UNIVERSITÄT REGENSBURG

Informationsstrukturierung für die syntaktische Annotation eines diachronen Korpus des Deutschen

Magisterarbeit im Fach Informationswissenschaft

Institut für Medien-, Informations- und Kulturwissenschaft

von:	Michael Heilemann
Adresse:	Josef-Schlicht-Str. 13 94330 Oberpiebing
Matrikelnummer:	114 207 2
Erstgutachter:	Prof. Dr. Christian Wolff
Zweitgutachter:	Prof. Dr. Rainer Hammwöhner
Laufendes Semester:	Sommersemester 2008
Abgabedatum:	01. Juni 2008

Zusammenfassung

Diese Arbeit beschreibt für das Projekt *Diachrone Syntax Deutsch* (DiSynDe) die Informationsstrukturierung für ein diachrones Korpus des Deutschen. Das Korpus soll auf unterschiedlichen linguistischen Ebenen annotiert werden. Da dadurch überlappende Hierarchien auftreten, die nicht in einem XML-Dokument repräsentiert werden können, wird unter anderem auf das Konzept der Stand-Off-Annotation eingegangen.

Für die morphosyntaktische Annotation wird das *Stuttgart Tübingen Tagset* (STTS) vorgestellt und an die Annotation historischer Texte angepasst. Der Schwerpunkt liegt auf der syntaktischen Annotation, für deren Kodierung Standards wie die *Text Encoding Initiative* (TEI), der *Corpus Encoding Standard* (XCES), das *TIGER-Projekt* und das *Syntactic Annotation Framework* (SynAF) gesichtet werden. Außerdem werden die *DiSynDe*-Annotationsvorgaben mit den syntaktischen Annotationsebenen nach EAGLES (*Expert Advisory Group on Language Engineering Standards*) in Beziehung gesetzt und überarbeitet. Das TIGER-Tagset zur syntaktischen Annotation wird an die *DiSynDe*-Annotationsvorgaben angepasst. Im Bereich der textgrammatischen Annotation wird auf die *Rhetorical Structure Theory* (RST) eingegangen.

Abstract

This thesis describes information structuring for the project *Diachrone Syntax Deutsch* (DiSynDe) in order to create a diachronic corpus of German. The corpus will be annotated on different linguistic levels. Thus there are overlapping hierarchies, which cannot be represented in one XML-document; therefore the concept of stand-off-annotation is examined.

A typical case of corpus annotation is morphosyntactic annotation; therefore the *Stuttgart Tübingen Tagset* (STTS) is adapted to the annotation of historical texts. But the focus lies on syntactic annotation. For the encoding of syntactic annotation the following standards are discussed: *Text Encoding Initiative* (TEI), *Corpus Encoding Standard* (XCES), *TIGER-Project* and *Syntactic Annotation Framework* (SynAF). Furthermore the annotation guidelines of *DiSynDe* will be related to the syntactic annotation levels of EAGLES (*Expert Advisory Group on Language Engineering Standards*) and revised. The syntactic TIGER-Tagset will be adapted to the annotation guidelines of *DiSynDe*. Textgrammatical annotation is discussed concerning *Rhetorical Structure Theory* (RST).

Inhaltsverzeichnis

1. Einleitung

Die Korpuslinguistik stellt keinen eigenen Teilbereich der Sprachwissenschaft wie Syntax oder Pragmatik dar, sondern eine neue Methode innerhalb der Linguistik: „(…) this modern phenomenon, corpus linguistics, has come to be an increasingly prevalent methodology in linguistics (…).“[1]

Anhand von Beispielen aus texttechnologischen Korpora werden Annahmen und Theorien empirisch belegt. Linguisten benutzen Korpora demnach als Werkzeug, um neue Erkenntnisse über Sprache zu gewinnen und zu belegen. Bevor man jedoch mit Hilfe eines Korpus sprachwissenschaftliche Fragestellungen beantworten kann, steht – sofern kein existierendes Korpus als geeignet erscheint – eine Phase der Korpuserstellung. Während die Linguistik in dieser Phase beispielsweise entscheiden muss, welche Texte aufgenommen und inwiefern diese Texte aufbereitet werden, fällt die technische Umsetzung in den Aufgabenbereich der Informationswissenschaft.

Dazu gehört neben der Bereitstellung von Software-Werkzeugen – zum Beispiel der Erstellung eines geeigneten Annotationstools – die Informationsstrukturierung für die Annotation der Korpustexte; ein wichtiger Gesichtspunkt dabei ist, sich mit der Verwendung von korpuslinguistischen Standards auseinanderzusetzen, denn „[s]tandardization of annotation practices can ensure that an annotated corpus can be used to its greatest potential.“[2]

1.1. Themenstellung und Zielsetzung

Eine Forschergruppe mit dem Arbeitstitel *Diachrone Syntax Deutsch* (DiSynDe) beschäftigt sich mit der Erstellung eines diachronen, syntaktisch annotierten Korpus des Deutschen. Dahinter steht die Motivation, ein neues Standardwerk zur geschichtlichen Entwicklung der deutschen Syntax zu verfassen, denn

> „[d]ie letzte große Gesamtdarstellung der historischen Syntax des Deutschen, nämlich die von OTTO BEHAGHEL, ist in vier Bänden zwischen 1923 und 1932 erschienen. Mittlerweile ist auf dem Gebiete der historischen Syntax seither viel geleistet worden. (…) [D]ie Fortschritte der Quellenerschließung, des methodischen Zugriffs und der theoretischen Fundierung in mehr als einem dreiviertel Jahrhundert lassen das Wagnis einer neuen Gesamtdarstellung, die neben den ehrwürdigen ‚Behaghel' tritt (…) als gerechtfertigt erscheinen.“[3]

[1] McEnery und Wilson 2003: 1.
[2] Kahrel et al. 1997: 231.
[3] Schmid 2007: 51.

Aus einer Studie des Projektes *Deutsch Diachron Digital* (DDD)[4] geht hervor, dass zwar viele ältere Texte des Deutschen in digitalisierter Form zugänglich sind, man aber diachrone Untersuchungen nur schwer durchführen kann. Dies liegt vor allem daran, dass eine Zusammenführung der unterschiedlichen Korpora nicht ohne weiteres möglich ist, da „[b]isher (…) anerkannte Standardisierungen für historische Korpora [fehlen] (es gibt zu viele Korpora, die sich nicht an die TEI o. ä. halten).“[5] In der Studie wird besonders kritisiert, dass sich „viele (insbesondere deutsche) Projekte (…) nicht an korpuslinguistische Standards halten.“[6]

Aus der Studie lässt sich außerdem ableiten, dass es kein diachrones Korpus des Deutschen gibt, das auch syntaktisch annotiert ist. Bei *DiSynDe* besteht jedoch ein Konsens darüber, „dass die Grundlage ein tragfähiges, elektronisch aufbereitetes, d.h. annotiertes Textcorpus sein muss, das die Zeit vom Beginn der deutschen Überlieferung bis zum 18. Jahrhundert repräsentativ dokumentieren soll.“[7]

Am Lehrstuhl für Informationswissenschaft in Regensburg setzen sich derzeit zwei Magisterarbeiten mit der technischen Seite der Korpuserstellung auseinander. Während Manuel Burghardt Annotationstools für diachrone Korpora evaluiert, beschäftigt sich die vorliegende Arbeit mit der Informationsstrukturierung für die Annotation des zu erstellenden Korpus. Ziel der Arbeit ist die Zusammenstellung eines Annotationsschemas, mit dem die geforderten Beschreibungsebenen und Phänomene der historischen Sprachstufen beschrieben werden können. Das Hauptaugenmerk der Arbeit liegt dabei auf der Annotation syntaktischer Information; darüberhinaus wird auch auf die Annotation von Metadaten, morphosyntaktischer und textlinguistischer Aspekte eingegangen.

Das Annotationsschema soll erweiterbar und wiederverwendbar sein, also potentiell auf andere Korpora, Sprachen oder Sprachstufen übertragen werden können. Damit diese Anforderungen erfüllt werden können, ist es notwendig, dass das Schema soweit wie möglich theorieunabhängig gehalten wird und dass korpuslinguistische Standards verwendet werden; die in dieser Arbeit diskutierten Standards sind das Stuttgart Tübingen Tagset (STTS), die Text Encoding Initiative (TEI), die XML-Version des Corpus Encoding Standard (XCES), das TIGER-Annotationsschema und das Syntactic Annotation Framework (SynAF).

---

[4] Vgl. Kroymann et al. 2004.
[5] Ebd.: 41.
[6] Ebd.: 41.
[7] Schmid 2007: 51.

1.2. Aufbau der Arbeit

Als erstes wird der Hintergrund der Arbeit – das Projekt *Diachrone Syntax Deutsch* – näher beleuchtet. Vor allem aus der Vorstellung der Analysegruppen und der potentiellen Suchanfragen an das annotierte Korpus werden Vorgaben für das Annotationsschema abgeleitet.

Der zweite Punkt unterteilt sich in die Abschnitte *Die Annotation syntaktischer Information* und *Die Kodierung syntaktischer Annotation*. Dabei werden zuerst grundlegende Begriffe wie Korpus und Annotation definiert. Im weiteren Verlauf motiviert der Punkt den Gebrauch syntaktischer Annotation, erörtert das Konzept *Theorieneutralität* und stellt die verschiedenen Ebenen syntaktischer Annotation der *Expert Advisory Group on Language Engineering Standards* (EAGLES) vor. Anschließend werden die Annotationsvorgaben des *DiSynDe*-Projekts auf diese Beschreibungsebenen übertragen.

Der Punkt *Die Kodierung syntaktischer Annotation* stellt unterschiedliche Formate für die Repräsentation syntaktischer Informationen vor; außerdem werden Lösungsansätze diskutiert, wie sich überlappende Hierarchien in XML kodieren lassen.

Im Abschnitt *Die Annotation historischer Texte* werden als erstes die Begriffe Diachronie und Synchronie definiert. Desweiteren werden aus Entwicklungslinien der deutschen Sprache Besonderheiten abgeleitet, die bei der Annotation historischer Texte zu beachten sind.

Der vierte Punkt stellt ausgewählte korpuslinguistische Standards vor und diskutiert ihre Einsatzmöglichkeiten für *DiSynDe* hinsichtlich der Bereiche Metadaten, morphosyntaktische, syntaktische und textgrammatische Annotation. Als erstes wird der CES-Metadatensatz erörtert, daran anschließend das STTS-Tagset für die morphosyntaktische Annotation. Der Schwerpunkt liegt auf der syntaktischen Annotation; hierfür werden die Standards TEI, XCES, TIGER und SynAF vorgestellt. Im Rahmen der textgrammatischen Annotation wird auf die Rhetorical Structure Theory (RST) eingegangen.

Im abschließenden Punkt der Arbeit wird auf das *Potsdamer Austauschformat für linguistische Annotation* (PAULA) hingewiesen. PAULA stellt eine Möglichkeit dar, die unterschiedlichen *DiSynDe*-Annotationsebenen zusammenzuführen und mit der linguistischen Datenbank ANNIS Korrelationen zwischen den Ebenen zu analysieren.

2. Das Projekt *Diachrone Syntax Deutsch*

2.1. Diachrone Korpora des Deutschen allgemein

Das Projekt *Diachrone Syntax Deutsch* verfolgt das Ziel, eine neue Gesamtdarstellung zur historischen Entwicklung der deutschen Syntax zu schreiben. Ausgangspunkt dazu soll ein syntaktisch annotiertes, diachrones Korpus des Deutschen sein, „das die Zeit vom Beginn der deutschen Überlieferung bis zum 18. Jahrhundert repräsentativ dokumentieren soll."[8]

DiSynDe ist nicht das einzige Projekt seiner Art, das momentan in diesem Themenkomplex angesiedelt ist; das Interesse an der Erstellung diachroner Korpora ist weit verbreitet. An der Universität Regensburg wird am Lehrstuhl für slavische Sprachwissenschaft an der Erstellung eines diachronen Korpus des Russischen gearbeitet. Aber auch in Bezug auf das Deutsche gibt es zahlreiche Initiativen: Das *Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts* (DWDS) etwa soll hinsichtlich seiner Einteilung der deutschen Sprache in Dekaden in Richtung Vergangenheit erweitert werden.

Das Projekt *Deutsch Diachron Digital* (DDD) hat ebenfalls „die Konzeption eines diachronen Korpus des Deutschen zum Gegenstand. (…) Zu entwickeln ist es im entstehenden Projekt DeutschDiachronDigital (…), einer bundesweiten Initiative von Forscherinnen und Forschern aus der (historischen) Philologie, der (historischen) Sprachwissenschaft sowie aus Literaturwissenschaft, Korpuslinguistik und Informatik."[9] *DiSynDe* arbeitet mit *Deutsch Diachron Digital* zusammen, wobei die Modalitäten der Zusammenarbeit noch nicht konkret sind, wie folgendes Zitat zeigt:

> „Eine Zeit lang war daran gedacht, mit dem Projekt ‚Deutsch Diachron Digital' zusammenzuarbeiten, da hier bereits Digitalisierungs- und Annotationsarbeiten geleistet wurden. (…) Ob und in welchem Umfang auf vorhandene digitalisierte Corpora zurückgegriffen werden kann, ist zum gegenwärtigen Zeitpunkt noch offen."[10]

DDD stellt dem Projekt *DiSynDe* nach Möglichkeit auf Ebene der Wortarten annotierte Texte zur Verfügung; allerdings soll das geplante *DiSynDe*-Pilotkorpus von diesen Überlegungen nicht betroffen sein.

An dieser Stelle werden weitere Projekte angeführt, die sich mit älteren Sprachstufen des Deutschen beschäftigen. Die Daten sind der in der Einleitung bereits erwähnten Studie von DDD entnommen. Weitere historische Korpora des Deutschen sind

---

[8] Schmid 2007: 51.
[9] Lüdeling et al. 2004: 1.
[10] Schmid 2007: 51f.

das *Bonner Frühneuhochdeutsch Korpus*, das *Bochumer Mittelhochdeutsch Korpus*, das *Digitale Mittelhochdeutsche Textarchiv*, der *Thesaurus Indogermanischer Text- und Sprachmaterialien* (TITUS), die *Codices Electronici Ecclesiae Coloniensis* (CEEC) und die *Mittelhochdeutsche Begriffsdatenbank*.

Die Sprachstufen des Deutschen sind hinsichtlich ihrer Digitalisierung unterschiedlich abgedeckt; während das Althochdeutsche fast komplett verfügbar ist – was auch an der geringen Überlieferungsdichte dieses Zeitraums liegt –, sind für das Frühneuhochdeutsche nur wenige Texte in digitalisierter Form zugänglich. Das Mittelhochdeutsche ist zu einem großen Teil erfasst, jedoch nicht komplett.[11]

Demzufolge ist viel Material vorhanden, doch diachrone Studien sind nicht ohne weiteres möglich, weil eine Zusammenführung der unterschiedlichen Korpora wegen unterschiedlicher Formate und fehlender Standards schwierig ist. So entstehen mit den unterschiedlichen Korpora Wissensinseln, die nicht zusammen eingesetzt werden können. Deshalb ist der Einsatz von Standards auch bei kleinen Projekten wichtig und wird für *DiSynDe* zur Pflicht.

2.2. Pilotkorpus

Das *DiSynDe*-Korpus soll nicht nur verschiedene Sprachstufen des Deutschen umfassen, sondern auch unterschiedliche Textsorten integrieren. Mit Hilfe des Korpus sollen Fragen des Syntaxwandels beantwortet werden; der Korpus stellt sicher, dass die Sprachwissenschaftler eine verlässliche Menge an Beispielsätzen – also an empirischen Daten – zum Beispiel zur Belegung von Argumenten auf effektive und effiziente Weise gewinnen können. Die Zusammensetzung des Korpus bezüglich der Texte ist noch nicht endgültig bestimmt, aber eine Vorauswahl für ein Pilotkorpus steht fest.

Das Pilotkorpus enthält Texte aus sieben Jahrhunderten (11. – 17. Jahrhundert). Die Texte können den Bereichen Chronistik, Fachliteratur, Geistliche Prosa, Privatschriften (Brief, Tagebuch, Reisebericht), Rechts- und Verwaltungsprosa, Unterhaltungsprosa und Übersetzungsliteratur zugeordnet werden. Das Pilotkorpus enthält Beispieltexte aus allen Jahrhunderten, außerdem sind unterschiedliche Regionalsprachen und verschiedene Textsorten vorhanden. Das Annotationsschema kann bei der Annotation des Pilotkorpus erprobt werden; wesentliche Analyse- und Annotationsprobleme sollten durch die Abbildung im Kleinen erkennbar und dadurch für die Anwendung auf das eigentliche *DiSynDe*-Korpus vermieden werden können. Wenn zum Beispiel eine Wortart,

[11] Vgl. Kroymann et al. 2004: 41.

eine bestimmte Nebensatzart oder ein syntaktisches Phänomen nicht kodiert werden können, kann das Annotationsschema um die fehlenden Elemente ergänzt werden.

2.3. Arbeitsgruppen und Bearbeitungsreihenfolge

Die Mitglieder von *DiSynDe* sind in fünf Gruppen aufgeteilt. Während die Annotationsgruppe für die technische Seite der Korpuserstellung zuständig ist, sind vier Analysegruppen für die Annotation ihrer jeweiligen Beschreibungsebene zuständig. Die Analysegruppen bestehen aus jeweils zwei Personen, sogenannten *Tandems*: „Jeweils ein ‚Tandem' aus Projektbeteiligten zeichnet für eine dieser Beschreibungsebenen verantwortlich."[12] Auf diesen Ebenen soll das Korpus linguistisch annotiert werden:

> „1. Der Text als größte syntaktisch relevante strukturbedingende Entität
> 2. Der komplexe Satz
> 3. Der einfache Satz
> 4. Die Wortgruppe
> 5. Die Wortarten als kleinste syntaktisch relevante strukturbedingende Entitäten."[13]

Ursprünglich war für die Annotation der Wortarten eine eigene Analysegruppe vorgesehen, die nun aber herausgenommen wird. Die Annotation der Wortarten wird möglicherweise von DDD übernommen.

Die *Wortgruppe* ist zum Beispiel für den Phraseninnenbau, etwa von Präpositionalphrasen, zuständig. Die Gruppe *Einfacher Satz* beschäftigt sich mit der Struktur von Elementarsätzen, die Gruppe *Komplexer Satz* mit der Gesamtstruktur von Sätzen, und die *Textgruppe* ist für den Bereich Textgliederung und Textgrammatik zuständig.

Die Analysegruppen erstellen für ihre jeweilige Beschreibungsebene abstrakte Annotationsvorschriften in Tabellenform. Diese Vorschriften der Sprachwissenschaftler zeigen, welche konkreten Phänomene kodiert werden sollen, und stellen die Grundlage für die Übertragung der verschiedenen Ebenen in ein geeignetes Tagset dar.

Da die Mitglieder der Analysegruppen auf verschiedene Universitäten und Institute verteilt sind und eine gemeinsame Datenbasis auf unterschiedlichen Beschreibungsebenen annotieren sollen, tritt das Problem der Koordination und der Bearbeitungsreihenfolge auf. Es existieren drei Vorschläge für die Zusammenarbeit der Analysegruppen: parallele Bearbeitung, gestaffeltes Vorgehen und *Bottom-up*.

---

[12] Schmid 2007: 53.
[13] Ebd.: 52.

Für das parallele Bearbeiten der Ebenen spricht das komfortable Arbeiten für die Analysegruppen. Die Tandems können jederzeit ihre Ebenen ändern und müssen nicht warten, bis eine andere Gruppe ihre Ebene an einem bestimmten Text fertig annotiert hat. Ein Nachteil dabei ist, dass es zu Inkonsistenzen kommen kann; beispielsweise können manche syntaktischen Phänomene von der vorhergehenden Annotation durch die Wortgruppe abhängen, sodass es für beide Satzgruppen nicht möglich ist, ihre Überlegungen im Text konsistent zu annotieren, solange sie nicht wissen, auf welche Weise die Wortartengruppe entscheidet. Außerdem stehen automatische Annotationshilfen nicht immer zur Verfügung oder sind zumindest fehlerträchtig, wenn eine dafür nötige Beschreibungsebene noch nicht von ihrer Gruppe bearbeitet wurde.

Ähnlichen Überlegungen ist das gestaffelte Vorgehen geschuldet. Die Gruppe *Einfacher Satz* braucht als Vorgabe die Ergebnisse der Gruppe *Komplexer Satz*, nämlich die Zerlegung von Gesamtsätzen in Teilsätze. Die Ergebnisse der Gruppe *Einfacher Satz* werden an die Gruppen *Komplexer Satz* und *Wortgruppe* weitergeleitet, da hier Überlappungsbereiche anzutreffen sind.

Bei der Bottom-up-Vorgehensweise beginnt man mit der Annotation der Wortarten und arbeitet sich Ebene für Ebene nach oben, bis die Gruppe für Textgrammatik schließlich die letzten Annotationen hinzufügt. Ein Vorteil dabei ist die einfache Implementierung des Annotationstools, da auf diese Weise ein komplexer Regelbaum hierarchisch abgearbeitet werden kann; dadurch ist die Möglichkeit von Annotationskonflikten geringer. Dazu muss das Annotationsschema die Ebenen konsequent voneinander trennen und in einer modulartigen Bauweise zur Verfügung stellen. Weitere Schlussfolgerungen für die Vorgehensweise bei der Annotation des *DiSynDe*-Korpus finden sich im Abschnitt *3.1.5. Annotationsebenen für Diachrone Syntax Deutsch* (vgl. Seite 32f).

2.4. Potentielle Suchanfragen an das annotierte Korpus

> „Die einzelnen Annotationsaspekte von der untersten bis zur obersten Annotationsebene müssen in einem automatisierten Abfrageverfahren in Kombination abfragbar sein."[14]

Im Sinne der Informationsstrukturierung ist es von besonderer Bedeutung, sich damit auseinanderzusetzen, welche Informationen man anhand des fertigen Korpus abfragen will, um eine neue Grammatik der historischen Syntax des Deutschen erstellen zu können. Es stellt sich also die Frage, welche potentiellen Suchanfragen das annotierte Korpus beantworten soll. Da sich diese Fragestellungen auch im Laufe des Projektes verändern können, ist dies ein weiteres Argument für die Erweiterbarkeit und Flexibilität des Annotationsschemas. Aus einem Aufsatz[15] von Hans Ulrich Schmid und dem Arbeitsblatt einer *DiSynDe*-Projektsitzung vom 23. Juni 2007 ergeben sich folgende potentielle Suchanfragen:

1. Wie ist die Verbstellung in Hauptsätzen, denen ein Konditionalsatz vorausgeht im 14. Jahrhundert?
2. Welche Struktur weisen Konditionalsätze „links" von der Trägerstruktur in Rechtstexten, und zwar des 14. Jahrhunderts auf?
3. Welche formalen Konditionalsatztypen werden in Gefügen mit mehr als einem Konditionalsatz in Rechtsprosa und Fachliteratur verwendet?
4. Welche Konstituentenabfolgen in Nominalgruppen gelten innerhalb von verbalen Klammern in Textsorte XY?
5. Welche Verben bilden Genitiv-Valenz aus innerhalb von Rechtstexten diachron vom 13. bis zum 17. Jahrhundert?
6. Auf welcher Sprachstufe und im Bereich welcher Textsorte erscheint in komplexen Prädikaten relativer Attributsätze das Finitum in Letztstellung?
7. Welche Subjunktionen kommen im ersten Jahrzehnt des 15. Jahrhunderts in der Rechtsprosa in vor- und nachgestellten Konditionalsätzen in welcher Häufigkeit vor?

[14] Schmid 2007: 57.
[15] Vgl. ebd.: 57.

Daraus lässt sich ableiten, dass für die Beantwortung der linguistischen Suchanfragen vor allem folgende Informationen von Bedeutung sind: Im Bereich der Metadaten sind dies die Klassifizierung der Texte nach den Textsorten und der Sprachstufe bzw. der Entstehungszeit. Im Rahmen der morphosyntaktischen Annotation muss das Abfragen von Wortarten wie Subjunktion möglich sein.

Nicht nur die grammatische Funktion von Teilsätzen (etwa Hauptsatz oder Konditionalsatz) soll abfragbar sein, sondern auch deren Struktur und Position innerhalb des Satzgefüges. Desweiteren ist die Stellung des finiten Verbs im Satz von Bedeutung. Außerdem sollen komplexe Prädikate und verbale Klammern (*Peter **hat** gestern Gemüse **eingekauft**.*[16]) identifiziert werden können – zur Erklärung: „Das Deutsche weist (…) die Eigenheit auf, die weiteren Prädikatsteile (z. B. infinite Verben oder trennbare Verbzusätze) an die letzte Position im Satz zu stellen. Man spricht hier von der verbalen Klammer des Deutschen."[17] Während das Auffinden von Verben mit einer bestimmten Valenz vor allem die Kategorisierung von Phrasen bzw. die Subkategorisierung von Verben betrifft, spielt bei der Abfolge von Konstituenten in Nominalgruppen sowohl die Annotation auf Ebene der Wortarten als auch auf Ebene der Phrasen (*Wortgruppe*) eine Rolle.

[16] Kessel und Reimann 2005: 9.
[17] Ebd.: 9.

3. Informationsstrukturierung und Annotation

In diesem Abschnitt werden Begriffe wie Korpus, syntaktische Annotation und Informationsstrukturierung definiert; zugleich erfolgt eine Einführung in die betreffenden Problemstellungen des Themas.

3.1. Die Annotation syntaktischer Information

3.1.1. Begriffsbestimmung von *Korpus* und *Annotation*

Die Definition des Begriffes *Korpus* ist nicht so einfach, wie es auf den ersten Blick scheint, denn „[i]m Prinzip kann ein Stapel alter Zeitungen oder eine Sammlung handschriftlicher Briefe einer bestimmten Autorin als Korpus angesehen werden."[18] Da *DiSynDe* die Erstellung eines texttechnologischen Korpus anstrebt, wird die Definition weiter eingeschränkt:

> „Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen. Die Daten des Korpus sind digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar." [19]

Das *DiSynDe*-Korpus wird keine gesprochenen Äußerungen, sondern nur schriftliches Datenmaterial enthalten, da die Aufzeichnung von Ton eine Erfindung der Neuzeit ist und es somit keine Quellen gibt, die beispielsweise mittelhochdeutsche Texte als Tonmaterial zur Verfügung stellen. Das *DiSynDe*-Korpus umfasst jedoch nicht nur die verschiedenen Sprachstufen des Deutschen, sondern es kann vorkommen, dass die aufgenommenen Texte auch fremdsprachiges Material, zum Beispiel lateinische Stellen, aufweisen. Solche Textstellen sind insofern wichtig, da es vor allem bei althochdeutschen Texten der Fall sein kann, dass syntaktische Konstruktionen des Lateinischen für das Deutsche übernommen wurden.

Ein texttechnologischer Korpus besteht aus bestimmten Komponenten:

> „Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind."[20]

Die „Daten selbst", die schriftlichen Sprachdaten bzw. der Text, der in das Korpus in digitalisierter Form aufgenommen wird, bezeichnet man auch als Primärdaten.

---

[18] Sasaki und Witt 2004: 195.
[19] Lemnitzer und Zinsmeister 2006: 40.
[20] Ebd.: 40.

Ein Korpus besteht darüberhinaus aus Metadaten und linguistischen Annotationen. Metadaten sind „*Daten über Daten*“[21], sie beschreiben „verschiedene Aspekte einer Informationsressource“[22]. Im Fall von *DiSynDe* ist eine Informationsressource ein Manuskript, dessen Text in das Korpus aufgenommen wird. Damit rekonstruiert werden kann, welche Handschrift benutzt wurde, wird der digitalisierte Text mit Metadaten versehen.

Dies stellt nur einen möglichen Aspekt dar, weitere „Aspekte, unter denen eine Informationsressource beschrieben werden kann, sind z.B. ihr Inhalt, das Trägermedium, die Art der Kodierung, die Autoren und andere bei der Produktion beteiligte Personen, der Zeitpunkt der Entstehung.“[23] Für *DiSynDe* besonders wichtig ist der Entstehungszeitpunkt eines Textes; dadurch lässt sich bei Suchanfragen das annotierte Korpus auf ein Teilkorpus eingrenzen, beispielsweise auf frühneuhochdeutsche Texte oder nur auf Texte aus der zweiten Hälfte des 14. Jahrhunderts. Während Metadaten also „[i]n einem guten Korpus (…) über die Herkunft dieser Äußerungen und Texte und über einiges mehr Auskunft geben“[24] und somit vom Terminus *Annotation* begrifflich getrennt werden, versteht man unter Annotation folgendes:

> „It can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/ or written language data.“[25]

Das bedeutet, dass ein Korpus im Prozess der Annotation mit zusätzlichen, linguistischen Informationen angereichert wird oder dass „die linguistischen Einheiten (…) mit ihren linguistischen Beschreibungen verbunden“[26] werden. Der Begriff *Annotation* kann sich nicht nur auf den Vorgang, sondern auch auf das Ergebnis dieses Prozesses beziehen:

> „‚Annotation‘ can also refer to the end-product of this process: the linguistic symbols which are attached to, linked with, or interspersed with the electronic representation of the language material itself.“[27]

Geoffrey Leech betont in seinen Ausführungen, dass linguistische Annotation stets eine Interpretation der vorliegenden Texte ist:

> „(…) by calling annotation ‚interpretative‘, we signal that annotation is, at least in some degree, the product of the human mind´s understanding of the text. There is

[21] Lemnitzer und Zinsmeister 2006: 45.
[22] Ebd.: 45.
[23] Ebd.: 45f.
[24] Ebd.: 44.
[25] Leech 1997: 2.
[26] Lemnitzer und Zinsmeister 2006: 44.
[27] Leech 1997: 2.

> no purely objective, mechanistic way of deciding what label or labels should be applied to a given linguistic phenomenon."[28]

In diesem Abschnitt wurde der Begriff *Korpus* als eine Sammlung schriftlicher oder gesprochener Äußerungen einer oder verschiedener Sprachen definiert; die Daten eines Korpus liegen in digitalisierter Form vor und sind maschinenlesbar. Außerdem setzt sich ein Korpus aus den drei Bestandteilen Primärdaten, Metadaten und linguistische Annotation zusammen. Unter *Annotation* versteht man das Hinzufügen von linguistischen Informationen zu den Primärdaten, wobei dadurch auch stets eine Interpretation der Daten erfolgt.

3.1.2. Der Nutzen syntaktischer Annotation

Um verständlich zu machen, welchen Nutzen eine syntaktische Annotation leisten kann, soll erst der Sinn von Annotation im Allgemeinen erläutert werden. Um einen Text mit linguistischen Informationen anzureichern, ist es notwendig, die Primärdaten zu interpretieren. Dazu darf man sie nicht isoliert untersuchen, sondern muss sie in ihrem Kontext betrachten, um die richtige Lesart zu erhalten:

> „Linguistische Informationen entstehen erst bei der Interpretation der Daten. Viele Wortformen oder Strukturen sind aus dem Kontext herausgenommen mehrdeutig. Sie müssen zunächst mit ihrem Kontext in Bezug gesetzt und so disambiguiert werden."[29]

Die Wortform *einen* beispielsweise kann je nach Lesart auf drei unterschiedliche Weisen interpretiert werden: als Artikel (*Diese Perspektive ermöglicht **einen** neuen Blick auf gesellschaftliche Verhältnisse.*), als Indefinitpronomen (*Was für die **einen** nur dekadenter Zeitvertreib, ist für die anderen blutiger Ernst.*) oder als Infinitiv eines Verbs (*Sie wollten von Bremen aus die Republik wieder **einen**.*).[30]

Wenn man in einem Korpus ohne linguistische Annotationen die Wortform *einen* nach der Lesart des Artikels finden will, erhält man eine große Anzahl relevanter Treffer; der Aufwand erscheint vertretbar, aus der Trefferliste die nicht relevanten Lesarten auszusondern. Wenn man jedoch daran interessiert ist, Korpusstellen mit *einen* als Verb zu ermitteln – eine relativ seltene Lesart –, ist der Benutzer eines nicht annotierten Korpus mit einem Problem konfrontiert: Er „(…) muss eine große Anzahl von irrelevanten Bei-

[28] Leech 1997: 2.
[29] Lemnitzer und Zinsmeister 2006: 60.
[30] Vgl. ebd.: 60f.

spielen sichten, was mühsam und zeitaufwändig ist. Diese Arbeit wird enorm reduziert, wenn den einzelnen Wortformen ihre jeweilige Wortart zugeordnet ist. Man kann dann gezielt nach Vorkommen von *einen* als Verb suchen."[31]

Wenn man auch daran interessiert ist, mit welchen Objekten das Verb *einen* im Korpus zusammen auftritt – „wie der Nominalphrase *die Republik*"[32] im obigen Beispielsatz –, „kann man nur dann effizient suchen, wenn entweder syntaktische Phrasen annotiert oder sogar die grammatischen Funktionen der Wortgruppen angegeben sind."[33] Je abstrakter das linguistische Phänomen ist, das im Korpus gefunden werden soll, umso deutlicher tritt der Sinn von Annotation hervor. Wenn man unabhängig von der Wortform wissen möchte, welche Prädikativkonstruktionen im Genitiv im Korpus auftreten, kann eine Suchanfrage nur bei entsprechender Annotation beispielsweise zu folgenden Ergebnissen führen: „*der Ansicht sein, der Meinung sein, guten Mutes sein*."[34] Der Nutzen syntaktischer Annotation wird demnach deutlich, wenn man nicht nur linguistische Phänomene auf Wortebene, sondern auch auf Ebene der Syntax mit Hilfe eines Korpus untersuchen möchte.

Die Syntax stellt einen Teilbereich der Grammatik dar und wird folgendermaßen definiert: „Syntax (…) ist die Lehre vom Bau der Sätze."[35] Der Begriff stammt von dem griechischen Wort σύνταξις (sýntaxis) und bedeutet *Zusammenordnung*; dies bezieht sich darauf, wie Wörter zu einem Satz zusammengeordnet werden.[36] Syntax oder die Satzlehre ist also ein „System von Regeln, die beschreiben, wie aus einem Inventar von Grundelementen (Morphemen, Wörtern, Satzgliedern) durch spezifische syntaktische Mittel (Morphologische Markierung, Wort- und Satzgliedstellung, Intonation u.a.) alle wohlgeformten Sätze einer Sprache abgeleitet werden können."[37]

Der Satz steht im Mittelpunkt der Syntax: „Der Satz ist als kleinster, relativ selbständiger Bestandteil der Rede bzw. des Textes die Grundeinheit, die mit Hilfe verschiedener linguistischer Modelle analysiert wird."[38] An dieser Stelle soll davon ausgegangen werden, dass ein Satz ein Prädikat beinhaltet – von Sonderfällen wie Ellipsen wird abgesehen. Ellipsen sind „Sätze, in denen Teile weggelassen werden, die jedoch aus dem Text oder einer vorhergehenden Äußerung ergänzt werden können. Hierher gehört

[31] Lemnitzer und Zinsmeister 2006: 61.
[32] Ebd.: 61.
[33] Ebd.: 61.
[34] Ebd.: 61.
[35] Kessel und Reimann 2005: 1.
[36] Vgl. Bußmann 2002: 676.
[37] Ebd.: 676.
[38] Volmert 2001: 115.

das Beispiel *Heute (kommt mein Besuch).* auf die Frage *Wann kommt dein Besuch?*"[39] Dies führt zu folgender Satz-Definition:

> „Ein Satz ist eine sprachliche Konstruktion aus verschiedenen Satzgliedern, in deren Zentrum ein Prädikat steht."[40]

Während der Satz als kleinster Bestandteil eines Textes angesehen wird, stellt ein Wort die kleinste Einheit dar, die im Bereich der Syntax untersucht werden kann. Sobald man die Strukturen innerhalb eines Wortes untersucht, befindet man sich im Bereich der Morphologie. Das Wort ist dabei die Schnittstelle zwischen Morphologie und Syntax: morphologisch gesehen repräsentiert das Wort die größte Einheit – syntaktisch gesehen ist das Wort die kleinste Einheit, die nicht weiter zerlegt werden kann. In Abgrenzung zur Morphologie tritt der Aufgabenbereich der Syntax deutlich hervor:

> „Während bei der Syntax die Analyse des Aufbaus von Satzstrukturen und der Zusammenfügung von Wörtern zu größeren Einheiten im Mittelpunkt steht (…), befasst sich die Morphologie mit dem Aufbau von Wortstrukturen und dem Zusammenfügen von kleineren bedeutungstragenden Bestandteilen zu Wörtern."[41]

Wie bereits gezeigt wurde, kann Sprache auf der Wortebene Mehrdeutigkeiten aufweisen, die in einem Korpus mit Hilfe von Annotation aufgelöst werden können. Solche Mehrdeutigkeiten können auch auf der syntaktischen Ebene auftreten; zum Beispiel kann der unklare Bezug von Präpositionalphrasen zu Missverständnissen führen:

> „Unter Linguisten ist in diesem Zusammenhang ein Zitat von Groucho Marx berühmt: ‚Last night I shot an elephant in my pajamas and how he got in my pajamas I'll never know'. Diese Ambiguität des Bezugs von *in my pajamas* ist normalerweise für den Leser eine Falle, da sie im Folgesatz in die weniger wahrscheinliche Lesart aufgelöst wird. Im Korpus kann sie durch syntaktische Annotation eindeutig gemacht werden, indem die Präpositionalphrase *in my pajamas* der nominalen Struktur von *an elephant* zugeordnet wird."[42]

Dadurch wird deutlich, dass syntaktische Annotation eine wortübergreifende Analyse ist. Bevor man die Primärdaten eines Korpus syntaktisch annotiert, wird normalerweise eine Annotation der Wortarten durchgeführt, das sogenannte *Part-of-Speech Tagging* (*POS-*

[39] Kessel und Reimann 2005: 1.
[40] Ebd.: 1.
[41] Flohr und Pfingsten 2002: 107.
[42] Lemnitzer und Zinsmeister 2006: 61f.

Tagging). Darauf kann eine syntaktische Annotation aufsetzen. *POS-Tagging* wird folgendermaßen definiert:

> „Im Englischen heißt die Annotation morphosyntaktischer Merkmale auch *Grammatical Tagging*, *Part-of-Speech Tagging* (kurz: *POS Tagging*) oder einfach *Tagging*. Ein *Tag* (…) ist ein Label, das dem einzelnen Wort zugeordnet wird und dessen grammatikalische Klasse angibt. Das Wortartenlabel erlaubt die Disambiguierung mehrdeutiger Wortformen (Homographen), insofern sie verschiedenen Wortarten angehören. Die Liste aller verwendeten Wortartenlabel ist ein *Tagset*."[43]

Auf der morphosyntaktischen Ebene findet nicht nur die Klassifikation der Wortarten statt, sondern auch die Analyse der Flexionsmorphologie. Dazu werden Informationen zu Kategorien wie „*Kasus, Genus, Numerus, Person, Tempus* und *Modus*"[44] erfasst. Die Annotation von Flexionsmorphologie wird dabei „(…) vom reinen Wortarten-Tagging unterschieden. Hierzu wird das Token analysiert und auf seine Grundform, das Lemma zurückgeführt, die Analyse liefert gleichzeitig morphologische Informationen."[45] Auf der nächsthöheren Ebene erfolgt die syntaktische Annotation; hierzu gehört, wie sich Wörter zu größeren Einheiten wie Phrasen und Satzglieder zusammenfügen. Satzübergreifende Beziehungen betreffen den Bereich der Textgrammatik.

Bei der syntaktischen Annotation kann man zwischen einer partiellen und einer vollständigen Analyse unterscheiden. Dies bezieht sich auf die Tiefe der Annotation. Das auch als *Chunking* bezeichnete *Partial Parsing* ermöglicht es, „Teilstrukturen mit relativ hoher Qualität zu analysieren, ohne dass man über die Gesamtstruktur des Satzes spekulieren muss."[46] Das Chunking kann nicht nur als „automatischer Vorverarbeitungsschritt einer vollständigen syntaktischen Analyse"[47] angesehen werden, sondern auch als eigenständige Annotation fungieren, wie beispielsweise im *Tübinger Partiell Geparsten Korpus des Deutschen*.

Ein Korpus, dessen Primärdaten syntaktisch annotiert sind, wird als *Treebank* bzw. *Baumbank* bezeichnet. Das liegt daran, dass die Struktur der Sätze oft durch Baumgraphen repräsentiert wird:

> „Baumbanken als spezielle Form von Textkorpora (…) sind ein fester Bestandteil der Computerlinguistik geworden, da sie detaillierte linguistische Information kodieren. Unter einer Baumbank wird eine Sammlung von Einheiten (meist

[43] Lemnitzer und Zinsmeister 2006: 66.
[44] Ebd.: 71.
[45] Ebd.: 71.
[46] Ebd.: 79.
[47] Ebd.: 79.

> Sätzen) verstanden, deren syntaktische Satzstruktur (…) annotiert ist. Der Begriff Baumbank verweist zudem darauf, dass die Satzstruktur meist in Form einer Baumstruktur kodiert ist."[48]

Jeder Baum hat einen eindeutigen Wurzelknoten (*root node*). Die Äste verzweigen sich normalerweise wohlgeordnet, das heißt, keine Äste oder Kanten (*edges*) überschneiden sich; außerdem besitzt jeder Knoten (*node*) einen eindeutigen Mutterknoten. Wenn sich die Kanten eines Baumes überkreuzen dürfen, spricht man genau genommen nicht mehr von Baumgraphen, sondern von allgemeineren Graphenstrukturen – doch auch für solche Korpora ist der Begriff Baumbank üblich (beispielsweise werden in der TIGER-Baumbank *überkreuzende Kanten* zugelassen).

Die einzelnen Wörter eines Satzes bilden in einem Baumgraph die Blätter des Baumes, die auch als *terminale Knoten* bezeichnet werden, da sie sozusagen die finalen Elemente der Struktur sind. Die weiteren Knoten oder Verzweigungspunkte heißen *nicht-terminale Knoten*. Um syntaktische Beziehungen darstellen zu können, die sich nicht durch eine wohlgeordnete Verzweigung in einem Korpus repräsentieren lassen, ist es möglich, eine zusätzliche Ebene für *sekundäre Kanten* einzuführen, die nicht Bestandteil der eigentlichen Baumstruktur ist, sondern gewissermaßen quer zu dieser verläuft. Als Beispiel für den Einsatz von sekundären Kanten dient wieder die TIGER-Baumbank.[49]

Wie an verschiedenen Beispielen gezeigt wurde, wird durch Annotation dem Korpus-Benutzer die Arbeit erleichtert, indem mehrdeutige Lesarten vor allem auf den grammatischen Ebenen der Morphologie und Syntax disambiguiert werden:

> „Die Beispiele haben gezeigt, dass es sinnvoll ist, Korpusdaten mit linguistischen Interpretationen anzureichern, indem man zum Beispiel Wortarten, syntaktische Phrasen oder grammatische Funktionen annotiert. Diese Annotationen machen Korpusuntersuchungen effizienter, indem präzisere Anfragen gestellt werden können und komplexere Phänomene überhaupt erst abfragbar gemacht werden."[50]

Letztlich dient syntaktische Annotation dazu, nicht nur die Existenz bestimmter syntaktischer Konstruktionen nachzuprüfen, son-dern auch neue oder unbekannte Konstruktionen aufspüren zu können.

---

[48] Lezius 2001: 414.
[49] Vgl. Tylman und Hinrichs 2004: 75.
[50] Lemnitzer und Zinsmeister 2006: 62.

3.1.3. Abhängigkeit von grammatischen Theorien

Bei der Planung von Annotationsprojekten wird meistens der Terminus *Theorieneutralität* aufgeworfen. Auch das Projekt *Diachrone Syntax Deutsch* „(…) ist keiner bestimmten momentan aktuellen theoretischen Richtung verpflichtet."[51] Eine Forderung nach absoluter Theorieneutralität kann jedoch in der Praxis nicht erreicht werden, denn unvermeidbar „fließen in das zu Grunde gelegte Annotationsschema die Grundannahmen einer bestimmten Theorie ein, die eine Verwendung der annotierten Daten unter anderen theoretischen Vorzeichen erschwert."[52] Statt Theorieunabhängigkeit sollte das pragmatische Ziel *Wiederverwendbarkeit* formuliert werden: „Diese ist gewährleistet, solange ein linguistisch annotiertes Korpus so umgewandelt werden kann, dass es ganz oder teilweise die Anforderungen erfüllt, die eine andere zu Grunde liegende Theorie an die linguistische Annotation stellt (…)."[53] Auch Wolfgang Lezius spricht über Theorieneutralität im Sinne von Wiederverwertbarkeit:

> „Für den Entwurf eines Datenmodells und der darauf basierenden Annotation sollte stets berücksichtigt werden, inwieweit diese mit möglichst vielen syntaktischen Theorien verträglich sind. Theorieunabhängigkeit bedeutet Wiederverwertbarkeit: und die hohen Kosten für die Erstellung von Baumbanken sind nur dann zu rechtfertigen, wenn möglichst viele Forschungsrichtungen davon profitieren können."[54]

Syntaktische Annotation ist also stets abhängig von einer zu Grunde gelegten, sprachwissenschaftlichen Theorie, da bei der Gestaltung eines Annotationsschemas unweigerlich theoretische Grundannahmen einfließen. Die Strukturen eines Satzes können auf unterschiedliche Weise betrachtet werden. Das Ergebnis einer syntaktischen Annotation hängt davon ab, welche linguistische Theorie bzw. welche linguistischen Theorien man zur Unterstützung heranzieht:

> „Syntaktische Annotation beschreibt die *syntaktische Struktur* sprachlicher Äußerungen. Da hier die Größe der einzelnen Bausteine der Strukturen weniger fest ist als beim POS-Tagging, bei dem jeweils ein Wort einer Klasse zugewiesen wird, ist bei der syntaktischen Annotation das Ergebnis weit abhängiger von der zu Grunde gelegten Theorie."[55]

---

[51] Schmid 2007: 52.
[52] Tylman und Hinrichs 2004: 224.
[53] Ebd.: 224.
[54] Lezius 2001: 416.
[55] Tylman und Hinrichs 2004: 221.

Im Folgenden soll der Unterschied zwischen einer syntaktischen Annotation, die der Konstituentenstrukturanalyse folgt, und einer Annotation nach der Dependenztheorie erläutert werden; außerdem wird eine Mischform der beiden Theorien – ein sogenanntes hybrides Modell – vorgestellt. Als letztes wird eine Annotation nach der Theorie der topologischen Felder diskutiert.

Für syntaktische Annotationen lassen sich generell zwei strukturelle Modelle unterscheiden: die Konstituentenstruktur und die Dependenzstruktur. Beide Ansätze gehen davon aus, dass Sätze hierarchisch strukturiert sind; sie unterscheiden sich jedoch hinsichtlich der Elemente, die in einer Hierarchie angeordnet werden. Während man nach der Konstituentenstruktur „Konstituenten, also abstrakte Einheiten, die jeweils ein oder mehrere Wörter repräsentieren (z.B. VP, NP (…))“[56], betrachtet, konzentriert man sich nach der Dependenzstruktur auf die Abhängigkeiten der einzelnen Wörter. Die Tatsache, dass die beiden Ansätze Unterschiedliches leisten, ist unbestritten – doch die Frage, was genau die Leistung von Konstituenz und Dependenz unterscheidet, kann nur vage beantwortet werden:

> „Konstituenz und Dependenz benennen (…) allgemeine Prinzipien syntaktischer Beschreibung. Es bleibt eine offene Frage, wieweit diese Ansätze kompatibel sind, oder ob gar Theorien eines Formats in das andere konvertiert werden können. Die Tatsache, daß immer wieder dependentielle Teile in konstitutionellen Syntaxen emuliert werden (und umgekehrt), spricht eher dafür, daß die beiden Ansätze letztlich Unterschiedliches leisten.“[57]

Die Konstituentenstrukturanalyse lässt sich auf den amerikanischen Strukturalismus zurückführen – Zellig Harris ist ein bekannter Vertreter dieser theoretischen Richtung – und basiert auf folgender Grundannahme:

> „Man nimmt an, dass Sätze aus hierarchisch geschachtelten Untereinheiten bestehen, die man zum Beispiel durch Klammerung markieren kann. Diese Untereinheiten sind Sequenzen von zusammenhängenden Wörtern, die als *Konstituenten* bezeichnet werden.“[58]

Als Beispiel für ein Korpus, das nach der Struktur der Konstituenten annotiert ist, lässt sich die amerikanische *Penn Treebank* „im Repräsentationsformat der ersten Projektphase“[59] anführen. Die Konstituentenstrukturanalyse des Teilsatzes *ein einfaches Beispiel geben* sieht folgendermaßen aus:

---

[56] Lemnitzer und Zinsmeister 2006: 76.
[57] Heringer 1996: 28.
[58] Lemnitzer und Zinsmeister 2006: 76.
[59] Ebd.: 76.

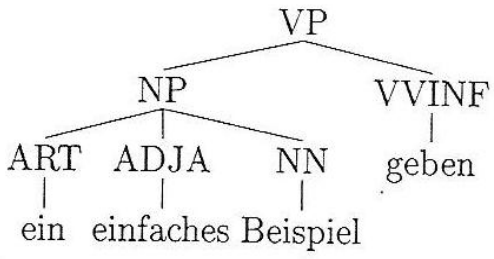

Abbildung 1. Baumstruktur einer Konstituentenstrukturanalyse.[60]

Die Nominalphrase (NP) besteht aus einem Artikel (ART), einem Adjektiv (ADJA) und einem Nomen (NN); die Verbalphrase (VP) setzt sich aus der NP und einem infiniten Vollverb zusammen.

Nach einer Dependenzgrammatik hingegen wird die Hierarchie eines Satzes aus Abhängigkeiten (*Dependenzen*) von einzelnen Wörtern gebildet. Diese Dependenzen modelliert man, indem man jeweils zwei Wörter miteinander verknüpft. Die Verknüpfungen sind gerichtet, das bedeutet, dass es „immer ein *Regens* und ein davon abhängiges *Dependens*“[61] gibt. In der folgenden Abbildung regiert das Verb *geben* das Substantiv *Beispiel*; *Beispiel* wiederum regiert den Artikel (Determinator) *ein* und das Attribut *einfach*. Die Pfeilspitzen zeigen jeweils auf das Regens:

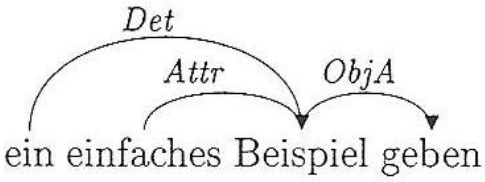

Abbildung 2. Baumstruktur einer Dependenzstrukturanalyse.[62]

Mit dem Konzept der Dependenzstruktur ist die Angabe von grammatischen Funktionen verbunden: „Normalerweise stehen die abhängigen Elemente in einer bestimmten *grammatischen Funktion* zum Regens, im Beispiel sind es *Det(erminator), Attr(ibut)* und *Akkusativobjekt (ObjA)*.“[63] Dies nennt man auch eine funktionale Analyse. Als Beispiel für ein Korpus, das nach der Dependenzstruktur annotiert ist, lässt sich die tschechische *Prague Dependency Treebank* anführen.

---

[60] Lemnitzer und Zinsmeister 2006: 76.
[61] Ebd.: 76.
[62] Ebd.: 77.
[63] Ebd.: 76.

Da in einer Konstituentenstrukturanalyse normalerweise keine grammatischen Funktionen angegeben werden, sondern nur syntaktische Kategorien, spricht man von einem *hybriden Modell*, wenn in einem Korpusprojekt Konstituentenstrukturanalyse und die Angabe grammatischer Funktionen gemeinsam auftreten. „Viele der Baumbanken, die eine konstituentenbasierte Grundarchitektur besitzen, fallen in die Klasse der hybriden Modelle, weil sie auf die eine oder andere Art auch funktionale Informationen kodieren."[64] Als Beispiel für ein Korpus, das mit einer hybriden Baumstruktur annotiert ist, lässt sich die *Penn Treebank* ab der zweiten Phase ihrer Erstellung anführen: „Ab Phase 2 wird auch dort ein hybrides Annotationsschema verwendet, das z.B. Subjekte und adverbiale PPs mit funktionalen oder semantischen Labeln auszeichnet."[65] In Abbildung 3 wird der Teilsatz *ein einfaches Beispiel geben* in einer hybriden Baumstruktur dargestellt. Die syntaktischen Kategorien erscheinen dabei an den Knotenpunkten, während die Kanten mit den grammatischen Funktionen versehen sind. „Die Köpfe der VP und NP sind [dabei] zusätzlich als *H(ead)d* markiert."[66] Der Kopf einer Phrase entspricht dem Regens: „Ein Wort bildet den Kopf der Phrase, es gibt ihr auch den Namen. So sprechen wir von Nominalphrasen (mit Nomen als Kopf), Präpositionalphrasen (mit Präposition als Kopf) usw. Die übrigen Wörter hängen vom Kopf ab: Er regiert sie, ist ihr Regens, sie sind die Dependentien (…) des Kopfes."[67]

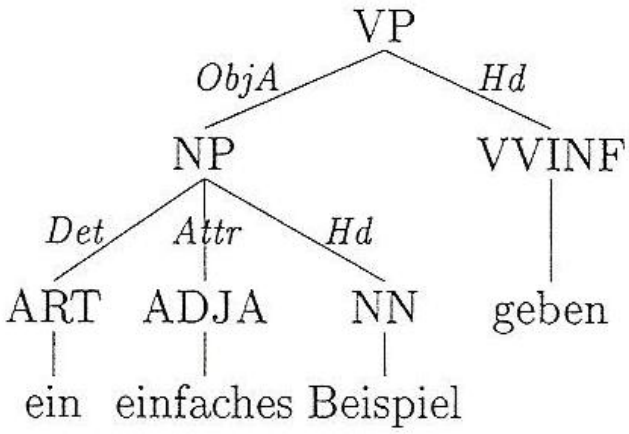

Abbildung 3. Hybride Baumstruktur.[68]

[64] Lemnitzer und Zinsmeister 2006: 77.
[65] Ebd.: 77.
[66] Ebd.: 77.
[67] Engel 2004: 15.
[68] Lemnitzer und Zinsmeister 2006: 77.

Das *DiSynDe*-Korpus wird nach einer Mischform aus konstituentenbasiertem Ansatz mit der Angabe grammatischer Funktionen annotiert. Die Gliederung der Sätze nach Konstituenten und die Angabe der grammatischen Funktionen sind Bestandteil der Annotationsvorgaben. Ferner hat die Konstituentenstrukturanalyse auf dem Gebiet der Computerlinguistik eine längere Tradition als der dependenzorientierte Ansatz – „vor allem aufgrund der richtungsweisenden Arbeiten von Noam Chomsky (…).“[69] Aus Sicht der Sprachwissenschaft bietet die Konstituentenstrukturanalyse darüberhinaus eine tiefere Aufbereitung der Primärdaten:

> „Die Konstitutionssyntax beschreibt eine Struktur, indem sie Strings stufenweise in Teile segmentiert. Sie basiert auf der Teil-Ganzes-Relation; ihre Strukturen sind tief (…). Die Dependenzsyntax basiert auf der Dependenz der Teile. Sie beschreibt die Satzstruktur über die Abhängigkeit der Teile; ihre Strukturen sind flacher (…).“[70]

Für *DiSynDe* ist es außerdem von Bedeutung, dass man anhand des Korpus die Stellung des Prädikats erfahren kann. Es stellt sich die Frage, ob eine explizite Annotation der Sätze nach der Theorie der topologischen Felder erforderlich ist:

> „Für das Deutsche existiert (…) eine Beschreibung der Satzstruktur, die von den Beziehungen der einzelnen Konstituenten eines Satzes zunächst abstrahiert: die Beschreibung der Gliederung des Satzbaus in *topologische Felder*. Diese Art der Charakterisierungen deutscher Satztypen folgt einer langen Tradition empirischer Untersuchungen deutscher Syntax und erlaubt es, unabhängig von der relativ freien Wortstellung des Deutschen Generalisierungen über die Struktur von Sätzen vorzunehmen.“[71]

In deutschen Aussagesätzen steht das finite Verb immer an der zweiten Position im Satz, während infinite Verben oder abtrennbare Bestandteile eines Verbs auf die letzte Stelle des Satzes verschoben werden. „Man spricht hier von der verbalen Klammer des Deutschen“[72] bzw. von einem Satzrahmen, der vor allem durch ein komplexes – also aus mehreren Wörtern bestehendes – Prädikat sichtbar wird.

Aus der verbalen Satzklammer leitet sich die Terminologie zur Einteilung des deutschen Satzes in topologische Felder ab: *Vorvorfeld* (wird zum Beispiel von Konjunktionen besetzt), *Vorfeld* (Position vor der linken Satzklammer), *linke Satzklammer* (zum Beispiel ein finites Verb), *Mittelfeld* (Position zwischen den Satzklammern), *rechte*

[69] Langer 2004: 234.
[70] Heringer 1996: 27f.
[71] Tylman und Hinrichs 2004: 221.
[72] Kessel und Reimann 2005: 9.

Satzklammer (zum Beispiel ein infinites Verb), *Nachfeld* (Position nach der rechten Satzklammer, entspricht einer Ausklammerung).[73]

Als Beispiel für ein deutsches Korpus, das nach der Theorie der topologischen Felder annotiert ist, dient die *Tübinger Baumbank des Deutschen / Zeitung (TüBa-D/Z)*. Abbildung 4 zeigt den Beispielsatz *Wir sind begeistert!*, dessen Struktur in Vorfeld (VF), linke Verbklammer (LK) und Mittelfeld (MF) gegliedert ist; erst daran werden die Konstituenten (NX für Nominalphrase, VXFIN für finite Verbalphrase, ADJX für Adjektivphrase) angebunden, an deren Kanten wiederum die grammatischen Funktionen (ON für Nominativobjekt, HD für Kopf, PRED für Prädikat) ausgezeichnet sind:

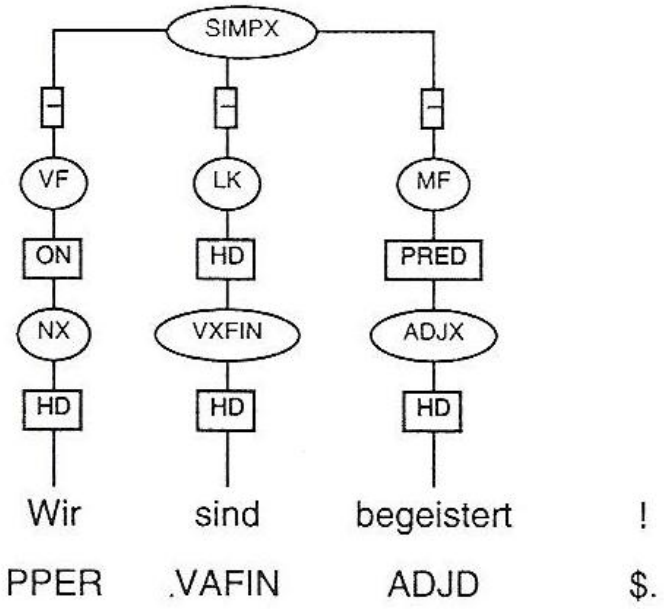

Abbildung 4. Baumstruktur mit topologischen Feldern.[74]

Der feste Satzrahmen des Deutschen entwickelt sich erst über die Jahrhunderte hinweg; erst seit dem 14. bzw. 15. Jahrhundert besteht „eine zunehmende Tendenz zum Gebrauch des vollen Satzrahmens"[75]. Aus diesem Grund ist es nicht zu empfehlen, sämtliche Sätze des *DiSynDe*-Korpus in das starre Schema eines Satzrahmens einfügen zu wollen. Dadurch verlagert sich das Problem auf die Seite der Korpusanfragesprache: Wenn man etwa im *DiSynDe*-Korpus Belege für *komplexe Prädikate mit dem Finitum in Letztstellung* finden will, muss man ein entsprechend mächtiges Suchwerkzeug benutzen, das es ermöglicht, in annotierten Sätzen nach solchen Mustern zu suchen.

---

[73] Vgl. Kessel und Reimann 2005: 10.
[74] Lemnitzer und Zinsmeister 2006: 83.
[75] Schmidt 2007: 138.

Als Beispiel für eine Mustersuche soll die verbreitete Software *Corpus Query Processor* (CQP) dienen: „Mit Hilfe geeigneter Suchausdrücke kann (…) versucht werden, linguistisches Wissen aus dem Korpus zu gewinnen."[76] CQP erlaubt die Suche nach einzelnen Tokens; zur näheren Spezifizierung werden Attribut-Wert-Paare durch boolesche Ausdrücke miteinander verknüpft. Auf Ebene der Tokens ist auch der Einsatz regulärer Ausdrücke möglich. Die erste Beispielanfrage sucht nach Phrasen, „die aus einem optionalen Artikel (ART), beliebig vielen Adjektiven (ADJA) und dem Nomen *Mann* (NN) bestehten"[77]:

```
[pos="ART"]? [pos="ADJA"]* [pos="NN" & word="Mann"];
```

Die zweite Beispielanfrage ermittelt Ergebnisse für Verb-Nomen-Kollokationen. „Die Einschränkung des Suchkontextes auf einen Satz (durch *within 1 s*) verhindert, dass sich der Match über eine Satzgrenze hinaus ausdehnt. Mit der Wortart VVFIN wird ein finites Verb bezeichnet. Nach dem finiten Verb können beliebig viele Token folgen, die kein Nomen bezeichnen. Als Abschluss der gesuchten Struktur folgt ein Nomen."[78]

```
[pos="VVFIN"] [pos!="NN"]* [pos="NN"] within 1 s;
```

Auf ähnliche Weise können mit CQP auch Anfragen für die Suche nach komplexen Prädikaten formuliert werden, bei denen das finite Verb an einer bestimmten Position im Satz steht.

CQP kann darüberhinaus auch Anfragen an syntaktisch annotierte Korpora leisten – eine letzte Beispielanfrage demonstriert die Suche nach Artikel-Nomen-Paaren, die als Nominalphrase (NP) ausgezeichnet sind:[79]

```
<np> [pos="ART"] [pos="NN"] </np>;
```

Doch die Software weist bezüglich syntaktischer Annotation auch Beschränkungen auf, denn „CQP erlaubt keine Abfrage rekursiver Strukturen, d.h. beispielsweise von Nominalphrasen *<np>*, die wieder Nominalphrasen einbetten."[80] Die Dissertation *Ein*

[76] Lezius 2002: 35.
[77] Ebd.: 34.
[78] Ebd.: 35.
[79] Vgl. ebd.: 35.
[80] Ebd.: 35.

Suchwerkzeug für syntaktisch annotierte Korpora von Wolfgang Lezius stellt weitere Suchwerkzeuge für Baumbanken vor.[81]

Im Rahmen der Dissertation wird das Suchwerkzeug *TIGERSearch* zur „Nutzbarmachung der TIGER-Baumbank"[82] entwickelt; zusätzlich entsteht zur Formulierung von Suchanfragen eine eigene Korpusbeschreibungssprache. Um das Arbeiten mit der Beschreibungssprache zu vereinfachen, besteht die Möglichkeit, komplexe und häufig gebrauchte Suchanfragen als *Templates* zu definieren:

> „Wiederkehrende Ausdrücke können gekapselt und über einen stellvertretenden Ausdruck nutzbar gemacht werden. Für die TIGER-Beschreibungssprache werden Templates als Relationen zur Verfügung gestellt. (…) Die Schreibweise und Verarbeitung orientiert sich an der Programmiersprache Prolog."[83]

Beim Download der frei erhältlichen Software *TIGERSearch*[84] findet sich ein Beispiel-Template (*VerbPositions.tig*), das die einfache Abfrage von Verbpositionen nach topologischen Feldern ermöglicht; das Template findet sich auf der folgenden Seite. Um Sätze ohne festen Satzrahmen bzw. Sätze, deren Verbstellung nicht der Satzklammer entspricht, als Suchergebnisse anzeigen zu lassen, müssen entsprechende Suchanfragen formuliert werden.

Abschließend zu diesem Punkt lässt sich sagen, dass eine vergleichbare Vorgehensweise – nämlich der Verzicht auf die explizite Annotation der Verbstellung – für *DiSynDe* anzustreben ist, sowohl aus linguistischen, als auch aus technischen Gründen: Sätze ohne festen Satzrahmen müssen nicht nach einem möglicherweise unpassenden Schema annotiert werden; außerdem erhöht sich die Wiederverwendbarkeit bzw. die Theorieneutralität des Annotationsschemas.

---

[81] Vgl. Lezius 2002: 33-66.
[82] Ebd.: 10.
[83] Ebd.: 96.
[84] TIGERSearch 2003.

```
//      File: VerbPositions.tig
//    Author: Esther Koenig, George Smith
//   Purpose: Basic topological structure: verb positions
//   Created: Thu Sep  6 13:38:54 2001 (esther)
//  Modified: Wed Nov  6 17:05:47 2002 (esther)

////////////////////////////////////////////////
// Primitive notion of verb first sentence
//
// Literally, nothing comes before the finite verb.

verbFirst(#s) <-
   #s:[cat = "S"] &
   #v:[pos = finite] &
   #s >@l #v ;

////////////////////////////////////////////////
// Primitive notion of verb last sentence
//
// Literally, nothing comes after the finite verb.

verbLast(#s) <-
   #s:[cat = "S"] &
   #v:[pos = finite] &
   #s >@r #v ;

////////////////////////////////////////////////
// Primitive notion of verb second sentence
//
// Literally, something comes before and after the finite verb.

verbSecond(#s,#v) <-
   #s:[cat = "S"] &
   #v:[pos = finite ] &
   #s > #v &
   #s !>@l #v &
   #s !>@r #v ;


////////////////////////////////////////////////
// Verb second, accusative object in Vorfeld
verbSecondVorfeldAcc(#s) <-
   verbSecond(#s,#v) &
   #s >OA #oa &
   #oa .* #v ;

////////////////////////////////////////////////
// Verb second, dative in Vorfeld
verbSecondVorfeldDative(#s) <-
   verbSecond(#s,#v) &
   #s >DA #da &
   #da .* #v ;

////////////////////////////////////////////////
// Verb second, subject not in Vorfeld
verbSecondVorfeldNonSubj(#s) <-
   verbSecond(#s,#v) &
   #s >SB #sb &
   #v .* #sb;
```

3.1.4. Ebenen der syntaktischen Annotation nach EAGLES

Die *Expert Advisory Group on Language Engineering Standards* (EAGLES) ist eine Initiative der Europäischen Kommission im Rahmen des Programmes *Linguistic Research and Engineering*.[85] EAGLES hat nicht nur den *Corpus Encoding Standard* (CES) entwickelt, sondern auch Empfehlungen für die syntaktische Annotation von Textkorpora aufgestellt (*Recommendations for the Syntactic Annotation of Corpora*).[86] In diesem Dokument werden verschiedene Ebenen der syntaktischen Annotation unterschieden und hierarchisch angeordnet. Im Anschluss daran werden Empfehlungen auf drei Stufen (obligatorische, empfohlene und optionale Annotationen) formuliert; dadurch entsteht kein starrer Standard, sondern lässt Raum für die Entwicklung neuer Möglichkeiten.

(1) Bracketing of segments;
(2) Labelling of segments;
(3) Marking of dependency relations (…);
(4) Indicating functional labels, such as Subject and Object;
(5) Marking subclassification of syntactic segments;
(6) Deep or 'logical' information;
(7) Information about the rank of a syntactic unit (e.g. Clause, Phrase, Word);
(8) Special syntactic characteristics of spoken language.[87]

Im Folgenden werden die verschiedenen Ebenen der syntaktischen Annotation erläutert. Ebene 1, die Klammerung von Segmenten, entspricht auch der Visualisierung von Satzstrukturen als Baumgraphen. Dadurch wird die komplexe, hierarchische Struktur, die sich hinter der Linearität von Sprache verbirgt, sichtbar und maschinell lesbar gemacht:

„[[He] [walked [into [the garden]]]]"[88]

[85] EAGLES 2003.
[86] EAGLES 1996a.
[87] Vgl. Kahrel et al. 1997: 237f.
[88] Ebd.: 238.

Auf der zweiten Ebene werden die Segmente oder Konstituenten, die durch die Klammerung entstanden sind, nach ihrer formalen Kategorie wie Nominalphrase, Verbalphrase oder Relativsatz gekennzeichnet.[89] Daraus ergibt sich folgender Beispielsatz:

„[S [NP He NP] [VP walked [PP into [NP the garden NP] PP] VP] S]"[90]

Auf der dritten Ebene wird die Annotation nach einer Dependenzgrammatik angesiedelt. Bei dieser Vorgehensweise bleiben die gebildeten Konstituenten außer Acht, und man modelliert Abhängigkeiten zwischen einzelnen Wörtern. Es wird darauf hingewiesen, dass man Ebene 2 und Ebene 3 als Alternativen behandeln kann; allerdings besteht auch die Möglichkeit, beide Ebenen in dieselbe Annotation aufzunehmen, da sie unterschiedliche Arten von linguistischen Informationen kodieren.

Die vierte Ebene betrifft die Auszeichnung von Annotationseinheiten mit grammatischen Funktionen wie Subjekt oder Objekt. Auf Ebene 5 erfolgt die Subklassifizierung der Segmente: „This means assigning attribute values to constituents such as clauses or phrases, e.g. marking a Noun Phrase as singular, or a Verb Phrase as past tense."[91] Die Subklassifizierung betrifft also unter anderem auch morphosyntaktische Informationen, die bereits auf Ebene der Wortarten markiert werden können.

Auf der sechsten Ebene werden unterschiedliche Arten von logischen Beziehungen annotiert: „This includes a variety of syntactic phenomena, such as coreferentiality (for example in control structures), cross-reference (or substitution), ellipsis, traces, and syntactic discontinuity."[92] Ebene 7 bestimmt die Rangfolge syntaktischer Einheiten; ein Satz etwa hat einen höheren Rang als ein Teil- oder Gliedsatz, danach folgen Phrasen und Wörter. Allerdings ergibt sich eine solche Rangfolge bei einer XML-Kodierung implizit. Als Beispiel für diese Ebene führt EAGLES das SUSANNE-Korpus an: „These ranks are found, for example, in the SUSANNE annotation scheme and are used as a basis for deciding whether to represent one-word constituents in the labelled bracketing notation (…)."[93]

Die achte und letzte Ebene betrifft die speziellen syntaktischen Eigenschaften gesprochener Sprache: „Spoken language corpora show a range of phenomena that do not normally occur in written language corpora, such as blends, false starts, reiterations, and

[89] Vgl. Kahrel et al. 1997: 238.
[90] Ebd.: 238.
[91] Ebd.: 238.
[92] Ebd.: 239.
[93] EAGLES 1996b.

filled pauses. In syntactic annotation, it has to be decided whether to include such phenomena in a parse tree, and if so, how."[94]

Nun werden die drei Stufen (obligatorisch, empfohlen und optional) der syntaktischen Annotations-Richtlinien vorgestellt. Diese dreistufige Unterscheidung ergibt sich aus einem ähnlichen Dokument von EAGLES zur morphosyntaktischen Annotation:

> „EAGLES has so far undertaken to propose sets of provisional guidelines for the morphosyntactic and syntactic annotation of corpora. To counter the danger of overrigidity (…) three levels of constraint on annotation practices have been suggested. These three levels, obligatory, recommended, and optional annotations, are naturally different for morphosyntactic and syntactic annotation, but in both types of annotation the three levels are distinguished."[95]

Auf der obligatorischen Stufe der syntaktischen Annotation werden keine Richtlinien ausgesprochen, da es zu verschiedene Arten und zu viele Kombinationsmöglichkeiten von syntaktischen Annotationen gäbe. Die erste Ebene, nämlich die Klammerung von Segmenten, könne man als obligatorisch ansehen, und sie sei es auch für Annotationen nach einer Konstituentenstrukturgrammatik. „However, as we have seen, there are dependency-based schemes that do not actually group together the words making up constituents (…), and these must still undoubtedly be regarded as a useful form of syntactic annotation."[96]

Auf der zweiten Stufe, den empfohlenen Annotationen, werden Kategorien für nicht-terminale Segmente aufgeführt: Satz (S), Teilsatz (CL für Clause), Nominalphrase (NP), Verbalphrase (VP), Präpositionalphrase (PP), Adverbialphrase (ADVP) und Adjektivphrase (ADJP). Außerdem wird das Problem von Koordinations-Phänomenen angesprochen; EAGLES schlägt vor, koordinierte Phrasen auf der gleichen Ebene zu annotieren, die Konjunktion dazwischen:

„[NP [NP John NP] and [NP Mary NP] NP]"[97]

Es werden auch weitere Lösungsmöglichkeiten erwähnt, nämlich die Indizierung der koordinierten Phrasen mit CO (für Coordination), oder die Einführung eines eigenen CO-Tags. In der TIGER-Baumbank werden entsprechende Tags und sekundäre Kanten für die

---

[94] Kahrel et al. 1997: 239.
[95] Ebd.: 235.
[96] Ebd.: 239f.
[97] EAGLES 1996c.

Annotation von Koordination eingesetzt.[98] Zum Abschluss der empfohlenen Annotationen wird auf ein grundsätzliches Problem hingewiesen:

> „Although these non-terminal categories are widely recognized, it is not easy to agree on precisely how they are instantiated in texts. The documentation accompanying a corpus should therefore give a clear account of how these constituents are defined, with sufficient attention to problem cases."[99]

Die dritte Stufe der EAGLES-Richtlinien umfasst zahlreiche optionale syntaktische Annotationen: die Subkategorisierung von Sätzen nach Satztypen wie Aussage-, Imperativ- und Fragesätzen; die Subkategorisierung von Gliedsätzen nach syntaktischen Gesichtspunkten (zum Beispiel Relativ- oder Adverbialsätze); die Subkategorisierung von Phrasen nach Person, Numerus, Kasus, Tempus, Modus und Aspekt; und die Angabe grammatischer Funktionen wie Subjekt oder Objekt.[100] Desweiteren wird die Möglichkeit der semantischen Subkategorisierung von Phrasen und Teilsätzen aufgeführt. Eine Nominalphrase könne beispielsweise als definit oder indefinit markiert werden, oder auch als temporal oder lokal – wenn ein Korpus allerdings morphosyntaktisch annotiert wäre, befänden sich entsprechende Annotationen oft bereits auf Ebene des POS-Tagging. Ebenso können Gliedsätze nach semantischen Gesichtspunkten weiter untergliedert werden; beispielsweise können als Adverbialsätze klassifizierte Gliedsätze darüberhinaus als temporal oder konditional eingestuft werden.[101] Als letzte Art von optionalen Annotationen wird die Kodierung von logischen Beziehungen (nach Ebene 6) aufgeführt. Dies betrifft beispielsweise Bewegungsphänomene, bei denen Elemente nicht an ihrer eigentlichen Position aufzufinden sind:

> „A man was at the door who wanted to speak to you."[102]

In der TIGER-Baumbank ist die Annotation ähnlicher Phänomene im Deutschen mit kreuzenden Kanten möglich.[103]

Das Feld der syntaktischen Annotation ist in diesem Punkt durch die Vorstellung der *Recommendations of Syntactic Annotation of Corpora* von EAGLES näher beleuchtet worden. Im nächsten Abschnitt sollen die verschiedenen Ebenen von Annotation mit den Annotationsvorschriften von *Diachrone Syntax Deutsch* in Verbindung gebracht werden.

---

[98] Vgl. Lezius 2002: 8f.
[99] Kahrel et al. 1997: 240.
[100] Vgl. ebd.: 240.
[101] Vgl. EAGLES 1996d.
[102] EAGLES 1996e.
[103] Vg. Lezius 2002: 8f.

3.1.5. Annotationsebenen für *Diachrone Syntax Deutsch*

Text	textelement	zitat/direkte rede/appell	
	diskurs		
	referenz		
	textanschluss		
Komplexer Satz	makrostrukt	matrix	
	einleit		
	position		
	verbstellung	erst	
	modus	imp	
Einfacher Satz		prädsimpl	obj.akk
Wortgruppe	phrase grob	vp	np
	phrase fein1		
	phrase fein2		
	vk	voll	
	flexion	2sg.imp	akk.sg.
Wortart		vb	p.pers1
Quelle		*Hor*	*mich*

Abbildung 5. Ausschnitt der *DiSynDe*-Annotationsvorschriften.[104]

In diesem Punkt werden die nach den Analysegruppen gegliederten Annotationsvorschriften auf die einzelnen Annotationsebenen übertragen, da bei der Aufteilung in *Wortgruppe, Einfacher Satz, Komplexer Satz* und *Text* teilweise Überlappungen auftreten, welche die Strukturierung der Informationen für die Kodierung der Annotation erschweren. Ferner müssen die Annotationsebenen voneinander abgegrenzt werden, damit durch das resultierende Schema keine linguistischen Merkmale doppelt – das heißt auf mehreren Ebenen bzw. von verschiedenen Analysegruppen – kodiert werden. Beispielsweise wird nach den Annotationsvorschriften der Modus *Imperativ* des Verbs *Hor* sowohl von der Wortgruppe als auch von der Gruppe Komplexer Satz markiert (siehe Abbildung 5).

Auf Ebene der morphosyntaktischen Annotation erfolgt sowohl die Bestimmung der Wortarten als auch die flexionsmorphologische Analyse der Wörter; eng verknüpft mit der flexionsmorphologischen Analyse ist die Lemmatisierung (das Verknüpfen

[104] Die vollständigen Annotationsvorschriften (mit weiteren annotierten Sätzen) finden sich im Anhang 1.

flektierter Wörter mit ihrer Grundform). *Flexion*, ein Unterpunkt der Wortgruppe (vgl. Abbildung 5), wird der morphosyntaktischen Annotationsebene zugeordnet.

Die Wortgruppe, Einfacher Satz und Komplexer Satz betreffen die eigentliche syntaktische Annotation. Die Wortgruppe unterteilt sich in die vier Ebenen *Phrase grob, Phrase fein1, Phrase fein2* und *Verbalkomplex*. Unter *Phrase grob* erfolgt die Klammerung und Beschriftung der Konstituenten (Verbal-, Nominal-, Adverbial-, Präpositionalphrase). Dies entspricht den ersten beiden Ebenen der syntaktischen Annotation nach EAGLES.

Unter *Phrase fein1* und *Phrase fein2* erfolgt eine Subklassifikation der Konstituenten; beispielsweise wird eine Präpositionalphrase unterteilt in die Präposition und eine Nominalphrase. Die Struktur der Konstituenten muss sich demnach rekursiv annotieren lassen, das bedeutet, dass Phrasen sowohl Segmente des gleichen als auch eines anderen Typs enthalten können müssen. Eine NP kann weiter in Kopf und Determinator gegliedert werden (grammatische Funktionen); desweiteren sind an dieser Stelle flexionsmorphologische Angaben wie *finites Verb* zu finden, die auf die morpho-syntaktische Annotationsebene verschoben werden. Der Unterpunkt *Verbalkomplex* enthält ebenfalls morphosyntaktische Markierungen wie *Voll-, Modal-* und *Hilfsverb*.

Die Gruppe Einfacher Satz annotiert nach der Annotationstabelle auch unterschiedliche Informationen auf einer Ebene. In erster Linie werden grammatische Funktionen wie Subjekt, Prädikatsnomen, Dativ-, Akkusativ- und Präpositionalobjekt sowohl für Phrasen als auch für Gliedsätze angegeben (dies entspricht der vierten Ebene nach EAGLES). Die Annotation, ob ein einfaches oder ein komplexes Prädikat (*Prädsimpl* bzw. *Prädcompl*) vorliegt, muss nicht separat annotiert werden, sondern ergibt sich aus einer Kombination der Informationen auf Ebene der Wortarten und der Konstituenten. Außerdem werden Konstituenten nach syntaktischen und semantischen Gesichtspunkten klassifiziert (Adverbiale modal, Adverbiale instrumental, Adverbiale lokal); dies geschieht bei EAGLES auf der optionalen Stufe.

Die Gruppe Komplexer Satz annotiert unter dem Punkt *Makrostruktur* Gliedsätze (Clauses) sowohl nach Satztypen als auch nach ihrer syntaktischen Funktion. Dabei bezeichnet *Matrix* die Trägerstruktur eines komplexen Satzes bzw. den Hauptsatz – eine Information, die bei entsprechender Klammerung nicht explizit annotiert werden muss. Gliedsätze werden ferner als *indirekter Fragesatz, Relativsatz, Kausalsatz* oder *Attributsatz* markiert.

Die Auszeichnung von Satzeinleitungen ist kein einfacher Fall: teilweise kann die entsprechende Information der morphosyntaktischen Ebene (*Subjunktion*) zugeordnet werden, teilweise der Ebene der grammatischen Funktionen – beispielsweise besteht auch die Möglichkeit von komplexen Satzeinleitungen (*und darumb das*). Das bedeutet, dass

auch auf der syntaktischen Annotationsebene ein Tag zur Auszeichnung von Satzeinleitungen zur Verfügung stehen sollte.

Position bezeichnet die relative Stellung von untergeordneten Gliedsätzen (vorangestellt, eingeschaltet, nachgestellt) zum Hauptsatz; diese Information muss nicht explizit annotiert werden, da sie durch die Klammerung der Konstituenten und Gliedsätze markiert wird. Für die Ermittlung von beispielsweise vorangestellten Nebensätzen können auf der Anfrageseite passende Suchanfragen formuliert werden. Entsprechendes gilt für das Problem der *Verbstellung*, das bereits weiter oben erörtert wurde (vgl. Seite 21ff). Die Annotation des *Modus* erfolgt auf morphosyntaktischer Ebene.

Die Erläuterungen dieses Abschnitts führen zu der Übersicht auf Seite 34. Die linguistischen Informationen, die im *DiSynDe*-Korpus annotiert werden sollen, werden nach Annotationsebenen gegliedert dargestellt. Die Liste der grammatischen Funktionen der Teilsätze ist um weitere Angaben eines ausführlichen Annotationsrasters der Gruppe Komplexer Satz ergänzt.[105]

Aus dieser Einteilung lassen sich Schlussfolgerungen für die Vorgehensweise bei der Annotation ableiten. Ein paralleles Bearbeiten der morphosyntaktischen und der syntaktischen Annotationsebene scheint nicht praktikabel, da Informationen der unteren Ebene die Voraussetzung für die Annotation der höheren Ebene darstellen. Deshalb ist bezüglich dieser beiden Annotationsebenen die *Bottom-up*-Lösung vorzuziehen. Auf die morphosyntaktische Annotation kann die syntaktische durch die drei Syntaxgruppen (Wortgruppe, Einfacher Satz, Komplexer Satz) aufsetzen.

Bei der Bildung eines Syntaxgraphen ist es am zweckmäßigsten, den Satz von außen nach innen zu klammern; das heißt, die Gruppe Komplexer Satz zerteilt als erstes Sätze in Teilsätze. Anschließend annotiert die Wortgruppe die Struktur der Phrasen und weist den enthaltenen Wörtern grammatische Funktionen zu; sie markiert zum Beispiel den Determinator und den Kopf einer Nominalphrase. Als letzten Arbeitsschritt markieren die Gruppen Einfacher Satz und Komplexer Satz die fehlenden grammatischen Funktionen (zum Beispiel Subjekt für eine NP und Satzeinleitungen).

Das Problem bei der Einteilung in diese drei Gruppen besteht darin, dass für jeden Satz ein konsistenter Syntaxgraph gebildet werden muss; dabei gehen die Klammerung der Segmente und die Annotation der grammatischen Funktionen Hand in Hand. Deshalb sollte man in Betracht ziehen, die Aufgabengebiete der Analysegruppen bei der Annotation zu ändern und stattdessen von *Annotationsgruppen* zu sprechen. Das bedeutet, dass diese Änderung die Einteilung in Analysegruppen bei der Auswertung und

[105] Vgl. Schmid 2007: 53f.

Interpretation des annotierten Korpus nicht betreffen muss. Eine Möglichkeit wäre, der Wortgruppe die morphosyntaktische Ebene zuzuweisen. Die beiden anderen Gruppen erstellen für unterschiedliche Korpusteile vollständige Syntaxgraphen. Die Ergebnisse der beiden Gruppen werden ausgetauscht, diskutiert und gegebenenfalls überarbeitet.

Eine weitere Möglichkeit wäre, dass eine zweifache Annotation von Sätzen stattfindet: Die zwei Gruppen erstellen für jeden Satz des Korpus komplette Syntaxgraphen. Diese werden in einem weiteren Schritt verglichen und unterschiedliche Ergebnisse diskutiert, ehe man sich für eine Variante entscheidet, die in das *DiSynDe*-Korpus einfließt – es besteht allerdings auch die Möglichkeit, alternative Annotationen (beispielsweise mit XCES) zu kodieren. Diese Vorgehensweisen würden die Konsistenz und Qualität der syntaktischen Annotation erhöhen.

Da auf der textgrammatischen Annotationsebene eine von den anderen beiden Ebenen unabhängige Segmentierung der Korpustexte in minimale Texteinheiten erfolgt, kann die Annotation durch die Textgruppe parallel zu den anderen Gruppen erfolgen (vgl. *5.4. Zur textgrammatischen Annotation*, Seite 104ff).

(1) Morphosyntaktische Annotation

a) Wortarten-Klassifikation
   - *Präposition*
   - *Subjunktion*
   - *finites Modalverb*

b) Flexionsmorphologische Analyse
   - *2. Person Singular Indikativ Präsens*
   - *Akkusativ Plural*
   - *Instrumental Singular*

(2) Syntaktische Annotation

a) (rekursive) Konstituentenstruktur
   - *Satz*
   - *Teilsatz*
   - *Verbalphrase*
   - *Nominalphrase*
   - *Adverbialphrase*
   - *Präpositionalphrase*

b) Dependenzstruktur (grammatische Funktionen)

I. von Teilsätzen:
   - *Adverbialsatz (Kausal-, Temporal-, Konditional-, Konsekutiv-, Konzessiv-, Lokal-, Modal-, Adversativ-, Finalsatz)*
   - *Attributsatz (Relativsatz)*
   - *Inhaltssatz (Subjektsatz, Objektsatz)*

II. von Konstituenten:
   - *Satzeinleitung*
   - *Subjekt*
   - *Prädikatsnomen*
   - *Dativobjekt*
   - *Akkusativobjekt*
   - *Präpositionalobjekt*
   - *Adverbiale (modal, lokal, instrumental)*

III. von Konstituenten-Bestandteilen:
   - *Kopf*
   - *Determinator*

3.2. Die Kodierung syntaktischer Annotation

3.2.1. Annotationsformate

Man unterscheidet generell zwischen der Darstellung der Annotation am Bildschirm und ihrer Repräsentation in einem Dateiformat. Während die Darstellung „unabhängig vom Dateiformat [ist] und (…) sich primär an den Bedürfnissen des Betrachters [orientiert]“[106], muss die Repräsentation der Annotation „in einer für Maschinen lesbaren Art“[107] definiert werden. Das bedeutet, dass die Benutzer einer Annotation „(sei es der Ersteller oder der Leser) nicht notwendigerweise mit dem Dateiformat in Berührung kommen“[108].

Dabei wird das Annotations- oder Dateiformat aus der logischen Struktur der Informationen, die annotiert werden sollen, abgeleitet. Man stellt sich demnach als erstes die Frage, was für eine Struktur die linguistischen Informationen haben, die man den Primärdaten hinzufügen möchte:

> „Ist es die Struktur von Wörtern, Phrasen oder Sätzen? Eignet sich eine Darstellung als einfacher Baum, oder sind allgemeinere Graphen notwendig? Aus der logischen Struktur ergibt sich dann eine Kodierung der Daten in einer Datei, das *Dateiformat*. Die Bandbreite logischer Strukturen reicht dabei von einfachen Strukturen wie einer Sequenz von Wörtern in einem Text bis hin zu hoch komplexen Strukturen wie sich überkreuzenden Kanten in syntaktischen Strukturen oder sich überlappenden Äußerungen in einem Dialog.“[109]

Das Mittel der Wahl für baumartige Strukturen stellt die Auszeichnungssprache XML dar; es gibt jedoch auch andere Annotationsformate, die „sich aus historischen Gründen (…) etabliert“[110] haben. Im Folgenden werden die Vor- und Nachteile unterschiedlicher Formate vorgestellt, wodurch letztlich die Vorteile, die XML bietet, deutlicher hervortreten.

Unter *Vertikalformat* versteht man zeilenorientierte Annotationsformate. Die sprachlichen Informationseinheiten und die dazugehörige linguistische Annotation werden dabei auf die einzelnen Zeilen einer Datei aufgeteilt, wobei die „einzigen strukturierenden Mittel (…) Zeilenumbruch und Leerzeichen“[111] sind. Diese Vorgehensweise bietet sich vor allem für die alleinige Annotation morphosyntaktischer Information an: „Ein typisches Beispiel ist die Darstellung von POS-Tags für die Wörter eines Textes.

---

[106] Tylman und Hinrichs 2004: 225.
[107] Ebd.: 225.
[108] Ebd.: 225.
[109] Ebd.: 225.
[110] Ebd.: 225.
[111] Ebd.: 226.

Eine Zeile beginnt hier typischerweise mit einer Wortform und wird von einem einzigen POS-Tag abgeschlossen."[112]

Für syntaktische Annotation sind Vertikalformate nur eingeschränkt verwendbar, denn die Datei „ist als Liste strukturiert und eignet sich daher zunächst nur zur Darstellung sequenzieller Information und weniger für hierarchische Strukturen."[113] Ein weiterer Nachteil besteht darin, dass eine „Überprüfung der Gültigkeit einer Datei (…) speziell für diesen Zweck zu schaffenden Werkzeugen überlassen"[114] ist, während für XML zahlreiche solcher Werkzeuge frei zur Verfügung stehen. Im Gegensatz zu XML kann bei einem zeilenorientierten Forma Information, die im Ausgangstext implizit vorhanden ist, verloren gehen: „Wählt man z. B. eine Repräsentation des Ausgangsmaterials, bei der ein Wort pro Zeile annotiert wird, geht meist die im Ausgangstext enthaltene Information über Leerzeichen verloren, die um die Wörter herum vorhanden waren. Sucht man im annotierten Text nun Wörter, die ‚direkt' von Anführungszeichen umschlossen waren, ist dies nun nicht mehr möglich."[115] Dies spricht für ein „ausdrucksstärkeres Kodierungsfomat wie XML (…), das sehr leicht solche Informationen darstellen kann."[116]

Doch Vertikalformate haben auch Stärken, „die sie für viele Zwecke prädestinieren:"[117] Ein Vorteil liegt darin, dass diese Darstellungsform besonders einfach verarbeitet werden kann. „Dateien dieses Formats sind einfach per Programm zu lesen und zu generieren, insbesondere durch Unix-Betriebssysteme, die mächtige Werkzeuge für die Verarbeitung listenbasierter Dateien bereithalten (…)."[118] Auch für syntaktische Annotation lässt sich das Vertikalformat einsetzen, indem man eine tabellarische Struktur mit eingebetteten Verweisen erzeugt:

> „Die Struktur von Datenbanken lässt sich leicht als Liste darstellen, in der einer Spalte die besondere Funktion des Schlüssels zukommt. Das *export*-Format (Brants, 1997) z. B. nutzt ein erweitertes Vertikalformat zur Darstellung der syntaktischen Annotation, bei dem jeweils ein terminaler oder nichtterminaler Knoten eine Zeile belegt und auf den Mutterknoten verweist (…)."[119]

Dennoch liegt die eigentliche Stärke von zeilenbasierten Formaten „bei weniger komplexer Annotation, die die Position einzelner Wörter (oder zumindest Sequenzen

[112] Tylman und Hinrichs 2004: 226.
[113] Ebd.: 226.
[114] Ebd.: 226.
[115] Ebd.: 225f.
[116] Ebd.: 226.
[117] Ebd.: 226.
[118] Ebd.: 226f.
[119] Ebd.: 227.

sprachlicher Äußerungen) als Ankerpunkte verwendet."[120] Anhand des *Negra-Formats* soll allerdings demonstriert werden, dass mit Vertikalformaten auch syntaktische Annotation möglich ist.

Wortform/ Nummer	Wortart/ Kategorie	morphologische Information	Kanten-label	Eltern-knoten	sek. Kante	Ziel-knoten
Er	PPER	3.Sg.Masc.Nom	SB	500	SB	502
kauft	VVFIN	3.Sg.Pres.Ind	HD	500		
und	KON	-	CD	503		
verkauft	VVFIN	3.Sg.Pres.Ind	HD	502		
Äpfel	NN	Masc.Akk.Pl	CJ	501		
und	KON	-	CD	501		
Birnen	NN	Fem.Akk.Pl	CJ	501		
.	$.	-	-	0		
#500	S	-	CJ	503		
#501	CNP	-	OA	502	OA	500
#502	S	-	CJ	503		
#503	CS	-	-	0		

Abbildung 6. Beispielsatz im Negra-Format.[121]

Das Negra-Format funktioniert folgendermaßen:

> „Zunächst werden zeilenweise alle Wortknoten, anschließend die übergeordneten Knoten aufgelistet. In jeder Spalte ist ein Attribut kodiert. Die Reihenfolge der Attribute (Wortform, Wortart, morphologische Markierung für äußere Knoten und syntaktische Kategorie für innere Knoten) ist dabei fest vorgegeben. Die angegebene Kantenbeschriftung gilt für die Kante, die vom Elternknoten zum aktuellen Knoten führt. Die Nummer des Elternknoten ist in einer eigenen Spalte angegeben. Durch diese Verweisspalte wird die Graphstruktur repräsentiert. Sekundäre Kanten werden im Anschluss an den Elternknoten aufgelistet. Sie sind durch die Kantenbeschriftung und den Zielknoten eindeutig festgelegt."[122]

Vorteile des Negra-Formats sind, dass es nicht nur sehr kompakt ist, sondern dass es „in Grenzen sogar für den menschlichen Betrachter lesbar"[123] ist. Obwohl diese Eigenschaft sekundär erscheinen mag, sollte bedacht werden, „dass es immer noch Anwendungsfälle gibt, in denen der menschliche Benutzer eingreifen muss: Programmierung von Konvertierungsskripten, kurzfristige manuelle Korrekturen an der Annotation usw."[124]

[120] Tylman und Hinrichs 2004: 227.
[121] Lezius 2002: 16.
[122] Ebd.: 15.
[123] Ebd.: 16f.
[124] Ebd.: 17.

Eine Korpusdeklaration im Header einer jeden Datei legt die Eigenschaften des jeweiligen Korpus genau fest:

> „Sehr hilfreich ist die Deklaration des Korpus zu Beginn einer Korpusdatei. Alle verwendeten Attribute und Attributwerte werden hier aufgelistet und mit Kurzbeschreibungen erläutert. Eine solche Dokumentation kann von weiterverarbeitenden Anwendungen benutzt werden."[125]

Ein Nachteil des Negra-Formats ist der verwendete Zeichensatz, nämlich ISO-Latin1. Dadurch können Zeichen aus anderen Zeichensätzen nicht benutzt werden. „Zwar ließe sich hier eine Ersatzkodierung auf der Grundlage von Unicode einbringen (z.B. \u0937 für das griechische Ω), doch ist (…) diese (…) nicht standardisiert."[126] Weil das Negra-Format an sich nicht standardisiert ist, treten Probleme hinsichtlich der Weiterverarbeitung auf:

> „Die Trennung von Inhalt und Struktur (durch Trennzeichen wie Tabulatoren oder Leerzeichen) ist zwar eindeutig definiert und das Format damit in eindeutiger Weise parsebar. Doch es stehen keine Konvertierungsprogramme für andere Formate wie das Klammerstrukturformat zur Verfügung, so dass jeder Anwender eigene Konverter bzw. Parser programmieren muss."[127]

Außerdem ist das Negra-Format nicht erweiterbar: „Weitere Attribute wie beispielsweise Lemma auf Tokenebene oder Kasus können nicht ausgedrückt werden. (…) Das Negra-Format würde die einsetzbaren Korpora auf einen festgelegten Typ einschränken."[128]

Klammerstrukturformate sind für die syntaktische Annotation besser geeignet als rein zeilenbasierte Formate. Der Beispielsatz *Etta chased a bird* wird in Abbildung 7 als Baumstruktur und anschließend im Klammerstrukturformat dargestellt.

Eine Einschränkung des Formats besteht darin, dass jedem Knotenpunkt nur ein Attribut hinzugefügt werden kann. „Zwar ist das Format prinzipiell erweiterbar, indem zusätzliche Attributwerte durch ein Trennsymbol abgetrennt an das Annotationssymbol angehängt werden. Doch ist eine solche Lösung nachteilig, da die einzelnen Attributwerte bei der Weiterverarbeitung wieder isoliert werden müssen."[129] Bei XML ist eine leichtere Trennung von Inhalt und Struktur gegeben; entsprechende Werkzeuge zur Weiterverarbeitung müssen nicht an nicht-standardisierte Trennsymbole angepasst werden.

---

[125] Lezius 2002: 15.
[126] Ebd.: 17.
[127] Ebd.: 17.
[128] Ebd.: 17.
[129] Ebd.: 20.

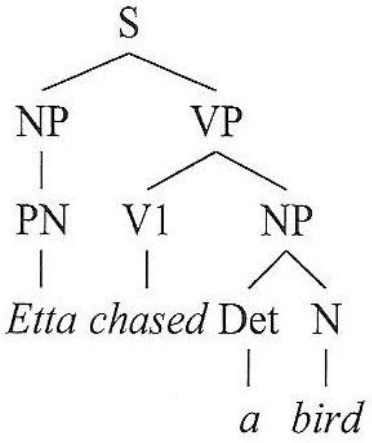

Abbildung 7. Baumstruktur für *Etta chased a bird*.[130]

„[S [NP [PN Etta]] [VP [V1 chased] [NP [Det a][N bird]]]]"[131]

Obwohl Wolfgang Lezius das Klammerstrukturformat als kompakt und platzsparend bewertet und feststellt, dass es „bei entsprechender Einrückung sogar für den menschlichen Betrachter gut lesbar"[132] ist, bemängelt er auch hier die Verwendung des Zeichensatzes ISO-Latin1 und die prinzipiell zwar mögliche, aber schwierige Erweiterbarkeit des Formats. Als Nachteil wird außerdem festgestellt, dass das Format auf Baumstrukturen beschränkt ist. „Graphstrukturen lassen sich nur durch aufwändige Konvertierungen mit zusätzlichen Informationen ausdrücken."[133]

Eine flexiblere Lösung stellen *Annotationsgraphen* dar. Diese Graphen bieten allerdings kein einheitliches Datenformat an, sondern sind vielmehr eine „Daten*struktur*, die mächtig genug ist, möglichst viele Korpustypen (…) repräsentieren zu können."[134] Annotationsgraphen werden hauptsächlich zur Annotation von gesprochener Sprache eingesetzt, „da es in Dialogen häufig vorkommt, dass mehrere Stränge sprachlicher Information gleichzeitig auftreten bzw. sich überlappen."[135]

Annotationsgraphen funktionieren nach folgendem Grundprinzip:

> „Zwischen in zeitlicher Reihenfolge gegliederten Ankerpunkten werden gerichtete Graphen gespannt. Auf diese Weise lassen sich alle erdenklichen Phänomene

[130] Lezius 2002: 18.
[131] Ebd.: 17.
[132] Ebd.: 20.
[133] Ebd.: 20f.
[134] Ebd.: 28.
[135] Tylman und Hinrichs 2004: 228.

annotieren: hierarchische Strukturen, Informationen über Zeitspannen und Zeitpunkte, sich überlappende Äußerungen oder Kombinationen von all dem."[136]

Annotationsgraphen sind demnach ein sehr mächtiges Annotationsformat, das vor allem zur Auszeichnung von komplexen Strukturen geeignet ist. Doch diese Mächtigkeit wirkt sich negativ auf den Aufwand aus, „der zur Erstellung, Verwaltung und Suche in Annotationsgraphen nötig ist."[137] Lezius zeigt in seiner Dissertation, wie man Annotationsgraphen auch für syntaktische Annotation verwenden kann. Wichtigster Bestandteil eines Annotationsgraphen ist die Zeitachse. Bei einer syntaktischen Annotation orientiert sich die Zeitachse an der linearen Reihenfolge der Wörter. Während die morphosyntaktischen Informationen an den Kantenübergängen festgeschrieben werden, stehen die übergeordneten Kanten für die Konstituenten – beispielsweise für die Nominalphrase (NP) *Ein Mann*. „Da in den Wurzelknoten keine Kante führt, ist die fehlende Beschriftung durch das Symbol ‚–' repräsentiert. (…) Die exakte Konstituenten-struktur ergibt sich dann implizit aus den Beziehungen der Kanten zueinander. Im Beispiel umfasst die S-Konstituente die Nominalphrase und das finite Verb, während Artikel und Nomen zur Nominalphrase gehören."[138]

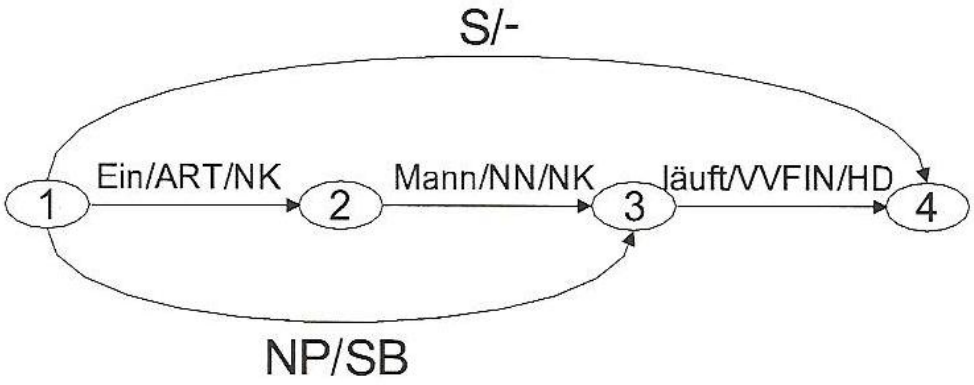

Abbildung 8. Annotationsgraph für einen einfachen Beispielssatz.[139]

[136] Tylman und Hinrichs 2004: 228.
[137] Ebd.: 228.
[138] Lezius 2002: 28f.
[139] Ebd.: 28.

Es gibt noch keine Standardisierung für eine entsprechende XML-Repräsentation der Annotationsgraphen. Lezius lehnt letztlich den Einsatz von Annotationsgraphen für syntaktische Annotation ab:

> „Ein Brückenschlag zwischen den Bereichen Baumbanken auf der einen Seite und Sprachkorpora auf der anderen Seite ist mit den *annotation graphs* sicher gelungen. Doch ist dieser Ansatz zu stark auf die Bedürfnisse der Sprachkorpora ausgerichtet, um auch im Bereich der Baumbanken als zentrales Kodierungsformat agieren zu können."[140]

Im Bereich der Auszeichnungssprache XML finden sich dagegen viele standardisierte Annotationsformate wie TEI, (X)CES, TUSNELDA oder TIGER. „Allen letztgenannten XML-basierten Annotationsstandards ist gemeinsam, dass sie sich der Mechanismen von XML bedienen, um die Kodierung von Primärdaten und ihrer Annotation zu vereinfachen."[141] Einerseits löst XML durch die Unterstützung von Unicode die Zeichensatzproblematik auf, andererseits stehen sowohl genügend Werkzeuge für die Weiterverarbeitung zur Verfügung als auch „Standards zur Beschreibung der Dokumentstruktur, die neben der Wohlgeformtheit auch die Überprüfung der Gültigkeit eines Dokuments erlauben (…)."[142] Die zunehmende Verbreitung dieser Werkzeuge und Standards findet ihre Ursache in „der weit reichenden Akzeptanz von XML als Textkodierungsformat."[143]

Doch man sollte sich bei der Wahl eines Annotationsformates auch an den Informationen orientieren, die man annotieren will. Wenn nämlich „die gewählte Dateistruktur den zu kodierenden Informationen [ähnelt], dann vereinfacht sich deren Kodierung."[144] Im folgenden Punkt wird gezeigt, welche Eigenschaften die Markupsprache XML für die Annotation linguistischer Informationen prädestinieren.

---

[140] Lezius 2002: 30.
[141] Tylman und Hinrichs 2004: 230.
[142] Ebd.: 230.
[143] Ebd.: 230.
[144] Ebd.: 231.

3.2.2. Die Strukturierung linguistischer Informationen mit XML

Andreas Witt setzt in seiner Dissertation den Begriff der Informationsstrukturierung implizit dem Begriff der Informationsmodellierung gleich:

> „Im Zentrum [des Titels *Multiple Informationsstrukturierung mit Auszeichnungssprachen*] steht der Begriff der Informationsmodellierung oder genauer der Begriff der ‚textuellen Informationsmodellierung', wie er von Lobin (…) beschrieben wird, nämlich als Bezeichnung für eine regelgeleitete Anordnung von textuellen Informationseinheiten."[145]

Henning Lobin definiert textuelle Informationsmodellierung als „die Modellierung von Information anhand von Gesetzmäßigkeiten, die wir in der Sprache finden."[146] Damit will er die Vorgehensweise, Markupsprachen wie SGML oder XML als „Instrumente für die Modellierung von strukturierter Information"[147] einzusetzen, vom Konzept der Datenbanken abgrenzen: „Datenbanken basieren auch in ihren neuesten Ausprägungen auf dem Konzept der Tabelle, durch die Daten in einen mehrdimensionalen Zusammenhang eingeordnet werden."[148]

Um die Möglichkeiten, die SGML und XML bieten, genauer zu fassen, führt Lobin drei Beobachtungen an Texten auf. Erstens können in einem Text verschiedene Ebenen unterschieden werden; als Beispiel nennt er auf der einen Seite die Ebene der sprachlichen Zeichen, auf der anderen Seite die Ebene der typographischen Auszeichnung, die etwa Überschriften vom eigentlichen Textkörper abgrenzt.[149] Als zweite Beobachtung stellt er fest, dass Informationseinheiten sowohl bei der Bildung von Sätzen als auch in Markupsprachen nach bestimmten Regeln angeordnet werden:

> „Die Regeln spezifizieren einerseits das hierarchische Verhältnis von abstrakten Informationseinheiten zu untergeordneten abstrakten oder konkreten Informationseinheiten, andererseits die lineare Abfolge gleichrangiger Informationseinheiten. Man kann diese Regeln zu einer Grammatik der Informationseinheiten zusammenfassen."[150]

Die dritte Beobachtung besteht darin, dass diese Grammatiken normalerweise so angelegt sind, dass die Informationseinheiten in einer Baumform angeordnet werden:

---

[145] Witt 2002: 7.
[146] Lobin 2001: 3.
[147] Ebd.: 3.
[148] Ebd.: 3.
[149] Vgl. ebd.: 3.
[150] Ebd.: 3f.

> „Ganz oben gibt es ein Wurzelelement, das den Text als Ganzes repräsentiert, die Töchter darunter repräsentieren die Teile, aus denen sich der Text auf oberer Ebene zusammensetzt, und diese Zerteilung wird solange fortgesetzt, bis man auf der Ebene der elementaren Texteinheiten angelangt ist."[151]

In SGML und XML sieht Lobin diese Beobachtungen hervorragend zur Anwendbarkeit gebracht. Für einen Anwendungsbereich lassen sich Typen von abstrakten und konkreten Informationseinheiten definieren, eindeutig benennen und mit weiteren Eigenschaften versehen. Durch Regeln, die in einer Grammatik zusammengefasst sind, können die unterschiedlichen Typen von Informationseinheiten zueinander in Beziehung gesetzt, mit realen Informationseinheiten verknüpft und in Baumform angeordnet werden.[152] Abschließend definiert Lobin strukturierte Information als „regelgeleitete Anordnung von Informationseinheiten, genauso wie wir korrekt strukturierte Sätze als regelgeleitete Anordnung von Wörtern verstehen können."[153]

XML wird dabei gegenüber SGML präferiert, denn XML stellt eine neuere Entwicklung dar, die sich im Vergleich zu SGML durch ihre Einfachheit auszeichnet und sich unter anderem dadurch im Bereich der Auszeichnungssprachen zum allgemein anerkannten Standard entwickelt hat:

> „Die angesprochene Standard Generalized Markup Language (SGML) (…) wurde (…) seit seiner 1986 erfolgreichen Standardisierung nicht nur weiterentwickelt, sondern es wurde mit der Extensible Markup Language (XML) im Jahr 1998 eine wesentlich einfachere Untermenge dieser Sprache definiert, die zudem das derzeitige Zentrum weiterer Entwicklungen auf dem Gebiet der Auszeichnungssprachen darstellt."[154]

Obwohl der Einsatz von XML zahlreiche Vorteile mit sich bringt, ergibt sich ein grundsätzliches Problem, das sich vor allem bei der Strukturierung von linguistischen Informationen auf mehreren Ebenen offenbart:

> „Eine der Stärken von XML (und SGML) besteht darin, dass Informationen hierarchisch strukturiert werden. Dies erlaubt u.a. eine effiziente maschinelle Verarbeitung, allerdings ist diese Stärke zugleich eine der Schwächen. Informationen, die sich nicht in bestimmten Hierarchien (mittels mathematischer Bäume) strukturieren lassen, können nicht in natürlicher Weise in diesem Formalismus repräsentiert werden."[155]

---

[151] Lobin 2001: 4.
[152] Vgl. ebd.: 4.
[153] Ebd.: 4.
[154] Witt 2002: 7.
[155] Ebd.: 9.

Das bedeutet, dass Informationseinheiten, die mit XML-Elementen ausgezeichnet werden, sich nicht partiell überlappen dürfen. Da auch das *DiSynDe*-Korpus auf verschiedenen linguistischen Ebenen annotiert wird, werden derartige Überlappungen sehr häufig auftreten:

> „Renear et al. (1996) diskutieren eine der Grundannahmen der Strukturierung von Textdaten: die ‚OHCO-These'. Diese besagt, dass ein Text eine ‚ordered hierarchy of content objects' bildet. Diese angenommene Hierarchie rechtfertigt eine grundlegende Einschränkung von XML, die sich selbst in der schwächsten Form der Dokumentannotation – den wohlgeformten XML-Dokumenten – wieder findet: Durch Elemente ausgezeichnete Umgebungen dürfen sich nicht partiell überlappen. Erfolgt jedoch eine direkte Zuordnung von linguistischen Phänomenen zu der Verwendung von XML-Elementen, zeigt sich, dass derartige Überlappungen ausgesprochen häufig auftreten."[156]

Zur besseren Illustration des Problems wird nach Andreas Witt ein kurzes Beispiel skizziert. Der Satz *A cat sat on the mat* soll nicht nur auf Ebene der Phrasenstruktur annotiert werden, sondern es soll auch die kontrastive Betonung des Satzes kenntlich gemacht werden. Die XML-Annotation für die Phrasen sieht folgendermaßen aus:[157]

```
<np>
    <det>the</det>
    <n>cat</n>
</np>
<vp>
    <v>sat</v>
    <pp>
      <p>on</p>
      <np>
        <det>the</det>
        <n>mat</n>
      </np>
    </pp>
</vp>
```

Die Annotation für die kontrastive Betonung hingegen hat diese Form:[158]

```
<contrastive>the cat sat</contrastive> on the mat
```

Wenn man nun den Beispielsatz auf beiden Ebenen gleichzeitig annotiert, überlappen sich die Informationseinheiten partiell und stören die eindeutige hierarchische Baumstruktur der Auszeichnungselemente. An folgender Darstellung kann man die partielle Über-

[156] Sasaki und Witt 2004: 209.
[157] Vgl. Witt 2002: 46.
[158] Vgl. ebd.: 46.

lappung einzelner Elemente erkennen; das Element <vp> beginnt, bevor das Element <contrastive> geschlossen wird:

```
<contrastive>
<np>
    <det>the</det>
    <n>cat</n>
</np>
<vp>
    <v>sat</v>
</contrastive>
    <pp>
      <p>on</p>
      <np>
        <det>the</det>
        <n>mat</n>
      </np>
    </pp>
</vp>
```

Das augenscheinlichste Beispiel für überlappende Hierarchien bietet eine Diskussion, in der eine Äußerung beginnt, während eine andere Äußerung noch nicht zu Ende geführt wurde.[159] Zudem können sich überlappende Strukturen sogar innerhalb eines Wortes befinden; wenn man sowohl die Silbenstruktur als auch die morphologische Struktur annotieren möchte, kann es vorkommen, dass sich diese Strukturen überschneiden:[160] „Betrachten wir das Verb *sagen*: Es besteht aus den Morphemen {sag-} und {-en} und den Silben *sa-* und *–gen*.“[161] Für die Kodierung überlappender Hierarchien in XML existieren verschiedene Lösungsmöglichkeiten, die im anschließenden Punkt vorgestellt werden.

3.2.3. Die Kodierung überlappender Hierarchien in XML

Die Grundannahme der bereits erwähnten OHCO-These besteht darin, dass ein Text stets aus einer geordneten Hierarchie von Inhaltsobjekten besteht (*ordered hierarchy of content objects*). Durch diese angenommene Hierarchie rechtfertigt man eine grundlegende Einschränkung von XML-Dokumenten, „die sich auch in der schwächsten Form der Dokumentannotation – den wohlgeformten XML-Dokumenten – wiederfindet: die durch Elemente ausgezeichnete Umgebungen dürfen sich nicht überlappen.“[162] Wenn man Texte jeglicher Art jedoch auf mehreren linguistischen Ebenen untersucht, finden sich überlappende Hierarchien sehr häufig:

---

[159] Vgl. Sasaki und Witt 2004: 209.
[160] Vgl. ebd.: 209.
[161] Kessel und Reimann 2005: 95.
[162] Witt 2002: 41.

> „Es kann vermutet werden, dass sich die Hierarchie-These nur entwickeln konnte, da ihr ein sehr enger Textbegriff zugrunde liegt. Texte bestehen aus Kapiteln, Abschnitten, Überschriften etc. Naheliegend war dies insbesondere deshalb, da die Wurzeln der Textannotation im Buchdruck lagen – gedruckte Textteile sind (meist) genau einer derartigen Basishierarchieebene zuzuordnen."[163]

Wenn man linguistische Informationen auf mehreren Ebenen zu textuellen Primärdaten hinzufügen möchte, ist es nötig, für diese Einschränkung von XML eine gangbare Lösung zu finden. „Die Bedeutung von Lösungen für dieses Problem rückt (…) immer stärker in den Fokus der Informationsmodellierungsforschung (…)."[164] Das europäische Verbundprojekt MATE (*Multilevel Annotation, Tools Engineering*) setzt sich beispielsweise mit diesem Problem auseinander. „Die dort verwendeten Lösungen zur Annotation von Informationen auf verschiedenen Ebenen erlauben die Markierung von nichthierarchischen Umgebungen, also Texten die der OHCO-These nicht genügen."[165] Im Folgenden werden drei verschiedene Ansätze vorgestellt, im Rahmen von XML konfligierende Hierarchien zu annotieren, nämlich die Verwendung von Meilensteinen, der Einsatz einer primären Annotationsebene und die sogenannte separate Annotation.

Eine simple und häufig eingesetzte Möglichkeit, in XML-Dokumenten weitere Hierarchieebenen zu kennzeichnen, besteht in der Verwendung von *Meilensteinen*; dazu fügt man leere Elemente ein, „die ausschließlich dazu dienen, eine oder mehrere zusätzliche Strukturierungsebenen einzufügen."[166] Leere Elemente bilden keine Behälter für Inhaltsobjekte, sondern markieren einfache Punkte in der Struktur des Dokuments und dienen dazu, indirekt Umgebungen auszuzeichnen; ein Beispiel ist das HTML-Element `<br>` für die Markierung von Zeilenumbrüchen:

> „In der HTML-DTD wird z. B. ein Element
 definiert, das die Position eines Zeilenumbruchs markiert. Zwischen zwei dieser Zeilenumbrüche befindet sich, falls dieses Element konsistent als Strukturierungsinstrument benutzt wird, der Inhalt der Zeilen."[167]

Ein Vorschlag der Text Encoding Initiative (TEI), weitere Hierarchien in einem XML-Dokument definieren zu können, besteht in der Definition des Elements <milestone>: „Diese Meilensteine werden ähnlich wie die beschriebene Verwendung des HTML-Elementes `<br>` als indirekte Umgebungsmarkierungen benutzt. Die TEI gibt einige

[163] Witt 2002: 41.
[164] Ebd.: 42.
[165] Ebd.: 42.
[166] Ebd.: 52.
[167] Ebd.: 53.

allgemeine und einige spezielle Meilensteinelemente an.“[168] Man kann diese leeren Elemente so einsetzen, dass man multiple Hierarchien in einem XML-Dokument annotieren kann – „und dies nicht separat, sondern an den Stellen, an denen sie auftreten.“[169] Durch die Spezifizierung von leeren Elementen mit Attributen, die Anfangs- und Endpunkte bezeichnen, kann man alternative Hierarchien in einem Dokument auch direkt auszeichnen:

> „Einen Schritt weiter als die einfache Markierung von Meilensteinen gehen Barnard et al. (1995). Sie führen in Analogie zu `<milestone>` die Elemente `<action>`, `<move>`, `<gesture>` etc. ein, um parallel zum Text ablaufende nichtsprachliche Ereignisse zu annotieren. Diese Elemente können in einer direkteren Form zur Markierung von Umgebungen gebraucht werden, da sie nicht nur eine mit einer Beschreibung versehene Marke bilden, sondern erlauben, den Beginn und das Ende einer Umgebung auszuzeichnen. Dies geschieht, indem ein Attribut eingeführt wird, dessen Wert entweder *start* oder *end* sein kann. Sie bilden somit das Analogon zu den speziell hierfür vorgesehenen Markierungen des Umgebungsbeginns (< . . . >) bzw. des Umgebungsendes (</ . . . >).“[170]

Meilensteine bringen allerdings auch Nachteile mit sich:

> „So kann nicht (…) überprüft werden, ob es für die jeweiligen Startmarken auch Endmarken gibt oder ob die Startmarke vor der Endmarke auftritt. Darüber hinaus ist es nicht möglich, Beziehungen zu den anderen Elementen des Textes zu spezifizieren, da es in der Natur leerer Elemente liegt, keine Inhaltsmodelle zu besitzen, die restringiert werden können.“[171]

Eine zweite Möglichkeit, ein Dokument mit XML in mehrere Hierarchien zu strukturieren, besteht darin, „eine primäre Annotationsebene zur Markierung der wesentlichen strukturellen Einheiten zu verwenden und zusätzliche Ebenen der Annotation mit dieser Ebene zu verknüpfen (…).“[172] Das europäische Verbundprojekt MATE wendet diese Vorgehensweise an. Als erstes muss man sich entscheiden, welche Ebene als primär eingestuft wird. Auf dieser Annotationsebene erhalten alle Elemente eindeutige Identifikatoren:[173]

---

[168] Witt 2002: 53.
[169] Ebd.: 54.
[170] Ebd.: 54.
[171] Ebd.: 54f.
[172] Sasaki und Witt 2004: 209.
[173] Vgl. Witt 2002: 47.

```
<np>
    <det id='x1'>the</det>
    <n id='x2'>cat</n>
</np>
<vp>
    <v id='x3'>sat</v>
    <pp>
      <p id='x4'>on</p>
      <np>
        <det id='x5'>the</det>
        <n id='x6'>mat</n>
      </np>
    </pp>
</vp>
```

Durch die Identifikatoren „wird ermöglicht, dass die Elemente potentiell als Verweisziele für zusätzliche Annotationsebenen dienen können.“[174] Die endgültige Annotation schließlich besteht aus der primären Annotationsebene mit den zugewiesenen Identifikatoren und den „zusätzlichen, mit ihnen durch Hypertextverknüpfungen verbundenen Annotationen:“[175]

```
<np>
    <det id='x1'>the</det>
    <n id='x2'>cat</n>
</np>
<vp>
    <v id='x3'>sat</v>
    <pp>
      <p id='x4'>on</p>
      <np>
        <det id='x5'>the</det>
        <n id='x6'>mat</n>
      </np>
    </pp>
</vp>
 <contrastive href='id(x1)..id(x3)'>
```

Das Attribut *href* des Elements <contrastive> erhält als Zielwert die Adressen des ersten und des letzten Identifikatoren für den Bereich, den dieses Element auszeichnen soll.

„Dieses Attribut kann auch zum Verweis auf andere Dokumente verwendet werden, so dass die unterschiedlichen Auszeichnungen auch physikalisch separiert sein können.“[176] Das bedeutet, dass man die verschiedenen Annotationsebenen auch in eigene Dateien auslagern kann; dies ermöglicht eine leichtere Verwaltung und Pflege der Annotationsebenen. Diese separate Speicherung nennt man *Stand-Off-Annotation* und steht im Gegensatz zur sogenannten *Inline-Annotation*.

---

[174] Witt 2002: 46.
[175] Ebd.: 46.
[176] Ebd.: 47.

Die Annotationsebenen müssen dabei so gewählt werden, dass innerhalb einer Ebene keine partiellen Überlappungen in der Dokumentstruktur auftreten können – diese Restriktion von XML bleibt bestehen. „Überlappungen [hingegen], die zwischen getrennten Ebenen auftreten, sind – aus der Sicht von XML betrachtet – unproblematisch."[177]

Ein Nachteil der Methode, weitere Annotationsebenen mit einer primären Ebene zu verlinken, besteht darin, „dass sich – mit Ausnahme der Basisannotationsebene – die einzelnen Annotationsebenen nicht aus sich heraus erklären."[178] Ein weiteres Problem dieses Verfahrens liegt bereits in der Entscheidung, welche Ebene die primäre Annotationsebene sein soll. Denn in der Basisannotation müssen sämtliche potentiellen Verweisziele mit eigenen Identifikatoren ausgestattet sein – dies ist nicht unbedingt der Fall, solange die Verweisziele der primären Annotationsebene weiter unterteilt werden können. Wenn sich die Identifikatoren auf die Wortebene beschränken, ist es zu einem späteren Zeitpunkt nicht mehr einfach möglich, die Morphem- und die Silbenstruktur der Wörter auf weiteren Ebenen hinzuzufügen.

Der Optimalfall besteht darin, jedem Zeichen einen eigenen Identifikator zuzuweisen. So können auch unterschiedliche Fassungen eines Textes aligniert werden – *Deutsch Diachron Digital* beispielsweise unterscheidet zwischen einer eng-diplomatischen und einer weit-diplomatischen Textfassung, „welche im Unterschied zu ersterer nicht mehr alle allographischen Feinheiten beibehält, sondern etwa Kürzelstriche oder Schlußbuchstabenformen normalisierend auflöst bzw. vereinheitlicht."[179] Eine alternative Möglichkeit zum Einsatz der Identifikatoren stellen Verweistechniken wie XPointer oder XPath dar. Diese Techniken erlauben es, Verweise auf Textabschnitte zu erstellen, die nicht annotiert sind – „allerdings wird hierdurch die Interpretation der Annotation wesentlich komplexer, sowohl für den Menschen als auch für die Maschine."[180]

Trotz der Schwierigkeiten überwiegen die Vorteile der Stand-Off-Annotation. Die Primärdaten und die linguistischen Annotationen bleiben voneinander unabhängig und können in eigenständige Dateien ausgelagert werden, während die Annotationen nur auf die Primärdaten verweisen: Die „ursprüngliche, zu annotierende sprachliche Äußerung [bleibt] eine separate Einheit, auf die die linguistische Annotation nur verweist."[181] Die XML-Version des *Corpus Encoding Standard* (XCES) bietet auch die Möglichkeit, „dem ursprünglichen Text linguistische Annotation hinzuzufügen, auch wenn die ursprüngliche

[177] Sasaki und Witt 2004: 210.
[178] Ebd.: 210.
[179] Lüdeling et al. 2004: 15.
[180] Sasaki und Witt 2004: 210.
[181] Tylman und Hinrichs 2004: 231.

sprachliche Äußerung bereits Annotation enthält, die sich zusammen mit der linguistischen Annotation nicht mehr direkt in einem XML-Baum darstellen ließe."[182] Darüberhinaus können durch die Stand-Off-Annotation auch verschiedene Annotationsformate – zum Beispiel TEI und XCES – kombiniert auf dieselben Primärdaten angewandt werden. Dies ermöglicht es einem Korpusprojekt zum Beispiel

> „einen möglichst allgemeinen, austauschorientierten Standard wie die TEI-Richtlinien erst dann einzusetzen, wenn Texte tatsächlich zwischen Annotatoren und Nutzern ausgetauscht werden. Dies ist eine gängige Praxis, um die problemorientierten und aufgabenspezifischen Formate während der Verarbeitung nutzen zu können, ohne auf die Vorteile eines allgemeineren Standards zu verzichten (…)."[183]

Eine dritte Möglichkeit, konfligierende Hierarchien zu annotieren, stellt die *separate Annotation* dar, die nach folgendem Grundprinzip funktioniert: „Jede Annotationsebene wird unabhängig von anderen Ebenen annotiert. Hierbei werden die zu annotierenden Daten entsprechend der Anzahl der Annotationsebenen dupliziert."[184] Bei der separaten Annotation kann man somit auf den Einsatz komplizierter Verweistechniken verzichten, denn die Verknüpfung der Annotationsebenen erfolgt implizit über die Primärdaten:

> „Die zugrundeliegende Idee besteht darin, dass es nicht notwendig ist, elaborierte Verknüpfungsmechnismen zu entwickeln und die entsprechenden Hyperlinks in die Annotationen einzubauen, sondern die annotierten Primärdaten als Basis der Verknüpfung zu verwenden. Da die annotierten Texte jeweils Kopien der Ausgangsdaten sind, wird für jede der separaten Annotationen derselbe Text verwendet. Die (unterschiedlich) annotierten Texte selbst ermöglichen bzw. bilden die Verknüpfung."[185]

Die separate Annotation erfordert allerdings eine konsequente Überwachung der Datenhaltung: „Die zu annotierenden Daten müssen in jeder einzelnen der separaten Auszeichnungsebenen identisch sein. Dies bedeutet insbesondere, dass Veränderungen der Primärdaten in allen Annotationsebenen vorgenommen werden müssen, damit die (impliziten) Verknüpfungen aufrecht erhalten werden können."[186] Das Problem der konsistenten Datenhaltung findet allerdings seine Entsprechung auch beim Einsatz einer primären Annotationsebene: Sobald Änderungen erfolgen, welche die ursprüngliche

---

[182] Tylman und Hinrichs 2004: 231.
[183] Ebd.: 231.
[184] Sasaki und Witt 2004: 212.
[185] Ebd.: 212.
[186] Ebd.: 213.

Reihenfolge der Identifikatoren zerstören, verlieren auch potentielle Verweise anderer Annotationsebenen ihre Gültigkeit.

> „Werden diese Punkte beachtet, könnte die multiple Annotation ein sehr mächtiges und adäquates Mittel zur Repräsentation der Informationen über multiple Annotationsebenen und somit eine geeignete Technik zur Auszeichnung linguistischer Korpora bilden, insbesondere im Hinblick auf den immer mehr an Bedeutung gewinnenden Aspekt der Wiederverwendung (linguistischer) Ressourcen."[187]

Für *Diachrone Syntax Deutsch* bietet sich letztlich die Entwicklung einer Mehrebenenarchitektur mit Stand-Off-Annotation an, bei der die verschiedenen Annotationsebenen durch eine primäre Annotationsebene verbunden sind.

[187] Sasaki und Witt 2004: 213.

4. Die Annotation historischer Texte

4.1. Synchronie und Diachronie

Die Begriffe Informationsstrukturierung, Korpus und syntaktische Annotation wurden bereits erläutert; als letzter wichtiger Bestandteil des Titels der Arbeit ist noch der Begriff *Diachronie* zu definieren. Was bezeichnet ein *diachrones Korpus*?

> „Die Unterscheidung zwischen Synchronie und Diachronie ist in der Sprachwissenschaft spätestens seit 1916, dem Erscheinungsjahr des Buches *Cours de linguistique générale* von Ferdinand de Saussures geläufig (…)."[188]

Diachronie ist das Gegenstück zur Synchronie. Während sich eine synchrone Analyse mit dem Zustand von Sprache zu einem bestimmten Zeitpunkt auseinandersetzt, befasst sich eine diachrone Analyse mit Fragen des Sprachwandels – ausgewählte Sprachphänomene werden betrachtet, wie sie sich über einen Zeitraum hinweg entwickeln.

Der Begriffe stammen aus dem Griechischen; *Diachronie* setzt sich aus den Bestandteilen δια (dia) für *durch, hindurch* und χρονος (chronos) für *Zeit* zusammen. Der Wortbestandteil *σύν* in Synchronie bedeutet *zusammen, mit, gemeinsam*. Die Begriffe kann man mit „Gleichzeitigkeit"[189] für Synchronie und mit „Aufeinanderfolge" [190] für Diachronie übersetzen.

Synchronie und Diachronie bieten zwei unterschiedliche methodische Zugriffsarten auf das Untersuchungsobjekt *Sprache*. Eine synchrone Betrachtung ist mit einem zeitlichen Querschnitt der Sprache gleichzusetzen – die Funktionalität des Sprachsystems wird sichtbar. Wie wird Groß- und Kleinschreibung behandelt? Wie funktioniert die Substantivflexion zu diesem Zeitpunkt?

Eine diachrone Betrachtung dagegen entspricht einem sprachlichen Längsschnitt. Die Entwicklung sprachlicher Phänomene über einen Zeitraum wird sichtbar. Wie verhält es sich mit der Groß- und Kleinschreibung über die Jahrhunderte hinweg betrachtet? Wie verändert sich die Flexion der Substantive?[191]

Man spricht also nicht unbedingt von einer diachronen Analyse, sobald das Alt- oder Mittelhochdeutsche untersucht wird. Jeder Sprachzustand kann als historisch angesehen werden – auch das gegenwärtige Neuhochdeutsche ist eine historische Sprachstufe, da es das Produkt einer geschichtlichen Entwicklung ist. Ältere Sprachstufen des Deutschen lassen sich auch synchron untersuchen, wenn man kein geschichtliches

[188] Wirrer 2002: 243.
[189] Wolff 2004: 20.
[190] Ebd.: 20.
[191] Vgl. Volmert 2001: 27.

Wissen einbringt, sondern nur den Zustand des Sprachsystems zu diesem Zeitpunkt betrachtet. Diachronische Analyse bedeutet zu fragen, wie es zu den sprachlichen Erscheinungen gekommen ist. Man beschreibt einzelne Phänomene, „deren Entwicklung/Veränderung beobachtet, beschrieben und erklärt wird."[192]

Beide Methoden – Synchronie und Diachronie – können auch als sich ergänzende Vorgehensweisen betrachtet werden, die nur gemeinsam ein vollständiges Bild ergeben.

> „Synchronie und Diachronie sind nur unter methodischem Aspekt als unterschiedliche Herangehensweisen relativ klar voneinander zu trennen. Diese Differenzierung gilt jedoch nicht für das Objekt der Untersuchung, für die Sprache. Jeder Sprachzustand ist selbst auch dynamisch, ist ein Zustand in der Bewegung, und jede Entwicklungsperiode ist insofern auch statisch, als sie strukturiert ist und Systemcharakter besitzt. Stabilität und Variabilität bilden als Wesensmerkmale der Sprache eine dialektische Einheit, die man auch als dynamische Stabilität bezeichnet. Daher können sprachliche Veränderungen geradezu als eine ständige Systematisierung charakterisiert werden. Das wiederum erfordert und ermöglicht in der Regel die Anwendung der synchronischen und diachronischen Methode. Die Sprachgeschichtsforschung muss beide Vorgehensweisen miteinander verbinden, auch wenn bei der Untersuchung und Darstellung längerer Zeiträume oder gar der gesamten Geschichte einer Sprache die diachronische Methode dominiert."[193]

Johannes Volmert stellt fest, dass diachronische Analysen nicht durchgeführt werden können, ohne dass vorher synchronische Beschreibungsschritte der Sprache stattfinden:

> „Strenggenommen kann diachronische Sprachwissenschaft nicht betrieben werden ohne synchronische; denn wenn man Funktion und Stellenwert eines Phänomens nicht in einem ganzen Sprachsystem erkennt, kann man auch keine Aussagen über seine historischen Veränderungen machen. Andererseits können viele Phänomene eines Systems (zu einem best. Zeitpunkt) nicht erklärt werden, wenn man ihre historische Herkunft und Entwicklung nicht kennt (z.B. bei Wort-Entlehnungen aus anderen Sprachen oder bei der Klassifizierung der deutschen Verben in starke und schwache Konjugationsklassen)."[194]

Demnach ist der erste Schritt einer diachronen Sprachanalyse ein synchronischer. Das Sprachsystem wird in seiner Funktion innerhalb eines zeitgleichen Rahmens beschrieben, bevor es mit einer weiteren synchronischen Beschreibung der Sprache zu einem anderen Zeitpunkt verglichen werden kann. Mit diesen synchronischen Beschreibungen des Sprachsystems lässt sich der Prozess der Annotation gleichsetzen. Dabei strebt *Diachrone Syntax Deutsch* an, Sprache als Kontinuum zu untersuchen:

---

[192] Volmert 2001: 27.
[193] Schmidt 2007: 16.
[194] Volmert 2001: 27f.

> „Auf jeden Fall soll das Corpus eine konsequent diachronische Anlage ermöglichen. Das heißt: es soll kein sprachstufenbezogenes Patchwork aus über- oder nebeneinander liegenden Synchronien entstehen. Zentrale Aspekte der historischen Syntax sollen kontinuierlich, also möglichst entlang einer Zeitachse verfolgt werden."[195]

Die sprachliche Entwicklung verläuft nicht einsträngig. Nicht alle Teilsysteme einer Sprache entwickeln sich bis zu einem gewissen Punkt weiter, ab dem man von einer neuen Sprachstufe sprechen kann. Sprache unterliegt einer Kontinuitätsbeziehung, während die Periodisierung von Sprache stets nur ein Hilfsmittel darstellt.

Dabei besteht über die Einteilung des Deutschen in unterschiedliche Epochen keineswegs Einigkeit. „Wie die Geschichte anderer Sprachen zerfällt auch die des Deutschen nicht natürlich und säuberlich in Perioden. Konkurrierende Einteilungsvorschläge spiegeln daher die verschiedenen Kriterien und Ziele der Sprachhistoriker wider."[196]

Zwar existiert ein allgemein anerkannter Konsens – aber auch dieser Konsens unterliegt einer Entwicklung. In der älteren Sprachgeschichtsforschung teilt man die deutsche Sprache nach Jacob Grimm in folgende drei Abschnitte ein, wobei die Jahreszahlen nur als ungefähre Grenzen zu verstehen sind:

> „Althochdeutsch: von den Anfängen bis 1100,
> Mittelhochdeutsch: von 1100 bis 1500,
> Neuhochdeutsch: von 1500 bis zur Gegenwart."[197]

Diese Ansicht ändert sich im 19. Jahrhundert aufgrund der Sprachgeschichte von Wilhelm Scherer. Das Frühneuhochdeutsche wird als eigene Sprachstufe anerkannt:

> „Althochdeutsch: von den Anfängen bis 1050,
> Mittelhochdeutsch: von 1050 bis 1350,
> Frühneuhochdeutsch: von 1350 bis 1650,
> Neuhochdeutsch: von 1650 bis zur Gegenwart."[198]

Die mittelhochdeutsche Sprachstufe nach Grimm verliert gewissermaßen 150 Jahre an Boden. Ferner fällt auf, dass sich auch das Ende der althochdeutschen Epoche um 50 Jahre verschiebt. Im Sinne der Theorieneutralität oder Wiederverwendbarkeit sollten *DiSynDe*-Texte demnach den Sprachstufen nicht durch bestimmte Tags zugeordnet werden,

---

[195] Schmid 2007: 52.
[196] Wells 1990: 25.
[197] Schmidt 2007: 18.
[198] Ebd. 18.

sondern Korpus-Benutzer sollten aufgrund von Zeitangaben zur Entstehung der Texte die Möglichkeit haben, eigene Subkorpora zu definieren.

Außerdem stellt sich die Frage, inwieweit die mehrsträngige Entwicklung von Sprachphänomenen über Periodisierungsgrenzen hinweg einen Einfluss auf die Gestaltung des Annotationsschemas hat. Es ist nicht praktikabel, für jede Sprachstufe ein eigenes Schema zu erstellen, sondern es wird ein Schema angestrebt, mit dessen Hilfe man linguistische Phänomene sprachstufen-übergreifend beschreiben kann. So können auch Übergangstexte, die sozusagen zwischen den Sprachstufen stehen, annotiert werden; zum Beispiel kann man sowohl althochdeutsche als auch mittelhochdeutsche Erscheinungen in einem Text auszeichnen, ohne dass zwei voneinander unabhängige Schemata angewandt werden müssen.

Sprache unterliegt also einem ständigen Wandel. Dabei verändert sich beispielsweise das Lexikon einer Sprache schneller als syntaktische Regeln. „Die Syntax ist das Grundgerüst der Sprache. Veränderungen gehen hier nur langsam vonstatten.“[199] Aber auch die Syntax ändert sich im Lauf der Jahrhunderte – es gestaltet sich allerdings schwieriger, entsprechende Entwicklungslinien nachzuzeichnen. Texttechnologische Korpora stellen dabei ein Werkzeug dar, den Wandel syntaktischer Gesetzmäßigkeiten effektiv und effizient zu untersuchen.

Die Bezeichnung *diachrones Korpus* bedeutet, dass der Korpus Texte verschiedener historischer Sprachstufen enthält und diachronische Untersuchungsmethoden bei der Interpretation von Suchergebnissen angewandt werden können – ein bestimmter Zeitraum kann analysiert werden. *Synchrone Korpora* untersuchen Sprache zu einen Zeitpunkt, können aber auch *historische Korpora* sein, wenn sie keine Texte der Gegenwartssprache enthalten:

> „Der Unterschied besteht darin, dass bei einem synchronen Vorgehen die Gemeinsamkeiten der Texte im Korpus erforscht werden, wohingegen bei diachroner Betrachtung der Sprachwandel, d.h. die Veränderungen, die sich in den Texten des Korpus zeigen, im Mittelpunkt des Interesses stehen.“[200]

[199] Eroms 2000: 13.
[200] Scherer 2006: 27.

4.2. Entwicklungslinien der deutschen Sprache

Dieser Punkt zeigt verschiedene Entwicklungslinien der deutschen Sprache vom Althochdeutschen zum Neuhochdeutschen auf; aus diesen Tendenzen werden einzelne Besonderheiten abgeleitet, welche die Annotation historischer Texte betreffen.

Wilhelm Schmidt stellt in seiner *Geschichte der deutschen Sprache* fest, dass „das Grundsystem der Satztypen, Satzarten und Satzglieder sowie die grundlegenden Möglichkeiten der Verknüpfung bereits im Germ. vorhanden waren. Auf diesen Voraussetzungen baut das Ahd. auf und entwickelt sie in vielfältiger Weise weiter."[201] Eine entscheidende Veränderung im Übergang vom Germanischen zum Althochdeutschen besteht darin, dass sich das System der Flexionsmorpheme wandelt:

> „Durch die Veränderung der Flexionsmorpheme (Reduzierung, Schwund) ist die grammatische Eindeutigkeit vieler Formen nicht mehr gewährleistet, so dass Ersatzformen erforderlich werden (Artikel beim Substantiv, Personalpronomen beim Verb, Umschreibungen von Kasus und Tempora)."[202]

Im Bereich der Substantivgruppe – „eine der wichtigsten Entwicklungen der Struktur des deutschen Satzes"[203] – entstehen die bestimmten und unbestimmten Artikel, welche allmählich die Funktionen verlorengegangener Flexionsendungen übernehmen.

Die Festlegung des Wortakzents auf das Stammmorphem in den germanischen Sprachen ist die Ursache dieser grammatischen Veränderungen: „Damit standen alle Flexionssuffixe in unbetonter Stellung, was zu deren Abschwächung geführt hat."[204] Durch die Abschwächung der Endsilben werden manche Kasusendungen identisch.

Es kommt also „im Bereich der Substantiva zu einem fast duchgängigen Formensynkretismus (…), so daß der Artikel die Funktionen (…) [übernimmt], die (…) dem Flexionssuffix zukommen, nämlich Genus, Kasus und Numerus anzuzeigen."[205]

> „Die Nebensilbenabschwächung verstärkt die Tendenz vom synthetischen zum analytischen Sprachbau."[206]

Die Tendenz zum analytischen Sprachbau wird ebenfalls durch den allmählichen Wegfall des fünften Kasus *Instrumental* deutlich. Verschiedene Konstruktionen, vor allem

---

[201] Schmidt 2007: 266.
[202] Ebd.: 266.
[203] Ebert 1978: 43.
[204] Wirrer 2002: 247.
[205] Ebd.: 247.
[206] Schmidt 2007: 104.

Präpositionen in Verbindung mit anderen Kasus, übernehmen Funktionen des Instrumentals. „Dagegen halten sich die Instrumentalformen bei den demonstrativen und interrogativ-relativen Pronomen länger (…).“[207] Dies bedeutet, dass der Instrumental zwar langsam verschwindet, aber dennoch im Althochdeutschen zu finden ist:

> „Das Deutsche ist die einzige der neueren germanischen Sprachen, die die vier im Altgermanischen lebendigen Kasus erhalten hat. Im Ahd. wie im Westgerm. überhaupt finden sich auch noch Reste eines alten Instrumentals, die aber bald im Dativ aufgehen.“[208]

In ähnlicher Weise verhält es sich mit einer Kategorie des Numerus. Neben Singular und Plural existiert auch der *Dual*, eine Zweizahlform: „Der Dual (…) ist (als Nominal- und Verbalform) im Schwinden begriffen.“[209]

Für das Annotationsschema bedeutet dies, dass auf morphosyntaktischer Ebene die Möglichkeit bestehen muss, entsprechende Formen als *Instrumental* bzw. als *Dual* zu markieren.

Eine weitere Entwicklungslinie der deutschen Sprache hat unter anderem die stärkere Ausprägung der Hypotaxe zur Folge:

> „Das komplexe Erfassen vielfältiger Sachverhalte erfordert eine genauere Kennzeichnung der kausalen, konditionalen, modalen, relativen, demonstrativen Beziehungen durch entsprechende Sprachmittel (stärkere Ausprägung der Hypotaxe, Weiterentwicklung der satzverknüpfenden Pronomina und Konjunktionen, stärkere Festigkeit in der Satzgliedstellung).“[210]

Bereits im Althochdeutschen stehen die sprachlichen Mittel zur Bildung von hypotaktischen Gefügen zwar grundsätzlich zur Verfügung, allerdings herrscht bei der Textproduktion noch die Verwendung der Parataxe vor. Es „ist (…) nicht zu leugnen, daß das Ahd. eine Vorliebe für die Parataxe zeigt.“[211] Auch bei Übersetzungen aus dem Lateinischen werden hypotaktische Konstruktionen parataktisch ins Althochdeutsche übertragen.[212] Die sprachlichen Mittel zur Verbindung von Sätzen verfeinern sich im Laufe der Zeit; diese Entwicklung setzt sich auch im Mittelhochdeutschen fort:

---

[207] Meineke und Schwerdt 2001: 314.
[208] Dal 1966: 4.
[209] Schmidt 2007: 47.
[210] Ebd.: 266.
[211] Ebert 1978: 21.
[212] Vgl. ebd.: 21.

> „Wie im Nhd. gibt es die Möglichkeit der parataktischen und hypotaktischen Verbindung von Sätzen. Allerdings sind die Mittel zur Kennzeichnung der Fügungsart – koordinierend oder subordinierend – wie auch zur semantischen Charakterisierung der Beziehung zwischen den Sätzen – temporal, kausal usw. – zum größten Teil noch vielfältiger verwendbar, im Anwendungsbereich noch nicht so stark festgelegt. Das System der Konjunktionen hat noch keine so klaren Konturen wie im Nhd., es ist in Herausbildung begriffen. In diesen wie in anderen Bereichen kann das Mhd. ‚als eine Epoche des Suchens und Versuchens gelten (…) Es geht darum, die Tauglichkeit unterschiedlicher Ausdrucksmittel zu erproben, weil das Bedürfnis nach semantischer Differenzierung immer größer wird.'"[213]

Eine andere Entwicklungslinie des Deutschen findet sich in der allmählichen Herausbildung des Satzrahmens bzw. der verbalen Klammer:

> „Durch die beginnende Differenzierung des Konjugationssystems durch umschreibende und zusammengesetzte Tempusformen zunächst vor allem in der Übersetzungsliteratur (…) wird der Satzbau dahingehend verändert, dass der verbale prädikative Rahmen stärker hervortritt (…)."[214]

Eine Eigenschaft des Neuhochdeutschen besteht darin, dass im Nebensatz das finite Verb stets am Ende zu finden ist, doch vor dem 17. Jahrhundert war die Stellung des Verbs im deutschen Nebensatz noch nicht eindeutig festgelegt. Mit der Herausbildung des Satzrahmens und der Fixierung der Verbstellung wird auch die Entwicklung der Hypotaxe gefördert: „Neben den Konjunktionen dient auch die Wortstellung immer mehr zur Bezeichnung der Opposition von Haupt- und Nebensatz."[215]

Desweiteren finden im Bereich der Prädikate und Ergänzungen große Veränderungen statt. Die Valenztheorie „beschäftigt sich mit der Eigenschaft von Wörtern (vor allem von Verben, aber auch von Substantiven und Adjektiven), andere Wörter an sich zu binden."[216] Als Begründer der Valenztheorie gilt Lucien Tesnière (1893-1954).[217] Die Valenz lässt sich gut „mit der Wertigkeit eines Atoms vergleichen, welches nur eine festgelegte Anzahl an Bindungspartnern haben kann."[218] Jedes Verb hat demnach eine bestimmte Wertigkeit, die quantitative Valenz: „(…) *lieben* (x liebt y) ist ein zweiwertiges Verb, *geben* (x gibt y z) ein dreiwertiges Verb."[219] Die Wertigkeit eines Verbs kann sich in diachroner Sicht verändern:

---

[213] Schmidt 2007: 341.
[214] Ebd.: 267.
[215] Ebert 1978: 32.
[216] Kessel und Reimann 2005: 14.
[217] Vgl. ebd.: 14.
[218] Ebd.: 14.
[219] Ebert 1978: 50.

> „Das Verb *verquicken* kam im Frnhd. als zweiwertiges Prädikat in der Bedeutung ‚ein Metall mit Quecksilber verbinden' auf. Als nicht-notwendige Angabe muß zu dieser Zeit ein pleonastisches *mit Quecksilber* hinzugetreten sein, das sich auf die Bedeutung auswirkte. Mit der Entwicklung dieser Angabe zur obligatorischen Ergänzungsbestimmung trat ein Wandel der Bedeutung von *verquicken* zu ‚sich innig verbinden' ein. Das Ergebnis war die Erhöhung der Wertigkeit zu *verquicken* als dreiwertigem Verb: *jemand1 verquickt etwas2 mit etwas3* (…)."[220]

Der Wandel der quantitativen Valenz äußert sich vor allem in Veränderungen im Bereich der Objekte; beispielsweise geht der Gebrauch des Genitivs zurück: „Als Objektkasus ging der Genitiv, der im Dt. bis ins Mhd. zunehmend gebraucht wurde, im Nhd. stark zurück und gilt in der dt. Gegenwartssprache als Relikterscheinung."[221] Als Gründe benennen historische Sprachwissenschaftler hierfür meistens „die Abschwächung der Flexionsendungen (…) und (…) [den] Verlust der Sonderbedeutung des Genitivobjekts gegenüber dem Akkusativ (…)."[222] Im *DiSynDe*-Korpus soll nicht mit Hilfe von Subkategorisierungsrahmen[223] annotiert werden, welche Phrasen freie Angaben, obligatorische oder fakultative Ergänzungen sind – dieser Bereich wird auf Seite der Korpusauswertung erforscht.

Das Projekt *Deutsch Diachron Digital* (DDD) strebt neben der Lemmatisierung der Wörter innerhalb einer Sprachstufe auch eine Lemmatisierung über die Sprachstufen hinweg an: „Zusätzlich ist ein Hyperlemma-System vorgesehen, das die Lemmata der verschiedenen Sprachstufen miteinander in Beziehung setzt."[224] Eine einfache Verbindung über die neuhochdeutsche Übersetzung herzustellen ist nicht möglich, da „etymologische und semantische Beziehungen zwischen den Lemmata sich widersprechen können"[225].

Abschließend zu diesem Punkt soll festgehalten werden, dass „Zahl, Art und Umfang der Satzglieder und Gliedteile (…) den heute vorhandenen Gegebenheiten"[226] entsprechen. Außerdem kann jede „Sprachstufe des Dt. (…) mit allen Syntaxtheorien beschrieben werden, die auch auf die dt. Gegenwartssprache oder andere Einzelsprachen anwendbar sind."[227] Das bedeutet, dass auch korpuslinguistische Standards, die ursprünglich nur zur Annotation der neuhochdeutschen Sprache vorgesehen waren, bei entsprechenden Anpassungen auch andere Sprachstufen des Deutschen beschreiben können.

---

[220] Ebert 1978: 50f.
[221] Ebd.: 51.
[222] Ebd.: 51.
[223] Vgl. Kermanidis et al. 2004.
[224] Lüdeling et al. 2004: 9.
[225] Ebd.: 9.
[226] Schmidt 2007: 272.
[227] Meineke und Schwerdt 2001: 307.

5. Standards

> „Standardization is difficult and, especially in the case of syntactic annotation, controversial to the extent that it will be impossible to formulate one agreed ‚consensus' standard."[228]

Im Folgenden werden ausgewählte korpuslinguistische Standards vorgestellt und daraufhin diskutiert, inwieweit sie für die Annotation des *DiSynDe*-Korpus eingesetzt werden können.

Diese Standards sind das *Stuttgart Tübingen Tagset* (STTS), die *Tübinger Sammlung nutzbarer empirischer linguistischer Datenstrukturen* (TUSNELDA), das TIGER-Annotationsschema, die *Text Encoding Initiative* (TEI), die XML-Version des *Corpus Encoding Standard* (XCES) und das *Syntactic Annotation Framework* (SynAF). Dabei wird als erstes auf den Bereich der Metadaten eingegangen, als zweites auf die morpho-syntaktische Annotation. Daran schließt sich ein Schwerpunkt zur syntaktischen Annotation an. Abschließend wird die textgrammatische Annotationsebene erörtert.

5.1. Metadaten

Metadaten sind definiert als *Daten über Daten* bzw. als „Daten, die verschiedene Aspekte einer Informationsressource beschreiben."[229] Die Angabe von Metadaten ist umso wichtiger, je schwerer zugänglich die Primärdaten sind.[230] Für die Erstellung eines diachronen Korpus wie für *Diachrone Syntax Deutsch* sind Metadaten somit unerlässlich.

Metadaten erfüllen verschiedene Funktionen. In erster Linie fungieren sie als Dokumentation zu Entstehung und Entwicklung der beschriebenen Informations-ressource. „Zu den dokumentierten Aspekten etwa eines Textes gehören die Entstehungs-zeit, die Druck- bzw. Publikationszeit, Publikationsort, beteiligte Personen usw."[231] Überdies stellen Metadaten den Schlüssel zu den Primärdaten dar – erst durch die präzise Kodierung von Metadaten wird es möglich, individuelle Subkorpora zu definieren:

> „Der dokumentierte Entstehungszeitpunkt von Texten (oder Tonaufnahmen) erlaubt es Teilkorpora zusammenzustellen, die die Sprache einer bestimmten Epoche bzw. Sprachstufe dokumentieren (*die deutsche Sprache der Goethezeit, die deutsche Sprache der Wendezeit*, etc.) oder die Sprache einer bestimmten Region (*das Oberschwäbische, die Sprache in der DDR*). Der Fokus kann auf bestimmte Textsorten

[228] Kahrel et al. 1997: 242.
[229] Lemnitzer und Zinsmeister 2006: 45.
[230] Vgl. ebd.: 46.
[231] Ebd.: 46.

> oder Genres gelegt werden (*die Sprache von Gebrauchsanweisungen, Formen der Höflichkeit in Erpresserbriefen*).“[232]

Bei der Erstellung von Metadaten für digitale Korpora ist zu beachten, „dass möglicherweise zwei Informationsobjekte beschrieben werden müssen“[233]. Das erste Informationsobjekt ist der Text in seiner digitalen Form. Aus dem zweiten Informationsobjekt ist die digitale Repräsentation entstanden, „z.B. durch Abtippen, Einscannen oder Einlesen eines Druckereidatenträgers.“[234] Für *DiSynDe* ist es von besonderer Bedeutung, möglichst genau auf die Quelle eines digitalisierten Textes hinweisen zu können, denn bei älteren Texten des Deutschen existieren meist verschiedene Abschriften oder unterschiedliche Editionen, sodass es hierdurch zu Missverständnissen bei der Auswertung von linguistischen Daten kommen könnte.

> „Beide Informationsobjekte führen ein getrenntes Dasein, und streng genommen beziehen sich die Metadaten, die wir hier meinen, nur auf das erste Informationsobjekt. Der digitalisierte Text kann zum Beispiel die Abschrift einer Geschichte aus einem Kinderbuch sein, das seitdem in einer neuen Auflage in neuer Rechtschreibung herausgegeben wurde. Es ist deshalb wichtig, in den Metadaten zu einem digitalisierten Text möglichst genau auf die Quelle dieses Textes, das Original, hinzuweisen.“[235]

Im Corpus Encoding Standard (CES) wird ein Satz an Metadaten definiert, der sehr gut geeignet ist, linguistische Korpora zu beschreiben.

> „Der CES wurde federführend von der *Expert Advisory Group on Language Engineering Standards* (EAGLES) entwickelt. Wie der Name dieses von der EU geförderten Gremiums vermuten last, ist dieser Metadatenstandard für Korpora in sprachtechnologischen Projekten entwickelt worden. Dennoch ist der von CES definierte Metadatensatz auch für die Beschreibung linguistischer Korpora geeignet. Dies hängt unter anderem damit zusammen, dass dieser Standard sich an die Konventionen anlehnt, die die *Text Encoding Initiative* (TEI) für ein breiteres Spektrum an Texten und Korpora aufgestellt hat. Die Kategorien des *Corpus Encoding Standard* sind im Großen und Ganzen eine Teilmenge der von der TEI definierten Kategorien, mit einigen wenigen für die Sprachtechnologie relevanten Erweiterungen.“[236]

---

[232] Lemnitzer und Zinsmeister 2006: 46.
[233] Ebd.: 47.
[234] Ebd.: 47
[235] Ebd.: 47.
[236] Ebd.: 49.

Im Folgenden wird der Aufbau der CES-Metadaten erläutert.[237] Damit ein Metadatensatz standardkonform ist, müssen nicht alle Felder ausgefüllt werden – nur wenige bestimmte Elemente sind obligatorisch.[238]

Das Wurzelelement der CES-Metatdaten ist der sogenannte <cesHeader>, der durch bestimmte Attribute spezifiziert wird. Mit *type* gibt man an, ob sich die Metadaten auf einen einzelnen Text oder auf ein ganzes Korpus beziehen; *creator* steht für denjenigen, der das Informationsobjekt beschreibt. Ferner kann die Version der CES-Metadaten, der Entstehungszeitpunkt des Headers und das Datum des letzten Updates angegeben werden. Der <cesHeader> enthält wiederum vier Elemente, die folgendermaßen kodiert werden:

```
<cesHeader>
     <fileDesc></fileDesc>
     <encodingDesc></encodingDesc>
     <profileDesc></profileDesc>
     <revisionDesc></revisionDesc>
</cesHeader>
```

Der erste Teil <fileDesc> ist der einzige obligatorische Abschnitt der CES-Metadaten. Er enthält bibliographische Angaben zu einem Korpus oder einem Korpustext; an dieser Stelle werden Informationen zum Titel (<titleStmt>), zur Version (<editionStmt>), zur Speichergröße (<extent>), zur öffentlichen Zugänglichkeit (<publicationStmt>) und zur Quelle des digitalen Textes (<sourceDesc>) kodiert.

Der zweite Teil <encodingDesc> enthält Angaben zur Kodierung der Datei und beschreibt dabei hauptsächlich das Verhältnis des digitalen Informationsobjekts zu seinem Original. Hier kann das Korpusprojekt detailliert beschrieben werden (<projectDesc>), die Vorgehensweise bei der Zusammenstellung der Texte (<samplingDecl>) und für Parallelkorpora das Referenzsystem zur Alignierung der Paralleltexte (<refsDecl>). Außerdem kann man die allgemein angewandten Prinzipien für die Annotation der Informationsobjekte erläutern (<editorialDecl>), Informationen über das tatsächliche Auftreten von Tags einfügen (<tagsDecl>) und eine Typologie der Textsorten definieren, die im Korpus auftreten (<classDecl>). Im *DiSynDe*-Pilotkorpus werden beispielsweise die Textsorten *Chronistik, Fachliteratur, Geistliche Prosa, Privatschriften, Rechtsprosa, Unterhaltungsprosa* und *Übersetzungsliteratur* unterschieden.

---

[237] Vgl. CES 1996a.
[238] Vgl. Lemnitzer und Zinsmeister 2006: 49.

Mit dem dritten Teil <profileDesc> kann ein Korpus detaillierter beschrieben werden. Teilweise werden dadurch manche Informationen mehrfach kodiert; das Element <creation> beispielsweise enthält erneut Angaben zum Ursprung des digitalen Textes: „This element is used to record details concerning the origination of the text, whether or not covered elsewhere."[239] Darüberhinaus können an dieser Stelle der CES-Metadaten Informationen über die im Text auftretenden Sprachen, Sprachstufen oder Dialekte gespeichert werden (<langUsage>), zudem Informationen über den verwendeten Zeichensatz (<wsdUsage>), ausführliche Angaben zur Textsorte (<textClass>), Hinweise auf existierende Übersetzungen (<translations>) und Verweise auf weitere Dateien, die linguistische Annotationen zum vorliegenden Text enthalten (<annotations>).

Im vierten Teil <revisionDesc> „kann schließlich die Revisionsgeschichte der Informationsressource verzeichnet werden, sofern Revisionen an dieser vorgenommen wurden."[240] Für jede Änderung an einzelnen Texten oder am Korpus wird ein zusätzliches <change>-Element eingefügt, welches wiederum das Datum der Änderung (<changeDate>), den Namen der für die Änderung verantwortlichen Person (<respName>) und die Art der Änderung (<h.item>) angibt.

Die Vorteile des CES-Metadatensatzes bestehen also darin, dass er einerseits eine „sehr reichhaltige Beschreibung von Korpora und von einzelnen Texten"[241] ermöglicht, andererseits müssen nicht alle zur Verfügung stehenden Felder ausgefüllt werden, damit die Metadaten auch standardkonform sind. Ein Nachteil liegt in der Beschreibung diachroner Korpora. Bei historischen Texten ist es oft nicht möglich, den exakten Entstehungszeitpunkt zu bestimmen. Damit man bei unterschiedlichen Datierungen nicht zu einem Kompromiss hinsichtlich der Angabe des Entstehungszeitpunktes gezwungen ist, schlägt TUSNELDA als Lösung vor, einen Zeitraum zur Entstehung eines Textes anzugeben.

> „Der für die *Tübinger Sammlung nutzbarer empirischer linguistischer Datenstrukturen* (TUSNELDA) entwickelte Annotationsstandard ist eine Erweiterung des XCES mit dem Ziel, Richtlinien für die Annotation einer breiten Zahl linguistischer Phänomene zu schaffen, die anhand von Korpora untersucht werden sollen. (…) TUSNELDA baut dabei auf dem CES und den TEI-Richtlinien auf und definiert Erweiterungen dort, wo die Richtlinien den gewünschten Grad an Detail nicht hergeben."[242]

---

[239] CES 1996a.
[240] Lemnitzer und Zinsmeister 2006: 50.
[241] Ebd.: 50.
[242] Tylman und Hinrichs 2004: 229.

TUSNELDA definiert für das Element <creation> im dritten Teil der CES-Metadaten (<profileDesc>) drei neue Attribute. Das CES-Attribut *date* wird ersetzt durch *earliest*, *latest* und *place*. Dadurch kann nicht nur der früheste und späteste mögliche Entstehungszeitpunkt eines historischen Textes kodiert werden, sondern auch sein Entstehungsort (zum Beispiel ein bestimmtes Kloster).[243] Eine entsprechende Modifikation des Elements <creation> sieht folgendermaßen aus:

```
<!--          Creation element                                    -->

<!ELEMENT creation      - -  (#PCDATA )                             >
<!ATTLIST creation           %a.header;
          earliest           CDATA                #REQUIRED
          latest             CDATA                #REQUIRED
          place              CDATA                #REQUIRED      >
```

Die vollständige Datei *header.elt* (ohne Modifikation) findet sich im Internet,[244] das entsprechende XML-Schema ebenfalls.[245] Auf den folgenden zwei Seiten ist ein Beispiel für die Anwendung des CES-Metadatensatzes eingefügt, welches der CES-Homepage entnommen ist und wegen der besseren Übersichtlichkeit leicht formatiert wurde.[246]

[243] Vgl. Kallmeyer und Wagner 2000.
[244] CES 1996b.
[245] CES 2003.
[246] CES 1996a.

```
<cesHeader version="2.0">


      <fileDesc>


            <titleStmt>
                <h.title>Machine-readable version of 1984, ch. 1</h.title>
                <respStmt>
                     <respType>typed in and marked with CES tags
                     </respType>
                     <respName>A. Student</respName>
                </respStmt>
            </titleStmt>

            <extent>
                <wordcount>6571 </wordcount>
                <bytecount units="bytes">6571 </bytecount>
            </extent>

            <publicationStmt>
                <distributor>Laboratoire Parole et Langage, CNRS
                </distributor>
                <pubAddress>29, avenue Robert Schuman
                          Aix-en-Provence, France</pubAddress>
                <telephone>+33 42 95 36 33</telephone>
                <fax>+33 42 59 50 96</fax>
                <eAddress>phonetic@univ-aix.fr</eAddress>
                <availability status=restricted>
                    internal use only--cannot be distributed
                </availability>
                <pubDate>6571</pubDate>
            </publicationStmt>

            <sourceDesc>
                <biblStruct>
                     <monogr>
                          <h.title>Nineteen Eighty-four</h.title>
                          <h.author>George Orwell</h.author>
                          <imprint>
                               <pubPlace>New York</pubPlace>
                               <publisher>New American Library</publisher>
                               <pubDate>1949; reprinted 1961</pubDate>
                          </imprint>
                     </monogr>
                </biblStruct>
            </sourceDesc>


      </fileDesc>
```

```
<encodingdesc>

    <projectdesc>
      This English version of the first chapter of Orwell's 1984
      is encoded for use in the MULTEXT-EAST project. The English
      is to serve as the base for the parallel corpus, and will be
      aligned to versions of the text in Romanian, Bulgarian,
      Estonian, Slovenian, Czech, and Hungarian.
    </projectdesc>

    <editorialdecl>
        <conformance level=1>CES Level 1</conformance>
        <correction status=medium method=silent></correction>
        <quotation marks=none form=std>Rendition attribute values
            on Q and QUOTE tags are adapted from ISOpub and ISOnum
            standard entity set names
        </quotation>
        <segmentation>Marked up to the level of paragraph plus
              marking of particular sub-paragraph elements: NAME,
              DATE, FOREIGN.
        </segmentation>
    </editorialdecl>

    <tagsdecl>
        <tagusage gi=body occurs=1></tagusage>
        <tagusage gi=date occurs=5></tagusage>
        <tagusage gi=div occurs=2></tagusage>
        <tagusage gi=foreign occurs=4></tagusage>
        <tagusage gi=hi occurs=4></tagusage>
        <tagusage gi=name occurs=149></tagusage>
        <tagusage gi=note occurs=1></tagusage>
        <tagusage gi=num occurs=2></tagusage>
        <tagusage gi=p occurs=41></tagusage>
        <tagusage gi=ptr occurs=1></tagusage>
        <tagusage gi=q occurs=22></tagusage>
        <tagusage gi=quote occurs=3></tagusage>
    </tagsdecl>


</encodingdesc>



<profiledesc>

    <langusage>
        <language id="fr" iso639="fr">French</language>
        <language id="en" iso639="en">English</language>
        <language id="la" iso639="la">Latin</language>
        <language id="ns">Newspeak</language>
    </langusage>

</profiledesc>



</cesHeader>
```

5.2. Morphosyntaktische Annotation mit STTS

Das Stuttgart-Tübingen Tagset (STTS) ist ein Tagset für die Annotation des Deutschen auf Ebene der Wortarten. Es wurde entwickelt am Institut für maschinelle Sprachverarbeitung der Universität Stuttgart und am Seminar für Sprachwissenschaft der Universität Tübingen. Ein weiteres morphosyntaktisches Tagset für das Deutsche stellt das Münsteraner Tagset[247] dar, das am Arbeitsbereich Linguistik der Universität Münster erstellt wurde. An dieser Stelle wird jedoch näher auf das STTS eingegangen, da dieses aufgrund seiner großen Verbreitung als Standard angesehen werden kann.

Das STTS-Tagset ist hierarchisch aufgebaut: „Die *tags* bestehen aus möglichst selbsterklärenden Buchstabensequenzen, die von links nach rechts gelesen zuerst die Hauptwortart und dann die Unterwortart kodieren, also von der allgemeinen Information zur spezifischeren hinführen."[248] Die Hauptwortarten sind „weitgehend nach allgemein anerkannter linguistischer Terminologie in den *tags* kodiert"[249] und „orientieren sich am ‚TEI Starter Set of Grammatical-Annotation Tags'"[250].

Die 11 Hauptworten sind *Nomina* (N), *Verben* (V), *Artikel* (ART), *Adjektive* (ADJ), *Pronomina* (P), *Kardinalzahlen* (CARD), *Adverbien* (ADV), *Konjunktionen* (KO), *Adpositionen* (AP), *Interjektionen* (ITJ) und *Partikeln* (PTK). Für jede Hauptwortart sind eigene Subklassifizierungen definiert. „So werden z.B. die Pronomina in weitere 8 Untergruppen unterschieden, wobei die Untergruppen wieder unterteilt sein können, je nachdem ob sie NP-ersetzende (substituierend, *tag*: S), nomenbegleitende (attribuierend, *tag*: AT) oder adverbiale (*tag*: AV) Funktion innehaben."[251] Abbildung 9 stellt eine strukturierte Übersicht der Pronomina dar.

STTS besteht aus insgesamt 54 Tags. Davon dienen 48 der Klassifizierung von Wortarten; die restlichen 6 sind zur Auszeichnung von fremdsprachlichem Material (FM), Kompositions-Erstgliedern (TRUNC), Satzzeichen ($,, $., $() und von Nichtworten und Sonderzeichen (XY). Für jede Wortform wird genau ein Tag zugewiesen. Der Begriff *Wortform* „umfaßt neben ‚echten' Wortformen auch Zahlen in Ziffern, Satzzeichen, Sonderzeichen (wie z.B. §, $), abgetrennte Wortteile oder Kompositions-Erstgliedern (wie z.B. **Ein-** und Ausgang) etc."[252] Eine komplette Übersicht der STTS-Tags findet sich in Anhang 2.[253]

---

[247] Vgl. Steiner 2004: 80f.
[248] Schiller et al. 1999: 4.
[249] Ebd.: 4.
[250] Ebd.: 4.
[251] Ebd.: 5.
[252] Ebd.: 5.
[253] STTS 2001.

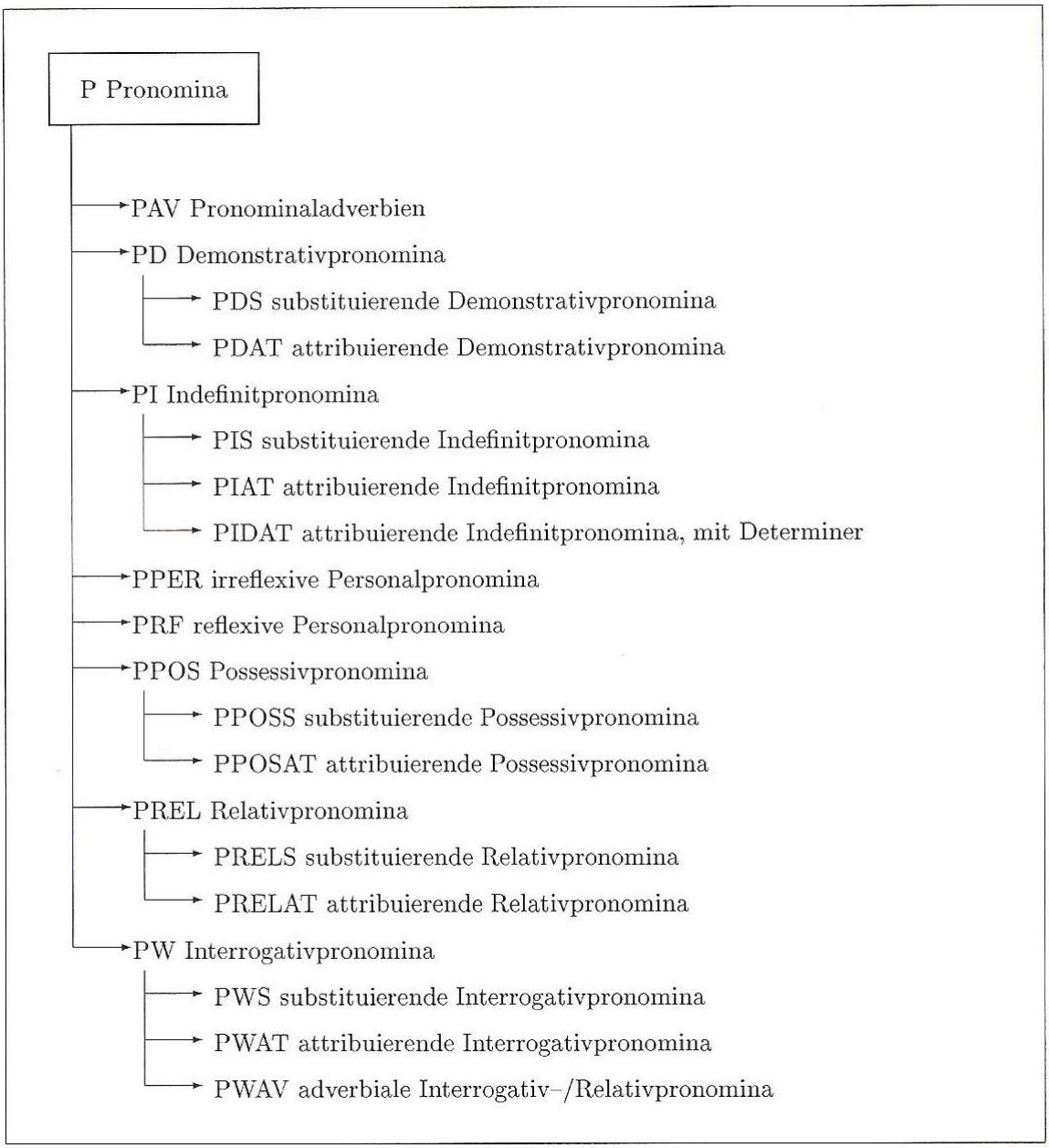

Abbildung 9. Pronomina in STTS.[254]

„Die Klasse der Pronomen *P* wird am stärksten unterteilt, was sich auch in den zusammengesetzten Tagnamen widerspiegelt. Je nach Funktion werden sie zu *D* (Demonstrativ), *I* (Indefinit), *PER* (PERsonal), *POS* (POSsessiv), *REL* (RELativ), *RF* (ReFlexiv), *W* (interrogativ oder relativ) oder *AV* (AdVerbial). Zusätzlich werden die meisten Pronomen noch nach ihrer Distribution spezifiziert: *S* (Substituierend) bzw. *AT* (Attribuierend). Ganz systematisch entstehen so die Tagnamen, z.B. *PPOSS* steht für ein *P*ronomen, *POS*sessiv, *S*ubstituierend und *PPOSAT* für ein *P*ronomen, *POS*sessiv, *At*tribuierend.“[255]

[254] Schiller et al. 1999: 36.
[255] Lemnitzer und Zinsmeister 2006: 68.

Petra Steiner zieht folgendes Fazit: „Das STTS erfüllt die Anforderungen der Vollständigkeit und der Erweiterbarkeit ohne weiteres.“[256] Durch die hierarchische Strukturierung kann das STTS auf einfache Weise eigenen Bedürfnissen angepasst werden, indem man zum Beispiel neue Tags definiert oder bestehende Tags weiter subklassifiziert. Außerdem wird „eine gewisse Flexibilität erreicht, die dem Benutzer erlaubt, je nach Anspruch nur auf die Hauptwortarten oder auf wortartenspezifische Informationen zuzugreifen.“[257]

In einer einfachen XML-Kodierung lässt sich der Beispielsatz *Diese Fischer haben viel geangelt* folgendermaßen annotieren (das Attribut *pos* steht dabei für *Part-of-Speech*):[258]

```
<word pos=”PDAT”>Diese</word>
<word pos=”NN”>Fischer</word>
<word pos=”VAFIN”>haben</word>
<word pos=”PIS”>viel</word>
<word pos=”VVPP”>geangelt</word>
<word pos=”$.”>.</word>
```

Das STTS wird darüberhinaus in ein kleines und ein großes Tagset unterschieden. Das eben beschriebene *kleine Tagset* lässt nur die Klassifizierung von Wortarten zu, während das *große Tagset* zudem die Annotation der Flexionsmorphologie ermöglicht. „Die beiden Versionen des STTS sind damit ein Beispiel für Tag-Sets, die eine gröbere Aufteilung durch eine strenge Hierarchie mit einer feineren Aufteilung verbinden (…).“[259]

Mit dem großen Tagset des STTS lassen sich folgende Kategorien der Flexionsmorphologie beschreiben: *Kasus*, *Genus*, *Numerus*, *Person*, *Tempus*, *Modus*, *Grad*, *Definitheit* und *Flexion*. Die Kategorie *Grad* bezeichnet die Komparation von Adjektiven und Adverbien, *Definitheit* unterscheidet zwischen bestimmten und unbestimmten Artikeln, und *Flexion* markiert Nomen und Adjektive nach starker (St), schwacher (Sw) oder gemischter (Mix) Flexionsklasse.[260] Ein Nomen wird zum Beispiel folgendermaßen annotiert:

„*der Beamte*/NN:Masc.Nom.Sg.Sw“[261]

---

[256] Steiner 2004: 79.
[257] Schiller et al. 1999: 4.
[258] Vgl. Dipper 2007.
[259] Tylman und Hinrichs 2004: 221.
[260] Vgl. Lemnitzer und Zinsmeister 2006: 71.
[261] Ebd.: 71.

Das Nomen *Beamte* ist demnach Maskulinum (Masc), steht im Nominativ (Nom) Singular (Sg) und wird schwach flektiert (Sw). Die Beschreibung der flexionsmorphologischen Kategorien findet sich den jeweiligen Wortarten zugeordnet in *Schiller et al. 1999*.

Eine Möglichkeit mit Wortarten-Ambiguitäten umzugehen, bieten die *Portmanteau Tags*: „Im *British National Corpus*, dem britischen Referenzkorpus, sind *Portmanteau Tags* erlaubt, die aus einer Kombination von zwei Tags bestehen, zum Beispiel *heard&VVD-VVN*; zeigt, dass das Token *heard* entweder in der einfachen Vergangenheit (VVD) steht oder als Partizip Perfekt (VVN) verwendet wird."[262] Für das *DiSynDe*-Korpus könnte man sich eine ähnliche Vorgehensweise überlegen. Außerdem wäre es möglich, Informationen zur Wortbildung zu annotieren; zum Beispiel kann markiert werden, dass das Nomen *Armer* eine deadjektivische Ableitung von *arm* ist:

„*ich Armer*/NN<ADJ:Masc.Nom.Sg.St"[263]

Wenn man das STTS-Tagset an den Einsatz für ein diachrones Korpus anpassen will, stellt sich die Frage, ob man auch besondere Verbklassen wie die Präterito-Präsentien markieren möchte. Diese sind folgendermaßen definiert:

> „Die Präterito-Präsentien sind Verben, deren Präsens die Form eines ablautenden Präteritums zeigt. Sie sind aus dem idg. Perfekt entstanden (…) Die alten Präsensformen gingen verloren; die Präteritalformen übernahmen die Präsensfunktion, flektierten aber weiter wie Präteritalformen ahd. starker Verben. (…) Als Ersatz für die ins Präsens übergegangenen Präteritalformen entstanden neue Präterita (…). Sie stehen (…) als Mischform zwischen den starken und den schwachen Verben."[264]

In historischen Grammatiken werden Wortarten meist nach ihrer Formenbildung unterschieden. Wilhelm Schmidt beispielsweise klassifiziert die Verben des Althochdeutschen in die Gruppen starke Verben, schwache Verben, Präterito-Präsentien, athematische Verben und das Verb *wellen* (entspricht neuhochdeutsch *wollen*).[265] Jede Untergruppe weist typische Konjugations-Merkmale auf.

Tagsets wie das STTS hingegen teilen die Wortarten nach ihrer syntaktischen Verwendung ein, sind darin allerdings nicht durchgehend konsequent. Petra Steiner kritisiert am STTS diesbezüglich: „Manche der semantischen Klassifikationskriterien

---

[262] Lemnitzer und Zinsmeister 2006: 70.
[263] Ebd.: 71.
[264] Schmidt 2007: 252.
[265] Vgl. ebd.: 237-254.

wären (…) vermeidbar."[266] Bei den Hauptwortarten etwa „fällt die semantisch motivierte Klasse der Kardinalzahlen (…) auf."[267] Da mit Hilfe des *DiSynDe*-Korpus der Wandel der deutschen Syntax untersucht werden soll, ist es zweckmäßiger, sich bei der Einteilung der Wortarten am STTS zu orientieren und die Klassifikation nach der Formenbildung zu vernachlässigen, wie es auch in den *DiSynDe*-Annotationsvorschriften angelegt ist.

Die flexionsmorphologische Kategorie *Kasus* muss um den fünften Fall Instrumental (*Instr*) erweitert werden. Bei der Kategorie *Numerus* ist zu überlegen, ob man den Dual mit aufnimmt. Denn falls duale Formen zu selten auftreten, können sie auch mit Plural markiert werden. Lateinische Textstellen können als fremdsprachliches Material (FM) annotiert werden.

Der zweite Satz aus der Tabelle der *DiSynDe*-Annotationsvorschriften (*So wil ich dir weisen, mit we du den teuffel überwindest.*)[268] wird im Folgenden nach dem leicht modifizierten großen STTS-Tagset (Instrumental) in einer einfachen XML-Kodierung annotiert. Dabei steht ADV für Adverb, VMFIN für modales finites Verb, PPER für irreflexives Personalpronomen, VVINF für infinites Vollverb, APPR für Präposition, PWS für substituierendes Interrogativpronomen, ART für Artikel, NN für normales Nomen und VVFIN für finites Vollverb. Bei den flexionsmorphologischen Angaben (Attribut *mor*) steht der Stern für nicht eindeutig bestimmbare Eigenschafen. Der Unterstrich beim Nomen zeigt, dass die Flexionsklasse nicht als stark, schwach oder gemischt eingeordnet werden kann.

```
<s>
   <word pos="ADV">So</word>
   <word pos="VMFIN" mor="2.Sg.Ind.Pres">wil</word>
   <word pos="PPER" mor="1.Sg.*.Nom">ich</word>
   <word pos="PPER" mor="2.Sg.*.Dat">dir</word>
   <word pos="VVINF">weisen</word>
   <word pos="$,">,</word>
   <word pos="APPR" mor="Instr">mit</word>
   <word pos="PWS" mor="*.Instr.Sg">we</word>
   <word pos="PPER" mor="2.Sg.*.Nom">du</word>
   <word pos="ART" mor="Def.Masc.Akk.Sg">den</word>
   <word pos="NN" mor="Masc.Akk.Sg._">teuffel</word>
   <word pos="VVFIN" mor="2.Sg.*.Pres">überwindest</word>
   <word pos="$.">.</word>
</s>
```

[266] Steiner 2004: 79.
[267] Ebd.: 78.
[268] Vgl. Anhang 1.

5.3. Syntaktische Annotation

5.3.1. Text Encoding Initiative (TEI)

Die Text Encoding Initiative ist eine Organisation namens TEI-Konsortium, die zum Ziel hat, einen Standard zur Kodierung und zum Austausch von Dokumenten und Texten jeglicher Art zu schaffen.[269]

> „Der wohl wichtigste Beitrag der TEI ist die Definition von umfangreichen Richtlinien für die Kodierung jeglicher Art elektronischer Texte: ‚They are addressed to anyone who works with any text in electronic form.' Eingeschlossen sind Texte aus Gegenwart und Vergangenheit, jeder Sprache und Schrift und neben Fließtexten auch nicht-kontinuierliche Texte wie Wörterbücher oder linguistisch annotierte Korpora (…). Die aktuelle Version der Richtlinien nutzt XML als Auszeichnungssprache und die zu XML gehörigen Mechanismen, um unterschiedliche typographische und inhaltliche Merkmale zu kodieren und Texte zu strukturieren."[270]

Der TEI-Standard ist demnach für eine Vielzahl unterschiedlicher Textsorten geeignet. Die TEI bietet auch Empfehlungen für die Annotation linguistischer Korpora an, allerdings ist sie nicht zentral dafür ausgerichtet. Um ihrem Anspruch der Allgemeingültigkeit gerecht zu werden, ist die *Document Type Definition* (DTD) der TEI modular aufgebaut. Für die unterschiedlichen Textsorten werden jeweils verschiedene Tagsets angeboten. Dabei wird unterschieden zwischen obligatorischen Modulen (*core tag-sets*), Basismodulen (*base tag-sets*) und zusätzlichen Modulen (*additional/auxiliary tag-sets*).[271]

Mit Hilfe der obligatorischen Module wird die allgemeine Dokumentenstruktur festgelegt, „die der von HTML sehr ähnlich ist, allerdings erlaubt der TEI-Ansatz eine weitaus stärker an den Korpustyp angepasste Annotation, was durch die Auswahl eines sog. *base tag-sets* ermöglicht wird."[272] Wenn man beispielsweise ein Korpus von Gedichten erstellt, bietet sich das Basismodul *Verse* an, für ein Korpus mit den Transkriptionen gesprochener Sprache das Modul *Transcription of Speech*.[273]

Zur Ergänzung der allgemeinen Dokumentstruktur wählt man für Korpora zusätzlich das Modul *Language Corpora* aus, das linguistische Annotationen ermöglicht: „Da einer der generellen Ansprüche der TEI darin bestand, die Annotationen in möglichst geringem Ausmaß theorieabhängig zu spezifizieren, entstand ein sehr allgemeines DTD-

[269] Vgl. TEI 2008a.
[270] Tylman und Hinrichs 2004: 228.
[271] Vgl. Sasaki und Witt 2004: 208.
[272] Ebd.: 208.
[273] Vgl. TEI 2008b.

Modul, in dem sechs Elemente definiert sind, die es erlauben, linguistische Einheiten zu kategorisieren."[274]

Diese sechs Elemente dienen der Auszeichnung von Sätzen (<s>), Teilsätzen (<cl>, engl.: *clause*), Phrasen (<phr>), Wörtern (<w>), Morphemen (<m>) und einzelnen Zeichen (<c>, engl.: *characters*).[275]

Allen sechs Elementen sind die Attribute *type* und *function* zugewiesen. Für Wörter lässt sich mit dem Attribut *lemma* zusätzlich die jeweilige infinite Form annotieren. Wörter wiederum setzen sich aus Morphemen zusammen, den „kleinsten bedeutungstragenden Einheiten der Sprache"[276]. Weil Morpheme auch als sogenannte Allomorphe in einer veränderten Form auftreten können, kann die Grundform eines Allomorphs in dem Attribut *baseForm* des Elements <m> angegeben werden. *Wäld* in *Wäldchen* ist beispielsweise ein Allomorph zum Morphem *Wald*.[277] Für *Diachrone Syntax Deutsch* ist allerdings nach den Annotationsvorschriften die Auszeichnung der Morphemstruktur von Wörtern nicht von Bedeutung.

Das Element <s> darf nicht ineinander verschachtelt sein. „Das hat z. B. zur Folge, dass mit <s> ausgezeichnete Einheiten nicht als Koordination zweier weiterer <s>-Einheiten repräsentiert sein können. Jedoch steht mit <cl> ein Element zur Verfügung, welches u.a. für diese Zwecke verwendet werden kann."[278] Ebenso kann das Element <m> keine weiteren <m>-Elemente enthalten. „Dies hat zur Folge, dass Zirkumfixe und Infixe nicht in einer intuitiven Form annotiert werden können (…)."[279]

Nach Andreas Witt ist der Beispielsatz *Wenn der Toast fertig ist, springt der Schiebeschalter hoch, und das Gerät wird automatisch ausgeschaltet* in „einer TEI-konformen partiellen linguistischen Annotation"[280] markiert. Witt verwendet alle sechs Elemente und verzichtet zur besseren Übersichtlichkeit auf eine vollständige Annotation. Die verwendeten Attributwerte sind nicht standardisiert. Das Attribut *type* der <w>-Elemente könnte mit Werten des STTS-Standards (*Part-of-Speech*) gefüllt werden. Da Tags des STTS teilweise auch die Funktion der Wortform enthalten (das Tag KON markiert eine nebenordnende Konjunktion), fiele *function="coord"* weg. Für flexionsmorphologische Angaben steht kein eigenes TEI-Attribut zur Verfügung.

---

[274] Witt 2002: 138.
[275] Vgl. TEI 2008c.
[276] Kessel und Reimann 2005: 92.
[277] Vgl. ebd.: 92.
[278] Witt 2002: 138.
[279] Ebd.: 138f.
[280] Ebd.: 139.

```
<s>
   <cl>
      <cl>Wenn der Toast fertig ist</cl>
      , springt der Schiebeschalter hoch
   </cl>
   <c>,</c>
   <w type="conj" function="coord">und</w>
   <cl>
      <phr type="np" function="subject">
         <w type="det">das</w>
         <w type="n">Gerät</w>
      </phr>
      wird automatisch
      <m>aus</m>
      <m>ge</m>
      <m>schalt</m>
      <m>et</m>
   </cl>.
</s>
```

Witt zieht als Fazit, dass die Mittel, die der TEI-Standard zur Verfügung stellt, linguistische Annotationen nur mit einem begrenzten Detailliertheitsgrad ermöglichen.[281]

> „Die TEI hat eher zum Ziel, die elektronische Kodierung von Texten zu unterstützen, und weniger, ein Kodierungsschema für sehr detaillierte linguistische Annotation zu definieren. (…) Ihr Anliegen ist vielmehr, alle jene Mechanismen zur elektronischen Kodierung von Texten zu beschreiben, die in wahrscheinlich jedem Text erwünscht sind (…).“[282]

Die TEI kann somit als *allgemeiner Standard* angesehen werden. Für eine detaillierte Annotation des *DiSynDe*-Korpus reichen die Vorgaben des fakultativen TEI-Moduls *Language Corpora* nicht aus.

Die TEI gestattet jedoch auch die freie Definition neuer Elemente und neuer Attribute, und vorhandene Elemente können verändert und angepasst werden. Durch entsprechende Eingriffe können sich so aus dem allgemeinen Standard spezialisierte Standards entwickeln. „Der unter anderem zur Annotation linguistischer Korpora geeignete Corpus Encoding Standard (CES) ist ein gutes Beispiel hierfür (…).“[283] Im folgenden Abschnitt werden die Möglichkeiten des CES aufgezeigt.

---

[281] Witt 2002: 139.
[282] Tylman und Hinrichs 2004: 229.
[283] Sasaki und Witt 2004: 208f.

5.3.2. Corpus Encoding Standard (CES)

Der CES wird ihm Rahmen von EAGLES (*Expert Advisory Group on Language Engineering Standards*) entwickelt. Ziel ist es, „einen einheitlichen Kodierungsstandard für linguistisch annotierte Korpora zu entwickeln."[284] SGML ist die ursprüngliche Grundlage des *Corpus Encoding Standard*; inzwischen existiert als XCES eine Implementierung in XML.

Der CES bietet nicht nur einen für Korpora angepassten TEI-Metadatensatz an, sondern verbessert auch – von der Grundlage des TEI-Moduls *Language Corpora* ausgehend – die Möglichkeiten zur linguistischen Annotation. Der CES stellt also eine Spezialisierung und Erweiterung der TEI dar. „Die Einbettungsmöglichkeiten in die durch die TEI vorgegebene Dokumentstruktur werden (…) genutzt, so dass eine komplette Integration in die modulare TEI-DTD möglich wird."[285]

Der CES-Standard bietet hinsichtlich linguistischer Annotation Richtlinien für die „Kodierung paralleler Texte, also von Texten gleichen Inhalts in verschiedene Sprachen"[286] an. Für *Diachrone Syntax Deutsch* bedeutender sind die Möglichkeiten des XCES sowohl zur morphosyntaktischen als auch zur syntaktischen Annotation.

Um einen Satz auszuzeichnen, steht wie in der TEI das Element `<s>` zur Verfügung, allerdings „darf `<ces:s>` auch verschachtelt sein, d. h. ein `<ces:s>`-Element darf selbst `<ces:s>`-Elemente enthalten. `<ces:s>` bekommt als eine Defaultattributierung `broken='no'` zugeordnet, was ausdrückt, dass ein vollständiger Satz in dieser Umgebung repräsentiert ist."[287] Mit dem Element `<chunk>` können mehrere Sätze zu einer Gruppe zusammengefasst werden. Die `<chunk>`-Elemente wiederum werden in das Element `<chunkList>` eingebettet.

Mit CES können die morphosyntaktischen Eigenschaften der Wörter detailliert beschrieben werden. Einem Token (`<tok>`) kann seine orthographische Form (`<orth>`) zugeordnet werden, die Wortart wird mit `<ctag>` angegeben, eine Lemmatisierung erfolgt mit Hilfe des Elements `<base>`, und die flexionsmorphologischen Angaben bezeichnet das Element `<msd>`.

Im Folgenden wird der Beispielsatz *MEMORY STOP arbeitet nicht bei REC RETURN* morphosyntaktisch annotiert:[288]

---

[284] Lezius 2002: 21.
[285] Witt 2002: 140.
[286] Tylman und Hinrichs 2004: 229.
[287] Witt 2002: 140.
[288] Vgl. ebd.: 140.

```
<chunkList>
   <chunk>
      <s>
         <tok>
            <orth>MEMORY STOP</orth>
         </tok>
         <tok>
            <orth>arbeitet</orth>
            <disamb>
               <ctag>verb</ctag>
               <msd>3PsSg</msd>
            </disamb>
            <lex>
               <base>arbeiten</base>
               <ctag>verb</ctag>
            </lex>
         </tok>
         <tok>
            <orth>nicht</orth>
            <ctag>adverb</ctag>
         </tok>
         <tok>
            <orth>bei</orth>
         </tok>
         <tok>
            <orth>REC RETURN</orth>
         </tok>
      </s>
   </chunk>
</chunkList>
```

Das Element <lex> kann beispielsweise bei einer automatischen Annotation der Wortarten dazu dienen, verschiedene Interpretationsmöglichkeiten einer Wortform aufzunehmen – das Element <disamb> kommt schließlich zum Einsatz, wenn ein Annotator die Interpretationsmöglichkeiten disambiguiert, indem er sie zu ihrem Kontext in Bezug setzt.[289]

Für die Kodierung der morphosyntaktischen Informationen in den Elementen <msd> und <ctag> merkt Andreas Witt an: „Diese Informationen sollen in einer in EAGLES definierten Form erfolgen. Hierauf wurde in dem hier gegebenen Beispiel verzichtet. Stattdessen wurde schlicht ‚Verb' bzw. ‚3PsSg' angegeben."[290] Für *DiSynDe*

[289] Vgl. Witt 2002: 141.
[290] Ebd.: 141.

kann man ebenfalls von den EAGLES-Empfehlungen abweichen und die Tags des STTS einsetzen. Witt bewertet den CES als einen Standard, der

> „besonders auf die Annotation der lexikalischen Ebene fokussiert ist. Es gibt kein Äquivalent zu den <tei:phr>-, <tei:cl>- oder <tei:m>-Elementen – ein Manko für eine linguistische Annotation. Der Zweck dieses Standards bestand jedoch darin, eine Möglichkeit der Annotation auf der Ebene der Wörter herzustellen. Während die zu dieser Ebene gehörenden sprachlichen Einheiten innerhalb der TEI-DTD schlicht mit <tei:w> annotiert werden, stellt der CES ein unvergleichbar feingliedrigeres Annotationsinventar zur Verfügung."[291]

Für die morphosyntaktische Annotation ist der CES-Standard demnach hervorragend geeignet. Ein Äquivalent zu <tei:m> (Annotation der Morphemstruktur) wird im Rahmen von *DiSynDe* nicht benötigt. Ein Äquivalent zu <tei:cl> (Annotation von Teilsätzen) ist implizit durch die Möglichkeit gegeben, <ces:s>-Elemente ineinander zu verschachteln. Das Fehlen des <tei:phr>-Elements (Annotation der Konstituentenstruktur) wird schließlich durch die Implementierung des CES in XML ausgeglichen, da der XCES auch um Möglichkeiten zur syntaktischen Annotation erweitert wird. Nancy Ide beschreibt das Ziel eines Meta-Models für syntaktische Annotation:

> „Because of the common practice for syntactic annotation utilizing trees, together with the natural tree-structure of markup in XML documents, we provide a meta-model for syntactic markup that follows this approach. The model is instantiated using the following tags:"[292]

Bei der Annotation von Baumstrukturen repräsentiert das Element <struct> jeweils einen Knoten. <struct>-Elemente können rekursiv benutzt werden. Das Attribut *id* erlaubt eine eindeutige Identifizierung der Knoten.[293]

Der XCES-Standard stellt keine Attributwerte zur Verfügung, mit deren Hilfe den Knoten spezifische Datenkategorien wie Satz oder Nominalphrase zugewiesen werden können, „da sonst eine Dokumentspezifikation durch DTD oder Schema zu stark korpusabhängig wäre."[294] Für die Spezifizierung dieser Werte ist das Element <feat> vorhanden, welches von <struct> umschlossen wird:

---

[291] Witt 2002: 141.
[292] Ide und Romary 2003: 289.
[293] Vgl. ebd. 289.
[294] Lezius 2002: 115.

> „A type attribute on the <feat> element identifies the data category of the feature. The tag may contain a string that provides an appropriate value for the data category (e.g., for type=CAT the value might be 'NP') or <feat> can be recursively refined to describe complex structures."[295]

Für alternative Annotationen steht das Element <alt> zur Verfügung.

Das Element <rel> definiert die Beziehungen zwischen den Knotenpunkten bzw. die Kanten des Baumes. Dies unterstützen folgende Attribute: *type* legt die Art der Beziehung oder die Kantenbeschriftung fest (zum Beispiel *Subjekt*); *head* und *dependent* verweisen auf den Kopf oder das Dependens einer Relation; *introducer* spezifiziert das einleitende Wort oder die einleitende Phrase eines Knotens; mit *initial* kann man thematische oder semantische Rollen einer Komponente angeben; *target* schließlich dient als Verweis auf zugehörige Knoten – dieses Element kann für die Annotation von kreuzenden Kanten in Baumstrukturen benutzt werden.[296]

XCES empfiehlt den Gebrauch der Stand-Off-Annotation; dafür steht das Element <seg> zur Verfügung: „A *target* attribute on the <seg> element uses XML Pointers (Xpointer) (…) to specify the location of the relevant data."[297] Dazu wird in einer primären Datei der Text gespeichert, in einer weiteren Datei werden die morphosyntaktischen Angaben annotiert, und in einer dritten Datei die syntaktischen Beziehungen. „Durch Referenzen auf die primäre Korpustextdatei werden dann die Bezüge zwischen den Annotationsebenen hergestellt (…)."[298] Ein für *DiSynDe* wichtiger Vorteil liegt dabei auch darin, dass man die Dateien für die verschiedenen Annotationsebenen zeitgleich bearbeiten kann. Außerdem kann das Korpus jederzeit um weitere Annotationsebenen erweitert werden. „Eine zusätzliche Annotationsebene muss dabei nicht notwendigerweise Bezug zu den Basisdaten nehmen, sondern kann auch an eine bereits bestehende Annotationsebene angehängt werden (…)."[299]

Anhand des einfachen Beispielsatzes *Ein Mann läuft* soll die Stand-Off-Annotation mit XCES vorgestellt werden. Der Korpustext wird in einer Basisdatei gespeichert. Der Text wird dabei nur in Sätze und Absätze gegliedert:

„ <s id="s1">ein Mann läuft</s> "[300]

---

[295] Ide und Romary 2003: 289.
[296] Vgl. ebd.: 289f.
[297] Ebd.: 290.
[298] Lezius 2002: 23.
[299] Ebd.: 23.
[300] Ebd.: 23.

Der Beispielsatz weist folgende Baumstruktur auf:

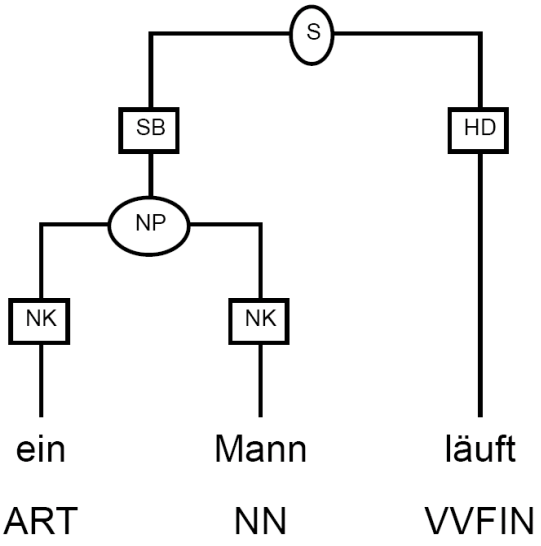

Abbildung 10. Baumstruktur des Satzes *ein Mann läuft*.[301]

Die Annotation der Wortarten (ART, NN, VVFIN) erfolgt in einer zweiten Datei. Erst auf dieser Ebene findet durch die XPointer eine Strukturierung des Textes in Tokens statt:

```
<p xlink:href="xptr(substring(//p[1]))">
  <s xlink:href="xptr(substring(//p[1]/s[1]))">
    <tok id="t1" xlink:href="substring(//p[1]/s[1]/text(),1,3)">
       <disamb>
          <pos>ART</pos>   <!-- ein -->
       </disamb>
    </tok>
    <tok id="t2" xlink:href="substring(//p[1]/s[1]/text(),5,4)">
       <disamb>
          <pos>NN</pos>    <!-- Mann -->
       </disamb>
    </tok>
    <tok id="t3" xlink:href="substring(//p[1]/s[1]/text(),10,5)">
       <disamb>
          <pos>VVFIN</pos> <!-- läuft -->
       </disamb>
    </tok>
  </s>
</p>
```

Abbildung 11. Stand-Off-Annotation mit XCES (Wortarten-Ebene).[302]

[301] Lezius 2002: 23.

Das Token *ein* wird zum Beispiel festgelegt, indem der XPointer auf eine Zeichenkette im Korpustext verweist, „die beim ersten Buchstaben im ersten Satz im ersten Paragraphen beginnt und drei Buchstaben lang ist [*xlink:href="substring(//p[1]/s[1]/text(),1,3)"*].“[303]

In einer dritten Datei wird eine weitere Sicht auf den Korpustext gespeichert, nämlich die Ebene der syntaktischen Annotation. Die Werte der Elemente `<feat>` und `<rel>` entnimmt Wolfgang Lezius dem TIGER-Annotationsschema. S steht für Satz, NP für Nominalphrase, SB für Subjekt, HD beschreibt den Kopf einer Phrase und NK den Bestandteil einer Nominalphrase (vgl. Abbildung 10, Seite 79).

```
<struct id="s0">

  <feat type="cat">S</feat>
  <rel type="SB" target="s1" />
  <rel type="HD" target="t2" />

  <struct id="s1">

    <feat type="cat">NP</feat>
    <rel type="NK" target="t0" />
    <rel type="NK" target="t1" />

    <struct id="t0">  <!-- ein -->
      <seg target="xptr(substring(//p[1]/s[1]/text(),1,3))" />
    </struct>

    <struct id="t1">  <!-- Mann -->
      <seg target="xptr(substring(//p[1]/s[1]/text(),5,4))" />
    </struct>

  </struct>

  <struct id="t2">    <!-- läuft -->
    <seg target="xptr(substring(//p[1]/s[1]/text(),10,5))" />
  </struct>

</struct>
```

Abbildung 12. Stand-Off-Annotation mit XCES (Ebene der syntaktischen Annotation).[304]

[302] Ebd.: 24.
[303] Lezius 2002: 24.
[304] Ebd.: 25.

Da der Beispielsatz keine kreuzenden Kanten aufweist, können die Knoten der Baumstruktur als sich einbettende <struct>-Elemente kodiert werden. Die von einem Knoten ausgehenden Kanten werden durch <rel>-Elemente modelliert und verweisen auf weitere Konstituenten oder Wörter.[305]

Das <seg>-Element zeigt wiederum mit XPointer auf die primäre Datei, die den Korpustext enthält. „Bei dieser Beispielmodellierung ist zu diskutieren, ob die Blätter aus linguistischer Sicht nicht besser an die Wortarten-Annotationsebene angebunden werden sollten, die bereits eine Tokenisierung vornimmt. Denn eine Änderung der Tokenisierung hätte sonst Anpassungen an zwei Stellen zur Folge.“[306]

Wenn man einen Satz syntaktisch annotieren will, der kreuzende Kanten aufweist, „so trägt die Einbettung als Modellierung der Teil-Ganzes-Beziehung nicht weit genug.“[307] Das bedeutet, dass eine Verschachtelung von <struct>-Elementen nicht ausreicht, die komplexe Struktur eines Satzes wie *Ein Mann kommt, der lacht* zu kodieren. Bei diesem Beispielsatz „liegt ein extraponierter Relativsatz vor, der über eine kreuzende Kante mit der Nominalphrase des Bezugsnomens verknüpft wird.“[308]

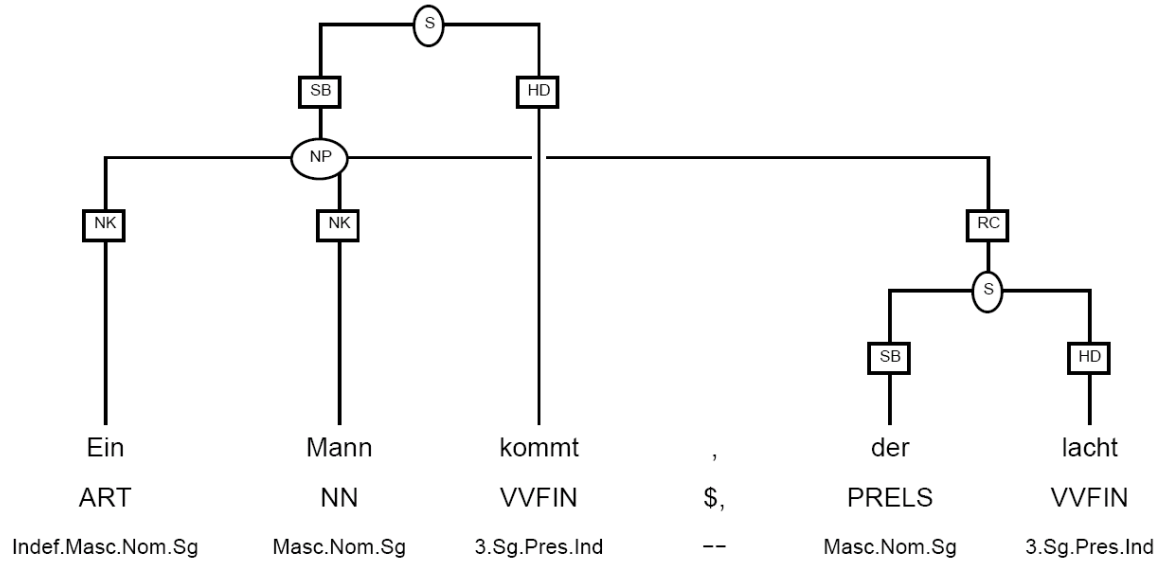

Abbildung 13. Baumstruktur mit kreuzenden Kanten.[309]

Wolfgang Lezius stellt fest, dass „eine Einbettung (…) eine Anordnung der inneren Knoten modellieren [würde], die hier nicht gegeben ist.“[310] Der XCES-Standard empfiehlt

[305] Vgl. Lezius 2002: 24.
[306] Ebd.: 24f.
[307] Ebd.: 25.
[308] Ebd.: 8.
[309] Ebd.: 19.

dazu, „alle Knoten als Geschwisterelemente anzugeben, wodurch sich die Satzstruktur durch entsprechende Referenzen ergibt (…).“[311] Sämtliche Referenzen werden mit dem Element `<rel>` und seinem Attribut *target* modelliert. Abbildung 14 zeigt die Kodierung des Satzes *Ein Mann kommt, der lacht* mit Hilfe des XCES-Referenzmechanismus:

```
<struct id="sentence_id">

  <struct id="s0">
    <feat type="cat">S</feat>
    <rel type="SB" target="s1" />
    <rel type="HD" target="t2" />
  </struct>
  <struct id="s1">
    <feat type="cat">NP</feat>
    <rel type="NK" target="t0" />
    <rel type="NK" target="t1" />
    <rel type="RC" target="s2" />
  </struct>
  <struct id="s2">
    <feat type="cat">S</feat>
    <rel type="SB" target="t3" />
    <rel type="HD" target="t4" />
  </struct>

  <struct id="t0">  <!-- Ein -->
    <seg target="xptr(substring(/p/s[1]/text(),1,3))" />
  </struct>

  <struct id="t1">  <!-- Mann -->
    <seg target="xptr(substring(/p/s[1]/text(),5,4))" />
  </struct>

  <struct id="t2">  <!-- kommt -->
    <seg target="xptr(substring(/p/s[1]/text(),10,5))" />
  </struct>

  <struct id="t3">  <!-- der -->
    <seg target="xptr(substring(/p/s[1]/text(),17,3))" />
  </struct>

  <struct id="t4">  <!-- lacht -->
    <seg target="xptr(substring(/p/s[1]/text(),21,5))" />
  </struct>

</struct>
```

Abbildung 14. XCES-Kodierung des Beispielsatzes *Ein Mann kommt, der lacht*.[312]

---

[310] Ebd.: 25.
[311] Lezius 2002: 25f.
[312] Ebd.: 26.

Demnach kann man syntaktische Annotationen im Rahmen von XCES auf zwei unterschiedliche Arten kodieren, nämlich durch Einbettung oder durch Referenzen.

> „Baumbanken wie die Penn Treebank werden sinnvollerweise den Einbettungsmechanismus verwenden, Baumbanken wie die TIGER-Baumbank hingegen den Referenzmechanismus. Auch ein Mischen der Ansätze ist zulässig."[313]

Für *Diachrone Syntax Deutsch* ist ebenfalls der Referenzmechanismus zu empfehlen, weil auf jeder Sprachstufe des Deutschen kreuzende Kanten auftreten. Gegen ein Mischen der Ansätze spricht die technische Erschwernis, dass stets beide Arten der Graph-Traversierung berücksichtigt werden müssen, zum Beispiel von verarbeitenden XSLT-Stylesheets.[314] Somit lässt sich sagen, dass der XCES-Standard ein mächtiges und durch die Möglichkeiten der Stand-Off-Annotation flexibles Werkzeug zur Kodierung syntaktischer Annotation darstellt.

[313] Lezius 2002: 27.
[314] Vgl. ebd.: 27.

5.3.3. Das TIGER-Projekt

5.3.3.1. Das TIGER-XML-Format

> „*Tiger* ist ein weiterer XML-basierter Standard zur Kodierung linguistisch annotierter Texte (…).“[315]

TIGER,[316] ein Gemeinschaftsprojekt der Universitäten Saarbrücken, Stuttgart und Potsdam, hatte zum Ziel, ein syntaktisch annotiertes Zeitungs-Korpus zu erstellen. TIGER stellt eine Weiterentwicklung des Negra-Projekts[317] dar. TIGER ist also nicht von der TEI abgeleitet, sondern will die syntaktischen Phänomene, die ursprünglich im Klammerstrukturformat des Negra-Projektes annotiert wurden, in einem eigenen XML-Format kodieren.[318] Darüberhinaus übernimmt das TIGER-Projekt von Negra zwei syntaktische Tagsets, mit denen sich sowohl Konstituenten als auch grammatische Funktionen kennzeichnen lassen. Als erstes wird das Design des TIGER-XML-Formats erläutert.

> „Die Anforderungen an das Design des XML-Formats sind klar definiert: Es geht um eine zum ursprünglichen Korpusdefinitionsformat semantisch äquivalente Korpusdefinition, die die Kodierung von Syntaxgraphen erlaubt.“[319]

Das Wurzelelement `<corpus>` wird mit einem *id*-Attribut versehen. Ein TIGER-XML-Dokument weist einen Header (`<head>`) auf; die Definition der Syntaxgraphen erfolgt im `<body>`-Element. Der Header setzt sich zusammen aus den Metadaten (`<meta>`) und der Attributdeklaration (`<annotation>`). Die Metadaten umfassen den Korpusnamen, den Autor des Korpus, das Datum der Korpuserstellung, eine Beschreibung des Korpus und das Format, in dem der Korpus ursprünglich kodiert wurde (zum Beispiel im Negra-Format). Die Deklaration der Attribute im Element `<annotation>` ist unter anderem für weiterverarbeitende Werkzeuge gedacht, die Informationen über die im Korpus auftretenden Kategorien des TIGER-Tagsets benötigen.[320]

Auf der folgenden Seite ist ein Dokumentheader für ein Korpus zu sehen, das allein aus dem Beispielsatz *Ein Mann läuft* (vgl. Abbildung 10, Seite 79) besteht.

---

[315] Tylman und Hinrichs 2004: 230.
[316] Vgl. TIGER 2007.
[317] Vgl. Negra 2005.
[318] Vgl. Lezius 2002: 418.
[319] Ebd.: 121.
[320] Vgl. ebd.: 121.

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<corpus id="DEMO">

<head>

  <meta>
    <name>Demokorpus</name>
    <author>Wolfgang Lezius</author>
    <format>TIGER-XML-Format</format>
  </meta>

  <annotation>

    <feature name="word" domain="T" />

    <feature name="pos" domain="T">
      <value name="ART">Artikel</value>
      <value name="NN">normales Nomen</value>
      <value name="VVFIN">finites Verb</value>
    </feature>

    <feature name="cat" domain="NT">
      <value name="NP">Nominalphrase</value>
      <value name="S">Satz</value>
    </feature>

    <edgelabel>
      <value name="HD">Kopf</value>
      <value name="NK">Teil des NP-Kerns</value>
      <value name="SB">Subjekt</value>
    </edgelabel>

  </annotation>

</head>
```

Abbildung 15. Dokumentheader des TIGER-XML-Formats.[321]

[321] Lezius 2002: 122.

Die linguistische Annotation des Satzes *Ein Mann läuft* wird im TIGER-XML-Format folgendermaßen kodiert:

```
<body>

<s id="graphid">

  <graph root="n2">

    <terminals>
      <t id="w1" word="ein"   pos="ART" />
      <t id="w2" word="Mann"  pos="NN" />
      <t id="w3" word="läuft" pos="VVFIN" />
    </terminals>


    <nonterminals>
      <nt id="n1" cat="NP">
        <edge label="NK" idref="w1" />
        <edge label="NK" idref="w2" />
      </nt>
      <nt id="n2" cat="S">
        <edge label="SB" idref="n1" />
        <edge label="NK" idref="w3" />
      </nt>
    </nonterminals>

  </graph>

</s>

...
</body>
```

Abbildung 16. Syntaxgraph im TIGER-XML-Format.[322]

Das TIGER-XML-Format entspricht einem hybriden Modell, welches „Konstituentenstrukturen mit funktionalen Abhängigkeiten verknüpft, wie sie Dependenzstrukturen ausdrücken."[323] Die Struktur und die Funktionen des Formats werden im Folgenden erläutert.[324]

[322] Lezius 2002: 128f.
[323] Ebd.: 8.
[324] Vgl. ebd.: 123-129.

Die Blätter einer Baumstruktur, welche die Wörter eines Satzes darstellen und auch als *terminale Knoten* bezeichnet werden, werden in der Gruppe <terminals> zusammengefasst. Ihnen wird das Element <t> zugewiesen; seine Attribute erlauben eine eindeutige Identifizierung des terminalen Knotens (*id*), enthalten die Zeichenkette des Wortes (*word*) und die zugeordnete Wortart (*pos*). An dieser Stelle können auch weitere Attribute hinzugefügt werden wie *morph* für flexionsmorphologische Angaben oder *lemma* für die Grundform eines Wortes – damit deckt dieser Teil des TIGER-XML-Formates die morphosyntaktische Annotationsebene ab. Durch die Anordnung der terminalen Knoten in einer Gruppe muss ihre Präzedenz, also ihre tatsächliche Reihenfolge im Korpustext, nicht explizit annotiert werden:

> „Da die Kinder eines XML-Elements in eindeutiger Weise angeordnet sind, kann die Präzedenz anhand der Position abgelesen werden: Das *n*-te Kind des <terminals>-Elements (also der n-te terminale Knoten) steht hinter dem (*n* – 1)-ten und vor dem (*n* + 1)-ten Kind."[325]

Die *nicht-terminalen Knoten* der Baumstruktur werden als <nt>-Elemente kodiert und unter der Bezeichnung <nonterminals> gruppiert. Um die Dominanz-Beziehung zwischen den einzelnen Knoten zu kodieren, enthält jeder nicht-terminale Knoten eine Liste aller ausgehenden Kanten, welche mit dem Element <edge> beschrieben sind.

In Abbildung 16 beschreibt der erste nicht-terminale Knoten die Nominalphrase *ein Mann*; seine erste Kante führt vom Knotenpunkt NP bzw. *n1* zum terminalen Knoten *w1*, welcher dem Artikel *ein* entspricht. Das Attribut *label* kodiert die Beschriftung der Kante; der Attributwert NK bestimmt dabei den Artikel als Bestandteil einer Nominalphrase.

Zur Kodierung des Wurzelknotens wird um die terminalen und nicht-terminalen Knoten das Element <graph> gespannt, welches mit dem Attribut *root* auf den Knoten der Baumstruktur verweist, der alle anderen dominiert.

Eine Besonderheit des TIGER-XML-Formates stellt die Möglichkeit dar, sekundäre Kanten zu annotieren. „Dazu werden analog zu den <edge>-Elementen für primäre Kanten <secedge>-Elemente für sekundäre Kanten eingeführt. Ein fiktives Beispiel einer sekundären Kante, die vom NP-Knoten zur Wortform läuft führt, sähe dann folgendermaßen aus:"[326]

---

[325] Lezius 2002: 125.
[326] Ebd.: 127.

```
<nt id="n1" cat="NP">

  <edge label="NK" idref="w1" />
  <edge label="NK" idref="w2" />

  <secedge label="LABEL" idref="w3" />

</nt>
```

Abbildung 17. Kodierung einer sekundären Kante.[327]

Sekundäre Kanten werden eingesetzt, wenn koordinierte Elemente auftreten. In Abbildung 18 findet sich eine koordinierte Nominalphrase (CNP, *Äpfel und Birnen*) und ein koordinierter Teilsatz (CS, *Er kauft und verkauft*). „Dabei kann es vorkommen, dass Konstituenten in einem der Konjunkte fehlen.“[328] Zu diesem Zweck benutzt man sekundäre Kanten. „Im vorliegenden Beispiel ist das Personalpronomen *Er* sowohl Subjekt des ersten [*kauft*] als auch des zweiten Konjunkts [*verkauft*]; *Äpfel und Birnen* ist entsprechend das Akkusativobjekt beider Konjunkte.“[329]

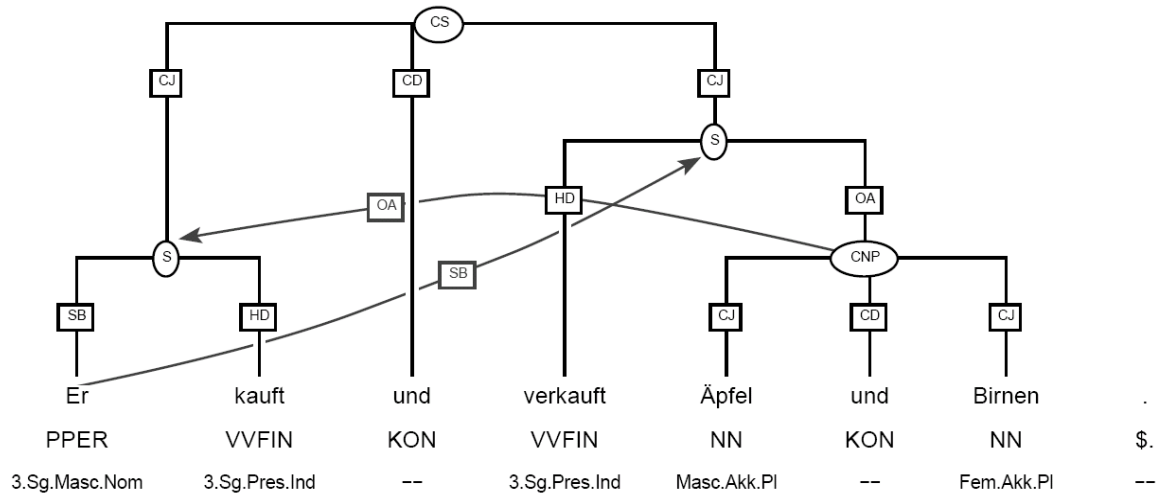

Abbildung 18. Baumstruktur mit sekundären Kanten.[330]

[327] Lezius 2002: 127.
[328] Ebd.: 8.
[329] Ebd.: 8.
[330] Ebd.: 9.

Damit Sätze in einem Korpus eindeutig identifiziert werden können, wird jedem Syntaxgraphen in einem weiteren umschließenden <s>-Element ein Identifikator zugewiesen. Außerdem dient das <s>-Element dazu, Ergebnisse von Suchanfragen aufzunehmen; dazu wird das Element <matches> als Geschwisterelement von <graph> eingefügt.[331] Für die Validierung des TIGER-XML-Formats steht ein öffentlich zugängliches XML-Schema zur Verfügung.[332]

> „In diesem Abschnitt ist die Konzeption des TIGER-XML-Formate beschrieben worden. Doch wie ist diese Konzeption im Vergleich zu XML-basierten Kodierungsansätzen wie XCES zu bewerten?“[333]

Der Nachteil des TIGER-XML-Formats ist die Tatsache, dass es keine Stand-Off-Annotation vorsieht, sondern sämtliche linguistischen Informationen in einer Datenstruktur zusammenfasst. Dadurch kann das Format nicht bzw. nur schwierig um andere Annotationsebenen erweitert werden. Da *Diachrone Syntax Deutsch* neben der morphosyntaktischen und syntaktischen auch eine textgrammatische Annotationsebene anstrebt, bietet der XCES-Standard mehr Flexibilität hinsichtlich der Kodierung. Nichtsdestotrotz ist das TIGER-Projekt ein wichtiger Ausgangspunkt für die Entwicklung des SynAF-Standards, der im folgenden Punkt vorgestellt wird. Das TIGER-XML-Format stellt eine auf die Software TIGERSearch[334] zugeschnittene Lösung dar; die Verwaltung der Annotation in einer Datei bringt dabei eigene Vorteile mit sich:

> „Daneben ist die Verwaltung aller Korpusdaten innerhalb einer einzigen Datei aus technischer Sicht die einfachste Lösung. Anwendungen wie der Korpusimport oder die Konvertierung in andere Format mit XSLT-Stylesheets können auf genau einem XML-Dokument arbeiten und müssen sich nicht um die Auflösung von Referenzen auf sekundäre Dateien kümmern.“[335]

Bei „entsprechend hoher Akzeptanz“[336] des XCES-Vorschlags zur Kodierung syntaktischer Annotation soll für TIGERSearch entweder ein Importfilter entwickelt oder „eine direkte Unterstützung als zweites Korpuseingangsformat“[337] umgesetzt werden.

Im anschließenden Punkt wird das TIGER-Annotationsschema beschrieben. Um zu zeigen, dass diese Tagsets einen guten Ausgangspunkt für die Entwicklung des syntaktischen *DiSynDe*-Schemas darstellen, wird der zweite Satz der *DiSynDe*-

---

[331] Vgl. Lezius 2002: 129-131.
[332] TIGER 2003.
[333] Lezius 2002: 132.
[334] TIGERSearch 2003.
[335] Lezius 2002: 132.
[336] Ebd.: 132.
[337] Ebd.: 132.

Annotationsvorschriften (*So wil ich dir weisen, mit we du den teuffel überwindest.*) nach dem TIGER-Annotationsschema beschrieben. Für die morphosyntaktische Annotation wird das große STTS-Tagset benutzt, die Kodierung erfolgt im TIGER-XML-Format. Abgesehen von der textgrammatischen Annotationsebene sind sämtliche Phänomene der *DiSynDe*-Annotationsvorschriften explizit kodiert bzw. implizit wieder abfragbar.

```
<s id="graphid">
   <graph root="n1">
      <terminals>
         <t id="w1" word="So" pos="ADV" />
         <t id="w2" word="wil" pos="VMFIN" morph="2.Sg.Ind.Pres" />
         <t id="w3" word="ich" pos="PPER" morph="1.Sg.*.Nom" />
         <t id="w4" word="dir" pos="PPER" morph="2.Sg.*Dat" />
         <t id="w5" word="weisen" pos="VVINF" />
         <t id="w6" word="," pos="$," />
         <t id="w7" word="mit" pos="APPR" morph="Instr" />
         <t id="w8" word="we" pos="PWS" morph="*.Instr.Sg" />
         <t id="w9" word="du" pos="PPER" morph="2.Sg.*.Nom" />
         <t id="w10" word="den" pos="ART" morph="Def.Masc.Akk.Sg" />
         <t id="w11" word="teuffel" pos="NN" morph="Masc.Akk.Sg._" />
         <t id="w12" word="überwindest" pos="VVFIN"
            morph="2.Sg.*.Pres" />
         <t id="w13" word="." pos="$." />
      </terminals>
      <nonterminals>
         <nt id="n1" cat="S">
            <edge label="MO" idref="w1" />
            <edge label="HD" idref="w2" />
            <edge label="SB" idref="w3" />
            <edge label="OC" idref="n2" />
         </nt>
         <nt id="n2" cat="VP">
            <edge label="DA" idref="w4" />
            <edge label="HD" idref="w5" />
            <edge label="OC" idref="n3" />
         </nt>
         <nt id="n3" cat="S">
            <edge label="OP" idref="n4" />
            <edge label="SB" idref="w9" />
            <edge label="OA" idref="n5" />
            <edge label="HD" idref="w12" />
         </nt>
         <nt id="n4" cat="PP">
            <edge label="AC" idref="w7" />
            <edge label="NK" idref="w8" />
         </nt>
         <nt id="n5" cat="NP">
            <edge label="NK" idref="n1" />
            <edge label="NK" idref="n2" />
         </nt>
      </nonterminals>
   </graph>
</s>
```

5.3.3.2. Ableitung des *DiSynDe*-Tagsets vom TIGER-Annotationsschema

Wie STTS auf Ebene der morphosyntaktischen Annotation Tags zur einheitlichen Beschreibung von Wortarten und flexionsmorphologischen Angaben bereitstellt, kommt im Rahmen des TIGER-Projekts auf Ebene der syntaktischen Annotation ebenfalls ein Annotationsschema zum Einsatz. Das Schema besteht aus zwei Tagsets, die als Basis für die Entwicklung spezieller *DiSynDe*-Tagsets dienen können, indem sie mit den *DiSynDe*-Annotationsvorschriften verglichen und um entsprechende Tags erweitert werden.

Das erste TIGER-Tagset besteht aus 48 Tags und dient der Beschreibung von grammatischen Funktionen; dadurch können die Kanten von Syntaxgraphen einheitlich beschriftet werden.[338]

AC adpositional case marker	MW way (directional modifier)
ADC adjective component	NG negation
AMS measure argument of adj	NK noun kernel modifier
APP apposition	NMC numerical component
AVC adverbial phrase component	OA accusative object
CC comparative complement	OA2 second accusative object
CD coordinating conjunction	OC clausal object
CJ conjunct	OG genitive object
CM comparative concjunction	OP object prepositional
CP complementizer	PAR parenthesis
DA dative	PD predicate
DH discourse-level head	PG pseudo-genitive
DM discourse marker	PH placeholder
EP expletive *es*	PM morphological particle
GL prenominal genitive	PNC proper noun component
GR postnominal genitive	RC relative clause
HD head	RE repeated element
JU junctor	RS reported speech
MC comitative	SB subject
MI instrumental	SBP passivised subject (PP)
ML locative	SP subject or predicate
MNR postnominal modifier	SVP separable verb prefix
MO modifier	UC (idiosyncratic) unit component
MR rhetorical modifier	VO vocative

[338] Vgl. Negra 1998a. Die Tags EP, PAR und OP sind durch die Weiterentwicklung des NEGRA-Tagsets im TIGER-Projekt entstanden (vgl. Albert et al. 2003).

Unter Bezugnahme auf die Einteilung der *DiSynDe*-Annotationsvorschriften in Annotationsebenen (vgl. Seite 34) fällt auf, dass zwar ein Tag für den Kopf von Phrasen existiert (HD), aber kein Tag für einen Determinator. Die Bestandteile einer Nominalphrase (*ein Mann*) werden als NK (*noun kernel modifier*) markiert, denn die „genauere Unterteilung kann aufgrund der Part-of-Speech bzw. kategorialen Information vorgenommen werden, so daß sich eine Unterscheidung auf der Ebene der Funktionslabels erübrigt."[339] *Head* wird eingesetzt, um zum Beispiel den Kopf einer Adjektivphrase zu markieren:

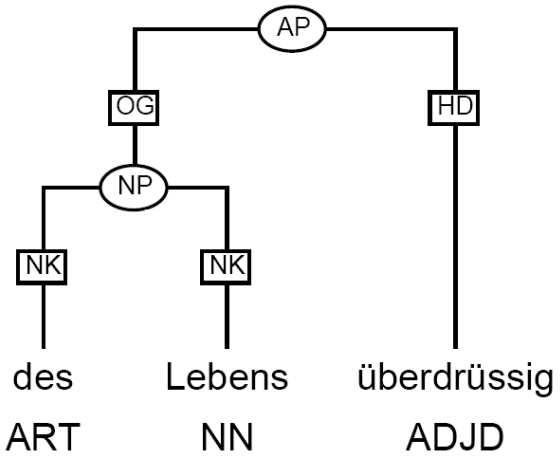

Abbildung 19. Kopf einer Adjektivphrase.[340]

Subjekte annotiert man als SB, während Prädikatsnomen bzw. Prädikative als PD (*predicate*) markiert werden können. Falls keine eindeutige Unterscheidung von Subjekt und Prädikativ möglich ist, wird SP (*subject or predicate*) eingesetzt.

Für Dativobjekte steht der Tag DA zur Verfügung. TIGER sieht dabei keine Unterscheidung von echten Dativobjekten und freien Dativen vor (*Jemand hat [ihm]$_{DA}$ sein Auto geklaut.*)[341]. Nach Überlegungen der Gruppe Einfacher Satz ist eine solche Spezifizierung wünschenswert. Im Sinne eines hierarchischen Aufbaus von Tags (vgl. *5.2. Morpho-syntaktische Annotation mit STTS,* Seite 67f) wird DAF (*freier Dativ*) vorgeschlagen.

Akkusativobjekte annotiert man mit OA; für ein weiteres Akkusativobjekt im Satz gibt es OA2 (*der Tanzlehrer lehrt [den Schüler]$_{OA}$ [einen Tanz]$_{OA2}$*). Freie Akkusative werden als MO markiert; die Kasus-Information der Konstituente ergibt sich über die morpho-syntaktische Ebene (*Paul hat [den ganzen Tag]$_{MO}$ [den Rasen]$_{OA}$ gemäht*).[342] Um diesen freien

[339] Albert et al. 2003: 9.
[340] Ebd.: 38.
[341] Vgl. ebd.: 55.
[342] Vgl. ebd. 2003: 52.

Akkusativ auch als Zeitangabe zu kennzeichnen, kann der Tag MT für temporale Modifikationen eingeführt werden (in Anlehnung an den bereits vorhandenen Tag ML für lokale Modifikationen).

Mit OP (*object prepositional*) und MO (*modifier*) kann zwischen Präpositionalobjekten und Präpositionalphrasen unterschieden werden, „die als Modifikatoren (MO) fungieren. (…) Ein Präpositionalobjekt zeichnet sich dadurch aus, daß seine Präposition infolge eines Abstraktionsprozesses an das Verb gebunden ist. Dabei verliert sie ihren lexikalischen Gehalt und nimmt funktionalen Charakter an (…).“[343]:

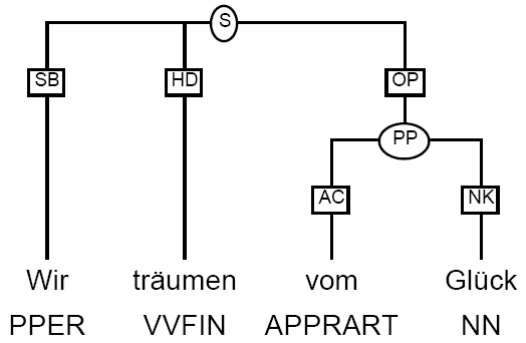

Abbildung 20. Präpositionalobjekt.[344]

Die modifizierenden Präpositionalphrasen werden in den *DiSynDe*-Annotationsvorschriften als Adverbiale bezeichnet:

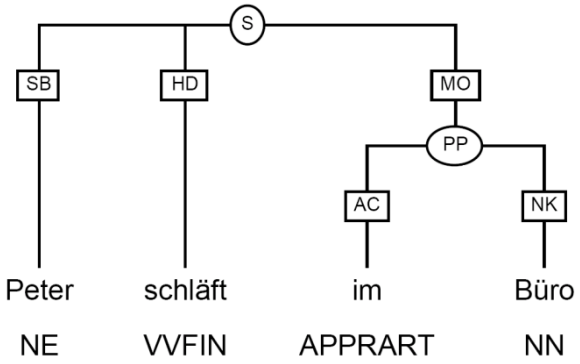

Abbildung 21. Lokale Adverbialbestimmung.[345]

In diesem Beispielsatz liegt eine *lokale* Adverbialbestimmung vor. Für *DiSynDe* sollen entsprechende Spezifizierungen annotiert werden. Für diese Zwecke muss deshalb der bereits angelegte Tag ML (*locative*) konsequent angewandt werden.

[343] Albert et al. 2003: 56.
[344] Ebd.: 29.
[345] Ebd.: 48.

Desweiteren werden Adverbialsätze nach dem TIGER-Tagset als Modifikatoren (MO) von Hauptsätzen annotiert. Die Unterscheidung von modifizierenden Präpositionalphrasen (PP, vgl. Abbildung 21) und adverbialen Nebensätzen (S, vgl. Abbildung 22) erfolgt dabei durch die Knotenlabels.

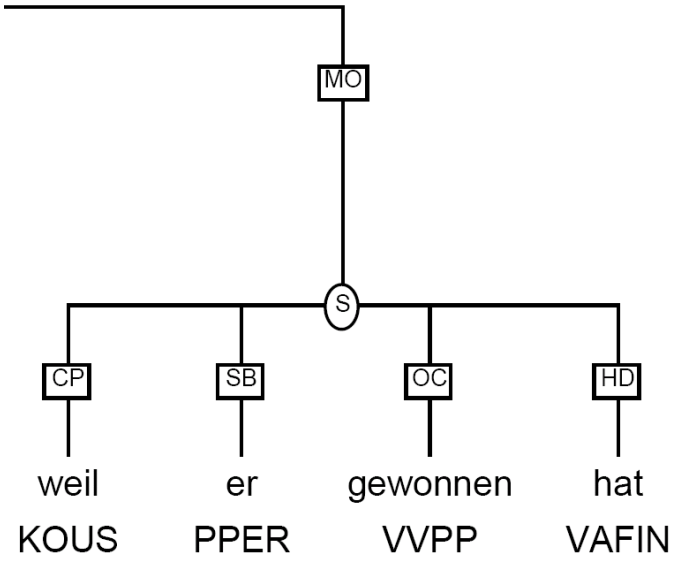

Abbildung 22. Kausalsatz.[346]

Wenn man den Adverbialsatz als Kausalsatz annotieren will, muss man das Tagset um den Tag MK (kausale Modifikation) erweitern. In Abbildung 22 wird ebenfalls ersichtlich, dass Satzeinleitungen als CP (*complementizer*) markiert werden. Man könnte also auch statt MO den Tag CP direkt erweitern. Allerdings könnte man bei einer solchen Vorgehensweise uneingeleitete Nebensätze nicht spezifizieren:

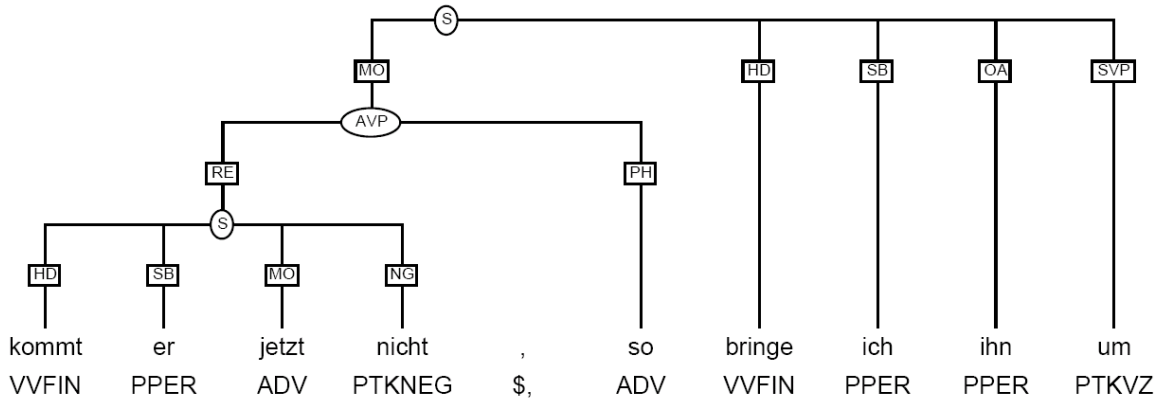

Abbildung 23. Durch Inversion eingeleiteter Konditionalsatz.[347]

[346] Albert et al. 2003: 65.
[347] Ebd.: 68.

Diese Überlegungen führen zu folgenden Erweiterungen des TIGER-Tagsets:

MA (adversative Modifikation)
MF (finale Modifikation)
MK (kausale Modifikation)
MN (konditionale Modifikation)
MS (konsekutive Modifikation)
MT (temporale Modifikation)
MZ (konzessive Modifikation)

Der Tag MI für instrumentale Modifikationen ist im TIGER-Tagset bereits angelegt.

Bei der Annotation von Adverbialsätzen kann es vorkommen, dass man nicht eindeutig entscheiden kann, ob beispielsweise ein Kausal- oder ein Temporalsatz vorliegt; die mittelhochdeutsche Konjunktion *sît (daz)* zur Einleitung von Kausalsätzen etwa „kann auch zur Einleitung temporaler Nebensätze verwendet werden (…).“[348] In solchen Fällen kann man entweder nur MO zuweisen oder mehrere Tags durch einen Punkt getrennt zusammenfügen (MK.MT, vgl. die Vorgehensweise bei der Annotation von flexionsmorphologischer Information mit STTS, Seite 71).

Relativsätze werden mit dem Label RC als abhängiger Knoten an die entsprechende Nominalphrase angebunden:

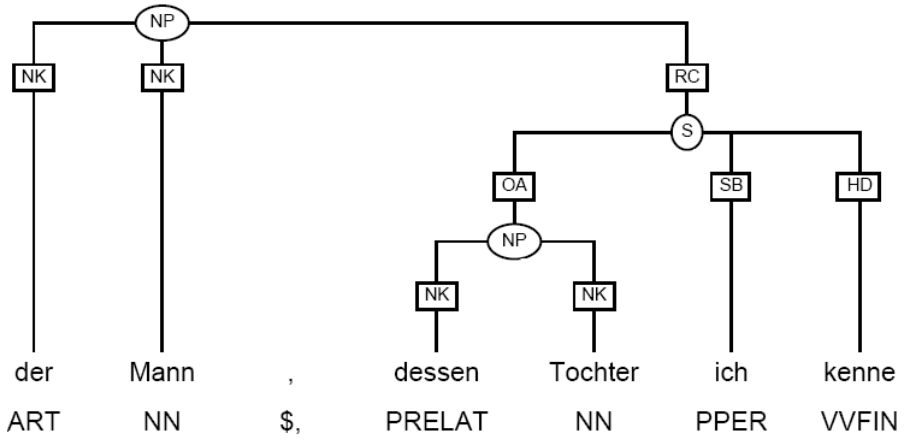

Abbildung 24. Relativsatz.[349]

[348] Hennings 2003: 194.
[349] Albert et al. 2003: 30.

Subjektsätze werden auf Ebene der Teilsätze mit dem SB-Tag annotiert (*[Daß der Duden immer Recht hat]*$_{SB}$*, ist unumstritten.*), während Objektsätze als OC (*clausal object*) markiert werden:

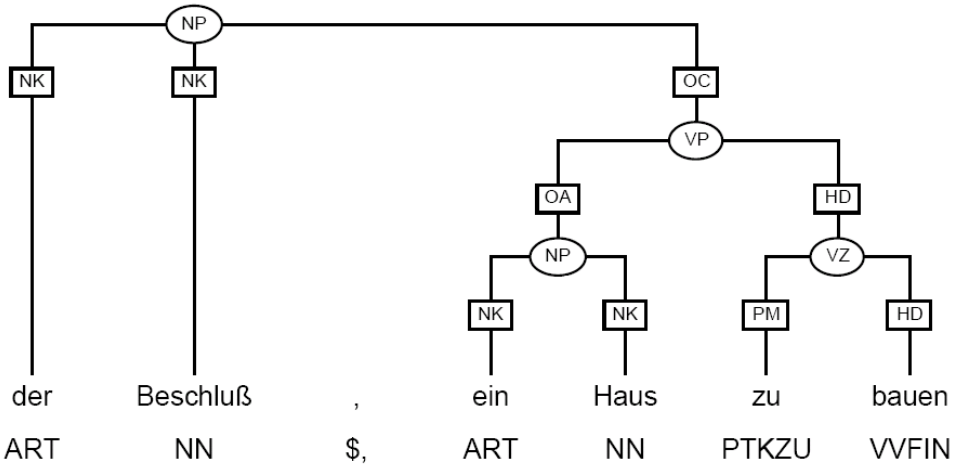

Abbildung 25. Objektsatz (Infinitivkonstruktion).[350]

Während das expletive *es* mit EP gekennzeichnet wird (*weil [es]*$_{EP}$ *heute regnet*), sind für die Annotation von *es*-Korrelaten die Labels PH (*placeholder*) und RE (*repeated element*) gedacht:

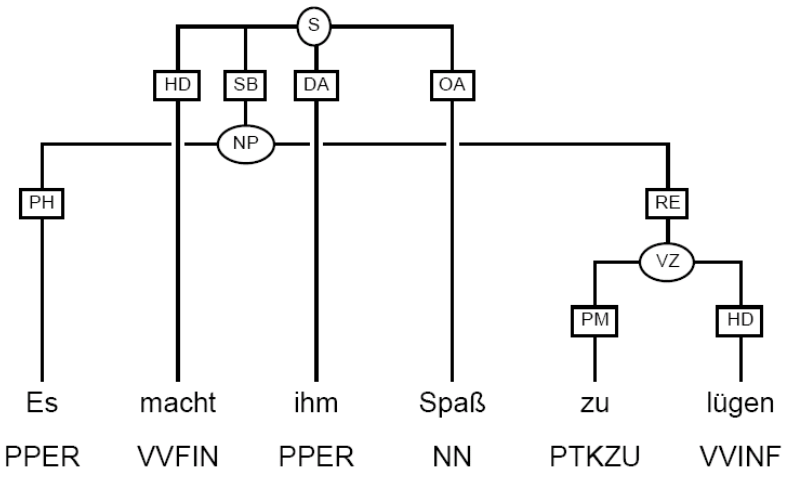

Abbildung 26. Annotation von *es*-Korrelaten.[351]

[350] Albert et al. 2003: 27.
[351] Ebd.: 77.

Das zweite TIGER-Tagset stellt 25 Bezeichner zur Verfügung, mit denen man Phrasenkategorien bzw. nicht-terminale Knoten kennzeichnen kann:[352]

AA superlative phrase with "am"	CVZ coordinated zu-marked infinitive
AP adjektive phrase	DL discourse level constituent
AVP adverbial phrase	ISU idiosyncratis unit
CAC coordinated adposition	MPN multi-word proper noun
CAP coordinated adjektive phrase	MTA multi-token adjective
CAVP coordinated adverbial phrase	NM multi-token number
CCP coordinated complementizer	NP noun phrase
CH chunk	PP adpositional phrase
CNP coordinated noun phrase	QL quasi-language
CO coordination	S sentence
CPP coordinated adpositional phrase	VP verb phrase (non-finite)
CS coordinated sentence	VZ zu-marked infinitive
CVP coordinated verb phrase (non-finite)	

Für *DiSynDe* müssen an dieser Stelle keine Erweiterungen vorgenommen werden. Sämtliche Konstituenten können angemessen annotiert werden (S für Satz und Teilsatz, VP für Verbalphrase, NP für Nominalphrase, AVP für Adverbialphrase, PP für Präpositionalphrase). Für koordinierte Phrasen stehen entsprechende Tags zur Verfügung (beispielsweise CNP für koordinierte Nominalphrasen oder CCP für koordinierte Satzeinleitungen). Mit CH (*chunk*) wird fremdsprachliches Material markiert; als grammatische Funktion wird dabei der Tag UC (*unit component*) vergeben.[353]

In Anhang 3 findet sich eine Zusammenstellung des ersten, erweiterten TIGER-Tagsets (*Dependenzstruktur*) und des zweiten TIGER-Tagsets (*Konstituentenstruktur*). Bei jeder Erweiterung des Annotationsschemas sollte bedacht werden, ob dadurch nicht eine Überspezifizierung des Tagsets stattfindet und ob eine entsprechende Interpretation nicht auch auf Seite der Auswertung von Korpusbelegstellen vorgenommen werden kann.

[352]Vgl. Negra 1998b.
[353] Vgl. Albert et al. 2003: 18.

5.3.4. Syntactic Annotation Framework (SynAF)

Das Projekt *Linguistic Infrastructure for Interoperable Resources and Systems* (LIRICS)[354] beschäftigt sich auf europäischer Ebene und in Zusammenarbeit mit der *International Organization for Standardization* (ISO)[355] mit der Standardisierung von Sprachtechnologien; unter anderem arbeitet man an einem ISO-Standard für syntaktische Annotation. „Ein entsprechendes Work Item wurde dazu dem ISO Commitee TC 37/SC4 vorgelegt und bereits akzeptiert."[356]

Bei der Erstellung von Korpora besteht das grundsätzliche Problem, dass die verschiedenen Projekte für ihre Zwecke spezielle Datenformate und Annotationsrichtlinien definieren. Dadurch entstehen Wissensinseln, deren Zusammenführung und gemeinsame Nutzung nur schwierig möglich ist. Bei einer entsprechenden Standardisierung könnte man zum Beispiel einzelne syntaktisch annotierte historische Korpora zu einem diachronen Korpus verbinden. Doch die vorhandenen „idiosynkratischen Strukturen erlauben es (…) nicht, Daten zwischen Applikationen oder Nutzern auszutauschen, ohne vorher eine detaillierte Analyse und Transformation durchzuführen, selbst wenn Teilstrukturen und Informationsgehalt direkt vergleichbar wären."[357]

SynAF strebt folgendes Ziel an:

> „SynAF is dealing with the description of a meta-model for syntactic annotation, which means that SynAF will describe elementary linguistic (in fact syntactic) abstractions that support the construction and the interoperability of (syntactic) annotations and resources as well as the procedure for the creation of data categories for syntactic annotation. SynAF will thus not propose a tagset for syntactic annotation, but is dedicated to proposing a (possibly hierarchical) list of data categories, which is much easier to update and extend, and which will represent a point of reference for particular tagsets used for the syntactic annotation of various languages, also in the context of various application scenarios."[358]

SynAF soll also für verschiedene Sprachen eine syntaktische Annotation ermöglichen. Da es nicht praktikabel ist, sämtliche syntaktischen Phänomene unterschiedlichster Sprachen in ein allgemein gültiges Tagset zu fassen, wird ein Metamodell für syntaktische Annotation entwickelt. Das bedeutet, dass eine hierarchische Liste von Datenkategorien erstellt wird, die als Grundlage für eine einheitliche syntaktische Annotation dienen kann.

---

[354] Vgl. LIRICS 2005.
[355] Vgl. ISO 2008.
[356] Trippel et al. 2005: 22.
[357] Ebd.: 17f.
[358] Declerck 2006: 1.

Man stellt aus dieser Liste von Datenkategorien ein eigenes Tagset zusammen; umgekehrt sollte es möglich sein, ein vorhandenes Tagset auf die Liste abzubilden und so zur besseren Wiederverwendbarkeit des eigenen Projektes standardisierte Tag-Definitionen einzusetzen. Theorie- und Sprachunabhängigkeit sollen erreicht werden, indem möglichst viele oder idealerweise alle syntaktischen Phänomene in die erweiterbare Liste von Datenkategorien integriert werden. Da jede „Sprachstufe des Dt. (…) mit allen Syntaxtheorien beschrieben werden [kann], die auch auf die dt. Gegenwartssprache oder andere Einzelsprachen anwendbar sind"[359], müssen sich mit SynAF prinzipiell auch jegliche syntaktischen Phänomene historischer und diachroner Korpora annotieren lassen.

Bei der Konzeption des Standards sichtete man als erstes syntaktisch annotierte Korpora verschiedener Sprachen wie Tschechisch, Englisch, Französisch, Deutsch, Italienisch, Japanisch und Türkisch.[360] Es kristallisierte sich die Vorgabe heraus, dass das SynAF-Metamodell die „zwei Haupttypen von syntaktischen Annotationen"[361] abdecken muss, nämlich Konstituenten- und Dependenzstrukturen. Deshalb wurden als Ausgangspunkt für den Standardisierungsprozess die Korpusprojekte TIGER und ISST (*Italien Semantic-Syntactic Treebank*) festgelegt:

> „We found some approaches (e.g. the Negra/Tiger initiatives in Germany, or the ISST, Italian Semantic-Syntactic Treebank, framework for Italian) proposing coherent frameworks accounting for both (hierarchical) constituency and dependency phenomena in syntactic representation. We consider for the time being those 2 initiatives as the starting point for SynAF, which will abstract over the particular annotation strategies and tagsets proposed."[362]

Der Begriff *Konstituentenstrukturen* bezieht sich auf strukturierte Sequenzen morphosyntaktisch annotierter Einheiten (zum Beispiel Nominalphrasen); dabei können auch nicht-benachbarte Elemente Konstituenten bilden. Mit Dependenzstrukturen wird die Beziehung zwischen syntaktischen Elementen bezeichnet; dabei wird zwischen interner und externer Dependenz unterschieden („we speak of an *internal dependency* and (…) of an *external dependency*"[363]).

Interne Dependenz bezeichnet die Beziehung der Elemente innerhalb von Phrasen, beispielsweise die Abhängigkeit zwischen Adjektiv und Nomen einer Nominalphrase: das Adjektiv modifiziert das Nomen, den Kopf der Phrase. Externe Dependenz bezeichnet Beziehungen zwischen Konstituenten auf der Ebene von Teilsätzen oder

---

[359] Meineke und Schwerdt 2001: 307.
[360] Vgl. Declerck 2006: 2.
[361] Trippel et al. 2005: 23.
[362] Declerck 2006: 2.
[363] Ebd.: 2.

Sätzen (zum Beispiel die grammatische Funktion einer Nominalphrase als *Subjekt* des Satzes).[364]

Ferner soll SynAF kein isolierter Standard werden, sondern in das *Linguistic Annotation Framework* (LAF), ebenfalls ein ISO-Standard, eingebettet werden.[365] LAF versucht ein Metaformat zu definieren und damit „eine einheitliche Grundlage für die Annotation von linguistischen Daten zu legen. Dabei liegt ein Schwerpunkt auf höheren Annotationsebenen, etwa morphosyntaktische, syntaktische und semantische Annotation, die auf tieferen Ebenen aufsetzen, ohne dabei gegenüber anderen Bereichen abgeschlossen zu sein."[366]

Das *Morpho-Syntactic Framework* (MAF) ist eine weitere Grundlage für den SynAF-Standard:

> „SynAF will build on the ISO MAF proposal (…). MAF (Morpho-Syntactic Framework) is dealing with the morpho-syntactic annotation of specific segments of textual documents. The morpho-syntactic annotation framework is about *part of speech* (noun, adjective, verb, etc.), *morphological* and *grammatical* features (such as number, gender, person, mood, verbal tense)."[367]

Als tatsächlicher Startpunkt für SynAF fungiert das TIGER-Projekt, sowohl hinsichtlich des TIGER-XML-Formats als auch hinsichtlich des TIGER-Annotationsschemas:

> „The Tiger annotation framework foresees 2 types of annotation: for constituency (represented than by a *node* label in the annotation framework) and for dependency (represented as an *edge* label in the annotation framework). This annotation strategy has reached in the meantime a kind of consensus within the corpus linguistics. We consider this to be a good basis for starting our standardization work in SynAF."[368]

Die italienische Baumbank ISST, eine weitere Basis für SynAF, stellt kein hybrides Modell wie TIGER dar, sondern annotiert Dependenzstrukturen von Wörtern und nicht von Konstituenten. Es ist noch eine offene Frage, wie sich dieser Ansatz innerhalb von SynAF mit anderen Modellen in Einklang bringen lässt:

> „An important feature of ISST is that it annotates word with dependency information, and not the syntactic constituents. We will have to see how to

---

[364] Vgl. Declerck 2006: 1f.
[365] Vgl. ISO/TC 37/SC 4 N421 2007: 11f.
[366] Trippel et al. 2005: 19.
[367] Declerck 2006: 1.
[368] Ebd.: 2.

> accommodate this with that approaches (like Tiger), which associate dependency mostly to constituents."[369]

Die aktuelle Liste der SynAF-Datenkategorien wird an dieser Stelle nicht abgebildet, da sie nicht abgeschlossen ist („In this document we present the actual list of candidates, as they have been detected in annotation initiatives like TIGER, ISST, Sparkle and EAGLES, and modified/harmonized for the purpose of this document."[370]). Die TIGER-Tagsets bieten derzeit einen besseren Ausgangspunkt für ein *DiSynDe*-Annotationsschema.

Das SynAF-XML-Format entspricht in seinen Grundzügen dem TIGER-XML-Format. Im Header erfolgt die Definition der einsetzbaren Tags; sie bezeichnen entweder Konstituenten (`<nodelabel>`) oder Dependenzstrukturen (`<edgelabel>`). Der Syntaxgraph wird ebenfalls in terminale (`<terminals>`) und nicht-terminale Knoten (`<nonterminals>`) eingeteilt. Ein Unterschied zum TIGER-Format besteht darin, dass die Knoten (`<node>`) und Kanten (`<edge>`) als Geschwisterelemente kodiert sind. Die `<node>`-Elemente verweisen mit den Attributen *from* und *to* auf ihren Inhalt; die `<edge>`-Elemente zeigen mit dem Attribut *t_node* oder *nt_node* auf von ihnen abhängige Knotenpunkte. Bei den terminalen Knoten fällt das Attribut *lemma* auf.

Auf den folgenden zwei Seiten ist der Header und der Syntaxgraph des Beispielsatzes *Für Angaben in unseren Listen wurde grundsätzlich die weitestgehende Bilanz zugrunde gelegt* im SynAF-XML-Format annotiert.[371]

[369] Declerck 2006: 4.
[370] ISO/TC 37/SC 4 N421 2007: 15.
[371] Vgl. ebd.: 20ff.

```
<SynAF>
   <head>
      <annotation>

         <nodelabel>
            <feature name="wordForm" domain="T" />
            <feature name="pos" domain="T" />
            <value name="adjective" />
            <value name="subordinatingConjunction" />
            <value name="particle" />
            <value name="pronoun" />
            <value name="reflexivePronoun" />
            <value name="verb" />
            <value name="auxiliaryVerb" />
            <value name="modalVerb" />
            <value name="coordinatingConjunction" />
            <value name="definiteDeterminer" />
            <value name="indefiniteDeterminer" />
            <value name="adverb" />
            <value name="prefix" />
            <value name="ordinalNumeral" />
            <value name="interjection" />
            <value name="person" />
            <value name="noun" />
            <value name="conjunction" />
            <value name="properNoun" />
            <value name="punctuation" />
            <value name="possessive" />
            <value name="numeral" />
            <value name="cardinal" />
            <value name="preposition" />
            <value name="relPronoun"/>
            <feature name="wordForm" domain="T" />
            <value name="NP" />
            <value name="PP"/>
            <value name="AP"/>
            <value name="ADVP"/>
            <value name="VG"/>
            <value name="SUBORDCLAUSE"/>
            <value name="C"/>
            <value name="sentence"/>
            <value name="clause"/>
            <value name="S"/>
            <value name="DA"/>
         </nodelabel>

         <edgelabel>
            <value name="subject"/>
            <value name="deepObject"/>
            <value name="directObject"/>
            <value name="indirectObject"/>
            <value name="prepositionPhraseAdjunct"/>
            <value name="predicativeAdverbial"/>
            <value name="xComp"/>
            <value name="head"/>
            <value name="mod"/>
            <value name="spec"/>
         </edgelabel>

      </annotation>
   </head>
```

```
   <body>
   <graph>

         <nonterminals>
         <sentence id="1">
            <edge id="s1_3" label="DEEP_OBJ" nt_node="8" />
            <edge id="s1_4" label="PP_ADJUNCT" nt_node="5" />
            <node id="s1_5" label="PP" from="14" to="24" />
            <node id="s1_6" label="VG" from="26" to="26" />
            <node id="s1_7" label="AdvP" from="28" to="28" />
            <edge id="s1_9" label="head" t_node="34" />
            <edge id="s1_10" label="spec" t_node="30" />
            <edge id="s1_11" label="mod" t_node="32" />
            <node id="s1_12" label="VG" from="36" to="38" />
            <node id="s1_13" label="S" from="39" to="39" />
         </sentence>
         </nonterminals>

         <terminals>
            <t id="14" wordForm="Fuer" lemma=" fuer" pos="preposition" />

            <t id="18" wordForm="Angaben" lemma=" angabe"
             pos="definiteDeterminer" />

            <t id="20" wordForm="in" lemma=" in" pos="preposition" />

            <t id="22" wordForm="unseren" lemma=" unser" pos="possessive"
             />

            <t id="24" wordForm="Listen" lemma=" list" pos="noun" />

            <t id="26" wordForm="wurde" lemma=" werd" pos="auxiliaryVerb"
             />

            <t id="28" wordForm="grundsaetzlich" lemma=" grundsaetzlich"
             pos="adverb" />

            <t id="30" wordForm="die" lemma="" pos="definiteDeterminer"
             />

            <t id="32" wordForm="weitestgehende" lemma=" weitestgehend"
             pos="adjective" />

            <t id="34" wordForm="Bilanz" lemma=" bilanz" pos="noun" />

            <t id="36" wordForm="zugrunde" lemma=" zugrunde"
             pos="particle" />

            <t id="38" wordForm="gelegt" lemma=" leg" pos="verb" />

            <t id="39" wordForm="." lemma=" ." pos="punctuation" />
         </terminals>

   </graph>
   </body>
</SynAF>
```

5.4. Zur textgrammatischen Annotationsebene

Das *DiSynDe*-Korpus soll auch textgrammatisch annotiert werden, da auch die linguistische Ebene der Textgrammatik auf den Wandel der deutschen Syntax einen Einfluss hat:

> „Auch textlinguistische und sprachpragmatische Gesichtspunkte (…) sollen berücksichtigt werden, denn mittlerweile gilt es als communis opinio, dass Grammatik und mithin die Syntax kein hermetisches, autonomes System darstellt, das sich ausschließlich aufgrund innerer Faktoren verändert."[372]

Die Textlinguistik betrachtet nicht den Satz, sondern den Text als größte sprachliche Einheit. Für den Begriff *Text* gibt es zahlreiche Definitionen; eine weitgehend akzeptierte Definition ist der *integrative Textbegriff* von Klaus Brinker:[373]

> „Der Terminus ‚Text' bezeichnet eine begrenzte Folge von sprachlichen Zeichen, die in sich kohärent ist und die als Ganzes eine erkennbare kommunikative Funktion signalisiert."[374]

Ein Text zeichnet sich also durch seine Kohärenz aus. Wenn man die Kohärenz eines Textes untersucht, will man die „Beziehungen zwischen den Sprachzeichen eines Textes (…) erfassen" [375] und die Mittel analysieren, mit denen Kohärenz im Text hergestellt wird. Brinker spricht dabei von *verbaler Kohärenz*. Dieser Begriff umfasst sowohl die semantische Kohärenz, die sich auf die Bedeutung, das Thema und den Sinnzusammenhang eines Textes bezieht, als auch die grammatische Kohärenz, welche die Verknüpfungsmittel auf grammatischer Ebene zur Herstellung von Textkohärenz betrachtet.[376]

Ein Mittel zur Herstellung von verbaler Kohärenz ist zum Beispiel die *Referenz*, welche die Beziehung zwischen einem Sprachzeichen (dem Referenzmittel) und einem außersprachlichen Bezugsobjekt (dem Referenzobjekt) beschreibt. „Von Referenzidentität spricht man, wenn sich zwei Referenzmittel auf dasselbe Referenzobjekt beziehen."[377]

Dabei gilt in der Textgrammatik das transphrastische Prinzip, das besagt, „dass verbale Kohärenz oberhalb der Satzebene zu untersuchen ist."[378] Wenn man also innerhalb eines Satzes zweimal auf dasselbe Bezugsobjekt referiert, bewegt man sich auf der syntaktischen Ebene (***Theresa*** *geht mit* ***ihrem*** *Hund spazieren.*). Wenn sich jedoch die Referenzmittel in verschiedenen Sätzen befinden, liegt ein textgrammatisches Phänomen

---

[372] Schmid 2007: 52.
[373] Vgl. Kessel und Reimann 2005: 199.
[374] Brinker (2005): 17.
[375] Kessel und Reimann 2005: 203f.
[376] Vgl. ebd. 199f.
[377] Ebd.: 204.
[378] Ebd.: 204.

vor: ***Theresa*** *hat einen Hund.* ***Sie*** *geht mit ihm spazieren.* Daraus lässt sich folgern, dass man einen zu analysierenden Text als erstes in minimale textgrammatische Einheiten aufteilt.[379]

Beim Prozess der textgrammatischen Annotation entspricht dies der Segmentierung eines Textes in *elementary discourse units* (EDUs).[380] Dabei entsprechen minimale Texteinheiten bzw. EDUs nicht einfach Sätzen. Parataktisch verbundene Hauptsätze etwa werden als eigenständige Sätze behandelt, sodass diese unter das transphrastische Prinzip fallen.[381] Allerdings muss bei der Annotation ebenfalls eine satzinterne Segmentierung erfolgen, um beispielsweise auch die Verbindung (Konnexion) zwischen intraphrastisch verbundenen Teilsätzen untersuchen zu können. Im folgenden Beispielsatz sind die Teilsätze durch die Subjunktion *weil* verbunden: *Weil ich morgen mündliche Prüfung in Sprachwissenschaft habe, kann ich heute nicht mit euch ins Kino gehen.* Bei der Behandlung intraphrastischer Relationen befindet man sich wieder auf der Syntaxebene.[382] Dies bedeutet, dass man sich bei der textgrammatischen Annotation fließend zwischen Syntax und Textlinguistik bewegt.

Die satzinterne Segmentierung wird beim DISCOR-Projekt (*Discourse Structure and Coreference Resolution*) folgendermaßen gehandhabt:

> „Punctuation and discourse markers are good surface syntactic cues for detecting sentence internal EDUs. Segment boundaries are placed *after* punctuation – including periods, commas, hyphens, colons and semi-colons – but *before* discourse connectors such as 'and,' 'or,' etc. and complementizers such as 'that,' 'if,' 'wether,' etc."[383]

Das Projekt *Diachrone Syntax Deutsch* will die grammatischen Mittel, die Textkohärenz bewirken, von den Anfängen bis zur Gegenwart untersuchen. In den *DiSynDe*-Annotationsvorschriften (siehe Anhang 1) finden sich dazu vier Bereiche, die annotiert werden sollen, nämlich Textelement, Referenz, Textanschluss und Diskurs. Auf der folgenden Seite sind diese Bereiche mit ihren jeweiligen Ausprägungen in den Annotationsvorschriften dargestellt.

---

[379] Vgl. Kessel und Reimann 2005: 204.
[380] Vgl. Reese et al. 2007: 2.
[381] Vgl. Kessel und Reimann 2005: 204.
[382] Vgl. ebd.: 207.
[383] Reese et al. 2007: 3.

Textgrammatische Annotation

1. **Textelement**
    - *Zitat*
    - *Direkte Rede*
    - *Appell*
    - *Quellennachweis*
    - *Kommentar*

2. **Referenz**
    - *Anaphorisch*
    - *Kataphorisch*
    - *Rekurrent antonym*

3. **Textanschluss**
    - *Adverb*
    - *Pronominal*
    - *Konjunktion*
    - *Subjunktion*
    - *Expliziter Rückverweis*
    - *Rekurrent*

4. **Diskurs**
    - *Konsequenz*
    - *Elaboration*
    - *Continuation*
    - *Result*

Manfred Stede behandelt die Annotation von Textgrammatik im Sinne einer Mehrebenenannotation: „Die einzelnen Annotationen sollten strikt voneinander getrennt sein, damit sie separat recherchiert und ggf. auch verändert werden können."[384] Es werden also auf der textgrammatischen Ebene eigene Annotationsebenen unterschieden. Stede behandelt dabei die Ebene der Referenziellen Struktur, der Thematischen Struktur und deren Verbindung zur Informationsstruktur von Texten; desweiteren werden die Temporale Struktur, die Illokutionsstruktur, die Argumentationsstruktur und die Rhetorische Struktur von Texten ausgearbeitet. In einem weiteren Punkt wird die Verknüpfung von minimalen Texteinheiten durch Konnektoren thematisiert.

Wenn man die textgrammatischen Ebenen von *DiSynDe* auf diese Einteilung überträgt, entspricht Ebene 1 (Textelement) der Illokutionsstruktur, Ebene 2 (Referenz) der Referenziellen Struktur, Ebene 3 (Textanschluss) der Verknüpfung minimaler Texteinheiten durch Konnektoren (Konnexion) und Ebene 4 (Diskurs) der Rhetorischen Struktur.

Die Illokution eines Textes bezeichnet seine dominierende Sprechhandlung. In Ebene 1 wird als Annotation eines Textelements *Apell* angegeben. Eine Werbeanzeige hat beispielsweise die dominierende Illokution „APPELLIEREN, INFORMIEREN"[385].

Folglich hat ein Text eine bestimmte kommunikative Funktion. „Diese kommunikative Funktion muss sich innerhalb des Textes manifestieren bzw. aus dem Text ableiten lassen."[386] Eine wichtige Ebene von Texten ist somit die Illokutionsstruktur:

> „Für jeden Text lässt sich (…) eine übergeordnete Illokution, eine Metaillokution, identifizieren, die ihrerseits durch eine Reihe untergeordneter Illokutionen gestützt wird. Die Illokutionen der einzelnen Sätze, Teiltexte und des gesamten Textes stehen somit in einem hierarchischen Verhältnis zueinander (…)."[387]

Stede weist daraufhin, dass es noch keine Korpora gibt, die mit Illokutionsstruktur annotiert sind. Es fehlen vor allem konkrete Vorschläge für Annotationsschemata.[388] Um „Annotationsrichtlinien für die Gewinnung von intersubjektiv nachvollziehbaren Analysen"[389] entwickeln zu können, muss ein Inventar von Illokutionstypen definiert werden, welche auf die Texte des *DiSynDe*-Korpus angewandt werden können. Dabei ist auch eine Beschränkung auf einzelne Textsorten zu empfehlen, zum Beispiel auf geistliche

---

[384] Stede 2007: 16.
[385] Kessel und Reimann 2005: 202.
[386] Ley 2005: 69.
[387] Ebd.: 69.
[388] Vgl. Stede 2007: 128.
[389] Ebd.: 128.

Prosa oder Fachliteratur. Martin Ley geht von fünf illokutiven Grundtypen aus (Assertion, Direktiv, Expressiv, Kommissiv, Deklaration); von diesen übernimmt er für den Bereich der technischen Kommunikation nur Assertionen und Direktiva, die er auch weiter untergegliedert (beispielsweise werden als vier assertive Untertypen Feststellungen, Beschreibungen, Erklärungen und Indikationen benannt).[390] Zur Annotation von Illokutionsstrukturen empfiehlt Stede das *RSTTool*, denn ein „Software-Werkzeug zur Erstellung von Analysen der Illokutions- oder der Argumentationsstruktur muss die Bildung von größeren Segmenten und deren Hierarchisierung unterstützen und anschaulich darstellen, ggf. auch die Benennung von Relationen zwischen Segmenten."[391]

Die zweite textgrammatische *DiSynDe*-Ebene beschäftigt sich mit der Annotation der referenziellen Textstruktur bzw. der Koreferenz:

> „Eine zentrale Säule der Kohärenz von Texten ist die Koreferenz: sprachliche Ausdrücke verweisen aufeinander (mit Pro-Formen) bzw. auf dieselben Diskursgegenstände. Die referenzielle Struktur bildet dies mit *referenziellen Ketten* ab (…)."[392]

Anhand von Referenzketten eines Textes lässt sich beispielsweise erkennen, „wie ausführlich ein Teilthema erläutert wird."[393] Bei der Annotation von Koreferenz müssen im Text Bezugs- und Verweisausdrücke identifiziert werden: Verweisausdrücke (häufig Pronomina: *sie*) beziehen sich anaphorisch oder kataphorisch auf einen autosemantischen Bezugsausdruck (*eine Katze*).[394] Außerdem werden die Relationen zwischen den Ausdrücken benannt:[395] Das *Potsdam Coreference Scheme*, das für das Deutsche Richtlinien für die Annotation von Koreferenz entwickelt, unterscheidet nominale und nicht-nominale Relationen.[396]

Für die Annotation von referenziellen Strukturen empfiehlt Stede das Software-Werkzeug *MMAX2*: „Es gestattet das komfortable Markieren von referenziellen Ausdrücken (…) und das Verbinden von koreferenten Ausdrücken mit der Maus. Entstehende Relationen können mit einer Bezeichnung versehen werden."[397] Die Annotation

---

[390] Vgl. Ley 2005: 111-113.
[391] Stede 2007: 128.
[392] Ebd.: 181.
[393] Kessel und Reimann 2005: 205.
[394] Vgl. ebd.: 205.
[395] Vgl. Stede 2007: 69.
[396] Vgl. Krasavina und Chiarcos 2007: 5.
[397] Stede 2007: 69.

von referenzieller Struktur ist „besonders nützlich, wenn sie (…) mit weiteren Analyse-Ebenen in Beziehung gesetzt werden kann."[398]

Die dritte Ebene *Textanschluss* behandelt die an der Textoberfläche sichtbare Verknüpfung von Textsegmenten. Hierbei müssen vor allem Fragen hinsichtlich der Segmentierung von Texten in EDUs geklärt werden, denn neben asyndetischen Verknüpfungen können auch Substantive oder Verben Konnektoren sein:[399]

a) *Gestern tobte ein heftiger Sturm. Ein Baum wurde entwurzelt.*
b) *Gestern tobte ein heftiger Sturm. Dadurch wurde ein Baum entwurzelt.*
c) *Der gestrige heftige Sturm verursachte einen entwurzelten Baum.*

Dies führt zu folgender Frage: „Wenn ein Verb oder ein Substantiv die Funktion übernimmt, den inhaltlichen Zusammenhang zwischen zwei Sachverhalten zu übermitteln, wie entscheiden wir dann die Frage, ob es sich um eine oder um zwei EDUs handelt?"[400] Außerdem können sich nicht nur minimale Texteinheiten aufeinander beziehen, „vieles [gilt] auch für den rekursiven Fall der Verbindung zwischen größeren Segmenten."[401]

Für das Deutsche liegt ein Entwurf zur Entwicklung von Richtlinien für die Segmentierung von Radionachrichten und Zeitungskommentaren vor, insbesondere bezüglich ihrer rhetorischen Struktur (Jasinskaja et al. 2006).

Als Korpusprojekte, die sich mit der Markierung von Konnektoren beschäftigen, führt Stede die *Penn Discourse Treebank* und das Projekt *HyTex* an.[402] Außerdem enthält das *Potsdamer Kommentarkorpus* eine entsprechende Annotionsebene, für deren „Erstellung (…) gezielt das halbautomatisch arbeitende Werkzeug *ConAno* implementiert [wurde]."[403]

Die vierte Ebene *Diskurs* behandelt die Annotation der rhetorischen Textstruktur. Diese kann mit Hilfe der „in den vergangenen 20 Jahren stets weiterentwickelte[n] und in verschiedenen computerlinguistischen Anwendungen wie z. B. der Sprachgenerierung oder der Sprachverarbeitung implementierte[n] Rhetorical Structure Theory (RST)"[404] beschrieben werden. Die RST verfolgt das Ziel, rhetorische Relationen zwischen Sätzen und Teiltexten sichtbar zu machen.[405] Das Annotationsschema, das für das DISCOR-Projekt entwickelt wurde, stellt dazu ein Inventar von 14 rhetorischen Beziehungen zur

[398] Stede 2007: 69.
[399] Vgl. ebd.: 166.
[400] Ebd.: 166f.
[401] Ebd.: 165.
[402] Vgl. ebd.: 179.
[403] Ebd.: 179.
[404] Ley 2005: 70.
[405] Vgl. ebd.: 71.

Verfügung *(Continuation, Narration, Result, Contrast, Parallel, Precondition, Consequence, Alternation, Background, Elaboration, Explanation, Commentary, Source, Attribution)*.[406] Durch die Bereitstellung sehr unterschiedlicher Kohärenzrelationen erhebt die rhetorische Struktur „eine Art ‚Alleinvertretungsanspruch' (…). Man kann (…) entweder den illokutiven Zusammenhang herausstellen, oder den inhaltlich-semantischen, oder den rein thematischen (mit einer Relation wie *Elaboration*)."[407]

Die Beziehung *Elaboration*, die auch in den *DiSynDe*-Annotationsvorschriften benutzt wird, bedeutet folgendes:

> „*Elaboration (α, β)* holds when *β* provides further information about the eventuality introduced in *α*; for example, if the main eventuality of *β* is a subtype or part of the eventuality mentioned in *α*. *Elaboration* implies that a relation of temporal inclusion holds between the related eventualities. Discourse markers like *for instance, for example,* or the explicit listing of sub-events (*first, second,* etc.), are good cues for *Elaboration*."[408]

Eine RST-Analyse ergibt eine hierarchische Baumstruktur. Dazu wird der Text in minimale Texteinheiten aufgeteilt, welche die Knoten des Baumes darstellen. In einem weiteren Schritt werden sie zu komplexeren Textsegmenten zusammengefasst, die wiederum in die Baumstruktur eingefügt werden. Meistens nehmen manche Sätze oder Teiltexte eine hervorgehobene Position ein; diese zentralen Elemente der rhetorischen Textstruktur werden als *Nuklei* bezeichnet, von denen die weniger bedeutenden *Satelliten* abhängen.[409]

Bei der Erstellung von RST-Baumstrukturen wird deutlich, dass „ein hohes Maß an Interpretation notwendig ist und die Ermittlung der rhetorischen Struktur eines Textes somit zu einem großen Teil auf der Urteilskraft und dem subjektiven Empfinden der Analysierenden beruht."[410] Dies betrifft sowohl die Festlegung, welche Textsegmente Nuklei und welche Satelliten sind, als auch die Bestimmung der Relation zwischen den Segmenten. Das subjektive Empfinden nimmt vor allem dann starken Einfluss auf die Annotation, wenn Oberflächenmerkmale wie Konnektoren im Text nicht vorhanden sind.[411]

Für das Deutsche existieren nach RST annotierte Texte im *Potsdamer Kommentarkorpus*.[412] Das Software-Werkzeug *RSTTool* ist speziell für die Annotation von RST-

---

[406] Vgl. Reese et al. 2007: 6.
[407] Stede 2007: 183.
[408] Reese et al. 2007: 7.
[409] Vgl. Ley 2005: 71.
[410] Ebd.: 71.
[411] Ebd.: 71f.
[412] Vgl. Stede 2007: 151.

Strukturen entwickelt. „Die Menge der Kohärenzrelationen ist [dabei] nicht durch RSTTool festgelegt, sondern wird in einer separaten Datei spezifiziert (…).“[413] Für *DiSynDe* muss also in erster Linie ein Inventar von Relationen zusammengestellt werden. Da sich mit RST nicht nur rhetorische, sondern zum Beispiel auch illokutive Zusammenhänge annotieren lassen, kann durch eine geeignete Auswahl der Relationen eventuell die Anzahl der Annotationsschichten für die textgrammatische Analysegruppe noch reduziert werden.

[413] Stede 2007: 150.

6. Schluss

Für die syntaktische Annotation des *DiSynDe*-Korpus sind verschiedene Entscheidungen zu treffen. Als erstes müssen geeignete Annotationsschemata ausgewählt werden. Die morphosyntaktische Ebene lässt sich am besten mit einem für historische Texte modifizierten STTS-Tagset beschreiben. Für die syntaktische Ebene eignet sich das für *DiSynDe* erweiterte TIGER-Annotationsschema (siehe Anhang 3). Wenn bei der Annotation des Pilotkorpus bestimmte linguistische Phänomene nicht eindeutig beschrieben werden können, müssen die Schemata um entsprechende Tags erweitert werden.

Die zweite Entscheidung betrifft die Kodierung der Annotation. XCES bietet hierfür zahlreiche Vorteile. Mit diesem Standard können Metadaten, morphosyntaktische und syntaktische Informationen in einer Stand-Off-Annotation gespeichert werden. Das TIGER-XML-Format ist vor allem für die syntaktische Annotation geeignet; ein Nachteil ist, dass die morphosyntaktischen und syntaktischen Informationen nicht auf verschiedene Dateien aufgeteilt werden können. Allerdings stellt das TIGER-XML-Format eine wichtige Grundlage für den sich in Entwicklung befindlichen ISO-Standard SynAF dar. Man muss sich dabei nicht auf ein Kodierungsformat beschränken, sondern es können auch zur besseren Wiederverwendbarkeit des Korpus Konvertierungen in verschiedene Formate umgesetzt werden.

Für die Annotation der verschiedenen textgrammatischen Strukturen stehen keine standardisierten Schemata und Formate zur Verfügung, allerdings existieren Projekte wie DISCOR oder das Potsdamer Kommentarkorpus, an denen man sich orientieren kann. Für *DiSynDe* müssen Richtlinien für die Segmentierung der Texte in minimale Texteinheiten aufgestellt und ein Inventar von Relationen für die Annotation von rhetorischen Strukturen (RST) definiert werden.

Für die textgrammatischen Annotationsebenen existieren desweiteren spezielle Software-Werkzeuge, welche die Annotationen in eigenen einfachen XML-Formaten speichern. Erst durch die Möglichkeiten der Stand-Off-Annotation ergibt sich der große Vorteil, „dass man für die verschiedenen Analyse-Ebenen jeweils spezielle Werkzeuge benutzen kann, die auf die zugrunde liegenden Strukturen zugeschnitten sind und damit ein möglichst einfaches und effektives Arbeiten erlauben.“[414]

Um letztlich bei der linguistischen Analyse der unterschiedlichen *DiSynDe*-Annotationsebenen auch Korrelationen zwischen den Ebenen aufdecken zu können, müssen die Daten zusammengeführt werden. Stede weist hierfür auf das *Potsdamer Austauschformat für linguistische Annotation* (PAULA) und die linguistische Datenbank

[414] Stede 2007: 16.

ANNIS hin. Mit Konvertierungsskripten können Annotationen, die mit *MMAX*, *RSTTool*, *Exmaralda* oder *annotate* erstellt wurden, in das Austauschformat PAULA umgewandelt werden, das „ein sehr allgemeines, ‚generisches' XML-Format"[415] darstellt. Auch das TIGER-XML-Format kann zu PAULA konvertiert werden. Die Software ANNIS dient schließlich dazu, sämtliche Annotationsebenen mit einem Werkzeug durchsuchen zu können:[416]

> „Ein wichtiges Merkmal von ANNIS ist nun, dass (…) [man] Suchanfragen formulieren kann, die mehrere Annotationsebenen miteinander verbinden. Angenommen, zu den gespeicherten Texten liegen Annotationen zur Syntax, zum Informationsstatus der Diskursgegenstände und zur rhetorischen Struktur vor, so ist es beispielsweise möglich, alle Textstellen zu finden, in denen (i) eine Präpositionalphrase am Satzanfang steht, (ii) der in der eingebetteten NP denotierte Diskursgegenstand *brand-new* ist, und (iii) die PP als Satellit der Kohärenzrelation *Concession* verwendet wird. (Ein entsprechender Satz könnte lauten: *Trotz einer Einladung zum DFB-Pokalfinale hatte Leonie heute sehr schlechte Laune.*)"[417]

Dahingehend lassen sich auch die potentiellen Suchanfragen an das annotierte *DiSynDe*-Korpus erweitern, die noch keine Bezüge zur textgrammatischen Annotationsebene beinhalten (vgl. *2.4 Potentielle Suchanfragen an das annotierte Korpus*, Seite 8). So können die morphosyntaktischen, syntaktischen und textgrammatischen Annotationsschichten des *DiSynDe*-Korpus nach der Bearbeitung durch die Analysegruppen in einer Datenbank zusammengeführt und Suchanfragen formuliert werden, welche die verschiedenen Annotationsebenen umfassen.[418]

[415] Stede 2007: 202.
[416] Vgl. ebd.: 203.
[417] Ebd.: 203.
[418] Vgl. ebd.: 201-203.

Abbildungsverzeichnis

Literaturverzeichnis

- **Albert, Stefanie; Anderssen, Jan; Bader, Regine [u.a.] (2003).** *TIGER Annotationsschema.* <http://www.ifi.uzh.ch/CL/volk/treebank_course/tiger_annot.pdf> (April 2008).
- **Brinker, Klaus (2005).** *Linguistische Textanalyse.* Eine Einführung in die Grundbegriffe und Methoden. 6., überarbeitete und erweiterte Auflage. Berlin: Erich Schmidt Verlag.
- **Bußmann, Hadumond (Hrsg.) (2002).** *Lexikon der Sprachwissenschaft.* 3., aktualisierte und erweiterte Auflage. Stuttgart: Alfred Kröner Verlag.
- **Dal, Ingerid (1966).** *Kurze deutsche Syntax auf historischer Grundlage.* 3., verbesserte Auflage. Tübingen: Niemeyer Verlag.
- **Declerck, Thierry (2006).** *SynAF: Towards a Standard for Syntactic Annotation.* <http://lirics.loria.fr/doc_pub/SynAF_LREC2006.pdf> (Februar 2008).
- **Dipper, Stefanie (2007).** *POS-Tagging.* CL-Einführung. Präsentation. <http://www.ling.uni-potsdam.de/~dipper/teaching/ss07_cl/slides/posTagging.pdf> (April 2008).
- **CES (1996a).** *CES Part 3. Corpus Encoding Standard – Document CES 1 Part 3. The CES Header.* <http://www.cs.vassar.edu/CES/CES1-3.html> (April 2008).
- **CES (1996b).** *header.elt. The CES Header.* <http://www.cs.vassar.edu/CES/sgml/header.elt> (April 2008).
- **CES (2003).** *XCES Schemas.* <http://www.xces.org/schema/2003/> (April 2008)
- **EAGLES (1996a).** *Recommendations for the Syntactic Annotation of Corpora.* <http://www.ilc.cnr.it/EAGLES96/segsasg1/segsasg1.html> (April 2008).
- **EAGLES (1996b).** *Information about the rank of a syntactic unit: layer (g).* <http://www.ilc.cnr.it/EAGLES96/segsasg1/node24.html> (April 2008).
- **EAGLES (1996c).** *Coordination.* <http://www.ilc.cnr.it/EAGLES96/segsasg1/node37.html> (April 2008).
- **EAGLES (1996d).** *Semantic annotation.* <http://www.ilc.cnr.it/EAGLES96/segsasg1/node40.html> (April 2008).
- **EAGLES (1996e).** *Annotation of deep/logical information.* <http://www.ilc.cnr.it/EAGLES96/segsasg1/node42.html> (April 2008).
- **EAGLES (2003).** *The essentials of EAGLES.* <http://www.ilc.cnr.it/EAGLES/intro.html> (April 2008).
- **Ebert, Robert Peter (1978).** *Historische Syntax des Deutschen.* Stuttgart: Metzler Verlag.
- **Engel, Ulrich (2004).** *Deutsche Grammatik. Neubearbeitung.* München: Iudicium Verlag.
- **Eroms, Hans-Werner (2000).** *Syntax der deutschen Sprache.* Berlin [u.a.]: Walter de Gruyter.
- **Flohr, Horst; Pfingsten, Friederike (2002).** *Die Struktur von Wörtern: Morphologie.* In: Müller, Horst M. (Hrsg.) (2002). Arbeitsbuch Linguistik. Paderborn [u.a.]: Verlag Ferdinand Schöningh. 102-124.
- **Hennings, Thordis (2003).** *Einführung in das Mittelhochdeutsche.* 2. Auflage. Berlin [u.a.]: Walter de Gruyter.

- **Heringer, Hans-Jürgen (1996).** *Deutsche Syntax dependentiell.* Tübingen: Stauffenburg Verlag.
- **Ide, Nancy; Romary, Laurent (2003).** *Encoding Syntactic Annotation.* In: Abeillé, Anne (Hrsg.) (2003). Treebanks. Building and Using Parsed Corpora. Dordrecht [u.a.]: Kluwer Academic Publishers. 281-296.
- **ISO (2008).** *ISO - International Organization for Standardization.* <http://www.iso.org/iso/home.htm> (April 2008).
- **ISO/TC 37/SC 4 N421 (2007).** *Language resource management – Syntactic Annotation Framework (SynAF).* <http:// lirics.loria.fr/doc_pub/N421_SynAF_CD_ISO_24615.pdf> (Februar 2008).
- **Jasinskaja, Katja; Mayer, Jörg; Boethke, Jutta; Neumann, Annika; Peldszus, Andreas; Rodriguez, Kepa Joseba (2006).** *Discourse Tagging Guidelines for German Radio News and Newspaper Commentaries.* <http://www.ling.uni-potsdam.de/~stede/Kursmaterial/KorpTA/GermanDiscTagging.pdf> (Mai 2008).
- **Kahrel, Peter; Barnett, Ruthanna; Leech, Geoffrey (1997).** *Towards Cross-Linguistic Standards or Guidelines for the Annotation of Corpora.* In: Garside, Roger; Leech, Geoffrey; McEnery, Tony (Hrsg.) (1997). Corpus Annotation. Linguistic Information from Computer Text Corpora. London [u.a.]: Longman Addison-Wesley. 231-242.
- **Kallmeyer, Laura; Wagner, Andreas (2000).** *Der TUSNELDA-Annotationsstandard.* Kolloquium des SFB 441. <http://www.sfb441.uni-tuebingen.de/c1/handout-8-5-2000.ps> (Januar 2008).
- **Kermanidis, Katia Lida; Fakotakis, Nikos; Kokkinakis, George (2004).** *Automatic acquisition of verb subcategorization information by exploiting minimal linguistic resources.* In: International Journal of Corpus Linguistics. Volume 9, Number 1, 2004. Amsterdam [u.a.]: John Benjamins Publishing Company.
- **Kessel, Katja; Reimann, Sandra (2005).** *Basiswissen Deutsche Gegenwartssprache.* Tübingen [u.a.]: Narr Francke Attempto Verlag.
- **Krasavina, Olga; Chiarcos, Christian (2007).** *PoCoS – Potsdam Coreference Scheme.* <http://www.ling.uni-potsdam.de/~stede/Kursmaterial/KorpTA/pocos07.pdf> (Mai 2008).
- **Kroymann, Emil; Thiebes, Sebastian; Lüdeling, Anke; Leser, Ulf (2004).** *Eine vergleichende Analyse von historischen und diachronen digitalen Korpora.* Technical Report 174 des Instituts für Informatik der Humboldt-Universität zu Berlin. <http://www.deutschdiachrondigital.de/publikationen/TRHistorischeKorpora.pdf> (Januar 2008).
- **Langer, Hagen (2004).** *Syntax und Parsing.* In: Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Ralf; Langer, Hagen (Hrsg.) (2004). Computerlinguistik und Sprachtechnologie. Eine Einführung. 2., überarbeitete und erweiterte Auflage. München: Elsevier. 232-275.
- **Leech, Geoffrey (1997).** *Introducing Corpus Annotation.* In: Garside, Roger; Leech, Geoffrey; McEnery, Tony (Hrsg.) (1997). Corpus Annotation. Linguistic Information from Computer Text Corpora. London [u.a.]: Longman Addison-Wesley. 1-18.
- **Lemnitzer, Lothar; Zinsmeister, Heike (2006).** *Korpuslinguistik. Eine Einführung.* Tübingen: Narr Francke Attempto Verlag.

- **Ley, Martin (2005).** *Kontrollierte Textstrukturen. Ein (linguistisches) Informationsmodell für die Technische Kommunikation.* <http://geb.uni-giessen.de/geb/volltexte/2006/2713/pdf/LeyMartin-2006-01-30.pdf> (Mai 2008).
- **Lezius, Wolfgang (2001).** *Baumbanken.* In: Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Ralf; Langer, Hagen (Hrsg.) (2004). Computerlinguistik und Sprachtechnologie. Eine Einführung. 2., überarbeitete und erweiterte Auflage. München: Elsevier. 414-422.
- **Lezius, Wolfgang (2002).** *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora.* In: Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), 2002, vol. 8, no. 4. <http://www.ims.uni-stuttgart.de/projekte/corplex/paper/lezius/diss/disslezius.pdf> (Februar 2008).
- **LIRICS (2005).** *Linguistic Infrastructure for Interoperable Resources and Systems.* <http://lirics.loria.fr/> (Februar 2008).
- **Lobin, Henning (2001).** *Informationsmodellierung in XML und SGML.* Korrigierter Nachdruck. Berlin [u.a.]: Springer Verlag.
- **Lüdeling, Anke; Poschenrieder, Thorwald; Faulstich, Lukas (2004).** *DeutschDiachronDigital – Ein diachrones Korpus des Deutschen.* In: Jahrbuch für Computerphilologie 2004. <http://www.deutschdiachrondigital.de/publikationen/ddd-computerphilologie.pdf> (Januar 2008).
- **McEnery, Tony; Wilson, Andrew (2003).** *Corpus Linguistics. An Introduction.* 2. Edition. Edinburgh: Edinburgh University Press.
- **Meineke, Eckhard; Schwerdt, Judith (2001).** *Einführung in das Althochdeutsche.* Paderborn [u.a.]: Verlag Ferdinand Schöningh.
- **Negra (1998a).** *Kanten.* <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/> (April 2008).
- **Negra (1998b).** *Knoten.* <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/knoten.html> (April 2008).
- **Negra (2005).** *NEGRA Corpus.* <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/> (April 2008).
- **Reese, Brian; Denis, Pascal, Asher, Nicholas; Baldridge, Jason; Hunter, Julie (2007).** *Reference Manual for the Analysis and Annotation of Rhetorical Structure (Version 1.0).* Technical Report. <http://comp.ling.utexas.edu/discor/manual.pdf> (Mai 2008).
- **Sasaki, Felix; Witt, Andreas (2004).** *Linguistische Korpora.* In: Lobin, Henning; Lemnitzer, Lothar (Hrsg.) (2004). Texttechnologie. Perspektiven und Anwendungen. Tübingen: Stauffenburg Verlag. 195-216.
- **Scherer, Carmen (2006).** *Korpuslinguistik.* Heidelberg: Universitätsverlag Winter.
- **Schiller, Anne; Teufel, Simone; Stöckert, Christine (1999).** *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).* <http://www.ifi.unizh.ch/cl/siclemat/man/SchillerTeufel99STTS.pdf> (Februar 2008).
- **Schmid, Hans Ulrich (2007).** *Das Projekt einer Historischen Syntax des Deutschen*. In: Bochmann, Klaus (Hrsg.) (2007). Theorie(n) und Methoden der Sprachgeschichte. Stuttgart [u.a.]: S. Hirzel Verlag. 51-57.

- **Schmidt, Wilhelm (2007).** *Geschichte der deutschen Sprache.* 10., verbesserte und erweiterte Auflage. Stuttgart: S. Hirzel Verlag.
- **Stede, Manfred (2007).** *Korpusgestützte Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik.* Tübingen: Gunter Narr Verlag.
- **Steiner, Petra (2004).** *Wortarten und Korpus. Automatische Wortartenklassifikation durch distributionelle und quantitative Verfahren.* Aachen: Shaker Verlag.
- **STTS (2001).** *STTS Tag Table.* <http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html> (März 2008).
- **TEI (2008a).** *TEI: Text Encoding Initiative.* <http://www.tei-c.org/index.xml> (April 2008).
- **TEI (2008b).** *TEI Guidelines TOC.* <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html> (April 2008).
- **TEI (2008c).** *17.1 Linguistic Segment* Categories. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AI.html#AILC> (April 2008).
- **TIGER (2003).** *TigerXML.xsd.* <http://www.ims.uni-stuttgart.de/projekte/TIGER/public/TigerXML.xsd> (April 2008).
- **TIGER (2007).** *The TIGER Project.* <http://www.ims.uni-stuttgart.de/projekte/TIGER/> (April 2008).
- **TIGERSearch (2003).** *TIGERSearch - tools for linguistic text exploration.* <http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/> (April 2008).
- **Trippel, Thorsten; Declerck, Thierry; Heid, Ulrich (2005).** *Sprachressourcen in der Standardisierung.* In: LDV-Forum 2005, Band 20 (2). <http://www.ldv-forum.org/2005_Heft2/Thorsten_Trippel_and_Thierry_Declerck_and_Ulrich_Heid.pdf> (Januar 2008).
- **Tylman, Ule; Hinrichs, Erhard (2004).** *Linguistische Annotation.* In: Lobin, Henning; Lemnitzer, Lothar (Hrsg.) (2004). Texttechnologie. Perspektiven und Anwendungen. Tübingen: Stauffenburg Verlag. 217-243.
- **Volmert, Johannes (2001).** *Grundkurs Sprachwissenschaft.* 4. Auflage. München: Wilhelm Fink Verlag.
- **Wells, Christopher J. (1990).** *Deutsch: eine Sprachgeschichte bis 1945.* Tübingen: Niemeyer Verlag.
- **Wirrer, Jan (2002).** *Historisch-Vergleichende Sprachwissenschaft: Der Wandel von Sprache.* In: Müller, Horst M. (Hrsg.) (2002). Arbeitsbuch Linguistik. Paderborn [u.a.]: Verlag Ferdinand Schöningh. 241-262.
- **Witt, Andreas (2002).** *Multiple Informationsstrukturierung mit Auszeichnungssprachen. XML-basierte Methoden und deren Nutzung für die Sprachtechnologie.* <http://xml.coverpages.org/WittHabil200207.pdf> (Januar 2008).
- **Wolff, Gerhart (2004).** *Deutsche Sprachgeschichte von den Anfängen bis zur Gegenwart.* 5., überarbeitete und aktualisierte Auflage. Tübingen [u.a.]: Narr Francke Attempto Verlag.

Anhang 1: *DiSynDe*-Annotationsvorschriften

Verwendete Abkürzungen:

adv	Adverb (ggf. mit Index $_{i, j}$ usw.)
adverbial	Adverbiale
advp	Adverbialphrase
akk	Akkusativ
anaph	anaphorisch
art.def	bestimmter Artikel
art.indef	unbestimmter Artikel
attr	Attribut
det	Determinator (Artikel, Pronomen)
erst	Verberststellung
expl	explizit
fem	Femininum
fragew	Fragewort
imp	Imperativ
ind	Indikativ
indiff	indifferent (z.B. bzgl. Modus)
indir	indirekt
inf	Infinitiv
instr	Instrumental
kataph	kataphorisch
knj	Konjunktiv
kompar	Komparativ
konju	Konjunktion
koord	koordinierend
korr	Korrelat (Index $_{i, j}$ usw. verweist auf das korrespondierende Element)
letzt	Verbletztstellung (Zusatzziffer: an wievielter numerischer Stelle)
lok	lokal
mod	modal
msk	Maskulinum
nach	nachgestellt
neg	Nagator
nom	Nominativ
np	Nominalphrase
ntr	Neutrum
num.kard	Kardinalzahl
num.ord	Ordinalzahl
obj.akk	Akkusativobjekt
obj.dat	Dativobjekt
obj.präp	Präpositionalobjekt
p.pers	Personalpronomen (Zusatzziffer: Person)
p.rel	Relativpronomen
part	Partikel
pp	Präpositionalphrase
präd	Prädikat

prädcompl	Teil eines Prädikatskomplexes (mit Index $_{i, j}$ usw.)
prädsimpl	einfaches Prädikat
präp	Präposition
pronom	pronominal
prs	Präsens
prt	Präteritum
pt1	Partizip I = Partizip des Präsens
pt2	Partizip II = Partizip des Präteritums
rekurr	rekurrent
sb	Substantiv
sb.prop	Substantiv – Sonderfall Eigenname
sg	Singular
subj	Subjekt
subju	Subjunktion
sw	schwach flektiert
unfl	unflektiert
vaux	Hilfsverb
vb	Verb
vfin	finites Verb
vinf	infinites Verb
vk	Verbalkomplex
vmod	Modalverb
vor	vorangestellt
vp	Verbalphrase (ggf. mit Index $_{i, j}$ usw.)
vvoll	Vollverb
zweit	Verbzweitstellung

Text	textelement	zitat/direkte rede / appell	
	diskurs		
	referenz		
	textanschluss		
kpl.S.	makrostrukt	matrix	
	einleit		
	position		
	verbstellung	erst	
	modus	imp	
einf. Satz		prädsimpl	obj.akk
wortgruppe	phrase grob	vp	np
	phrase fein1		
	phrase fein2		
	vk	voll	
	flexion	2sg.imp	akk.sg.
wortart		vb	p.pers1
Quelle		*Hor*	*mich*

Text	textelement	zitat/direkte rede										
	diskurs	konsequenz										
	referenz	anaph										
	textanschluss	adv										
komplexer Satz	makrostrukt	matrix					indir. fragesatz					
	einleit						präp + fragew					
	position	vor					nach					
	verbstellung	zweit										letzt.4
	modus	ind										indiff
einf. Satz	satzglied	adverbialmod	prädcompl$_i$	subj	obj.dat	präd. compl$_i$	obj.akk					
							adverbial.instr		subj	obj.akk		präd
Wortgruppe	phrase grob	advp	vp$_i$	np	np	vp$_i$	pp		np	np		vp
	phrase fein1		vfin			vinf						vfin
	phrase fein2									det	kopf	
	vk		vmod			vvoll						vvoll
	flexion		2.sg.ind.präs	nom.sg	dat.sg	inf		instr.sg	sg.nom	akk.sg.msk	akk.sg	2sg.prs
Wortart		adv.mod	vb	p.pers1	p.pers2	vb	präp	p.rel	p.pers2	art.def	sb.	vb
Quelle		*So*	*Wil*	*ich*	*dir*	*weisen,*	*mit*	*we*	*du*	*Den*	*teuffel*	*überwindest*

Text	textelement	Quellennachweis										
	diskurs	Elaboration										
	referenz	anaph										
	textanschluss	adv										
kpl.S.	makrostrukt	matrix										
	einleit											
	position											
	verbstellung	zweit										
	modus	ind										
einf.Satz	satzglied	adverbial mod	prädcompl$_i$	prädcompl$_i$	adverbial lok				adverbial lok			
wortgruppe	phrase grob	advp	vp$_i$	vp$_i$	pp				pp			
	phrase fein1		vfin	vinfin		np				np		
	phrase fein2					det.	kopf	attr		det	attr	kopf
	vk		vvoll	vvoll								
	flexion		3.sg.ind.präs	pt2		dat.sg.ntr	dat.sg	gen.sg		dat.sg.ntr	dat.sg	dat.sg
wortart		adv.mod	vb	vb	präp	art.def	sb.	sb.prop	präp	art.def	num.ord	sb.
Quelle		*Also*	*stet*	*geschriben*	*In*	*dem*	*puch*	*Tobie*	*an*	*Dem*	*6.*	*capitel*

Text	textelement	kommentar															
	diskurs	elaboration															
	reference	anaph															
	textanschluss	pronom															
kpl.S.	makrostruk	matrix															
	einleit																
	position																
	verbstellung	zweit															
	modus	ind															
einf.S		subj.		prädcompl$_i$	adverbial mod				prädcompl$_i$	prädcompl$_i$		obj.präp					
wortgruppe	phrase grob	np			adv.p							pp					
	phrase fein1											np					
	phrase fein2	det	kopf		adv$_i$	koord		adv$_i$				det	attr	titel	kopf		
	vk			vaux					vvoll	vaux							
	flexion	nom.pl.ntr	nom.pl	3pl.ind.prs					pt2	pt2		akk.sg.msk	akk.sg.msk				
wortart		p.dem	sb	vb	adv	konju	neg	adv	vb	vb	part	präp	art.def	adj	sb	adj	sb.prop
Quelle		*Diese*	*wort*	*sein*	*göttlichen*	*vnd*	*Nit*	*menschlichen*	*geret*	*worden*	*wann*	*durch*	*Den*	*namhaften*	*furstengel*	*sand*	*Raphael*

Text	textelement	kommentar (weiter)																								
	diskurs	continuation															result									
	referenz		kataph		anaph												anaph									
	textanschluss	konju	subju		pronom																					
kompl.Satz	makrostr	Kausalsatz															matrix									
	einleit	subju.kompl/korr$_i$															korr$_i$									
	position	vor															nach									
	vbstell											letzt 3					zweit									
	modus											indiff					ind									
einf.S					subj		obj.präp					prädcompl.$_i$	prädcompl.$_i$	prädcompl.$_i$		prädcompl.$_j$	subj	adverbial.mod		obj.präp		prädcompl.$_j$	prädcompl.$_j$			
wortgr.	phrase grob				np		pp										np	advp		pp						
	phrase fein1						np																			
	phrase fein2						det	attr			kopf															
	vk											vaux	vvoll	vaux		vmod						vvoll	vaux			
	flexion				pl.nom		akk.sg.msk		akk.sg.msk..sw		akk.sg	3pl.ind.präs	pt2	pt2		3.pl.prs.ind	nom.pl		kompar		dat.pl	pt2	inf			
wortart		konju	adv	subju	ppers3	präp	art.indef	advmod	adj		sb.	vb	vb	vb	adv	vb	ppers3	adv	adjadv	präp	ppers1	vb	vb			
Quelle		*vnd*	*darvmb*	*das*	*sy*	*durch*	*ainen*	*so*	*wirdigen*		*engel*	*sein*	*Gespr*	*worden*	*darumb*	*syllen*	*sy*	*dester*	*vleissiger*	*von*	*vns*	*gemerckt*	*werden*			

Text	textelement	kommentar (weiter)																		siehe Fortsetzung
	diskurs	cotinuation																		
	referenz				kataph				anaph		anaph									
	textanschluss	konju							expl. Rückverw.	rekurr		rekurr								
kompl.Satz	makrostrukt	matrix					inhaltssatz													
	Einleit						subju													
	Position	vor					nach													
	verbstellung	erst					dritt													
	Modus	ind					ind													
einf.S			$prädcompl_i$			$prädcompl_i$		subj												
								subj			obj.akk		$prädcompl._j$	$prädcompl._j$	obj.präp					
wortgruppe	phrase grob		vp_i			vp_i		np			np		vp_j	vp_j	pp					
	phrase fein1															np				
	phrase fein2							det	attr	kopf	det	kopf				det	attr	kopf		
	vk		vaux			vvoll							vvoll	vaux						
	flexion		3sg.ind.prs			inf *zu*		nom.sg.msk	nom.sg.msk	nom.sg	akk.pl.ntr	akk.pl	$pt2_i$	3sg.ind.ps_i.perf		dat.sg.msk	kompar. dat.sg.mask. sw	unfl		
wortart		konju	vb	part	adv	vb	subju	art.def	adj.pt2	sb.	art.def	sb.	vb	vb	präp	art.def	adj	sb.prop		
Quelle		*Vnd*	*ist*	*auch*	*dauon*	*zewissen*	*das*	*der*	*genant*	*engel*	*dy*	*wort*	*gesprochen*	*hat*	*zu*	*dem*	*jungeren*	*Thobias*		

Fortsetzung	Text	textelement							
		diskurs	elaboration						
		referenz	anaph					rekurrent antonym	
		textanschluss	pronom						
	kompl.Satz	makrostrukt	attributsatz						
		einleit	rel						
		position	nach						
		verbstellung	dritt						
		modus	ind						
	einf. Satz	subj (Fortsetzung)							
		obj.präp (Fortsetzung)							
			subj	$präd.nom_i$		präd	$präd.nom_i$		
	wortgruppe	phrase grob	np	np_i		vp	np_i		
		phrase fein1		$kopf_i$			attr.gen_i		
		phrase fein2		det	kopf		det	attr	kopf
		vk							
		flexion	nom.sg.msk	nom.sg.msk	nom.sg	3sg.ind.prät	gen.sg.msk	kompar.gen.sg.msk.sw	unflekt
	wortart		p.rel	art.indef	sb	vb	art.def	adj	sb.prop
	Quelle		*der*	*ein*	*sun*	*was*	*des*	*eltern*	*Thobias*

Anhang 2: STTS-Tagset

POS	DESCRIPTION	EXAMPLES
ADJA	attributives Adjektiv	[das] große [Haus]
ADJD	adverbiales oder prädikatives Adjektiv	[er fährt] schnell, [er ist] schnell
ADV	Adverb	schon, bald, doch
APPR	Präposition; Zirkumposition links	in [der Stadt], ohne [mich]
APPRART	Präposition mit Artikel	im [Haus], zur [Sache]
APPO	Postposition	[ihm] zufolge, [der Sache] wegen
APZR	Zirkumposition rechts	[von jetzt] an
ART	bestimmter oder unbestimmter Artikel	der, die, das, ein, eine
CARD	Kardinalzahl	zwei [Männer], [im Jahre] 1994
FM	Fremdsprachliches Material	[Er hat das mit ``] A big fish [" übersetzt]
ITJ	Interjektion	mhm, ach, tja
KOUI	unterordnende Konjunktion mit ``zu'' und Infinitiv	um [zu leben], anstatt [zu fragen]
KOUS	unterordnende Konjunktion mit Satz	weil, daß, damit, wenn, ob
KON	nebenordnende Konjunktion	und, oder, aber
KOKOM	Vergleichskonjunktion	als, wie
NN	normales Nomen	Tisch, Herr, [das] Reisen
NE	Eigennamen	Hans, Hamburg, HSV
PDS	substituierendes Demonstrativpronomen	dieser, jener
PDAT	attribuierendes Demonstrativpronomen	jener [Mensch]
PIS	substituierendes Indefinitpronomen	keiner, viele, man, niemand
PIAT	attribuierendes Indefinitpronomen ohne Determiner	kein [Mensch], irgendein [Glas]
PIDAT	attribuierendes Indefinitpronomen mit Determiner	[ein] wenig [Wasser], [die] beiden [Brüder]
PPER	irreflexives Personalpronomen	ich, er, ihm, mich, dir

PPOSS	substituierendes Possessivpronomen	meins, deiner
PPOSAT	attribuierendes Possessivpronomen	mein [Buch], deine [Mutter]
PRELS	substituierendes Relativpronomen	[der Hund ,] der
PRELAT	attribuierendes Relativpronomen	[der Mann ,] dessen [Hund]
PRF	reflexives Personalpronomen	sich, einander, dich, mir
PWS	substituierendes Interrogativpronomen	wer, was
PWAT	attribuierendes Interrogativpronomen	welche [Farbe], wessen [Hut]
PWAV	adverbiales Interrogativ- oder Relativpronomen	warum, wo, wann, worüber, wobei
PAV	Pronominaladverb	dafür, dabei, deswegen, trotzdem
PTKZU	``zu'' vor Infinitiv	zu [gehen]
PTKNEG	Negationspartikel	nicht
PTKVZ	abgetrennter Verbzusatz	[er kommt] an, [er fährt] rad
PTKANT	Antwortpartikel	ja, nein, danke, bitte
PTKA	Partikel bei Adjektiv oder Adverb	am [schönsten], zu [schnell]
TRUNC	Kompositions-Erstglied	An- [und Abreise]
VVFIN	finites Verb, voll	[du] gehst, [wir] kommen [an]
VVIMP	Imperativ, voll	komm [!]
VVINF	Infinitiv, voll	gehen, ankommen
VVIZU	Infinitiv mit ``zu'', voll	anzukommen, loszulassen
VVPP	Partizip Perfekt, voll	gegangen, angekommen
VAFIN	finites Verb, aux	[du] bist, [wir] werden
VAIMP	Imperativ, aux	sei [ruhig !]
VAINF	Infinitiv, aux	werden, sein
VAPP	Partizip Perfekt, aux	gewesen
VMFIN	finites Verb, modal	dürfen
VMINF	Infinitiv, modal	wollen
VMPP	Partizip Perfekt, modal	gekonnt, [er hat gehen] können
XY	Nichtwort, Sonderzeichen enthaltend	3:7, H2O, D2XW3
\$,	Komma	,
\$.	Satzbeendende Interpunktion	. ? ! ; :
\$(	sonstige Satzzeichen; satzintern	- [,]()

Anhang 3: Syntaktische Tagsets für *DiSynDe*

Dependenzstruktur:

AC adpositional case marker	MR rhetorical modifier
ADC adjective component	MS (konsekutive Modifikation)
AMS measure argument of adj	MT (temporale Modifikation)
APP apposition	MW way (directional modifier)
AVC adverbial phrase component	MZ (konzessive Modifikation)
CC comparative complement	NG negation
CD coordinating conjunction	NK noun kernel modifier
CJ conjunct	NMC numerical component
CM comparative concjunction	OA accusative object
CP complementizer	OA2 second accusative object
DA dative	OC clausal object
DAF (freier Dativ)	OG genitive object
DH discourse-level head	OP object prepositional
DM discourse marker	PAR parenthesis
EP expletive *es*	PD predicate
GL prenominal genitive	PG pseudo-genitive
GR postnominal genitive	PH placeholder
HD head	PM morphological particle
JU junctor	PNC proper noun component
MA (adversative Modifikation)	RC relative clause
MC comitative	RE repeated element
MF (finale Modifikation)	RS reported speech
MI instrumental	SB subject
MK (kausale Modifikation)	SBP passivised subject (PP)
ML locative	SP subject or predicate
MN (konditionale Modifikation)	SVP separable verb prefix
MNR postnominal modifier	UC (idiosyncratic) unit component
MO modifier	VO vocative

Konstituentenstruktur:

AA superlative phrase with "am"	CVZ coordinated zu-marked infinitive
AP adjektive phrase	DL discourse level constituent
AVP adverbial phrase	ISU idiosyncratis unit
CAC coordinated adposition	MPN multi-word proper noun
CAP coordinated adjektive phrase	MTA multi-token adjective
CAVP coordinated adverbial phrase	NM multi-token number
CCP coordinated complementizer	NP noun phrase
CH chunk	PP adpositional phrase
CNP coordinated noun phrase	QL quasi-language
CO coordination	S sentence
CPP coordinated adpositional phrase	VP verb phrase (non-finite)
CS coordinated sentence	VZ zu-marked infinitive
CVP coordinated verb phrase (non-finite)	

Eidesstattliche Erklärung

Ich versichere an Eides statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbstständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe. Die Arbeit hat in dieser oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Regensburg, 01. Juni 2008 ..……