortagentenmodell, einem wortzentrierten und prozeduralen Ansatz zur Sprachbeschreibung [Helbig, Mertens 94a], das im System LINAS der FernUniversität Hagen praktisch eingesetzt wird (s. Abschn. 5). Die Grundidee besteht darin, daß das Wort und die von ihm ausgehenden Erwartungen im Mittelpunkt des Sprachverarbeitungsprozesses stehen. Das Wechselspiel zwischen Eröffnung von Erwartungen (Valenzen)² und ihrer Sättigung bestimmt die Analyse mit Wortklassenagenten. Wird ein neues Wort in die Analyse einbezogen, so löst dieses syntaktische und semantische Erwartungen aus, die erst im weiteren Verlauf erfüllt werden können (Eröffnungsphase eines Wortklassenagenten). Nach Einbeziehung weiterer Wörter in die Verarbeitung, werden die hinzukommenden Informationen in Form partiell analysierter ("verständener") Strukturen gesammelt, die daraufhin überprüft werden, ob sie zu den ausgelösten Erwartungen passen. In dieser Phase, der sogenannten Komplettierungsphase, werden syntaktische und semantische Strukturen aufgebaut bzw. vervollständigt. Mit dieser Zweiteilung in Eröffnungs- und Komplettierungsphase unterscheidet sich der hier verwendete Ansatz deutlich von anderen Wortexpertenmodellen (vgl. [Helbig, Mertens 94b]).

Das Wortagentenmodell stützt sich auf kognitive Verhaltensweisen, wie sie beim menschlichen Sprachverstehen vorzufinden sind. Der Mensch ist

²Den Valenzbegriff verwenden wir im Sinne von Tesnière (vgl. [Tesnière 59]). Die Erwartungen, die von einem Konzept bzw. Lexem ausgehen, werden auf semantischer Ebene durch kognitive Rollen in der Bedeutungsstruktur dieser Konzepte ausgelöst und formal durch sogenannte Tiefenrelationen beschrieben. In KI-Terminologie werden die eröffneten Valenzen *Slots* bezeichnet, und die sie sättigenden Konstituenten mit *Fillers*.

offensichtlich bestrebt, möglichst frühzeitig – also noch bevor der Satz vollständig analysiert ist – begriffliche Strukturen bzw. Vorstellungen herauszubilden, die wir *mentale Kerne* nennen. Diese stellen bildlich gesprochen die Säulen dar, auf denen das gesamte Satzgebäude aufgebaut wird. Im Wortagentenmodell werden diese kognitiven Strukturen – nun aber auf analysetechnischer Ebene – durch das Konzept der elementaren bzw. komplexen Kerne ausgedrückt. Die Kerne repräsentieren abgeschlossene semantische Einheiten, von denen im Sinne eines valenzorientierten Sprachverarbeitungsmodells keine Erwartungen ausgehen. Einfache elementare Kerne können im Verlauf des Sprachverarbeitungsprozesses durch Herstellung von Beziehungen zu anderen Kernen oder durch funktionelle Komposition zu komplexen Kernen zusammengefaßt werden (vgl. [Helbig 94]).

2 Analyseprinzipien und Disambiguierungsprobleme

Es ist ein wesentliches Merkmal natürlichsprachlicher Ausdrücke, daß sie mehrdeutig sein können. Sprachverstehen bedeutet deshalb immer auch *Auflösung* dieser Mehrdeutigkeiten. Für die automatische Verarbeitung natürlicher Sprache sind verschiedene Prinzipien zur Disambiguierung struktureller Mehrdeutigkeiten vorgeschlagen worden (z.B. Minimal Attachment, Right Association und Lexical Preference). Sie sind in praktisch angewandten Systemen insbesondere für die Entscheidung relevant, ob eine Konstituente dem Verb oder einer anderen Konstituente zuzuordnen ist. Manche der verwendeten Prinzipien sind eng an bestimmte grammatische Konzeptionen gebunden (wie z.B. Minimal Attachment an Konstituentenbäume), andere wieder werden durch experimental-psychologische Untersuchungen gestützt (vgl. z.B. [Hemforth, Konieczny 93], [Hemforth, Konieczny, Scheepers 94]).

2.1 Analyseprinzipien

Von den Analyseprinzipien, die dem Wortagentenmodell zugrundeliegen, werden nachfolgend die für die Disambiguierung relevanten vorgestellt. Zunächst einmal sind drei Prinzipien zu nennen, die im Valenzsystem der natürlichen Sprache wurzeln. **Kompatibilitäts-** und **Inkompatibilitätsprinzip** regeln die Kombinierbarkeit von Konstituenten mit Hilfe der im Lexikon bei

valenzeröffnenden Lexemen unter den Merkmalen SELECT und KOMPAT (vgl. dazu den Lexikoneintrag des Verbs *schreiben* in Abschn. 3) eingetragenen Tiefenbeziehungen.³

Durch das Kompatibilitätsprinzip wird sichergestellt, daß nur solche Konstituenten zusammengefaßt werden, deren Deutungsspektren hinsichtlich der beteiligten Tiefenrelationen einen nicht-leeren Durchschnitt besitzen. Das Inkompatibilitätsprinzip schließt umgekehrt eine Zusammenfassung von Lexemen zu einer höheren syntaktischen bzw. semantischen Einheit aus, wenn in deren Lexikoneinträgen keine miteinander verträglichen Tiefenbeziehungen vorgefunden werden. Die beiden Prinzipien setzen voraus, daß die Lexikoninformation vollständig ist und insbesondere bei den Verben nicht nur die Tiefenkasusrelationen im engeren Sinne umfaßt, sondern auch den Bereich der freien Fügungen einschließt.

Schließlich sind die Valenzangaben nicht als gleichrangig anzusehen, sondern nach Prioritäten geordnet. Wenn konkurrierende Selektionsanforderungen mehrerer Wörter (wie z.B. Verb vs. Nomen, Verb vs. gleichzeitig vorhandenes Partizip) einander gegenüberstehen, wird das **Prioritätsprinzip** wirksam. Dabei sind die wichtigsten Ordnungskriterien die höhere Priorität von obligatorischen Valenzen gegenüber fakultativen Valenzen und von fakultativen Valenzen gegenüber freien Fügungen.

Kann mit Hilfe der valenzbasierten Prinzipien keine Entscheidung bezüglich der Unterordnung einer nominalen Konstituente unter eine vorangehende nominale Konstituente oder unter das Verb getroffen werden, dann wird das **Rechtsassoziationsprinzip** als wirksame Default-Regel zur Disambiguierung eingesetzt. Darunter verstehen wir die Tendenz, eine neue Konstituente eher dem zuletzt analysierten Kern als einem bereits vorher gebildeten Kern zuzuordnen (vgl. auch *Right Association* z.B. in [Allen 87]).⁴

Darüber hinaus kann für die Disambiguierung das Kriterium der **Lesartenpräferenz** eingesetzt werden. Es trägt der Tatsache Rechnung, daß bestimmte Wortbedeutungen häufiger verwendet werden als andere oder auch durch den außersprachlichen, situativen Kontext bevorzugt werden. Andere

³Beispiele für Tiefenrelationen sind: Handlungsträger (AGT), Objekt der Zuwendung (DAT), Richtungsangabe (DIRC), Instrument (INSTR), Ortsangabe (LOK), beteiligtes Objekt (OBJ), Zeitangabe (TEMP) etc.

⁴Psychologisch läßt sich das Rechtsassoziationsprinzip mit der beschränkten Kapazität des menschlichen Kurzzeitgedächtnisses und mit dem abnehmenden Aktivierungspotential der kognitiven Kerne erklären, je weiter diese im zeitlichen Verstehensablauf zurückliegen.

speziellere Kriterien betreffen z.B. die besondere Behandlung der Eigennamen und der Negatoren.

2.2 Disambiguierungsprobleme

Für die Arbeit der syntaktisch-semanticen Analyse in einem praktisch angewandten System ist die Lösung folgender Disambiguierungsprobleme⁵ von besonderer Bedeutung:

- morphologische Mehrdeutigkeiten (*produzierte*: Imperfekt bzw. Partizip II);
- lexikalische Mehrdeutigkeiten (Polysemie: *grün*, Homonymie: *Kiefer*, Homographie: *Tenor*);
- unterschiedliche Zuordnungsmöglichkeiten von syntaktischen Funktionen zu verschiedenen Tiefenkasus (*Die Mutter ruft das Kind.*);
- Pronomeninterpretation (z.B. Antezedentenfindung für Relativpronomen);
- Wirkungsbereich von Partizipien (*Der alte Meister mit Engagement betreuende Restaurator*);
- Wirkungsbereich von Quantoren (*Ein Buch hat jeder Schüler gelesen.* vs. *Jeder Schüler hat ein Buch gelesen.*);
- Konstituenten- vs. Satznegation (*Er kam nicht pünktlich.* vs. *Er kam nicht.*);
- Konstituenten- vs. Satzkoordination (*Er sah Peter mit dem Hund und dem Kind.* vs. *Er sah Peter mit dem Hund und das Kind mit der Katze.*);
- Zahl- vs. Datumsangaben (*Hat die Firma im Mai 1990 Rechner verkauft?*);
- Zusammensetzung von Appositionen (*... der Albrecht den Bären {aufband | noch kannte}...*);
- Funktionen von Satzzeichen (z.B. Komma als Satzzeichen vs. *und*);

⁵Zur Disambiguierungsproblematik vgl. z.B. auch [Hirst 87], [Schmitz, Quantz 93] und [Hindle, Rooth 93].

- Strukturelle Mehrdeutigkeiten, die mit der Rolle der Präposition im Satz und mit der Zuordnung von Konstituenten zum Verb bzw. zu einer vorhergehenden Konstituente zusammenhängen.

Die letzteren werden im Abschnitt 4 ausführlich erläutert. Sie werden wegen ihrer praktischen Bedeutung im Zentrum unseres Beitrags stehen, sollen aber zunächst anhand von Beispielsatz 1 verdeutlicht werden (vgl. auch Abbildung 2).

- (1) *Der unbekannte Schriftsteller schrieb ein sehr interessantes Buch über die Mammuthöhle in Kentucky.*

In diesem Beispielsatz läßt sich die Konstituente *in Kentucky* syntaktisch zum einen dem Verb *schreiben* unterordnen, und zum anderen kann sie die Konstituenten *ein sehr interessantes Buch* oder *die Mammuthöhle* näher bestimmen. Analoges gilt auch für die Präposition *über*, die sich entweder auf das Verb oder auf die vorausgehende Konstituente *ein sehr interessantes Buch* beziehen kann.

3 Lexikoninformation

Im System LINAS der FernUniversität Hagen werden im Lexikon zu jedem Lexem sowohl morphologische und syntaktische als auch semantische Informationen bereitgestellt. Die morphologische Information unter dem Attribut MORPH umfaßt insbesondere die Angabe des zuständigen Wortklassenagenten (WCA) und Flexionscharakteristika (FLEX). KOMPAT drückt bei Verben die semantische Kompatibilität im Hinblick auf freie Fügungen aus, und bei Präpositionen überhaupt die Kompatibilität mit den jeweils für eine Deutung in Frage kommenden Tiefenbeziehungen. Unter SELECT ist der Kasusrahmen angegeben, der insbesondere auch semantische Selektionsanforderungen enthält. SEMSORT ordnet den Lexemen Sorten einer lexikalischen Taxonomie zu (s. Abb. 1). SEMREL stellt elementare semantische Beziehungen zwischen den Lexemen her.

3.1 Verben

Verben und andere Wortklassen, die Leerstellen im Satz eröffnen, liefern den Schlüssel für die Disambiguierung von strukturellen Mehrdeutigkeiten.

Solche Lexeme verlangen die "Sättigung" bestimmter Valenzen in ihrem Kasusrahmen (vgl. [Fillmore 68]).

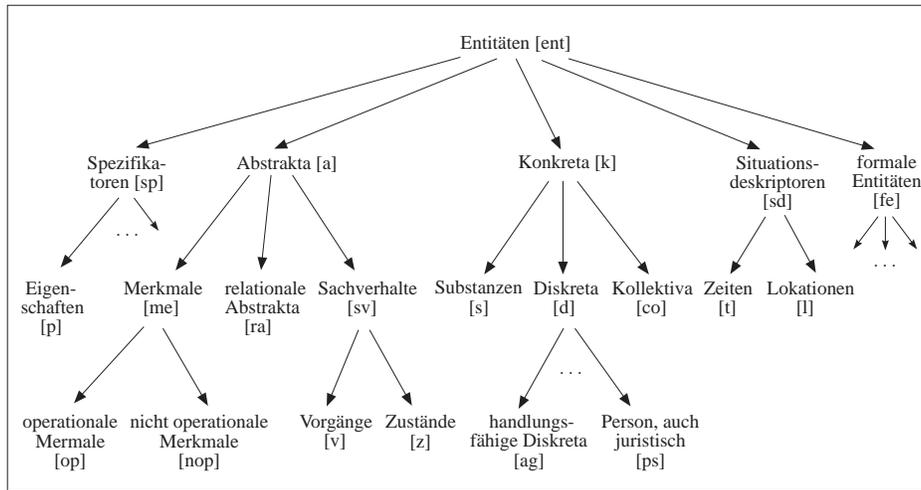


Abbildung 1: Ausschnitt aus der Sorteneinteilung

Ein vereinfachter Beispieleintrag für *schreiben* hat die nachfolgend angegebene Form. Das Lexem *schreiben* verlangt die Besetzung von mindestens zwei obligatorischen Argumenten. Das eine drückt den Tiefenkasus AGT (aktiver Bewirker einer Handlung) aus. Ferner muß es von einer der unter SEMSORT angegebenen Sorten sein und im Oberflächenkasus Nominativ (CAS 1) stehen. Das zweite obligatorische Argument kann zum einen durch den Tiefenkasus OBJ (Objekt, das von der Handlung betroffen wird) besetzt werden. In diesem Fall wird es im Oberflächenkasus Akkusativ realisiert und seine Sorte unterliegt den unter dem entsprechenden SEMSORT-Attribut aufgeführten Beschränkungen. Die andere Alternative für das zweite obligatorische Element besteht in einer Dativ-Ergänzung, die den Tiefenkasus DAT (Objekt, dem sich die Handlung zuwendet) bezeichnet. Auch beide Alternativen sind möglich. Ferner selegiert *schreiben* noch ein fakultatives Argument, das THM (Thema) ausdrückt. Es wird syntaktisch durch eine mit *über* eingeleitete Konstituente realisiert.⁶ *Schreiben* ist neben LOK und

⁶Runde Klammern bedeuten Optionalität, geschweifte schließen Alternativen ein und gerade Striche symbolisieren eine Menge von Selektionsrestriktionen.

TEMP z.B. mit INSTR und (im Gegensatz zu *schlafen*) mit DIRC kompatibel.

schreiben:

$$\left[\begin{array}{l}
 \text{MORPH} \quad \left[\begin{array}{l} \text{WCA} \quad *VB \\ \text{FLEX} \quad PG,NS,VK1 \end{array} \right] \\
 \text{KOMPAT} \quad LOK,DIRC,TEMP,INSTR,\dots \\
 \text{SELECT} \quad \left\{ \begin{array}{l} \left[\begin{array}{l} \text{SEMREL} \quad AGT \\ \text{CAS} \quad 1 \\ \text{SEMSORT} \quad AG,PS \end{array} \right] \\ \left[\begin{array}{l} \text{SEMREL} \quad OBJ \\ \text{CAS} \quad 4 \\ \text{SEMSORT} \quad IF \end{array} \right] \\ \left[\begin{array}{l} \text{SEMREL} \quad DAT \\ \text{CAS} \quad 3 \\ \text{SEMSORT} \quad PS \end{array} \right] \end{array} \right\} \\
 \left(\left[\begin{array}{l} \text{SEMREL} \quad THM \\ \text{WCA} \quad *PRÄP \\ \text{LEX} \quad ÜBER \\ \text{CAS} \quad 4 \end{array} \right] \right) \\
 \text{SEMREL} \quad [\text{SYNM} \quad \text{VERFASSEN}] \\
 \text{SEMSORT} \quad V
 \end{array} \right]$$

3.2 Präpositionen

Die Behandlung der durch Präpositionen eröffneten Valenzen spielt eine wichtige Rolle bei der Disambiguierung. Ein und dieselbe Präposition kann die unterschiedlichsten Tiefenkonzepte ausdrücken:

- (2) *Er liest das Buch über Wortklassen.* (THM: Thema)
- (3) *Es wurden über hundert Exemplare davon verkauft.* (GREATER: Größerrelation)
- (4) *Das Buch steht im Regal über dem Schreibtisch.* (LOK: Lokation)
- (5) *Einer flog über das Kuckucksnest.* (VIA: lokale Erstreckung)
- (6) *Er hat das Haus über den Makler gekauft.* (AUW: Art und Weise)

Diese Verträglichkeiten werden im Rahmen des Wortagentenmodells unter dem Lexikonmerkmal KOMPAT eingetragen. Ferner ist für die semantische Interpretation (vgl. DEUTG) jeder Präposition ein gesonderter Wortagent zuständig. In diesem Falle handelt es sich um den Agenten *ÜBER:

$$\text{über:} \left[\begin{array}{l} \text{MORPH} \quad \left[\begin{array}{ll} \text{WCA} & *PRÄP \\ \text{FLEX} & NFL \end{array} \right] \\ \text{KOMPAT} \quad \text{THM, GREATER, LOK, VIA, AUW} \\ \text{DEUTG} \quad *ÜBER \end{array} \right]$$

Für die Deutung von *über* in Beispielsatz 2 ergibt sich z.B. die Tiefenrelation THM, da das vorausgehende Nomen *Buch* der Sorte "Informationsträger" (IF) angehört. Diese Deutung geht mit der Selektion einer nominalen Ergänzung im Akkusativ einher. Die durch den Präpositionsagenten während der Satzanalyse aus dem Spektrum der möglichen Bedeutungen von *über* ermittelte Lexikoninformation ist nachstehend aufgeführt:

$$\text{über:} \left[\begin{array}{l} \text{MORPH} \quad \left[\begin{array}{ll} \text{WCA} & *PRÄP \\ \text{FLEX} & NFL \end{array} \right] \\ \text{SELECT} \quad \left[\begin{array}{ll} \text{SEMREL} & \text{THM} \\ \text{CAS} & 4 \end{array} \right] \end{array} \right]$$

3.3 Nomen

Neben den Verbvalenzen sind auch die durch Nomen eröffneten Leerstellen für die Analyse wichtig.⁷ Zwar handelt es sich hier häufig um fakultative Argumente; dennoch tragen sie in erheblichem Maße zur Klärung von Disambiguierungsproblemen bei. Nomen besitzen obligatorische und fakultative Valenzen, im Falle von nominalisierten Verben auch freie Fügungen.⁸ Daher spielt das Merkmal KOMPAT bei Nomen eine untergeordnete Rolle:

$$\text{Schriftsteller:} \left[\begin{array}{l} \text{MORPH} \quad \left[\begin{array}{ll} \text{WCA} & *NOM \\ \text{GEN} & M \\ \text{FLEX} & S04 \end{array} \right] \\ \text{SEMREL} \quad \left[\begin{array}{ll} \text{SUB} & \text{VERFASSEER} \\ \text{SYNM} & \text{AUTOR} \end{array} \right] \\ \text{SEMSORT} \quad \text{AG, PS} \end{array} \right]$$

⁷vgl. [Pustejovsky 91, S. 410] "[...] the lexicon is not just verbs. [...] we have little insight into the semantic nature of adjectival predication, and even less into the semantics of nominals."

⁸Obligatorische Valenzen sind allerdings seltener wie bei *Errichtung*, das immer ein Objekt verlangt.

Buch:	<table style="border-collapse: collapse; width: 100%;"> <tr> <td style="padding-right: 10px; vertical-align: top;"><i>MORPH</i></td> <td style="border-left: 1px solid black; border-right: 1px solid black; padding: 5px;"> <table style="border-collapse: collapse; width: 100%;"> <tr> <td style="padding-right: 10px;"><i>WCA</i></td> <td><i>*NOM</i></td> </tr> <tr> <td><i>GEN</i></td> <td><i>N</i></td> </tr> <tr> <td><i>FLEX</i></td> <td><i>S11</i></td> </tr> </table> </td> </tr> <tr> <td style="padding-right: 10px; vertical-align: top;"><i>SELECT</i></td> <td style="border-left: 1px solid black; border-right: 1px solid black; padding: 5px;"> <table style="border-collapse: collapse; width: 100%;"> <tr> <td style="padding-right: 10px;"><i>SEMREL</i></td> <td><i>THM</i></td> </tr> <tr> <td><i>WCA</i></td> <td><i>*PRÄP</i></td> </tr> <tr> <td><i>LEX</i></td> <td><i>ÜBER</i></td> </tr> <tr> <td><i>CAS</i></td> <td><i>AKK</i></td> </tr> </table> </td> </tr> <tr> <td style="padding-right: 10px; vertical-align: top;"><i>SEMREL</i></td> <td style="border-left: 1px solid black; border-right: 1px solid black; padding: 5px;">[<i>SUB</i> <i>DRUCKERZEUGNIS</i>]</td> </tr> <tr> <td style="padding-right: 10px; vertical-align: top;"><i>SEMSORT</i></td> <td style="border-left: 1px solid black; border-right: 1px solid black; padding: 5px;"><i>IF</i></td> </tr> </table>	<i>MORPH</i>	<table style="border-collapse: collapse; width: 100%;"> <tr> <td style="padding-right: 10px;"><i>WCA</i></td> <td><i>*NOM</i></td> </tr> <tr> <td><i>GEN</i></td> <td><i>N</i></td> </tr> <tr> <td><i>FLEX</i></td> <td><i>S11</i></td> </tr> </table>	<i>WCA</i>	<i>*NOM</i>	<i>GEN</i>	<i>N</i>	<i>FLEX</i>	<i>S11</i>	<i>SELECT</i>	<table style="border-collapse: collapse; width: 100%;"> <tr> <td style="padding-right: 10px;"><i>SEMREL</i></td> <td><i>THM</i></td> </tr> <tr> <td><i>WCA</i></td> <td><i>*PRÄP</i></td> </tr> <tr> <td><i>LEX</i></td> <td><i>ÜBER</i></td> </tr> <tr> <td><i>CAS</i></td> <td><i>AKK</i></td> </tr> </table>	<i>SEMREL</i>	<i>THM</i>	<i>WCA</i>	<i>*PRÄP</i>	<i>LEX</i>	<i>ÜBER</i>	<i>CAS</i>	<i>AKK</i>	<i>SEMREL</i>	[<i>SUB</i> <i>DRUCKERZEUGNIS</i>]	<i>SEMSORT</i>	<i>IF</i>
<i>MORPH</i>	<table style="border-collapse: collapse; width: 100%;"> <tr> <td style="padding-right: 10px;"><i>WCA</i></td> <td><i>*NOM</i></td> </tr> <tr> <td><i>GEN</i></td> <td><i>N</i></td> </tr> <tr> <td><i>FLEX</i></td> <td><i>S11</i></td> </tr> </table>	<i>WCA</i>	<i>*NOM</i>	<i>GEN</i>	<i>N</i>	<i>FLEX</i>	<i>S11</i>																
<i>WCA</i>	<i>*NOM</i>																						
<i>GEN</i>	<i>N</i>																						
<i>FLEX</i>	<i>S11</i>																						
<i>SELECT</i>	<table style="border-collapse: collapse; width: 100%;"> <tr> <td style="padding-right: 10px;"><i>SEMREL</i></td> <td><i>THM</i></td> </tr> <tr> <td><i>WCA</i></td> <td><i>*PRÄP</i></td> </tr> <tr> <td><i>LEX</i></td> <td><i>ÜBER</i></td> </tr> <tr> <td><i>CAS</i></td> <td><i>AKK</i></td> </tr> </table>	<i>SEMREL</i>	<i>THM</i>	<i>WCA</i>	<i>*PRÄP</i>	<i>LEX</i>	<i>ÜBER</i>	<i>CAS</i>	<i>AKK</i>														
<i>SEMREL</i>	<i>THM</i>																						
<i>WCA</i>	<i>*PRÄP</i>																						
<i>LEX</i>	<i>ÜBER</i>																						
<i>CAS</i>	<i>AKK</i>																						
<i>SEMREL</i>	[<i>SUB</i> <i>DRUCKERZEUGNIS</i>]																						
<i>SEMSORT</i>	<i>IF</i>																						

Bei den Nomen wird – wie auch bei den Verben⁹ – zusätzlich zu den im Kasusrahmen enthaltenen Informationen jedes Lexem selbst durch SEMREL und SEMSORT semantisch klassifiziert. SEMREL ermöglicht eine thesaurusartige Vernetzung der Lexeme. Dabei bedeutet SYN M Synonymie und SUB begriffliche Unterordnung. Durch SEMSORT werden dem Lexem Sorten einer vorgegebenen Ontologie zugeordnet (vgl. Abb. 1). Es können auch mehrere Sorten zur Beschreibung herangezogen werden. Beispielsweise ist *Schriftsteller* sowohl als handlungsfähig (AG) als auch als Person (PS) klassifiziert. Ferner unterscheiden wir die Entitäten zusätzlich danach, ob sie Artefakte sind. Eine Subsorte von "Artefakt" stellt "Informationsträger" (IF) dar.¹⁰

⁹vgl. [Lenders 93, S. 432] "In der Computerlinguistik ist diese Frage [nach dem Designatum der Verben] bisher stark vernachlässigt worden, so daß man meinen könnte, Verben seien bezüglich ihres Inhaltes ärmer als Nomina, die mit vielen anderen Nomina vernetzt sind."

¹⁰Auf die Problematik der Ontologiebildung kann hier nicht näher eingegangen werden. Vgl. dazu z.B. [Lang 92].

4 Disambiguierung struktureller Mehrdeutigkeiten

4.1 Die Verweislisten OList und CList

Die Zuordnung einer Konstituente zum Verb bzw. zu einer vorausgehenden Konstituente ist durch die Gegenüberstellung von z.T. miteinander konkurrierenden offenen Valenzen (Slots) und von geschlossenen Kernen geprägt, die als potentielle Filler für diese Slots in Frage kommen. Die Wortagentenmaschine verwaltet die Träger der Valenzen in einer Verweisliste OList von offenen Elementen und die potentiellen Filler, d.h. die geschlossenen Kerne, in einer Verweisliste CList (vgl. Abb. 2).

- (7) *Der unbekannte Schriftsteller schrieb ein sehr interessantes Buch über die Mammuthöhle in Kentucky.*

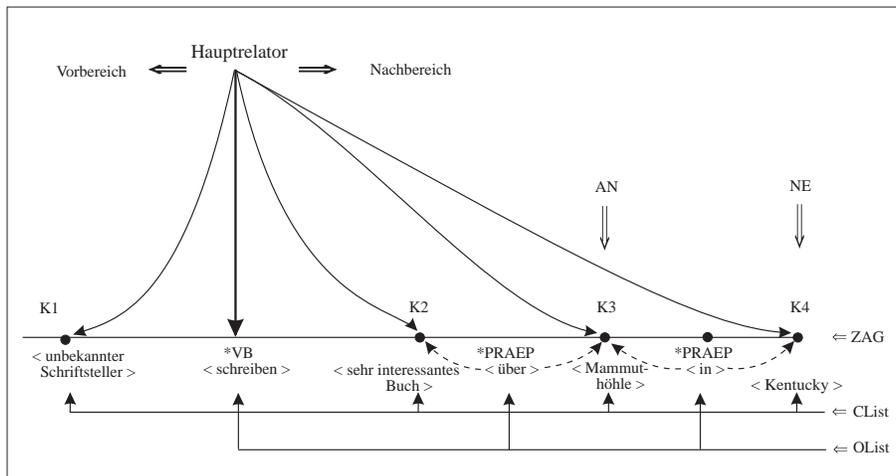


Abbildung 2: Verweislisten für die Disambiguierung struktureller Mehrdeutigkeiten

Als Ausgangspunkt der Disambiguierung werden die Position und die Merkmale des Hauptrelators des Satzes – im Beispielsatz 7 handelt es sich dabei um das Vollverb *schreiben* – ermittelt. Damit ist auch eine Zuordnung der Konstituenten zu Vorbereich und Nachbereich des Verbs gegeben. Diese

Information ist wichtig, da im Deutschen im Hauptsatz vor dem finiten Verb nur eine Konstituente, d.h. vor dem Hauptrelator nur ein (elementarer oder komplexer) Kern stehen darf, wodurch eine wichtige Disambiguierungshilfe gegeben ist. In Nebensätzen gilt das nicht, da dort (im Deutschen) das finite Verb ans Satzende tritt, und damit natürlich i.a. nicht alle Kerne im Vorbereich des Verbs zu einem komplexen Kern zusammenzufassen sind.

Die schwierigste Aufgabe besteht in der Entscheidung, ob die abgeschlossenen elementaren Kerne einer Kette vom Typ <Kern> <Präposition> <Kern>... jeweils mit Hilfe des durch die Präposition beschriebenen Relators einem davorstehenden Kern oder dem Verb (dem Hauptrelator des Satzes) unterzuordnen sind. Die elementaren Kerne, die in CList vorkommen, werden von rechts nach links – also entgegengesetzt zur Reihenfolge ihres Auftretens im Satz – analysiert und (soweit zulässig) zusammengefaßt.

4.2 Unterordnungsmöglichkeiten

In den folgenden Ausführungen bezeichnen wir mit NE das aktuell untersuchte Element von CList, mit AN das vor NE in CList stehende Element und mit ZAG das zentrale Arbeitsgedächtnis. Wenn unmittelbar vor NE ein elementarer Kern AN steht und zwischen NE und AN im ZAG eine Präposition P vorkommt, werden die folgenden vier Fälle unterschieden:

- NE ist eindeutig VB unterzuordnen (Fall 1)
- NE ist eindeutig AN unterzuordnen (Fall 2)
- NE ist einer anderen Konstituente aus CList unterzuordnen (Fall 3)
- Keine Entscheidung möglich (Fall 4)

Diese Fallunterscheidung ergibt sich als Konsequenz aus den Kombinationsmöglichkeiten der Bindungsstärken von P+NE an den Hauptrelator VB bzw. von P+NE an AN. Wenn NE an VB zu binden ist, dient die Präposition P als Rollenmarkierung in der Oberflächenstruktur des Satzes, und wenn NE dem AN unterzuordnen ist, wirkt die Präposition P als Funktor bzw. Relator mit eigenständiger semantischer Deutung.¹¹

¹¹Wir verwenden bewußt nicht den Begriff der *Präpositionalphrase*, da es sich dabei um eine rein syntaktische Kategorie handelt. Semantisch gehört die Präposition in Sätzen wie

Die oben genannten Kombinationen und die daraus resultierenden Fälle sind in Tabelle 1 zusammengefaßt. Die Bezeichnungen für die Bedingungen der Unterordnung sind in dieser Tabelle etwas verkürzt angegeben. So müßte die Angabe [NE+AN] bzw. "NE sicher zu AN" genauer lauten: "der Kern NE ist vermöge der Präposition P mit Sicherheit dem Kern AN unterzuordnen" (für die übrigen Bezeichnungen gilt Analoges). Der Fall 4 in der Mitte der Tabelle kann in heutigen Sprachverarbeitungssystemen i.a. nur durch Rückfragen an den Nutzer zu einem der anderen Fälle disambiguiert werden.

Es gibt eine Reihe wirksamer Kriterien, die es zu entscheiden gestatten, welche der in der Tabelle 1 verzeichneten Kombinationsmöglichkeiten ([NE+VB], [NE%VB], ..., [NE-AN]) vorliegen. In der nachfolgenden Auflistung ist für die einzelnen Kombinationsmöglichkeiten jeweils eine Auswahl an Kriterien und Beispielen angegeben (vgl. dazu auch die Abschnitte 2.1 und 3):

[NE+VB] :

- Eine obligatorische Valenz von VB läßt sich nur durch P+NE füllen (z.B. *Er glaubte in der Jugend an Gott.*);
- Konstituente NE steht am Satzanfang bzw. unmittelbar hinter VB (z.B. Kern K1 bzw. K2 in Abb. 2);

[NE%VB] :

- P+NE erfüllt fakultative Valenz von VB (z.B. Kern K3 in Abb. 2);

[NE-VB] :

- VB und die semantisch möglichen Deutungen von P sind inkompatibel (z.B. wird nach Ersetzen von *schreiben* durch *besprechen* in Abb. 2 die Unterordnung von K3 unter VB unzulässig, da unter den Lexikonmerkmalen SELECT und KOMPAT von *besprechen* keine Tiefenrelation eingetragen ist, die mit einer bei *über* verzeichneten Relation verträglich ist);

"*Peter* glaubt an *Gott*." und "*Maria* spricht mit *dem Nachbarn*." eher zum Verb als zur nachfolgenden Nominalgruppe.

[NE+AN] :

- Von AN gehen obligatorische Valenzen aus, die nur von P+NE erfüllt werden können (z.B. *Die Stadtverwaltung verbot die Errichtung* von Zäunen.);
- P=GENATTR¹² und VB besitzt keine Genitivreaktion (z.B. führt das Ersetzen von *in Kentucky* in Abb. 2 durch *des Staates Kentucky* zur obligatorischen Unterordnung von K4 unter K3);
- NE ist zweite Konstituente nach AN in CList im Vorbereich von VB (im Hauptsatz) (z.B. ist *aus Oklahoma* nicht VB sondern K1 unterzuordnen, wenn man *unbekannter Schriftsteller* in Abb. 2 zu *unbekannter Schriftsteller aus Oklahoma* erweitert);

[NE%AN] :

- P+NE erfüllt fakultative Valenzen von AN (z.B. *Trotz Kritik bewahrte sich der Schauspieler seinen Stolz* auf den Erfolg.);

[NE-AN] :

- AN und semantisch mögliche Deutungen von P sind inkompatibel (z.B. *Er fiel im Schlaf* aus dem Bett.);
- AN ist Personalpronomen und P \notin {mit, ohne} oder AN ist Relativpronomen (z.B. führt das Ersetzen von *Mammuthöhle* durch *sie* in Abb. 2 zur obligatorischen Unterordnung von K4 unter VB).

Wenn die (mitunter heuristischen) Unterordnungskriterien keine endgültige Entscheidung in mehrdeutigen Konstruktionen erlauben (Fall 4 in Tabelle 1), und kein weiteres Hintergrundwissen für die Disambiguierung zur Verfügung gestellt werden kann, dann liefert in vielen Fällen das Rechtsassoziationsprinzip (vgl. Abschnitt 2.1) eine gute Grundlage für die Ermittlung der richtigen Unterordnung. Im Beispielsatz 7 aus Abbildung 2 kann der Kern K4 sowohl dem Kern K3 als auch dem Verb VB untergeordnet werden. Diese Mehrdeutigkeit könnte eventuell durch Hintergrundwissen aufgelöst

¹²Jeder sicher erkannte Genitiv einer Nominalgruppe wird durch die künstliche Präposition GENATTR beschrieben, die etwa die Funktion des englischen *of* besitzt. Dadurch lassen sich Genitivattribute ähnlich wie präpositionale Fügungen behandeln.

werden (z.B. durch die Kenntnis einer Mammothöhle in Kentucky, von der die Rede ist). Dieses Wissen ist aber automatisch nur sehr schwer zu nutzen. Einfacher ist die Default-Annahme, daß bei solchen mehrdeutigen Sätzen die Zusammenfassung (Klammerung) von rechts her beginnt, daher der Name des Prinzips. Im Beispielsatz 7 ergibt sich nach dem Rechtsassoziationsprinzip auch die präferierte Deutung. Wollte man die andere Unterordnung K4 unter VB erzwingen, so müßte die Phrase *in Kentucky* im Satz unmittelbar nach dem Verb folgen. Dadurch würde der in der Auflistung unter dem Punkt [NE+VB] als zweites angegebene Fall wirksam, was zu einer eindeutigen Unterordnung der Konstituente führt.

	[NE+AN] NE sicher zu AN	[NE%AN] NE möglicher- weise zu AN	[NE-AN] NE sicher nicht zu AN
[NE+VB] NE sicher zu VB	UN- GRAMMATISCH	Fall 1 "Er gedachte in der Hütte <i>des Freundes</i> ."	Fall 1 "Man verbannte Na- poleon <i>nach Elba</i> ."
[NE%VB] NE möglicher- weise zu VB	Fall 2 "Er erinnerte sich des Gedenkens <i>an Goethe</i> ."	Fall 4 "Er traf den Mann <i>mit dem Stock</i> ." RÜCKFRAGE	Fall 1 "Peter flog im Traum <i>nach Rio</i> ."
[NE-VB] NE sicher nicht zu VB	Fall 2 "Ihn bewegte das Gedenken <i>an den Freund</i> ."	Fall 2 "Er schlief im Zug <i>nach Köln</i> ."	Fall 3 "Der Mann aus Köln <i>mit dem Lo- denmantel</i> schläft."

Tabelle 1: Fallunterscheidung für die Unterordnungsmöglichkeiten

Wenn die Kerne K4 und K3 durch die Arbeit des Agenten der Präposition *in* zu dem Kern K3' *Mammothöhle in Kentucky* zusammengefaßt sind, tritt noch einmal eine analoge Mehrdeutigkeit auf. Der Kern K3' kann dann zusammen mit *über* sowohl dem Verb *schreiben* als auch dem Kern K2 *interessantes Buch* untergeordnet werden. Auch hier liefert das Rechtsassoziationsprinzip eine legitime Entscheidung (Unterordnung von K3' unter K2).

Insgesamt liefern die hier behandelten Disambiguierungsmethoden gute Ergebnisse bei der Auflösung von Mehrdeutigkeiten. In den wenigen verbleibenden Fällen, in denen Rückfragen an den Nutzer gestellt werden müssen, stößt dies auf keine Akzeptanzprobleme.

5 Praktische Anwendung und Ausblick

Die praktische Anwendung der hier vorgestellten Disambiguierungsmethoden erfolgt im Rahmen der Wortagentenanalyse des Literaturrecherchesystems LINAS an der FernUniversität Hagen. LINAS ist ein Informationssystem, das dem Benutzer gestattet, Anfragen an eine Literaturdatenbank in natürlicher Sprache einzugeben. Zur Übersetzung einer natürlichsprachlichen Benutzeranfrage in eine semantische Tiefenstruktur greift die lexikalisch-morphologische Analyse auf das hier in Ausschnitten vorgestellte merkmalsbasierte Computerlexikon zu, das rechnergestützt mit Hilfe einer Werkbank für den Computerlexikographen aufgebaut wird. Die Bedeutungsdarstellung der natürlichsprachlichen Ausdrücke wird mit Hilfe der zur Wissensrepräsentation bereitgestellten Darstellungsmittel für mehrschichtige, erweiterte semantische Netze durchgeführt.

Das Wortagentenmodell wird zur Zeit dahingehend weiterentwickelt, daß die strukturellen Disambiguierungsprobleme in der konkurrierenden Wechselwirkung autonomer Agenten gelöst werden. Dabei führen die Wort- bzw. Wortklassenagenten derjenigen Lexeme, die die Valenzen eröffnen, sozusagen einen Wettbewerb um die jeweils vorhandenen abgeschlossenen Kerne (d.h. um die Filler). Die im LINAS-Lexikon unter den Merkmalen SELECT, KOMPAT, SEMSORT und SEMREL verzeichneten Tiefenbeziehungen stellen eine wesentliche Voraussetzung für die Lösung dieser Aufgabe dar.

6 Literaturverzeichnis

- [Allen 87] Allen, J.: Natural Language Understanding. Menlo Park, California: Benjamin/Cummings 1987
- [Boves, Nijholt 94] Boves, L., Nijholt, A. (ed.): Speech and Language Engineering, Proceedings of the eighth Twente Workshop on Language Technology, Enschede, NL: Faculteit Informatica, Universiteit Twente 1994
- [Bröker, Hahn, Schacht 93] Bröker, N., Hahn, U., Schacht, S.: Ein Plädoyer für Performanzgrammatiken. In: [DGfS-CL 93], S. 6–11

- [DGfS–CL 93] Deutsche Gesellschaft für Sprachwissenschaft, Sektion Computerlinguistik: Deklarative und prozedurale Aspekte der Sprachverarbeitung, Tagungsband der 4. Fachtagung, Universität Hamburg 1993
- [Eimermacher 88] Eimermacher, M.: Wortorientiertes Parsen. Dissertation. Fachbereich Informatik. Technische Universität Berlin 1988
- [Fillmore 68] Fillmore, Ch. J.: The case for case. In: Bach, E.; Harms, R. (Hrsg.): *Universals in linguistic theory*. New York: Holt, Rinehart & Winston 1968, S. 1–88
- [Helbig 86] Helbig, H.: Syntactic–semantic analysis of natural language by a new word–class controlled functional analysis (WCFA). *Bratislava: Computers and AI* 5 (1986) 1, S. 53–59
- [Helbig, Mertens 94a] Helbig, H., Mertens, A.: Der Einsatz von Wortklassenagenten für die automatische Sprachverarbeitung. Teil I – Überblick über das Gesamtsystem. *Informatik Berichte* 158. FernUniversität Hagen 1994
- [Helbig, Mertens 94b] Helbig, H., Mertens, A.: Word Agent Based Natural Language Processing. In: [Boves, Nijholt 94], S. 65–74
- [Helbig 94] Helbig, H.: Der Einsatz von Wortklassenagenten für die automatische Sprachverarbeitung. Teil II – Die vier Verarbeitungsstufen. *Informatik Berichte* 159. FernUniversität Hagen 1994
- [Hemforth, Konieczny 93] Hemforth, B., Konieczny, L.: Zum Zeitpunkt syntaktischer und thematischer Anbindung. In: [DGfS–CL 93], S. 47–52
- [Hemforth, Konieczny, Scheepers 94] Hemforth, B., Konieczny L., Scheepers, C.: Principle–based or Probabilistic Approaches to Human Parsing: How Universal is the Human Language Processor? In: [Trost 94], S. 161–170
- [Hindle, Rooth 93] Hindle, D., Rooth, M.: Structural Ambiguity and Lexical Relations. In: *Computational Linguistics* 19, S. 103–120
- [Hirst 87] Hirst, G.: *Semantic interpretation and the resolution of ambiguity*. London: Cambridge University Press 1987

- [Lang 92] Lang, E.: Linguistische vs. konzeptuelle Aspekte der LILOG-Ontologie. In: Klose, G., Lang, E., Pirlein, Th. (Hrsg.): *Ontologie und Axiomatik der Wissensbasis von LILOG*. Informatik-Fachberichte 307. Berlin etc.: Springer 1992, S. 23–45
- [Lenders 93] Lenders, W.: Strukturelle Eigenschaften von Bedeutungserklärungen deutscher Verben. In: Mattheier, K. J., Wegera, K.-P., Hoffmann, W., Macha, J., Solms, H.-J. (Hrsg.): *Vielfalt des Deutschen*. Festschrift für Werner Besch. Frankfurt a.M. etc.: Peter Lang 1993, S. 429–459
- [Pustejovsky 91] Pustejovsky, J.: The generative lexicon. In: *Computational Linguistics* 17 (1991), S. 409–441
- [Schmitz, Quantz 93] Schmitz, B., Quantz, J. J.: *Defaults in Machine Translation*. KIT-Report 106. Technische Universität Berlin 1993
- [Small 81] Small, S.: Viewing word expert parsing as linguistic theory. *Proceedings IJCAI-81*. Vancouver 1981, S. 70–76
- [Tesnière 59] Tesnière, L.: *Éléments de syntaxe structurale*. Paris 1959
- [Trost 94] Trost, H. (Hrsg.): *KONVENS '94 – Verarbeitung natürlicher Sprache*, Wien, 28.–30. September 1994. Berlin: Springer 1994

7 Abbildungsverzeichnis

1	Ausschnitt aus der Sorteneinteilung	9
2	Verweislisten für die Disambiguierung struktureller Mehrdeutigkeiten	13

8 Tabellenverzeichnis

1	Fallunterscheidung für die Unterordnungsmöglichkeiten	17
---	---	----