Statistische Verfahren zur Zuordnung von Präpositionalphrasen

Mehl, S ; Langer, H ; Volk, Martin

Abstract: Zahlreiche neuere Arbeiten für das Englische zeigen, daß statistische Analysen großer Korpora und Treebanks gute Heuristiken für die Zuordnung von Präpositionalphrasen liefern können. Entsprechende Untersuchungen für das Deutsche scheitern bisher an den fehlenden Daten. Wir zeigen jedoch, daß durch Einbeziehung weiterer Faktoren auch für das Deutsche mit guten Ergebnissen zu rechnen ist. Betrachtet werden der Einfluß unterschiedlicher Gewichte für Verben und Nomina, die Auswirkungen einer vorgeschalteten lexikalischen Disambiguierung sowie die Kopplung lexikalischer und grammatischer Präferenzen. Recent proposals have shown that statistical analyses of large English corpora and treebanks provide good heuristics for the attachment of prepositional phrases. Similar proposals for German have failed since such resources have not been available. We show that by using some additional factors we can achieve similar results for German. We demonstrate the influence of different weights for verbs and nouns, the influence of lexical disambiguation and the combination of lexical and grammatical preferences.

Statistische Verfahren zur Zuordnung von Präpositionalphrasen

Stephan Mehl, Hagen Langer, Martin Volk

Zahlreiche neuere Arbeiten für das Englische zeigen, daß statistische Analysen großer Korpora und Treebanks gute Heuristiken für die Zuordnung von Präpositionalphrasen liefern können. Entsprechende Untersuchungen für das Deutsche scheitern bisher an den fehlenden Daten. Wir zeigen jedoch, daß durch Einbeziehung weiterer Faktoren auch für das Deutsche mit guten Ergebnissen zu rechnen ist. Betrachtet werden der Einfluß unterschiedlicher Gewichte für Verben und Nomina, die Auswirkungen einer vorgeschalteten lexikalischen Disambiguierung sowie die Kopplung lexikalischer und grammatischer Präferenzen.

Recent proposals have shown that statistical analyses of large English corpora and treebanks provide good heuristics for the attachment of prepositional phrases. Similar proposals for German have failed since such resources have not been available. We show that by using some additional factors we can achieve similar results for German. We demonstrate the influence of different weights for verbs and nouns, the influence of lexical disambiguation and the combination of lexical and grammatical preferences.

1 Disambiguierung von Präpositionalphrasen

Die Zuordnung von Präpositionalphrasen (PPs) gilt als Paradebeispiel für eine strukturelle Mehrdeutigkeit. In einer von uns durchgeführten Untersuchung von 710 PP-Belegen waren nicht weniger als 502 (=70,7%) von ihrer syntaktischen Position her nicht eindeutig zuzuordnen. Im einfachsten Fall steht eine Präpositionalphrase hinter einem Nomen im Mittelfeld und kann entweder diesem oder dem Verb zugeordnet werden. Häufig stehen jedoch mehrere Nomina als Anbindungspunkte zur Verfügung, so daß eine kombinatorische Explosion droht.[1]

PP-Zuordnung ist ein typisch computerlinguistisches Problem, weil zu seiner Lösung komplexes semantisches Wissen erforderlich ist, das in keinem sprach-

[1] Da sie selten auftreten bzw. leicht zu erkennen sind, verzichten wir hier auf die Betrachtung von Satzadverbialen sowie von PPs, die Adjektiven oder Adverbien zuzuordnen sind.

verarbeitenden System zur Verfügung steht. Nur bei thematisch stark eingeschränkten Diskursbereichen (z.B. Fahrplanauskünften) ist eine semantische Disambiguierung denkbar. Zwar gibt es Fälle, in denen eine fehlende oder falsche PP-Zuordnung nicht auffällt, etwa bei der maschinellen Übersetzung bestimmter syntaktischer Strukturen, die in Quell- und Zielsprache gleichermaßen mehrdeutig sind. Dies gilt jedoch keineswegs immer. Die folgende Übersetzung durch den Personal Translator plus 98 etwa würde deutlich verbessert, wenn auch die letzte PP korrekt zugeordnet wäre:

> (1) *Mit der Ausgliederung [...] wird eine Voraussetzung für die geplante Kooperation mit Bertelsmann, America Online und dem Springer-Verlag geschaffen.*
> (1') *With the exclusion [...] a prerequisite for the planned cooperation is made with Bertelsmann, America online and the jumper publishing house.*

Als Alternative zur semantischen Analyse bietet sich die Einbeziehung von Valenzinformationen an. Beispiel (1) läßt sich disambiguieren, wenn man die Valenzrahmen der potentiellen Zuordnungspartner vergleicht: Nach Götz et al. 1993 besitzt *Kooperation* eine fakultative Komplementstelle für eine *mit*-PP, *schaffen* und *Voraussetzung* dagegen nicht. Allerdings ergeben sich bei einer solchen Valenzanalyse verschiedene Probleme: Erstens, wann eine Komplementstelle vorliegt und wann nicht, ist umstritten, besonders bei Nomina (vgl. Jacobs 1994). Zweitens, möglicherweise bieten beide Zuordnungspartner eine passende Valenzstelle an. In diesem Fall haben obligatorische den Vorrang vor fakultativen Valenzstellen. In seltenen Fällen sind die Erwartungen jedoch gleichrangig. Und drittens, es gibt keine vollständigen maschinenlesbaren Valenzlexika, besonders nicht für das Deutsche. Die umfangreichsten Valenzinformationen für ca. 9400 Verben bietet das CELEX. Selbst gedruckte Wörterbücher bieten selten Valenzinformationen; rühmliche Ausnahmen sind der dtv-Wahrig (Wahrig 1978) sowie Wörterbücher für Deutsch als Fremdsprache wie Götz et al. 1993.[2]

2 Quantitative Ermittlung von Bindungsstärken

Alle drei Probleme gleichzeitig lassen sich lösen, wenn man statt eines theoretisch definierten Valenzbegriffs die quantitativ ermittelte Wahrscheinlichkeit heranzieht, daß ein bestimmtes Lexem eine PP mit einer bestimmten Präposition

[2] Valenzwörterbücher wie Schumacher (1986) oder Helbig/Schenkel (1969) decken nur einen Bruchteil des Gesamtwortschatzes ab.

(eventuell auch noch einer bestimmten Inhaltsklasse) nach sich zieht. Ein Beispiel (Langer et al. 1997): Das Nomen *Zusammenarbeit* tritt in 44,5% seiner Belege in der COMPUTER ZEITUNG mit einer *mit*-PP zusammen auf, das Nomen *Server* dagegen nur in 3,8%, das Verb *arbeiten* in 38% seiner Belege. Aufgrund dieser Werte lassen sich die *mit*-PPs in (5) und (6) korrekt zuordnen:

(2) *Seit 1982 arbeitete das Standardisierungs-Gremium in enger Zusammenarbeit mit IEEE an den Basisnormen.*
(3) *Die Computer arbeiten auch als Server mit Microsofts Hochleistungsbetriebssystem Windows NT.*

Allerdings reicht eine solche einfache Statistik oft nicht aus, und zwar aus folgenden Gründen: Viele Lexeme sind so selten belegt, daß keine signifikanten statistischen Aussagen möglich sind (Problem der „spärlichen Daten"). Strenggenommen müßte man die Häufigkeit von Quadrupeln aus Verb, Nomen, Präposition und regiertem Nomen untersuchen. Dies würde jedoch das Problem der spärlichen Daten noch enorm vergrößern.

In jüngster Zeit haben sich besonders im englischsprachigen Raum zahlreiche Arbeiten recht erfolgreich mit möglichen Lösungen für diese Probleme beschäftigt. Leider sind die Ergebnisse dieser Arbeiten selten vergleichbar, weil einige sich speziell mit dem PP-Zuordnungsproblem beschäftigen[3], andere dagegen mit der Akquisition von Valenzrahmen im allgemeinen[4]. Bei letzteren fallen die berichteten Ergebnisse günstiger aus, weil die Erkennung von PP-Valenzstellen schwieriger ist als die Erkennung anderer Valenzstellen. Die meisten Autoren beschränken sich - wie im obigen Beispiel der Sätze (2) und (3) - auf den Vergleich der Tupel <Verb, Präposition> bzw. <Nomen, Präposition>. Collins/Brooks 1995 dagegen versuchen, soweit wie möglich mindestens die Tripel heranzuziehen, zu denen genügend Belege existieren. Damit erreichen sie das beste bisher publizierte Ergebnis für PP-Zuordnungen, nämlich 84,5%.

Das Problem der falsch positiven Belege wird üblicherweise vermieden, indem statt eines Korpus eine Treebank mit eindeutigen syntaktischen Analyseergebnissen als Datengrundlage verwendet wird, nämlich die Penn Treebank.[5]

---

[3] z.B. Hindle/Rooth 1993, Brill/Resnik 1994, Collins/Brooks 1995, Franz 1996, Basili et al. 1997, Merlo et al. 1997, Zavrel et al. 1997.
[4] z.B. Brent 1991, Manning 1993, Ersan/Charniak 1996, Briscoe/Carroll 1997.
[5] Allerdings sind die Einträge der Penn Treebank so konstruiert, daß nicht alle fehlerhaften PP-Zuordnungen auffallen (vgl. Manning/Carpenter 1997).

Einen anderen Weg gehen Hindle/Rooth 1993, indem sie zunächst die syntaktisch eindeutigen Belege aus einem nicht disambiguierten Korpus auswerten und die daraus gewonnenen Erwartungswerte als Heuristik zur Verarbeitung der mehrdeutigen Belege nehmen.

3 Eigene Untersuchung weiterer Faktoren

Das Ziel unserer Untersuchungen war es zunächst, zu überprüfen, ob statistische PP-Zuordnungsverfahren für das Deutsche ebenso gute Ergebnisse liefern wie für das Englische. Durch die unterschiedliche Wortstellung entstehen im Deutschen in vielen Fällen syntaktische Mehrdeutigkeiten, die im Englischen vermieden werden, z.B. bei einer Verschiebung des Subjekts hinter das Verb wie in (4) oder durch die Verbendstellung in Nebensätzen wie in (5):

(4) *Darüber sprach die COMPUTER ZEITUNG mit Charles B. Wang, ...*
(5) *Obwohl sich das Unternehmen mit strammen Wachstumsraten [...] auf Platz 67 der Weltrangliste der umsatzstärksten DV-Unternehmen hochgearbeitet hat ...*

Umgekehrt können im Deutschen PP-Komplemente problemlos an den Satzanfang gestellt werden, so daß die Zuordnung zum Verb eindeutig wird. Die Untersuchungsergebnisse für das Englische lassen sich also nicht ungeprüft auf das Deutsche übertragen. Wir haben uns für diese Überprüfung auf einen exemplarischen Fall beschränkt, nämlich PPs mit der Präposition *mit*.

Leider steht für das Deutsche derzeit noch keine genügend große Treebank zur Verfügung, um die oben geschilderten genaueren Methoden einsetzen zu können. Wir mußten uns daher darauf beschränken, mit Hilfe eines partiellen Parsings die potentiellen Anbindungspartner von *mit*-PPs zu ermitteln. Dabei stellten wir fest, daß sich wesentlich bessere Resultate erzielen lassen, wenn die statistisch auszuwertenden Belege vorher genauer untersucht werden, anstatt alle Belege in einen Topf zu werfen. Wir haben deshalb verschiedene Korpusausschnitte manuell disambiguiert und überprüft, welchen Einfluß die einzelnen Faktoren einer statistisch orientierten PP-Zuordnung haben. Im folgenden stellen wir unsere Ergebnisse aus den folgenden drei Bereichen vor: die unterschiedliche Behandlung von Nomina und Verben, die Auswirkungen einer lexikalischen Disambiguierung sowie die Kopplung von lexikalischen und grammatischen Präferenzen.

3.1 Nomenausgleichsfaktor und Mindesthäufigkeiten

3.1.1 Berechnung der Bindungsstärke

Um die Brauchbarkeit der Bindungsstärke für die Disambiguierung der PP-Anbindung zu untersuchen, haben wir die Werte zur Bindungsstärke zwischen Nomen und Präposition (N+P) und zwischen Verb und Präposition (V+P) aus einem Korpus von 2 Jahrgängen der COMPUTER ZEITUNG (2,7 Mio. Token) berechnet und evaluiert.

Die Berechnung der Bindungsstärke N+P ist relativ einfach, da die Präposition in den meisten Fällen unmittelbar auf das Nomen folgt. Um also die Bindungsstärke eines Nomens zur Präposition *mit* zu ermitteln, zählen wir einerseits die Auftretenshäufigkeit des Nomens unmittelbar vor *mit* (relative Häufigkeit) und andererseits die gesamte Auftretenshäufigkeit des Nomens (absolute Häufigkeit). Die Division der relativen Häufigkeit durch die absolute Häufigkeit ergibt die Bindungsstärke. Wenn also beispielsweise ein Nomen insgesamt 100mal im Korpus vorkommt und 60mal von *mit* gefolgt wird, dann erhält es den N+P Bindungswert 0,6.

Bei der Berechnung von N+P Bindungswerten wurden alle Nomen mit Hilfe von Gertwol lemmatisiert und die Bindungsstärke wurde jeweils über das Lemma berechnet. Außerdem wurde die Präzision der Berechnung dadurch gesteigert, daß das gesamte Korpus zunächst mit einem Wortarten-Tagger getaggt wurde. Auf diese Weise konnten diejenigen Fälle ausgeschlossen werden, in denen auf ein Nomen die Wortform *mit* folgt, diese aber nicht Präposition, sondern z.B. abgetrenntes Verbpräfix ist. Die höchsten Bindungsstärken ergaben sich nach diesem Verfahren für die Nomina *Umgang* (0,95), *Zusammenarbeit* (0,44), *Beschäftigung* (0.34), und *Zusammenhang* (0,31). Nomen mit einer relativen Häufigkeit kleiner 10 wurden in diese Liste nicht aufgenommen.

Die Berechnung der Bindungsstärke V+P ist dagegen kompliziert, da das Verb und die Präpositionalphrase an unterschiedlichen Positionen stehen können. Die einzige Bedingung für eine Bindung ist, daß Verb und Präposition in derselben Clause stehen. Es ist also notwendig, Clause-Grenzen (CG) zu identifizieren. Zu diesem Zweck haben wir einen Clause-Grenzen-Erkenner (CG-Erkenner) entwickelt. Er arbeitet mit regulären Ausdrücken über einem getaggten Korpus. Bei der Entwicklung wurde besonders darauf geachtet, eine hohe Präzision des CG-Erkenners zu erzielen, wenn dies auch zu Lasten der Vollständigkeit geht. Denn

nicht erkannte Clause-Grenzen führen zu „Einheiten“ mit mehreren Verben, diese Einheiten werden anschließend ausgefiltert und gehen damit nicht in die Berechnung ein. Unser CG-Erkenner wurde manuell evaluiert über 500 Sätze (mit einem Durchschnitt von 1,93 Clauses pro Satz) und erzielte 96,4% Präzision bei 87,7% Vollständigkeit.

Durch Tagging und Clause-Grenzen-Erkennung ermittelten wir rund 167'000 Clauses in unserem Korpus. Vor der Berechnung der V+P Werte wurden sämtliche Verben lemmatisiert (unter Berücksichtigung evtl. abgetrennter Verbpräfixe). Wir erhalten so 11'788 Verbformen mit 5031 unterschiedlichen Lemmata. Wir berechnen die relative Häufigkeit V+P als gemeinsames Vorkommen von Verblemma und Präposition in derselben Clause und andererseits die absolute Häufigkeit des Verbs als die gesamte Auftretenshäufigkeit des Verbs in allen Clauses. Die Verben mit den höchsten Bindungsstärken waren *behaften* (1,00), *begnügen* (1,00), *takten* (0,97), *versehen* (0,96), und *einhergehen* (0,92), wobei wiederum Verben mit einer relativen Häufigkeit kleiner 10 nicht berücksichtigt wurden.

3.1.2 Evaluierung der Bindungsstärke

Um den Nutzen der berechneten Werte zu überprüfen, haben wir rund 500 Sätze aus unserem Korpus, die alle die Präposition *mit* in ambiger Position enthalten, manuell disambiguiert. D.h. alle diese Sätze enthalten diese Präposition unmittelbar hinter einem Nomen, und die Präposition bezieht sich entweder auf dieses Nomen oder auf das Verb. Sätze, in denen die Präposition sich auf ein Nomen bezieht, das weiter vorn steht, oder Sätze, in denen die Präposition in einer idiomatischen Wendung vorkam, wurden manuell ausgeschlossen. Von 502 Sätzen, die wir auf diese Weise erhalten, ist die *mit*-PP in 266 Sätzen an das Nomen anzubinden und in 236 Sätzen an das Verb. Tabelle 1 zeigt in der ersten Spalte die Anbindungsraten, die ein erster Test mit den berechneten Werten ergab.

		NAF 4	NAF 4, $f \geq 15$
Anzahl der evaluierten Sätze:	473	473	236
Anzahl der korrekten V-Anbindungen:	203	164	75
Anzahl der korrekten N-Anbindungen:	88	187	106
Rate der korrekten V-Anbindungen:	91,85%	73,87%	72,81%
Rate der korrekten N-Anbindungen:	34,92%	74,50%	79,69%
Rate aller korrekten Anbindungen:	61,52%	74,20%	76,69%

Tabelle 1: Einfluß von Nomenausgleichsfaktor (NAF) und Mindesthäufigkeit

Von den 502 Sätzen wurden 29 nicht evaluiert, was immer dann vorkommt, wenn das Morphologie-Programm Gertwol zu einem Nomen oder einem Verb kein Lemma finden konnte. Das tritt insbesondere bei englischen Lehnwörtern (*User, Screen*) auf. Bei den 473 evaluierten Sätzen wurden 61,52% richtige Anbindungen erreicht. Das ist kein überzeugendes Ergebnis, wenn man bedenkt, daß durch zufällige Anbindung 50% erzielt würden. Es fällt jedoch auf, daß die Raten für V-Anbindung sehr gut, aber für N-Anbindung sehr schlecht aussehen. Das läßt die Vermutung zu, daß die von uns berechneten Werte zur V-Anbindung tendieren.

Wenn wir dementsprechend die Werte für die N-Anbindung mit einen Nomenausgleichsfaktor anheben, sollten sich zumindest die Raten für V-Anbindung und N-Anbindung angleichen. In unseren Experimenten haben wir deshalb die Werte für die N-Anbindung schrittweise mit den Faktoren 2, 3, 4, ... multipliziert, um herauszufinden, welcher Ausgleichsfaktor die besten Resultate liefert. Dabei stellt sich heraus, daß die Rate der korrekten Anbindungen bei Faktor 4 am besten ist und daß auch dort die Raten für korrekte V-Anbindung und N-Anbindung fast ausgeglichen sind (vgl. Tabelle 1, zweite Spalte). Der Nomenausgleichsfaktor 4 spiegelt offensichtlich die Tatsache wider, daß Nomen eine geringere Bindungsintensität mit Präpositionalphrasen eingehen als Verben.

Bei der obigen Evaluation haben wir alle Werte berücksichtigt, die wir vorher berechnen konnten, auch solche, die auf sehr geringen Häufigkeiten beruhen. Aber statistische Werte, die auf wenigen Instanzen basieren, gelten als nicht sehr verläßlich. Wir haben deshalb auch überprüft, ob sich die Anbindungsrate ändert, wenn wir Werte, die auf einer geringen relativen Häufigkeit beruhen, nicht berücksichtigen. Die dritte Spalte von Tabelle 1 zeigt die Ergebnisse, wenn wir eine relative Häufigkeit von mindestens 15 fordern (und den Nomenausgleichsfaktor bei 4 belassen).

Es sind jetzt nicht mehr 473 Sätze evaluiert worden, sondern nur noch 236 Sätze, da für die anderen die geforderte Mindesthäufigkeit nicht gegeben war. Jedoch steigt die Rate der korrekten Anbindungen nochmals um rund 2,5% an. Interessant ist, daß sich dieser Anstieg durch Anhebung der Mindesthäufigkeit nicht beliebig steigern läßt, sondern, im Gegenteil wieder leicht abnimmt. So erhalten wir für die Mindesthäufigkeit 25 nur noch eine Gesamtrate von 73,86%. Das läßt den Schluß zu, daß es auch unter den Verben und Nomen mit geringer Häufigkeit einige gibt, deren Bindungsstärke zur Präposition *mit* eine Entscheidung möglich macht, während hochfrequente Verben und Nomen durch viele unterschiedliche Lesarten das Ergebnis verwischen.

3.2 Lexikalische Disambiguierung

Theoretisch könnte die Bindungsstärke bei Verben mit obligatorischem PP-Komplement einen Wert von 100% annehmen. Tatsächlich ist dies fast nie der Fall, einerseits weil auch obligatorische Komplemente unter bestimmten Bedingungen entfallen können, vor allem aber, weil die meisten Verben polysem sind und nur in einem Teil ihrer Lesarten obligatorische Valenzstellen besitzen. Unter 268 Verben mit einer *mit*-Komplementstelle gemäß Götz et al. 1993, die wir genauer untersucht haben (Mehl 1998), fanden sich nur 25 eindeutige Verben mit obligatorischem *mit*-Komplement, also nur ein knappes Zehntel. Für die übrigen könnte die Effizienz der statistischen Verfahren deutlich verbessert werden, wenn die PP-Bindungsstärken nicht einem Lexem, sondern einzelnen Lesarten zugeordnet würden. Immerhin ist es für eine Reihe von Verben möglich, diese Lesarten aufgrund ihrer übrigen Valenzeigenschaften oder aufgrund von Selektionsrestriktionen zu identifizieren (vgl. Resnik/Hearst 1993). Wir untersuchten daher für einige polyseme Verben, welche Auswirkungen eine solche Zuordnung von Bindungsstärken zu Lesarten hat.

Um genügend Belege zu haben, wurden aus den obengenannten 268 Verben solche Polyseme ausgewählt, die jeweils mindestens 30mal in der ComputerZeitung belegt waren; zusammen sind dies 17 Verben mit 1644 Belegen. Diese 1644 Belege wurden manuell einer Lesart des jeweiligen Verbs zugeordnet. 498 der 1644 Belege enthielten eine *mit*-PP. Besonders interessant ist, wieviele dieser PPs eine obligatorische Komplementstelle füllen (was nichts besonderes wäre) oder aber eine fakultative, was zu einem charakteristischen Erwartungswert führt. Tabelle 2 schlüsselt dies auf:

Belege aus Lesarten mit obligatorischem mit-Komplement	164	
(alle mit *mit*-PP)		164
Belege aus Lesarten mit fakultativem mit-Komplement	777	
davon mit *mit*-Komplement		290
stattdessen mit *mit*-Adjunkt		4
Belege aus Lesarten ohne *mit*-Komplement	703	
davon mit *mit*-Adjunkt		40
Gesamtzahl der Belege	1644	
Gesamtzahl der *mit*-PPs		498

Tabelle 2: Funktion der Präpositionalphrasen

In 290 von 777 Belegen für fakultative Komplemente (=37,3%) fanden sich also tatsächlich *mit*-PPs in Komplementfunktion. Mischt man alle Lesarten und alle Arten von Valenzstellen zusammen, so ergeben sich im Durchschnitt 30% Belege mit *mit*-PPs, von denen 91% tatsächlich *mit*-Komplemente sind. Dieser Unterschied (30% vs. 37,3%) erscheint gering angesichts des Aufwands für eine lexikalische Disambiguierung. Die folgende Tabelle zeigt jedoch, daß die Unterschiede zwischen der separaten Betrachtung der Lesarten mit *mit*-Komplementstelle und der gängigen Durchschnittsberechnung in einigen Fällen wesentlich gravierender sind:

Verb	Bindungsstärke, bezogen auf Lesarten mit *mit*-Komplement	Bindungsstärke, bezogen auf alle Lesarten
unterhalten	0,33	0,02
drohen	0,59	0,17
überraschen	0,91	0,21
beschäftigen	0,94	0,19
sich aufhalten	1,00	0,05
handeln	1,00	0,15

Tabelle 3: Veränderung der Bindungsstärken durch lexikalische Disambiguierung

Hier führt eine Berechnung der Bindungsstärke ohne vorherige Disambiguierung offenbar zu groben Fehlbewertungen, die sich im praktischen Einsatz rächen. Beispielsweise würde (6) je nach Berechnung der Bindungsstärke für das Verb anders bewertet: Ohne lexikalische Disambiguierung liegt die Bindungsstärke für das Nomen (0,02 * NAF 4) höher als die des Verbs.

(6) *Wer sich angesichts dieser Entwicklung mit zwischenstaatlichen Querelen aufhält, bekommt schnell die Rechnung präsentiert.*

Wie leicht oder schwer diese Disambiguierung ist, hängt vom Einzelfall ab. Bei *sich aufhalten* etwa erfordern die alternativen Lesarten eine Orts- bzw. Zeitangabe, so daß die verschiedenen Lesarten anhand ihrer unterschiedlichen Valenzrahmen leicht zu identifizieren sind. Bei *unterhalten* können grobe semantische Constraints eingesetzt werden (*etw. unterhalten* [z.B. *eine Geschäftsbeziehung*] vs. *jd. unterhalten*). Belege für *überraschen* im Sinn von *verwundern* und im Sinn von *ertappen* sind dagegen nur schwer zu trennen.

3.3 Praktischer Einsatz der Bindungsstärken in einem probabilistischen Parser

Verläßliche Disambiguierungsstrategien für PP-Anbindung sind für Parsingsysteme aus zwei Gründen von Interesse: Zum einen erlauben sie eine präzisere Analyse, indem die Strukturbeschreibungen ambiger Konstruktionen nicht nur als (möglicherweise sehr große) ungeordnete Alternativenmenge oder ersatzweise als unterspezifizierte partielle Strukturbeschreibung ausgeben werden, sondern eine Ordnung über den Analyseresultaten definiert wird. Zum anderen kann eine frühzeitige verläßliche Disambiguierung die Analyseeffizienz eines Parsers massiv erhöhen[6].

In einem weiteren Experiment überprüften[7] wir die Auswirkungen von Bindungsstärken auf die Analysequalität eines Parsers. Dabei benutzten wir ein anderes Maß für die Berechnung der Bindungsstärke. Zu diesem Zweck extrahierten wir zunächst aus einem Korpus von insgesamt 1,1 Millionen Wortformen alle Muster des Typs *N mit-PP*. Das Verhältnis der relativen Häufigkeit eines Nomens in diesem Kontext zu seiner relativen Gesamthäufigkeit im Korpus nahmen wir als Maß für seine Bindungsstärke **np_attach**. Dieses Maß hat den neutralen Wert 1, wenn die relative Häufigkeit des jeweiligen Nomens im Kontext einer *mit*-PP sich nicht von seiner relativen Gesamthäufigkeit unterscheidet. Hat ein Nomen z.B. den **np_attach**-Wert 10, so kommt es im Kontext einer *mit*-PP zehnmal häufiger vor, als aufgrund seiner Gesamthäufigkeit erwartbar wäre. Die Resultate von insgesamt 73 häufigeren Nomina mit besonders hohen und niedrigen **np_attach**-Werten (z.B. *Umgang* (57.17), *Interview* (33.24), ..., *Jahre* (0.31), ...) integrierten wir in einen probabilistischen Parser, dessen Grammatik einen relativ breiten Bereich deutscher Syntax abdeckt. Der Parser verwendet eine kontextfreie Grammatik mit Merkmalsannotationen, deren Regeln im Sinne einer traditionellen probabilistischen Grammatik nach ihrer Häufigkeit relativ zur Kategorie der linken Regelseite gewichtet sind. Die Gesamtwahrscheinlichkeit einer Strukturbeschreibung ist das Produkt der Gewichte jener Regeln, die in einer Ableitung eines Satzes verwendet werden.

---

[6] So enthalten z.B. alle in Kuhn/Rohrer 1997 diskutierten Beispielsätze, deren Analyse mit der PARGRAM-Grammatik und dem XEROX-LFG-Parser länger als 10 Sek. dauert, mehrfache PP-Anbindungsambiguitäten. Auch unsere eigenen Erfahrungen bestätigen, daß massive globale syntaktische Ambiguität fast immer mit Präpositionalkonstruktionen zusammenhängt.

[7] Eine etwas detailliertere Beschreibung des Testverfahrens (allerdings lediglich mit Resultaten aus einer erheblich kleineren Testsatzmenge) findet sich in Langer et al. 1997.

Mit diesem Berechnungsverfahren lassen sich prinzipiell nur solche Ableitungen differenzieren, in denen verschiedene Regeln vorkommen. Das ist bei vielen Typen von PP-Anbindungs-Ambiguität nicht der Fall, da sich die Ableitungen nicht durch die darin vorkommenden Regeln, sondern nur durch deren Reihenfolge unterscheiden (z.B. [[*Gespräche der Regierung*] [*mit der Delegation*]] vs. [[*Gespräche*] [*der Regierung mit der Delegation*]]). Zur Bewertung solcher Anbindungsalternativen wurde der **np_attach**-Wert des jeweiligen Kopfnomens (im obigen Fall: *Gespräche* vs. *Regierung*) als zusätzlicher Faktor integriert, indem die aus den Regelgewichten geschätzte Ableitungswahrscheinlichkeit mit den **np_attach**-Werten multipliziert wurden. In diesem probabilistischen Modell waren zwar lediglich 73 Nomina (und überhaupt keine Verben) enthalten. Bei der Prüfung einer Stichprobe von 681 hinsichtlich der Zuordnung einer *mit*-PP ambigen Sätzen zeigten sich aber dennoch bereits positive Effekte auf die Disambiguierungsleistung des Parsers, obwohl in lediglich 47 von 197 Fällen von N-Anbindung ein **np_attach**-Wert im Modell enthalten war.

Die Testsatzmenge wurde zunächst manuell nach Fällen von 1. N-Anbindung einer *mit*-PP, 2. *mit*-PP als Adverbialphrase und 3. *mit*-PP als Präpositionalobjekt klassifiziert. Anschließend wurde die Testsatzmenge mit der Grundvariante des Parsers und der um das Anbindungsmodell erweiterten Version geparst. Eine Zuordnung des Parsers wurde dann als korrekt gewertet, wenn sie mit der manuellen Grobklassifikation übereinstimmte, d.h., Fehlzuordnungen – z.B. innerhalb komplexer NPn – wurde nicht als Fehler gewertet, solange sie nicht zu einer Fehlanbindung an ein Verb führten. Die Resultate sind in Tabelle 4 wiedergegeben.

	N-Anbindung	Adverbial	Präp. obj.	Summe
manuell	197	282	202	681
davon korrekt Parser mit Anbindungsmodell	155	236	90	481
davon korrekt Parser ohne Anbindungsmodell	149	237	89	475

Tabelle 4: Einbeziehung von Bindungsstärken

Die relativ geringe Steigerung der korrekten Erkennung im Fall von N-Anbindungen resultiert aus der Tatsache, daß lediglich in 47 von 197 Sätzen Anbindungen an ein Nomen vorkamen, für das das Anbindungsmodell einen

np_attach-Wert enthielt. Lediglich in 5 von diesen 47 Fällen wich die Analyse des Parsers mit Anbindungsmodell von der manuellen Analyse ab; darüber hinaus zeigte sich bei einer Detailanalyse, daß die jeweils korrekte Zuordnung in keinem Fall daran scheiterte, daß die **np_attach**-Werte die fehlerhafte Zuordnung präferiert hätten. Die Fehler waren ausnahmslos auf Lücken in der Grammatik zurückzuführen. So fehlten in der Grammatik z.B. Regeln zur Behandlung der Extraktion von PPn wie in (7):

(7) *Mit hiesigen Gewohnheiten und Bestimmungen hat das Unternehmen dagegen einige Probleme.*

Da die korrekte Zuordnung der *mit*-PP zu *Probleme* nicht möglich war, wurde *Mit hiesigen Gewohnheiten und Bestimmungen* vom Parser unkorrekt als Adverbial analysiert. Wo das Anbindungsmodell des Parsers den **np_attach**-Wert des Nomens mit angebundener *mit*-PP enthielt und die korrekte Analyse im Rahmen der verwendeten Grammatik möglich war, wurde die Anbindung jedoch auch korrekt erkannt.

Ein Blick auf die Analyseresultate zeigt, daß sowohl der Parser mit Anbindungsmodell als auch die Grundversion die deutlichsten Schwächen im Bereich der Erkennung von Präpositionalobjekten haben. Der Fehler besteht fast immer darin, daß statt dessen eine Klassifikation als Adverbial vorgenommen wird, und auch die Fälle von Nichterkennung von N-Anbindungen sind zumeist Fehlklassifikationen als Adverbial (36 N-Anbindungen wurden vom Parser mit Anbindungsmodell als Adverbiale analysiert, 42 vom Parser ohne Anbindungsmodell und nur jeweils 6 N-Anbindungs-Fälle wurden fälschlich als Präpositionalobjekt geparst). Ein Teil der Fehler war auf Lücken im Lexikon und im Regelapparat zurückzuführen. Der Rest beruht zum einen auf dem Faktum, daß das Gros der Verben mit einer *mit*-PP systematisch ambig ist, da ja gerade *mit*-PPs in der Regel fakultative Komplemente sind (s.o.), und zum anderen darauf, daß VP-Strukturen mit Akkusativ-Objekt und Adverbial erheblich häufiger sind als VPn mit Akkusativ- und Präpositionalobjekt.

Nun ist zumindest bei diesem Disambiguierungsproblem die Annahme, daß die häufigere Struktur auch mit einer entsprechend höheren Wahrscheinlichkeit die korrekte ist, empirisch falsch: Präpositionalobjekte sind zwar relativ selten, wenn sie aber durch den Subkategorisierungsrahmen eines Verbs zugelassen sind, ist die ansonsten häufigere Adjunkt-Lesart praktisch ausgeschlossen (s.o.). Mit die-

ser zusätzlichen Heuristik erzielte der Parser bei einer Stichprobe von insgesamt 238 Sätzen[8] eine korrekte Grobklassifikation für 91% der *mit*-PPn.

4 Ausblick

Unsere Untersuchungen haben ergeben, daß die statistische Ermittlung von Bindungsstärken eine zuverlässige Zuordnung von PPs mit Komplementfunktion erlauben dürfte - selbst dann, wenn mehrere potentielle Zuordnungspartner eine entsprechende PP erwarten, was selten vorkommt. Die Ergebnisse verbessern sich noch, wenn eine lexikalische Disambiguierung hinzukommt. Das entscheidende Problem liegt bei den Adjunkten. Listen typischer Verbadjunkte sowie die Erkennung von Orts- und Zeitangaben werden hier voraussichtlich weiterhelfen; dies ist ein Gesichtspunkt, an dem wir zur Zeit weiterarbeiten.

Sowohl der Bereich der Adjunkte als auch die Relevanz bestimmter Lesarten ist im übrigen stark durch die Textsorte geprägt. Wir haben die Analyseergebnisse für die fachsprachlichen COMPUTER ZEITUNG-Texte mit solchen für Tageszeitungstexte und Mischkorpora verglichen und erhebliche Unterschiede festgestellt. Beispielsweise liegt die Bindungsstärke für (*jd*) *unterhalten mit* im allgemeinsprachlichen Bereich wesentlich höher als im fachsprachlichen, wo die Lesart (*etw*) *unterhalten* dominiert. Solche fachsprachlichen Aspekte statistischer Verfahren werden daher einer unserer weiteren Forschungsschwerpunkte sein.

Literaturverzeichnis

Basili, Roberto; Candito, Marie-Hélène; Pazienza, Maria-Teresa; Velardi, Paola (1997): Evaluating the information gain of probability-based PP-disambiguation methods. In: D.B. Jones, H.L. Somers (Hrsg.): New Methods in Language Processing. London: UCL Press.

Brent, Michael R. (1991): Automatic Acquisition of Subcategorization Frames from untagged Text. Proc. 29. ACL.

Briscoe, Ted / Carroll, John (1997): Automatic Extraction of Subcategorization from Corpora. Proc. of ANLP-97.

Collins, Michael / Brooks, James (1995): Prepositional Phrase Attachment through a Backed-Off Model. Proc. Third Workshop on Very Large Corpora.

[8] Die Stichprobe bestand ausschließlich aus Sätzen, bei denen die korrekte Lesart mit dem symbolischen Kern der Grammatik grundsätzlich ableitbar war. Abgesehen von diesem Kriterium handelte es sich um eine zufällige Auswahl.

Ersan, Murat / Charniak, Eugene (1996): A statistical syntactic disambiguation program and what it learns. In: S. Wermter; E. Riloff; G. Scheler (Hrsg.): Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing. Berlin: Springer, S. 188-202.

Franz, Alexander (1996): Automatic Ambiguity Resolution in Natural Language Processing. Berlin: Springer.

Götz, Dieter; Hänsch, Günther; Wellmann, Hans (Hrsg.) (1993): Langenscheidts Großwörterbuch Deutsch als Fremdsprache. Berlin: Langenscheidt.

Helbig, Gerhard / Schenkel, Wolfgang (1969): Wörterbuch zur Valenz und Distribution deutscher Verben. Leipzig: VEB Verlag Enzyklopädie.

Hindle, Donald / Rooth, Mats (1993): Structural ambiguity and lexical relations. Computational Linguistics 19(1), S. 103-120.

Jacobs, Joachim (1994): Kontra Valenz. Trier: Wissenschaftlicher Verlag Trier.

Kuhn, Jonas / Rohrer, Christian (1997): Approaching ambiguitiy in real-life sentences - the application of an Optimality Theory-inspired constraint ranking in a large-scale LFG grammar. DGfS-CL-Tagung.

Langer, Hagen; Mehl, Stephan; Volk, Martin (1997): Hybride NLP-Systeme und das Problem der PP-Anbindung. In: Busemann, S.; Harbusch, K.; Wermter S.: Berichtsband des Workshops „Hybride konnektionistische, statistische und symbolische Ansätze zur Verarbeitung natürlicher Sprache“ auf der 21. Deutschen Jahrestagung für Künstliche Intelligenz. Freiburg.

Manning, Christopher D. (1993): Automatic Acquisition of a large Subcategorization Dictionary from Corpora. Proc. 31. ACL.

Manning, Christopher D. / Carpenter, Bob (1997): Probabilistic Parsing Using Left Corner Language Models. Proc. 5th International Workshop on Parsing Technologies, Boston.

Mehl, Stephan (1998): Semantische und syntaktische Disambiguierung durch fakultative Verbkomplemente. In: Ludewig, P.; Geurts, B.: Lexikalische Semantik aus kognitiver Sicht. Tübingen: Narr Verlag.

Merlo, Paola; Crocker, Matthew W.; Berthouzoz, Cathy (1997): Attaching Multiple Prepositional Phrases: Generalized Backed-off Estimation. Proc. of the Second Conference on Empirical Methods in Natural Language Processing. Brown University,RI.

Resnik, Philip / Hearst M. (1993): Structural Ambiguity and Conceptual Relations. In: Proceedings of the Workshop on very large Corpora: Academic and Industrial Perspectives S. 58-64, Ohio State University, June 22.

Schumacher, Helmut (Hrsg.) (1986): Verben in Feldern. Berlin: de Gruyter.

Wahrig, Gerhard (Hrsg.) (1978): dtv-Wörterbuch der deutschen Sprache. München: dtv.

Zavrel, Jakub; Daelemans, Walter; Veenstra, Jorn (1997): Resolving PP attachment Ambiguities with Memory-Based Learning. In: Mark Ellison (Hrsg.): Proc. of the Workshop on Computational Natural Language Learning (CoNLL '97), Madrid.

Statistische Verfahren zur Zuordnung von Präpositionalphrasen

Stephan Mehl
Gerhard-Mercator-Universität
Gesamthochschule Duisburg
FB3 - Computerlinguistik
Geibelstr. 41
47048 Duisburg
he234me@unidui.uni-duisburg.de

Hagen Langer
Universität Osnabrück
FB7
Neuer Graben 40
49069 Osnabrück
hlanger@cip.lili.uni-osnabrueck.de

Martin Volk
Universität Zürich
Institut für Informatik
Winterthurerstr. 190
CH-8057 Zürich
volk@ifi.unizh.ch