Erschienen in: Zeitschrift für Germanistische Linguistik (ZGL) Jg. 38 (2010) H. 2, S. 224-248.

Ines Rehbein

Der Einfluss der Dependenzgrammatik auf die Computerlinguistik

Abstract
In 1959, Lucien Tesnière wrote his main work *Éléments de syntaxe structurale*. While the impact on theoretical linguistics was not very strong at first, 50 years later there exist a variety of linguistic theories based on Tesnière's work. In computational linguistics, as in theoretical linguistics, dependency grammar was not very influential at first. The last 10–15 years, however, have brought a noticeable change and dependency grammar has found its way into computational linguistics. Syntactically annotated corpora based on dependency representations are available for a variety of languages, as well as statistical parsers which give a syntactic analysis of running text describing the underlying dependency relations between word tokens in the text. This article gives an overview of relevant areas of computational linguistics which have been influenced by dependency grammar. It discusses the pros and cons of different types of syntactic representation used in natural language processing and their suitability as representations of meaning. Finally, an attempt is made to give an outlook on the future impact of dependency grammar on computational linguistics.

0. Einleitung

Vor 50 Jahren schrieb Lucien Tesnière, der als Begründer der modernen Dependenzgrammatik gilt, sein Hauptwerk *Éléments de syntaxe structurale*. Seitdem haben sich viele Ausprägungen der Dependenzgrammatik entwickelt, die sich zuweilen deutlich unterscheiden. Allen gemeinsam ist jedoch die Annahme, dass sich syntaktische Strukturen als binäre, asymmetrische Beziehungen zwischen lexikalen Elementen analysieren lassen, als sogenannte Dependenzrelationen. Während der Einfluss der Dependenzgrammatik nicht nur auf die theoretische Linguistik, sondern auch auf die Computerlinguistik anfänglich eher gering war, lässt sich hier in

den letzten 10–15 Jahren eine Veränderung ausmachen. Die Repräsentation von syntaktischen Informationen in der Form von Dependenzen hat ihren Einzug in den Bereich der Verarbeitung natürlicher Sprache gehalten.

Die Kodierung von syntaktischen Strukturen in Form von Dependenzrelationen wurde in vielen Bereichen der Verarbeitung natürlicher Sprache aufgegriffen. Allen voran ist das Gebiet der Korpuslinguistik zu nennen, in welchem mit der Erstellung von syntaktisch annotierten Korpora im Rahmen der Dependenzgrammatik begonnen wurde. Die Existenz von Dependenzbaumbanken wiederum ermöglichte die Generierung von probabilistischen Sprachmodellen, die in vielen Gebieten der Verarbeitung natürlicher Sprache eingesetzt wurden, vor allem im Bereich des syntaktischen Parsings. Dort fand die wohl fruchtbarste Entwicklung statt, die sich einerseits in einer Weiterentwicklung der Parsingtechnologie selbst manifestierte, andererseits aber auch bestimmt war durch eine grundsätzliche Diskussion über die Frage, welche Informationen ein syntaktischer Parser zurückgeben soll, und wie eine geeignete Form der Evaluation aussehen sollte.

Die Entwicklung von robusten, statistischen Dependenzparsern, die auf neuen, unbekannten Texten eine breite Abdeckung erzielen, ermöglichte es, syntaktische Informationen in Form von Dependenzrelationen als Komponenten in verschiedene Systeme zur Verarbeitung natürlicher Sprache zu integrieren. Beispielhaft genannt seien Anwendungsbereiche wie Question Answering, Semantic Role Labelling, Diskursparsing oder Maschinelle Übersetzung. Aber auch für Gebiete der Linguistik wie z.B. die Zweitspracherwerbsforschung bildeten sich innovative Herangehensweisen heraus, wie die Erstellung von Lernerkorpora, in denen die von den Lernenden entwickelte Interimsprache, die sogenannte Interlanguage, auf Basis von Dependenzen repräsentiert wird.

In diesem Artikel soll untersucht werden, auf welche Bereiche der Verarbeitung natürlicher Sprache die Dependenzgrammatik prägend Einfluss nahm, und wie heute, 50 Jahre nach Begründung der modernen Dependenzgrammatik, die bisherige und weitere Entwicklung einzuschätzen ist. Dazu wird zuerst die Entwicklung in der Korpuslinguistik beschrieben, die durch die Bereitstellung von Ressourcen in Form von Dependenzbaumbanken die Voraussetzung für weitere Forschung vor allem im Bereich des syntaktischen Parsings bot. Anschließend werden ein Überblick über den aktuellen Forschungsstand des Dependenzparsings gegeben und wichtige Unterschiede zum Parsen auf Basis von Phrasenstrukturgrammatiken herausgearbeitet. Darauf aufbauend wird auf Fragen der Evaluation eingegangen, die zu einer grundsätzlichen Diskussion über angemessene Formen der Repräsentation von sprachlichem Wissen führten. Abschließend werden die wichtigsten Aspekte zusammengefasst und bewertet.

1. Dependenzen in der Korpuslinguistik

Anfang der 60er Jahre hielten digitale Korpora ihren Einzug in die theoretische Linguistik. Nicht-digitale Korpora waren schon lange vorher bekannt und wurden in verschiedenen Bereichen der theoretischen Linguistik auch viel genutzt, so z.B. in der Spracherwerbsforschung (Preyer, 1889; Stern, 1924), der Feldforschung des Amerikanischen Strukturalismus (Boas, 1940), in Studien zur Verteilung der Häufigkeit von Buchstabensequenzen in deutschen Texten (Kaeding, 1897) oder in der Lexikografie (Das Deutsche Wörterbuch, (Bahr, 1984)). Das Vorhandensein von digitalen, maschinell lesbaren Korpora eröffnete jedoch vorher ungeahnte Möglichkeiten des Zugriffs, der Verbreitung und der Arbeit mit diesen Ressourcen.

Die ersten digitalen Korpora enthielten vorwiegend geschriebene Texte in englischer Sprache, zum Teil ganz ohne linguistische Annotation, zum Teil mit Auszeichnung der Wortarten. Die frühen Korpora wurden sowohl in der theoretischen Linguistik als auch in der Computerlinguistik mit Misstrauen betrachtet. In der theoretischen Linguistik lag der Fokus auf der Sprachkompetenz, und die Untersuchung quantitativer Aspekte natürlicher Sprache, wie sie mit Hilfe von Korpora vorgenommen werden konnte, galt als uninteressant, wie das folgende Zitat von Chomsky (1969; Seite 57) belegt.

> It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term.

Doch auch in der Computerlinguistik, die diese Vorbehalte gegenüber quantitativen Methoden nicht teilte, wurden Korpora nicht sogleich akzeptiert. Das folgende Zitat von Kilgarriff and Grefenstette (2003; Seite 334) beschreibt die Ursachen hierfür.

> Corpora crashed into computational linguistics at the 1989 ACL meeting in Vancouver: but they were large, messy, ugly objects clearly lacking in theoretical integrity in all sorts of ways, and many people were skeptical regarding their role in the discipline.

Nach Überwindung dieser Anfangsschwierigkeiten konnten sich digitale Korpora nicht nur in der theoretischen Linguistik, sondern auch in der Computerlinguistik etablieren und waren bald als Ressource für die anwendungsorientierte statistische Sprachverarbeitung nicht mehr wegzudenken. Korpora erwiesen sich als nützliche Hilfsmittel zur Überprüfung von linguistischen Theorien (siehe die Arbeiten von Müller und Meurers (2006); Meurers und Müller (2009) zur mehrfachen Vorfeldbesetzung, zum Subjazenzprinzip oder zur Voranstellung von Verbpartikeln im Deutschen). Im Bereich der automatischen Verarbeitung natürlicher Sprache hingegen werden die Ressourcen zur Erstellung von statistischen Sprachmodellen und zum Trainieren und zur Evaluation von Systemen zur Verarbeitung natürlicher Sprache benötigt. Dies soll am Beispiel des Baumbank-basierten probabilistischen Parsens sowie der Statistischen Maschinellen Übersetzung illustriert werden.

In der Computerlinguistik steht der Begriff *Parsen* für die Analyse von Texten mit Hilfe eines automatischen Systems zur Verarbeitung natürlicher Sprache, dem *Parser*. Meist handelt es sich hier um die syntaktische Analyse von Daten, es gibt aber auch semantische Parser, die Argumentstruktur und semantische Rollen im Text auszeichnen (Gildea und Jurafsky, 2002; Pradhan u.a., 2008), oder Diskursparser, die einem Dokument eine Diskursstruktur zuweisen (Marcu, 1999; Sagae, 2009).

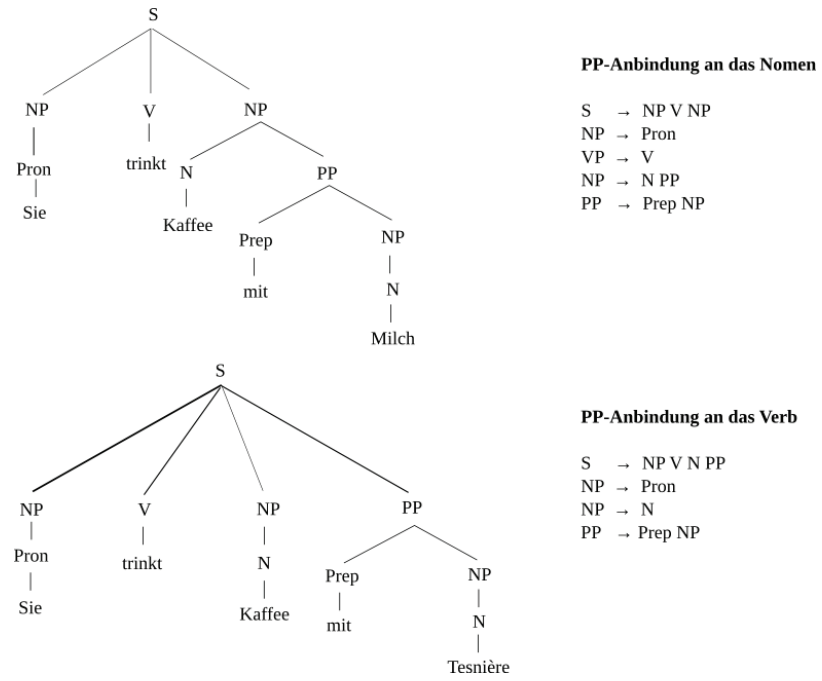

Abb. 1: Zwei Phrasenstrukturbäume mit den dazugehörigen Regeln

Baumbank-basiertes probabilistisches Parsen bezeichnet die syntaktische Analyse von Texten, bei der die dazu benötigte Grammatik nicht von Hand geschrieben, sondern direkt von den annotierten Syntaxbäumen einer Baumbank abgelesen wird (siehe Cahill (2008) für einen gut lesbaren Überblick über Baumbank-basiertes probabilistisches Parsen). Abbildung 1 zeigt zwei konstituentenbasierte Syntaxbäume mit den dazugehörigen Grammatikregeln. Jeder dieser Regeln wird eine Wahrscheinlichkeit zugewiesen, mit deren Hilfe zwischen verschiedenen möglichen Analysen disambiguiert werden kann. Die Wahrscheinlichkeiten basieren auf den Häufigkeiten, mit denen die syntaktischen Konstruktionen in der Baumbank vorkommen. Ist in den in der Baumbank enthaltenen Texten z.B. Nomen-Anbindung von PPs häufiger enthalten als Verb-Anbindung, so erhält die Regel $S \rightarrow NP\ V\ NP$ aus Abbildung 1 eine höhere Wahrscheinlichkeit als die zweite Regel S →

NP V NP PP. Dies kann dazu führen, dass ein und derselbe Parser, wenn er auf unterschiedlichen Baumbanken trainiert wurde, eine Präferenz für die eine oder andere syntaktische Konstruktion haben kann (Volk, 2006). Deshalb ist die Größe der Baumbank ein entscheidendes Kriterium für die Qualität der darauf trainierten automatischen Systeme, denn nur ein ausreichend großes Korpus stellt sicher, dass wichtige syntaktische Phänomene repräsentativ vertreten sind. Weitere wichtige Faktoren sind Genre und Domäne der in der Baumbank enthaltenen Texte, da diese einen großen Einfluss auf die syntaktische Struktur und damit auf die vom Parser gelernten Regeln haben. Während handgeschriebene Grammatiken mehrere Personenjahre an Arbeitszeit und ein großes Maß an linguistischer Expertise benötigen, kann mittels der automatischen Grammatikextraktion in kurzer Zeit eine robuste Grammatik gewonnen werden, die auf häufig vorkommenden syntaktischen Phänomenen gute Ergebnisse liefert und außerdem eine gute Abdeckung auf unbekannten Texten aufweist.

Als zweites Beispiel für die Verwendung von Baumbanken in der Verarbeitung natürlicher Sprache soll der Bereich der Statistischen Maschinellen Übersetzung genannt werden. Hier gibt es in den letzten Jahren einen immer stärkeren Trend, syntaktisches Wissen in den Übersetzungsprozess einzubeziehen. Das Vorhandensein von parallelen Baumbanken, die die Übersetzung eines Texts in verschiedenen Sprachen beinhalten, ermöglicht über die Alignierung von Worten hinaus die Verknüpfung von ganzen Phrasenpaaren (Abbildung 2). Basierend auf diesen Auszeichnungen kann die Häufigkeit und somit die Wahrscheinlichkeit des Auftretens geeigneter Kandidaten für die Übersetzung von Phrasen berechnet und in das statistische Modell mit einbezogen werden (Lavie u.a., 2008; Tinsley u.a., 2009). Hierfür werden ungleich größere Mengen an Daten benötigt als für das Baumbank-basierte probabilistische Parsen. Da der erforderliche Zeitaufwand es nicht zulässt, solche Datenmengen von Hand zu annotieren, werden automatische Systeme eingesetzt, die Baumbanken für parallele Texte erstellen und alignieren (Volk und Samuelsson, 2004; Zhechev und Way, 2008). Zwar sind die von den automatischen Systemen erzeugten Daten fehlerbehaftet, dennoch kann die Einbindung einer solchen syntaktischen Komponente zu einer Verbesserung der Übersetzungsqualität führen (Tinsley u.a., 2009).

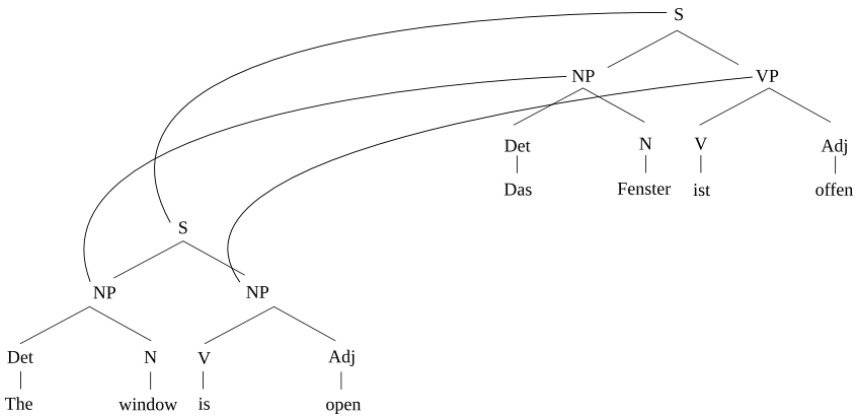

Abb. 2: Alignierung von Phrasen in einer parallelen Baumbank

1.1 Dependenzbaumbanken

Die Erstellung von syntaktisch annotierten Korpora, sogenannten Baumbanken, war eine der Kernvoraussetzungen für die Entwicklung von probabilistischen, datengesteuerten Parsern. Allerdings war die Mehrzahl aller frühen, syntaktischen Korpora auf der Basis von Phrasenstrukturgrammatiken ausgezeichnet worden.[1] Die bekannteste und wohl auch größte konstituentenbasierte Baumbank ist die englische Penn Treebank (PTB), ein im Rahmen der Phrasenstrukturgrammatik ausgezeichnetes syntaktisches Korpus mit Texten aus dem Wall Street Journal (Marcus u.a., 1993), das heute noch als Referenzbaumbank für das Englische gilt.

Das erste wirklich große Projekt zur Erstellung einer Dependenzbaumbank wurde in Prag durchgeführt, wo man, aufbauend auf der starken strukturalistischen Tradition, ein syntaktisch annotiertes Korpus mit verschiedenen Analyseebenen erstellte. Die Annotation der Prager Dependenzbaumbank folgt der Theorie der Funktionalen Generativen Beschreibung (Sgall u.a., 1986) und umfasst drei Ebenen der Annotation:

1. Morphologische Annotation (Morphologische Ebene)
2. Oberflächen-Syntax (Analytische Ebene)
3. Tiefenstruktur (Tektogrammatikalische Ebene).

Ähnlich wie nach dem Vorbild der englischen Penn Treebank eine Vielzahl von Konstituentenbaumbanken für weitere Sprachen erstellt wurden (so z.B. die Penn

1 Eine Ausnahme unter den frühen Baumbankprojekten ist die schwedische *Talbanken76* (Einarsson, 76a; Einarsson, 76b), deren Annotationsschema Konstituenten, Dependenzen und topologische Felder umfasst.

Chinese Treebank (Xue u.a., 2005) oder die Penn Arabic Treebank (Maamouri u.a., 2004)), so wirkte die Prager Dependenzbaumbank besonders im Bereich der slavischen Sprachen als Vorbild und Auslöser für zahlreiche Korpusprojekte. Beispiele hierfür sind die Kroatische Dependenzbaumbank (Tadić, 2007), die Russische Dependenzbaumbank (Boguslavsky u.a., 2000), und die Slovenische Dependenzbaumbank (Džeroski, 2006).

Doch auch andere, nicht-slavische Sprachen entdeckten Dependenzrelationen als eine vorteilhafte Form der syntaktischen Repräsentation. Für eine Vielzahl von europäischen Sprachen existieren inzwischen Dependenzbaumbanken. Stark vertreten sind die skandinavischen Länder, die sich im Nordischen Baumbank-Netzwerk (*Nordic treebank network*) zusammengeschlossen haben.[2] Aber auch für das Niederländische (Beek, 2002) und das Italienische (Bosco und Lombardo, 2004) sowie für einige asiatische Sprachen wie Japanisch (Lepage u.a., 1998), Hindi (Bharati u.a., 2002) oder Chinesisch (Liu, 2009) gibt es inzwischen Dependenzbaumbanken. Auch für das Arabische (Hajič u.a., 2004) wurde nach dem Vorbild der Prager Baumbank ein syntaktisch annotiertes Korpus erstellt. Weiter zu nennen sind die türkische (Oflazer u.a., 2003) und die baskische Dependenzbaumbank (Aduriz u.a., 2003) sowie mit Dependenzen annotierte Korpora für klassische Sprachen wie Latein (Bamman und Crane, 2006) und Griechisch (Bamman u.a., 2009).

1.2 Konstituenten versus Dependenzen in der Baumbankannotation

Die Arbeit an Dependenzbaumbanken fand vorwiegend im nicht-englisch-sprachigen Raum statt. Dies ist ebenso wenig ein Zufall wie die Tatsache, dass auch heute, 50 Jahre nach Erscheinen von Tesnière's *Éléments de syntaxe structurale* noch keine Übersetzung ins Englische vorliegt. Während sich das Englische mit seiner konfigurationellen Wortfolge im Rahmen der Phrasenstrukturgrammatik adäquat darstellen lässt, scheint für Sprachen mit einer variableren Wortfolge eine Repräsentation mit Hilfe von Dependenzrelationen angemessener zu sein. Dies soll anhand eines Beispiels illustriert werden. Vergleicht man den englischen Satz *Man bites dog* (Mann beißt Hund) in Abbildung 3 mit der deutschen Übersetzung, so zeigt sich, dass im Englischen die im Phrasenstrukturbaum enthaltenen Informationen ausreichen, um zwischen Subjekt und direktem Objekt zu unterscheiden. Die an den Satzknoten angehängte NP trägt die Subjekt-Funktion, während die an die VP angehängte NP das direkte Objekt des Satzes bildet. Im Deutschen jedoch kann sowohl die Subjekt-NP als auch die Objekt-NP sowohl an den Satz-

2 Nicht alle der nordischen Baumbanken sind reine Dependenzbaumbanken. Das norwegische Trepil-Projekt folgt der Lexikalisch-Funktionalen Grammatik, und die estnische Baumbank wurde im Rahmen der Constraint Grammar erstellt; jedoch spielen in beiden Theorien Dependenzen eine wichtige Rolle.

knoten als auch innerhalb der VP angehängt werden, da hier nicht die Wortstellung, sondern die morphologische Ausprägung der Wortformen entscheidungstragend ist. Deshalb ist eine Annotation, die die grammatikalische Funktion der Konstituenten unberücksichtigt lässt, für die Interpretation der Bedeutung des Satzes nicht hinreichend. Aus diesem Grund weisen alle drei deutschen Baumbanken eine hybride Form der Annotation auf, in der zusätzlich zur Konstituentenstruktur auch Dependenzrelationen annotiert werden.

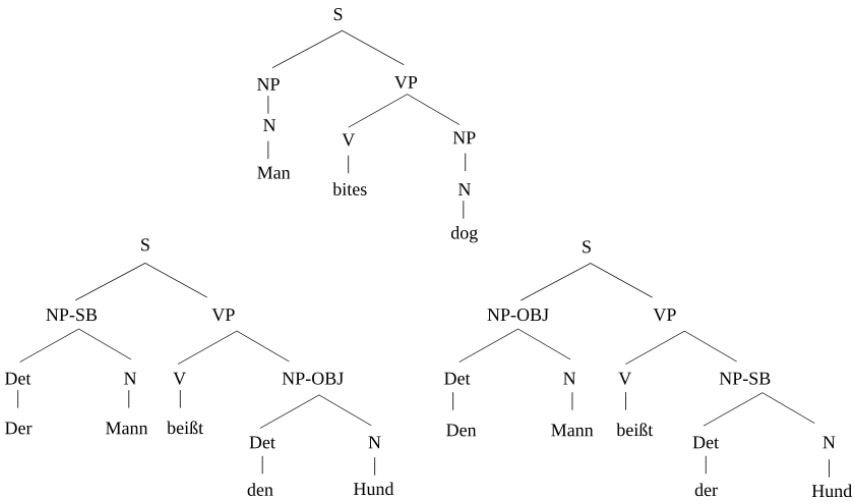

Abb. 3: Konfigurationelle Wortfolge im Englischen und variable Wortfolge im Deutschen

1.3 Hybride Baumbanken

In diesem Abschnitt soll auf hybride Formen von Baumbanken eingegangen werden, die Elemente der Phrasenstrukturgrammatik mit Dependenzinformationen vereinen. Nachdem anfänglich konstituentenbasierte Baumbanken für viele Sprachen die Norm darstellten, wurden in den letzten Jahren Dependenzrelationen als Repräsentation für syntaktische Strukturen immer populärer. Heute geht der Trend deutlich zu Baumbanken mit multiplen Repräsentationsformen. So beschreiben Xia u.a. (2009) die Anforderungen an die nächste Generation von Baumbanken als eine flexible Kombination von Konstituenten- und Dependenzrepräsentationen, mit der Möglichkeit zur automatischen Konvertierung in das jeweils gewünschte Format.

Beispiele für hybride Annotationsschemata, die Informationen aus beiden Welten vereinen, gibt es schon heute. Die Annotation in den drei deutschen Baumbanken, NEGRA (Skut u.a., 1997), TiGer (Brants u.a., 2002) und TüBa-D/Z

(Telljohann u.a., 2005), basiert auf einer Konstituentenstruktur, deren Kanten zusätzlich mit grammatikalischen Funktionslabeln versehen wurden. Trotz dieser Gemeinsamkeiten gibt es beträchtliche Unterschiede zwischen den Annotationsschemata von NEGRA/TiGer und TüBa-D/Z.

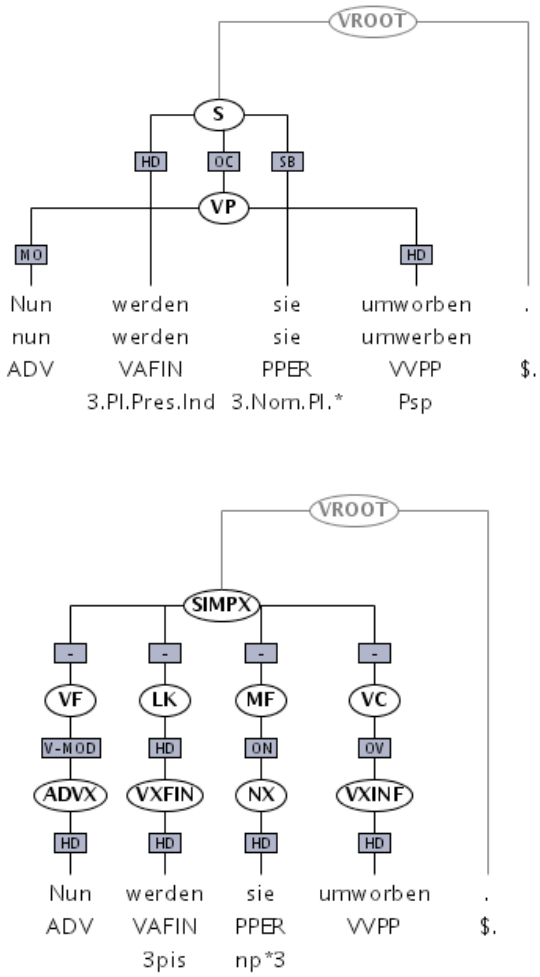

Abb. 4: Beispielbaum im TiGer- (oben) und TüBa-D/Z-Annotationsschema (unten); VROOT: virtueller Wurzelknoten, S/SIMPX: Satz, VF: Vorfeld, LK: Linke Satzklammer, MF: Mittelfeld, VC: Verbkomplex, HD: Kopf, SB/ON: Subjekt, MO: Modifizierer, V-MOD: Verb-Modifizierer, OC: Satz-Objekt, OV: verbales Objekt, ADVX: Adverbalphrase, VXFIN: finite Verbalphrase, NX: Nominalphrase, VXINF: infinite Verbalphrase

TiGer, dessen Annotationsschema auf dem der NEGRA-Baumbank basiert, rückt die Kodierung von Prädikat-Argument-Struktur ins Zentrum der Annotation. Zusammengehörige Argumente werden am gleichen Mutterknoten angehängt, so dass die Baumstruktur als eine direkte Repräsentation der Prädikat-Argument-Struktur betrachtet werden kann. Bei nicht-lokalen Abhängigkeiten führt dies allerdings zu kreuzenden Kanten (Abbildung 4, oben) und verletzt somit die Anforderungen einer kontextfreien Grammatik. Als Folge daraus handelt es sich bei den TiGer-Bäumen nicht mehr um Bäume im formalen Sinn, sondern vielmehr um gerichtete, azyklische Graphen.

Die Annotation in der TüBa-D/Z hingegen folgt der deskriptiven Theorie der Topologischen Felder (Drach, 1937; Höhle, 1986), was zu einer zusätzlichen Annotationsebene in der Baumbank führt (Abbildung 4, unten).

Die Annotation von nicht-lokalen Abhängigkeiten erfolgt hier nicht durch kreuzende Kanten, sondern mit Hilfe der Kantenlabel, die die grammatikalische Funktion spezifizieren. Als Folge dieser Unterschiede sind die Bäume in TiGer sehr flach, während die Annotation in der TüBa-D/Z zu hierarchischeren Baumstrukturen führt. Dies hat Auswirkungen auf die von der jeweiligen Baumbank extrahierte Grammatik und ihre Eignung für das probabilistische Parsen.[3]

1.4 Konvertierung von Phrasenstrukturbaumbanken zu Dependenzen

Das Fehlen einer hinreichend großen Dependenzbaumbank für die englische Sprache bei gleichzeitigem Vorhandensein einer großen Konstituentenbaumbank führte zur Entwicklung von zahlreichen Konvertierungsalgorithmen, mit dem Ziel, vorhandene Phrasenstrukturbaumbanken in Dependenz-Repräsentationen zu überführen, um die Daten dann zum Trainieren von statistischen Dependenzparsern zu benutzen. Als eine der wichtigsten Vorarbeiten auf diesem Gebiet ist Magerman (1994) zu nennen, der eine Tabelle mit Heuristiken erstellte, mit dem Ziel, für alle Konstituenten im Phrasenstrukturbaum die präferierten Köpfe zu finden. Hier folgt ein (vereinfachtes) Beispiel für eine solche Kopf-Heuristik (*head finding rule* oder *head percolation table*):

> *Für jede Nominalphrase, wähle das am weitesten rechts stehende Nomen und markiere es als Kopf der Phrase.*

Die Kopf-Heuristiken von Magerman (1994) wurden in vielen späteren Arbeiten aufgegriffen und modifiziert (Collins, 1999; Charniak, 2000; Yamada und Matsumoto, 2003; Nivre, 2005b). Im Folgenden wird der Prozess der Konstituenten-zu-Dependenz-Konvertierung beschrieben.

---

3 Für eine Diskussion über den Einfluss von Baumbank-Annotationsschemata auf die Performanz von probabilistischen Parsern siehe Kübler (2005); Maier (2006); Kübler u.a. (2006); Rehbein und van Genabith (2007); Kübler u.a. (2008).

Abbildung 5 zeigt einen Konstituentenbaum ohne grammatikalische Funktionslabel. Das Markieren der Köpfe bildet den ersten Schritt einer Konstituenten-zu-Dependenz-Konvertierung (Abbildung 6). Hierbei müssen wichtige Entscheidungen getroffen werden, wie z.B.: Was ist der Kopf einer PP? Wie sollen Koordinationen repräsentiert werden? Nachdem alle Köpfe markiert sind, werden im nächsten Schritt alle Kindknoten im Baum, die nicht Kopf der Konstituente sind, mit ihrer grammatikalischen Funktion ausgezeichnet (Abbildung 7). Verfügt die Konstituentenbaumbank über keine solchen grammatikalischen Funktions-Label, oder wird eine detailliertere Annotation gewünscht als im Konstituentenbaum vorhanden, so müssen die jeweiligen Dependenzrelationen der einzelnen Knoten mit Hilfe von Heuristiken bestimmt werden. Abbildung 8 zeigt die Dependenzdarstellung des fertig konvertierten Baums.

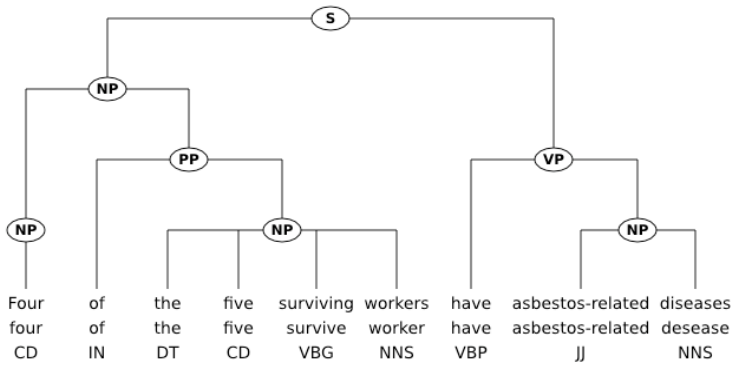

Abb. 5: Konstituentenbaum vor der Konvertierung

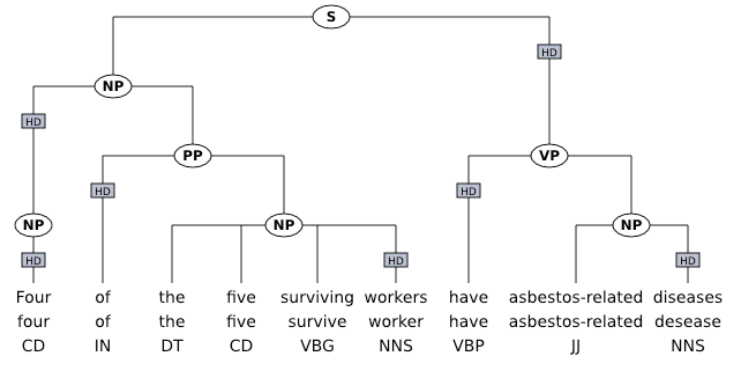

Abb. 6: Schritt 1: Markieren der Köpfe

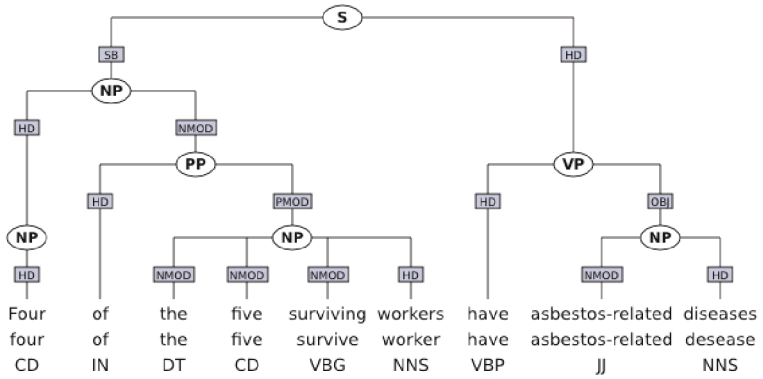

Abb. 7: Schritt 2: Auszeichnung der Dependenzrelationen

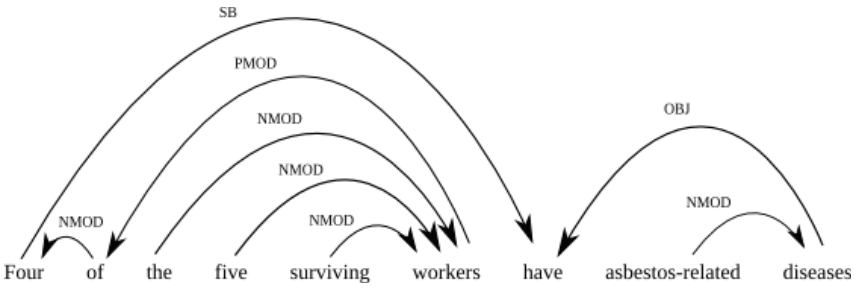

Abb. 8: Fertig konvertierter Dependenzbaum

Auf der Basis ihrer Modifikation der Kopf-Heuristiken von Magerman und angereichert mit weiteren Heuristiken und Regeln, stellen Yamada und Matsumoto (2003) und Nivre (2005a) Software zur Konvertierung der PTB zu Dependenzen bereit.[4] Allerdings weisen diese Ansätze zur Konvertierung Schwächen auf. Sie erzeugen Dependenz-Repräsentationen, die weniger aussagekräftig sind als die ursprüngliche, konstituenten-basierte Annotation der PTB, die z.B. nicht-lokale Abhängigkeiten, bedingt durch Topikalisierung, Wh-Bewegung, Gapping und andere Phänomene, mit Hilfe von leeren Argumenten oder Spuren indiziert. Die konvertierten Dependenzen hingegen lassen diese Informationen vermissen und annotieren darüber hinaus nur eine beschränkte Menge an Dependenzrelationen, die für eine tiefe linguistische Analyse unzureichend sind.

4 Die Programme sind frei erhältlich und können unter folgender URL heruntergeladen werden: http://www.jaist.ac.jp/~h-yamada/ (Yamada und Matsumoto, 2003); http://w3.msi.vxu.se/~nivre/research/idp.html (Nivre, 2005)

Dies ist ein Hauptkritikpunkt von Johansson und Nugues (2007), die sich zum Ziel nehmen, eine „semantisch sinvollere“ Art der Konvertierung bereitzustellen. Ihr Konvertierungsalgorithmus annotiert eine reichhaltigere Menge an feinkörnigeren Dependenzrelationen und führt außerdem eine Methode ein, um nichtlokale Phänomene wie Wh-Bewegung und Topikalisierung zu kodieren. Die resultierenden Dependenzbäume weisen dadurch eine komplexere Struktur auf als die Bäume aus früheren Konvertierungen. Dadurch wird die Aufgabe, die neuen Strukturen zu lernen und automatisch zuzuweisen, um einiges schwieriger und führt zu einer höheren Fehlerrate der auf diesen Daten trainierten Dependenzparser. Dennoch können sich die nun reichhaltigeren Informationen in den darauffolgenden Schritten der automatischen Sprachverarbeitung als nützlich erweisen. Johansson und Nugues (2007) zeigen dies am Beispiel eines Systems zur automatischen Annotation von semantischen Rollen (*Semantic Role Labelling*). Das System, das syntaktisch annotierte Daten als Eingabe bekommt und auf Basis dieser Informationen semantische Rollen zuweist, zeigt auf den neuen Dependenzrepräsentationen eine niedrigere Fehlerrate als auf den früheren, einfacheren Dependenzrepräsentationen.

2. Statistisches Dependenzparsing

So wie die Erstellung der englischen, konstituentenbasierten Penn Treebank (PTB) intensive Forschung im Bereich des statistischen Parsings mit probabilistischen kontextfreien Grammatiken ausgelöst hat, so hat auch die Erstellung von Dependenzbaumbanken zu einer Zunahme an wissenschaftlichen Arbeiten im Bereich des statistischen, datengesteuerten Dependenzparsings geführt. Bevor wir auf den Stand der Forschung im Bereich des datengesteuerten Dependenzparsings eingehen, müssen noch einige Begriffe geklärt werden.

2.1 Dependenzparsing: Konzepte und Strategien

Bei der Beschreibung von syntaktischen Parsingsystemen unterscheidet man zwischen *regelbasierten* Parsern und *datengesteuerten* Parsern. Regelbasierte Parser verfügen über eine zugrunde liegende, meist handgeschriebene Grammatik, und die syntaktische Analyse kann entweder als Suchproblem über alle in der Grammatik möglichen Ableitungen definiert werden, oder als Problem der Erfüllung von in der Grammatik beschriebenen Bedingungen (*constraint satisfaction problem*), so dass während des Parsings alle Analysen, die nicht die gestellten Bedingungen erfüllen, nacheinander eliminiert werden, bis nur noch eine einzige, nämlich die korrekte Analyse übrig bleibt. Im Gegensatz dazu verfügen datengesteuerte Parser über

keine Grammatikregeln, sondern „lernen“ die richtige syntaktische Analyse aus den annotierten Daten, der Baumbank.

Die Entwicklung von handgeschriebenen Grammatiken ist ein langwieriger Prozess, der viel Zeit und linguistisches Wissen erfordert. Datengesteuerte Parser hingegen lassen sich, wenn auf eine hinreichende Menge an annotierten Trainingsdaten zurückgegriffen werden kann, direkt auf diesen Daten trainieren. Im Weiteren zeichnen sie sich durch eine größere Robustheit und durch höhere Abdeckung auf neuen, unbekannten Texten aus. Im Gegensatz dazu sind regelbasierte Parser besser geeignet für Aufgaben, bei denen die Akkuratheit der Analyse im Vordergrund steht, so z.B. für die Beurteilung der Grammatikalität von Sätzen. Je nach Anwendungsgebiet wird man dem einen oder dem anderen Parser den Vorzug geben. Heutzutage jedoch lässt sich die Trennung zwischen datengesteuerten und regelbasierten Verfahren nicht mehr vollkommen aufrecht erhalten. Viele regelbasierte Parser verwenden zusätzlich statistische Modelle, so z.B. der WCDG-Parser (*Weighted Constraint Dependency Grammar*) für das Deutsche (Foth u.a., 2004), und umgekehrt können auch datengesteuerte Systeme durch die Anreicherung mit syntaktischen Informationen aus der Ausgabe von regelbasierten Parsern verbessert werden (Øvrelid u.a., 2009). Zusätzlich wurde viel Arbeit darauf verwendet, die Robustheit und Effizienz von regelbasierten Parsern zu verbessern (Riezler, 2002; Cahill, 2008), so dass die oben genannten Argumente nur noch bedingt gelten. Im Weiteren wird vorwiegend von datengesteuerten Dependenzparsern die Rede sein.

Beim datengesteuerten Dependenzparsing unterscheidet man zwischen *graphbasierten* und *transitionsbasierten* Parsern. Die Unterschiede zwischen beiden Modellen betreffen im Wesentlichen das Problem des Lernens und der Inferenz. Beim Lernen (oder Trainieren) wird versucht, die Parameter des statistischen Modells zu bestimmen, so dass sie die Trainingsdaten gut beschreiben. Inferenz hingegen bezeichnet die Aufgabe, aus einer Vielzahl von möglichen Parsebäumen die nach dem statistischen Modell beste syntaktische Analyse auszuwählen. Hier handelt es sich also um ein Suchproblem.

Graphbasierte und transitionsbasierte Modelle unterscheiden sich darin, wie sie das Problem des Lernens und der Inferenz angehen (Nivre und McDonald, 2008). Beim transitionsbasierten Parsen wird jede einzelne Aktion des Parsers bewertet, in Abhängigkeit zur bereits aufgebauten syntaktischen Struktur. Während der transitionsbasierte Parser lokale Entscheidungen trifft, kann er dabei auf eine reichhaltige Menge an Informationen über den globalen syntaktischen Kontext zugreifen. Der graphbasierte Parser hingegen berechnet die globale Wahrscheinlichkeit für den gesamten Parsebaum, wobei er eine erschöpfende Suche (*exhaustive search*) vornimmt, jedoch nur eingeschränkt auf Kontextinformationen zugreifen kann, da die Berücksichtigung einer größeren Anzahl an Informationen bei der erschöpfenden Suche zu Berechenbarkeitsproblemen führen würde. Der transitionsbasierte Parser hingegen benutzt einen gierigen Suchalgorithmus (*greedy*

search), der terminiert, sobald ein lokales Optimum gefunden wurde. Dies erlaubt die Berücksichtigung einer Vielzahl an Informationen bei der Suche nach der optimalen syntaktischen Analyse.

Beide Modelle weisen eine etwa gleich hohe Performanz auf, jedoch werden von den einzelnen Systemen sehr unterschiedliche Fehler gemacht. McDonald und Nivre (2007) führen die spezifischen Stärken und Schwächen beider Parsingmodelle auf ihre zugrunde liegenden Eigenschaften zurück, der Kombination von globalem Training und exhaustiver Inferenz auf Seiten des graphbasierten Parsers, und dem Zugriff auf reiche, aussagekräftige Informationen auf Seiten des transitionsbasierten Modells.

2.2 Stand der Forschung

Das Vorhandensein von Dependenzbaumbanken war die Voraussetzung für die Entwicklung von statistischen, datengesteuerten Dependenzparsern. Als treibende Kraft wirkte die CoNLL (Conference on Computational Natural Language Learning) *shared task*, ein Wettbewerb, bei dem die Teilnehmenden die Möglichkeit haben, Systeme zur automatischen Verarbeitung natürlicher Sprache auf den gleichen Datensets zu testen und systematisch zu vergleichen. In den Jahren 2006 und 2007 widmete sich dieser Wettbewerb dem datengesteuerten Dependenzparsing. Für verschiedene Sprachen aus unterschiedlichen Sprachfamilien wurden Trainingsdaten in einem einheitlichen Dependenzformat bereitgestellt, erzeugt durch die Konvertierung aus bestehenden Konstituenten- und Dependenzbaumbanken. Die gemeinsame Evaluation von verschiedenen Parsern auf den gleichen Daten erlaubt einen fairen Vergleich der Performanz verschiedener Parsingmodelle, aber auch eine Einschätzung, welche Sprachen leichter zu parsen sind, und welche Parsingstrategien für bestimmte Sprachtypen geeignet sind.

In der CoNLL shared task 2006 wurden 19 Parsingsysteme auf 13 verschiedene Sprachen angewendet und evaluiert (Buchholz und Marsi, 2006), in der zweiten CoNLL shared task in 2007 waren es 23 Systeme und 10 Sprachen (Nivre u.a., 2007). Als überraschendes Ergebnis zeigte sich, dass die besten Systeme nur geringe Unterschiede in der Performanz aufwiesen, erstaunlich vor allem deshalb, weil diese Systeme sich in Bezug auf Parsingstrategie und Architektur deutlich unterschieden. Unter den besten Systemen waren sowohl graphbasierte als auch transitionsbasierte Parser, sowie Ensemble-Systeme, die die Ausgabe von mehreren verschiedenen Parsern kombinieren.

Große Unterschiede in der Performanz waren auch für die verschiedenen Sprachen zu beobachten. Die besten Ergebnisse wurden für das Japanische erzielt (91,7% Labelled Attachment Score; LAS[5]), während die niedrigste Performanz

5 Labelled Attachment Score: die Prozentzahl an Token, für die das System den korrekten Kopf und die korrekte Dependenzrelation bestimmt hat.

für Türkisch mit 65,7% LAS deutlich abfiel. Es wäre aber vereinfacht anzunehmen, dass nur sprachtypologische Eigenschaften für diese Ergebnisse verantwortlich sind. Weitere Faktoren, die in Betracht gezogen werden müssen, sind die Menge an vorhanden Trainingsdaten, sowie Domäne und Genre dieser Daten, die einen großen Einfluss auf Struktur und Komplexität der Texte haben; aber auch das Annotationsschema der jeweiligen Baumbank kann die Parse-Qualität beeinflussen. Eine feinkörnigere Annotation macht die Analyse schwieriger, und es besteht die Möglichkeit, dass manche Annotationsschemata sich für die Anwendung von Methoden des Maschinellen Lernens besser eignen als andere.

Zusammenfassend kann gesagt werden, dass sowohl durch die Erstellung von Dependenzbaumbanken für verschiedene Sprachen als auch durch die Konvertierung vorhandener konstituentenbasierter Baumbanken zu Dependenzen viele Forschungsvorhaben im Bereich des datengesteuerten Dependenzparsings inspiriert wurden, so dass inzwischen für viele Sprachen statistische Dependenzparser vorhanden sind. Im Folgenden soll auf die Qualität dieser Parser eingegangen werden. Dabei sind die folgenden Fragen relevant: Wie lässt sich die Qualität von datengesteuerten Dependenzparsern im Vergleich zu konstituentenbasierten Parsern beurteilen? Wie kann eine sinnvolle Parser-Evaluation aussehen? Welches sind die relevanten Informationen, die ein guter syntaktischer Parser bereitstellen sollte?

3. Fragen der Evaluation

Ein direkter Vergleich der Performanz von Dependenzparsern mit konstituentenbasierten Parsern ist, bedingt durch die unterschiedlichen Ausgabeformate, nur schwer möglich. Konstituentenbasierte statistische Parser werden standardmäßig mit der PARSEVAL-Metrik evaluiert (Black u.a., 1991). Dazu benötigt man einen Goldstandard, also manuell annotierte bzw. von Hand korrigierte Daten, und gemessen an der Anzahl an Phrasen-Spannen, die im Goldstandard und in der Ausgabe des Parsers übereinstimmen, wird ein Wert für die Parse-Qualität berechnet.[6]

Bei Dependenzparsern hingegen wird meist der LAS (Labelled Attachment Score), also die Prozentzahl an Token, für die der Parser sowohl Kopf als auch Dependenzrelation richtig bestimmt hat, als Maß der Parse-Qualität angegeben. Möchte man Dependenzparser und Konstituentenparser vergleichen, so muss ei-

6 Diese Art der Evaluation ist stark abhängig vom jeweiligen Annotationsschema der Baumbank, da die Werte in Abhängigkeit zur Gesamtanzahl an Konstituenten im Baum ermittelt werden. Für hierarchische Annotationsschemata, die eine hohe Anzahl an unären Knoten aufweisen, wird die Anzahl an Fehlern, also nicht übereinstimmenden Phrasen-Spannen, geringer gewichtet, da sie durch eine höhere Anzahl an Konstituenten im Parsebaum geteilt wird. Die PARSEVAL-Metrik ist demnach ungeeignet für den Vergleich von Parsern, die auf verschiedenen Baumbanken trainiert wurden, und demnach auch für den Vergleich von Parsern für verschiedene Sprachen (Rehbein und van Genabith, 2007).

nes der beiden Ausgabeformate ins jeweils andere Format umgewandelt werden. Lin (1995) beschreibt eine Methode zur Konvertierung von Konstituentenstrukturen zu Dependenzrelationen. Allerdings ist eine solche Konvertierung, wie in Kapitel 1.4 beschrieben, oft fehlerbehaftet. Schiehlen (2004) und Rehbein und van Genabith (2007) präsentieren eine Evaluation eines nicht lexikaliserten, deutschen Konstituentenparsers mit Hilfe der PARSEVAL-Metrik und einer dependenzbasierten Evaluation nach der Methode von Lin (1995) auf den gleichen Daten und zeigen, dass die beiden Evaluationsverfahren zu unterschiedlichen Ergebnissen führen können. Unklar ist jedoch, welche der beiden Methoden ein verlässlicheres Maß der Parse-Qualität bietet: die konstituentenbasierte PARSEVAL-Metrik oder die dependenzbasierte Evaluation.

Die konstituentenbasierte Evaluation mit PARSEVAL wurde in der Literatur vielfach kritisiert (Carroll u.a., 1998; Briscoe u.a., 2002, Sampson und Babarczy, 2003, Rimell u.a., 2009). Als ein Hauptkritikpunkt gilt, dass PARSEVAL keine linguistisch bedeutsame Beurteilung der Parse-Qualität wiedergibt. Deshalb argumentierte schon Lin (1995; Seite 1425) für die dependenzbasierte Evaluation als „more intuitively meaningful than other scores but also more relevant to semantic interpretation".

Diese Auffassung, die auch eine Antwort auf die Fragen gibt, wie eine sinnvolle Parser-Evaluation aussehen kann und welche Informationen die Analyse eines syntaktischen Parsers beinhalten sollte, hat sich in den letzten Jahren stark verbreitet und nahm einen großen Einfluss auf den gesamten Bereich der automatischen Sprachverarbeitung. So wird die Evaluation auf Basis von Dependenzrelationen nicht nur im Bereich des syntaktischen Parsens immer üblicher, sondern greift auch auf andere Gebiete in der Verarbeitung natürlicher Sprache über, so z.B. auf die Evaluation von Systemen zur Maschinellen Übersetzung (Liu und Gildea, 2005; Owczarzak, 2007) oder von automatisch erstellten Zusammenfassungen (Owczarzak, 2009). Dies bildet die Überleitung zu der Frage, auf welche anderen Bereiche der automatischen Verarbeitung natürlicher Sprache die Dependenzgrammatik Einfluss genommen hat.

4. Dependenzen in der automatischen Sprachverarbeitung

Die Repräsentation von syntaktischer Information in Form von Dependenzen hat sich in den letzten Jahren in vielen Bereichen der automatischen Verarbeitung natürlicher Sprache durchgesetzt. Eine vollständige Aufzählung würde den Rahmen dieses Artikels sprengen, so sollen zur Illustration nur einige Beispiele genannt werden.

Besonders großen Einfluss hatte die Dependenzgrammatik auf alle Gebiete der automatischen Sprachverarbeitung, die einen mehr oder weniger direkten Zusammenhang mit der Extraktion von semantischen Informationen aufweisen.

Dazu zählt das Feld der Informationswiedergewinnung *(Information Retrieval)*, das die automatische Suche und Extraktion von Informationen aus Datenbeständen wie z.B. einer großen Datenbank oder dem WWW zum Ziel hat. Zu den bekanntesten Anwendungen der Informationswiedergewinnung zählen Internetsuchmaschinen wie Yahoo!, Google oder Bing.[7] Ein Unterbereich der Informationswiedergewinnung ist das sogenannte *Question Answering*, bei dem nicht nur mögliche Antworten zu Fragen in natürlicher Sprache gefunden, sondern diese auch nach Relevanz und Korrektheit geordnet werden müssen. Dies erfordert eine tiefe linguistische Analyse der Daten. Hier gibt es immer mehr Ansätze, die sich dazu der Dependenzgrammatik bedienen (Punyakanok u.a., 2004; Bouma u.a., 2005).

Bei der automatischen Annotation von semantischen Rollen (*Semantic Role Labelling*) werden ebenfalls immer häufiger syntaktische Repräsentationen in der Form von Dependenzgraphen genutzt (Hacioglu, 2004; Fürstenau und Lapata, 2009), die für diese Aufgabe einen direkteren Zugriff auf relevante Informationen bieten als Repräsentationen im Rahmen der Phrasenstrukturgrammatik. Eine systematische Untersuchung des Einflusses der syntaktischen Repräsentation auf Anwendungen des Semantic Role Labelling wird von Johansson und Nugues (2008) vorgestellt. Die Autoren zeigen, dass beide Formen der Repräsentation als Komponente in einem Semantic Role Labelling-System in etwa gleiche Ergebnisse liefern, obwohl die Forschung im Bereich der konstituentenbasierten Parser viel weiter vorangeschritten ist. Das heißt, dass man in Zukunft mit Verbesserungen der Technologie von Dependenzparsern auch auf weitere Verbesserungen im Bereich des Semantic Role Labelling hoffen kann.

Sagae (2009) präsentiert ein System zur Diskursanalyse im Rahmen der Rhetorical Structure Theory (Mann und Thompson, 1988). Ziel ist die Extraktion von Diskursrelationen innerhalb und zwischen Sätzen von Dokumenten. Dabei nutzt Sagae als Komponenten der syntaktischen Verarbeitung sowohl einen konstituentenbasierten Parser als auch einen Dependenzparser und erzielt eine höhere Akkuratheit als vergleichbare Arbeiten bei gleichzeitiger Verbesserung der Effizienz seines Systems.

In der statistischen Maschinellen Übersetzung, einem großen und wichtigen Zweig der automatischen Verarbeitung natürlicher Sprache, gibt es immer mehr Bestrebungen, statistische Modelle mit syntaktischen Informationen anzureichern. Auch hier gibt es immer mehr Arbeiten, die dazu Dependenzrelationen nutzen. Ein Beispiel sind Bojar und Hajič (2008), die auf syntaktische Analysen im Rahmen der Funktionalen Generativen Beschreibung (Sgall u.a., 1986) zurückgreifen.

Als letztes Beispiel für den Einfluss der Dependenzgrammatik auf die Computerlinguistik soll die Auszeichnung eines Lernerkorpus auf Basis von Dependenzen genannt werden (Ragheb und Dickinson, 2009). Die syntaktische Annotation der von den Lernenden entwickelten Interimsprache, der sogenannten Interlanguage, umfasst zwei Analyseebenen (Oberflächen- und Tiefenstruktur) und

7 Yahoo!: http://www.yahoo.com; Google: http://www.google.com; Bing: http://www.bing.com

ähnelt damit der Annotation in der Prager Dependenzbaumbank. Als Ziel ihrer Arbeit sehen Ragheb und Dickinson (2009) die Erstellung einer Ressource für die Zweitspracherwerbsforschung, die aber auch als Testdaten für die Entwicklung von robusten, fehlertoleranten Parsern dienen kann.

5. Fazit

Ähnlich wie in der theoretischen Linguistik, in der sich viele verschiedene Ausprägungen der Dependenzgrammatik entwickelten, so gibt es auch in der automatischen Sprachverarbeitung weder eine einheitliche Theorie für die Annotation von Dependenzen in Korpora noch für die Zielrepräsentationen von Dependenzparsern. Im Bereich der syntaktisch annotierten Korpora zeigt sich die ganze Bandbreite der Dependenzgrammatik, von monostratalen Baumbanken mit nur einer Ebene der syntaktischen Analyse bis hin zu multistratalen Baumbanken mit mehreren Analyseebenen, wie zum Beispiel die Prager Dependenzbaumbank. Auf dem Gebiet des datengesteuerten Dependenzparsings hingegen sind Repräsentationen mit nur einer Ebene der syntaktischen Analyse vorherrschend (Nivre, 2005a).

Während in den ersten Jahrzehnten nach Erscheinen von Tesnière's Hauptwerk *Éléments de syntaxe structurale* die Dependenzgrammatik wenig Beachtung fand, so hat sich in der letzten Dekade gerade in der Korpuslinguistik viel getan. Die Erstellung von Dependenzkorpora für eine Vielzahl an Sprachen hat vor allem im Bereich des datengesteuerten Dependenzparsings große Fortschritte gebracht. Wenn auch ein Vergleich von Dependenzparsern und Konstituentenparsern aufgrund der unterschiedlichen Ausgabeformate schwierig ist, so lässt sich doch sagen, dass die Performanz von Dependenzparsern inzwischen für Sprachen wie Englisch mit der von konstituentenbasierten Parsern vergleichbar ist (Mcdonald, 2006). Für Sprachen mit einer komplexen Morphologie und relativ freier Wortfolge, wie z.B. dem Tschechischen, liefern manche Dependenzparser inzwischen bessere Ergebnisse als herkömmliche konstituentenbasierte Parser wie die von Collins und Charniak (Mcdonald, 2006).[8] Diese Ergebnisse sind im Einklang mit der oft diskutierten Behauptung, dass Dependenzgrammatiken besser geeignet sind für die Kodierung und Verarbeitung von Sprachen mit freier Wortfolge, da sie nicht versuchen, die Oberflächenstruktur eines Satzes abzubilden, sondern die zugrunde liegenden funktionalen Abhängigkeiten zwischen den lexikalen Einheiten im Satz. Dies gilt jedoch nicht für alle Varianten der Dependenzgrammatik

8 Die Unterscheidung zwischen dependenzbasierten und konstituentenbasierten Parsern ist ein wenig vereinfacht, da auch die Parsingmodelle von Collins (1999) und Charniak (2000) internen Gebrauch von bilexikalen Dependenzen machen. Es handelt sich jedoch um Parser mit einer generativen Komponente, die als Ausgabe Phrasenstrukturbäume erzeugen, während die beschriebenen datengesteuerten Dependenzparser auf diskriminativen Methoden des Maschinellen Lernens beruhen.

und auch nicht für alle Dependenzparser, sondern nur für solche, die nicht-projektive Strukturen zulassen. Auch hier sind jedoch in den letzten Jahren bedeutende Fortschritte erzielt worden (Nivre, 2007, Hall und Nivre, 2008, Nivre, 2009).

Ein wichtiger Vorteil der Dependenzgrammatik ist die Einfachheit der syntaktischen Repräsentation, die durch die direkte Kodierung von Prädikat-Argument-Struktur intuitiv verständlich ist. Dadurch sind syntaktische Strukturen in Form von Dependenzrelationen besonders geeignet für Anwendungen, die die automatische Extraktion von Bedeutung zum Ziel haben. Als weiterer Vorteil ist die Effizienz von Dependenzparsern zu nennen, deren Algorithmen durch die Beschränkung der Anzahl an möglichen syntaktischen Knoten im Parsebaum, die durch die Anzahl an lexikalen Elementen in der Eingabe bestimmt wird, eine geringere Komplexität aufweisen (Nivre, 2005a).

Insgesamt lässt sich sagen, dass die Dependenzgrammatik im Bereich der automatischen Verarbeitung natürlicher Sprache angekommen ist und dass durch sie wichtige Impulse gesetzt wurden, nicht nur durch die Entwicklung von neuen Technologien und Systemen, sondern auch durch die theoretische Debatte über Fragen der geeigneten Repräsentation von syntaktischen Informationen und eine sinnvolle Evaluation von Systemen zur automatischen Extraktion sprachlichen Wissens.

6. Literatur

Aduriz, Itzair; Aranzabe, Maria J.; Arriola, Jose M.; Atutxa, Aitziber; Ilarraza, Arantza D. de; Garmendia, Aitzpea; Oronoz, Maite: Construction of a Basque Dependency Treebank. In: *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT-2)*. Växjö, Schweden, 2003.

Bahr, Joachim: Eine Jahrhundertleistung historischer Lexikographie: Das Deutsche Wörterbuch, begründet von Jacob und Wilhelm Grimm. In: Werner Besch, Oskar Reichmann, Stefan Sonderegger (Hrsg.): *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* Bd. 1. Halbband. Handbücher zur Sprach- und Kommunikationswissenschaft; 2. Berlin/New York, 1984, S. 492–501.

Bamman, David; Crane, Gregory: The Design and Use of a Latin Dependency Treebank. In: *Proceedings of the 5th Workshop on Treebanks and Linguistic Theories (TLT-5)*. Prag, Tschechische Republik, 2006, S. 67–68.

Bamman, David; Mambrini, Francesco; Crane, Gregory: An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. In: *Proceedings of the 8th Workshop on Treebanks and Linguistic Theories (TLT-8)*. Mailand, Italien, 2009, S. 5–16.

Beek, Leonoor Van D.; Bouma, Gosse; Malouf, Robert; Noord, Gertjan V.: The Alpino Dependency Treebank. In: *The 12th Meeting of Computational Linguistics in the Netherlands (CLIN-02)*. Enschede, Niederlande, 2002, S. 1686–1691.

Bharati, Akshar; Sangal, Rajeev; Chaitanya, Vineet; Kulkarni, Amba; Sharma, Dipti M.; Ramakrishnamacharyulu, K.V.: AnnCorra: Building Treebanks in Indian Languages. In: *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardiza-*

tion (COLING-02). Morristown, NJ : Association for Computational Linguistics, 2002, S. 1–8.

Black, Ezra W.; Abney, Steven; Flickinger, Dan; Gdaniec, Claudia; Grishman, Ralph; Harrison, Philip; Hindle, Donald; Ingria, Robert; Jelinek, Fred; Klavans, Judith; Liberman, Mark; Marcus, Mitch; Roukos, Salim; Santorini, Beatrice; Strzalkowski, Tomek: A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In: *In Proceedings of the DARPA Speech and Natural Language Workshop*. San Mateo, CA, 1991, S. 306–311.

Boas, Franz: *Race, Language, and Culture*. New York, NY: Macmillan, 1940.

Boguslavsky, Igor; Grigorieva, Svetlana; Grigoriev, Nikolai; Kreidlin, Leonid; Frid, Nadezhda: Dependency Treebank for Russian: Concept, Tools, Types of Information. In: *Proceedings of the 18th Conference on Computational Linguistics*. Morristown, NJ, 2000, S. 987–991.

Bojar, Ondřej; Hajič, Jan: Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In: *Proceedings of the 3rd Workshop on Statistical Machine Translation*. Columbus, OH, June 2008, S. 143–146.

Bosco, Christina; Lombardo, Vincenzo: Dependency and Relational Structure in Treebank Annotation. In: *Proceedings of Workshop on Recent Advances in Dependency Grammar at COLING-04*. Genf, Schweiz, 2004.

Bouma, Gosse; Mur, Jori; Noord, Gertjan van; Plas, Lonneke van der; Tiedemann, Jörg: Question Answering for Dutch Using Dependency Relations. In: *Proceedings of the Cross-Language Evaluation Forum Workshop (CLEF-05)*. Wien, Österreich, 2005, S. 370–379.

Brants, Sabine; Dipper, Stefanie; Hansen, Silvia; Lezius, Wolfgang; Smith, George: The TIGER Treebank. In: *Proceedings of the 1st Workshop on Treebanks and Linguistic Theories (TLT-1)*. Sozopol, Bulgarien, 2002, S. 24–42.

Briscoe, Ted; Carroll, John; Graham, Jonathan; Copestake, Ann: Relational Evaluation Schemes. In: *Proceedings Workshop 'Beyond Parseval – towards improved evaluation measures for parsing systems', 3rd International Conference on Language Resources and Evaluation (LREC-02)*. Las Palmas, Kanarische Inseln, 2002, S. 4–38.

Buchholz, Sabine; Marsi, Erwin: CoNLL-X Shared Task on Multilingual Dependency Parsing. In: *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*. Morristown, NJ, 2006, S. 149–164.

Cahill, Aoife: Treebank-Based Probabilistic Phrase Structure Parsing. In: *Language and Linguistics Compass* 2 (2008), Nr. 1, S. 18–40.

Cahill, Aoife; Burke, Michael; O'Donovan, Ruth; Riezler, Stefan; van Genabith, Josef; Way, Andy: Wide-coverage Deep Statistical Parsing Using Automatic Dependency Structure Annotation. In: *Computational Linguistics* 34 (2008), Nr. 1, S. 81–124.

Carroll, John; Briscoe, Ted; Sanfilippo, Antonio: Parser Evaluation: a Survey and a New Proposal. In: *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC-98)*. Granada, Spanien, 1998.

Charniak, Eugene: A Maximum-entropy-inspired Parser. In: *Proceedings of the Conferences and Proceedings of the ANLP-NAACL 2000 Student Research Workshop*. Seattle, WA, 2000, S. 132–139.

Chomsky, Noam: Quine's Empirical Assumptions. In: Davidson, Donald (Hrsg.); Hintikka, Jaakko (Hrsg.): *Words and objections. Essays on the work of W. V. Quine*. Dordrecht : Reidel, 1969, S. 53–68.

Collins, Michael: *Head-Driven Statistical Models for Natural Language Parsing*. Philadelphia, PA, University of Pennsylvania, Dissertation, 1999.

Dickinson, Markus; Ragheb, Marwa: Dependency Annotation for Learner Corpora. In: *Proceedings of the 8th Workshop on Treebanks and Linguistic Theories (TLT-8)*. Mailand, Italien, 2009, S. 59–70.

Drach, Erich: *Grundgedanken der Deutschen Satzlehre*. Diesterweg, Frankfurt/M.: reprint Darmstadt, Wissenschaftliche Buchgesellschaft, 1963, 1937.

Džeroski, Sašo; Erjavec, Tomaž; Ledinek, Nina; Pajas, Petr; Žabokrtsky, Zdenek; Žele, Andreja: Towards a Slovene Dependency Treebank. In: *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC-06)*. Genua, Italien, 2006.

Einarsson, Jan: *Talbankens skriftspråkskonkordans*. Lund University, Department of Skandinavian Languages. 1976.

Einarsson, Jan: *Talbankens talspråkskonkordans*. Lund University, Department of Skandinavian Languages. 1976.

Foth, Kilian A.; Daum, Michael; Menzel, Wolfgang: A Broadcoverage Parser for German Based on Defeasible Constraints. In: *Beiträge zur 7. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS-04)*. Wien, Österreich, 2004, S. 45–52.

Fürstenau, Hagen; Lapata,Mirella: Semi-Supervised Semantic Role Labeling. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*. Athen, Griechenland, 2009, S. 220–228.

Gildea, Daniel; Jurafsky, Daniel: Automatic Labeling of Semantic Roles. In: *Computational Linguistics* 28 (2002), Nr. 3, S. 245–288.

Hacioglu, Kadri: Semantic Role Labeling Using Dependency Trees. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*. Morristown, NJ, 2004, S. 1273–1277.

Hajič, Jan; Smrž, Otakar; Zemánek, Petr; Šnaidauf, Jan; Beška, Emanuel: Prague Arabic Dependency Treebank: Development in Data and Tools. In: *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*. Kairo, Ägypten, 2004, S. 110–117.

Hall, Johan; Nivre, Joakim: Parsing Discontinuous Phrase Structure with Grammatical Functions. In: *Advances in Natural Language Processing (GoTAL-08)*. Göteburg, Schweden, 2008, S. 169–180.

Höhle, Tilman: Der Begriff ‚Mittelfeld', Anmerkungen über die Theorie der topologischen Felder. In: *Akten des Siebten Internationalen Germanistenkongresses*. Göttingen, Deutschland, 1986, S. 329–340.

Johansson, Richard; Nugues, Pierre: Extended Constituent-to-dependency Conversion for English. In: *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA-07)*. Tartu, Estland, 2007, S. 105–112.

Johansson, Richard; Nugues, Pierre: The Effect of Syntactic Representation on Semantic Role Labeling. In: *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*. Manchester, United Kingdom, August 18-22 2008, S. 393–400.

Kaeding, Friedrich W.: *Häufigkeitswörterbuch der Deutschen Sprache*. Steglitz bei Berlin: Selbstverlag des Herausgebers, 1897.

Kilgarriff, Adam; Grefenstette, Gregory: Introduction to the Special Issue on the Web as Corpus. In: *Computational Linguistics* 29 (2003), Nr. 3, S. 333–347.

Kübler, Sandra; Hinrichs, Erhard W.; Maier, Wolfgang: Is it Really that Difficult to Parse German? In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-06)*. Morristown, NJ, 2006, S. 111–119.

Kübler, Sandra; Maier, Wolfgang; Rehbein, Ines; Versley, Yannick: How to Compare Treebanks. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-08)*. Marrakesch, Marokko, 2008, S. 2322–2329

Kübler, Sandra: How do Treebank Annotation Schemes Influence Parsing Results? or How Not to Compare Apples and Oranges. In: *Proceedings of Recent Advances in Natural Language Processing (RANLP-05)*. Borovets, Bulgaria, 2005, S. 293–300.

Lavie, Alon; Parlikar, Alok; Ambati, Vamshi: Syntax-driven Learning of Sub-sentential Translation Equivalents and Translation Rules from Parsed Parallel Corpora. In: *Proceedings of the 2nd Workshop on Syntax and Structure in Statistical Translation (SSST-2)*. Columbus, OH, 2008.

Lepage, Yves; Ando, Shin-Ichi; Akamine, Susumu; Iida, Hitoshi: An Annotated Corpus in Japanese Using Tesnière's Structural Syntax. In: *Proceedings of the Coling-ACL'98 Workshop on Processing of Dependency Grammars*. Montreal, Canada, 1998, S. 109–115.

Lin, Dekang: A Dependency-based Method for Evaluating Broad-Coverage Parsers. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-95)*. Montreal, QC, 1995, S. 1420–1425.

Liu, Ding; Gildea, Daniel: Syntactic Features for Evaluation of Machine Translation. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, MI, 2005.

Liu, Haitao: Probability Distribution of Dependencies Based on a Chinese Dependency Treebank. In: *Journal of Quantitative Linguistics* 16 (2009), Nr. 3, S. 256–273.

Maamouri, Mohamed; Bies, Ann; Buckwalter, Tim; Mekki, Wigdan: The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In: *NEMLAR Conference on Arabic Language Resources and Tools*. Kairo, Ägypten, 2004, S. 102–109.

Magerman, David M.: *Natural Language Parsing as Statistical Pattern Recognition*. Stanford, CA, Stanford University, Dissertation, 1994.

Maier, Wolfgang: Annotation Schemes and their Influence on Parsing Results. In: *Proceedings of the COLING/ACL 2006 Student Research Workshop*. Sydney, Australien, July 2006, S. 19–24.

Mann, William C.; Thompson, Sandra A.: Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. In: *Text* 8 (1988), Nr. 3, S. 243–281.

Marcu, Daniel: A Decision-based Approach to Rhetorical Parsing. In: *The 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*. College Park, MD, 1999, S. 365–372.

Marcus, Mitchell P.; Marcinkiewicz, Mary A.; Santorini, Beatrice: Building a Large Annotated Corpus of English: the Penn Treebank. In: *Computational Linguistics* 19 (1993), Nr. 2, S. 313–330.

McDonald, Ryan: *Discriminative Training and Spanning Tree Algorithms for Dependency Parsing*. Philadelphia, PA, University of Pennsylvania, Dissertation, 2006.

Mcdonald, Ryan; Nivre, Joakim: Characterizing the Errors of Data-driven Dependency Parsing Models. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*. Prag, Tschechische Republik, 2007, S. 122–131.

Meurers, Walt D.; Müller, Stefan: Corpora and Syntax (Article 42). In: Lüdeling, Anke; Kytö, Merja (Hrsg.): *Corpus linguistics* Bd. 2. Berlin : Mouton de Gruyter, 2009, S. 920–933.

Müller, Stefan; Meurers, Walt D.: Corpus Evidence for Syntactic Structures and Requirements for Annotations of Tree Banks. In: *Proceedings of the International Conference on Linguistic Evidence*. Tübingen, 2006.

Nivre, Joakim: Dependency Grammar and Dependency Parsing / Växjö University: School of Mathematics and Systems Engineering. Växjö, Schweden, 2005. – Forschungsbericht.

Nivre, Joakim: *Inductive Dependency Parsing of Natural Language Text*. Växjö, Schweden, Växjö University, Dissertation, 2005.

Nivre, Joakim: Incremental Non-Projective Dependency Parsing. In: *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL-07)*. Rochester, NY, 2007, S. 396–403.

Nivre, Joakim: Non-Projective Dependency Parsing in Expected Linear Time. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapur, 2009, S. 351–359.

Nivre, Joakim; Hall, Johan; Kübler, Sandra; McDonald, Ryan; Nilsson, Jens; Riedel, Sebastian; Yuret, Deniz: The CoNLL 2007 Shared Task on Dependency Parsing. In: *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL-07*. Prag, Tschechische Republik, 2007, S. 915–932.

Nivre, Joakim; McDonald, Ryan: Integrating Graph-Based and Transition-Based Dependency Parsers. In: *The 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08:HLT)*. Columbus, OH, 2008, S. 950–958.

Oflazer, Kemal; Say, Bilge; Hakkani-Tür, Dilek Z.; Tür, Gökhan: Building A Turkish Treebank. (2003), S. 261–277.

Øvrelid, Lilja; Kuhn, Jonas; Spreyer, Kathrin: Improving Data-driven Dependency Parsing Using Large-scale LFG Grammars. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapur, August 2009, S. 37–40.

Owczarzak, Karolina: DEPEVAL(summ): Dependency-based Evaluation for Automatic Summaries. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference *on Natural Language Processing of the AFNLP*. Suntec, Singapur, August 2009, S. 190–198.

Owczarzak, Karolina; van Genabith, Josef; Way, Andy: Labelled Dependencies in Machine Translation Evaluation. In: *Proceedings of the Second Workshop on Statistical Machine Translation (StatMT-07)*. Morristown, NJ, 2007, S. 104–111.

Pradhan, Sameer S.; Ward, Wayne; Martin, James H.: Towards Robust Semantic Role Labeling. In: *Computational Linguistics* 34 (2008), Nr. 2, S. 289–310.

Preyer, William T.: *The Mind of the Child*. New York, NY : Appleton, 1889.

Punyakanok, Vasin; Roth, Dan; Yih, Wen-tau: Mapping Dependencies Trees: An Application to Question Answering. In: *Proceedings of the 8th International Symposium on Artificial Intelligence and Mathematics*. Fort Lauderdale, FL, 2004.

Rehbein, Ines; van Genabith, Josef: Treebank Annotation Schemes and Parser Evaluation for German. In: *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP-07)*. Prag, Tschechische Republik, 2007, S. 630–639.

Riezler, Stefan; King, Tracy H.; Kaplan, Ronald M.; Crouch, Richard; III, John T. M.; Johnson, Mark; Johnson, Iii M.: Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In: *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*. Philadelphia, PA, 2002, S. 271–278.

Rimell, Laura; Clark, Stephen; Steedman, Mark: Unbounded Dependency Recovery for Parser Evaluation. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-09)*. Suntec, Singapur, August 2009, S. 813–821.

Sagae, Kenji: Analysis of Discourse Structure with Syntactic Dependencies and Data-Driven Shift-Reduce Parsing. In: *Proceedings of the 11th International Conference on Parsing Technologies (IWPT-09)*. Paris, Frankreich, 2009, S. 81–84.

Sampson, Geoffrey; Babarczy, Anna: A Test of the Leaf-ancestor Metric for Parse Accuracy. In: *Natural Language Engineering* 9 (2003), Nr. 4, S. 365–380.

Schiehlen, Michael: Annotation Strategies for Probabilistic Parsing in German. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*. Morristown, NJ, 2004, S. 390.

Sgall, Petr; Hajičová, Eva; Panevová, Jarmila: *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht : Reidel, 1986.

Skut, Wojciech; Krenn, Brigitte; Brants, Thorsten; Uszkoreit, Hans: An Annotation Scheme for Free Word Order Languages. In: *Proceedings of the 5th Applied Natural Language Processing Conference (ANLP-97)*. Washington, D.C., 1997, S. 88–95.

Stern, William: *Psychology of Early Childhood up to the Sixth Year of Age*. New York, NY : Henry Holt, 1924.

Tadić, Marko: Building the Croatian Dependency Treebank: the Initial Stages. In: *Suvremena lingvistika* 33 (2007), Nr. 63, S. 85–92.

Telljohann, Heike; Hinrichs, Erhard W.; Kübler, Sandra; Zinsmeister, Heike: *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Deutschland: Universität Tübingen (Veranst.), 2005.

Tinsley, John; Hearne, Mary; Way, Andy: Parallel Treebanks in Phrase-Based Statistical Machine Translation. In: *Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. Mexico City, Mexico, 2009, S. 318–331.

Volk, Martin: How Bad is the Problem of PP-Attachment? A Comparison of English, German and Swedish. In: *Proceedings of 3rd ACL-SIGSEM Workshop on Prepositions*. Trient, Italien, 2006.

Volk, Martin; Samuelsson, Yvonne: Bootstrapping Parallel Treebanks. In: *Proceedings of the COLING'04 5th International Workshop on Linguistically Interpreted Corpora*. Genf, Schweiz, 2004, S. 63–70.

Xia, Fei; Rambow, Owen; Bhatt, Rajesh; Palmer, Martha; Sharma, Dipti M.: Towards a Multi-representational Treebank. In: *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT-7)*. Groningen, Netherlands, 2009, S. 159–170.

Xue, Naiwen; Xia, Fei; Chiou, Fu-dong; Palmer, Martha: The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. In: *Natural Language Engineering* 11 (2005), Nr. 2, S. 207–238.

Yamada, Hiroyasu; Matsumoto, Yuji: Statistical Dependency Analysis With Support Vector Machines. In: *Proceedings of 8th International Workshop on Parsing Technologies (IWPT-03)*. Nancy, Frankreich, 2003, S. 195–206.

Zhechev, Ventsislav; Way, Andy: Automatic Generation of Parallel Treebanks. In: *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08*. Morristown, NJ, USA, 2008, S. 1105–1112.

Adresse der Verfasserin:
Ines Rehbein, Department for Computational Linguistics & Phonetics, Saarland University, D-66041 Saarbruecken.
E-Mail: rehbein@coli.uni-sb.de