

Effizientes Parsing extraponierter Phrasen

Exposé des Dissertationsvorhabens

Jan Strunk

26. März 2005

1 Einleitung

In meiner Dissertation möchte ich mich mit dem effizienten Parsing extraponierter Phrasen beschäftigen. Dabei werde ich mich hauptsächlich auf das Deutsche und die im Deutschen häufig vorkommende Extraposition von Relativsätzen konzentrieren.

Im Folgenden werde ich kurz das Phänomen der Extraposition beschreiben und in den Abschnitten 2–4 den Stand der theoretischen und computerlinguistischen Forschung zu diesem Thema umreißen. In Abschnitt 5 leite ich daraus mein eigenes Dissertationsvorhaben ab. Zum Schluss berichte ich in Abschnitt 6 über bereits geleistete Vorarbeiten und gebe in Abschnitt 7 einen Zeitplan für das weitere Vorgehen an. Das Exposé schließt mit einer ausführlichen Literaturliste.

Beispiele (1) und (2) enthalten zwei Varianten desselben komplexen Satzes. In Beispiel (1) folgt der durch eckige Klammern markierte Relativsatz unmittelbar auf das von ihm modifizierte Kopfnomen *Reformkräfte*. In Beispiel (2) dagegen ist der Relativsatz durch mehrere Wörter von dem modifizierten Kopfnomen getrennt. Da der Relativsatz in Beispiel (2) rechts vom infiniten Verb *werden* und damit außerhalb der durch das finite Verb *sollen* und das infinite Verb *werden* gebildeten Satzklammer steht, bezeichnet man diese Wortstellungsvariante als Extraposition (Rechtsversetzung).

- (1) Sollen bis zum Tag seiner Auslieferung *all jene demokratischen Reformkräfte*, [die Kroatien binnen weniger Jahre zu einem Hoffnungsträger der Region gemacht haben], in politische Geiselhaft genommen werden?
(Vom Autor geänderte Version von Beispiel 2)
- (2) Sollen bis zum Tag seiner Auslieferung *all jene demokratischen Reformkräfte* in politische Geiselhaft genommen werden, [die Kroatien binnen weniger Jahre zu einem Hoffnungsträger der Region gemacht haben]?
(Westdeutsche Allgemeine Zeitung, 17. März 2005)

Es wird allgemein angenommen, dass die beiden Varianten die gleichen Wahrheitsbedingungen haben und somit bedeutungsgleich sind (Stucky 1987, S. 379; vgl. auch Kiss 2003, S. 5). Die Version des Satzes in (1) wird traditionell als kanonische Form mit der zugrundeliegenden Wortstellung betrachtet, während

die Version in (2) mit dem extrapониerten Relativsatz als abgeleitete Variante mit “besonderer” Wortstellung gilt (Baltin 2001, S. 1; vgl. auch Culicover und Rochemont 1990, S. 23). Dies liegt daran, dass die moderne Syntaxtheorie davon ausgeht, dass Elemente, die semantisch zusammengehören, im Regelfall auch syntaktisch eine Konstituente bilden und somit adjazent zueinander stehen (vgl. Hawkins 1994). Dieses Prinzip hat schon Behaghel in seiner Deutschen Syntax als oberstes Gesetz formuliert:

Das oberste Gesetz ist dieses, daß das geistig eng Zusammengehörige auch eng zusammengestellt wird.
(Behaghel 1932, S. 4)

Dies ist in (2) nicht der Fall, da der Relativsatz keine direkte semantische Verbindung zum Verb *werden* oder gar zur Phrase *in politische Geiselhaft genommen werden* eingeht. Stattdessen modifiziert der Relativsatz in (2) genau wie der in (1) das Kopfnomen *Reformkräfte* und schränkt die Referenz der von diesem Nomen projizierten Nominalphrase ein. Erst die Tatsache, dass der Relativsatz in Beispiel (2) so interpretiert wird als stünde er adjazent zum Kopfnomen *Reformkräfte*, lässt die Beispiele (1) und (2) als syntaktische Varianten erscheinen und definiert Extraposition als eine syntaktische Alternation.

2 Theoretische Ansätze zur Relativsatzextraposition

Die moderne Syntaxtheorie hat verschiedene Ansätze entwickelt, um Extraposition formal beschreiben zu können und trotz der diskontinuierlichen Stellung von Kopfnomen und extrapониertem Relativsatz eine kompositionelle semantische Analyse zu ermöglichen. Im Folgenden möchte ich die drei wichtigsten in der Literatur vorgeschlagenen Ansätze zur syntaktischen Analyse extrapониierter Phrasen vorstellen.

2.1 Der Bewegungsansatz

Die in der generativen Grammatik am häufigsten vertretene Analyse (Ross 1967; Akmajian 1975; Baltin 1978, 1981) nimmt an, dass die beiden Sätze in (1) und (2) die gleiche zugrundeliegende Struktur haben, aber dass sich Satz (2) von Satz (1) darin unterscheidet, dass eine Bewegungsoperation den Relativsatz aus seiner kanonischen Position in der Nominalphrase heraus nach rechts bewegt hat.¹ Meist wird dabei angenommen, dass der bewegte Relativsatz eine Spur (englisch *trace*) in der kanonischen Position zurücklässt, vgl. das etwas kürzere Beispiel (3).

(3) Im Katalog habe ich [einen Artikel *t*] gesehen [der sich im Internet nicht bestellen lässt]. ↑

Im Folgenden werde ich diese Art von Analyse als *Bewegungsansatz* bezeichnen. Dabei schließe ich auch Ansätze wie z.B. den von Keller (1995) im Rahmen deklarativer Grammatiktheorien wie der Head-Driven Phrase Structure

¹Die genaue Landeposition des Relativsatzes ist umstritten. Ich werde hier nicht weiter darauf eingehen.

Grammar (HPSG) mit ein, die keine wirklichen Bewegungsoperationen annehmen, sondern Extraposition mit Hilfe nichtlokaler Merkmalsperkolation als einer Art von Bewegungssimulation modellieren (vgl. Kiss 2003). In solchen Ansätzen werden – ähnlich wie bei der Analyse von Linksversetzungen wie z.B. der Topikalisierung – leere Kategorien oder lexikalische Regeln benutzt, um nichtlokale Abhängigkeiten einzuführen. Die Information über diese Abhängigkeiten wird dann in der syntaktischen Struktur auch über die maximale Projektion eines Kopfes hinaus weitergegeben und kann so benutzt werden, um auch zwischen nicht adjazenten Phrasen einerseits morphosyntaktische Merkmale abzugleichen² und andererseits eine kompositionelle semantische Struktur aufzubauen.

2.2 Der Diskontinuitätsansatz

Der zweite wichtige Ansatz zur Analyse von Extraposition wird hauptsächlich in neueren Arbeiten im Rahmen der HPSG vertreten (z.B. Reape 1994; Nerbonne 1994; Kathol und Pollard 1995); siehe auch Kathol (2000, S. 29–46) für einen geschichtlichen Überblick. Diese Autoren gehen davon aus, dass Sätze neben der für die kompositionelle Semantik wichtigen, stark hierarchischen “tektogrammatismatischen” Struktur auch eine mitunter weniger hierarchische, d.h. flachere, “phänogrammatismatische” Struktur haben (Curry 1961; Dowty 1996), über die die lineare Abfolge der einzelnen Wörter und Phrasen in einem Satz in sogenannten Word-Order-Domains modelliert wird. Eine tektogrammatismatische Konstituente wie – z.B. eine Nominalphrase einschließlich eines Relativsatzes – kann dabei entweder als *ein* Element in einer Word-Order-Domain erscheinen oder die Unterbestandteile dieser Konstituente können als separate Elemente in der Word-Order-Domain stehen. In ersterem Fall kann die tektogrammatismatische Konstituente auch phänogrammatismatisch nur als zusammenhängende Phrase (also kontinuierlich) realisiert werden. In letzterem Fall können die separaten Elemente mit anderen in der Word-Order-Domain unter Beachtung eventueller Linearisierungsbeschränkungen vermischt werden, so dass eine tektogrammatismatische Konstituente phänogrammatismatisch, d.h. im Satz als Kette von Wörtern, auch diskontinuierlich realisiert werden kann. Schematisch ist dies in (4) und (5) mit einem Beispiel aus Kathol und Pollard (1995) dargestellt. Beide Strukturen in (4) und (5) stellen die mögliche phänogrammatismatische Struktur einer Verbalphrase dar, die eine Nominalphrase mit einem Relativsatz enthält. In (4) ist die Nominalphrase als Ganzes Teil der Word-Order-Domain der Verbalphrase, so dass sie nur kontinuierlich linearisiert werden kann: *einen Hund, der Hunger hat, füttern*. In (5) dagegen ist der Relativsatz sozusagen aus der Nominalphrase freigelassen worden und kann in der Word-Order-Domain der Verbalphrase separat und damit auch diskontinuierlich linearisiert werden: *einen Hund füttern, der Hunger hat*.

(4) < <einen, Hund, <der, Hunger, hat>>, füttern >

(5) < <einen, Hund>, <der, Hunger, hat>, füttern >

²Ein Beispiel ist die Kongruenz in Genus und Numerus zwischen dem Kopfnomen und dem Relativpronomen auch in extrapolierten Relativsätzen wie dem in Beispiel (2).

Analysen wie Nerbonne (1994) und Kathol und Pollard (1995) nehmen somit an, dass tektogrammatische Konstituenten phonologisch diskontinuierlich realisiert werden können. Deshalb werde ich sie im Folgenden unter den Begriff *Diskontinuitätsansatz* fassen.

2.3 Der anaphorische Ansatz

Die beiden bisher beschriebenen Ansätze gehen davon aus, dass ein extraponierter Relativsatz zwar in der phonologischen Wortkette nicht adjazent zu der von ihm modifizierten Nominalphrase erscheint, aber trotzdem in irgendeiner zugrundeliegenden *syntaktischen* Struktur eine Konstituente mit dieser Nominalphrase bildet – sei es in der Tiefenstruktur in transformationellen Theorien oder in der tektogrammatischen Struktur in Diskontinuitätsanalysen. Der dritte Ansatz, den ich im Folgenden den *anaphorischen Ansatz* nennen möchte (Wittenburg 1987; Stucky 1987; Haider 1996, 1997; Kiss 2003), versucht hingegen nicht, die kanonische Position von extraponierten Relativsätzen in der Syntax zu “rekonstruieren”. Stattdessen nehmen die Verfechter dieses Ansatzes an, dass ein extraponierter Relativsatz syntaktisch in der rechtsperipheren Position generiert wird, aber durch anaphorische Prozesse trotzdem semantisch wie ein nicht extraponierter Relativsatz interpretiert wird. Dadurch ergibt sich für Sätze mit extraponierten Relativsätzen im Vergleich zu den anderen beiden Analyseansätzen eine weniger komplizierte syntaktische Struktur. Gleichzeitig wird das Problem der Diskontinuität auf die Semantik verlagert.

Die Annahme, dass ein extraponierter Relativsatz anaphorisch an die von ihm modifizierte Nominalphrase angeschlossen wird, ist insofern plausibel, als Relativsätze (im Deutschen und vielen anderen Sprachen) von Relativpronomina³ eingeleitet werden, die die gleiche Form haben wie andere pronominale Elemente, z.B. Demonstrativpronomina oder Interrogativpronomina.⁴ Dabei geht Kiss (2003) davon aus, dass sich Relativpronomina von anderen Pronomina darin unterscheiden, dass sie referentiell defekt sind und deshalb unbedingt ein geeignetes Antezedens im selben (möglicherweise eingebetteten) Satz benötigen.

Kiss (2003) zeigt, dass die Extraposition von Relativsätzen anderen Beschränkungen unterworfen ist, als die Extraposition von syntaktischen Komplementen. Er argumentiert gegen eine Bewegungsanalyse für Relativsatzextraposition, da die Extraposition von Relativsätzen nicht den allgemein in der generativen Grammatik angenommenen Bewegungsbeschränkungen unterworfen zu sein scheint (Kiss 2003, S. 1–7). Er wendet sich auch gegen den Diskontinuitätsansatz, da dieser Extraposition im wesentlichen als ein rein phonologisches Phänomen ohne semantische Konsequenzen ansieht und deshalb nicht dazu geeignet ist, Interaktionen zwischen Extraposition und Variablenbindung (Kiss 2003, S. 33–39) zu modellieren. Stattdessen schlägt er eine HPSG-Analyse vor, derzufolge Relativsätze syntaktisch nicht nur nominale Phrasen sondern auch z.B. Präpositionalphrasen, Verbalphrasen und sogar Sätze modifizieren können, solange diese Phrasen ein geeignetes Antezedens für das Relativpronomen enthalten (Prinzip der *verallgemeinerten Modifikation*):

³Auch der traditionelle Terminus für diese Elemente legt eine anaphorische Funktion nahe.

⁴Interessanterweise geht auch ein prominenter Ansatz in der Psycholinguistik, der sogenannte *attachment-binding dualism*, davon aus, dass die Interpretation von Relativsätzen sowohl von syntaktischen als auch von anaphorischen Prozessen bestimmt wird (Hemforth et al. 2000).

[...] relative clauses can simply be adjoined to a wide categorical range of phrases, provided that these phrases contain suitable antecedents for the relative clause.

(Kiss 2003, S. 40)

Die anaphorische Identifikation des Relativpronomens mit seinem Antezedens modelliert er mittels der ohnehin in der HPSG angenommenen semantischen Indexmerkmale für referentielle Ausdrücke, wobei er die HPSG-Theorie so modifiziert, dass diese Indexmerkmale auch über die maximale Projektion eines nominalen Ausdrucks hinaus in der syntaktischen Struktur perkolieren können und so auch für einen diskontinuierlich realisierten Relativsatz zugänglich sind.

3 Relativsatzextraposition als lohnendes Problem für die praktische Computerlinguistik

Bevor ich die Vorzüge und Nachteile der drei vorgestellten theoretischen Ansätze in Bezug auf das praktische, computerlinguistische Parsing von Sätzen diskutiere, möchte ich kurz ausführen, warum es sich überhaupt lohnt, sich mit Relativsatzextraposition aus Sicht der praktischen Computerlinguistik auseinanderzusetzen. Dazu möchte ich eine Korpusauszählung aus Uszkoreit et al. (1998) zitieren:

Von den 11.996 vollständig annotierten Sätzen des Korpus enthalten 12% einen Relativsatz (1394) [...] Rund 24% der Relativsätze (340) sind extraponiert, 32% (449) nicht extraponiert. Für den Fall, daß die NP und der Relativsatz zwar adjazent sind, beide Konstituenten sich aber bereits am Ende des Satzes befinden, treffen wir keine Entscheidung bezüglich der Extraposition. Diese Fälle (ca. 44%) bleiben im folgenden unberücksichtigt.

(Uszkoreit et al. 1998, S. 131)

Dies bedeutet, dass 2,83% aller Sätze in dem von Uszkoreit et al. (1998) untersuchten Zeitungskorpus extraponierte Relativsätze enthalten. In einer von mir selbst durchgeführten Untersuchung eines Korpus von niederdeutschen Texten (Strunk 2004) ergab sich ein Anteil von 26% extraponierter (335), und 23 % nichtextraponierter Relativsätze (295) an der Gesamtzahl aller Relativsätze (1285). Für den Rest der Relativsätze (655) war wiederum nicht entscheidbar, ob Extraposition vorlag oder nicht. Der im Vergleich zu der Untersuchung von Uszkoreit et al. (1998) relativ hohe Anteil von extraponierten Relativsätzen in meinem niederdeutschen Korpus, könnte darauf hindeuten, dass Relativsatzextraposition in der gesprochenen Sprache näheren Registern häufiger vorkommt als in eher formelleren Zeitungstexten, da für das moderne Niederdeutsche keine in der Schule vermittelte Schriftsprache existiert. Relativsatzextraposition ist somit durchaus kein marginales Phänomen. Ein System, das nicht in der Lage ist, extraponierte Relativsätze zu analysieren, entweder weil die verwendete Grammatik sie nicht modelliert oder der verwendete Parser nicht mächtig genug ist, verliert demnach ungefähr 3% Coverage, d.h. ca. 3% aller Sätze eines typischen deutschen Zeitungskorpus können nicht geparkt werden. Somit scheint es sich durchaus zu lohnen, nach effizienten Möglichkeiten zum Parsen extraponierter Relativsätze zu suchen.

Obwohl Relativsatzextraposition im Deutschen also relativ häufig ist und darüber hinaus auch in anderen Sprachen vorkommt, gibt es laut Crysmann (2004) zur Zeit keine umfangreiche HPSG-Grammatik außer Stefan Müllers Babel-Grammatik (Müller 1996), die Sätze mit extraponierten Relativsätzen parsen kann.

4 Parsing mit den verschiedenen Ansätzen

In diesem Abschnitt werde ich für jeden der drei vorgestellten theoretischen Ansätze Vorteile und Nachteile beim praktischen Einsatz in automatischen Parsern diskutieren. Dabei werde ich mich auf die heute üblicherweise verwendeten Bottom-Up-Chart-Parser beschränken.

4.1 Parsing und der Bewegungsansatz

Das Parsing mit Grammatiken, die annehmen, dass ein extraponierter Relativsatz aus seiner kanonischen Position nach rechts bewegt worden ist, ist insofern ineffizient, als leere Kategorien wie die in den meisten Bewegungsanalysen angenommenen Spuren (traces) zu einer großen Zahl von lokalen Ambiguitäten während des Parsings führen. So muss beim Bottom-Up-Parsing für jede Nominalphrase mindestens eine mögliche Spur angesetzt werden (Müller 2004, S. 222–224),⁵ da sich zu einem späteren Zeitpunkt im Parsingprozess ein extraponierter Relativsatz finden könnte, der die Nominalphrase modifiziert und mit der Spur koindiziert werden muss; vgl. das folgende Zitat aus dem Handbuch zum ALE Parser, einem Bottom-Up-Parser für Grammatiken mit komplexen Merkmalsstrukturen wie z.B. der HPSG:

Empty categories are expensive to compute under a bottom-up parsing scheme such as is used in ALE. The reason for this is that these categories must be inserted at every position in the chart during parsing (with the same begin and end points). If the empty categories cause local structural ambiguities, parsing will be slowed down accordingly as these structures are calculated and then propagated. [...] Thus empty categories should be used sparingly, and preferably in environments where their effects will not propagate.
(Carpenter und Penn 1994, S. 43, 44)

Das gleiche gilt für Analysen, die Bewegung mit Hilfe nicht-lokaler Abhängigkeiten simulieren, die durch lexikalische Regeln eingeführt werden (Crysmann 2004, S. 2), wie z.B. den Ansatz von Keller (1995).

4.2 Parsing und der Diskontinuitätsansatz

Das Parsing von Grammatiken, die diskontinuierliche Konstituenten zulassen, ist ein komplexeres Problem als das Parsing von herkömmlichen, kontext-freien

⁵Wenn man annimmt, dass mehrere Relativsätze, die das gleiche Kopfnomen modifizieren, gleichzeitig extraponiert werden können, müsste sogar mehr als eine Spur pro Nominalphrase angenommen werden. Wenn man die Zahl extraponierter Relativsätze pro Nominalphrase nicht künstlich beschränkt, führt dies dazu, dass ein Bottom-Up-Parser (ohne spezielle Heuristiken) nicht mehr terminiert (Müller 2004, S. 222–224).

Grammatiken, da die lineare Reihenfolge der Wörter nicht mehr direkt an den terminalen Knoten des Strukturbaums ablesbar ist. Außerdem können geparsete Konstituenten nicht mehr nur durch die zwei Parameter *Anfangsposition in der Kette* und *Endposition in der Kette* bestimmt werden, da eine Konstituente diskontinuierlich sein und somit “Löcher” enthalten kann. Nach Müller (2004, S. 210) ist die Parsingkomplexität solcher Grammatiken im schlimmsten Fall mindestens exponentiell.

Daniels und Meurers (2002) haben jedoch gezeigt, dass relativ effiziente Algorithmen für das Parsing von Grammatiken mit diskontinuierlichen Konstituenten konstruiert werden können, die je nach dem Grad der Diskontinuität der benutzten Grammatik nur graduell ineffizienter als herkömmliche Parsingalgorithmen werden. Müller (2004) versucht sogar zu zeigen, dass der Diskontinuitätsansatz für eine Sprache wie das Deutsche mit vielen diskontinuierlichen Phänomenen theoretisch effizienter sein kann als der Bewegungsansatz, da bei Letzterem wegen der Annahme zahlreicher leerer Kategorien oder lexikalischer Regeln sehr viele lokale Ambiguitäten (durch passive Kanten) entstehen.

Trotz dieser Überlegungen besteht jedoch das praktische Problem, dass es zur Zeit keine effizienten Implementationen von Systemen gibt, die diskontinuierliche Konstituenten verarbeiten können (Crysmann 2004, S. 2). Das einzige implementierte System mit einer umfangreichen Grammatik für das Deutsche – das Babel-System (Müller 1996) – kann in der Verarbeitungsgeschwindigkeit nicht mit anderen HPSG-Parsern wie z.B. PET (Callmeier 2000) konkurrieren.

4.3 Parsing und der anaphorische Ansatz

Der Vorteil des anaphorischen Ansatzes für das Parsing ist, dass er keine zusätzlichen Anforderungen an den Parser stellt, da ein extraponierter Relativsatz weder mit einer leeren Kategorie koindiziert werden muss noch eine diskontinuierliche Konstituente mit der von ihm modifizierten Nominalphrase bildet. Bereits existierende, effiziente Parser für HPSG-Grammatiken, wie z.B. PET oder TRALE (Carpenter et al. 2003), können somit ohne Probleme zur Implementation des anaphorischen Ansatzes verwendet werden (vgl. Crysmann 2004, S. 14).

Ein besonderer Vorteil des Ansatzes von Kiss (2003) ist, dass der zusätzliche Aufwand durch extraponierte Relativsätze sich beim Parsing nur bemerkbar macht, wenn ein Satz tatsächlich einen extraponierten Relativsatz enthält, wohingegen der Parser beim Bewegungsansatz auch beim Parsing von Sätzen ohne extraponierten Relativsatz hypothetisch leere Kategorien annehmen würde. Der Kiss’sche Ansatz funktioniert also datengetrieben und ist dadurch besonders effizient (Crysmann 2004, S. 5, 17).

Crysmann (2004) berichtet über einen direkten Vergleich einer Implementation der Analyse von Kiss (2003) mit einer Implementation des Bewegungsansatzes. Seine Evaluation ergibt eine deutlich höhere Effizienz des anaphorischen Ansatzes nach Kiss (2003) gegenüber einem Bewegungsansatz im Stile von Keller (1995):

[...] the performance losses associated with the movement approach are considerable, increasing the number of executed tasks by a factor between 1.3 and 1.5. The anaphoric approach, however, features an increase in executed tasks of at most 12.7%. (Crysmann 2004, S. 16)

Crysmann verwendet jedoch eine leicht modifizierte Version der Analyse von Kiss (2003), bei der durch spezielle zusätzliche Merkmale Scheinambiguitäten verhindert werden. Diese Scheinambiguitäten entstehen dadurch, dass der referentielle Index eines möglichen Antezedens an mehreren Stellen in der syntaktischen Struktur mit dem referentiellen Index des Relativpronomens identifiziert werden kann, ohne dass dies einen semantischen Unterschied machen würde. In Beispiel (6) kann der extraponierte Relativsatz wegen des Prinzips der *verallgemeinerten Modifikation* entweder die Konstituente V' , die Konstituente V'' oder sogar noch eine höhere Konstituente modifizieren. In allen Fällen würde aber der Index des Relativpronomens mit dem über die NP hinaus projizierten Index des Nomens *Artikel* identifiziert werden, so dass sich für die verschiedenen syntaktischen Analysen weder ein semantischer Unterschied noch ein Unterschied in der Wortstellung ergeben würde.

- (6) Im Katalog habe [V'' ich [V' [NP einen Artikel $_i$] gesehen]] [der $_i$ sich im Internet nicht bestellen lässt].

Dieses Problem ungewollter Ambiguitäten bekommt Crysmann (2004) durch die Einführung zusätzlicher Merkmale in den Griff. Nur leider führt dies zu einer etwas komplizierteren Analyse und ausschließlich aus parsingtechnischen Gründen verwendeten Ad-Hoc-Merkmalen.

5 Vorschlag einer getrennten Verarbeitung des extraponierten Relativsatzes

Auf Grund der Überlegungen in den vorhergehenden Abschnitten gehe ich davon aus, dass der anaphorische Ansatz von Kiss (2003) sowohl aus praktischen als auch aus theoretischen Gründen der geeignetste Kandidat als Grundlage für eine effiziente Verarbeitung extraponierter Relativsätze ist. Meine Dissertation wird dabei auch deswegen auf dem Ansatz von Kiss (2003) aufbauen, weil dieser die Modifikation der Nominalphrase durch den extraponierten Relativsatz mit einer einfachen Identifikation der referentiellen Indizes des Kopfnomens und des Relativpronomens modelliert. Diese unkomplizierte Verknüpfung der beiden nicht adjazenten Konstituenten ermöglicht ein Verarbeitungskonzept, bei dem der extraponierte Relativsatz und der Rest des Satzes in einem ersten Schritt zunächst getrennt verarbeitet werden und die daraus resultierenden syntaktisch-semantischen Strukturen dann in einem zweiten Schritt zusammengesetzt werden.

Von dieser zweistufigen Vorgehensweise erwarte ich mir mehrere Vorteile. Erstens hat sich in mehreren Studien gezeigt, dass eine geteilte Vorverarbeitung komplexer Sätze einen Performanzgewinn erbringen kann (z.B. Foth und Menzel 2003; Frank et al. 2003). Zweitens ermöglicht eine solche Vorgehensweise, die Identifikation eines Antezedens für einen extraponierten Relativsatz mit speziell dafür entwickelten Verfahren vorzunehmen und dabei Erkenntnisse aus verschiedenen anderen computerlinguistischen Problembereichen anzuwenden. Drittens kann das von Crysmann (2004) beschriebene Problem der Scheinambiguitäten gelöst werden, ohne dass in der Grammatik Ad-Hoc-Merkmale angenommen werden müssen, indem man die Identifikation des Index des modifizierten Nomens mit dem Index des Relativpronomens als wichtigstes Problem auffasst und

die eigentliche syntaktische Anbindung des Relativsatzes entweder unterspezifiziert lässt oder durch bestimmte Prinzipien und Heuristiken so regelt, dass keine Scheinambiguitäten mehr auftreten.

Ein Performanzgewinn durch geteilte Verarbeitung kann sich dadurch ergeben, dass ein globales Problem in wohldefinierte Subprobleme geteilt wird, zwischen denen keine großen Abhängigkeiten bestehen, vgl. folgende Überlegung aus Foth und Menzel (2003), die sich sehr gut auf das Problem extraponierter Relativsätze anwenden lässt:

[...] this paper investigates different possibilities to break down large parsing problems into smaller subtasks whose results can later be recombined into a solution for the original problem. The rationale behind such an approach is that many attachment problems which have to be solved when parsing a sentence can also be treated in a fairly restricted context. Solving these local problems as local optimization subtasks might therefore save precious processing time which is needed badly to tackle the notoriously difficult attachment problems on the global sentence level.

(Foth und Menzel 2003, S. 92)

In der Analyse von Kiss (2003) besteht zwischen dem Nomen als Antezedens und dem Relativsatz lediglich eine anaphorische Beziehung. Darüber hinaus sind in der Regel keine weiteren Abhängigkeiten zwischen dem extraponierten Relativsatz und dem Rest des komplexen Satzes vorhanden. Insofern können diese beiden Teile des komplexen Satzes mit einer leicht modifizierten Grammatik zunächst getrennt geparkt und dann nachträglich zu einer kompletten Struktur zusammengesetzt werden. Der Parser muss so während der tiefen Verarbeitung nicht nach syntaktischen Abhängigkeiten zwischen dem Relativsatz und dem Rest des komplexen Satzes suchen und spart dadurch Verarbeitungszeit ein.

Ein besonders wichtiger Vorteil, der sich durch die getrennte Vorverarbeitung des Relativsatzes und des restlichen komplexen Satzes ergibt, ist die Möglichkeit, grammatik- und parsingexterne Verfahren zur Rekombination der beiden Teile zu verwenden. Diese Verfahren könnten dabei sowohl auf regelbasiertes als auch auf statistisches Wissen zurückgreifen. Die anaphorische Natur von Relativsätzen in der Analyse von Kiss (2003) legt dabei die Adaption von Methoden der Anaphernresolution nahe; siehe Mitkov (1999) für einen Überblick. Die computerlinguistische Anaphernresolution hat zum Ziel, die Referenz von Pronomina und anderen unselbstständigen referentiellen Ausdrücken zu bestimmen, indem ein korrektes Antezedens innerhalb desselben Satzes oder in einem vorhergehenden Satz gefunden wird. Die dabei zumeist verwendeten Kriterien der syntaktischen oder informationsstrukturellen Prominenz – Subjekte sind bessere Antezedenten als Objekte, usw. – und des linearen Abstandes – nähere referentielle Ausdrücke sind bessere Antezedenten als weiter entfernte – (vgl. Lappin und Leass 1994), dürften sich auch als relevant für die Identifikation des Antezedens eines extraponierten Relativsatz innerhalb eines komplexen Satzes herausstellen (vgl. auch Siddharthan 2002b, S. 45). Weitere relevante Forschungsgebiete der aktuellen Computerlinguistik sind Verfahren zur Disambiguierung der Anbindung von Präpositionalphrasen (PP attachment ambiguity resolution) wie z.B. Hindle und Rooth (1993) und Ratnaparkhi und Roukos (1994) und neuerdings auch Verfahren zur Disambiguierung der Anbindung von Relativsätzen

(allerdings zumeist nicht im Zusammenhang mit Extraposition) wie z.B. Siddharthan (2002a) und Siddharthan (2002b). Auch Studien aus der kognitiven Linguistik und der Psycholinguistik wie Uszkoreit et al. (1998), Hemforth et al. (2000), Konieczny (2000), u.a. ergeben nicht kategorische Beschränkungen zur Anbindung von extraponierten Relativsätzen und sind somit direkt relevant für die Suche nach dem Antezedens eines extraponierten Relativsatzes.

Ich möchte daher aufbauend auf der B-Ger-Gram⁶ – einer von Stefan Müller implementierten deutschen HPSG-Grammatik für das TRALE-System (Carpenter et al. 2003) – ein System implementieren und evaluieren, das den extraponierten Relativsatz und den Rest des komplexen Satzes zunächst getrennt verarbeitet und danach den Relativsatz mit Hilfe von noch zu erarbeitenden Verfahren an die übrige Struktur anbindet. Dazu ist es zunächst nötig die B-Ger-Gram-Grammatik so zu modifizieren, dass Relativsätze als eigenständige Fragmente geparkt werden können. Außerdem muss nach geeigneten grammatikexternen Methoden gesucht werden, um einen extraponierten Relativsatz zunächst überhaupt finden und separat verarbeiten zu können. Eine Möglichkeit dazu ist die Implementation eines speziellen Relativsatzchunkers, der mit heuristischen Methoden arbeitet; eine andere Möglichkeit ist die Benutzung eines topologischen Parsers (wie z.B. Becker und Frank 2002), der deutsche Sätze in die traditionelle topologische Felderstruktur einteilt, ohne detailliertere syntaktische Strukturen aufzubauen. Der nächste Schritt ist die grundlegende Implementation der Rekombination der beiden Teile unter Vermeidung von Scheinambiguitäten aber zunächst ohne ausgeklügelte statistische oder regelbasierte Verfahren zur Bestimmung der wahrscheinlichsten (bzw. geeignetsten) Antezedenten, um feststellen zu können, ob die getrennte Verarbeitung zu einer Erhöhung der Parsingeffizienz führt. Dazu werde ich die Performanz einer direkten Implementation des Kiss'schen Ansatzes mit der des hier vorgeschlagenen zweistufigen Systems mit getrennter Vorverarbeitung vergleichen. Als letzten Schritt möchte ich dann auch nichtkategorische Beschränkungen und statistisches Wissen in den Rekombinationsschritt einbauen, um idealerweise ein Rangfolge vom wahrscheinlichsten zum unwahrscheinlichsten Antezedens des extraponierten Relativsatzes erstellen zu können, die an handannotierten Korpusdaten und möglicherweise auch an experimentellen psycholinguistischen Daten überprüft werden kann.

Ich bin der Meinung, dass das hier vorgeschlagene System zum einen sehr effizient sein wird, weil es ohne leere Kategorien auskommt, keine diskontinuierlichen Konstituenten zulässt und "in sich abgeschlossene" Teile von komplexen Sätzen getrennt verarbeitet. Zum anderen vermeidet es Scheinambiguitäten und eröffnet die Möglichkeit, nicht kategorische Beschränkungen bei der Relativsatzanbindung zu untersuchen. Ich erwarte mir von meiner Arbeit also einen Performanz- und Robustheitsgewinn beim Parsing von Sätzen mit extraponierten Relativsätzen und interessante Ergebnisse durch die Erforschung statistischer Beschränkungen der Relativsatzanbindung.

⁶<http://www.cl.uni-bremen.de/Fragments/b-ger-gram.html> de

6 Schon geleistete Vorarbeiten

Die bisher von mir geleisteten Vorarbeiten bestehen im Wesentlichen in der theoretischen Vorbereitung und Konzeption des vorgeschlagenen zweistufigen Ansatzes zum Parsen von Sätzen mit extrapolierten Relativsätzen. Als Grundlage dafür habe ich eine gründliche Literaturrecherche zu den Themen Relativsatzextrapolation, Parsing, Anaphernresolution, PP-Anbindung und Relativsatzanbindung durchgeführt und auch nach relevanter Literatur aus der kognitiven Linguistik und der Psycholinguistik gesucht. Die von mir bis zu diesem Zeitpunkt aufgebaute bibliographische Datenbank umfasst schon 137 Titel. Die meisten dieser Titel und viele weitere, noch nicht bibliographierte, habe ich auch schon elektronisch oder als Kopie beschafft, was nicht immer einfach ist, da besonders die computerlinguistischen Aufsätze oft nur in Konferenzproceedings oder sogar (noch) gar nicht publiziert sind. Zusätzlich zu einer ersten Durchsicht der Literatur habe ich mich gründlich in die Theorie der Head-Driven Phrase Structure Grammar (HPSG) eingearbeitet und mit der Einarbeitung in die computerlinguistischen Tools und Systeme begonnen, die ich verwenden möchte. Dazu gehören die B-Ger-Gram-Grammatik von Stefan Müller und das TRALE-System (Carpenter et al. 2003), das sie als Parser verwendet, sowie die logische Programmiersprache PROLOG, in der das TRALE-System implementiert wurde.

7 Weiteres Vorgehen und Zeitplan

Bei meiner hier in Tabellenform angegebenen zeitlichen Planung zum weiteren Vorgehen gehe ich von einer Gesamtdauer meines Promotionsstudiums von zwei Jahren (24 Monaten) aus.

Monat	Geplantes Vorgehen
1-4	<ul style="list-style-type: none">• Weitere Literaturrecherche und eingehendes Studium der theoretischen und computerlinguistischen Forschungsliteratur• Durchsicht von Korpusdaten zur Überprüfung theoretischer Annahmen• Aufbau von Testkorpora durch das Sammeln authentischer Sätze mit extraponiertem Relativsatz• Weitere Einarbeitung in Grammatik, Parser und PROLOG• Verfassen einer ersten Version von Überblickskapiteln zum Stand der Forschung
5-8	<ul style="list-style-type: none">• Testen von Möglichkeiten zum grammatikexternen automatischen Finden von extraponierten Relativsätzen durch Implementation eines Relativsatzchunkers und Adaption eines topologischen Parsers• Vergleichende Evaluation dieser Systeme• Modifikation der verwendeten Grammatik, so dass diese einen Relativsatz als eigenständiges Fragment verarbeiten kann• Erste schriftliche Ausarbeitung dieser Schritte

Monat	Geplantes Vorgehen
9-12	<ul style="list-style-type: none"> • Implementation eines einfachen zweistufigen Systems mit getrennter Verarbeitung und darauffolgender Rekombination von extrapoliertem Relativsatz und dem Rest des komplexen Satzes (ohne statistisches Ranking) • Direkte Implementation des Ansatzes von Kiss (2003) ohne getrennte Verarbeitung von extrapoliertem Relativsatz und dem Rest des komplexen Satzes • Vergleichende Evaluation der beiden Systeme auf einem authentischen Korpus und auf einem speziell zusammengestellten Testkorpus zur Relativsatzextrapolation • Erste schriftliche Ausarbeitung dieser Schritte
13-16	<ul style="list-style-type: none"> • Noch einmal genauere Beschäftigung mit in der Literatur beschriebenen Verfahren der Anaphernresolution, der PP-Anbindung und der Relativsatzanbindung • Modifikation und Adaption solcher Verfahren für die Anbindung extrapoliierter Relativsätze • Empirische Untersuchungen mit Hilfe von Korpora zu Faktoren, die die Anbindung eines extrapolierten Relativsatzes beeinflussen könnten • Einarbeitung der gewonnenen Erkenntnisse in die Verfahren zur Bestimmung des besten Antezedens eines extrapolierten Relativsatzes • Erste schriftliche Ausarbeitung dieser Schritte
17-20	<ul style="list-style-type: none"> • Evaluation der statistischen und/oder regelbasierten Verfahren zur Bestimmung des besten Antezedens eines extrapolierten Relativsatzes mit Hilfe handannotierter Korpora • Eventuell auch Evaluation dieser Verfahren mit Hilfe psycholinguistischer Experimente • Erste schriftliche Ausarbeitung dieser Schritte
21-24	<ul style="list-style-type: none"> • Abschließende Evaluation • Überarbeitung und Zusammenfügung der bereits geschriebenen Kapitel zu einer kohärenten Dissertation

Literatur

- Akmajian, A. (1975). More evidence for an NP cycle. *Linguistic Inquiry* 6(1), 115–129.
- Baltin, M. R. (1978). *Toward a Theory of Movement Rules*. Dissertation, MIT, Cambridge, USA.
- Baltin, M. R. (1981). Strict bounding. In C. L. Baker und J. McCarthy (Hrsg.), *The Logical Problem of Language Acquisition*. MIT Press, Cambridge, USA, 257–295.
- Baltin, M. R. (Erster Entwurf von September 2001). Extraposition, the right roof constraint, result clauses, relative clause extraposition, and PP extraposition. In H. van Riemsdijk und M. Everaert (Hrsg.), *The Syntax Companion: An electronic encyclopaedia of case studies*. The LingComp foundation. <http://www.nyu.edu/gsas/dept/lingu/people/faculty/baltin/papers/extrapos.pdf>.
- Becker, M. und A. Frank (2002). A stochastic topological parser of German. Proceedings of COLING 2002, Taipei, Taiwan. 71–77.
- Behaghel, O. (1932). *Deutsche Syntax: Eine geschichtliche Darstellung, Band 4, Wortstellung. Periodenbau*. Carl Winters, Heidelberg.
- Callmeier, U. (2000). PET – A platform for experimentation with efficient HPSG processing techniques. *Journal of Natural Language Engineering* 6(1), 99–108.
- Carpenter, B. und G. Penn (1994). *ALE. The Attribute Logic Engine. User's Guide*. Computational Linguistics Program, Carnegie Mellon University, Pittsburgh, USA.
- Carpenter, B., G. Penn, und M. Haji-Abdolhosseini (Dezember 2003). *ALE. The Attribute Logic Engine. User's Guide with TRALE Extensions*. Version 4.0 Beta.
- Crysmann, B. (Version von Oktober 2004). Relative clause extraposition in German: An efficient and portable implementation. <http://www.coli.uni-sb.de/~crysmann/papers/>.
- Culicover, P. W. und M. S. Rochemont (1990). Extraposition and the complement principle. *Linguistic Inquiry* 21(1), 23–47.
- Curry, H. (1961). Some logical aspects of grammatical structure. In R. Jakobson (Hrsg.), *Structure of Language and its Mathematical Aspects*, Volume 7, Proceedings of Symposia in Applied Mathematics. American Mathematical Society, Providence, USA, 56–68.
- Daniels, M. und W. D. Meurers (2002). Improving the efficiency of parsing with discontinuous constituents. Datalogiske Skrifter No. 92, Proceedings of the 7th International Workshop on Natural Language Understanding and Logic Programming. Roskilde Universitetscenter, Kopenhagen, 49–68.

- Dowty, D. (1996). Towards a minimalist theory of syntactic structure. In H. Bunt und A. van Horck (Hrsg.), *Discontinuous Constituency*. Mouton de Gruyter, Berlin, 11–62.
- Foth, K. und W. Menzel (2003). Subtree parsing to speed up deep analysis. Proceedings of the 8th International Workshop on Parsing Technologies, IWPT-2003, April 2003, Nancy, Frankreich. 91–103.
- Frank, A., M. Becker, B. Crysmann, B. Kiefer, und U. Schäfer (2003). Integrated shallow and deep parsing: TopP meets HPSG. Proceedings of ACL 2003, Sapporo, Japan. 104–111.
- Haider, H. (1996). Downright down to the right. In U. Lutz und J. Pafel (Hrsg.), *On Extraction and Extraposition in German*, Linguistik Aktuell/Linguistics Today No. 11. John Benjamins Publishing, Amsterdam, 245–271.
- Haider, H. (1997). Extraposition. In D. Beerman, D. LeBlanc, und H. van Riemsdijk (Hrsg.), *Rightward Movement*. John Benjamins Publishing, Amsterdam, 115–151.
- Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge, UK.
- Hemforth, B., L. Konieczny, und C. Scheepers (2000). Syntactic attachment and anaphor resolution: The two sides of relative clause attachment. In M. Crocker, M. Pickering, und C. Clifton jr. (Hrsg.), *Architectures and Mechanisms for Language Processing*. Cambridge University Press, Cambridge, UK, 259–282.
- Hindle, D. und M. Rooth (1993). Structural ambiguity and lexical relations. *Computational Linguistics* 19(1), 103–120.
- Kathol, A. (2000). *Linear Syntax*. Oxford University Press, Oxford, UK.
- Kathol, A. und C. Pollard (1995). Extraposition via complex domain formation. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 26.–30. Juni, Cambridge, USA. 174–180.
- Keller, F. (1995). Towards an account of extraposition in HPSG. Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, 27.–31. März 1995, University College Dublin, Ireland. 301–306.
- Kiss, T. (Im Erscheinen. Version von November 2003). Semantic constraints on relative clause extraposition. *Natural Language and Linguistic Theory*. http://www.linguistics.rub.de/~kiss/publications/semconst_ss.pdf.
- Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research* 29(6), 627–645.
- Lappin, S. und H. Leass (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4), 535–561.

- Mitkov, R. (1999). Anaphora resolution: The state of the art. Arbeitspapier (basierend auf dem COLING98/ACL98-Tutorium über Anaphernresolution). <http://clg.wlv.ac.uk/papers/mitkov-99a.pdf>.
- Müller, S. (1996). The Babel-System – an HPSG Prolog implementation. In *Proceedings of the 4th International Conference on the Practical Application of Prolog*, London. 263–277. <http://www.cl.uni-bremen.de/~stefan/Pub/babel.html>.
- Müller, S. (2004). Continuous or discontinuous constituents? A comparison between syntactic analyses for constituent order and their processing systems. *Research on Language and Computation, Special Issue on Linguistic Theory and Grammar Implementation 2(2)*, 209–257. <http://www.cl.uni-bremen.de/~stefan/Pub/discont.html>.
- Nerbonne, J. (1994). Partial verb phrases and spurious ambiguities. In J. Nerbonne, K. Netter, und C. Pollard (Hrsg.), *German in Head-Driven Phrase Structure Grammar*. CSLI Publications, Stanford, USA, 109–150.
- Ratnaparkhi, A. und S. Roukos (1994). A maximum entropy model for prepositional phrase attachment. Proceedings of the ARPA Workshop on Human Language Technology.
- Reape, M. (1994). Domain union and word order variation in German. In J. Nerbonne, K. Netter, und C. J. Pollard (Hrsg.), *German in Head-Driven Phrase Structure Grammar*, CSLI Lecture Notes No. 46. CSLI Publications, Stanford, USA, 151–198.
- Ross, J. R. (1967). *Constraints on Variables in Syntax*. Dissertation, MIT, Cambridge, USA. Veröffentlicht als Ross, John R. 1986 Infinite Syntax! Ablex, Norwood.
- Siddharthan, A. (2002a). Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. Proceedings of the Student Research Workshop, 40th Meeting of the Association for Computational Linguistics (ACL 2002). 60–65.
- Siddharthan, A. (2002b). Resolving relative clause attachment ambiguities using machine learning techniques and WordNet hierarchies. Proceedings of the 5th National Colloquium for Computational Linguistics in the UK (CLUK 2002), University of Leeds. 45–49.
- Strunk, J. (2004). Relative clause extraposition in Low Saxon. Hausarbeit, Stanford University, <http://www.linguistics.rub.de/~strunk/RelClauseLS.pdf>.
- Stucky, S. U. (1987). Configurational variation in English: A study of extraposition and related matters. In G. J. Huck und A. E. Ojeda (Hrsg.), *Discontinuous Constituency*, Volume 20 of *Syntax and Semantics*. Academic Press, Orlando, USA, 377–404.
- Uszkoreit, H., T. Brants, D. Duchier, B. Krenn, L. Konieczny, S. Oepen, und W. Skut (1998). Studien zur performanzorientierten Linguistik: Aspekte der Relativsatzextraposition im Deutschen. *Kognitionswissenschaft 7*, 129–133.

Wittenburg, K. (1987). Extraposition from NP as anaphora. In G. J. Huck und A. E. Ojeda (Hrsg.), *Discontinuous Constituency*, Volume 20 of *Syntax and Semantics*. Academic Press, Orlando, USA, 428–445.