

Effizientes Parsing extraponierter Phrasen

Jan Strunk

28. April 2005

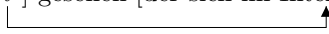
1 Das Problem der Extraposition

- (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?
(Modifizierte 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)
 - Die Sätze in den Beispielen (1) und (2) haben die gleiche Bedeutung (zumindest die gleichen Wahrheitsbedingungen).
 - Der Relativsatz in Beispiel (2) modifiziert das Kopfnomen *Reformkräfte*, obwohl er nicht adjazent dazu steht; er hat keine direkte semantische Beziehung zu dem Verbalkomplex *in politische Geiselhaft genommen werden*, an den er angrenzt.
 - Damit ist eine Grundannahme der Syntaxtheorie verletzt: semantisch Zusammengehöriges steht auch syntaktisch adjazent (z.B. Behaghel 1932, S. 4; Hawkins 1994).
 - Die adjazente Variante (Beispiel 1) wird traditionell als die kanonische Grundform betrachtet.

2 Lösungsansätze in der theoretischen Syntax

Ziel: Ermöglichung einer kompositionellen semantischen Analyse von Sätzen mit extraponierten Relativsätzen.

2.1 Bewegungsansatz

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

- Es wird angenommen, dass eine Bewegungsoperation den extraponierten Relativsatz aus seiner kanonischen Position innerhalb der NP heraus nach rechts bewegt hat (vgl. Beispiel 3).
- Dabei bleibt eine Spur (trace) zurück, die für die semantische Interpretation benötigt wird.
- Alternative: Der extraponierte Relativsatz wurde bei einer Bewegung nach links zurückgelassen (gestrandet).
- Vertreter des Bewegungsansatzes: u.a. Ross (1967); Akmajian (1975); Baltin (1978).
- In deklarativen Grammatiken wird nicht-lokale Merkmalsperkolation als “Bewegungssimulation” eingesetzt (z.B. Keller 1995).

2.2 Diskontinuitätsansatz

- Unterscheidung zwischen tektogrammatischer und phänogrammatischer Struktur (Curry 1961, Dowty 1996).
- Tektogrammatische Struktur: stark hierarchisch, wichtig für die kompositionelle Semantik
- Phänogrammatische Struktur: weniger hierarchisch, Wortstellungsdomänen, wichtig für die lineare Abfolge der Wörter im Satz
- Zusammenfassung mehrerer tektogrammatischer Konstituenten zu einer phänogrammatischen Konstituente ermöglicht phonologisch diskontinuierliche Phrasen.
- Vertreter des Diskontinuitätsansatzes vor allem in der HPSG: u.a. Nerbonne (1994); Kathol und Pollard (1995).
- (4) und (5) sind Beispiele für mögliche phänogrammatische Strukturen aus Kathol und Pollard (1995):

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

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

- In (4) ist die NP als ganzes Teil der Wortstellungsdomäne der VP und kann daher nur als ganzes linearisiert werden.
- In (5) stehen der Relativsatz und der Rest der NP separat in der Wortstellungsdomäne und können daher diskontinuierlich realisiert werden.
- Eine Linearisierungsregel der Form: [] < Relativsatz führt für Beispiel (5) zur extraponierten Stellung.

2.3 Anaphorischer Ansatz

- Keine Annahme einer zugrundeliegenden, kanonischen *syntaktischen* Position des extraponierten Relativsatzes und daher keine syntaktische Rekonstruktion der zugrundeliegenden Struktur.

→ Relativ unkomplizierte syntaktische Struktur

- Stattdessen Verlagerung des Problems der Diskontinuität auf die Semantik
- Integration des extraponierten Relativsatzes mittels anaphorischer Prozesse
- Plausible Analyse, da Relativsätze in vielen Sprachen von Demonstrativ- oder Interrogativpronomina eingeleitet werden.
- Vertreter: Wittenburg (1987), Stucky (1987), Kiss (2003)

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

- Kiss (2003) zeigt, dass Relativsatzextraposition bestimmten Bewegungsbeschränkungen nicht unterworfen ist. → Argument gegen den Bewegungsansatz

(7) * Man hat den Überbringer der Mitteilung beschimpft, dass die Erde rund ist.

(8) Man hat den Überbringer der Mitteilung beschimpft, die alle schockiert hat.

(9) * Hier habe ich bei den Beobachtungen faul auf der Wiese gelegen, dass die Erde rund ist.

(10) Hier habe ich bei den Beobachtungen faul auf der Wiese gelegen, die alle schockiert haben.

- Kiss (2003) argumentiert gegen den Diskontinuitätsansatz, weil dieser Extraposition als rein phonologisches Phänomen auffasst und deshalb keine Interaktionen mit Variablenbindung modellieren kann.

(11) Wir haben niemandem_i die Frage gestellt, auf die er_i sich vorbereitet hatte.

(12) * Wir haben die Frage niemandem_i gestellt, auf die er_i sich vorbereitet hatte.

- Kiss (2003) schlägt ein Prinzip der “verallgemeinerten Modifikation” vor:

[...] 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 Integration des Relativsatzes in die NP erfolgt über den von dem Nomen projizierten semantischen Index.

3 Relevanz des Problems für die Computerlinguistik

Lohnt es sich überhaupt, sich im Rahmen der Computerlinguistik mit der Relativsatzextraposition zu beschäftigen? Ist dies nicht ein relativ exotisches Phänomen?

Untersuchung von Uszkoreit et al. (1998):

- 12 % von insgesamt 11.996 vollständig annotierten Sätzen im Korpus mit Relativsatz (1394)
- Davon 24 % nachweisbar extrapониert (340) und 32 % nachweisbar nicht extrapониert (449) → 2,38 % aller Sätze enthalten extrapониerte RS

4 Computerlinguistische Umsetzungen der theoretischen Lösungsansätze

Obwohl extrapониerte Relativsätze im Deutschen sehr häufig vorkommen, gibt es laut Crysmann (2004) nur eine umfangreiche HPSG-Grammatik des Deutschen, die Sätze mit extrapониerten Relativsätzen parsen kann: Stefan Müllers Babel-Grammatik (Müller 1996).

4.1 Bewegungsansatz

- Grammatiken und Parser vorhanden (jedoch nur mit Bewegungssimulationen)
- Parsing ist sehr ineffizient, da an jeder möglichen Position in der Kette eine leere Kategorie, nämlich eine Spur, postuliert werden muss.
- Wenn man mehrere extrapониerte Relativsätze pro Nomen erlaubt, dann werden ohne spezielle Heuristiken unendlich viele leere Kategorien angesetzt und der Parser terminiert nicht mehr.

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)

→ Ansatz wirft Probleme für Standard-Bottom-Up-Parser auf.

4.2 Diskontinuitätsansatz

- Parsing mit dem Diskontinuitätsansatz ist ein viel komplexeres Problem als das Parsing herkömmlicher kontext-freier Grammatiken.
- Die lineare Reihenfolge der Wörter ist nicht mehr direkt im Strukturbaum ablesbar.
- Konstituenten können diskontinuierlich sein und somit Löcher enthalten.
- Parsingkomplexität im schlimmsten Fall mindestens exponentiell (Müller 2004).
- Das einzige umfangreiche System, Müllers Babel-Grammatik (Müller 1996), kann in puncto Geschwindigkeit nicht mit neueren Parsern konkurrieren.

→ Höhere Anforderungen an den Parser, komplexeres Parsingproblem.

4.3 Anaphorischer Ansatz

- Stellt keine zusätzlichen Anforderungen an den Parser.
- Keine Verletzung der Kontextfreiheit.

→ Bereits existierende effiziente Parser können verwendet werden.

- Hohe Effizienz des Ansatzes von Kiss (2003) wurde von Crysmann (2004) gezeigt.
- Datengetriebener Ansatz

→ Hohe Effizienz mit herkömmlichen Parsern, aber Probleme mit Scheinambiguitäten

5 Vorschlag einer getrennten Verarbeitung

Meine Dissertation baut auf dem anaphorischen Ansatz nach Kiss (2003) auf, da dieser theoretisch fundiert und sehr effizient ist und eine getrennte Verarbeitung von (extraponiertem) Relativsatz und dem Rest des Satzes erlaubt.

- Zweistufiges Verarbeitungskonzept:
 - Separate Analyse von (extraponiertem) Relativsatz und dem Rest des Satzes (nach Trennung durch geeignetes Verfahren)
 - Integration beider Strukturen mit einem speziell für diese Aufgabe entwickelten Modul

5.1 Effizienteres Parsing

- Eine geteilte Vorverarbeitung komplexer Sätze kann noch einmal einen Effizienzgewinn erbringen (Foth und Menzel 2003; Frank et al. 2003).
 - Ein globales Problem wird in lokale Subprobleme zerlegt, zwischen denen kaum Abhängigkeiten bestehen (Foth und Menzel 2003, S. 92).
 - Eine Kombination von flachen Parsingverfahren mit tiefen Parsingverfahren kann zu einer Steigerung der Verarbeitungseffizienz führen (Frank et al. 2003).
 - Durch die grammatikexterne Rekombination beider Teile kann das Problem der Scheinambiguitäten mittels geeigneter Heuristiken vermieden werden.

→ Weitere Effizienzgewinne gegenüber der von Crysmann (2004) implementierten Variante des Ansatzes von Kiss (2003) sind zu erwarten.

5.2 Chance zur Behandlung des Anbindungsproblems

- Die getrennte Verarbeitung ermöglicht die Benutzung eines speziellen Moduls zur Anbindung des (extrapolierten) Relativsatzes.
 - Dieses Modul kann *regelbasiertes* und/oder *statistisches* Wissen benutzen, um geeignete Antezedenten für den (extrapolierten) Relativsatz zu bestimmen.
 - Die geeigneten Antezedenten könnten sogar in *eine Rangfolge* gebracht werden, vom wahrscheinlichsten Antezedens zum unwahrscheinlichsten.
- Die Disambiguierung der Anbindung des (extrapolierten) Relativsatzes ist somit ein zweites wichtiges Ziel.
- Dabei plane ich auf Verfahren und Ergebnisse aus folgenden Bereichen der theoretischen Linguistik und der Computerlinguistik zurückzugreifen:
 - Anaphernresolution (siehe Mitkov 1999 für einen Überblick)
 - PP-Anbindungsdisambiguierung (z.B. Hindle und Rooth 1993)
 - Disambiguierung der Anbindung von Relativsätzen (z.B. Siddharthan 2002)
 - Psycholinguistische Studien zur Relativsatzanbindung (z.B. Uszkoreit et al. 1998 und Hemforth et al. 2000)
 - Eigene Korpusuntersuchungen und möglicherweise auch psycholinguistische Experimente zur Relativsatzanbindung
- Ziel ist die möglichst fehlerfreie (und dabei effiziente) Bestimmung der korrekten Anbindung (d.h. des korrekten Antezedenten) von extrapolierten Relativsätzen.

6 Konkrete Schritte

- Aufbau eines Korpus von Beispielsätzen mit extrapolierten Relativsätzen und Auswahl eines geeigneten zusammenhängenden Korpus
- Überprüfung der theoretischen Ansätze, insbesondere des anaphorischen Ansatzes nach Kiss (2003) an Hand dieser authentischen Daten
- Modifikation der B-Ger-Gram¹, einer von Stefan Müller auf dem TRALE-System (Carpenter et al. 2003) implementierten HPSG-Grammatik für die getrennte Verarbeitung von (extrapoliertem) Relativsatz und dem Rest des komplexen Satzes
- Entwicklung eines geeigneten Verfahrens zur Trennung von Relativsatz und Rest des komplexen Satzes:
 - heuristischer Chunker
 - topologischer Parser (Becker und Frank 2002)
- Implementation eines einfachen Rekombinationsmoduls
- Vergleichende Evaluation des Systems mit einer direkten Implementation des Ansatzes von Kiss (2003)
- Empirische Untersuchungen und Auswertung der Literatur zum Thema Relativsatzanbindung
- Entwicklung von Verfahren zur Disambiguierung der Relativsatzanbindung
- Evaluation dieser Verfahren mit Hilfe von handannotierten Korpora

¹<http://www.cl.uni-bremen.de/Fragments/b-ger-gram.html>

7 Relevanz für andere Bereiche der Linguistik und mögliche Zusatzfragestellungen

Relevanz meiner Arbeit für andere Bereiche der Computerlinguistik und der theoretischen Linguistik:

- Überprüfung der theoretischen Analysen an Hand natürlicher Daten
- Testen eines modularen Systems zur getrennten Verarbeitung von Satzfragmenten
 - Solche Systeme sind möglicherweise auch für andere Problemstellungen interessant.
 - Generelle Frage, ob mit solchen Systemen ein zusätzlicher Effizienzgewinn erzielt werden kann.
- Entwicklung von Verfahren zur Disambiguierung von Relativsätzenbindung
 - Ermittlung von relevanten Faktoren (“Gewicht”, Informationsstruktur, usw.), die auch für ähnliche Probleme relevant sein könnten.
 - Durch die Untersuchung von Korpora können sich auch für die theoretische Linguistik und die Psycholinguistik interessante empirische Ergebnisse zur Relativsätzenbindung ergeben.
- Mögliche Ausweitungen der Untersuchung:
 - Freie Relativsätze
 - Andere extraponierte Phrasen
 - Extraposition in anderen Sprachen als dem Deutschen

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.
- 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.
- 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/>.
- 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.
- 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.
- 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. 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.
- 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.
- 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. (2002). 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.
- 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.