Bochumer Linguistische Arbeitsberichte (Bla 2)

Kategorisierungsprobleme bei der Wortarten-Annotation von Textkorpora

Katja Keßelmeier
Anneli von Könemann

Bochumer Linguistische Arbeitsberichte

Herausgeber: Stefanie Dipper & Björn Rothstein

Die online publizierte Reihe "Bochumer Linguistische Arbeitsberichte" (BLA) gibt in unregelmäßigen Abständen Forschungsberichte, Abschluss- oder sonstige Arbeiten der Bochumer Linguistik heraus, die einfach und schnell der Öffentlichkeit zugänglich gemacht werden sollen. Sie können zu einem späteren Zeitpunkt an einem anderen Publikationsort erscheinen. Der thematische Schwerpunkt der Reihe liegt auf Arbeiten aus den Bereichen der Computerlinguistik, der allgemeinen und theoretischen Sprachwissenschaft und der Psycholinguistik.

The online publication series "Bochumer Linguistische Arbeitsberichte" (BLA) releases at irregular intervals research reports, theses, and various other academic works from the Bochum Linguistics Department, which are to be made easily and promptly available for the public. At a later stage, they can also be published by other publishing companies. The thematic focus of the series lies on works from the fields of computational linguistics, general and theoretical linguistics, and psycholinguistics.

Band 2 (November 2010)

Bandautor/in: Katja Keßelmeier & Anneli von Könemann

Bandherausgeber: Stefanie Dipper
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Björn Rothstein
Germanistisches Institut
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Erscheinungsjahr: 2010
ISSN **2190-0949**

Katja Keßelmeier
Anneli von Könemann

Kategorisierungsprobleme bei der Wortarten-Annotation von Textkorpora

2010

Bochumer Linguistische Arbeitsberichte
(Bla 2)

Inhalt

0 Einleitung

0.1 Fragestellung und Ziel

Am Sprachwissenschaftlichen Institut der Ruhr-Universität Bochum wurde im Rahmen eines PoS-Tagging-Projekts ein automatisch annotiertes Korpus auf Fehler bei den Wortarten-Tags überprüft und korrigiert. Dabei stellte es sich als notwendig heraus, zunächst die Problematik der Wortartenklassifikation genauer zu untersuchen. Zwei Problemfelder waren besonders auffällig: Zum einen waren einige Tags vom Tagger inkonsistent zugewiesen worden, zum anderen schien das zugrunde liegende Tag-Inventar, das Stuttgart-Tübinger Tagset (STTS)[1] (Schiller/Thielen 1999), in sich nicht immer schlüssig.

Dies warf die Frage auf: Was muss ein Tagset berücksichtigen, damit linguistisch plausibel und möglichst konsistent annotiert werden kann und die daraus resultierenden Daten statistisch weiterverarbeitet werden können? Denn häufig hängt eine Einteilung der Wortarten davon ab, mit welchem Ziel eine solche Kategorisierung vorgenommen wird (beispielsweise in einem computerlinguistischen Rahmen oder mit dem Ziel, eine Schulgrammatik zu schreiben).

Die vorliegende Arbeit[2] hat das Ziel, die Problematik der Wortartenklassifikation im Rahmen des automatischen Taggings genauer zu beleuchten und Lösungsvorschläge vor allem für die manuelle Klassifikation zu erarbeiten.

0.2 Hintergrund

Um Text mit statistischen Methoden zu verarbeiten, also statistische Aussagen beispielsweise über die in einem Text vorkommenden Wortarten machen zu können, müssen die einzelnen Wörter bzw. Wortarten eines Korpus zunächst eindeutig bestimmt werden. Dies geschieht mittels Part-of-Speech-Tagging (POS-Tagging), also Wortarten-Annotation. Dabei werden den einzelnen Wortformen so genannte Tags zugeordnet. Dies sind Etiketten, die Auskunft über die jeweilige Wortart geben, z.B.:

(1)

Der/ART	*Hund*/NN	*bellt*/VVFIN –
Der-Artikel	*Hund*-Nomen	*bellt*-finites Vollverb
(nach STTS[3])		

Wie der Annotation zu entnehmen ist, setzt sich der Beispielsatz aus einem Artikel, einem Nomen und einem Verb in seiner finiten Form zusammen.

An diesem Beispiel wird deutlich, dass nicht nur traditionelle Wortarten wie Nomen, Verb usw. zugewiesen werden, sondern auch Merkmale wie FIN für

---

1 STTS steht für ‚**S**tuttgart-**T**übinger **T**ag**s**et'. Es wurde 1995 von Anne Schiller, Christine Thielen, Simone Teufel und später (1999) auch von Christine Stöckert entwickelt und seitdem weiter modifiziert.

2 Bei der vorliegenden Arbeit handelt es sich um eine überarbeitete Fassung unserer gemeinsamen Bachelor-Arbeit vom 25.02.2005.

3 Wir beziehen uns, soweit nicht anders erwähnt, auf die Version von 1999.

Finitheit des Verbs. VVFIN wird von VVIMP oder VVINF unterschieden, da sie nicht gleich distribuiert sind.

Die traditionelle Wortartenlehre geht von den Kategorien Nomen, Verb, Adjektiv, Präposition und Adverb aus, während das STTS-Tagset auf elf Hauptwortarten basiert. Ausgehend von diesen finden sich feinere Unterteilungen der jeweiligen Tags, die insgesamt ein Tagset bilden: eine vordefinierte Menge von Bezeichnungen, mit denen nicht nur Wörter[4], sondern auch Nicht-Wörter wie Zahlzeichen, Sonderzeichen, Interpunktion usw. annotiert werden können.

Zum einen sollen mittels solcher Tagsets Wörter einer Wortklasse zugeordnet werden können, zum anderen sollen die Tags so differenzierte Aussagen über ein Wort erlauben, dass später statistische Verfahren herangezogen werden können, um das Verhalten anderer Wörter im jeweiligen Kontext so zuverlässig wie möglich vorherzusagen.

Wie wir in Abschnitt 2.2.2 zeigen werden, ist die Klassifizierung von Wörtern dabei nicht ganz unproblematisch, da sowohl semantische als auch syntaktische oder morphologische Kriterien bei der Einordnung zugrunde gelegt werden können (Manning/Schütze 2002:144). Dazu kommt, dass gerade im Deutschen eine eindeutige Bestimmung nicht immer möglich ist, wie an den vorgestellten Problemen in Kapitel 3 deutlich werden wird.

Wie ein Tagset beschaffen sein muss, um möglichst gute Dienste zu leisten, hängt davon ab, was erreicht werden soll. Eine möglichst präzise Vorhersage für die Wortart des nächsten Wortes in einem bestimmten Kontext gelingt am leichtesten formbasiert beziehungsweise aufgrund distributioneller Kriterien, die darüber Aufschluss geben, welche Wortarten statistisch gesehen besonders häufig in einer bestimmten messbaren Distanz zueinander stehen, doch werden in Tagsets auch häufig die oben genannten Gesichtspunkte eingearbeitet.

Zu berücksichtigen ist dabei der Umstand, dass eine automatische Vorhersage umso treffsicherer wird, je differenzierter die Wortformen unterschieden sind, also je mehr Informationen ein Tag über das entsprechende Wort liefert. Je feinkörniger unterschieden wird, desto genauer kann zwar annotiert werden, doch umso schwieriger wird auch die manuelle Klassifizierung, weil sehr viele Kriterien gleichzeitig erfüllt sein müssen (Manning/Schütze 2002:144f.). So kann ein Adverb ausschließlich als Adverb getaggt werden, aber auch eine weitere Unterteilung in lokales, temporales und modales Adverb ist denkbar.

0.3 Aufbau und Methodik

Als Grundlage für die vorliegende Arbeit dienten uns Ausschnitte aus den Jahrgängen 1993 bis 1999 der Neuen Zürcher Zeitung (NZZ) und das STTS-Taginventar. Zunächst stellen wir verschiedene Tagsets vor, um deutlich zu machen, dass nicht nur in verschiedenen Sprachen unterschiedlich annotiert

[4] Genauere Ausführung zum ‚Wort' siehe Abschnitt 2.1.1.

werden muss, sondern auch innerhalb einer Sprache unterschiedliche Annotationen ihre Berechtigung haben.

Es folgt eine kurze Darstellung der Wortartenklassifikation anhand verschiedener theoretischer Positionen.

Am Beispiel ausgewählter Probleme, die als besonders kontrovers aufgefallen sind, verdeutlichen wir die Relevanz der Wortarteneinteilung für ein effektives Tagging.

1 Vorstellung und Vergleich verschiedener Tag-Sets

Für verschiedene Sprachen sind unterschiedliche Tagsets nötig, da morphologische und distributionelle Eigenheiten der jeweiligen Sprachen beim Taggen berücksichtigt werden müssen. So unterscheiden sich z.B. agglutinierende und flektierende Sprachen erheblich voneinander: Das Finnische etwa kommt mit wenigen Adpositionen aus und drückt die entsprechenden Verhältnisse mittels Kasus aus, während das Deutsche sich in den Formen der Kasus kaum noch unterscheidet, dafür aber über viele Adpositionen verfügt. Auch die flektierenden Sprachen untereinander sind sehr vielgestaltig. Das Deutsche etwa verfügt über Artikel und wenig distinktive Kasus, das Slowenische hingegen hat distinktive Kasus und keine Artikel, wie am Beispiel *Katze – mačka* zu sehen ist:

(2) *die Katze* – Nominativ/Akkusativ
(3) *der Katze* – Genitiv/Dativ
(4) *mačka* – Nominativ
(5) *mačke* – Genitiv
(6) *mački* – Dativ
(7) *mačko* – Akkusativ
(8) *(pri) mački* – Lokativ
(9) *(z) mačko* - Instrumentalis

Daher sind Tagsets für Einzelsprachen grundsätzlich nicht auf andere Sprachen übertragbar.

In Abschnitt 1.2 stellen wir kurz die wichtigsten Tagsets für britisches (B.E.) bzw. US-amerikanisches Englisch (A.E.) und in 1.4.1 das wichtigste deutsche Tagset vor.

1.1 Problem der Sparse Data

Ein wesentliches Problem bei der Zuteilung von Tags besteht in den so genannten ‚Sparse Data'.

Relativ wenige Wörter kommen relativ häufig vor und finden sich in den meisten Texten wieder, die weitaus meisten Wörter jedoch sind eher selten und kommen daher auch in größeren Korpora unter Umständen gar nicht vor. Gibt es also zu viele spezialisierte Tags, ist die Wahrscheinlichkeit, dazu passende Wörter

in einem Korpus zu finden, sehr gering, d.h. es gibt zu wenige Daten (engl.: sparse data), auf denen eine statistische Auswertung begründet werden könnte.

Durch Reduzierung von Tags wird die Klassifikation der Wortarten vergröbert und die Gefahr des Sparse Data-Problems verringert.

1.2 Tagsets für Englisch (B.E. und A.E.)

Die historisch gesehen wichtigsten Tagsets waren eine Reihe von Sets, die von der University of Lancaster für die Bearbeitung des Lancaster-Bergen-Oslo-Korpus entwickelt wurden. Später wurde auch das British National Korpus damit bearbeitet. Diese Sets laufen unter dem Namen CLAWS (Constituent Likelihood Automatic Word-tagging System), Version 1 bis zurzeit Version 8. Das Tagset, mit dem das amerikanische Brown-Korpus getaggt wurde, ist das so genannte Brown Tagset.

1.2.1 CLAWS Tagset

CLAWS1 verfügte zunächst über 132 Tags in der ersten Version und wurde im Laufe der Zeit immer wieder angepasst, um die Genauigkeit der Taggingergebnisse und damit der statistischen Auswertung zu verbessern (Manning/Schütze 2002:140). CLAWS2 arbeitet mit einem erweiterten Set aus 166 Tags, CLAWS5 (auch C5 genannt) schließlich nur noch mit 62 Tags. Die zurzeit aktuelle Version 8 bedient sich 170 verschiedener Tags. Gegenüber C5 mit vier Tags für Nomina[5] verwendet C8 hier 22 Tags und unterscheidet damit deutlich genauer die Funktionen von Nomina, wie z.B. als Ausdruck von Zeit wie *day* oder als Numeral wie *dozen* etc. (Jurafsky/Martin 2000:837ff.).

1.2.2 Brown Tagset

Auch das Brown-Tagset wurde in der Anzahl der Tags immer wieder angepasst, um die Trefferquote zu verbessern. Mit ursprünglich 87 einzelnen Tags wurde 1961 das Brown-Korpus bearbeitet. Mit der Zeit wurde dieses Tagset immer weiter auf schließlich 179 Tags vergrößert. Damit konnte den in der englischen Sprache vorkommenden Spezialfällen (wie etwa in Bsp. (10) und (11)) Rechnung getragen werden.

(10) *you'll – you will*
(11) *can't - cannot*

Auch Fremdwörter sollten deutlicher hervorgehoben werden, und so wurde zusätzlich ein Tag für Fremdwörter eingeführt, um sie einerseits als solche kenntlich zu machen, aber gleichzeitig auch ihre Funktion in der fremden Sprache zeigen zu können. Beispielsweise würde nach dieser Systematik das deutsche Wort *Hund* in einem englischsprachigen Korpus wie folgt getaggt:

---

[5] Es handelt sich bei den vier Tags um: NN0 - noun (neutral for number), NN1 - singular noun, NN2 - plural noun, NP0 - proper noun.

(12) *Hund*/FW-NN

Dabei steht *FW* für Fremdwort und *NN* für fremdsprachiges Nomen im Singular, es würde also sowohl der Status als Fremdwort als auch die Wortart in der entsprechenden Sprache aufgezeigt. Der Trend zu einer solchen Doppelbenennung ist jedoch mittlerweile gegenläufig (Manning/Schütze 2002:143).

Auch im STTS, dem wichtigsten deutschen Tagset, wird auf die Doppelbenennung verzichtet und entweder nur FM oder als Eigennamen erkannte Fremdwörter auf NE oder die deutsche Wortart annotiert. Eine Doppelbenennung würde dazu führen, dass beim wortartengetreuen Tagging von Fremdwörtern Kenntnisse in der jeweiligen Fremdsprache nötig wären. Dies wäre nicht nur nicht wünschenswert, sondern in den meisten Fällen auch nicht möglich.

1.2.3 Penn Treebank Tagset

Im Rahmen des Penn Treebank Projektes der University of Pennsylvania in Philadelphia wurde aufbauend auf dem Brown-Tagset das kleinere Penn Treebank Tagset[6] entwickelt. Um syntaktische und lexikalische Redundanz auszuräumen, wurde das Tag-Inventar des Brown Tagsets im Laufe der Zeit stark verkleinert und auf 45 Tags reduziert.

Da die Sätze der Penn Treebank nicht nur getaggt, sondern auch geparst wurden, also Informationen über die syntaktischen Funktionen der Wörter zumindest zum Teil aus den Syntaxbäumen entnommen werden konnten, war es möglich, mit weniger Tags auszukommen (Jurafksy/Martin 2000:298). Zum Beispiel erhielten Präpositionen und subordinierende Konjunktionen denselben Tag. Ihre Unterscheidung ergab sich durch ihre syntaktische Position.

Korpora, die getaggt werden, werden vorher nicht zwangsläufig geparst, so dass die syntaktischen Funktionen nicht ohne Weiteres ersichtlich sind. Außerdem entfallen im Penn-Tagset aufgrund des geringen Umfangs an Tags einige lexikalische Informationen wie z.B. eine Unterscheidung zwischen Hilfs- und Vollverben oder zwischen den Zeitformen der Verben. Daher sind die 45 Tags des Penn-Tagsets häufig nicht spezifisch genug (Manning/Schütze 2002:140). Dennoch ist es in den letzten Jahren das dominierende Tagset für englischsprachige Texte geworden.

1.3 Unterschiede zwischen den Tagsets

Die einzelnen Sets unterscheiden sich deutlich in ihren Schwerpunkten.

Während das Brown Tagset beispielsweise die Formen der Hilfsverben *have*, *do* und *be* grundsätzlich als Hilfsverb taggt, auch wenn sie im konkreten Satz Vollverb sind, werden in der Penn Treebank Voll- und Hilfsverben gar nicht unterschieden (Taylor 2003:6ff.).

---

[6] Tagset siehe Anhang.

C8 (die achte Version von CLAWS) weist elf verschiedene Tags im Bereich der Interpunktion auf, Penn verwendet hier neun Tags. Während das Brown Tagset Pronomina in fünf verschiedene Unterkategorien einteilt, begnügt sich Penn mit einem einzigen Tag (Manning/Schütze 2002:141).

Auch in der Klassifikation der einzelnen Wörter lassen sich zwischen den verschiedenen Tagsets fundamentale Unterschiede feststellen. So zählt z.B. C8 die subordinierenden Konjunktionen zur Gruppe der Konjunktionen, das Penn Tagset jedoch ordnet sie bei den Präpositionen ein (Manning/Schütze 2002:143).

1.4 Tagsets für das Deutsche

1.4.1 STTS

Wie bereits angesprochen, haben verschiedene Tagsets unterschiedliche Schwerpunkte. Dies ist nicht nur abhängig vom Ziel, mit dem ein Korpus getaggt wird, sondern auch von der jeweiligen Sprache. So könnte beispielsweise im Englischen bei *-ing*-Formen eine Unterscheidung zwischen Gerundium einerseits und Partizip Präsens andererseits relevant sein. Im ersten Fall müsste zwischen adjektivischem und verbalem (Bsp. (13) bzw. (14)), im letzteren zwischen nominalem und verbalem (Bsp. (15) und (16)) Gebrauch unterschieden werden (Allen 1970:186).

(13) *An exciting story*
(14) *A burning house*
(15) *Swimming is a sport*.
(16) *Climbing mountains is a sport, too*.

Wenn eine solche Unterscheidung beim Tagging berücksichtigt werden soll, kann diese beispielsweise auf Basis der Distribution der Formen vorgenommen werden: In Abhängigkeit vom Inventar des Tagsets wird der entsprechende Tag je nachdem ausgewählt, in welchem Kontext die Wortform steht. Dass diese Unterscheidung nicht notwendigerweise getroffen werden muss, zeigt der folgende Satz aus dem Brown Corpus (Manning/Schütze 2002:143):

(17) *Fulton/*NP-TL *County/*NN-TL *Purchasing/*VBG *Department/*NN

Obwohl es sich bei *Purchasing* um ein Nomen handelt, werden im Brown Corpus auf *-ing* endende Formen durchgängig mit VBG, also als Verb, getaggt. Die Vernachlässigung scheint keine Verschlechterung für das Ergebnis des Taggings zu bedeuten. Vielmehr kann so das Sparse Data-Problem umgangen werden.

In anderen Sprachen, beispielsweise im Deutschen, würde sich die Frage dieser speziellen Unterscheidung erst gar nicht stellen, da es keine Form gibt, die sowohl Gerundium als auch Partizip Präsens zugleich sein könnte.

Aufgrund solcher sprachlichen Unterschiede möchten wir nun, nach der Vorstellung der wichtigsten britischen bzw. amerikanischen Tagsets, auf das bedeutendste deutsche Tagset eingehen, das STTS[7].

Entstanden ist das STTS aus zwei POS-Tagsets, die an der Universität Tübingen (SfS) bzw. an der Universität Stuttgart (IMS) entwickelt wurden. Ziel der Vereinigung war eine übereinstimmende Annotation der Korpora beider Universitäten und daraus resultierend eine Vereinfachung bei der gegenseitigen Nutzung bereits durchgeführter Korpusarbeit.

1.4.1.1 Struktur und Einteilung des STTS

Das STTS enthält 54 Tags. 48 davon sind reine POS-Tags, sechs zusätzliche Tags werden für fremdsprachliches Material (FM), Kompositions-Erstglieder (TRUNC), Nichtwörter (XY) und Satzzeichen verwendet. Die Tags bestehen aus möglichst selbsterklärenden Buchstabensequenzen und sind gruppiert. Durch die Gruppierung der Tags wird eine implizite Beziehung deutlich: Die Tags sind strukturiert in Haupt- und Unterwortart. Von links nach rechts gelesen gelangt man also von der allgemeinen zur spezifischeren Information, z.B.

(18) VVINF – **V**oll**v**erb **inf**init
(19) VVFIN – **V**oll**v**erb **fin**it

Das STTS unterscheidet zwischen Nomina (N), Verben (V), Artikeln (ART), Adjektiven (ADJ), Pronomina (P), Kardinalzahlen (CARD), Adverbien (ADV), Konjunktionen (KO), Adpositionen (AP), Interjektionen (ITJ) und Partikeln (PTK) und geht somit von elf Hauptwortarten aus, die wiederum subklassifiziert sind. So werden beispielsweise Konjunktionen nochmals in nebenordnende Konjunktionen (KON), unterordnende Konjunktionen (KOUS) sowie unterordnende Konjunktionen mit *zu* und Infinitiv (KOUI) eingeteilt. Mit Ausnahme der Kardinalzahlen und Konjunktionen richtet sich diese Klassifizierung nach dem TEI Starter Set Of Grammatical-Annotation Tags (TEI 1991).

1.4.1.2 Zuweisung von Tags

Grundsätzlich wird jeder Wortform genau ein Tag zugewiesen. Der Begriff ‚Wortform' bezeichnet hier neben echten Wortformen (wie *bin*, *bist*, *sind*) auch Zahlen in Ziffernschreibweise, Satzzeichen, Sonderzeichen, abgetrennte Wortteile (zum Beispiel abgetrennte Verbzusätze wie *ich sah <u>hin</u>*) oder Kompositions-Erstglieder (*<u>hin</u>- und hergerissen*). Dies geht von der Voraussetzung aus, dass der zu taggende Text tokenisiert ist. Das bedeutet, dass der Text in einzelne Wörter unterteilt wird, die üblicherweise durch Leerzeichen voneinander getrennt werden.

Die Quote korrekter Disambiguierungen liegt bei den besten Taggingverfahren zwischen 96% und 98%, somit scheint auf den ersten Blick die Genauigkeit bereits

[7] Tagset siehe Anhang.

relativ hoch zu sein. Diese beeindruckende Angabe ist jedoch nur bedingt aussagekräftig, denn ihr liegt eine wortbasierte Auswertung zugrunde.

Ein Vergleich soll diese Zahlen verdeutlichen: Ein 15-Wörter-Satz wird mit einer Wahrscheinlichkeit von nur 63% vollständig richtig getaggt, wenn die Trefferquote des Taggers bei 97% liegt. Liegt die Trefferquote bei 98%, erhöht sich die Chance bei demselben Satz bereits auf 74% (Manning/Schütze 2002:373).

1.4.2 UIS

Wie bereits in der Einleitung angesprochen, können beim Tagging verschiedene Ziele verfolgt werden. Je nach Zweck, zu dem das Korpus getaggt wird, bieten sich unterschiedlich feine Klassifikationen an. Wenn beispielsweise Vorhersagen über das wahrscheinlichste nächste Wort getroffen werden sollen, können bessere Ergebnisse erzielt werden, wenn das Taginventar möglichst fein strukturiert ist. Genau dies war auch Ziel des UIS. UIS steht für Universitäts-Informations-System der Uni Zürich, wo im Rahmen des UIS-Projektes mit einer leicht angepassten Variante des STTS ein Tagger für das Deutsche trainiert wurde.

1.4.2.1 Verfeinerungen im UIS

Verfeinerungen des STTS-Inventars werden am Beispiel der Abgrenzung von Vollverben (VV*) und Hilfsverben (VA*) deutlich. Im STTS werden Verben wie *sein* und *haben* grundsätzlich als VA* klassifiziert, das UIS klassifiziert hier in Abhängigkeit von der Funktion des Verbs: Ist eine Form von *sein* oder *haben* als Vollverb realisiert, wird sie auch so, mit VV*, getaggt.

Ein weiteres Indiz für eine genauere Vorgehensweise des UIS ist die Unterscheidung zwischen bestimmtem (ARTDEF) und unbestimmtem (ARTIND) Artikel statt des einheitlichen Tags ART.

Die Tags für Interpunktion weichen im UIS deutlich von den Konventionen des STTS-Tagsets ab. Während diese Tags im STTS aus Zusammensetzungen von Satz- bzw. Dollarzeichen bestehen ($, für Komma, $. für satzbeendende Interpunktion, $(für sonstige Satzzeichen, satzintern), benutzt das UIS hierfür selbsterklärende Tags (C für Comma, Ex für Exclamation Mark etc.).

2 Probleme der Klassifikation

2.1 Theorien zu Wort und Wortarten

2.1.1 Was ist ein Wort?

Um die einzelnen Wörter eines Korpus verschiedenen Wortarten zuordnen zu können, ist zunächst eine genauere Betrachtung des Terminus ‚Wort' nötig.[8]

Dieser ist in der Linguistik umstritten und findet mindestens drei verschiedene Verwendungsweisen:

- auf orthographischer Ebene als das, „was zwischen zwei Leerräumen geschrieben steht",
- auf lexikalischer Ebene als „die entsprechende sprachliche Einheit Lexem" und
- auf grammatischer Ebene als „sprachliche Einheit, die [..] innerhalb der sprachlichen Einheit Satz bestimmte syntaktische Positionen und Funktionen erfüllt" (Bergenholtz/Schaeder 1977:58).

Bergenholtz/Schaeder (1977) kritisieren zu Recht, dass dabei häufig nicht genau zwischen den Ebenen unterschieden wird und alle drei Beschreibungsebenen gleichgesetzt werden.

Den inhaltlichen Aspekt eines sprachlichen Zeichens ‚Wort' sehen sie hier einerseits durch das dazugehörige Lexem und andererseits durch die syntaktische Position und Funktion bestimmt, die das Wort innerhalb des Syntagmas innehat:

(1) *Es hat rund 9 Euro gekostet*
(2) *Das Programm läuft rund*.

Das Lexem *rund* und die jeweilige entsprechende Funktion im Satz bestimmen die Bedeutung als ‚ungefähr' beziehungsweise ‚problemlos'; das Lexem allein lässt ohne weiteren Kontext zunächst keine Rückschlüsse über die inhaltliche Seite zu.

Bußmann (2002) unterteilt die möglichen Beschreibungsformen in fünf verschiedene Ebenen.

- Auf der phonetisch-phonologischen Ebene ist ein Wort demnach eine Ansammlung von kleinsten, durch Wortakzent und Grenzsignale isolierbaren Lautsegmenten.
- Wörter werden auf orthographisch-graphemischer Ebene durch Leerstellen im Schriftbild voneinander getrennt. Diese Definition

[8] Unsere Überlegungen sollen die Begriffe ‚Wort' und ‚Wortart' für die deutsche Sprache klären. Für andere Sprachen wären zusätzliche Betrachtungen nötig. Fremdsprachige Beispiele dienen lediglich der Bewusstmachung der Problematik.

entspricht am ehesten der gebräuchlichen, nicht-sprachwissenschaftlichen Verwendung des Terminus ‚Wort'.

- Auf morphologischer Ebene wird das Wort als Grundeinheit grammatischer Paradigmen betrachtet, die durch Flexion gekennzeichnet und strukturell stabil sowie durch spezifische Wortbildungsregeln beschreibbar sind.
- Lexikalisch-semantisch gesehen ist ein Wort der kleinste relativ selbstständige Träger von Bedeutung, der ins Lexikon gestellt ist.
- Syntaktisch betrachtet handelt es sich um die kleinste Einheit, die innerhalb eines Satzes als Satzglied verschiebbar ist.

Über die verschiedenen Definitionen hinweg gelten akustische und semantische Identität sowie morphologische Stabilität und syntaktische Mobilität als Hauptkriterien für die Bestimmung des Terminus ‚Wort' (Bußmann 2002:750), doch ist die Terminologie nicht nur uneinheitlich, sondern es bleibt unklar, was genau z.B. mit morphologischer Stabilität gemeint ist. Aber auch die Einteilungen an sich – was genau macht z.B. die einzelnen Ebenen aus? – weichen voneinander ab.

Eisenberg (1999:15f.) betont, dass ‚Wort' umgangssprachlich nicht dasselbe bedeutet wie in der Sprachwissenschaft und schlägt eine Zweiteilung in ‚lexikalisches Wort' und, wenn es um Wörter im Satz geht, ‚Wortform' vor.

Das lexikalische Wort entspricht dabei dem Lemma[9] im Lexikon und steht für alle seine Wortformen und alle Bedeutungen aller seiner Formen. Diese Formen, also z.B. ein Substantiv in den verschiedenen Kasus und Numeri, und ihre jeweiligen Bedeutungen ergeben ein Wortparadigma oder auch syntaktisches Paradigma.

Die Wortform ist dagegen die einzelne Form des Lemmas in ihrer jeweiligen Kategorie, also etwa ein Substantiv im Nominativ Plural oder Akkusativ Singular. Ein Satz besteht immer aus Wortformen, nicht aus lexikalischen Wörtern. Darauf stellt auch Lyons (1971:198) ab, wenn er sagt, dass entsprechend der Funktion der Wörter im Satz diese Wörter verschiedene Formen annehmen, die zu Einheiten verbunden zu Sätzen zusammengefügt werden. Damit entspricht der Terminus ‚Wortform' von Eisenberg am ehesten dem intuitiven Begriff von ‚Wort', nämlich dem, „was zwischen zwei Leerräumen geschrieben steht" von Bergenholtz/Schaeder (1977:58).

Dabei ist zu berücksichtigen, dass eine Wortform unter Umständen mehrere grammatische Wörter repräsentieren kann, wie z.B. das englische *cut* für Präsens, Präteritum und Partizip Perfekt steht, also möglicherweise auch eine Unterscheidung zwischen orthographischen und grammatischen Wörtern notwendig sein könnte (Lyons 1971:200).

---

[9] Bergenholtz/Schaeder (1977) verwenden hier den Terminus ‚Lexem' wie Bußmann, die Lexem im weiteren Sinne als Synonym für Wort als lexikalische Einheit bzw. Element des Wortschatzes definiert (Bußmann 2002:400).

Der Genauigkeit halber müsste also immer gesagt werden, ob von Wortform oder lexikalischem Wort oder dem grammatischen Wort die Rede ist bzw. welche genaue Definition dem Terminus jeweils zugrunde liegt. Dies löst zwar keineswegs das Grundproblem, was genau ein Wort ist, macht aber transparent, auf welcher Bezugsebene argumentiert wird.

Das STTS (wie auch andere Tagsets) definiert an keiner Stelle, von welchem Wortverständnis es ausgeht und bezieht sich bei allen Kategorisierungen immer auf eine ‚Wortform'. Dies entspricht implizit der Definition von Eisenberg, der auch wir uns anschließen, denn ein zu annotierendes Korpus enthält in der Regel als Token genau die Einheiten, die zwischen den Leerzeichen stehen. Beim Taggen wird dann auf die Wortformen zugegriffen, wie sie dort vorkommen.

Wir benutzen in dieser Arbeit also den Terminus ‚Wort' im Sinne von ‚Wortform'.

2.1.2 Wort und Wortarten

Aus dem Problem der Definition des Terminus ‚Wort' ergibt sich entsprechend ein Problem bei der Einteilung in ‚Wortarten'.

Sowohl in der Wortlehre als auch in bekannten Wörterbüchern gehört eine Klassifikation der Wortarten zur Beschreibung des Wortbestandes. Wie bei der Erfassung des Terminus ‚Wort' sind auch hier unterschiedliche Bezeichnungen üblich, wie etwa ‚Redeteile', ‚Wortklassen', ‚Lexemklassen' oder ‚Wortarten', ‚parts of speech' und ‚partes orationis'.

Auch Einteilungen, die von mehrteiligen Wortbegriffen ausgehen, wie es bspw. bei mehrteiligen Präpositionen häufig der Fall ist, klammern wir aus, da beim Tagging in der Regel pro Wort tokenisiert wird.

Eisenberg (1999:14) bietet eine Einordnung der Wortarten zu den grammatischen Kategorien an und unterteilt sie in lexikalische Kategorien und Funktionswörter. Dabei haben die lexikalischen Kategorien eine „Wortbedeutung im eigentlichen Sinn", die Funktionswörter eine rein strukturelle Bedeutung. Problematisch wird diese Einteilung, wenn Wörter in beide Untergruppen eingeordnet werden können, wie beispielweise Präpositionen. So ließe sich *über* in (3) eine lexikalische Bedeutung zuweisen, nämlich die Angabe der Lokalität, in (4) jedoch nicht.

(3) *Die Lampe hängt über dem Regal.*
(4) *Er spricht schlecht über dich.*

Eine andere Möglichkeit wäre die Einteilung in offene, auf der Wortbildungsebene produktive, und geschlossene, nicht mehr produktive, Kategorien. Dabei wäre zwar relativ leicht festzustellen, welche Wortarten zu den jeweiligen Kategorien gehören, da z.B. primäre Präpositionen, Konjunktionen und einige andere nur begrenzt in unserem Wortschatz enthalten sind und damit den geschlossenen

Wortarten zuzuordnen sind, doch löst sich dadurch nicht das Problem der Abgrenzung einzelner Wortklassen untereinander.

Bergenholtz/Schaeder (1977:10ff.) unterscheiden zwischen den Wortarten als grammatisch-syntaktisch begründete und Lexemklassen als lexikalisch-morphologisch begründete Kategorie.

Wöllstein-Leisten ordnet die Wortarten den lexikalischen Kategorien zu und setzt sie von den Phrasentypen als phrasale Kategorien ab. Sie verweist dabei auf neun traditionelle Wortarten des Deutschen, die nach ihren distributionellen, morphologischen und syntaktischen Eigenschaften unterschieden werden (1997:19ff.).

Zifonun (1997:24ff.) verweist wie die anderen auf Funktion einerseits und Form andererseits und schlägt zur Klassifizierung Kriterien-Bündel vor, die neben funktionalen Kriterien auch semantische Kriterien mit einbeziehen.

2.2 Das Problem der Wortarten

2.2.1 Historie

Bereits Platon analysierte die Wörter und kam zu der Unterscheidung von Onoma (‚Namen': Nomen) und Rhema (‚Aussage': Verb). Aristoteles ergänzte diese um die Gruppe der ‚Undeklinierbaren'. Heutige Klassifizierungen basieren auf Dyonisios Trax (1. Jh. v. Chr.) und gehen größtenteils von acht Wortarten aus: Nomen, Verb, Adjektiv, Artikel, Pronomen, Präposition, Adverb, Konjunktion. Je nach theoretischem Hintergrund, auf den sich einzelne Vertreter jeweils stützen, wird auch von fünf, sieben oder zehn Wortarten ausgegangen. Diese Angaben zeigen, dass die Aufgabe einer einheitlichen Klassifizierung bislang weder theoretisch noch praktisch zufriedenstellend gelöst werden konnte.[10]

2.2.2 Klassifizierung der Wortarten

Die Einteilung von Wortarten ist nicht trivial, da Klassifizierungen zum einen auf verschiedenen Ebenen vorgenommen werden (Morphologie, Distribution, teilweise Semantik), zum anderen theorie- und nicht sprachabhängig sind (Zifonun 1997:23). Die Abhängigkeit zwischen Wortarten und ihrem theoretischen Rahmen wirft allerdings weitere Fragen auf:

> Zur Begründung und Definition einer Kategorie Wortarten ist zuallererst der sprachtheoretische Rahmen aufzuzeigen, innerhalb dessen sie etabliert werden soll, ehe sich im weiteren der Platz bestimmen läßt, den sie in diesem Rahmen einnimmt.
>
> (Bergenholtz/Schaeder 1977:12ff)

Eine Annäherung an das Problem der Wortartenklassifizierung scheint zu sein, dass sich alle unterschiedlichen Einteilungen weitgehend auf drei gemeinsame

[10] Außerdem ist bei Klassifizierungsansätzen zu berücksichtigen, dass die vorgenommenen Einteilungen vom Griechischen ausgehen. Ob sie auf nicht-indoeuropäische Sprachen übertragbar sind, bleibt zu prüfen.

Gliederungskriterien beziehen (Bußmann 2002:750), von denen der eine die Morphologie, der andere die Syntax und der dritte die Semantik betrifft. Die unterschiedlich starke Berücksichtigung der einzelnen Aspekte bewirkt jedoch voneinander abweichende Klassifizierungen sowie diverse Schwierigkeiten innerhalb der Klassifizierung.

2.2.2.1 Morphologisches Kriterium

Die Unterscheidung zwischen flektierenden und nicht flektierenden Wörtern (Substantiv, Adjektiv, Verb und Pronomen einerseits sowie Adverb, Konjunktion und Präposition andererseits) macht den morphologischen Aspekt aus. Die Unterschiede innerhalb der Gruppe der nicht flektierenden Wörter werden dadurch jedoch nicht erfasst. Deutlich wird dies beispielsweise am Vergleich von Präpositionen und Konjunktionen, die beide keinerlei Flexionsmerkmale wie z.B. Konjugierbarkeit, Deklinierbarkeit oder Komparierbarkeit haben. Also reicht die alleinige Betrachtung der Morphologie für eine vollständige Klassifizierung nicht aus.

2.2.2.2 Syntaktisches Kriterium

Wörter haben folgende Eigenschaften:

- Sie können als Satzglied verwendbar sein. Folgende Beispiele verdeutlichen dies[11]:

(5) *Während des Essens las er die Zeitung.*
(6) *Während er aß, las er die Zeitung.*

Hier stellt sich das Problem, dass *während* verschiedenen Wortarten zugeordnet werden kann, den Konjunktionen und den Präpositionen. In Bsp. (5) folgt dem zu kategorisierenden Lexem ein Substantiv im Genitiv, und das finite Verb steht nach dem ersten Satzglied – es handelt sich um eine Präposition. In (6) hingegen folgt dem Lexem ein nominales Element im Nominativ, und das finite Verb steht an der letzten Stelle im Satz – *während* ist hier eine Konjunktion. Position und Distribution helfen also dabei, eine Entscheidung zu treffen.

- Sie können nominale oder verbale Elemente modifizieren. Hinsichtlich dieser Eigenschaft sei auf die Erörterung der Unterscheidung von ADJD vs. ADV in dieser Arbeit verwiesen (Kap. 3.1).

- Sie können einen Artikel zu sich nehmen (Substantiv vs. Pronomen). Die Eigenschaft eines Wortes (typischerweise eines nominalen Elements), mit bzw. ohne Artikel aufzutreten, ermöglicht eine

[11] Die Beispiele sind entnommen aus Helbig 1977:98.

Unterscheidung zwischen Substantiv (welches mit Artikel vorkommen kann) und Pronomen (welches immer ohne Artikel vorkommt).

- Sie können als Substantiv oder Pronomen durch Rektion einen bestimmten Kasus erhalten (Präposition vs. Konjunktion). Dies kann entschieden werden, da eine Präposition Kasus fordert und somit das Wort innerhalb ihrer Rektionsdomäne kasusmarkiert sein muss. Auf eine Konjunktion trifft dies nicht zu.

Wenn auch der letzte Aspekt u.U. zu den morphologischen Kriterien gezählt werden könnte[12], wird deutlich, dass ein Hinzuziehen von syntaktischen Kriterien eine Klassifizierung erheblich erleichtert.

2.2.2.3 Semantisches Kriterium

Unter dem semantischen (oder auch begrifflich-kategorialen) Aspekt führt Bußmann (2002:750) auf, dass „die drei Grundwortarten Substantiv, Verb und Adjektiv [...] auf den logischen Kategorien ‚Substanz', ‚Prozess' und ‚Eigenschaft' [beruhen], während Konjunktion und Präposition durch die Kategorie der ‚Relation' begründet werden." Dass Wörtern einer bestimmten Wortart nicht immer die gleichen Typen von Sachverhalten in der Außenwelt direkt entsprechen, zeigen die Beispiele *Hoffnung, sein* bzw. *liegen, gestrig*. *Hoffnung* bezeichnet keine Substanz, *sein* bzw. *liegen* sind keine Prozesse und *gestrig* keine Eigenschaft. Dennoch spricht Jung (in Helbig 1977:103) von den Wortarten als „Abbild der Wirklichkeit". Basierend auf den oben genannten Entsprechungen von Wortarten und Sachverhalten in der Außenwelt „ordnen [sie] sprachlich die uns umgebende Welt...".

Bei Klassifikationen, die nur auf einem der drei Kriterien basieren, wird immer mindestens eine Eigenschaft von Wörtern vernachlässigt. Doch auch zwei oder drei Aspekte zu berücksichtigen, ist nicht unproblematisch, denn gerade eine unterschiedliche Gewichtung führt wieder zum Überwiegen eines Aspekts.

> Die Wortarten werden so angesetzt, daß jedes Wort mindestens einer und - besser noch - genau einer von ihnen angehört. (Eisenberg 2004:35)

Mit diesem Ziel vor Augen zu einer Entscheidung zu gelangen, welches Kriterium berücksichtigt und welches vernachlässigt wird, stellt die größte Schwierigkeit dar. Als bestes Beispiel für dieses Dilemma dienen die Numeralia. Unter dem Aspekt ihrer gemeinsamen lexikalischen Merkmale (sie bezeichnen Zahlen und Mengen) stellen sie eine selbstständige Gruppe dar. Bei Betrachtung ihrer einzelnen Vertreter unter syntaktischen Gesichtspunkten ergibt sich jedoch ein anderes Bild (Bußmann 2002:750):

---

[12] Die Kasusrektion an sich ist zwar der Syntax zuzuordnen, doch wird sie sichtbar anhand von morphologischen Merkmalen.

(7) *Tausende von Menschen*
(8) *ein Buch*
(9) *er ruft dreimal*

Die Numeralia in (7) bis (9) zeigen, dass sie sich wie Vertreter anderer Kategorien verhalten und somit auch anderen Kategorien zugeordnet werden könnten, wie beispielsweise das Zahlwort in (7) den Substantiven, in (8) den Adjektiven oder den Adverbien wie in (9).

2.2.3 Weitere Klassifizierungsprobleme

Andere Klassifizierungsprobleme können Wortartwechsel und Homonymie darstellen.

Wortartwechsel (oder auch Konversion) bezeichnet einen Prozess der Wortbildung, bei dem eine Stammkategorie in eine andere überführt wird (Bußmann 2002:380). Beispiele hierfür sind desubstantivische Verben des Deutschen *(frühstück(en))* und des Englischen *((to) bicycle)*, deverbale Substantive (dt. *Treff* beziehungsweise engl. *buy*) sowie deadjektivische Verben (*kürzen, to tidy*). Eine besondere Rolle spielt in diesem Zusammenhang die Klassifizierung von abgetrennten Verbzusätzen, weil diese oft einen Grammatikalisierungsprozess durchlaufen haben, der noch rekonstruierbar ist (*ich nehme teil*). In diesem Fall könnte *teil* beispielsweise auch als Nomen klassifiziert werden.[13]

2.2.4 Ziele einer Einteilung

Begonnen hat unsere Diskussion der Wortarten mit der Definition von ‚Wort'. Denn um überhaupt Wörter nach bestimmten Kriterien einteilen zu können, muss zunächst gesagt werden, welche Eigenschaften eines Wortes hierbei eine Rolle spielen.

Zum anderen ist der Punkt des theoretischen Status ausschlaggebend für die Einteilung der Wortarten. Dies ist bereits deutlich geworden.

Die von Bergenholtz/Schaeder (1977) angeführten Wortartensysteme beispielsweise wurden in einem computerlinguistischen Rahmen entwickelt und zur syntaktischen Analyse verwendet. Von semantischen Kriterien wurde dabei völlig abgesehen. Helbig (1977:104) fordert in diesem Zusammenhang Explizitheit und Einfachheit. Im muttersprachlichen Unterricht scheint die Explizitheit am wenigsten notwendig, da Kompetenz vorausgesetzt werden kann. Bei der automatischen Sprachverarbeitung ist die geforderte Explizitheit am größten, da Kompetenz ausgeschlossen wird. Der Fremdsprachenunterricht bewegt sich dazwischen.

Nach diesem Überblick über die Zuordnungsproblematik von Wörtern und Wortarten stellen wir im Folgenden ein Problem vor, das insbesondere für das Tagging eine der größten Herausforderungen darstellt.

---

[13] Diese Problematik wird im Rahmen des Tags PTKVZ in Kap. 3.4 diskutiert.

2.2.5 Ambiguität bzw. Mehrdeutigkeit

Bevor wir näher auf spezielle Fälle von Ambiguitäten bzw. Mehrdeutigkeiten eingehen, soll zunächst die Verwendung dieser beiden Begriffe geklärt werden.

Der Begriff ‚Mehrdeutigkeit' kann einerseits als Hyperonym zu ‚Ambiguitäten' und ‚Vagheit' (pragmatische Mehrdeutigkeit) verwendet werden. Andererseits können die Begriffe ‚Ambiguität' und ‚Mehrdeutigkeit' auch synonym gebraucht werden (Bußmann 2002:73,426). Da im Rahmen dieser Arbeit pragmatische Mehrdeutigkeiten vernachlässigt werden können, ist eine Differenzierung der beiden Termini hier nicht relevant. Somit folgen wir Bußmann und verwenden ‚Ambiguität' und ‚Mehrdeutigkeit' synonym.

Wenn einem Ausdruck mehrere Bedeutungen zukommen, handelt es sich um Ambiguität bzw. Mehrdeutigkeit. Laut Jurafsky/Martin (2000:299) können ambige Wörter in mehr als einer möglichen Art gebraucht sowie oft mehr als einer Wortart zugeordnet werden. Eine mögliche syntaktisch orientierte Erklärung ist, dass ein Ausdruck mehr als eine grammatische Beschreibung hat. Jurafsky/Martin (2000:4) sprechen hier von „multiple alternative linguistic structures".

Eine Vielzahl dieser Mehrdeutigkeiten eliminiert der Mensch unbewusst und somit völlig problemlos. Für die Computerlinguistik, die Sprache mithilfe von formalen Regelsystemen abzubilden versucht, stellt die Auflösung jedoch eine grundlegende Schwierigkeit dar.

Im Folgenden führen wir kurz in das Problem der Ambiguitäten ein, anschließend zeigen wir exemplarisch anhand von einigen Beispielen, wie sich verschiedene Arten von Ambiguitäten beim Tagging auswirken.

2.2.5.1 Homonymie

Homonymie ist eine Mehrdeutigkeit, die (u.a.) für das Tagging ein essenzielles Problem darstellt. Je nachdem, in welcher Bedeutung das jeweilige Wort auftritt, ist es möglicherweise unterschiedlichen Kategorien zuzuordnen. Daher ist in diesem Zusammenhang ausschließlich von kategorialer Homonymie die Rede.

(10) *Der*/ART *Ball*/NN *ist*/VVFIN *rund*/ADV *5*/CARD Kilo/NN *schwer*/ADJD.
(11) *Der*/ART *Ball*/NN *ist*/VVFIN *rund*/ADJD *5*/CARD *Kilo*/NN *schwer*/ADJD[14].

In den Beispielen (10) und (11) kann *rund* zum einen als Adverb (im Sinne von *zirka/ungefähr 5 Kilo schwer*) und zum anderen als Adjektiv (hier im Sinne von *ein runder Ball ist 5 Kilo schwer, ein ei-förmiger ist leichter*...) verstanden werden.

Ein anderes Beispiel, das verschiedene Arten von Mehrdeutigkeiten illustriert, nennen Jurafsky/Martin (2000:4):

(12) *I made her duck.*

[14] Getaggt nach STTS.

Zum einen sind die Wörter *duck* und *her* morphologisch beziehungsweise syntaktisch ambig, d.h. sie können jeweils unterschiedlichen Wortarten zugeordnet werden (*duck* kann sowohl Verb als auch Nomen, *her* entweder Personalpronomen im Dativ oder Possessivpronomen sein). Man könnte hier auch von lexikalischer Ambiguität sprechen. In Anlehnung an die Terminologie von Jurafsky/Martin (2000:288), die von „parts-of-speech“ als „word classes”, „morphological classes” sowie „lexical tags“ gleichermaßen sprechen, soll der Frage nach einer genaueren Klassifizierung von syntaktischen, morphologischen (oder lexikalischen) Mehrdeutigkeiten hier nicht weiter nachgegangen werden. Die Tatsache allein, dass manche Wörter verschiedenen Wortarten zugewiesen werden können, macht das Problem von Mehrdeutigkeiten für das POS-Tagging relevant - unabhängig davon, wie genau diese Ambiguität klassifiziert wird.

Zum andern ist das Wort *make* semantisch mehrdeutig (entweder kommt es in der Bedeutung *create* oder *cook* vor). Zu entscheiden, welche Bedeutung gemeint ist, ist Aufgabe der Word Sense Disambiguation (WSD)[15]. Hinzu kommt, dass *make* syntaktisch ambig ist: es tritt transitiv und ditransitiv auf.

Als ein ähnliches Beispiel für eine morphologische bzw. syntaktische Mehrdeutigkeit dient der folgende Satz, für den sowohl die Tagging-Variante in (13) als auch die in (14) (getaggt nach Brown/Penn) möglich ist (Manning/Schütze 2002:341):

(13) *The*-AT *representative*-NN *put*-VBD *chairs*-NNS *on*-IN *the*-AT *table*-NN.
(14) *The*-AT *representative*-JJ *put*-NN *chairs*-VBZ *on*-IN *the*-AT *table*-NN.

Auch eine solche Ambiguität wird vom menschlichen Sprecher intuitiv und mithilfe des Kontextes aufgelöst. Dies führt jedoch zu Problemen beim inter-annotator agreement, d.h. wenn verschiedene Personen dasselbe Textstück annotieren, führt die subjektive Intuition häufig zu unterschiedlichen Ergebnissen. Selbst wenn ein und dieselbe Person zu unterschiedlichen Zeitpunkten annotiert, wird nicht immer dieselbe Auswahl getroffen.

2.2.5.2 Disambiguierung

Nach der Vorstellung unterschiedlicher Formen von Mehrdeutigkeiten skizzieren wir nun verschiedene Möglichkeiten zur Auflösung (Disambiguierung) solcher Mehrdeutigkeiten.

Die meisten Wörter beispielsweise des Englischen sind nicht kategorial ambig (wenn auch aus jedem Verb ein Nomen gemacht werden kann). Doch gerade viele der am häufigsten vorkommenden Wörter können mehr als einer Wortart zugeordnet werden. So sind etwa im Brown Korpus nur 11,5% der Worttypen hinsichtlich ihrer Wortart ambig, aber das macht mehr als 40% der Token aus (Jurafsky/Martin 2000:299). Viele der ambigen Token sind jedoch leicht zu disambiguieren.

[15] WSD bezeichnet die Disambiguierung polysemer Ausdrücke durch Selektions-restriktionen und statistische Verfahren.

> The task of disambiguation is to determine which of the senses of an ambiguous word is invoked in a particular use of the word. This is done by looking at the context of the word's *use.* (Manning/Schütze 2002:229)

Diese Erklärung würde beispielsweise auf die Disambiguierung von Homonymie zutreffen, bei der die wahrscheinlichste Bedeutung im jeweiligen Kontext ermittelt werden soll. Der Kontext ist insofern hilfreich, als er die nötigen Informationen liefert, um die Wahrscheinlichkeit einer bestimmten Wortklasse eben in diesem spezifischen Kontext ermitteln zu können. Jurafsky/Martin verdeutlichen dies so:

> Knowing whether a word is a possessive pronoun or a personal pronoun can tell us what words are likely to occur in its vicinity (possessive pronouns are likely to be followed by a noun, personal pronouns by a verb). (Jurafsky/Martin 2000:288)

So arbeitet die syntagmatische Vorgehensweise: Betrachtet wird eine vorher festgelegte Anzahl von Tags vor und hinter dem fraglichen Wort und so die wahrscheinlichste Tagsequenz ausgewählt.

Modelle für das POS-Tagging arbeiten entweder syntagmatisch, d.h. mit dem lokalen Kontext, wie eben beschrieben, oder auf der Basis von lexikalischer Information. Bei der lexikalischen Vorgehensweise wird bei jeder Mehrdeutigkeit die wahrscheinlichste Wortklasse ausgewählt, jedoch ohne Kontext (Manning/Schütze 2002:343). Das englische Wort *book* beispielsweise kann entweder als Nomen oder als Verb vorliegen. Angenommen, die Häufigkeit, mit der es als Nomen vorkommt, läge bei 80%, und die Häufigkeit als Verb bei 20%, so würde die Wortklasse mit der größten Wahrscheinlichkeit zugewiesen, in dem Fall also ‚Nomen'. Dieses Verfahren hat eine Trefferquote von 90%, die sozusagen als Messlatte (‚Baseline') anzusehen ist und von keinem Verfahren unterschritten werden sollte (Manning/Schütze 2002:234). Selbst deutlich präzisere Verfahren führen zu erheblichen Fehlerquoten: Wird ein Zeitungssatz, der durchschnittlich mehr als 20 Wörter umfasst, mit einer Genauigkeit von 96% getaggt, so bedeutet dies, dass sich im Durchschnitt mehr als ein Taggingfehler pro Satz findet (Manning/Schütze 2002:342).

Üblich sind Tagger, die sowohl syntagmatische als auch lexikalische Informationen kombinieren (Manning/Schütze 2002:344).

Im nächsten Kapitel werden sechs Probleme näher vorgestellt, die bei der Arbeit am NZZ-Korpus Fragen aufgeworfen haben. In Bezug auf die Lösungen können grundsätzlich zwei verschiedene Ansätze verfolgt werden. Zum einen können Lösungen dahingehend erarbeitet werden, dass sie dem Tagger zu Gute kommen. Zum anderen kann das manuelle Tagging von Referenzkorpora durch solche Kriterien erleichtert werden.

Unsere Kriterien sollen als Leitfaden für die manuelle Annotation dienen, damit Tags auf linguistischer Basis sicher zugeordnet werden können und die Intuition weitgehend ausgeschaltet wird. Mit einem manuell konsistent getaggten Korpus

sollte der Grundstein für Verbesserungen des automatischen Taggings gelegt sein, denn linguistische Überlegungen zur automatischen Wortarten-Annotation sind die Basis, auf der Verfeinerungen am Tagger selbst vorgenommen werden, der sich nicht auf Intuition stützen kann.

3 Exemplarische Probleme

3.1 Adjektiv (ADJD) oder Adverb (ADV)

Ein Problem in der Klassifizierung von Wortarten ist die Unterscheidung zwischen Adjektiven und Adverbien.

Im STTS werden nur reine Adverbien wie *hier, gestern, darum* als solche getaggt (Tag: ADV), also „nicht von Adjektiven abgeleitete, nicht flektierbare Modifizierer von Verben, Adjektiven, Adverbien und ganzen Sätzen“ (Schiller/Thielen 1999:56). Adverbial gebrauchte Adjektive hingegen, z.B. *er fährt schnell*, werden als adverbiales Adjektiv (Tag: ADJD) gekennzeichnet.

Für das Tagging scheint diese Vorgehensweise von Vorteil, da die Bestimmung relativ eindeutig ist. Diese Einteilung ist rein kategorial orientiert, denn alles, was sich wie ein Adjektiv flektieren lässt, wird auch als solches behandelt, z.B.:

(1) *Peter lernt fleißig*/ADJD.
(2) *Willi ist ein fleißiges*/ADJA *Bienchen*.

In (1) handelt es sich um ein adverbial gebrauchtes, in (2) um ein attributiv gebrauchtes Adjektiv.

3.1.1 Distributionelle vs. kategoriale Sicht

Diese Einordnung ist nicht unumstritten, denn es gibt im Deutschen durchaus Überlegungen, auch Adjektive bei den Adverbien einzuordnen und zwar abhängig von ihrer Funktion im Satz. Dass semantisch gesehen zwischen Adjektiv und Adverb ein Zusammenhang besteht, weil beide „ihre Kern-Konstituente hinsichtlich bestimmter Eigenschaften modifizieren“ (Bußmann 2002:48) macht eine eindeutige Einteilung noch schwieriger:

(3) *lesbar schreiben*
(4) *lesbare Schrift*

Am Sprachwissenschaftlichen Institut der Ruhr-Universität Bochum wurde bereits versucht, durch ein stärkeres Gewicht auf die syntaktische Distribution der Adjektive und eine entsprechende Kennzeichnung von de-adjektivischen Adverbien günstigere Taggingergebnisse zu erzielen. Dies ließe sich dadurch rechtfertigen, dass Adjektive in Adverbstellung (de-adjektivisches Adverb) die gleiche Verteilung aufweisen wie echte Adverbien, z.B.:

(5) *Er schläft hier*. – Adverb
(6) *Er schläft lange*. – de-adjektivisches Adverb

Aus distributioneller Sicht würde also ein Tag genügen. Allerdings würde diese Lösung dazu führen, dass deadjektivische Adverbien als prädikativ verwendete Adjektive identifiziert würden:

(7) *Er isst schnell*. – de-adjektivisches Adverb
(8) *Er ist schnell*. – prädikatives Adjektiv

Der Versuch, Adjektive genauer einzuordnen, führte zu einer Einteilung in vier Stufen, die den Bezug auf Nomen und Verb zur Grundlage hatten. So sollte Stufe 1 abweichend vom STTS als ADV getaggt werden, wenn kein Bezug zum Nomen vorlag, Stufe 2 sollte starken Bezug zu Verb wie Nomen haben und damit als ADJD getaggt werden, Stufe 3 sollte eher auf das Nomen bezogen sein (ADJD) und Stufe 4 keinerlei Bezug zeigen (ADJA) (Stufe 2-4 wie im STTS vorgesehen):

(9) *Dies versteht er schwerlich*/ADV. – Stufe 1
(10) *Der Laster wiegt schwer*/ADJD. – Stufe 2
(11) *Er ist schwer*/ADJD. – Stufe 3
(12) *Der schwere*/ADJA *Mann trat ein*. – Stufe 4

Im Wesentlichen lief diese Einteilung also auf ähnliche Unterscheidungen hinaus, wie sie das STTS macht. Allerdings wurde hier zunächst die zusätzliche Überlegung einbezogen, alle Adjektive als ADV zu annotieren, die von ihrer Stellung her einem Adverb entsprächen.

Auch Helbig/Buscha (1997:337f.) unterstützen diese Einteilung, indem sie adjektivische Formen in adverbialer Stellung den Adverbien zurechnen und sie als ‚Adjektivadverbien' bezeichnen. Diese unterscheiden sich von reinen Adverbien insofern, als sie voll und nicht nur beschränkt graduierbar sind, z.B.:

(13) *Er arbeitet hier*. – echtes Adverb
(14) *Er arbeitet fleißiger*. – Adjektivadverb

Allerdings spricht das Argument der Graduierbarkeit auch dafür, dass es sich um ein Adjektiv handelt, denn die Steigerbarkeit ist ein wesentliches Merkmal von Adjektiven (Eisenberg 2001:183).

Die ‚Adjektivadverbien', auch ‚adjektivische Adverbien' genannt, beziehen sich meist auf einen verbalen Vorgang (Admoni 1970:198f.), wie in

(15) *Der Lehrer spricht schnell/schneller*.

Bezieht sich ein Adjektiv also auf einen verbalen Vorgang, so wird es wie im Beispiel (15) als Adverb betrachtet. Diese Form ist dann allerdings ambig zur prädikativen Verwendung:

(16) *Der Lehrer ist schnell/schneller*.

Trotz Formengleichheit ordnen Helbig/Buscha (1997:337) in (16) das prädikativ verwendete Wort den Adjektiven zu, da hier nicht das Verb, sondern ein Nomen modifiziert wird.

Bergenholtz und Schaeder gehen noch weiter. Sie rechnen die unflektierte Form sowohl in prädikativer wie in adverbialer Verwendung den Adverbien zu, mit der Begründung, nur die tatsächlich flektierten Formen gehörten zu den Adjektiven. Damit geben sie dem Kriterium der syntaktischen Funktion die größte Priorität bei der Einteilung (Bergenholtz/Schaeder 1977:108ff). Hier stoßen wir jedoch auf das Problem, dass Adverbien morphologisch gesehen grundsätzlich unflektierbar sein sollen (Bergenholtz/Schaeder 1977:109).

Bei der maschinellen Anwendung dieser Einteilung zeigte sich aber, dass eine Kennzeichnung von Adjektiven in Adverbstellung als adverbial gebrauchtes Adjektiv die Teffsicherheit bei Adverbien und Adjektiven nicht verbessern konnte, das Tagging also keine besseren Ergebnisse liefert als bei einer Kennzeichnung nur der reinen Adverbien als ADV.

3.1.2 Adverb als Nebenwort

Einer Zuordnung von Adjektiven zu den Adverbien wird besonders auch von Eisenberg (1999) widersprochen.

Zum einen hält er fest, dass sich Adverbien durchaus nicht nur auf Verben beziehen. Ganz im Gegenteil: Es herrscht allgemein keine Einigkeit darüber, was genau der Begriff ‚Adverb' bedeutet. In der Diskussion wird er sowohl als ‚Nebenwort' gedeutet, das in „dienender Funktion" steht, aber auch als „zum Verb tretend", wobei Verb im weiter gefassten Sinne von ‚Wort' und nicht im heutigen Sinne von Verb als Tätigkeitswort verstanden wird (Eisenberg 1999:205 nach Lyons 1980:331).

Würde man also die Bezeichnung ‚Adverb' für adverbiale Adjektive einführen, weil sie auf das Verb in unserem heutigen, engeren Sinn, bezogen sind, müssten, um Missverständnisse zu vermeiden, die reinen Adverbien umbenannt werden, da diese ja nun gerade nicht nur auf Verben im eigentlichen Sinn bezogen sind (Eisenberg 1999:220).

3.1.3 Flektierbarkeit

Auch die Zuordnung aufgrund der Unflektiertheit von Adjektiven in adverbialer Stellung lässt Eisenberg nicht gelten. Der wesentliche Unterschied zwischen Adverb und Adjektiv ist nicht die Frage, ob ein Adjektiv in einer bestimmten

Position flektiert ist, sondern ob es grundsätzlich flektiert werden kann (Eisenberg 1999:220).

Adverbien haben einelementige Paradigmen, die deshalb als nichtflektierbar bezeichnet werden. Die Kurzform des Adjektivs, wie sie in prädikativer und adverbialer Position erscheint, ist unflektiert, aber das Adjektiv an sich ist grundsätzlich flektierbar, z.B.

(17) *klein, kleine, kleiner, kleines* – flektierte Formen
(18) *Sie ist klein*. – prädikative Form
(19) *Sie schreibt klein*. – adverbiale Form

Wer Nichtflektiertheit in einer bestimmten Funktion zum ausschlaggebenden Kriterium für eine Zuweisung zu den Adverbien macht, muss auch das prädikative Adjektiv zu den Adverbien zählen (wie es Bergenholtz/Schaeder (1977) konsequenterweise tun).

Jedoch können fast alle Adjektive adverbial verwendet werden. Bezeichnete man sie in dieser Verwendung als Adverbien, wären die entsprechenden Adjektive lediglich Homonyme zu einer Teilklasse der Adverbien, und eine umfangreiche offene Kategorie, nämlich die der Adjektive, hätte ihre Eigenständigkeit damit verloren (Eisenberg 1999:221). Vielmehr scheint sinnvoll, die Adjektive nach ihrer syntaktischen Funktion zu beleuchten und den Wortklassencharakter (d.h. Adjektiv) von der Satzgliedfunktion (d.h. Prädikativ oder Adverbial) getrennt zu betrachten, also davon auszugehen, dass Adjektive in bestimmten syntaktischen Positionen unflektiert verwendet werden (Heidolph 1981:622).

Adjektive beschreiben Eigenschaften von etwas, und zwar sowohl in prädikativer als auch in attributiver Verwendung:

(20) *der kluge Professor*
(21) *Die Professorin ist klug*.

Wenn auch beide Verwendungsweisen syntaktisch verschieden voneinander sind, so sind sie doch offenbar semantisch gleich.

Auch der Nominalisierungstest zeigt, dass dasselbe Adjektiv genommen wird, um sowohl den vom Verb bezeichneten Vorgang näher zu beschreiben als auch den vom Nomen ausgedrückten Prozess:

(22) *Karl befragt den Minister sorgfältig*.
(23) *Karls sorgfältige Befragung des Ministers*

Dies beweist zwar nicht, dass beide Ausdrücke (adverbiales und attributives Adjektiv) zwingend derselben Kategorie angehören, doch ist es ein Beleg für eine enge Beziehung in systematischer Hinsicht (Eisenberg 1999:222).

Eisenberg ordnet die fraglichen Ausdrücke also der Kategorie der Adjektive zu. Er nennt sie ‚adverbiale Adjektive' im Gegensatz zu ‚adjektivischen Adverbien',

und betont damit die Zuordnung zu den Adjektiven, die lediglich auf adverbiale Art und Weise verwendet werden können.

3.1.4 Ausnahmen

Die unterschiedliche Semantik von Adjektiven wie *natürlich*, *rund* oder *ganz* kann dabei jedoch, wie auch im STTS argumentiert, nicht außer Acht gelassen werden.

(24) *Hast du das Buch ganz*/ADJD *gelesen*?
(25) *Das hast du ganz*/ADV *gut gemacht*.
(26) *Er ist schnell*/ADJD.
(27) *Er läuft schnell*/ADJD.
(28) *Er ist ein schneller*/ADJA *Läufer*.

In den Beispielen (26) bis (28) handelt es sich immer um dieselbe Bedeutung des Adjektivs *schnell* i.S.v. *Geschwindigkeit*, in (25) und (24) wird der Bedeutungsunterschied deutlich zwischen *ganz* i.S.v. *ziemlich* gegenüber *ganz* i.S.v. *komplett*.

Bei *ganz* handelt es sich also nicht nur um ein Adjektiv wie in Beispiel (24), sondern auch um ein homonymes Adverb (Beispiel (25)), das selbstverständlich als Adverb getaggt wird.

3.1.5 Zwischenfazit

In dieser Frage folgen wir Eisenbergs Argumenten und schlagen vor, beim Taggen dem STTS zu folgen. Damit werden nur reine Adverbien als solche mit dem Tag ADV versehen, adverbiale Adjektive jedoch als ADJD, also als Adjektiv in adverbialer Position. Dabei kann es allerdings wie beschrieben zu stellungsbedingten Ambiguitäten kommen, da bei dieser Klassifizierung ADV und ADJD in derselben syntaktischen Position und Funktion vorkommen können (Beispiel (27))

Der Übersichtlichkeit halber noch einmal die wichtigsten Argumente, die in Zweifelsfällen zur Entscheidung herangezogen werden:

wenn flektierbar → Adjektiv
wenn voll graduierbar → Adjektiv
wenn nominalisierbar → Adjektiv

3.2 Adjektiv (ADJD) oder Partizip Perfekt (VVPP)

Ähnliche Schwierigkeiten, ein Wort in eine Wortklasse einzuordnen, zeigen sich ganz besonders deutlich bei dem Versuch einer Abgrenzung von Adjektiven (ADJD) einerseits und Perfektpartizipien (VVPP) andererseits.

Neben eindeutigen Fällen von lexikalisierten Partizipien wie in (1) können Zweifel aufkommen bei Sätzen wie in (2), (3) und (4).[16]

(1) *Das Spiel ist verloren.*
(2) *Und schon ist scheinbar der Nachweis erbracht, dass die Kriminalität mit der Wohlstandsgesellschaft Schweiz konfrontiert ist.*
(3) *Je nach Ausgestaltung der Funktion wäre sie den Bundesräten [...] direkt unterstellt.*
(4) *Das Haus ist gebaut.*

Hier entscheidet bei der manuellen Annotation häufiger die Intuition, was hinsichtlich einer konsistenten Auflösung solcher Ambiguitäten natürlich nicht erstrebenswert ist.

Um das Problem zu verdeutlichen, stellen wir zunächst einige Positionen aus der Literatur vor und diskutieren diese kurz.

3.2.1 Unterscheidung von Adjektiv und Partizip Perfekt

Eine Abgrenzung von Partizip Perfekt und Adjektiv stellt sich deswegen so schwierig dar, weil eine Unterscheidung allgemein nur durch die Bedeutung im Kontext, nicht aber durch starre Regeln möglich erscheint (Helbig/Buscha 1997:175ff.).

Auch Eisenberg sieht es als schwierig an, Zustandspassiv und formgleiches Adjektiv zu unterscheiden. Oft ist das Partizip als Adjektiv lexikalisiert, dennoch gleichwohl auch als verbale Form anzusehen (Eisenberg 1999:132).

Je nach Verbart gibt es Abstufungen im Gebrauch: Zur Unterscheidung wird bei ihm u.a. auf die agentive Grundbedeutung des Verbs verwiesen, und zwar auch bei Verben mit nichtagentiver Lesart wie beispielsweise:

(5) *Die beiden Zimmer werden/sind durch einen Gang verbunden*/ADJD.[17]
(6) *Die beiden Zimmer werden vom Maurer verbunden*/VVPP.

Daraus folgt, dass die Präpositionalgruppe mit *durch* oder *mit* auf eine deagentivische Lesart hinweist, eine mit *von* eingeleitete Präpositionalgruppe auf eine agentivische. Im Fall der deagentivischen Lesart wird der Tag ADJD (5), bei agentivischer Lesart VVPP (6) zugewiesen (vgl. Eisenberg 1999:132).

---

16 Die Sätze sind dem NZZ-Korpus entnommen.
17 Das Beispiel enstammt Eisenberg (1999:131).

3.2.2 Kriterien

Um die jeweilige Bedeutung so wenig subjektiv und so eindeutig wie möglich bestimmen zu können, haben wir uns schließlich entschieden, bei dieser Problematik weitgehend dem STTS zu folgen. Das STTS legt Folgendes fest (Schiller/Thielen 1999:24ff.):

- Partizipien in adverbialer Stellung: ADJD

(7) *Er spielt gekonnt*/ADJD.
(8) *Die Mittel wurden gezielt*/ADJD *eingesetzt*.

- Attributiv oder modifizierend verwendete Partizipien werden als ADJD getaggt, ebenso Partizipien nach *wie* und *als*.

(9) *Er macht es wie geplant*/ADJD.
(10) *Die geplante*/ADJA *Sache*

- Lexikalisierte Partizipien: Problemfälle sind Passivpartizipien (Vorgangspassiv: mit *werden*, Zustandspassiv: mit *sein*), die je nach Kontext auch eine adjektivische Lesart zulassen (z.B. *verrückt*: Patiens = [+BELEBT] => ADJD).

(11) *Der Tisch wird verrückt*/VVPP.
(12) aber: *Der alte Mann wird verrückt*/ADJD.

Als Kriterien für die Disambiguierung der Kopulakonstruktionen mit ADJD und des Verlaufspassivs mit VVPP gibt das STTS (Schiller/Thielen 1999:24) die folgenden drei Tests an:

- Wenn der Satz mit gleicher Semantik ins Aktiv gesetzt werden kann: VVPP.

(13) *Wo Menschen selbst betroffen*/VVPP *seien*
→ *Wo es die Menschen selbst betrifft*

- Wenn es eine *von*-PP oder ähnliche PP gibt, die auf Semantik als Verb hinweist: VVPP.

(14) *Ihm wurden die Haare kurz geschnitten*/VVPP.
→ *Ihm wurden die Haare (vom Friseur) kurz geschnitten*/VVPP.

- Wenn eine Ersetzung durch ein semantisch nahes Adjektiv möglich ist: ADJD.

(15) *Der Tisch wird verrückt*/VVPP.
→ **Der Tisch wird wahnsinnig.*
→ keine Ersetzung möglich, somit kein ADJD

(16) *Der alte Mann wird verrückt*/ADJD.
→ *Der alte Mann wird wahnsinnig.*
→ Ersetzung möglich, also ADJD.

Da die drei Kriterien nicht immer ausreichen, muss in Zweifelsfällen noch der satzübergreifende Kontext herangezogen werden:

- Wenn das Perfekt mit *haben* einen Sinn ergibt: VVPP.

(17) *Der Schrank wird verrückt*/VVPP.
→ *Er hat den Schrank verrückt*.

- Wenn das Verb in seiner Grundform nicht mehr existiert: ADJD.

(18) *mit jemandem verwandt*/ADJD *sein*

Zwei weitere, syntaktische Kriterien vervollständigen die Tests zur Abgrenzung von Adjektiv und Partizip Perfekt:

- Wenn sich die Satzkonstruktion mit *worden* bei gleich bleibender Semantik ergänzen lässt: VVPP.

(19) *Der Tisch wird verrückt*/VVPP.
→ *Der Tisch ist verrückt worden*.

(20) *Der Mann wird verrückt/ADJD*.
→ **Der Mann ist verrückt worden*.

- Wenn das Partizip mit *un-* negierbar ist: ADJD.

(21) *Der Mann hat die Haare geschnitten*/ADJD.
→ *Der Mann hat die Haare ungeschnitten*.

(22) *Der Friseur hat die Haare geschnitten*/VVPP.
→ **Der Friseur hat die Haare ungeschnitten*.

3.2.3 Zwischenfazit

Diese Kriterien erleichtern eine manuelle Annotation, bei der das subjektive Urteil auf ein Minimum reduziert werden kann. Für den Tagger stellen sie allerdings keine direkte Hilfe dar, da ADJD und VVPP nicht komplementär verteilt sind und

daher nicht aus ihrer Distribution hervorgeht, um welchen Tag es sich handeln muss.

3.3 Eigenname (NE) oder Nomen (NN)

Die Substantive bilden mit mindestens 60% des Wortschatzes die größte Klasse der Wörter des Deutschen und sind in mehrere Subklassen mit jeweils charakteristischen Eigenheiten unterteilt (Eisenberg 1999:135).

Das STTS unterteilt die Klasse der Nomina jedoch nur in Gattungsnamen (‚nomina appellativa') und Eigennamen (‚nomina propria'). Eine weitere semantische Einteilung, z.B. der Konkreta in Stoffsubstantive oder in einzelne Subklassen der Abstrakta wird dort nicht vorgenommen, entsprechende Nomina werden unter die Appellativa subsumiert.

Diese grobe Einteilung genügt für das Tagging zunächst, jedoch wirft die Abgrenzung von Eigennamen gegenüber Appellativa einige Fragen auf.

Ein besonderes Problem sind mehrteilige Eigennamen wie *das Schwarze Meer* oder *Paulchen Panther*. Bei der automatischen Verarbeitung werden sie durch die Tokenisierung typischerweise auseinander gerissen, so dass sie anschließend einzeln getaggt werden. Das STTS schlägt hier vor, die einzelnen Teile entsprechend ihrer Wortform zu taggen, also z.B.

(1) *Das*/ART *Schwarze*/ADJA *Meer*/NN
(2) *Paulchen*/NE *Panther*/NE

Während in (2) weiterhin der Eigenname erkennbar ist, geht dies aus (1) nicht mehr hervor. Das ist besonders problematisch, weil das STTS bei fremdsprachlichen Eigennamen anders vorgeht[18]:

(3) *New*/NE *York*/NE statt *New*/ADJA *York*/NE

Aber auch eine Annotation wie in (4), aus der zu entnehmen wäre, dass der gesamte Ausdruck als Eigenname oder mehrere adjazente Eigennamen gilt, erscheint nicht einleuchtend.

(4) *Das/*NE *Schwarze/*NE *Meer/*NE

Eine weitere Unsicherheit tritt bei der Zuordnung von Akronymen auf. Sie werden weitgehend als NE getaggt, doch sind sicher nicht alle Akronyme, wie die Beispiele *NATO, EDV, WM* und *C&A* zeigen, tatsächlich als Eigennamen anzusehen.

Des Weiteren treten Einordnungsschwierigkeiten bei Wörtern wie *Bundesrepublik, die Grünen* oder *Sowjetrepubliken* auf sowie bei Markennamen**.**

[18] Genauere Ausführungen dazu siehe Teil FM, Kap. 3.6.

3.3.1 Eigennamen versus Gattungsnamen

Bei den Eigennamen handelt es sich um eine Klasse von Nomen, die einzelne Objekte und Sachverhalte im Kontext eindeutig und mit direkter Referenz, also ohne Zwischenschaltung deiktischer Gesten, identifizieren.

Gattungsnamen hingegen dienen zur Bezeichnung ganzer Klassen von Gegenständen und Sachverhalten sowie ihrer einzelnen Vertreter (Bußmann 2002:185,234).

(5) *Diederich*/NE *Heßling*/NE *war ein weiches Kind*.
(6) *Der Tisch*/NN *ist rund*.

Ein Eigenname wie in (5) bezeichnet genau ein Individuum. Seine Definitheit ist fest und braucht nicht z.B. mittels Artikel hergestellt zu werden. Das Individuum wird innerhalb einer gekennzeichneten Klasse (hier: der männlichen Wesen) identifiziert (Eisenberg 1999:161f.).

Ein Gattungsname wie in (6) bezeichnet keine Entität, sondern z.B. einen Vertreter einer Klasse von Gegenständen, der erst durch Hinzufügen des bestimmten Artikels definit wird.

3.3.2 Zur Frage der Bedeutung

Ob ein Eigenname eine eigene Bedeutung hat, wird dabei durchaus kontrovers diskutiert.

Laut Eisenberg ist er ein identifizierendes Etikett ohne semantische Implikation, obgleich er noch einige Begriffsreste enthält (so z.B. ein Männername als Bezeichnung für ein männliches Wesen) (2004:350). Aber er ist semantisch leer; er bezeichnet ein Individuum als Ganzes, ohne ihm bestimmte Eigenschaften zuzuordnen, denn ein Eigenname sagt nichts darüber aus, ob z.B. die Person dumm, groß oder hungrig ist.

Dennoch sind Eigennamen nicht immer völlig unmotiviert. Namen wie *Bäcker* oder *Schuster* sind aus Berufsbezeichnungen entstanden und tragen bzw. trugen daher auch einen Anteil an Charakterisierung in sich, in der Regel ist eine solche allgemeine Bedeutung eines Namenswortes für den Bezug zum Namensträger heutzutage jedoch irrelevant (Fleischer 1964:370).

Danach sind Namen von Parteien ein Grenzfall: Eine Bezeichnung wie *Bündnis 90/Die Grünen* benennt nicht nur, sondern charakterisiert eindeutig, welche Gesinnung in Verbindung mit diesem Namen zu erwarten ist. Damit wäre die jeweilige Bezeichnung eher dem Gattungsnamen gleichzusetzen, der immer auch eine Charakterisierung des bezeichneten Objektes in sich trägt und damit eine unmittelbare Information gibt, während der Eigenname nur ein Schlüssel zu dieser Information ist:

(7) *Wir haben die Löwen*/NN *im Zoo besucht*.
(8) *Herr Löwe*/NE *wohnt nebenan*.

Über die Eigenschaften des *Herrn Löwe* in (8) gibt der Name keine Auskunft, während *die Löwen* in (7) bereits auf eine bestimmte Bedeutung, nämlich i.S.v. Raubkatze o.Ä., abzielt (Fleischer 1964:369ff.).

Ein Eigenname benennt demnach ein Individuum, ohne eine eigene Bedeutung zu haben, während ein Appellativum eine Bedeutung hat und gleichzeitig Benennung ist (Knobloch 1992:454.).

3.3.3 Zur Frage der Abgrenzung

Die Abgrenzung von Gattungsnamen und Eigennamen ist dabei jedoch nicht so unproblematisch, wie es auf den ersten Blick erscheinen mag, denn es gibt zahlreiche Überlappungen. Besonders deutlich wird dies bei Produktnamen:

(9) *Ein Papiertaschentuch von Tempo*/NE
(10) *Hast du mal ein Tempo*/NN *für mich*?

Eigennamen wie in (9) können sich also im Laufe der Zeit mit wachsender Verwendung auch als Gattungsnamen wie in (10) etablieren.

Ein weiteres Problem sind monoreferentielle Ausdrücke, die sich auf genau ein Objekt beziehen, wie etwa

(11) *Erde, Himmel, Mond*

Diese werden i.d.R. nicht zu den Eigennamen gezählt, sondern als Monosemantika verstanden, deren Begrenzung in der Zahl der möglichen bezeichneten Objekte liegt. Wird ein solcher Ausdruck jedoch wie z.B. Mond auch in seiner fachspezifischen Bedeutung als ‚Trabant eines Planeten' verwendet, ließe sich in Abgrenzung dazu der Name Mond für den Erdmond als Eigenname verstehen (Fleischer 1964:372).

In diesem Bereich ist auch die Abgrenzung zu definiten Kennzeichnungen nicht ganz einfach.

(12) *Die Französische Revolution, die Lüneburger Heide, die Hebriden*

Alle Ausdrücke in (12) beziehen sich wie die zuvor genannten Monosemantika auf Entitäten, die es genau einmal gibt, doch sind sie von ihrer Form her weder Appellativa noch Eigennamen, sondern definite Kennzeichnungen. Dabei handelt es sich um Ausdrücke, die mithilfe eines bestimmten Artikels und eines Prädikats, das auf genau eine Entität zutrifft, ein bestimmtes Objekt bezeichnen (Bußmann 2002:341). Sprachliche Form und Bezeichnetes sind zwar hier genauso fest verbunden wie dies bei Eigennamen der Fall ist, doch wird mit dem Ausdruck deutlich mehr in Verbindung gebracht als nur ein Begriffsrest (Eisenberg 1999:162).

Trotz dieses Wechsels zwischen den Klassen gibt es einen wichtigen Unterschied zwischen Appellativa und Eigennamen: Letztere werden einmal, oft durch einen Akt der Namensgebung, zugeordnet und bleiben dann für immer bestehen. Dies gilt auch für Namen, die auf den ersten Blick charakterisierend wirken: Häuptling *Adlerauge* behält seinen Namen auch dann, wenn er nicht mehr so gut sehen kann. Namen sind also nicht völlig unmotiviert und können sogar einen recht hohen Symbolwert haben, wie etwa Neugründungen von Städten (Bsp. (13)) und die (aus namenstheoretischer Sicht ungewöhnlichen) Umbenennungen von Städten (Bsp. (14)) und Plätzen in der ehemaligen DDR belegen (Knobloch 1992:453f.):

(13) *Eisenhüttenstadt*
(14) *Chemnitz → Karl-Marx-Stadt*

Es lässt sich also feststellen, dass Appellativa auch als Eigennamen verwendet werden können und auch die umgekehrte Richtung möglich ist.

3.3.4 Mögliche Kriterien der Unterscheidung

3.3.4.1 Großschreibung

Eigennamen können anhand ihrer Schreibung der Klasse der Substantive zugeordnet werden, denn für die Schreibung von Eigennamen hat sich historisch die Großschreibung für alle ihre Bestandteile, mit Ausnahme von Artikel, Präposition und Konjunktion, entwickelt. Dies gilt aber nur zum Teil, denn dass der Status des Eigennamens untrennbar mit der Kategorie des Substantivs verbunden sei, wie es Knobloch (1992:459) behauptet, wird durch die schon erwähnten mehrteiligen Namen widerlegt.

Bei manchen Namen gehört der Artikel (wie in (15)) als fester Bestandteil dazu und wird ebenfalls groß geschrieben, es sei denn, die flektierte Form weicht von der Grundform ab (vgl. (16)):

(15) *Ich lese Die Zeit*.
(16) *Ich lese in der Zeit*.

Die Großschreibung eines zugehörigen Artikels gibt also nur begrenzt Hinweise, ob es sich um einen zusammengesetzten Eigennamen handelt oder nicht (Eisenberg 2004:343ff.).

3.3.4.2 Artikelgebrauch und Flexion

Als eine andere Möglichkeit, Eigenname von Appellativum zu unterscheiden, gilt häufig die Verwendung des Artikels, der bei Eigennamen nur beschränkt einsetzbar ist. Ein Großteil der Namen wird artikellos verwendet, doch gibt es zahlreiche Ausnahmen:

(17) *Ich lese Die Zeit*.
(18) *Ich liebe die Schweiz*.
(19) *Ich fahre in die Alpen*.

In (17) wird der Artikel als Bestandteil des Namens aufgefasst und somit großgeschrieben, in (18) und (19) gilt er nicht als Bestandteil im Sinne der Großschreibung.

Im Gegensatz zu Appellativa stehen Eigennamen in der Regel ohne Artikel:

(20) *Istanbul ist prächtig*.
(21) **Tisch ist hoch*.

Eigennamen mit sächsischem Genitiv können dem Kernsubstantiv einer Nominalgruppe vorausgehen, bei Appellativa ist dies erst nach Hinzufügen eines Artikels möglich:

(22) *Emils Mutter*
(23) **Mannes Mutter*
(24) *des Mannes Mutter*

Der Eigenname im Genitiv hat dabei Kopffunktion und tritt an dieselbe Stelle, die ein Artikel einnehmen würde. Allerdings gilt das Gesagte uneingeschränkt nur für Eigennamen ohne dazugehörigen Artikel und nicht für Eigennamen, die bereits einen Artikel enthalten, wie *die Eifel*:

(25) *?Der Eifel Vulkane*

Innerhalb einer erweiterten Nominalgruppe kann das Genitiv-*s* beim Eigennamen nicht mehr stehen, wenn der Eigenname mit Artikel steht:

(26) *Die Stücke Brechts*
(27) **Die Stücke des Brechts*
(28) *Der Zusammenbruch des Hauses*

Bei Gattungsnamen wie in (28) ist dies jedoch obligatorisch (Eisenberg 1999:160, 245ff.).

Bei den Eigennamen ist zudem die Tendenz zu beobachten, dass immer häufiger die Kasusendung des Genitivs ganz vermieden wird, während sie bei Appellativa, in deren Flexionsparadigma das Genitiv-s vorkommt, weiterhin üblich ist (Fleischer 1964:376):

(29) *Die Verfassung eines demokratischen Deutschland*/NE.
(30) *Die AGB eines großen Unternehmens*/NN.

Ein weiteres Kriterium stellt die Flexion innerhalb der Nominalphrase dar. Tendenziell bleiben Eigennamen unverändert und werden nicht flektiert, vgl. Beispiel (31).[19] Dies ist aber durchaus nicht die Regel, wie Beispiel (32) zeigt.

(31) *Rote Armee Fraktion → in der Rote Armee Fraktion*
(32) *Schwäbische Alb → auf der Schwäbischen Alb*

3.3.4.3 Pluralbildung

Pluralformen gibt es bei den Propria nur in geringem Umfang, da i.d.R. nur genau eine Entität mit dem Eigennamen bezeichnet wird. Gibt es mehrere Entitäten mit gleichem Namen und soll auf alle gleichzeitig referiert werden, weicht die Pluralform hier von der usuellen Bildung ab:

(33) *Die Bocks*/NE – Mitglieder der Familie Bock
(34) *Die Böcke*/NN – die Tiere

In (33) wird deutlich, dass nicht auf eine Menge von Entitäten Bezug genommen wird, sondern auf mehrere eigenständige Entitäten, die zufällig denselben Namen tragen.

Eine Ausnahme bilden hier natürlich solche Eigennamen, die regelmäßig mit Pluralartikel vorkommen, wie z.B. *die Niederlande*.

3.3.4.4 Übersetzbarkeit

Bei Eigennamen lässt sich im allgemeinen feststellen, dass sie sich nicht oder nur schwer übersetzen lassen und daher in der Regel auch nicht übersetzt werden. Sind die Namen etymologisch durchschaubar, wie in (35) kommt es gelegentlich zu Eins-zu-Eins-Übersetzungen.

(35) *Schwarzwald – Black Forest*

In manchen Fällen werden Eigennamen lautlich an die Zielsprache angepasst:

(36) *Nürnberg - Nuremberg*

Bei einigen Namen führt die Frage der Übersetzbarkeit jedoch nicht weiter. Laut STTS würde beispielsweise

(37) *Ostsee*/NN

getaggt, weil *See* kein Eigenname ist und Determinativkomposita nach ihrem Zweitglied eingeordnet werden. Für diesen Standpunkt spräche, dass *Ostsee* z.B. regelmäßig als *Baltic Sea* ins Englische übersetzt wird und es auch in anderen

[19] Diese Form wird in der gesprochenen Sprache allerdings häufig als markiert empfunden.

Sprachen eigene Namen gibt. Außerdem ist die Bezeichnung *Ostsee* durchaus mit bestimmten Vorstellungen verbunden, so dass der Name nicht lediglich identifiziert, sondern auch charakterisiert. Da jedoch nur eine einzige Entität mit diesem Namen existiert und damit die Identifizierung immer hundertprozentig gegeben ist, lässt sich auch die Kennzeichnung als NE rechtfertigen.

3.3.4.5 Prädikation

Eigennamen sind im Gegensatz zu anderen Nomina in Prädikationen auf die Subjektposition beschränkt und nicht durch Prädikation mit dem Nominatum verbunden. Sie können also kein Prädikativ mit den Kopulaverben bilden (mit Ausnahme von *heißen* oder *sein* in der Konnotation von *heißen* (42)) (Lyons 1971:343ff.):

(38) *Tschetschenien und Tadschikistan waren Sowjetrepubliken*/NN.
(39) **Tadschikistan war Sowjetrepublik*/NE.
(40) *Er wird Lehrer*/NN.
(41) **Er bleibt Sokrates*/NE.
(42) *Die beiden dort sind die Bocks*/NE.

3.3.5 Zwischenfazit

Die obigen Ausführungen machen deutlich, dass die Abgrenzung von Nomina propria und Appellativa nicht immer einfach zu treffen ist. Im Zweifel müssen die entsprechenden Nomina genau auf ihre Eigenschaften überprüft werden. Dabei helfen verschiedene Kriterien, die richtige Zuordnung zu treffen, wenn auch häufig nicht alle Unsicherheiten dadurch beseitigt werden können.

Der Übersichtlichkeit halber sind die Kriterien, die für den Tag NE sprechen, hier noch einmal aufgelistet:

Name semantisch eher unmotiviert?
Zugehöriger Artikel großgeschrieben?
Nomen völlig ohne Artikel verwendbar?
Flexionsendung bei artikellosem Gebrauch?
Sächsischer Genitiv ohne Artikel vor Kernsubstantiv einer Nominalgruppe?
Sächsischer Genitiv nicht nach Artikel?
Genitiv-Endung weglassbar?
Keine regulären Pluralformen?
Schwer übersetzbar?
Prädikation nicht möglich?

3.4 Abtrennbarer Verbteil (PTKVZ)

Der Tag PTKVZ nimmt bei der Wortarten-Annotation eine besondere Rolle ein, da er keine Wortart bezeichnet, sondern den Teil eines Verbs, der unter bestimmten strukturellen Gegebenheiten vom Verb abgetrennt auftritt. Je nachdem, um welche Art von abgetrenntem Verbzusatz es sich handelt, könnte er verschiedenen Wortarten zugeordnet werden (Beispiele (1) bis (5)):

(1) *mitkommen*
(2) *krankschreiben*
(3) *weggehen*
(4) *teilnehmen*
(5) *sitzenbleiben*

Die Fragestellung lautet daher: Sollte beim Wortarten-Tagging nicht konsequenterweise ausschließlich nach Wortarten getaggt werden? In diesem Fall wären abgetrennte Verbzusätze nach ihrer jeweiligen Kategorie zu annotieren, nicht nach ihrer Eigenschaft als Verbzusatz.

Zunächst soll geklärt werden, was ein Verbzusatz sein kann.

> Allgemein handelt es sich bei Verbzusätzen um Beziehungswörter (Partikeln), die feste und unfeste Zusammensetzungen mit Verben bilden können. (Doll 1967:4)

Eine Partikel, im STTS durch den Tag PTKVZ gekennzeichnet, kann die Form eines (teilweise verblassten) Substantivs haben (Beispiele in (6), einer Präposition (7), eines Adjektivs (8), eines Adverbs (9) oder – nach alter Rechtschreibung – eines Verbs (10)).

(6) *teilnehmen*, *bergsteigen*
(7) *mitkommen*, *aufstehen*
(8) *krankschreiben*, *warmhalten*
(9) *weggehen*, *zusammenbinden*, *zurücklassen*
(10) *kennenlernen*, *sitzenbleiben*

3.4.1 Substantivische Partikel

Der komplizierteste Typ ist der mit substantivischem Erstglied und verbalem Zweitglied (6). Diese Verben lassen sich in drei Gruppen unterteilen: in untrennbare Verbindungen, in Verbindungen, die in bestimmten Fällen trennbar sind, und Rückbildungen aus echten Substantivkomposita, bei denen der substantivische Bestandteil kein Objekt sein kann. Die letzte Gruppe kann demnach bei der Diskussion um PTKVZ oder einen nominalen Tag vernachlässigt werden. Hierbei handelt es sich um Infinitive, die durch Rückbildung oder Konversion aus echten Substantivkomposita und nicht aus substantivischen Infinitiven entstanden sind (Beispiele (11), (12) und (13) (Eisenberg 2004:340)).

(11) *Strafversetzung → strafversetzen*
(12) *Bergsteiger → bergsteigen*
(13) *Bausparer → bausparen*

Die substantivischen Bestandteile dieser Verben können deshalb nicht Objekt sein, weil sie nicht syntaktisch vom verbalen Bestandteil abtrennbar sind. So stellt sich auch nicht die Frage nach dem entsprechenden Tag.

Die erste Gruppe, die untrennbaren Verbindungen, stellen aus demselben Grund kein Problem für das Tagging dar. Hierbei handelt es sich um Verben wie in (14), (15) und (16) (Eisenberg 2004:339).

(14) *brandmarken*
(15) *lobpreisen*
(16) *wetteifern*

Problematisch sind also nur die Verben der zweiten Gruppe, die Verbindungen, die in bestimmten Fällen trennbar sind.[20] Dies sind Verben wie in (17) und (18):

(17) *biertrinken*, *geldwaschen*
(18) *klavierspielen*, *radiohören*, *autofahren*

Die Verben in (17) werden laut Eisenberg (2004:339) nur im reinen Infinitiv zusammengeschrieben, ansonsten handelt es sich dabei um Fügungen aus Verb und direktem Objekt. Der verbale Infinitiv *biertrinken* sei eine Konversion aus dem substantivischen *das Biertrinken*, welches selbst ein Inkorporationsprodukt sei (2004: 339). Es wurde also ein Stoffsubstantiv in der Funktion eines direkten Objekts inkorporiert. Andere Formen als der Infinitiv seien immer Teil eines Syntagmas. Das spricht unserer Meinung nach für einen nominalen Tag für diese Gruppe von Verbpartikeln.

Im Gegensatz hierzu sieht Eisenberg für Beispiel (18), dass beispielsweise *Klavier* nicht die Funktion eines direkten Objekts habe, denn das Substantiv ist hier nicht erweiterbar (vgl. aber (25)). So ist (19) etwas anderes als (20) und es stellt sich für Verben wie in (19) die Frage nach dem passenden Tag.

(19) *Klavier spielen (klavierspielen)* bzw. *Auto fahren (autofahren)*
(20) *auf dem Klavier spielen* bzw. *mit dem Auto fahren*

Das direkte Objekt der Verben in (17) steht laut Eisenberg (2004:339) in vielen Satztypen in einer Position, in der auch Partikeln stehen. Zu klären ist nun, ob es sich dabei (oder möglicherweise bei den substantivischen Bestandteilen der Verben in (18)) um Partikeln handelt. Die folgende Übersicht soll dies deutlich machen:

[20] Nach der neuen deutschen Rechtschreibung werden diese Verbindungen getrennt geschrieben und die substantivische Partikel groß. Das Problem stellt sich also nur nach der alten Rechtschreibung.

(21) *weil er heute Bier trinkt*
(22) *weil er heute Klavier spielt*
(23) *weil er heute abreist*

Daraus geht hervor, dass sowohl die Nomina in (17) (also hier in (21)) als auch die in (18) (hier (22)) zwar distribuiert sind wie Partikeln (23), was sich allerdings nicht mit der Schreibung deckt.

Im Folgenden soll eine weitere Gruppe von Verben zum Vergleich herangezogen werden, die den Partikelverben am nächsten kommen (Eisenberg 2004:341). Es handelt sich hierbei um Verben wie in (24):

(24) *eislaufen*, *haltmachen*, *teilnehmen*, *kopfstehen*

Eisenberg (2004:341) argumentiert, dass *eis*, *halt*, *kopf* usw. weder Verbpartikeln noch substantivische Bestandteile wie in (18) sind. Eine Rolle spielen die (Nicht)Produktivität der jeweiligen Form bzw. die Beweglichkeit der Partikel.

(25) *Sie spielt Klavier. – Sie spielt ein altes Klavier.*
(26) *Sie läuft eis. - *Sie läuft ein festes Eis.*

In (25) wird, anders als oben angenommen, durch die Produktivität die Annäherung von *Klavier* an ein direktes Objekt deutlich. In (26) wird die Nicht-Produktivität erkennbar, die laut Eschenlohr (in Eisenberg 2004:341) „auf die Unmöglichkeit zurück [zu führen ist], dass substantivische Bestandteile sich wie echte Partikeln verhalten.“. Die Beweglichkeit der Partikel wie in (25) würde eine Wiederannäherung an Substantive bedeuten und da möglicherweise einen nominalen Tag rechtfertigen.

Als weitere Kriterien, auf die alle vorgestellten Verbgruppen zu testen sind, nennt Eisenberg (2004:341) Vorfeldfähigkeit und Abtrennbarkeit des ersten Bestandteils durch Einschub von Adverbien. Beides ist bei Verbpartikeln nur bedingt möglich.

(27) *?eis laufen wir immer gern*
(28) **dass wir eis gern laufen*

Tests:

Kann der abgetrennte Verbzusatz direktes Objekt sein? → wenn ja, nominaler Tag (Bsp. (17))
Ist die Partikel beweglich/produktiv? → Wiederannäherung an Substantiv, nominaler Tag (18)
Besteht Vorfeldfähigkeit? → wenn ja, keine Partikel (27)
Ist ein Adverbeinschub möglich? → wenn Adverb möglich, keine Partikel (28)

Fragen hinsichtlich des Status sowie der Trennbarkeit (und somit letztlich der Wortart) von Partikeln berühren grundsätzliche Probleme bezüglich der Orthographie (Eisenberg 2004:333). Die Orthographie sollte eigentlich Strukturen deutlich machen. Sind Schwierigkeiten bei der Getrennt-/Zusammenschreibung bzw. Groß-/Kleinschreibung also nur ein Manko der Orthographie, oder basieren die Ungewissheiten auf dem, was die Orthographie sichtbar macht?

3.4.2 Präpositionale Partikeln

Dass die mit Präpositionen homonymen Partikeln keine Eigenschaften von Präpositionen aufweisen, illustrieren wir anhand des Partikelverbs *mitkommen*:

(29) *Ich komme mit*.
(30) *Ich komme zu dir mit*.
(31) *Ich komme mit dir*.

In (29) kann es sich bei dem Wort *mit* nicht um eine Präposition handeln. Eine Präposition tritt mit einem Bezugswort auf und weist diesem einen Kasus zu. Dies ist aber hier nicht der Fall. Um eine Präpositionalphrase zu bilden, ist eine Präposition erforderlich wie in (30). Abgesehen von der Präpositionalphrase sind die Sätze in (29) und (30) gleich, es handelt sich um das gleiche Verb, *mitkommen*, und *mit* ist eine Partikel. In (31) hingegen handelt es sich um das Verb *kommen*, *mit* ist Präposition.

Test:

Liegen Eigenschaften einer Präposition vor? → wenn nein, vermutlich Partikel

3.4.3 Adjektivische Partikeln

> Mit dem Adjektiv wird über das vom direkten Objekt Bezeichnete prädiziert und es hat dabei dieselbe Stellung wie eine Verbpartikel. (Eisenberg 2004:336)

Ein Adjektiv als Objektsprädikativ zu transitiven Verben bildet das produktivste Muster bei den Partikelverben. Dadurch, dass das Adjektiv als Verbpartikel inkorporiert wird, kann ein neues transitives Verb entstehen (*hochheben*)[21]. Bei einem solchen inkorporierten Adjektiv handelt es sich (noch) nicht notwendigerweise um eine Partikel. Es kann also, wie sonst auch, modifiziert werden (Beispiel (32)), und es ist vorfeldfähig (Beispiel (33)) (Eisenberg 2004:337):

[21] Dieses neue transitive Verb existiert neben der alten syntaktischen Konstruktion.

(32) *völlig grün streichen, ganz tot sein, ...*
(33) *Grün streicht Helga ihr Fahrrad.*

Dass diese beiden Eigenschaften auf (die meisten) Partikelverben eher nicht zutreffen (Beispiel (34)) (Eisenberg 2004:337), spricht für ihre starke Grammatikalisierung – und damit auch gegen die Wortart ‚Adjektiv'.

(34) **Fest legt er sich.*

Ein syntagmatisches Homonym zu dem komplexen Verb ist in diesen Fällen nicht zwingend, weil das Adjektiv obligatorisch ist ((35)). Im Gegensatz dazu stehen die weniger stark grammatikalisierten Verben, die zum einen sowohl mit als auch ohne Objektsprädikativ auftreten können (Beispiel (36)) und zum anderen über ein zum Wort homonymes Syntagma verfügen.

(35) *Sie schreibt ihn krank. – *Sie schreibt ihn.*
(36) *Sie putzt die Zähne blank. – Sie putzt die Zähne.*

Dies ist ein gutes Argument für Zusammenschreibung:

> Bei starker Idiomatisierung gibt es nur ein komplexes Verb, bei den transparenten Typen gibt es sowohl ein Verb als auch ein Syntagma. Daraus folgt, dass Zusammenschreibung erlaubt sein muss. (Eisenberg 2004:338)

Die Zusammenschreibung impliziert, dass die jeweiligen Adjektive eine Einheit mit dem Verbstamm bilden, was dafür spricht, sie als Partikel zu kennzeichnen.

Die Rechtschreibung bleibt auch nach der Neuregelung problematisch. Heute ist die Getrenntschreibung und somit eine Degrammatikalisierung zugelassen. So ist teilweise willkürlich in das bestehende (nicht weniger willkürliche) System der Getrennt- bzw. Zusammenschreibung eingegriffen worden.

Das Kernproblem lautet also: Wenn ein Adjektiv und ein Verb zusammengeschrieben werden, wie lange ist dann das Adjektiv als Adjektiv und ab wann ist es als Verbzusatz zu betrachten? Auch hier könnten einige Tests einen Hinweis geben, wenn auch nicht eine endgültige Entscheidung abnehmen.

Tests:

Vorfeldfähigkeit? → eher Hinweis auf Adjektiv, keine Partikel
Modifizierbarkeit? → eher Hinweis auf Adjektiv, keine Partikel
Kein homonymes Syntagma? → guter Grund zusammenzuschreiben, also Partikel

3.4.4 Adverbien als Verbzusätze

Šimečková (1994:68) klassifiziert die Kandidaten für Verbpartikeln dieser Art folgendermaßen:

- Verben mit solchen adverbialen Elementen, die nicht frei, sondern nur in der Verbindung mit einem Verb vorkommen: *anheim-, fürlieb-, inne-, überein-, überhand-, vorlieb-, zurecht-*;

- Verben mit präpositional-adverbialen Elementen *ab-, an-, auf-, aus-, bei-, durch-, entgegen-, entlang-, gegen-, gegenüber-, hinter-, nach-, ob-, über-, um-, unter-, vor-, wider-, zu-, zwischen-;*

- Verben mit den adverbialen Elementen *dar-, ein-, empor-, entzwei-, fort-, heim-, her-, hinten-, hin-, nieder-, weg-, zurück-.*

Diese (nicht vollständige) Einteilung zeigt, wie heterogen die Gruppe der Adverbien ist. Eine einheitliche Behandlung wird dadurch erheblich erschwert. Die Eigenschaft der Vorfeldfähigkeit ist jedoch sowohl den aufgeführten Adverbien als auch den Adjektiven gemein. Somit dient derselbe Test, der bei den Adjektiven möglicherweise Aufschluss gibt, auch bei den Adverbien als Anhaltspunkt. In (37) handelt es sich bei *zusammen* um ein Adverb, welches das Verb *arbeiten* modifiziert. In (38) liegt das Verb *zusammenarbeiten* vor.

(37) *Zusammen (in diesem Raum) arbeiten wir schon lange.*
(38) **Zusammen arbeiten die Firmen seit kurzem.*

Test:

Vorfeldfähigkeit? → eher Hinweis auf Adverb, keine Partikel

3.4.5 Verben als Verbzusatz

Diese Gruppe spielt eine besondere Rolle, denn in diesen Bereich hat die Rechtschreibreform stark eingegriffen. Verben wie *kennen lernen, verloren gehen, bekannt machen* wurden vor der Neuregelung zusammengeschrieben. Folgt man der alten Rechtschreibung, so stellt sich die Frage nach dem Status der verbalen abgetrennten Verbzusätze in demselben Zusammenhang wie die nach dem Status der bereits diskutierten. Setzt man jedoch die neue Rechtschreibung voraus, wie wir es hier tun, so stellt sich diese spezielle Frage nicht, da das fragliche Element nicht im Infinitiv der Verbform enthalten ist und somit gar kein abgetrennter Verbzusatz sein kann.

3.4.6 Partikel- vs. Präfixverben

Die bisher vorgestellten Verben gehören alle den Partikelverben an. Neben diesen gibt es noch die Gruppe der Präfixverben (*entstauben, durchleiden*). Wenn auch die Funktionsunterschiede schwierig zu erfassen sind, so sind Präfix- und Partikelverben zumindest formal eindeutig voneinander unterscheidbar.

Die Unterschiede äußern sich dahingehend, dass die Präfixverben in allen Formen untrennbar sind (39). Sie werden auch feste bzw. nicht-trennbare Verben genannt. Die unfesten (also die trennbaren) werden nur in den infiniten Formen sowie als Partizip und finit in Verbzweit- und Verbletztsätzen fest verbunden (40). In den finiten Formen können beide Bestandteile beliebig weit voneinander entfernt stehen.

(39) *durchleiden – er durchleidet*
(40) *anlegen – er legt* (...) *an*, *angelegt*, ..., *dass er anlegt*

Der phonologische Unterschied liegt darin, dass Präfixverben auf dem verbalen Glied betont werden, die Betonung der trennbaren Verben hingegen liegt auf der Partikel (Doll 1967:4). Die Präfixverben stellen den normalen Wortbildungstyp dar und sind somit uninteressant für die PTKVZ-Problematik. Da es hier um die trennbaren Verben und insbesondere um ihre Eigenschaft der speziellen Wortbildung gehen soll, werden die Präfixverben lediglich zum Vergleich herangezogen.

Die typische Verbpartikel (41) hat eine homonyme freie Form (beispielsweise eine Präposition), ist betont und wird in bestimmten Kontexten sowohl morphologisch (42) als auch syntaktisch (43) vom Stamm des Basisverbs getrennt.

(41) *ankleben*
(42) *angeklebt*
(43) *Sie klebt das an*.
(44) *aufstehen – aufgestanden – aufzustehen*
(45) *das Aufgestehe, Umgefalle*

Genauer gesagt bezeichnet die morphologische Trennung die Trennung der Partikel vom Basisstamm innerhalb einer Form (Eisenberg 2004:255), wie beispielsweise in (44) oder bei Substantivableitungen mit dem Zirkumfix *Ge-e* (in (45)). Eine syntaktische Trennung ist nur bei den finiten Formen von Partikelverben vorzufinden, wie z.B. in (43).

Trotzdem bleibt nach wie vor der Status der abgetrennten Verbzusätze unklar und problematisch, da sich jegliche Analysen im „Übergangsbereich von Wort- und Satzgrammatik“ bewegen (Eisenberg 2004:268). Einerseits ist morphologische bzw. syntaktische Trennbarkeit nicht typisch für Wörter. Andererseits haben die Partikelverben viele Eigenschaften mit Komposita gemeinsam und gehen als Ganze in die Wortbildungsprozesse ein ((46), Eisenberg 2004:268). Das Ganze hat

Eigenschaften einer Form, aber die Partikel ist vom Stamm getrennt. Trotzdem werden Partikel und Stamm zusammengeschrieben und als eine Form behandelt.

(46) *anlötbar, Ausführung, Eisläufer, Heimreise, Totschläger, Warmhalterei ...*

Bei den produktiven Typen – also bei Verbstämmen mit substantivischen und adjektivischen Partikeln – ist klar erkennbar, inwiefern sie sich als Wörter und inwiefern als syntaktische Phrasen verhalten; außerdem sind die morphologischen Prozesse über die Bildungsweise klar: Dabei handelt es sich nicht um Wortbildungs-, sondern Univerbierungsprozesse[22].

Auch in Bezug auf die Argumentstruktur bestehen Unterschiede zwischen Präfix- und Partikelverben. Eisenberg (2004:259) erläutert dies, indem er Diathesen mit Wortbildungsprozessen vergleicht:

> Diathesen sind an Verbformen gebunden, die zum selben Paradigma gehören. Die passivischen Formen sind Formen ‚desselben' Verbs wie die aktivischen, dafür gibt es formale und semantische Gründe [...] Deshalb enthält ein Passivsatz im Normalfall vielleicht weniger, aber nicht andere Aktanten als der Aktivsatz. Ein Wortbildungsprozess wie Präfigierung erzeugt dagegen nicht andere Verbformen innerhalb eines Paradigmas, sondern ein anderes Verb. (Eisenberg 2004:259)

Das basiert auf der Annahme, dass das Präfix als Kopf des abgeleiteten Wortes betrachtet wird. Seine Argumentation stützt er im Wesentlichen auf die Beispiele *streichen* bzw. *überstreichen*. Im Großen und Ganzen treffen diese Annahmen auch auf andere Präpositionen zu, im Einzelnen müsste dies jedoch noch genauer überprüft werden.

3.4.7 Zwischenfazit

Die Tatsache, dass in nicht wenigen Fällen sowohl zusammen als auch getrennt geschrieben wird, lässt ahnen, dass das Problem nicht so einfach bzw. gar nicht gelöst werden kann. Beispielweise existiert *holzfällen*, das genauso von *Holzfäller* rückgebildet wie *Bäcker* von *backen* abgeleitet ist, neben dem Syntagma *Holz fällen* (Eisenberg 2004:332).

> Ob eine Form tatsächlich als Partikel anzusehen ist und also in das Verb inkorporiert ist, bleibt schwer entscheidbar. Entsprechend groß und gut begründet sind die Unsicherheiten für die Getrennt- und Zusammenschreibung. (Eisenberg 2004:268)

Die vorgestellten Tests reichen nicht in jeder Hinsicht aus, um die unterschiedlichen Probleme gleichwertig zu lösen. Während der Präpositionstest sich als zuverlässig erweist, sind z.B. die Substantivkriterien nicht aussagekräftig genug.

---

[22] D.h. es entstehen keine neuen Wörter, sondern häufig nebeneinander stehende Wortformen wachsen zusammen. Dies bewirkt jedoch keine (wesentlichen) semantischen Unterschiede (Eisenberg 2004:332): *zu Hause/zuhause.*

Wir schlagen daher vor, bis auf Weiteres dem STTS zu folgen und sich weiterhin des Tags PTKVZ zu bedienen. So bleibt das eingangs formulierte Problem, dass PTKVZ keine Wortart bezeichnet, zwar bestehen. Doch sollte abschließend zur Kenntnis genommen werden, dass andere Tags des STTS ebenfalls keine Wortart bezeichnen (FM, XY, ABK ...). All diese an das Wortarten-System anzugleichen wäre jedoch überhaupt nicht möglich (z.B. XY).

3.5 Indefinitpronomina (PI*)

Die Klasse der Indefinitpronomina ist bei der Wortarteneinteilung nicht unproblematisch, weil die traditionell dazu gezählten Wörter ganz unterschiedliche Eigenschaften hinsichtlich Flexion, Kasus und Vorkommen von Pluralformen haben.

So werden *man* und *jemand* zwar wie ein Substantiv verwendet, doch wird *man* nicht dekliniert, hat keinen Genitiv und wird im Dativ und Akkusativ vom substantivierten unbestimmten Artikel ersetzt:

(1) *Man spricht über ihn.* → Nom
(2) *Das geht einem an die Nieren.* → Dat
(3) *Das interessiert einen gar nicht.* →Akk

Jemand hingegen verfügt über ein voll ausgebautes Deklinationsschema (Götze/Hess-Lüttich 1989:231ff.):

(4) *jemand klopft* → Nom
(5) *jemandes Haus* → Gen.
(6) *jemandem/jemand auf den Geist gehen* → Dat
(7) *jemanden/jemand lieben* →Akk

Irgendein- wiederum bedient sich der Flexionsaffixe des Possessivpronomens in (8), *irgendwer* flektiert wie das w-Pronomen *wer* in (9), jedoch ohne Genitiv, nur *irgendwelch-* hat auch Pluralformen, tritt sogar zumeist nur im Plural auf. *Irgendetwas* ist unflektierbar (Grammis):

(8) *irgendein, -eines, -einem*, *-einen*
(9) *irgendwer, - , -wem, -wen*

Daher könnten Wörter, die unter dem Begriff Indefinitpronomina geführt werden, zum Teil ebenso gut z.B. den Numeralia, Adjektiven, Adverbien wie den verschiedenen Pronomina zugeschlagen werden (Perl 1976:292). Die folgenden Beispiele möglicher Zuweisung verdeutlichen dies:

(10) *die vielen*/PIDAT od. ADJA *Menschen*
(11) *Sie hat andere*/PIAT od. ADJA *Wünsche*.
(12) *ein solcher*/PDAT od. PIDAT *Holzkopf*
(13) *viel*/ADV *gelacht* und *viel*/PIS *gesehen*

Die Einteilungen in verschiedenen Grammatiken sind entsprechend unterschiedlich. Götze/Hess-Lüttich (1989:231) ordnen die attributiv gebrauchten Indefinitpronomina wie in (14) den Artikelwörtern zu, die wie ein Substantiv gebrauchten (15) werden davon ausgenommen und als Indefinitpronomina geführt. Einige unbestimmte Zahlwörter (16) werden zu den unbestimmten Zahladjektiven gezählt.

(14) *etliche Kinder, mancher Mann* etc.
(15) *jemand, irgendwer, niemand*, *nichts*
(16) *viel, wenig, andere*, *ein paar*

Eichler/Bünting ordnen allesamt den Indefinitpronomina zu und weisen darauf hin, dass unbestimmte Zahlwörter und adjektivisch gebrauchte Indefinitpronomina identisch sind (1996:76,136).

Die übliche Einordnung ist semantisch bedingt, und auch Eisenberg zählt sie zu den Pronomina, merkt aber an, dass sie syntaktisch zusammen mit den Demonstrativ- und den Possessivpronomina zu den Determinativpronomina gezählt werden können. Die Bezeichnung ‚indefinit' ist dabei insoweit missverständlich, als die fraglichen Elemente nicht Indefinitheit signalisieren, sondern Quantitäten abgrenzen (2001:179ff.).

Bei allen Unterschieden haben jedoch diese Wörter zwei wichtige Gemeinsamkeiten. Zum einen haben alle unter dem Begriff Indefinitpronomina subsumierbaren Wörter eine unbestimmt quantitative Bedeutung. Dadurch unterscheiden sie sich von den Personal- und Demonstrativpronomina, haben aber Ähnlichkeit z.B. mit Adjektiven wie *häufig* oder *zahlreich*. Zum anderen sind sie „hinweisend, situations- oder kontextbezogen" und damit Prowörter im weiteren Sinne und verschieden von anderen Bezeichnern von unbestimmbarer Menge. Sie bezeichnen entweder ein Teil vom Ganzen oder eine quantitative Eigenschaft (Perl 1976:292f.).

3.5.1 Indefinitpronomina im STTS

Im STTS werden sowohl klassische Indefinitpronomina wie etwa *irgendein-, irgend(et)was* oder *man* als auch Quantifikativpronomina wie *einig-, etlich-, jed-, mehrer-, sämtlich-* sowie Adjektive wie *viel-, wenig-, reichlich-* u.a. unter der Klasse der Indefinitpronomina zusammengefasst.[23] Die oben genannten Gemeinsamkeiten machen die Einteilung in eine Klasse nachvollziehbar; daher scheint die Einteilung im STTS im Großen und Ganzen kein Problem zu sein, zumal sie nicht allzu stark von anderen gängigen Klassifizierungen abweicht.

[23] Den genauen Formenbestand siehe STTS.

3.6 Fremdsprachliches Material (FM)

3.6.1 Problemstellung

Der Tag FM bezeichnet im STTS fremdsprachliches Material. Laut STTS erhalten den Tag FM „größere Textstücke, die einer fremden Sprache angehören, und nicht als Eigennamen klassifiziert werden können [...]“. Das STTS führt folgende Beispiele an:

(1) *Er hat das mit “but*/FM *this*/FM *was*/FM *not*/FM *so*/FM“ *übersetzt*.
(2) *der spanische Film “mujer*/FM *de*/FM *Benjamin*/NE“
(3) *Sie hat ihn dann einfach “lazy*/FM“ *genannt*
(4) *New*/NE *York*/NE
(5) *University*/NE *of*/NE *Michigan*/NE

Während die Beispiele (1) und (2) eindeutig zu taggen sind, sind (3), (4) und (5) in dieser Hinsicht problematisch. Zwar ist das englische *lazy* kein Eigenname, doch es ist auch kein größeres Textstück. Festzuhalten ist also, dass auch einzelne fremdsprachliche Wörter den Tag FM erhalten. Bei Fällen wie in (3) könnten die Anführungszeichen ein Hinweis auf ein fremdes Wort sein. Als ausschlaggebendes Kriterium kann dies jedoch nicht verstanden werden, da nicht sicher ist, ob in allen zu taggenden Texten solche Fälle gekennzeichnet werden.

Dass Eigennamen nicht so einfach zu erkennen sind, wird sowohl im Abschnitt über Eigennamen (Seite 32) als auch in den Beispielen (4) und (5) deutlich. Beim Wort-für-Wort-Tagging lässt sich eine Annotation wie in (4) nicht begründen, denn *New* alleine ist kein Eigenname. *York* hingegen ist zufälligerweise auch der Name einer Stadt und damit auch allein stehend ein Eigenname, doch das ist nicht immer der Fall (bspw. *Bad Honnef, Los Angeles*). *York* würde dann natürlich auch nicht mehr *New York* bezeichnen.

Außerdem müsste die Frage geklärt werden, welchen Tag *New* erhalten würde, wenn *York* NE zugewiesen bekäme. *New* als Adjektiv zu taggen wäre in diesem Fall zwar möglich. Bei Textstücken, die aus weniger geläufigen Sprachen kommen, wäre diese Vorgehensweise jedoch keine Alternative, weil Kategorie bzw. Bedeutung unter Umständen nicht bekannt sind. Bei bekannten Sprachen nach der jeweiligen Wortart zu taggen und bei Sprachen, die nicht beherrscht werden, entsprechende Textstücke mit FM zu versehen, wäre wiederum inkonsistent. Hinzu käme in diesem Fall, dass unterschiedliche Sprachen unterschiedliche Wortarteneinteilungen haben können. Daher lässt sich nicht ohne weiteres beurteilen, welcher Tag ggf. in Frage käme.

Konsequenterweise müssten demnach alle fremdsprachlichen Bezeichnungen, die sich aus mehr als einem Wort zusammensetzen, den Tag FM erhalten. Hieraus ergibt sich jedoch, dass einelementige fremde Eigennamen als Eigennamen gekennzeichnet werden, mehrelementige jedoch als fremdsprachliches Material.

Dass die Klassifizierung von Eigennamen problematisch ist, wurde bereits im Abschnitt über Eigennamen deutlich. Dass insbesondere an der Grenze zu fremdsprachlichem Material besondere Schwierigkeiten entstehen können, ist nun klar geworden. Laut STTS „[...] ist das fremdsprachliche Material [auf keinen Fall] auf die deutsche Syntax zu übertragen“. Somit scheidet die Möglichkeit, dem fremdsprachlichen Material, das nicht den Eigennamen zuzuordnen ist, die jeweilige deutsche Wortart zuzuweisen, aus.

Als vorläufiger Kompromiss lautet der Vorschlag hier, Eigennamen, die als solche erkannt werden, auch mit NE zu kennzeichnen. Dies entspricht so auch dem STTS. Beispiel (5) müsste somit auf NE-Eigenschaften untersucht werden.

3.6.2 Fremdwort vs. fremdes Wort

Weitaus problematischer sind jedoch jene Wörter, die aus fremden Sprachen stammen und zu einem gewissen Grad ins Deutsche übernommen sind. Laut Duden (2002:122) werden Wörter aus anderen Sprachen „üblicherweise Fremdwörter genannt, obgleich sie zu einem großen Teil durchaus keine fremden, sondern seit langem bekannte und gebräuchliche Wörter für die deutsche Sprachgemeinschaft sind, die in der Sprache ihren festen Platz haben.“. Um diese so genannten Fremdwörter näher bestimmen zu können, nennt der Duden vier Merkmale, die ein Wort als nichtmuttersprachliches kennzeichnen (Duden 2002:122). Diese Merkmale lassen sich unterteilen in morphologische, phonologische, orthographisch/graphematische sowie das Lexikon betreffende Kriterien, die im Folgenden näher beleuchtet werden.

3.6.3 Morphologisches Kriterium

Das morphologische Kriterium bezieht sich auf bestimmte Affixe, die auf ein Fremdwort hindeuten können:

(6) *Apparat<u>schik</u>*, *<u>hypochondrisch</u>*[24]

Dieses Kriterium wird zwar relativiert aufgrund von Ausnahmen wie *ab-* (in *absolut*, aber auch in *abreisen*), kann jedoch grundsätzlich als Hinweis auf ein Fremdwort dienen. Laut Duden begünstigen fremde Suffixe, wie sie in Mischformen, so genannten hybriden Bildungen, vorkommen, die Zuordnung zu den Fremdwörtern (7). Umgekehrt werden Wörter mit fremdem Stamm und deutscher Morphologie als deutsche empfunden (8):

(7) *buchstab<u>ieren</u>*
(8) *<u>Direkt</u>heit*, *<u>temperament</u>voll*

[24] Die Beispiele (6) bis (18) wurden dem Duden (2002) entnommen.

3.6.4 Phonologisches Kriterium

Eine vom Deutschen abweichende Aussprache (9) oder Betonung ist ein weiterer Anhaltspunkt. Typischerweise liegt die Betonung von deutschen Wörtern auf der ersten oder der Stammsilbe. Ist dies nicht der Fall, könnte dies ein Hinweis auf ein Fremdwort sein. Doch auch hier gibt es Ausnahmen. Zum einen gibt es Fremdwörter mit einer für das Deutsche typischen Anfangsbetonung (10). Zum anderen ist die Betonung auch bei nativen deutschen Wörtern nicht so regelmäßig (11). Außerdem hat sich die Aussprache teilweise schon angeglichen (12).

(9) *engl. Boot* (dt. *Stiefel)*; *Friseur*
(10) *Atlas*, *Lyrik*, *Radio*
(11) *Forelle*, *lebendig*
(12) *Stil*, *Stadion*

3.6.5 Orthographisch-graphemisches Kriterium

Die Schreibung eines Wortes kann ebenfalls Aufschluss über die Herkunft geben. Laut Duden (2001:122) signalisieren bestimmte Buchstabenverbindungen für das Deutsche unübliche graphische Strukturen (13). Die Position bestimmter Buchstabenverbindungen kann auch ein Indikator sein (beispielsweise kommt <pt> in nativ deutschen Wörtern nicht im Anlaut vor). Bestimmte Fremdwörter sind jedoch nicht anhand ihrer Buchstabenverbindungen zu enttarnen, entweder weil diese in der Zielsprache nicht als nicht-nativ auffallen (14) oder weil sie ein der Zielsprache angepasstes Schriftbild erhalten haben (15).

(13) *Bodybuilder*, *Soutane*, *Osteoporose*
(14) *Sprinkler*
(15) *Keks* (statt *cakes*), *schocken (*statt *to shock)*

3.6.6 Lexikalisches Kriterium

Teilweise werden native, aber selten gebrauchte Wörter als fremd empfunden. Das Kriterium der Häufigkeit könnte mit dem Lexikon zusammenhängen: Wenn es im Deutschen einen Begriff für etwas Bestimmtes gibt, muss kein Wort aus einer fremden Sprache dafür geliehen werden. Auch Erbwörter[25], nicht entlehnt und nicht fremd, können extrem selten in der Alltagssprache sein und so möglicherweise für Fremdwörter gehalten werden (16). Zum Teil werden sogar Wörter für Fremdwörter beispielsweise lateinischen oder englischen Ursprungs gehalten, obwohl es sich um native Wörter des Deutschen handelt, deren Etymologie lediglich nicht mehr durchschaubar ist (17). Andererseits gibt es fremde Wörter, die völlig gebräuchlich und allgemein verständlich sind (18) (Duden 2002).

(16) *Buhne*, *Flechse*
(17) *Bovist*, *Quarz*, *blaken*
(18) *Auto*, *militärisch*, *Möbel*, *Salat*

[25] Vgl. Tabelle S.65.

Diese Kriterien können als Hinweis dienen, sie sind jedoch nicht ausschlaggebend. Meistens haben Fremdwörter mehr als eines dieser Merkmale, aber keines davon ist entscheidend. Trotz dieser Kennzeichen lässt sich nicht mit Sicherheit sagen, was ein Fremdwort ist, denn die Grenzen zwischen Fremdwort und eingebürgertem Wort sind oft fließend.

3.6.7 Fremdwort, Lehnwort und Erbwort

Einen Ansatzpunkt, um die Grenze zwischen eingebürgertem Wort und Fremdwort genauer zu spezifizieren, bieten die Begriffe Fremdwort, Lehnwort und Erbwort.

Laut Eisenberg (2004:38) sind „Fremdwörter [...] ganz oder in wesentlichen Bestandteilen aus anderen Sprachen übernommen und haben sich nicht vollständig an die Strukturen des Deutschen angepasst“, wie die Beispiele in (19).

(19) *Exot*, *Galaxis*, *Cholesterin*, *Thermostat*

Lehnwörter sind zwar auch fremder Herkunft, aber ins Deutsche integriert (z.B. Latinismen, vor allem Substantive wie *Fenster*, Adjektive wie *recht*, Verben wie *wollen* und Präpositionen wie *pro*, sowie Lehnwörter aus dem Griechischen (*Meter*), Französischen *(Soße)*, Englischen (*Boss*). Für Eisenberg sind Wörter, die in phonologischer, morphologischer und graphematischer Hinsicht unauffällig sind zum Kern des Wortschatzes zu rechnen (Eisenberg 2004:39).

Welche Unterscheidungen genau gemacht werden, geht aus der folgenden Tabelle hervor (nach Baurmann/Eisenberg 1984:16):

synchron	**diachron**	**Bezeichnung** nach Heller	**Beispiel**
fremd	entlehnt	Fremdwort	*Malheur*
fremd	nicht entlehnt	Pseudofremdwort	*Showmaster*
nicht fremd	entlehnt	Lehnwort	*Panne*
nicht fremd	nicht entlehnt	Erbwort	*Baum*

Diese Darstellung basiert auf der Annahme, dass „ein aus einer fremden Sprache im historischen Prozeß übernommenes Wort ‚entlehnt‘ und alle anderen Wörter ‚nicht entlehnt‘ genannt werden“ (Baurmann/Eisenberg 1984:16). Als fremd werden hier Wörter bezeichnet, die auf einer der Ebenen, die Baurmann/Eisenberg für relevant halten, von der Norm abweichen. Die verschiedenen Ebenen des Systems sind bei Baurmann/Eisenberg:

> [...] ein Wort hat einen Lautkörper, eine morphologische Struktur, eine graphemische Struktur, eine Bedeutung und eine syntaktische Charakteristik. Auf jeder dieser Ebenen kann das entlehnte Wort also mehr oder weniger stark von deutschen Wörtern abweichen.

> Dasselbe gilt für bestimmte Merkmale, die seinen Gebrauch betreffen wie Häufigkeit, Verbreitung und Zugehörigkeit zu bestimmten Stil-schichten.(Baurmann/Eisenberg 1984:16)

Entlehnt und fremd ist demnach ein Wort wie *Malheur*. Es stammt aus dem Französischen und weicht hinsichtlich der Schreibung vom Deutschen ab.

Nicht entlehnt, aber fremd sind so genannte Pseudofremdwörter wie *Showmaster, Handy* oder *Highlife*. Die äußere Form, hier beispielsweise Aussprache und Schreibung, lassen auf ein entlehntes Wort schließen. Tatsächlich sind solche Wörter aber im Deutschen gebildet. An diesem Fall wird deutlich, dass sowohl die Kriterien bzw. Ebenen als auch der diachronische Aspekt eine Rolle spielen.

Entlehnt und nicht fremd sind Wörter wie *Panne* oder *Keks*. Ihre äußere Form lässt nicht darauf schließen, dass sie aus anderen Sprachen übernommen sind.

Als nicht entlehnt und nicht fremd bzw. als Erbwörter werden die so genannten alten deutschen Wörter wie *Baum* bezeichnet. Baurmann/Eisenberg gehen nicht darauf ein, wie lange ein Entlehnungsvorgang vergangen sein muss, damit von einem alten deutschen Wort die Rede sein kann. Kleinpaul sagt hierzu:

> Fremdwörter, die so alt sind, dass man sich ihrer Ankunft gar nicht mehr erinnert, die inzwischen auch alle Leiden der Sprache geteilt haben, nennt man Lehnwörter. Der Ausdruck bezeichnet nichts grundsätzlich Verschiedenes, sondern nur einen höheren Grad der Einbürgerung, die stillschweigende Aufnahme in der Sprachschatz und in die Staatsangehörigkeit, meist auch die Gewöhnung an die Sache. Die Lehnwörter sind naturalisierte Fremde. (Kleinpaul 1905:74)

3.6.8 Zwischenfazit

Im Hinblick auf eine plausible Taggingregel folgen wir Baurmann/Eisenberg, die in Anlehnung an Polenz vorschlagen, den Fremdwortbegriff eng zu fassen:

> Er solle beschränkt bleiben auf Ausdrücke, die aus anderen Sprachen zitiert werden (pro forma; de facto; last but not least) und solche Wörter, die etwas bezeichnen, was es bei uns nicht gibt (College, Lord, Siesta). Alle anderen sollten Lehnwörter genannt und danach gruppiert werden, ob sie zum Bildungswortschatz, zum Fachwortschatz oder zum Gemeinwortschatz gehören. (Baurmann/Eisenberg 1984:16)

Da Fremdwörter die einzigen sind, die sowohl entlehnt als auch fremd (hinsichtlich ihrer Merkmale) sind, schlagen wir vor, dass ausschließlich diese den Tag FM erhalten.

Zusammenfassend lässt sich also sagen, dass die Abgrenzung zu Eigennamen zwar weiterhin schwierig bleibt, jedoch die richtungweisenden Kriterien der Eigennamen eine starke Tendenz vorgeben. Ebenso richtungweisend sind die Kriterien für die Unterscheidung von fremdsprachlichem und nativem Material, sowohl die synchronen Kriterien als auch diachrone Aspekte.

Dass dennoch die Intuition des kompetenten Sprechers, also des Muttersprachlers, als übergeordnetes Kriterium zu betrachten ist, bleibt unumstritten.

4 Fazit

Ziel der vorliegenden Arbeit war darzustellen, dass die Einteilung von Wörtern in bestimmte Wortklassen ein grundlegendes Problem sowohl für die manuelle als auch die automatische Wortarten-Annotation darstellt. Die verschiedenen Versionen der einzelnen dargestellten Tagsets zeigen, dass Veränderungen hinsichtlich nur eines bestimmten Tags sich auf das gesamte Inventar auswirken können. Um möglichst genau annotieren zu können und nicht mit Veränderungen auf der einen Seite Probleme auf der anderen Seite zu generieren, müssen die Veränderungen mittels automatischen Lernens überprüft werden, um so letztlich ein effizientes, in sich geschlossenes Set bestimmen zu können.

Je feinkörniger ein Tagset beschaffen ist, umso bessere Ergebnisse können damit bei der manuellen Annotation erzielt werden. Inwieweit dies im Einzelfall auch auf das automatische Tagging zutrifft, bleibt Gegenstand weiterer Untersuchungen.

Wir haben verschiedene Lösungsansätze für ausgewählte Einzelprobleme diskutiert und sind insgesamt zu dem Schluss gekommen, die Einteilung des STTS weitgehend zu übernehmen. Unsere Überlegungen hierzu sollen im Folgenden noch mal kurz zusammengefasst werden:

ADJA/ADJD vs. ADV: Die Annahme, Adjektive in adverbialer Stellung als Adverb zu taggen, könnte zu besseren Ergebnissen führen, bestätigte sich letztlich nicht.

ADJD vs. VVPP: Die dargestellten Kriterien erleichtern zwar die manuelle Annotation. Da ADJD und VVPP jedoch nicht komplementär verteilt sind, können die Kriterien nicht ohne Weiteres in den Tagger integriert werden.

NE vs. NN: Der entwickelte Kriterienkatalog für NE hilft dabei, die Zuordnung zumindest bei der manuellen Annotation zu erleichtern.

PTKVZ: Auch wenn es sich hierbei nicht um ein Wortartentag handelt, ist die Beibehaltung des Tags für die verschiedenen Phänomene sinnvoll. Die Diskussion der unter PTKVZ fallenden Phänomene hat gezeigt, dass vor allem unterschiedliche Schreibungen (z.B. aufgrund alter vs. neuer Rechtschreibung) einer einheitlichen, genaueren Einteilung im Wege stehen.

PI*: Die Gemeinsamkeiten der unter Indefinitpronomina erfassten Wörter haben ein stärkeres Gewicht als die Unterschiede, so dass die vom STTS vorgeschlagene Einteilung beibehalten werden sollte.

FM: Unsere Überlegungen zu fremdsprachlichem Material haben ergeben, dass der Fremdwortbegriff eng gefasst werden sollte: Nur Wörter, die sowohl entlehnt als auch fremd sind (hinsichtlich Phonologie, Graphematik, Lexikon), sollten den Tag FM erhalten.

5 Literaturangaben

Admoni, Wladimir (1970): *Der deutsche Sprachbau*. München: C.H.Beck.

Allen, W. Stannard (1970): *Living English Structure. A Practice Book for Foreign Students*. London: Longman.

Baurmann, Jürgen/ Eisenberg, Peter (1984): Fremdwörter – fremde Wörter. In: *Praxis Deutsch.* Zeitschrift für den Deutschunterricht 11 (67). Velbert: Friedrich Verlag; 15-26.

Bergenholtz, Henning/ Schaeder, Burkhard (1977): *Die Wortarten des Deutschen*. Stuttgart: Klett.

Bußmann, Hadumod (2002): *Lexikon der Sprachwissenschaft*. 3., aktual. und erweit. Auflage. Stuttgart: Kröner.

Doll, Ruth (1967): *Der deutsche Verbzusatz als Richtungsträger und seine Wiedergabe im Französischen und Italienischen*. Universität Tübingen.

Duden (2002): *Fremdwörterbuch*. 7., neu bearb. und erweit. Auflage, hrsg. von der Dudenredaktion. Mannheim: Dudenverlag. (Der Duden; Bd. 5).

Eichler, Wolfgang/ Bünting, Karl-Dieter (1996): *Deutsche Grammatik – Form, Leistung und Gebrauch der Gegenwartssprache.* 6. Aufl. Weinheim: Beltz Athenäum.

Eisenberg, Peter (1999): *Grundriß der deutschen Grammatik. Bd.2. Der Satz*. Stuttgart: Metzler.

Eisenberg, Peter (2004): *Grundriß der deutschen Grammatik. Bd.1. Das Wort*. 2. überarb. und aktual. Auflage. Stuttgart: Metzler.

Fleischer, Wolfgang (1964): Zum Verhältnis von Name und Appellativum im Deutschen. In: Der Rektor der Karl-Marx-Universität Leipzig (Hrsg.): *Wissenschaftliche Zeitschrift der Karl-Marx-Universität Leipzig. Gesellschafts- und Sprachwissenschaftliche Reihe.* 13 (2); 369-378.

Fleischhack, Erich u.a. (1981): *English G Grammatik.* Berlin: Cornelsen-Velhagen & Klasing.

Götze, Lutz/ Hess-Lüttich, Ernest W.B. (1989): *Knaurs Grammatik der deutschen Sprache. Sprachsystem und Sprachgebrauch.* München: Lexikographisches Institut.

grammis – das grammatische Informationssystem des IDS, URL: http://hypermedia.ids-mannheim.de [Letzter Zugriff 27.09.2010]

Heidolph, Karl Erich u.a. (1981): *Grundzüge der deutschen Grammatik*. Berlin: Akademie-Verlag.

Helbig, Gerhard (1977): Zu einigen Problemen der Wortartklassifizierung im Deutschen. In: Helbig, Gerhard (Hrsg.), *Linguistische Studien. Beiträge zur Klassifizierung der Wortarten*. Leipzig: Verlag Enzyklopädie; 90 – 118.

Helbig, Gerhard/ Buscha, Joachim (1997): *Deutsche Grammatik*. 19. Aufl. Leipzig: Langenscheidt.

Jurafsky, Daniel/Martin, James H. (2000): *Speech and Language Processing*. New Jersey: Prentice Hall.

Kleinpaul, Rudolf (1905): *Das Fremdwort im Deutschen*. Leipzig: G.J. Göschen'sche Verlagshandlung.

Knobloch, Clemens (1992): Eigennamen als Unterklasse der Nomina und in der Technik des Sprechens. In: Bergmann, Rolf u.a. (Hrsg.), *Sprachwissenschaft* 17; 451 - 471

Lyons, John (1980): *Einführung in die moderne Linguistik*. 5. Aufl. München: Beck.

Manning, Christopher M./ Schütze, Hinrich (2002): *Foundations for Statistical Natural Language Processing*. Boston: MIT.

Perl, N.E. (1976): Zum Problem der semantisch-syntaktischen Beziehungen von Prowörtern mit unbestimmt quantitativer Bedeutung. In: Herder-Institut der Karl-Marx-Universität (Hrsg.), *Deutsch als Fremdsprache* 13; Leipzig; 292 – 296.

Schiller, Anne/ Thielen, Christine u.a. (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).* Universität Stuttgart/Tübingen.

Šimečková, Alena (1994): *Untersuchungen zum ‚trennbaren' Verb im Deutschen I*. Prag.

Taylor, Ann u.a. (2003), *The Penn Treebank: An Overview*. URL: http://treebank.linguist.jussieu.fr/pdf/1.pdf [Letzter Zugriff 08.02.2005], vgl. Taylor, Ann/Marcus, Mitchell/Santorini, Beatrice (2003): The Penn Treebank: An Overview. In: Abeillé, Anne (Hrsg.): *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer.

TEI AI1W2 (1991): *List of Common Morphological Features For Inclusion in TEI Starter Set Of Grammatical-Annotation Tags*. URL: http://www.tei-c.org/Vault/AI/ai1w02.txt [Letzter Zugriff 27.09.2010]

Wahrig, Gerhard (2000): *Wörterbuch der deutschen Sprache.* 4. Aufl. der Neuausgabe, bearb. v. Dr. Renate Wahrig-Burfeind. München: Deutscher Taschenbuch Verlag.

Wöllstein-Leisten, Angelika u.a. (1997): *Deutsche Satzstruktur. Grundlagen der syntaktischen Analyse.* Tübingen: Stauffenberg.

Zifonun, Gisela u.a. (1997): *Grammatik der deutschen Sprache.* Schriften des Instituts für deutsche Sprache, Bd. 7.1. Berlin: de Gruyter.

6 Anhang

6.1 Links zum Thema

(letzter Zugriff auf alle URLs: 12.11.2010)

Beispiel für Brill Tagger:
http://cst.dk/online/pos_tagger/uk/index.html

Penn Treebank:
http://www.cis.upenn.edu/~treebank/home.html
http://www.scs.leeds.ac.uk/amalgam/tagsets/upenn.html

STTS:
http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html

UIS:
http://www.ifi.unizh.ch/CL/tagger/UIS-STTS-Diffs.html

6.2 Penn Treebank Tagset

Quelle: http://www.computing.dcu.ie/~acahill/tagset.html

CC	Coordinating conjunction e.g. and,but,or...
CD	Cardinal Number
DT	Determiner
EX	Existential *there*
FW	Foreign Word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List Item Marker
MD	Modal e.g. can, could, might, may...
NN	Noun, singular or mass
NNP	Proper Noun, singular
NNPS	Proper Noun, plural

NNS	Noun, plural
PDT	Predeterminer e.g. all, both ... when they precede an article
POS	Possessive Ending e.g. Nouns ending in 's
PRP	Personal Pronoun e.g. I, me, you, he...
PRP$	Possessive Pronoun e.g. my, your, mine, yours...
RB	Adverb Most words that end in -ly as well as degree words like quite, too and very
RBR	Adverb, comparative Adverbs with the comparative ending -er, with a strictly comparative meaning.
RBS	Adverb, superlative
RP	Particle
SYM	Symbol Should be used for mathematical, scientific or technical symbols
TO	*to*
UH	Interjection e.g. uh, well, yes, my...
VB	Verb, base form subsumes imperatives, infinitives and subjunctives
VBD	Verb, past tense includes the conditional form of the verb to be
VBG	Verb, gerund or persent participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
WDT	Wh-determiner e.g. which, and *that* when it is used as a relative pronoun
WP	Wh-pronoun e.g. what, who, whom...
WP$	Possessive wh-pronoun e.g.
WRB	Wh-adverb e.g. how, where why
#	#
$	$

"	"
(	(
)	)
,	,
.	.
:	:
``	``

6.3 STTS Tag Table (1995/1999)

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	ich, er, ihm, mich, dir

	Personalpronomen	
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	- [,]()