Robustes Chunkparsing mit variabler Analysetiefe*

Sandra Kübler, Erhard W. Hinrichs
Seminar für Sprachwissenschaft, Eberhard-Karls-Universität Tübingen

Abstract

Das Chunkparsing [1], [2] bietet einen besonders vielversprechenden Ansatz zum robusten, partiellen Parsing mit dem Ziel einer breiten Datenabdeckung. Ziel beim Chunkparsing ist eine partielle, nicht-rekursive syntaktische Struktur. Dieser extrem effiziente Parsing-Ansatz läßt sich als Kaskade endlicher Transducer realisieren.
In diesem Beitrag wird TüSBL vorgestellt, ein System, bei dem die Eingabe aus spontaner, gesprochener Spache besteht, die dem Parser in Form eines Worthypothesengraphen aus einem Spracherkenner zur Verfügung gestellt wird. Chunkparsing ist für eine solche Anwendung besonders geeignet, da es fragmentarische oder nicht wohlgeformte Äußerungen robust behandeln kann.
Des weiteren wird eine Baumkonstruktionskomponente vorgestellt, die die partiellen Chunkstrukturen zu vollständigen Bäumen mit grammatischen Funktionen erweitert. Das System wird anhand manuell überprüfter Systemeingaben evaluiert, da sich die üblichen Evaluationsparameter hierfür nicht eignen.

1 Einleitung

Die gegenwärtige Forschung zum Parsen natürlicher Sprache ist vom Spannungsfeld zwischen flacher bzw. partieller Strukturanalyse (mit dem Ziel einer breiten Datenabdeckung) einerseits und tiefer und möglichst vollständiger Strukturanalyse (unter Inkaufnahme einer engen Datenabdeckung) andererseits gekennzeichnet.
Das Chunkparsing [1], [2] stellt einen viel diskutierten und besonders erfolgversprechenden Ansatz für das effiziente Parsen großer Textmengen mit breiter Datenabdeckung dar. Die Strategie des Chunkparsing besteht darin, die Analyse nicht-rekursiver Teilkonstituenten vom Parsen größerer, rekursiver syntaktischer Einheiten zu separieren. Dies ermöglicht eine effiziente Chunkparsingarchitektur als Kaskade endlicher Transducer mit *rightmost, longest-match* Strategie [2].
Die bisherige Literatur zum Chunkparsing weist jedoch drei Lücken auf:

- Bisherige Studien zum Chunkparsing gehen von schriftsprachlichen Textkorpora als Eingabeketten aus. Zusätzliche Schwierigkeiten ergeben sich, wenn es sich bei der Eingabe um die von einem akustischen Spracherkenner verarbeiteten Worthypothesen spontansprachlicher Äußerungen handelt.
- Es gibt bisher keine Untersuchungen darüber, wie partielle Chunkanalysen zu Strukturanalysen chunkübergreifender, rekursiver syntaktischer Einheiten miteinander verbunden werden können.
- Es liegen bislang keine quantitativen Studien darüber vor, welche Datenabdeckung mit dem Chunkparsing erzielt werden kann.

Ziel dieses Beitrags ist es, diese Forschungslücken zu schließen und das System TüSBL (Tübingen Similarity Based Learning) vorzustellen.
Den Forschungskontext der hier beschriebenen Studie bildet das BMBF Verbundprojekt Verbmobil, das die maschinelle Übersetzung spontansprachlicher Äußerungen zwischen den Sprachen Deutsch, Englisch und Japanisch zum Ziel hat.
Die Anwendungsdomäne liegt in den Bereichen Terminvereinbarungen, Reiseplanung und PC-Wartung. Die nachfolgenden Beispiele sind daher aus dieser Anwendungsdomäne entnommen, die beschriebenen Techniken jedoch domänenunabhängig und generell anwendbar.

2 Parsing mit variabler Analysetiefe

TüSBL wurde in einer dreistufigen Systemarchitektur angelegt, um ein effizientes und robustes Parsing zu

* Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 701 M0 und von der Deutschen Forschungsgemeinschaft im Rahmen des Sonderforschungsbereich 441 gefördert. Die Verantwortung für den Inhalt der Arbeit liegt bei den Autoren.

ermöglichen. In den drei Stufen werden jeweils genau spezifizierte Teilprobleme angegangen: In der ersten Stufe wird das Part-of-Speech Tagging, d.h. die Zuweisung der Wortarten mittels des LIKELY-Taggers [5] durchgeführt; als Tagset dient das Stuttgart-Tübingen-Tagset (STTS) [9]. Die so ermittelten POS Tags dienen als Eingabe für den Chunkparser [1], [2], der an die systemspezifischen Anforderungen angepaßt wurde. Die dort ermittelte Chunkstruktur dient dann als Eingabe in die Baumkonstruktion, wo die Chunkstrukturen soweit möglich zu kompletten Bäumen erweitert werden. Eine mangelnde Abdeckung in der Chunk- oder in der Baumkonstruktionskomponente führt nicht zu einem Systemabbruch, die fragliche (Teil-)Struktur wird unverändert an die nächste Stufe weitergegeben.

2.1 Chunkparsing mit Spracherkenner-Input

In natürlich-sprachlichen Systemen, deren Input aus gesprochener Sprache besteht, wird dem syntaktischen Parsing die automatische Erkennung von Einzelwörtern vorgeschaltet. Der Output des Spracherkenners für die Weiterverarbeitung in einer NLP Anwendung besteht üblicherweise aus einem Worthypothesengraphen. Der syntaktische Parser hat dann die Aufgabe, aus diesem Graphen die gemäß einer zugrundeliegenden Grammatik beste Hypothese für die Gesamtstruktur der Äußerung zu ermitteln.

Im Falle von Verbmobil werden aus dem Worthypothesengraphen, den der Spracherkenner liefert, die n-besten Ketten von Worthypothesen für die Gesamtäußerung gebildet und dem Chunkparser als Eingabeketten zur Verfügung gestellt. Diese n-besten Ketten sind mit all jenen Fehlerquellen behaftet, die für akustische Spracherkenner tpyisch sind: das fehlerhafte Einfügen oder "Verschlucken" von kurzen Wörtern aus geschlossenen Wortklassen (z.B. Präpositionen, Artikel, Interjektionen), das "Verschlucken" von gebundenen Morphemen bzw. das fehlerhafte Abbilden von unbekannten Wörtern auf bekannte Wörter.

Ein typisches Beispiel einer derartig fehlerbehafteten Satzhypothese aus der Verbmobildomäne ist die Zeichenkette *ich könnte Ihnen ja aber zum Beispiel ein Dienstag mich den zwölften anbieten*, in der die Wörter *ein* und *mich* vom Spracherkenner eingefügt worden sind. Trotz des fehlerhaften Inputs gelingt es dem Chunkparser, eine partielle Analyse zu liefern:

Eingabe: *ich könnte Ihnen ja aber zum Beispiel ein Dienstag mich den zwölften anbieten*

Chunkausgabe:

```
[simpx
        [nx2     [pper       ich]]
        [vxfin   [vmfin      könnte]]
        [nx2     [pper       Ihnen]]
        [ptkant  ja]
        [advx    [adv        aber]]
        [px      [appr       zum]
                 [nx2        [nn        Beispiel]]]
        [nx3     [art        ein]
                 [nx2        [nn        Dienstag]]]
        [nx2     [pper       mich]]
        [nx4     [art        den]
                 [adja       zwölften]]
        [vvinf   anbieten]]
```

Wenn der Chunkparser eine fehlerhafte Hypothese zu verarbeiten hat, liefert er, wie im obigen Beispiel mehrere, lokal grammatische Teilstrukturen. Bei diesen Teilstrukturen handelt es sich um "islands of certainty" im Sinne von Abney [2], die sich für die Weiterverarbeitung durch weitere Systemmodule (im Fall von Verbmobil: die semantische Analyse und die maschinelle Übersetzung) eignet. Ein Chunkparser liefert somit einen wichtigen Beitrag zu einem robusten Sprachverarbeitungssystem, das auch defizitären Input weiterverarbeiten kann.

2.2 Ähnlichkeitsbasierte Baumkonstruktion

Die Baumkonstruktion basiert auf dem Lernverfahren des *memory-based reasoning* [10], das mit großem Erfolg auf diverse NLP Klassifikationsaufgaben angewandt worden ist, darunter POS Tagging, Graphem-Phonem-Konvertierung, Wortbedeutungsdisambiguierung und PP Attachment [4], [12], [13]. Dieser Ansatz geht davon aus, daß die Verarbeitung aktueller Information durch den Vergleich mit gespeicherter Information über andere, bereits gesehene Strukturen erfolgt. Es handelt sich dabei um "lazy learning" in dem Sinne, daß über gespeicherte Instanzen nicht wie in regel-basierten Systemen abstrahiert wird, sondern die Instanzen wie vorgefunden mit der aktuellen Eingabe abgeglichen werden. Im vorliegenden Fall besteht die Datengrundlage der bereits gesehenen Instanzen aus einer syntaktisch annotierten Baumbank [11], mit ca. 60.000 Bäumen[1]. Das Ähnlichkeitsmaß für die Baumkonstruktion berechnet sich aus den vorkommenden

[1] Die deutsche Baumbank ist in ca. 38.000 Dialogabschnitte (dialog turns) gegliedert, die jeweils aus einem oder mehreren Teilbäumen bestehen. TüSBL wurde für die Sprachen Deutsch und Englisch trainiert, für Englisch liegt eine Baumbank mit ca. 35.000 Bäumen vor [7].

Lexemen, den zugewiesenen POS-Tags, der Segmentierung in Chunks und der den Chunks zugewiesenen Kategorien.
In den Fällen, in denen der Spracherkenner dem Chunkparser gute Hypothesen für eine syntaktisch wohlgeformte Gesamtstruktur liefert, lassen sich die gechunkten Teilstrukturen im weiteren Verarbeitungsschritt der Baumkonstruktion zu satzüberspannenden Gesamtstrukturen verbinden. Für die folgende Eingabekette, zum Beispiel, liefert der Chunkparser zunächst folgende Teilstrukturen:

Eingabe: *genau ich sehe grade der würde zurückfliegen von Hannover nach München um sechzehn Uhr fünf zum Beispiel*

Chunkausgabe:

```
[simpx  [adjx   [adjd     genau]]
        [nx2    [pper     ich]]
                [vxfin    [vvfin   sehe]]
                [advx     [adv     grade]]
                [nx2      [art     der]]
                [vxfin    [vafin   würde]]]
[simpx  [vvinf  zurückfliegen]
                [px       [appr    von]
                  [nx2    [ne      Hannover]]]
                [px       [appr    nach]
                  [nx2    [ne      München]]]
                [px       [appr    um]
                  [time   [card    sechzehn]
                          [nn      Uhr]
                          [card    fünf]]]
                [px       [appr    zum]
                  [nx2    [nn      Beispiel]]]]
```

Diese Chunks werden dann in **Abbildung 1** zu Gesamtstrukturen zusammengefügt[2]. In dieser Struktur sind die folgenden neuen Arten von Information enthalten:

- Teilweise können POS Tags im größeren Kontext korrigiert werden. Bei der Baumkonstruktion hat sich z.B. herausgestellt, daß das als Artikel (art) getaggte Wort *der* eine Phrase für sich bildet, daher wird das Tag in Demonstrativpronomen (pds) geändert.
- Komplexe Nominal- oder Präpositionalphrasen, die wegen des nicht-rekursiven Charakters des Chunkparsers auf Chunkebene nicht erkannt werden können, werden in der Baumkonstruktion richtig gruppiert. So wird aus den beiden px-Chunks *von Hannov*er und *nach München* eine komplexe PX gebildet.
- Eine Unzulänglichkeit im Chunkparse ist das Fehlen von grammatischen Funktionen. Diese Unterspezifiziertheit ist auf der Chunkebene beabsichtigt, da sie es erlaubt, auch Korrekturen u.ä. in die Satzstruktur einzubinden.Wird jedoch eine komplette Baumstruktur angestrebt, ist die grammatische Funktion der Hauptkonstituenten eine unerläßliche Information. In Abbildung 1 sind die grammatischen Funktionen als Kantenlabels dargestellt, ON kennzeichnet das Subjekt, HD den Head des Satzes, MOD einen ambigen Modifikator, dessen Anbindung im Satz unterspezifiziert ist, OV ein verbales Objekt, OS einen Objektsatz und V-MOD einen Modifikator des Verbs.
- Weiterhin sind im Baum topologische Felder [6] eingesetzt worden. Diese Information ist direkt unter der Simplex-Ebene annotiert. Die Aufteilung des Simplex-Satzes in Felder ermöglicht eine flache Baumstruktur.

Ein maschinelles Lernverfahren läßt sich jedoch nicht ohne Modifikation anwenden. Betrachtet man die Erweiterung der Chunkstruktur zu kompletten Bäumen als Klassifizierung, erhält man eine extrem große Anzahl von Klassen, da jeder Baum aus der Baumbank eine eigene Klasse darstellen würde. Außerdem wäre die Generalisierungsfähigkeit des Verfahrens gleich null, da nur bereits gesehene Bäume klassifiziert werden könnten. Aus diesem Grund wurde in TüSBL ein *backing-off* Ansatz realisiert. Dieser Ansatz wird anhand des folgenden Beispiels exemplarisch dargestellt:

Eingabe: *ja das ist bei mir leider schlecht da bin ich auf einer Messe schon*

Chunkausgabe:

```
[simpx  [ptkant  ja]
        [nx2     [pds     das]]
        [vafin   ist]
        [px      [appr    bei]
                 [nx2     [pper    mir]]]
        [adjx    [adv     leider]
                 [adjd    schlecht]]
        [adv     da]]
[simpx  [vafin   bin]
        [nx2     [pper    ich]]
        [px      [appr    auf]
                 [nx2     [art     einer]
                          [nn      Messe]]]
        [advx    [adv     schon]]]
```

[2] Die Bäume in dieser und den folgenden Abbildungen folgen dem Format, das im NEGRA Projekt des SFB 378 an der Universität des Saarlandes entwickelt wurde [3]. Die Bäume wurden mit dem Annotate-Tool [8] ausgedruckt.

Im ersten Schritt werden durch einen Vergleich auf der lexikalischen Ebene "direkte Treffer" gesucht, d.h. Bäume, bei denen eine vollständige Übereinstimmung bei den Klassifikationsmerkmalen gefunden wird. Für die obige Eingabe ist dies für die erste Chunkstruktur der Fall, die Teilbäume sind in **Abbildung 2** dargestellt. Wie in Abbildung 2 gezeigt, werden in diesem Schritt Simplex-Chunks u.U. auch aufgeteilt, wenn keine Evidenz für den kompletten Chunk vorhanden ist, dafür aber direkte Treffer für Teilchunks gefunden werden. Auffällig ist hier, daß die Adverbialphrase *da* einen separaten Baum erhält. Dies liegt daran, daß wegen der Ambiguität von *da* beim Chunkparsing die longest-match Strategie greift und das Adverb daher dem ersten Chunk zugeordnet wird. Da außerdem ein direkter Treffer für *da* gefunden wurde, wird eine Gruppierung des Adverbs und des zweiten Chunks nicht versucht. Diese partiellen Bäume können im Verbmobil System in einem späteren Schritt zusammengefügt werden.

Da der zweite Chunk nicht als direkter Treffer gefunden wurde, wird in einem zweiten Schritt versucht, ebenfalls durch lexikalischen Vergleich, einen Teilbaum zu finden. Dies gelingt auch, der gefundene Baum umfaßt die Verbalphrase und das Subjekt *bin ich*. In einem dritten Schritt wird aufgrund eines Vergleichs der POS Tags eine Erweiterung des bisher gefundenen Baums versucht. Auf diese Weise kann die Präpositionalphrase *auf einer Messe* angehängt werden, wodurch sich der in Abbildung 2 gezeigte Baum ergibt. Lediglich für das nachgestellte *schon* wird keine Erweiterung gefunden; diese Adverbialphrase wird in einem weiteren Schritt mit interner Struktur versehen, wenn nach Teilphrasen in den Bauminstanzen gesucht wird.

Trotz des hier betriebenen Suchaufwands ist die Verarbeitung sehr effizient, da auch in diesem Schritt des Systems deterministisch gesucht wird. Da jedoch, im Vergleich zu anderen Parsern, die nur Beziehungen zwischen Müttern und direkten Töchtern betrachten, ein größerer Ausschnitt des Baumes zur Verfügung steht, ist die Qualität der Ausgabestrukturen sehr hoch.

In Fällen, in denen der Chunkparser jedoch nur defizitären Input erhält, kann i.a. nur eine partielle Analyse erfolgen. In solchen Fällen ist die Anzahl der partiellen Strukturen deutlich höher als bei syntaktisch wohlgeformten Eingaben. **Abbildung 3** zeigt das Parsing-Ergebnis für das oben genannte Beispiel *ich könnte Ihnen ja aber zum Beispiel ein Dienstag mich den zwölften anbieten.* Hier kann nur der erste Teil der Äußerung in eine partielle Simplex-Struktur gruppiert werden, die restliche Äußerung wird nur auf Phrasenebene analysiert. Dennoch eignen sich die auch in solchen Fällen analysierten Teilstrukturen zur Weiterverarbeitung, z.B. in der Maschinellen Übersetzung.

3 Evaluation

Für die quantitative Evaluation eines Chunkparsers sind die gängigen Kriterien von Precision und Recall hinsichtlich eines vollständig annotierten Testkorpus nur bedingt geeignet. Dies liegt zum einen an den häufig partiellen Ausgabestrukturen des Chunkparsers. Da eine partielle Analyse einer wahrscheinlich falschen vorgezogen wird, sollten dem Parser eben solche partiellen Strukturen nicht angelastet werden. Andererseits beruhen die traditionell ermittelten Precision und Recall Werte auf dem Abgleich mit einem eindeutig bestimmbaren und grammatisch wohlgeformten Input. Dies läßt sich auf die Evaluation des Chunkparsers, der häufig defizitären und nur begrenzt voraussagbaren Input weiterzuverarbeiten hat, nicht ohne weiteres übertragen. Im Fall von TüSBL müßten daher tatsächliche, aus dem Spracherkenner stammende Eingaben mit manuell transliterierten Sprachdaten aus der Baumbank verglichen werden. Ein Test mit Testdaten aus der Baumbank erscheint nicht sinnvoll, da dies kaum Aufschlüsse über die tatsächliche Performanz im System erlaubt. **Tabelle 1** enthält Evaluationsparameter, die für eine quantitative Evaluation eines Chunkparsers geeignet erscheinen.

durchschnittl. Anzahl v. Wörtern pro Satz	13,8
inkorrekter oder unvollständiger WHG	48,1 %
korrekte Teilbäume	93,2 %
korrekte Bäume bei korrektem Input	54,8 %

Tabelle 1 Quantitative Evaluation

Die in Tabelle 1 genannten Zahlen wurden aus einem Korpus von vier Verbmobil Testdialogen mit 81 Turns ermittelt, die den Verbmobil-Projektpartnern in Form von Spracherkennerausgaben als Testmaterial zur Verfügung gestellt wurden.

Daß der Input in der Tat in hohem Maße defizitär ist, läßt sich daran ablesen, daß der Worthypothesengraph (WHG) fast für jeden zweiten Satz unvollständig oder fehlerhaft ist. Bei dieser hohen Fehlerrate ist allerdings auch die durchschnittliche Satzlänge von fast 14 Wörtern zu berücksichtigen. Die beiden letzten Werte beziehen sich auf die Präzision der Ausgabestrukturen. Der hohe Prozentsatz korrekter Teilbäume belegt die Robustheit des Chunkparsers, die darin besteht, auch bei inkorrektem oder unvollständigem Input korrekte Teilstrukturen zu liefern. Die Tatsache, daß der Chunkparser bei wohlgeformten Input für mehr als jeden zweiten Satz einen korrekten, äußerungsüberspannenden Baum liefert, zeigt, daß partielle Chunkanalysen zu Strukturanalysen chunkübergreifender, rekursiver syntaktischer Einheiten miteinander verbunden werden können. Dies bedeutet, daß sich ein Chunkparser mit Gewinn auch bei solchen Anwendungen einsetzen läßt, in denen eine partielle Strukturanalyse nicht ausreicht.

4 Schlußbemerkung

Der Nachweis, daß mit diesem Verfahren aus Chunks vollständig(er)e Baumstrukturen erzeugt werden können, ist insofern bedeutsam, als es den Kreis der möglichen Anwendungsfelder für einen Chunkparser signifikant erweitert. In der bisherigen Literatur wird der Chunkparser ausschließlich im Zusammenhang mit solchen Anwendungen genannt, bei denen eine partielle und robuste Analyse ausreicht. Der hier beschriebene Einsatz des Chunkparsers zeigt auf, daß ein um eine Baumkonstruktionskomponente erweiterter Chunkparser sich sehr wohl auch für Anwendungen eignet, bei denen eine robuste variable Analysetiefe notwendig ist. Im Gegensatz zu stochastischen Parsern, die zum Ziel haben, immer eine Gesamtstruktur per Äußerung zu generieren, ist die Baumkonstruktion des Chunkparsers flexibler und erlaubt nach wie vor partielle Strukturen als Ausgabe. Strukturen werden nur zu größeren Einheiten zusammengefügt, wenn das Ähnlichkeitsmaß dafür hinreichende Evidenz liefert. Damit verfährt auch der erweiterte Chunkparser nach der Abneyschen Maxime des "islands of certainty"-Parsing [2] und liefert im Vergleich zu einem stochastischen Parser Strukturen mit einem hoheren Gütemaß an Korrektheit.

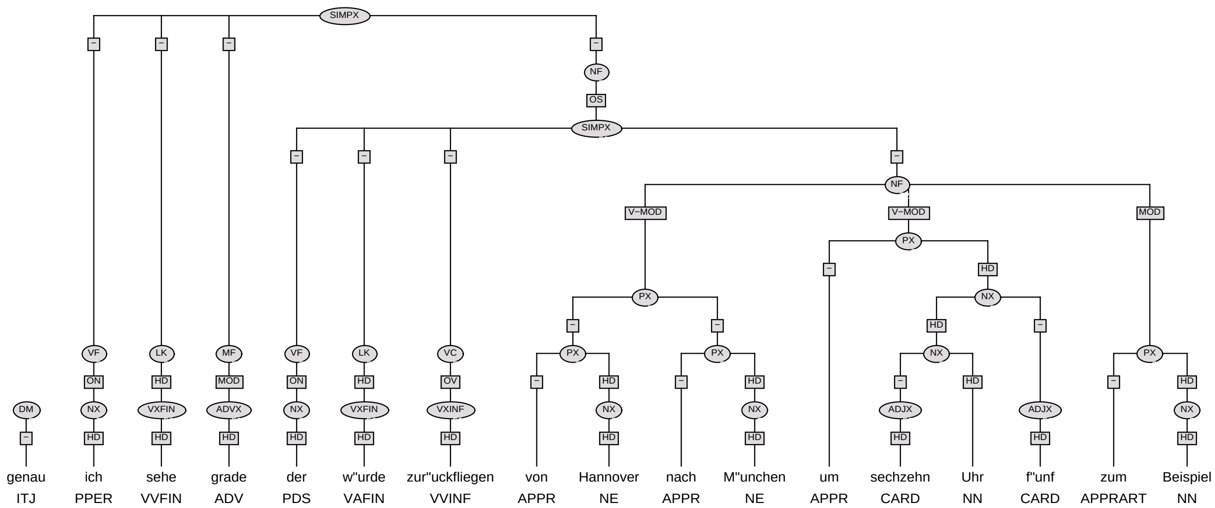

Abbildung 1 Ein aus Chunks zusammengefügter Baum

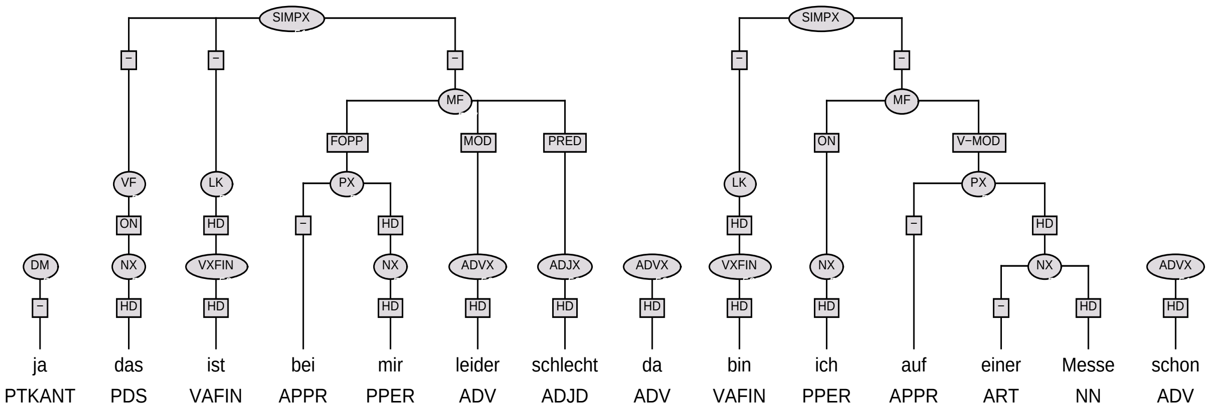

Abbildung 2 Baum nach der Baumkonstruktion

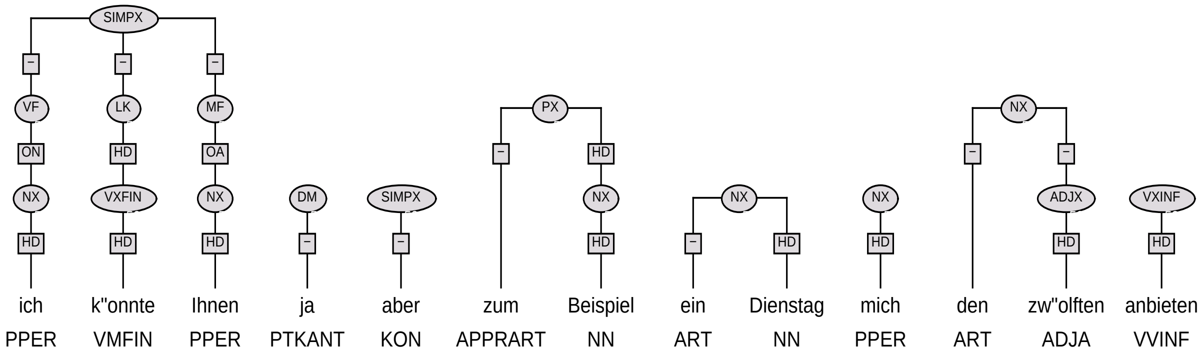

Abbildung 3 Ein unvollständiger Baum

5 Literatur

[1] Abney, S.: Parsing By Chunks. In: R. Berwick, S. Abney und C. Tenny (Hrsg.), Principle-Based Parsing. Dordrecht: Kluwer Academic Publishers, 1991

[2] Abney, S.: Partial Parsing via Finite-State Cascades. In: Proceedings of the ESSLLI '96 Robust Parsing Workshop, 1996

[3] Brants, T.; Skut, W.: Automation of treebank annotation. NeMLaP-3/CoNLL98, Sydney, Australien, 1998, pp. 49-58

[4] Daelemans, W.; van den Bosch, A.; Zavrel, J.: Forgetting exceptions is harmful in language learning. Machine Learning 34 (1999), pp. 11-43

[5] Feldweg, H.: Stochastische Wortartendisambiguierung mit dem robusten System LIKELY. Eberhard-Karls-Universität Tübingen, 1993

[6] Höhle, T. N.: Der Begriff "Mittelfeld". Anmerkungen über die Theorie der topologischen Felder. In: A. Schöne (Hrsg.), Kontroversen alte und neue. Akten des 7. Internationalen Germanistenkongresses Göttingen, 1985, S. 329-340

[7] Kordoni, V.: Stylebook for the English Treebank in VERBMOBIL. Eberhard-Karls-Universität Tübingen, 1998

[8] Plaehn, O.: *Annotate* – Bedienungsanleitung, Universität des Saarlandes, FR 8.7 Computerlinguistik, Projekt C3 Nebenläufige Grammatische Verarbeitung, Sonderforschungsbereich 378, Ressourcenadaptive Kognitive Prozesse, 1998

[9] Schiller, A.; Teufel, S.; Thielen, C.: Guidelines für das Tagging deutscher Textcorpora mit STTS. Universitäten Stuttgart und Tübingen, 1995

[10] Stanfill, C.; Waltz, D.: Toward memory-based reasoning. Communications of the ACM, 29(12), 1986, pp. 1213-1228

[11] Stegmann, R.; Telljohann, H. und Hinrichs, E.W.: Stylebook for the German Treebank in VERBMOBIL. Eberhard-Karls-Universität Tübingen, 1998

[12] Veenstra, J.; van den Bosch, A.; Buchholz, S.; Daelemans, W.; Zavrel, J.: Memory-based word sense disambiguation. Computers and the Humanities 34 (2000)

[13] Zavrel, J.; Daelemans, W.; Veenstra, J.: Resolving PP attachment ambiguities with memory-based learning. CoNLL97, Madrid