Komplexität von Anfragen für syntaktisch annotierte Korpora: Wh-Bewegung im Negra Korpus und in der Penn Treebank

Laura Kallmeyer, SFB 441, Universität Tübingen

Kurzfassung

Dieser Artikel behandelt die Komplexität von Anfragesprachen für syntaktisch annotierte Korpora. Insbesondere wird am Negra Korpus und an der Penn Treebank untersucht, ob sich die den beiden Korpora zugrundeliegenden unterschiedlichen Annotationsschemata auf die Komplexität von Anfragen auswirken. Am Beispiel von langen wh-Bewegungen werden mögliche Anfragen für beide Korpora verglichen. Es stellt sich heraus, daß trotz der Unterschiede in den Datenstrukturen in beiden Fällen die gleiche Anzahl von Variablen, d.h. die gleiche Quantorenschachtelungstiefe benötigt wird, um in einer Logik erster Ordnung eine adäquate Anfrage zur Extraktion von langen wh-Bewegungen zu formulieren. Dieses Ergebnis wird mit Hilfe von Ehrenfeucht-Fraïssé-Spielen bewiesen.

1 Lange wh-Bewegungen in Negra und Penn Treebank

Syntaktisch annotierte Korpora nehmen meist eine baumartige Datenstruktur als Annotation an. Bäume sind allerdings nicht ausreichend, wenn man Argumentstrukturen kodieren möchte, da dies bei diskontinuierlichen Konstituenten zu Konflikten führt. Aus diesem Grund wurde sowohl für das Saarbrücker Negra Korpus ([8]) als auch für die Penn Treebank ([1, 6]) die zugrundeliegende Baumstruktur etwas variiert bzw. mit zusätzlichen Koindizierungsrelationen angereichert. Im Negra Korpus sind überkreuzende Kanten erlaubt, und in der Penn Treebank gibt es Spuren, die mit bewegten Elementen koindiziert sein können. Diese beiden Möglichkeiten, die der Annotation zugrundeliegende Datenstruktur zu variieren, sollen im folgenden genauer verglichen und auf Komplexitätsunterschiede der jeweils benötigten Anfragesprachen hin untersucht werden.

Als Beispiel sollen lange wh-Bewegungen wie in (1) im Negra Korpus und in der Penn Treebank betrachtet werden. Dieses Phänomen ist charakteristisch für den Unterschied zwischen den beiden Annotationsschemata und daher für einen Vergleich besonders geeignet. In (1) ist jeweils das Fragepronomen *wen* bzw. *whom* ein Argument des eingebetteten Satzes. ((1)b. ist die englische Übersetzung von (1)a.) Die Standardanalyse in Government Binding (wie in [2]) nimmt an, dass das Fragepronomen aus dem eingebetteten Satz herausbewegt wurde. Im folgenden soll der Begriff "Bewegung" jedoch nicht im Sinne einer bestimmten Theorie sondern rein deskriptiv verstanden werden.

(1) a. Wen$_i$ meinst du liebt Peter t_i?
b. Whom$_i$ do you think Peter loves t_i?

Um Konstruktionen wie in (1) in einem Korpus finden zu können, benötigt man eine Annotation, die in irgendeiner Form Argumentstrukturen kodiert, z.B. durch Verwendung von gekreuzten Kanten wie in Negra oder durch Einführen von Spuren wie in der Penn Treebank. **Abb. 1** zeigt, wie die

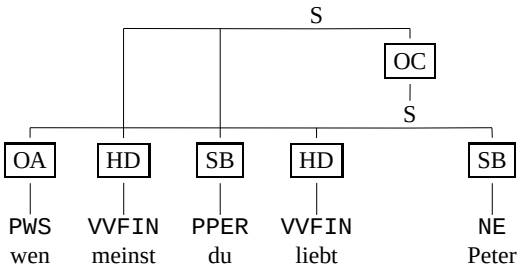

Abbildung 1: Annotation von (1)a. in Negra

Annotation von (1)a. im Negra Korpus aussieht. Als Part-of-Speech Tagset wurde in Negra das Stuttgart-Tübingen Tagset (STTS, [7]) verwendet, und PWS ist das POS-Tag für "substituierendes Interrogativpronomen". Das Fragepronomen steht zwar links von dem Matrixsatz, sein direkter Mutterknoten ist jedoch der Knoten des eingebetteten Satzes mit Kategorie S, und in diesem Satz hat das Fragepronomen die Funktion des Akkusativobjekts (OA).

```
(SBARQ (WHNP-1 whom)
  (SQ do
    (NP-SBJ you)
    (VP think
     (S (NP-SBJ Peter)
      (VP loves
        (NP *T*-1))))))
```

Abbildung 2: Annotation von (1)b. in der Penn Treebank

Abb. 2 zeigt die Annotation von (1)b. in der Penn Treebank. Diese Annotation nimmt an, dass die Ursprungposition der bewegten Nominalphrase in der VP des eingebetteten Satzes ist, daher gibt es hier eine Spur, ein leeres Element. Diese Spur ist mit dem Fragepronomen am Anfag des Satzes koindiziert. Im Gegensatz zu Negra wird in der Penn Treebank mehr als nur das Ergebnis der Bewegung kodiert,

da explizit eine Ursprungsposition für das bewegte Element festgelegt werden muss. Insofern ist Negra gewissermaßen theorieneutraler als das Penn Treebank Annotationsschema.

2 Anfragen für lange wh-Bewegungen

Eine Anfrage ist eine partielle Strukturbeschreibung, die in einer bestimmten Logik formuliert wird. Mit Hilfe dieser Anfrage können Strukturen aus einem Korpus extrahiert werden, nämlich die Strukturen, die die Anfrage in einem modelltheoretischen Sinn erfüllen. Zur Extraktion benötigt man natürlich ein geeignetes Suchprogramm (siehe z.B. [5]), dies ist aber nicht Gegenstand dieses Artikels.
Betrachtet man die Annotationen in Abb. 1 und Abb. 2, so könnte man vermuten, daß es aufwendiger ist, eine Anfrage nach wh-Bewegungen im Negra Korpus zu formulieren als eine entsprechende Anfrage für die Penn Treebank, da es bei der Penn Treebank über die Relation der Koindizierung eine direkte Verbindung zwischen Spur und bewegtem Element gibt. Im Negra Korpus ist dies nicht der Fall. Es gibt keine Spur, und man muß in der Anfrage die strukturelle Konfiguration beschreiben, die indiziert, daß ein wh-Element bewegt wurde.
Mögliche Anfragen in einer Logik erster Ordnung, um wh-Bewegungen zu extrahieren, sind in (2) für Negra und in (3) für die Penn Treebank gezeigt. In (3) wird eine Relation ind für Koindizierung angenommen. Die Mutter-Tochter Beziehung wird durch $\lhd$ und lineare Präzedenz durch $\prec$ denotiert, und ein Superskript "*" bezeichnet die reflexive transitive Hülle. Die Anfrage für die Penn Treebank hat nur vier Variablen während die für das Negra Korpus fünf Variablen enthält.

$$
\begin{array}{ll}
(2) & \exists x_1 \exists x_2 \exists x_3 \exists x_4 \exists x_5 \\
 & \quad [\mathsf{PWS}(x_1) \wedge x_2 \lhd x_1 \wedge \mathsf{OC}(x_4, x_2) \wedge \\
 & \quad \mathsf{HD}(x_4, x_3) \wedge x_1 \prec^* x_3 \wedge \\
 & \quad \mathsf{HD}(x_2, x_5) \wedge x_3 \prec^* x_5]
\end{array}
$$

$$
\begin{array}{ll}
(3) & \exists x_1 \exists x_2 \exists x_3 \exists x_4 \\
 & \quad [\mathsf{WHNP}(x_1) \wedge \mathsf{*T*}(x_2) \wedge ind(x_1, x_2) \wedge \\
 & \quad \mathsf{SBARQ}(x_3) \wedge x_3 \lhd x_1 \wedge \mathsf{S}(x_4) \wedge \\
 & \quad x_3 \lhd^* x_4 \wedge x_4 \lhd^* x_2]
\end{array}
$$

In beiden Fällen ist eine Reduzierung der Variablenzahl möglich (siehe (4) für Negra und (5) für Penn Treebank), und es zeigt sich, dass trotz der unterschiedlichen Annotationen für beide Anfragen drei Variablen genügen.

$$
\begin{array}{lll}
(4) & \exists x \exists y [\quad [\exists z & (\mathsf{PWS}(z) \wedge z \lhd x \wedge \\
 & & \quad z \prec^* y)] \\
 & \wedge \quad [\exists z & (\mathsf{OC}(z, x) \wedge \mathsf{HD}(z, y))] \\
 & \wedge \quad [\exists z & (\mathsf{HD}(x, z) \wedge y \prec^* z))]]
\end{array}
$$

$$
\begin{array}{lll}
(5) & \exists x \exists y [\quad [\exists z & (\mathsf{WHNP}(z) \wedge \mathsf{SBARQ}(y) \wedge \\
 & & \quad \mathsf{*T*}(x) \wedge y \lhd^* z \wedge \\
 & & \quad ind(z, x))] \\
 & \wedge \quad [\exists z & (\mathsf{S}(z) \wedge y \lhd^* z \wedge z \lhd^* x)]]
\end{array}
$$

Dass drei Variablen in beiden Fällen das Minimum ist, wird im folgenden mit Hilfe von Ehrenfeucht-Fraïssé Spielen (siehe [3]) bewiesen, einer Technik aus der endlichen Modelltheorie (vgl. auch [4] zu einer weiteren Anwendung).

3 Ehrenfeucht-Fraïssé Spiele

Ehrenfeucht-Fraïssé Spiele dienen dazu, Resultate über die Ausdrucksstärke bestimmter Logiken zu beweisen.
Die zugrundeliegende Idee ist wie folgt: Um festzustellen, ob eine gegebene Anfrage in einer bestimmten Logik ausgedrückt werden kann, wählt man zwei Strukturen über gleichen Signaturen, so dass nur eine der beiden die Anfrage erfüllt. Zwei Spieler spielen mit k Paaren von Spielsteinen. In jedem Zug wählt Spieler 1 ein Paar Steine aus und markiert mit dem einen der beiden ein Element in einer der Strukturen. Spieler 2 markiert anschließend ein anderes Element mit dem zweiten Stein des Paares. Markiert Spieler 1 ein Element, so entspricht dies einer existentiellen Quantifizierung. Wird ein Stein zum ersten Mal eingesetzt, so entspricht dies der Verwendung einer neuen Variablen. Spieler 1 versucht, seine Steine so zu setzen, daß sich die beiden von den markierten Elementen erzeugten Substrukturen unterscheiden, während Spieler 2 versucht, die beiden Substrukturen isomorph zu halten. Gelingt Spieler 2 dies in jedem Zug, so gewinnt er, und dies bedeutet, dass der Unterschied zwischen den beiden Strukturen mit nur k Variablen nicht ausgedrückt werden kann.

Definition 1 (Ehrenfeucht-Fraïssé Spiel)
Ein Ehrenfeucht-Fraïssé Spiel wird von zwei Spielern S und D mit k Paaren von Spielsteinen auf zwei Strukturen G und H gespielt. In jedem Zug wählt S (der "spoiler") ein Spielsteinpaar. Er setzt einen der beiden Steine auf ein Element in einer der beiden Strukturen. D (der "duplicator") setzt anschließend den anderen Stein des gewählten Paares auf ein Element in der jeweils anderen Struktur. D gewinnt eine Runde, wenn die von den markierten Elementen in G und H induzierten Substrukturen anschließend isomorph sind. D gewinnt das gesamte Spiel, wenn er jede einzelne Runde gewinnt.

Für die Anfragen in (4) und (5) wird im folgenden Abschnitt gezeigt werden, dass es jeweils Spiele gibt mit zwei Spielsteinpaaren, bei denen D eine Gewinnstrategie hat. Die verwendeten Strukturen sind dabei so gewählt, dass sie theoretisch im Korpus (zumindest als Teilstruktur) auftreten könnten. Dies bedeutet, dass in beiden Fällen zwei Variablen nicht genügen, um die Anfrage auszudrücken, die ja einen Unterschied zwischen den gewählten Strukturen beschreibt.

4 Drei Variablen als Minimum

4.1 Negra

Satz 1 *Um die Anfrage in (4) in einer Logik erster Ordnung auszudrücken über Strukturen mit binären Relationen*

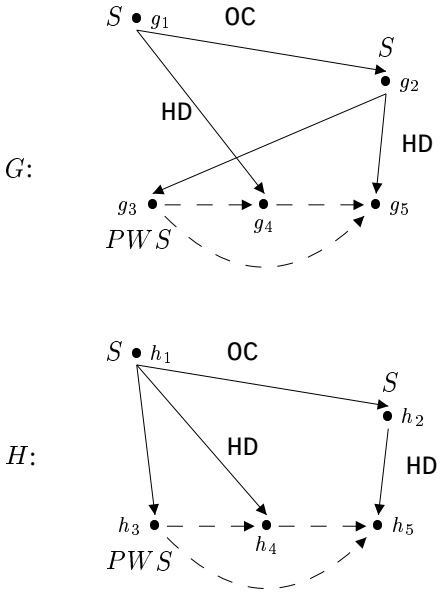

Abbildung 3: Graphen für Anfrage (4)

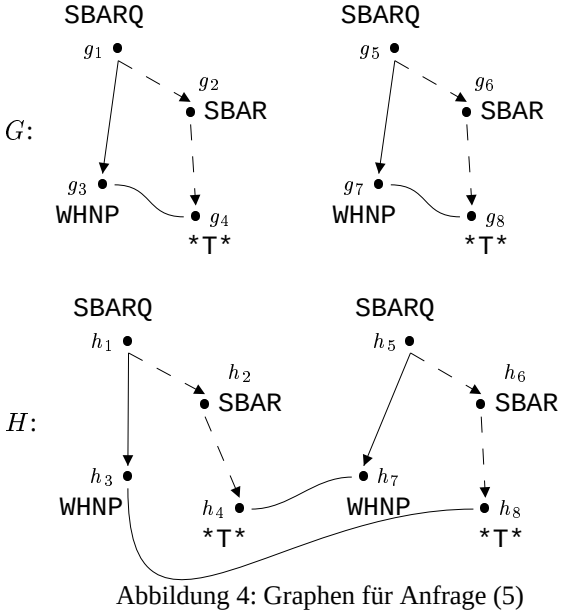

Abbildung 4: Graphen für Anfrage (5)

unmittelbare Dominanz (Mutter-Tochter-Beziehung) und lineare Präzedenz und mit Knoten- und Kantenkategorien, benötigt man mindestens drei Variablen.

Dieser Satz wird bewiesen mit Hilfe eines Ehrenfeucht-Fraïssé Spiels mit 2 Spielsteinpaaren. **Abb. 3** zeigt die beiden Strukturen G und H, auf denen das Spiel gespielt wird. Die gerichteten durchgezogenen Kanten stehen für unmittelbare Dominanz, während die gestrichelten gerichteten Kanten für lineare Präzedenz stehen. G erfüllt die Anfrage (4), während H dies nicht tut.
Bei drei Spielsteinpaaren gibt es auf G und H eine Gewinnstrategie für S, nämlich die von (4) widergespiegelte: In den ersten drei Zügen markiert S g_2, g_4 und g_3, im vierten Zug bewegt S den zuletzt verwendeten Stein auf g_1 und im fünften Zug bewegt er diesen Stein auf g_5. Spätestens dann hat S gewonnen.

Beweis von Satz 1 Es wird ein Ehrenfeucht-Fraïssé Spiel mit zwei Spielsteinpaaren auf G und H in Abb. 3 gespielt. Der Spieler D hat eine einfache Gewinnstrategie.
Wir definieren zunächst eine bijektive Relation R zwischen den Knoten in G und den Knoten in H:

$$R := \{\langle g_1, h_2\rangle, \langle g_2, h_1\rangle, \langle g_3, h_3\rangle, \langle g_4, h_5\rangle, \langle g_5, h_4\rangle\}$$

Die Gewinnstrategie von D sieht dann wie folgt aus: Für $1 \leq i \leq 5$: Setzt S einen Stein auf g_i, so antwortet D, indem er einen Stein auf $R(g_i)$ setzt. Setzt S einen Stein auf h_i, so setzt D als Antwort einen Stein auf $R^{-1}(h_i)$.
Diese Strategie führt dazu, dass nach jeder Runde eines der folgenden Strukturpaare in G und H markiert sind:

$\langle\{g_1, g_2\}, \{h_2, h_1\}\rangle$ $\langle\{g_1, g_3\}, \{h_2, h_3\}\rangle$
$\langle\{g_1, g_4\}, \{h_2, h_5\}\rangle$ $\langle\{g_1, g_5\}, \{h_2, h_4\}\rangle$
$\langle\{g_2, g_3\}, \{h_1, h_3\}\rangle$ $\langle\{g_2, g_4\}, \{h_1, h_5\}\rangle$
$\langle\{g_2, g_5\}, \{h_1, h_4\}\rangle$ $\langle\{g_3, g_4\}, \{h_3, h_5\}\rangle$
$\langle\{g_3, g_5\}, \{h_3, h_4\}\rangle$ $\langle\{g_4, g_5\}, \{h_5, h_4\}\rangle$

Wie leicht überprüft werden kann, sind die beiden Strukturen eines Paares jeweils isomorph.
Damit ist gezeigt, dass sich der Unterschied zwischen G und H (und somit die Anfrage (4)) mit nur zwei Variablen nicht ausdrücken lässt.

□

4.2 Penn Treebank

Satz 2 *Um die Anfrage in (5) in einer Logik erster Ordnung auszudrücken über Strukturen mit binären Relationen unmittelbare Dominanz, Dominanz und Koindizierung und mit Knotenkategorien, benötigt man mindestens drei Variablen.*

Dieser Satz wird ebenfalls durch ein Ehrenfeucht-Fraïssé Spiel mit zwei Spielsteinpaaren gezeigt, und zwar auf den Strukturen in **Abb. 4**. Die gerichteten durchgezogenen Kanten stehen wieder für unmittelbare Dominanz, die gerichteten gestrichelten Kanten für Dominanz und die ungerichteten Kanten für die (reflexive) Koindizierungsrelation. G erfüllt die Anfrage in (5), während H dies nicht tut.
In einem Spiel auf G und H mit drei Spielsteinpaaren hat S eine Gewinnstrategie, nämlich die Strategie, die von der Anfrage (5) widergespiegelt wird: Zunächst setzt S einen Stein auf g_4, dann einen g_1 und dann einen dritten auf g_3, und als letztes bewegt er den dritten Stein g_2.

Beweis von Satz 2 Es wird ein Ehrenfeucht-Fraïssé Spiel mit $k = 2$ auf G und H in Abb. 4 gespielt.
Wir betrachten zweielementige Teilstrukturen von G und H. Diese Strukturen werden in Äquivalenzklassen eingeteilt (bzgl. Knotenkategorien, unmittelbarer Dominanz, Dominanz und Koindizierung). Die Elemente der Teilstrukturen seien geordnet. Notation: $S(g_i, g_j)$ bezeichnet die Teilstruktur mit g_i als erstem und g_j als zweitem Element, $S(h_i, h_j)$ die Teilstruktur mit h_i als erstem und h_j als zweitem Element ($1 \leq i, j \leq 8$ und $i \neq j$).
Die Äquivalenzklassen zweielementiger geordneter Teilstrukturen von G und H sind:

1. $\{S(g_1, g_2), S(g_5, g_6), S(h_1, h_2), S(h_5, h_6)\}$
2. $\{S(g_1, g_3), S(g_5, g_7), S(h_1, h_3), S(h_5, h_7)\}$
3. $\{S(g_1, g_4), S(g_5, g_8), S(g_1, g_8), S(g_5, g_4),$
   $S(h_1, h_4), S(h_5, h_8), S(h_1, h_8), S(h_5, h_4)\}$
4. $\{S(g_1, g_5), S(g_5, g_1), S(h_1, h_5), S(h_5, h_1)\}$
5. $\{S(g_1, g_6), S(g_5, g_2), S(h_1, h_6), S(h_5, h_2)\}$
6. $\{S(g_1, g_7), S(g_5, g_3), S(h_1, h_7), S(h_5, h_3)\}$
7. $\{S(g_2, g_3), S(g_2, g_7), S(g_6, g_3), S(g_6, g_7),$
   $S(h_2, h_3), S(h_2, h_7), S(h_6, h_3), S(h_6, h_7)\}$
8. $\{S(g_2, g_4), S(g_6, g_8), S(h_2, h_4), S(h_6, h_8)\}$
9. $\{S(g_2, g_6), S(g_6, g_2), S(h_2, h_6), S(h_6, h_2)\}$
10. $\{S(g_2, g_8), S(g_6, g_4), S(h_2, h_8), S(h_6, h_4)\}$
11. $\{S(g_3, g_4), S(g_7, g_8), S(h_3, h_8), S(h_7, h_4)\}$
12. $\{S(g_3, g_7), S(g_7, g_3), S(h_3, h_7), S(h_7, h_3)\}$
13. $\{S(g_3, g_8), S(g_7, g_4), S(h_3, h_4), S(h_7, h_8)\}$
14. $\{S(g_4, g_8), S(g_8, g_4), S(h_4, h_8), S(h_8, h_4)\}$

Die Gewinnstrategie von Spieler D in einem Spiel mit zwei Spielsteinpaaren sieht wie folgt aus:
Setzt S im ersten Zug einen Stein auf das i-te Element der einen Struktur ($1 \leq i \leq 8$), so setzt D seinen Stein im Antwortzug auf das i-te Element der anderen Struktur.
In der zweiten Runde antwortet D auf den jeweiligen Zug von S so dass

(a) die beiden Teilstrukturen isomorph sind und
(b) die in der ersten Runde markierten Elemente entweder beide die ersten Elemente der Teilstrukturen oder beide die zweiten Elemente der beiden Teilstrukturen sind.

(Es gibt häufig mehr als eine Möglichkeit für D.) Dies ist möglich für Spieler D, da für alle Äquivalenzklassen C und für alle i mit $1 \leq i \leq 8$ und alle k, $1 \leq k \leq 2$ folgendes gilt: es gibt eine Struktur in C mit g_i als k-tes Element gdw. es in C eine Struktur mit h_i als k-tem Element gibt.
In jeder weiteren Runde kann D immer so antworten, dass (a) und (b) weiterhin gegeben sind. Dies ist der Fall, da für alle Äquivalenzklassen C_1, C_2 und alle k, $1 \leq k \leq 2$ gilt: wenn eine der Teilstrukturen in C_1 durch ändern des k-ten Elements überführt werden kann in eine Struktur in C_2, dann können alle Teilstrukturen in C_1 durch Ändern des k-ten Elements in Strukturen in C_2 überführt werden.

□

5 Zusammenfassung

Wie hier gezeigt wird, wirkt sich die unterschiedliche Annotation in Negra und Penn Treebank nicht auf die Anzahl der Variablen aus, die man benötigt, um in einer Logik erster Ordnung Anfragen zur Extraktion langer wh-Bewegungen zu stellen. Die minimale Variablenanzahl entspricht der benötigten Quantorenschachtelungstiefe und ist ein entscheidender Komplexitätsfaktor. Das Phänomen der langen wh-Bewegungen ist zwar nur ein Beispiel, aber es ist charakteristisch für den Unterschied zwischen den beiden Annotationsschemata, und daher ist das Resultat von allgemeiner Bedeutung für einen Vergleich zwischen Negra und Penn Treebank.
Man benötigt allerdings wahrscheinlich einen Quantor mehr für die Anfrage im Negra Korpus gegenüber der Penn Treebank. Dies liegt daran, dass die Penn Treebank eine reichere, allerdings auch theorieabhängigere Annotation aufweist. Aber da man die gleiche Anzahl Variablen benötigt, kann man vermuten, dass sich der Vorteil der größeren Theorieneutralität in Negra nicht wesentlich auf die Komplexität der benötigten Anfragesprache auswirkt.

Literatur

[1] A. Bies, M. Ferguson, K. Katz & R. MacIntyre. Bracketing Guidelines for Treebank II Style Penn Treebank Project. University of Pennsylvania, 1995.

[2] L. Haegeman. *Introduction to Government and Binding Theory*. Blackwell, Oxford UK, Cambridge USA, 2. Auflage, 1994.

[3] N. Immerman. *Descriptive Complexity*. Graduate texts in computer science. Springer, New York, 1999.

[4] L. Kallmeyer. On the Complexity of Queries for Structurally Annotated Linguistic Data. In *Proceedings of ACIDCA'2000*, S. 105–110, März 2000.

[5] L. Kallmeyer. A query tool for syntactically annotated corpora. In *Workshop on Syntactic Annotation of Electronic Corpora*, Tübingen, Juni 2000. Abstract.

[6] M. Marcus, G. Kim, M.A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz & B. Schasberger. The Penn Treebank: Annotating Predicate Argument Structure. In *ARPA '94*, 1994.

[7] A. Schiller, S. Teufel & C. Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. Manuskript Universität Stuttgart und Universität Tübingen, 1995.

[8] W. Skut, B. Krenn, T. Brants & Hans Uszkoreit. An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*, Washington, D.C., 1997.