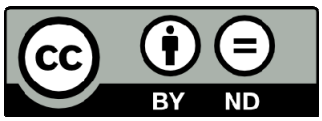
UNIVERSITAT DE BARCELONA

Parameter der akademischen Testproduktion: beiträge zur kontrastiven Analyse der Textproduktion deutscher und spanischer Studenten

Oliver Strunk



Universitat de Barcelona

Facultat de Filologia

Oliver Strunk

Parameter der akademischen Textproduktion.

Beiträge zur kontrastiven Analyse der Textproduktion deutscher und spanischer Studenten.

Tesi dirigida pel Dr. Fco. Javier Orduña

Tesi presentada per a l'obtenció del grau de Doctor en Filologia Alemanya

Departament de Filologia Anglesa i Alemanya

Secció de Filologia Alemanya

Barcelona, Desembre de 1998

Anerkennung

Die vorliegende Arbeit hat vielleicht nur einen Autor, aber zu ihrem Entstehen haben verschiedene Personen beigetragen, denen ich hier für ihre Anregungen und Unterstützung danken will.

Prof. Dr. Javier Orduña möchte ich sowohl für die Zeit und die Arbeit danken, die er unseren Diskussionen und dem Lesen der verschiedenen Versionen der Arbeit gewidmet hat, besonders aber auch wegen seiner Betreuung, die ich in Anbetracht seiner zahlreichen weiteren Verpflichtungen als vorbildlich bezeichnen muß.

Prof. Dr. Marisa Siguan bin ich verbunden wegen ihrer Anregungen zur Arbeit und der Unterstützung bei der nicht immer leichten Aufgabe, Seminare und Forschung miteinander zu verbinden, und vor allem wegen ihrer Zuversicht. Und Prof. Dr. Pedro Guardia hat mir geholfen mit seinen Tips, mit seinem Interesse, bei der Bewältigung zahlreicher bürokratischer Probleme und vor allem durch das „Vorbeischauen" um zu sehen, ob ich denn auch gut vorankomme.

Zur Entstehung der Arbeit haben einige Dozenten deutscher Universitäten und Institutionen beigetragen, die mir freundlicherweise textuelles Material von ihren Studenten überließen. Ohne diese oft aufwendige Hilfe von Dr. Martina Nicklaus, Dr. Elisabeth Coeffen, Dr. Ina Schreiter und anderen Dozenten und Forschern wäre diese Arbeit wohl unmöglich gewesen, und ich bin ihnen deswegen zutiefst dankbar, ebenso wie all den Studenten, im besonderen der Universität Barcelona, die mir ihre Texte überlassen haben. Auch sei hier die technische Unterstützung von Dr. Helmut Schmid und Dr. Wolfgang Lezius hervorgehoben, die mir freundlicherweise ihre Programme zur Verfügung stellten, technische Fragen klären halfen und auch ansonsten lebhaftes Interesse an der Arbeit zeigten, was besonders in Momenten der Verwirrung zur Weiterarbeit anspornte.

Unter den Freunden, die gewissenhaft für sie oft unverständliche Erklärungen über sich ergehen ließen, seien vor allem Virginia Trueba und Joan Castellví erwähnt. Joan, weil wir parallel an unseren Untersuchungen arbeiteten und fast zur gleichen Zeit zum Ende gekommen sind, was zu einer tiefen Freundschaft geführt hat. Und Virginia, weil sie trotz ihrer Vorliebe für die Literatur auch sprachwissenschaftlichen Fragen Interesse abringen konnte und mich zur Fertigstellung der Arbeit antrieb.

Doch die wichtigste Person für diese Arbeit war zweifellos Marta Fernández-Villanueva Jané. Dank ihr habe ich ein Thema gefunden, an dem ich hoffe weiter arbeiten zu können, mit ihrer Unterstützung habe ich diese Untersuchung vorantreiben und fertigstellen können. Auch in Momenten schierer Verzweiflung hat sie mir immer mit fachlichem Rat und vor allem durch ihre

persönliche Nähe geholfen. Ohne sie wäre diese Arbeit weder entstanden noch hätte sie die vorliegende Form.

Und schließlich möchte ich meinen Eltern, Renate und Jochen Strunk, danken für all die Jahre Unterstützung, für das Vertrauen und die Bitte, doch endlich fertig zu werden. All ihnen, besonders aber meiner Mutter, widme ich diese Arbeit.

INHALTSVERZEICHNIS

1 EINLEITUNG **11**

1.1 UNTERSUCHUNGSBEREICH UND UNTERSUCHUNGSOBJEKT **13**

Fehleranalyse, Kontrastive Analyse und Lernersprachenanalyse 17

Textproduktion in der Zweitsprache 19

Korpuslinguistik 24

1.2 METHODISCHE GRUNDLAGEN **26**

1.3 ZIELE DER UNTERSUCHUNG, HYPOTHESEN UND VORGEHENSWEISE **35**

Verfahrensweise und Phasen der Untersuchung 39

Die Texte der zusammengestellten Korpora 41

1.4 TERMINOLOGISCHE ASPEKTE **48**

1.5 GLIEDERUNG DER ARBEIT **50**

2 KORPUSDESIGN **53**

2.1 KORPUSBEZOGENE ASPEKTE **62**

2.1.1 ALLGEMEINE THEORETISCHE VORÜBERLEGUNGEN UND ZIELSETZUNG 63

Computerkorpora mit allgemeiner Zielsetzung 65

Computerkorpora mit spezifischer Zielsetzung 68

2.1.2 REPRÄSENTATIVITÄT, STRUKTURIERUNG UND VARIATIONSKRITERIEN 69

2.1.2.1 Produktionsform 78

2.1.2.2 Texteinteilung 81

2.1.2.3 Textursprung 88

2.1.2.4 Verfasser 92

Population 92

Idiolekte 93

2.1.2.5 Kommunikationsform 94

2.1.2.6 Zusammenfassung 95

2.1.3 KORPUSUMFANG 96

2.1.4 ABGESCHLOSSENHEIT 102

2.1.5 MASCHINELLE LESBARKEIT 104

2.2 TEXTBEZOGENE ASPEKTE 108

2.2.1 AUSWAHL 109

2.2.1.1 Verallgemeinerbarkeit 109

2.2.1.2 Textlänge 112

2.2.1.3 Textauswahl 114

2.2.2 AUSZEICHNUNGEN 115

2.2.2.1 Auszeichnungsformat 119

2.2.2.2 Auszeichnungsebenen 125

3 COMPUTERLERNERKORPUS (CLK) UND COMPUTERMUTTERSPRACHLERKORPUS (CMK) 131

3.1 KORPUSBEZOGENE VORAUSSETZUNGEN 138

3.1.1 ALLGEMEINE THEORETISCHE VORÜBERLEGUNGEN UND ZIELSETZUNG 138

3.1.2 REPRÄSENTATIVITÄT, STRUKTURIERUNG UND VARIATIONSKRITERIEN 139

3.1.2.1 Produktionsform 150

3.1.2.2 Texteinteilung 150

3.1.2.3 Der Sprachstand 160

3.1.2.4 Textursprung 161

3.1.2.5 Verfasser 162

3.1.2.6 Kommunikationsform 165

3.1.3 KORPUSUMFANG 165

3.1.4 ABGESCHLOSSENHEIT 168

3.1.5 MASCHINELLE LESBARKEIT 169

3.2 TEXTBEZOGENE ASPEKTE 172

3.2.1 AUSWAHL 172

3.2.1.1 Verallgemeinerbarkeit 173

3.2.1.2 Textlänge 174

3.2.1.3 Auswahlmethode 179

3.2.2 AUSZEICHNUNGEN 180

3.2.2.1 Auszeichnungsformat 183

3.2.2.2 Auszeichnungsschema 187

Der Header 188

Der Body 192

POS-Tagging 193

Das POS-Tagset von STTS 198

4 AUSWERTUNG DES EMPIRISCHEN MATERIALS 203

4.1 BESTIMMUNG UND AUSWAHL RELEVANTER PHÄNOMENE 205

Sprachunabhängige statistische Phänomene 209

Lexikalische Phänomene 209

Morphologische Phänomene 210

Grammatische oder POS-Phänomene 210

Syntaktische Phänomene 210

Argumentative, semantische und weitere theoriegebundene Phänomene 211

Nicht erkennbare Elemente oder manuell zu bestimmende Phänomene 211

4.2 AUSWERTUNG SPRACHLICHER PHÄNOMENE 213

4.2.1 ALLGEMEINE STATISTISCHE UNTERSUCHUNGEN 218

4.2.1.1 Wortlänge 219

4.2.1.2 Satzlänge 227

4.2.1.3 Rekurrente Wortkombinationen (Cluster) 230

Zweiwortcluster 232

Dreiwortcluster 234

4.2.1.4 Type-Token-Ratio: lexikalischer Variationsindex 236

4.2.1.5 Affixe 242

Allgemeines 245

Substantive: Präfixe 248

Substantive: Suffixe 249

Adjektivische Suffixe 253

Verbale Präfixe 255

4.2.2 WORTARTEN VON STTS 258

4.2.2.1 Adjektive (ADJA und ADJD) 260

4.2.2.2 Adverbien (ADV) 264

4.2.2.3 Adpositionen 265

Präpositionen (APPR) 267

APPRART 270

4.2.2.4 Artikel 270

4.2.2.5 Konjunktionen 272

Unterordnende Konjunktionen mit Infinitiv (KOUI) 274

Unterordnende Konjunktionen mit Satz (KOUS) 275

Nebenordnende Konjunktionen (KON) 279

Vergleichspartikel (KOKOM) 282

4.2.2.6 Substantive (NN) 284

4.2.2.7 Pronomina 287

Nicht reflexive Personalpronomina (PPER) 288

Relativpronomina (PRELAT und PRELS) 292

Demonstrativpronomina (PD) 295

Indefinitpronomina (PI) 298

Reflexive Personalpronomina (PRF) 303

Possessivpronomina (PPOS) 306

Interrogativpronomina (PW) 307

4.2.2.8 Pronominaladverbien (PAV) 310

4.2.2.9 Verben 314

Verben: Allgemeine Darstellung 315

Verben: Infinitive mit zu (VVIZU) 321

Verben: Partizipiale Formen (VVPP, VMPP, VAPP) 325

4.2.2.10 Weitere Wortartentags 327

Zu vor Infinitiv (PTKZU) 328

Negationspartikel (PTKNEG) 329

Abgetrennter Verbzusatz (PTKVZ) 329

Partikel bei Adjektiv oder Adverb (PTKA) 331

Kompositions-Erstglied (TRUNC) 332

5 INTERPRETATION DES EMPIRISCHEN MATERIALS 335

5.1 WECHSELBEZIEHUNGEN 338

5.1.1 TENDENZEN IN DER WORTBILDUNG 344

5.1.2 TENDENZEN AUF WORTEBENE (SYNTAKTISCHE WORTARTEN) 351

Überrepräsentation und Unterrepräsentation 352

Unterrepräsentation fakultativer Elemente 361

Unterrepräsentation komplexer lexikalischer Einheiten 364

5.1.3 TENDENZEN AUF SATZEBENE (SATZSTRUKTUREN) 366

5.1.4 WEITERE TENDENZEN 371

5.2 ERWEITERUNGEN, EINSCHRÄNKUNGEN UND MÖGLICHE FORSCHUNGSBEREICHE 378

6 BIBLIOGRAPHIE 385

7 ANHANG 409

1 Einleitung

1.1 Untersuchungsbereich und Untersuchungsobjekt

In der vorliegenden Untersuchung soll ausgehend von einer produktorientierten, korpuslinguistischen Methode die akademische Textproduktion von Deutschlernern zu vergleichbaren Textprodukten von Muttersprachlern dargestellt werden, um Abweichungen im Sprachgebrauch festzustellen. Die ermittelten Tendenzen können als empirisch gewonnenes Datenmaterial die Grundlage weiterführender Untersuchungen in theoretisch und praktisch orientierten Forschungsbereichen bilden. Die vorliegende Arbeit wird im Rahmen der Textproduktion und der frequentiellen Untersuchung sprachlicher Phänomene durchgeführt.

Bestehende Untersuchungen von Lernersprache haben hauptsächlich das Ziel, theoretische Erklärung der Lernphasen und des Sprachstands zu geben. Dies kann jedoch u.E. nur auf der Grundlage eines Ansatzes geschehen, der auf empirisch überprüfbare Daten zurückgreift und interdisziplinäre Antworten sucht. In diesem Sinne ermöglicht die Bestimmung von charakteristischen Phänomenen der schriftlichen Äußerungen fortgeschrittener Fremdsprachlerner in akademischen Texten

eine klare Abgrenzung zu den Phänomenen, die vergleichbare Textprodukte von Muttersprachlern aufweisen. Die frequentiell vergleichende Beschreibung von konkreten sprachlichen Phänomenen, wie z.B. die Wortlänge oder die Verwendung von Wortarten, dient als Grundlage praxisorientierter Untersuchungen. So kann sie beispielsweise (analog zur Lernersprachenanalyse, vgl. Kapitel 3) zur Differenzierung des erreichten Sprachstands beitragen, d.h. des Niveaus, das der Lerner in der Fremdsprache erreicht hat, oder in der Didaktik didaktische Schwerpunkte lernergruppenspezifisch neu definieren helfen.

Der Vorteil einer frequentiellen Beschreibung realisierter sprachlicher Phänomene liegt in der Art der Phänomene, die dadurch beschrieben werden können. Während in der Fehleranalyse (vgl. S. 17) hauptsächlich sprachlich nicht korrekte Äußerungen (Fehler) den Gegenstand der Untersuchungen bildeten, ermöglicht es der vorgeschlagene Ansatz, sowohl realisierte (und potentiell fehlerhafte) als auch nicht realisierte (und dementsprechend potentiell nicht-fehlerhafte aber vernachlässigte) sprachliche Phänomene hervorzuheben und auszuwerten.

Die theoretische Grundlage der Untersuchung ist die Variation, d.h. die durch bestimmbare Faktoren bedingte uneinheitliche Verwendung sprachlicher Mittel, die aus einem statistischen Gesichtspunkt in Textprodukten einerseits Individualität, gleichzeitig aber auch Regelmäßigkeit erzeugt. Für

unterschiedliche Sprechanlässe würden nach Biber (1995a) einem einzelnen Sprecher verschiedene linguistische Formen zur Verfügung stehen, und verschiedene Sprecher einer Sprache könnten auf unterschiedliche Formen zum Ausdruck eines gleichen Sachverhalts zurückgreifen (Biber 1995a). Dieser relativen Simplifizierung der Möglichkeiten sprachlichen Ausdrucks steht die statistische Regelmäßigkeit gegenüber, die Variation auszeichnet, wie u.a. soziolinguistische (Turell 1995b: 21) und stilistische (Tuldava 1995) Untersuchungen bewiesen haben. Variation drückt sich sprachlich im Bereich der Aussprache, der Wortwahl und der Grammatik und nicht sprachlich je nach Kommunikationsziel, Beziehung der Sprecher untereinander, Geschlecht, kommunikativer Situation u.a. aus.

Die Variation ist Teil aller sprachlicher Produkte und dementsprechend auch in der Textproduktion von Lernern zu erkennen. Die teilweise systematischen Eigenschaften, die sie hier aufweist (Marcos Marín 1988: 61), beziehen sich auf vergleichbare Variationsmerkmale unter Lernern eines ähnlichen Niveaus (Maritxalar et al. 1996). Im Vergleich zur Variation zwischen Muttersprachlern kann die Variation der Lernersprache in ihrer konkreten Form (als mündliches oder schriftliches Produkt) identisch sein (Übereinstimmung), nicht identisch sein (was traditionell als „fehlerhafte Verwendungen" bezeichnet wurde, vgl. Kasper 1995) oder innerhalb der Übereinstimmung eine Verlagerung des Schwerpunkts der sprachlichen Phänomene aufweisen. Liegt der Schwerpunkt von Lernersprache frequentiell über den Produkten von

Muttersprachlern, stellen wir dies als Überrepräsentation oder positive Abweichung (hier nicht vom Standard, sondern von der frequentiellen Verwendung) dar, liegt er darunter, als Unterrepräsentation oder negative Abweichung.

Diese Untersuchung benötigt einer interdisziplinären Annäherung, theoretisch basiert einerseits auf der Fehleranalyse, der empirisch orientierten kontrastiven Analyse und der Lernersprachenanalyse, andererseits auf den Untersuchungen zur Textproduktion in der Zweitsprache (vgl. weiter unten). All diese Forschungsbereiche arbeiten mit unterschiedlicher Intensität auf der Grundlage der systematischen Variation der Lernersprache, deren Existenz ausschlaggebend für die theoretische Erklärung von allgemeingültigen Lernphasen oder die Ausprägung sprachlicher Phänomene ist. Die Einbeziehung der theoretischen Aspekte der genannten Forschungsbereiche war selektiv, weil verschiedene Einschränkungen zu berücksichtigen waren. Dazu gehören die Prozeßorientiertheit der Fehleranalyse, der kontrastiven Analyse und der Lernersprachenanalyse (im Gegensatz zur Produktorientiertheit des korpuslinguistischen Ansatzes) und die abweichenden Untersuchungsobjekte der Textproduktion in der Zweitsprache, die sich im Gegensatz zu den anderen Forschungsbereichen mit den Lernphasen einer Zweitsprache beschäftigt, nicht aber mit der Darstellung der Lernphasen in der Fremdsprache.

In den folgenden Abschnitten stellen wir die Ansätze dar, die sich mit Lernersprache, insbesondere mit der textuellen Produktion, befaßt haben. Trotz der Verbindung zu unserer Untersuchung ist auf die empirische Methode hinzuweisen, der in der vorliegenden Arbeit besonderes Gewicht beigemessen wird, da für die Beschreibung und Untersuchung frequentieller Phänomene der Lernersprache konkrete Daten benötigt werden. Eine Vorgehensweise wie die der Transformationsgrammatik, basiert auf der Intuition des Forschers, ist hier vollständig ungenügend, da dieser nicht intuitiv bestimmen kann, welche Phänomene mit welcher Frequenz vorkommen (vgl. McEnery 1996).

Fehleranalyse, Kontrastive Analyse und Lernersprachenanalyse

Untersuchungen zu mündlichen und schriftlichen Textprodukten von Lernern sind nicht neu, und für den Versuch, Lernphasen zu erklären, sind in der Forschung immer wieder empirische Daten verwendet worden.

Ein Rückblick auf die durchgeführten Untersuchungen in diesem Bereich zeigt, daß die Textproduktion von Lernern zwischen den 50er und den 70er Jahren Gegenstand der Untersuchungen der Fehleranalyse und der kontrastiven Analyse war, aber aus unterschiedlichen Perspektiven. Die kontrastive Analyse ging von der Annahme aus, daß Lernphasen durch den Einfluß der Muttersprache bedingt sind und hatte dabei das Ziel, „Lernschwierigkeiten und Lernerleichterungen zu identifizieren, die sich in Fehlern bzw. zielsprachlich

korrekten Sprachverwendungen manifestierten" (Kasper 1995: 263). Die Fehleranalyse hingegen versuchte, fremdsprachliche Lernphasen aus fehlerhaften Verwendungen zu erschließen, oft komplementär mit kontrastiven Untersuchungen verbunden, wenn Lernphasen nicht praktisch, sondern theoretisch dargestellt werden sollen (vgl. Nickel 1972a und Nickel 1972b). Zudem hatte die Fehleranalyse das Ziel, verbesserte Verfahren zur Fehleridentifizierung vorzuschlagen und Fehler aus dem Erwerbskontext zu beschreiben zur Ermittlung allgemeiner und spezifischer Fehlertheorien (Raabe 1980: 63f).

Beide Ansätze scheiterten größtenteils an methodischen Unstimmigkeiten. In der kontrastive Analyse wurde nicht die Diskrepanz zwischen Forschungsziel und Forschungsmethode überwunden, denn dekontextualisiertes Datenmaterial sollte als Erklärung für individuelle, sprachpsychologische Lernphasen herangezogen werden. Das methodische Problem der Fehleranalyse hingegen lag in der Einschränkung des verwendeten Datenmaterials, für das fehlerhafte Äußerungen berücksichtigt wurden, nicht aber kontextuelle Elemente und sprachlich korrekte Äußerungen, die ebenfalls Teil des Spracherwerbs bilden. Beide Ansätze arbeiteten auf der Grundlage einer Variation sprachlicher Schwierigkeiten, Erfolge und Fehler, die systematisiert werden und zu Erklärungen des Spracherwerbs führen sollte.

Dieses Ziel wurde Anfang der 70er Jahre von der Lernersprachenanalyse übernommen, die in enger Anbindung an

den Erstsprachenerwerb die Phasen des Fremdsprachenerwerbs darzustellen versucht (vgl. Tarone 1989 und Kasper 1995) und teilweise Forschungsinteressen der kontrastiven Analyse und der Fehleranalyse aufnimmt[1]. Die Lernphasen werden anhand der Produktion von Lernern untersucht und in Verbindung mit den Variablen gebracht, die den Prozeß der Produktion beeinflussen können und somit Variation erklären helfen. Diese Variablen, die in der Lernersprachenanalyse hauptsächlich auf kontextuelle und psychologische Einflüsse zurückzuführen sind (Sharwood Smith 1997: 16), werden von anderen Disziplinen hinzugezogen, was dem interdisziplinären Charakter der Lernersprachenanalyse entspricht (Sharwood Smith 1997: 173).

Textproduktion in der Zweitsprache

Eine weitere Tendenz zur Untersuchung von Lernersprache ist die der Textproduktion schriftlicher Äußerungen in der Zweitsprache, die in Deutschland v.a. seit den 80er Jahren an Texten von Migrantenkindern im Schulalter durchgeführt wurde. Ziel dieser Untersuchungen war meistens die Bestimmung der Sprachkompetenz dieser Schüler auf grammatischer Ebene, seltener jedoch wurde versucht, qualitative Erkenntnisse über die Fähigkeit zur Bewältigung bestimmter pragmatisch-

[1] Vgl. die Forderung in Raabe (1980), für die Erklärung der Interimsprache oder Interlanguage nicht nur Fehler, sondern auch andere Prozesse zu berücksichtigen, womit das Ende der Unabhängigkeit der Fehleranalyse bescheinigt wird.

kommunikativer Kompetenzen zu erlangen (Knapp 1997: 24). Abgesehen von den Zielen sind diese Untersuchungen wegen ihrer theoretischen Grundlage und methodischer Verfahren bedeutend.

Bis auf wenige Ausnahmen handelt es sich um produktorientierte Studien. Zur Erfassung der Komplexität des untersuchten Phänomens wären dabei theoretische Grundlagen interdisziplinär zusammenzufügen. Daß dies nicht immer in ausreichendem Maße unter Berücksichtigung der betreffenden Variablen oder Einflüsse auf die Textproduktion von mündlichen oder schriftlichen Äußerungen geschieht, wird in Knapp (1997) kritisiert.

Methodisch werden für die Arbeiten der Textproduktion in der Zweitsprache quantitative Aspekte hinzugezogen, die auch in allgemeinen korpuslinguistischen Untersuchungen Anwendung finden. In Untersuchungen zur Textproduktion in der Zweitsprache werden beispielsweise quantitative Phänomene wie Textumfang, lexikalische Variation und Verwendung von Wortarten zur Darstellung der Sprachkompetenz verwendet (vgl. Knapp 1997: 28-38). Im allgemeinen werden quantitative Aspekte jedoch nicht multivariant überprüft, d.h. auf gegenseitige Beeinflussungen untersucht, sondern vereinzelt dargestellt.

Die hier genannten Untersuchungen zur Lernersprache haben sowohl die theoretische Grundlage als auch die Methode der vorliegenden Untersuchung beeinflußt, weisen jedoch bedeutende

Unterschiede hinsichtlich des Forschungsziels auf, die zusammenfassend zum einen in der Prozeßorientiertheit liegen (Fehleranalyse, kontrastive Analyse und Lernersprachenanalyse), zum anderen in der Auswahl des untersuchten empirischen Materials (Textproduktion in der Zweitsprache).

Fehleranalyse, kontrastive Analyse und Lernersprachenanalyse greifen für die Erklärung der Phasen des individuellen Spracherwerbs teilweise auf kollektive empirische Daten zurück, die dazu beitragen sollen, die individuellen Lernphasen herauszufiltern. Besonders in der Fehleranalyse und der empirischen kontrastiven Analyse weist die Sammlung dieser Daten jedoch bedeutende methodische Schwachstellen auf, die v.a. in der Auswahl des Materials und seiner Repräsentativität begründet liegen. Aufgrund dieser methodischen Probleme können die Ergebnisse nicht verallgemeinert werden (vgl. 3.1.2). Zudem ermöglicht die Beschränkung der Fehleranalyse auf unkorrekt realisierte sprachliche Äußerungen keine differenzierte Untersuchung der vom Lerner erreichten Kompetenz (vgl. Knapp 1997: 34), die sich auch in zielsprachlich korrekten Äußerungen widerspiegelt.

Die ausgeprägte Prozeßorientiertheit der Lernersprachenanalyse hingegen, d.h. der Versuch der Darstellung der Lernphasen, und die Einbettung in einen theoretischen Rahmen zur allgemeinen Erklärung des Erwerbs von Fremdsprachen erschwert die Anwendung ihrer theoretischen Grundlagen auf

produktorientierte Untersuchungen, d.h. Untersuchungen zur Darstellung von gegebenen sprachlichen Verhaltensweisen. Zudem wird dadurch ebenfalls die Verwendung der Ergebnisse für Anwendungsbereiche begrenzt, die kollektiv-statistisch ausgerichtet sind, wie die Bestimmung des Sprachstands, didaktische Anwendungen u.a. Des weiteren seien nach Knapp (1995: 35) viele Erklärungsversuche der Lernersprachenanalyse reduktionistisch. Der besondere Wert, der z.B. Simplifizierungsstrategien gegeben würde, ermögliche zwar die Einbettung der Ergebnisse in ein klar umrissenes Theoriegebilde, verstelle jedoch gleichzeitig oft den Blick auf komplexere Erklärungsmuster. So kann die Herauslösung aus dem produktorientierten Handlungszusammenhang dazu führen, daß an der Produktion beteiligte Faktoren, wie z.B. die Aufgabenstellung, die zu einem Text geführt hat, nicht für die Erklärung berücksichtigt werden.

Gemeinsam hatten die Ansätze (Fehleranalyse, kontrastive Analyse und Lernersprachenanalyse), daß die Phasen des Erwerbs von Kompetenz in der Zielsprache untersucht werden (Kasper 1995). Wie Leech (1998) im Vorwort zu *Learner English on Computer* darstellt, läßt die Zuwendung zu vorwiegend mentalistischen Erklärungsversuchen des Lernens kommunikative Aspekte des Lernens und Lehrens in den Hintergrund treten. Produktorientierte Ansätze, wie die korpuslinguistischen Methoden, würden jedoch einerseits den bisher vorwiegend negativen Gesichtspunkt in einen positiven umwandeln helfen (*what did the learner get wrong* wird zu *what did the learner*

right) und andererseits die Lernersprache aus der Perspektive der *Über-* und *Unterrepräsentation* betrachten. So kann beantwortet werden, welche sprachlichen Phänomene der Lerner öfter und welche er seltener als ein Muttersprachler verwendet; dadurch erweitert sich die Untersuchbarkeit der Variation von Überschneidung und Fehler auf Unter- und Überrepräsentation.

Die Textproduktion in der Zweitsprache stellt im Gegensatz zur Fehleranalyse, kontrastiven Analyse und Lernersprachenanalyse einen produktorientierten Ansatz dar, dessen Erklärungsversuche und Verbesserungsvorschläge interdisziplinär geprägt sind und mit korpuslinguistischen Fragestellungen zur Über- und Unterrepräsentation vereinbar sind. Die Einschränkung der Verwendbarkeit dieses Ansatzes für die in dieser Untersuchung gesetzten Ziele liegt in der Abweichung der Untersuchungsobjekte und in methodischen Aspekten.

Die Mehrzahl der uns bekannten Untersuchungen der Textproduktion in der Zweitsprache widmen sich dem Zweitsprachenerwerb von Migrantenkindern in Deutschland, wie z.B. Kuhs (1989). Diese Gruppe zeichnet sich durch ihr Alter aus und durch das sprachliche Umfeld. In diesem Sinne weichen die Variablen, die entscheidend an der Textproduktion beteiligt sind, deutlich von der hier untersuchten Gruppe ab. Sie haben Einfluß auf die Textsorte, da Schüler selbstverständlich noch keine akademischen Texte produzieren,

und enthalten andere sprachliche Eigenschaften, die Texte von Kindern von denen von Erwachsenen unterscheiden. Die sprachlichen Phänomene, die von der Textproduktion in der Zweitsprache ermittelt werden, können jedoch auch in unserer Untersuchung auf ihre Relevanz hin überprüft werden.

Korpuslinguistik

Die systematische Ausprägung bestimmter sprachlicher Phänomene in der Textproduktion von schriftlichen Äußerungen von Deutschlernern soll hier anhand korpuslinguistischer Methoden frequentiell und distributiv dargestellt werden, wobei sich die Phänomene auf eine Über- oder Unterrepräsentation in der Anwendung beziehen sollen und nicht auf die von der Fehleranalyse untersuchten sprachlichen Abweichungen zu „korrekten" Äußerungen (Königs 1995a: 268), die anhand des Vergleichs mit der natürlichen Sprachkompetenz eines Muttersprachlers (z.B. des Forschers) identifizierbar sind und anschließend theoriebedingt in Kategorien eingeteilt werden können.

Das Interesse an dieser korpuslinguistischen Form der Abweichung soll Tendenzen aufdecken, die selbst bei sprachlich korrekten Äußerungen Unterschiede zwischen fortgeschrittenen Lernern und Muttersprachlern erkennen lassen, häufig aber nur subjektiv von der Lehrkraft identifiziert werden. Diese Abweichung kann auch Hinweise auf Äußerungen geben, die den

Anlaß der Textproduktion stilistisch, pragmatisch oder auf eine andere Weise verfehlen.

Analog zur Variation in muttersprachlichen schriftlichen Äußerungen können diese sprachlichen Tendenzen, die keine Abweichung mehr vom grammatisch korrekten Sprachgebrauch, sondern vom statistisch mittleren Sprachgebrauch darstellen, identifiziert und in Kategorien eingeteilt werden, wie ebenso bei der Fehleranalyse früher die Fehler identifiziert und klassifiziert wurden. Grundlage für die Erkennung dieser Tendenzen ist die Zusammenstellung von zwei Computerkorpora authentischer Texte, für deren Zusammenstellung die Variablen berücksichtigt werden mußten, die die Variation in der Textproduktion gewährleisten.

Die Zusammenstellung von zwei Computerkorpora authentischer Texte als Arbeitsziel ist bescheiden, insofern es keinen theoretischen Erklärungsversuch der Ursachen der Variation darstellt, sondern die Beschreibung der Variation (tendenzen- und lernstadienbedingt) anstrebt und zu bestimmen versucht, welches Datenmaterial für Untersuchungen vielfältiger Art anzuwenden sind. Nur schwer können die Symptome beschrieben und die Ursachen diagnostiziert werden, wenn nicht das dafür notwendige Datenmaterial vorhanden ist. Unabdingbar ist danach jedoch ein theoretischer Erklärungsversuch, um die Anwendung von Kenntnissen ohne die Erkennung der Ursachen zu vermeiden.

Die methodischen Grundlagen für die Zusammenstellung der benötigten Computerkorpora und die Untersuchung der

sprachlichen Phänomene, die zu interdisziplinären Erklärungen der Ursachen der Variation führen kann, ergeben sich aus der Methode der Korpuslinguistik, die im folgenden Abschnitt geschichtlich und methodisch dargestellt werden soll.

1.2 Methodische Grundlagen

Ende der 50er Jahre wurde der Grundstein zur Veränderung der Methode der bisherigen Korpuslinguistik gelegt. 1959 stellte Randolph Quirk sein Projekt *Survey of English Usage* vor (Leech 1991a: 8f); zusammen mit dem kurz darauf an der amerikanischen Brown-University von W. Nelson Francis und H. Kucera begonnenen *Brown Korpus* verwendete es erstmals elektronische Speicher- und Abrufverfahren zur Bearbeitung großer Textmengen (vgl. CCL 1996: 2). Der Unterschied zu den vorherigen Korpora bestand in der maschinellen Bearbeitbarkeit. Diese bedeutete aber nicht nur eine quantitative Neuerung, denn es stellte sich heraus, daß auch qualitative Untersuchungsmöglichkeiten betroffen waren.

Der Begriff Korpus wird laut Bußmann (1990) in sprachwissenschaftlichen Untersuchungen aufgefaßt als „endliche Menge von sprachlichen Äußerungen, die als empirische Grundlage für sprachwissenschaftliche Untersuchungen dienen", laut Crystal (1995: 410) als „eine für linguistische Analysen zusammengestellte repräsentative Sprachprobe [...]. Ein Korpus ermöglicht objektive Aussagen über Gebrauchshäufigkeiten und liefert Daten, die auch anderen

Forschern zugänglich sind". Korpora waren dementsprechend z.B. die Untersuchungen zum Spracherwerb gegen Ende des 19. Jahrhunderts (z.B. Aufzeichnungen der sprachlichen Äußerungen von Kindern), wobei immer nur ein Informant beobachtet wurde; gegen Ende der 20er Jahre dieses Jahrhunderts begannen dann Untersuchungen, die statistische Techniken anwendeten, damit die Untersuchungen zum Spracherwerb verallgemeinert werden konnten (vgl. McEnery 1996: 3f). Diese Zeit vor dem Brown Korpus wird mittlerweile als die Vorzeit der modernen Korpuslinguistik bezeichnet (vgl. Leech 1991: 8f; ANNO 1996).

Die dabei benutzten statistischen Techniken wurden auch vom amerikanischen Strukturalismus zur Erstellung repräsentativer Korpora angewendet (vgl. Pelz 1996: 84 und 87), da Linguisten wie Bloomfield und Harris diese für die Grundlage der Untersuchungen einer Sprache hielten. Der Übergang von empirischen sprachwissenschaftlichen Untersuchungen zu rationalistischen Untersuchungen, die den Sprachproduktionsprozeß wiederzugeben versuchen, stellte eine Umorientierung hinsichtlich der Datengewinnung dar (Schlieben-Lange 1991: 115) und wurde von den Kritiken Chomskys an korpuslinguistischen Methoden begleitet, die für ihn nur die Beschreibung der Performanz zulassen, nicht aber die Erklärung der Kompetenz.

Diese Kritiken hatten zur Folge, daß korpuslinguistische Verfahren zur Zeit der ersten Generation moderner Computerkorpora in den 60er und 70er Jahren wenig populär

waren. Trotzdem wendeten empirisch ausgerichtete Forschungsbereiche, zu denen auch die Untersuchung des Spracherwerbs oder die Lexikographie gehören, weiterhin Korpora in ihren Untersuchungen an. Dadurch sollte vermieden werden, was auch später noch als frustrierend empfunden wird (vgl. Schafer 1981: 12): daß Untersuchungen aufgrund von Texten, die unter besonderen Einflüssen entstanden und von wenigen Personen verfaßt wurden, zu universellen Generalisierungen gelangen.

Mit der Zeit hat sich die Auffassung durchgesetzt (vgl. McEnery 1996: 5ff), daß Korpuslinguistik einerseits als Methode nur teilweise für bestimmte Forschungsbereiche anwendbar ist und mit anderen Erklärungsansätzen ergänzt werden muß; und daß andererseits, abgesehen von der Erklärung der der Sprache zugrunde liegenden Regeln, andere, auch empirische Forschungen betrieben werden, wie zur Terminologie und Lexikographie oder zur Lernersprache.

Diese beiden Aspekte, verbunden mit der Fertigstellung der ersten modernen Korpora in den 60er Jahren, führte zur Weiterentwicklung der modernen korpuslinguistischen Methode, deren Werkzeuge anderen Disziplinen entnommen werden, wie der Statistik, der Soziolinguistik (z.B. Labov 1983), der computationellen und der statistischen Sprachwissenschaft usw.

Wie oben angemerkt, bestand die Neuheit der Ansätze von Quirk und Svartvik in der Möglichkeit der maschinellen Bearbeitung des textuellen Materials. Die Methode zur Zusammenstellung des

Materials hat sich in ihren Grundzügen nicht verändert, denn es handelt sich dabei um Verfahren, die der Statistik entlehnt wurden.

Für die Zusammenstellung eines Korpus werden Texte oder Textfragmente verwendet. Die *Expert Advisory Group on Language Engineering Standards* (EAGLES 1996a: 4) benutzt dafür den Begriff *pieces of language*, da Fragmente aufgrund ihrer fehlenden Abgeschlossenheit nach EAGLES nicht als Texte aufzufassen seien. In der vorliegenden Arbeit werden wir weiterhin im allgemeinen Sinn von Texten sprechen, auch wenn es sich um Fragmente handelt. Die Texte für das Korpus werden so zusammengetragen, daß sie als repräsentativ für eine größere Gruppe angesehen werden können; somit bilden die Texte eines Korpus eine Stichprobe (Schlobinski 1996: 26), die einen Induktionsschluß auf die Grundgesamtheit zuläßt. Diese Grundgesamtheit kann nach vielfältigen Kriterien zusammengestellt werden, mit der Einschränkung, daß sie homogene Eigenschaften aufweisen muß, anhand der sie bestimmt werden kann. So kann ein Korpus beispielsweise eine literarische Epoche darstellen oder die Textproduktion von Zeitungen.

Eine weitere Eigenschaft der Erstellung eines modernen Korpus ist die Aufarbeitung und Speicherung der Texte im Computer. Es erfolgt die Umsetzung der Texte in ein computerlesbares Format durch Eintippen oder, heute üblicher, Scannen, Kopieren und Bearbeiten von schon vorhandenen Dateien. Zur Auswertung

werden die Texte einer weiteren Aufarbeitung unterworfen, die es einem oft spezifisch dafür entwickelten Softwareprogramm ermöglicht, Information aus dem Korpus zu entnehmen. Nach Abschluß dieses Prozesses steht dem Forscher das Korpus für Untersuchungen zur Verfügung.

Die Vorgehensweise zur Zusammenstellung eines modernen Computerkorpus im Vergleich zur Zusammenstellung eines traditionellen Korpus hat grundlegende Folgen. Während für umfangreichere, korpuslinguistische Untersuchungen vor den 60er Jahren noch viele Forscher mit der Auswertung eines Korpus beauftragt werden mußten, bot der Computer die Möglichkeit, diese Prozesse zu automatisieren, aber auch die Auswertung gegen die Fehler zu schützen, die der Mensch bei der Bearbeitung begehen kann[2]. Dieses Mittel zur schnellen Überprüfung und zum Vergleich von empirischen Daten bedeutet ursprünglich nur eine quantitative Verbesserung durch Zeitersparnis, hat aber letztendlich zu grundlegenden quantitativen Veränderungen der Forschungsrichtungen geführt, weil dadurch die Zahl der real verwirklichbaren Untersuchungen und der untersuchbaren Wechselbeziehungen zwischen Einzelphänomenen gesteigert wird[3].

---

[2] Zu den Vorteilen der maschinellen Untersuchung versus der manuellen, siehe Mair (1991) und Marcus et al. (1993: 318f).

[3] Die neuen Anwendungsmöglichkeiten von Computerkorpora spiegeln sich in automatisierten probabilistischen Systemen zur grammatischen und syntaktischen Analyse und im kommerziellen Bereich nicht zuletzt in den zwar immer noch in ihren Anwendungsmöglichkeiten stark eingeschränkten,

Eine weitere Folge dieser Methode ist die Möglichkeit, große Textmengen zu speichern. Das sieht McEnery (1996) als gerechtfertigte Erwiderung auf Chomskys Kritik, der beanstandete, daß aufgrund eines endlichen Systems, die Manifestationen der Sprache, kein potentiell unendliches System beschrieben werden könne, wie es die Grammatik sei. Nach McEnery könne jedoch eine große Textmenge, wie die 100.000.000 Wörter des *British National Corpus* (BNC) Korpus, zu einer statistischen Annäherung an die unendliche Anzahl möglicher Sätze führen.

Zusätzlich beinhalten Computerkorpora die Möglichkeit der Wiederverwendbarkeit. Dies bedeutet, daß sie unabhängig von dem Grad der Spezifizität, mit dem sie zusammengestellt wurden, für weitere Forschungen verwendbar sind, die nicht unbedingt die gleichen Forschungsziele wie die Untersuchungen haben müssen, für die das Korpus zusammengestellt wurde.

Im Gegensatz zu den Korpora der ersten Generation, die bis ugf. 1980 reicht, werden Korpora der zweiten Generation, die ab 1980 zusammengestellt wurden (ANNO 1996), üblicherweise noch mit Zusatzinformation versehen, wie z.B. Information zu den Wortarten, aber auch zum Verfasser des Textes oder zum Textursprung. Die Möglichkeit der maschinellen Bearbeitung der Zusatzinformation anhand spezifisch entwickelter Software kann

---

aber schon brauchbaren Übersetzungsprogrammen wider, die sich teilweise aufgrund probabilistischer, einem Korpus entnommener Daten für die eine oder andere Übersetzung entscheiden.

so zu komplexeren Analysen führen als die, die ein traditionelles, nur auf Papier vorliegendes Korpus bietet.

Für die Auswertung werden Verfahren verwendet, die aus anderen Wissensbereichen stammen, hauptsächlich aber, da es sich mehrheitlich um quantitative Verfahren handelt, aus dem Bereich der mathematischen Statistik. Nichtsdestotrotz bieten diese Verfahren nur die empirisch-quantitative Grundlage der Auswertung, und es muß dieser quantitativen Phase immer eine qualitativ-wertende folgen (vgl. Tuldava 1996: 7).

Aufgrund all dieser Eigenschaften kann die moderne Korpuslinguistik definiert werden als linguistische Methode, die anhand von Computerkorpora, statistischer Methoden und grammatikalischen Richtlinien quantitative und qualitative Regelmäßigkeiten eines sprachlichen Phänomens darstellt. Ein Computerkorpus ist in diesem Sinne eine maschinell lesbare Sammlung von gesprochenen oder geschriebenen Texten oder Textfragmenten, die anhand von Kriterien zusammengestellt werden, die ihrerseits vom Forschungsziel abhängig zu machen sind.

Textsammlungen oder Korpora im traditionellen Sinn, d.h. im allgemeinen nur gedruckt vorliegende Korpora, werden weiterhin in zahlreichen Forschungsbereichen verwendet, erfüllen aber nicht immer die genannten Kriterien der Repräsentativität und maschinellen Lesbarkeit. Solche Bereiche sind beispielsweise die Fehleranalyse oder die Lernersprachenanalyse.

Die langsame Evolution der Korpuslinguistik zur eigenen Disziplin, die Methoden von anderen Bereichen übernommen hat, und die Erstellung von Verfahren für die Auswertung, bilden den Rahmen für die Möglichkeiten und Einschränkungen des methodischen Ansatzes, der der vorliegenden Arbeit zugrunde liegt. So können anhand eines Korpus nur auf der Grundlage des materiell vorliegenden textuellen Materials quantifizierbare Phänomene untersucht werden. D.h., wenn ein Computerkorpus nur die Texte enthält, aber keine weitere Information, z.B. Auszeichnungen zu den Wortarten, können diese dementsprechend nicht quantifiziert werden. Ähnliches ist auf weitere textinterne Informationsinhalte anzuwenden.

Die Quantifizierung an sich ist aber ebenfalls eingeschränkt. Zunächst ist das zu untersuchende Objekt qualitativ zu bestimmen, d.h. der Forscher benötigt eine Hypothese, die den Untersuchungsansatz bildet, unabhängig davon, ob er im Laufe der Untersuchung weitere Phänomene entdeckt. Die so gewonnenen quantitativen Daten müssen einer qualitativen Analyse unterworfen werden, denn ohne Interpretation entbehren die reinen Zahlen jeder Bedeutung. Diese augenscheinlich überflüssige Bemerkung findet leider ihre Bestätigung in zahlreichen methodisch einwandfreien Aufsätzen, in denen aber aus den Daten keine klare Deutung gezogen werden kann.

Korpuslinguistik ist zudem eine empirisch fundierte Methode. Untersuchte Objekte sind historisch-konkrete Äußerungen, schriftliche oder mündliche Texte, die in einem

Kommunikationskontext entstanden sind. In diesem Sinne können selbst anhand probabilistisch-statistischer Formulierungen nur aufgrund schon bestehenden Materials Aussagen gemacht werden, die aber im Unterschied zu den Naturwissenschaften nie zu einem Gesetzeswissen führen können (Schlieben-Lange 1991: 130), selbst wenn die Wahrscheinlichkeit für das Zutreffen der Ergebnisse sich 100% nähert. Hier ist jede Überbewertung der Methode zu vermeiden, die es nicht erlaubt, Gesetze aufzustellen, wohl aber dazu beiträgt, Tendenzen darzustellen.

Trotz dieser Einschränkungen sind Korpuslinguistik (verstanden als Methode) und vor allem Computerkorpora (verstanden als in besonderer Form angelegtes empirisches Datenmaterial) in vielen Bereichen anwendbar. Sie finden Anwendung in Arbeiten zur *Natural Language Processing* (NLP), z.B. in den Bereichen der automatischen Übersetzung, der Spracherkennung und der theoretischen Sprachwissenschaft oder der automatischen Bestimmung von syntaktischen Strukturen (Marcus et al. 1993: 313). In der Lexikographie ist Korpuslinguistik zu einer der wichtigsten Methoden avanciert, und Verlagshäuser wie Harper-Collins gehören, zumindest im englischsprachigen Raum, zu den wichtigsten Förderern von großen Computerkorpusprojekten. Aber auch in der Didaktik erscheinen Korpora immer öfter. Im *Computer Assisted Language Learning* (CALL) bilden Korpora die Grundlage vieler Programme und Hilfsmaterialien, im *English as Foreign Language* (EFL) und *English as Second Language* (ESL) werden sie zum Selbstlernen und zur Bearbeitung von

grammatischen Phänomenen, darunter syntaktische Strukturen, verwendet.

Korpora können sowohl für qualitative als auch für quantitative Untersuchungen verwendet werden. Für ihre quantitativ-statistische Auswertung wird eine Auswertungsmethode benötigt, für die zumindest Grundkenntnisse in Statistik vonnöten sind. In der qualitativen Anwendung bieten sie die Möglichkeit, in natürlichen Texten Belege für morphologische, syntaktische, lexikalische und andere Phänomene zu suchen, je nachdem, wie komplex das angewendete Korpus gestaltet und mit Zusatzinformation versehen wurde.

1.3 Ziele der Untersuchung, Hypothesen und Vorgehensweise

Die Beschreibung von sprachlichen Phänomenen, die kennzeichnend für die schriftlichen Textprodukte von Lernern sind und im Gegensatz zu den sprachlichen Phänomenen stehen, die vergleichbare Textprodukte von Muttersprachlern auszeichnen, stellt das Strukturierungselement der vorliegenden Arbeit dar. Sprachliche Phänomene sollen hier allerdings nicht wie so oft in Untersuchungen zur Lernersprache als fehlerhafte Äußerungen aufgefaßt werden, sondern innerhalb von sprachlich als „korrekt" einzustufenden geschriebenen Textprodukten auf die Phänomene bezogen werden, die einem Lernertext beispielsweise seine fehlende „Muttersprachlichkeit" verleihen. Unsere Hypothese war, daß der fremdsprachliche Charakter auch in der Häufigkeit der

Verwendung von sprachlichen Phänomenen zu suchen ist, die öfter oder seltener als bei Muttersprachlern erscheinen.

Die methodische Vorgehensweise der Korpuslinguistik paßte sich in dieser Hinsicht besser dem Ziel an als ähnliche Ansätze, wie z.B. die Lernersprachenanalyse. Regelmäßigkeiten in der Verwendung konkreter sprachlicher Phänomene können zwar induktiv anhand nur weniger authentischer Texte belegt werden, doch ist bei einem solchen Ansatz die Untersuchung auf sprachliche Phänomene beschränkt, die der Forscher subjektiv festlegt. Zudem kann ohne einen statistischen Ansatz nicht festgestellt werden, welche Phänomene mit welcher Intensität Verwendung finden und so, um auf das vorhergehende Beispiel zurückzugreifen, auch für die fehlende Muttersprachlichkeit verantwortlich gemacht werden können.

Demgegenüber bietet ein korpuslinguistischer Ansatz die Möglichkeit, sprachliche Phänomene auf nicht subjektiver Grundlage zu entdecken, d.h. anhand festgelegter Methoden festzustellen, welche Phänomene abweichende Verwendungen finden. Das ist allerdings nicht mit vollständiger Objektivität gleichzusetzen, da Aspekte wie die Bestimmung der Wortarten nicht aufgrund einheitlicher Grundlagen erfolgen kann (vgl. 3.2.2).

Zudem ermöglicht die Korpuslinguistik eine quantitativ-statistische Darstellung der Phänomene, und somit die Ermittlung vergleichbarer Werte für jedes Phänomen. Die so erhaltenen Daten sind verallgemeinerbar, d.h. sie sind

repräsentativ für die angegebene Gruppe von Informanten und stehen dementsprechend anderen Forschungsbereichen zur Interpretation zur Verfügung.

Die Ziele der vorliegenden Arbeit lauten zusammengefaßt:

- Welche sprachlichen Phänomene, d.h. sprachliche Mittel, werden trotz korrekten Gebrauchs unterschiedlich in geschriebenen Textprodukten von Lernern und Muttersprachlern verwendet?
- Besteht die Möglichkeit einer quantitativ-statistischen Darstellung von diesen sprachlichen Phänomenen?
- Trägt eine solche quantitativ-statistische Darstellung zum Vergleich von geschriebenen Textprodukten von Lernern und Muttersprachlern bei?
- Welche Wechselbeziehungen bestehen zwischen den einzelnen sprachlichen Phänomenen, d.h., bedingen sie sich gegenseitig?
- In welchen Forschungsbereichen können die Erkenntnisse umgesetzt und produktiv angewendet werden?

Bei der Zusammenstellung des Materials sollte gewährleistet werden, daß die Daten allgemeinen Charakter hätten, und dementsprechend die Ergebnisse der Untersuchung, die Anwendung korrekter Untersuchungsmethoden vorausgesetzt, ebenfalls

verallgemeinerbar wären. Dafür war innerhalb der korpuslinguistischen Methode abzugrenzen, welche Variablen die Textproduktion beeinflussen und somit Variation in den Textprodukten von Lernern und Muttersprachlern erzeugen. Je größer die Übereinstimmung der Variablen, desto vergleichbarer sind die Texte und desto mehr sprachliche Phänomene werden sie teilen; je weniger Variablen übereinstimmen, desto unvergleichbarer werden sie sein. So sind z.B. die charakteristischen sprachlichen Phänomene eines Gedichtes nur schwer mit denen eines Gesetzestextes vergleichbar.

Zur Bestimmung der Variablen können allerdings keine Einzeltexte herangezogen werden, da diese potentiell alle möglicherweise vorkommenden Variablen enthalten können und der systematische Einfluß derselben nicht innerhalb des Textes abgegrenzt werden kann (vgl. 2.1.2). Demgegenüber ermöglicht es ein kollektiv-statistischer Ansatz, Texte untereinander zu vergleichen und den Einfluß der möglichen Variablen als Tendenzen auszudrücken. Zur näheren Bestimmung bietet ein statistischer Ansatz zudem die Möglichkeit, potentielle Variablen auszuschließen, also die Gruppe der in Frage kommenden Variablen zu reduzieren und ihren gegenseitigen Einfluß zu begrenzen. Dieser Prozeß erfolgt anhand textinterner und textexterner Aspekte (vgl. S. 74), wobei die ersten sich auf Eigenschaften des Textes beziehen und die letzteren auf bekannte äußere Umstände, normalerweise zum Verfasser des Textes, wie Alter, Bildung, aber auch auf Interaktion, Beziehung der Teilnehmer in einem Briefkontakt

(Meurman-Solin 1995) u.a. Die Anwendung von statistischen Methoden erlaubt ferner die Streichung individueller Variablen, die identifiziert werden können, aber nur einen Text betreffen.

Ein weiterer Vorteil des statistischen Ansatzes stellt die Möglichkeit dar, sprachliche Phänomene zu quantifizieren und innerhalb des untersuchten Kollektivs zu vergleichen. Dieser Prozeß der Untersuchung der Phänomene führt zur Profilierung, der Abgrenzung von relevanten und nicht relevanten sprachlichen Phänomenen.

Verfahrensweise und Phasen der Untersuchung

Als Grundlage für die Profilierung der Phänomene der Lernersprache wurde ein Korpus authentischer akademischer Texte von fortgeschrittenen Germanistikstudenten (im folgenden auch Deutschlerner) der Universität Barcelona verwendet (Analysekorpus oder im folgenden auch Computerlernerkorpus, CLK), das dem Vergleich mit einem als Norm agierenden Referenzkorpus (oder Computermuttersprachlerkorpus, CMK) diente. Für die Zusammenstellung der beiden Korpora wurden theoretisch und praktisch zu berücksichtigende Voraussetzungen erarbeitet, da dazu in der Literatur keine umfassende methodische Anleitung vorhanden war. Anschließend wurden anhand verschiedener Untersuchungsmethoden potentiell relevante Phänomene profiliert, anhand der Korpora untersucht und deskriptiv dargestellt. Schließlich wurde der Versuch

unternommen, die Wechselbeziehungen dieser Phänomene untereinander zu bestimmen.

Für die Untersuchung konnte kein gesprochenes Material verwendet werden, da die Bearbeitung von Tonaufnahmen vielfältige ethische, methodische und technische Probleme mit sich führt. So beeinflußt möglicherweise die Bitte um Erlaubnis, ein natürlich vorkommendes Gespräch aufnehmen zu können, die Form des Gesprächs, während z.B. für das Problem der Transkription und der Hinzufügung visueller Elemente in der mündlichen Kommunikation noch keine einheitlich akzeptierte Lösung gefunden worden ist (vgl. Menge 1993: 15f). Ein weiterer wichtiger Grund zur Ausschließung mündlicher Korpora in dieser Untersuchung ist der Arbeitsaufwand, der dazu führt, daß in korpuslinguistischen Untersuchungen die Zusammenstellung geschriebener Korpora weiterhin dominiert[4]. Des weiteren wäre es hinsichtlich der Authentizität des Materials, das sich für empirische Untersuchungen „auf tatsächlich und in natürlichen Situationen produzierte Texte" (Rothkegel 1995: 180) stützen sollte, wahrscheinlich problematisch gewesen, sprachliche Aufnahmen von Lernern zu machen, da anzunehmen ist, daß der Hinweis auf die Aufnahme einer mündlichen Kommunikation die Produktion des Lernenden vorbelastet.

[4] "Die Faustregel, daß für eine Minute Aufnahme eine Stunde Transkription benötigt wird, scheint sich vielerorts eher als zu günstig gerechnet erwiesen zu haben" (Menge 1993: 17)

Die Texte der zusammengestellten Korpora

Die Untersuchung der sprachlichen Phänomene, die die Textproduktion von Lernern auszeichnen, ist stark von verschiedenen Variablen beeinflußt, die in der Literatur nicht immer gebührend berücksichtigt werden. Stimmung, Interaktivität, Rollenbeziehungen zwischen den Sprechern, Produktionskontext, Thema und Kommunikationsziel sind in Biber (1994: 201) Variablen, die jede Textproduktion, sei es von einem Muttersprachler oder Lerner, beeinflussen. Zu diesen werden weitere lernerspezifische gerechnet, wie Dozent oder didaktisches Material (Marcos Marín/Sánchez Lobato 1988: 46) oder Persönlichkeit, Alter zu Beginn des Lernprozesses (*Age of Arrival*), Muttersprache und andere (Dulay/Burt/Krashen 1982: 74ff).

Zur größtmöglichen Einschränkung dieser Variablen wurde bei der Zusammenstellung des Korpus (*Korpusdesign*) die Textauswahl so getroffen, daß textinterne und textexterne Faktoren automatisch zur Reduzierung dieser Variablen beitrugen, wodurch der benötigte Umfang des Korpus eingeschränkt werden konnte.

Als textinterne Faktoren wurden alle Aspekte aufgefaßt, die sich auf die geschriebene Form des Textes beziehen, in ihr ausgeprägt sind und unabhängig von nicht im Text vorhandener Information erkannt und gedeutet werden können, wie Orthographie, grammatische Eigenschaften, Stil, Form, Struktur, Thema u.a. Demgegenüber stellen textexterne Faktoren

Aspekte dar, die nicht aus dem Text zu erschließen sind, für Untersuchungen jedoch notwendig sein können und potentiell Einfluß auf die textinternen Faktoren haben, wie Aspekte zum Autor (Alter, Beruf, Geschlecht u.a.), Kommunikationsziel, Produktionskontext, Gattung und Textsorte[5] u.a.

Die Einschränkung dieser Faktoren ermöglicht es, aus den Texten die Variablen auszuschließen, die mit dem explizit nicht angewendeten Faktor verbunden sind und in den anderen mit hoher Wahrscheinlichkeit nicht erscheinen. So impliziert die textexterne Reduzierung auf den Faktor der Gattung akademische Texte, daß es höchst unwahrscheinlich ist, daß beispielsweise eine Bedienungsanleitung in einem der Texte erscheint, obwohl diese Möglichkeit auch nicht gänzlich auszuschließen ist[6].

Als textexterne Faktoren wurden Eigenschaften des Autors berücksichtigt, des Themas, des Produktionskontextes und der Textsorte. So sollte es sich um Texte handeln, die den angegebenen akademischen Textsorten zugeschrieben werden konnten, deren Thema im humanistischen akademischen Bereich lag und die von Studenten verfaßt wurden in einem Produktionskontext, der im Rahmen der Ausbildung als Aufgabe

---

[5] Zur problematischen Unterscheidung zwischen Gattung und Textsorte, vgl. 2.1.2.2.

[6] In einem der schließlich nicht in den Korpus einbezogenen Texte erschien zum Beispiel ein Gedicht, das nicht als Zitat gekennzeichnet werden konnte, weil es vom Verfasser des Textes stammte.

aufzufassen ist, d.h. der Student hatte genug Zeit und Mittel, um den Text zu verfassen und befand sich nicht in einer Prüfung.

Aber auch textinterne Faktoren wurden auf die Texte des Computerlernerkorpus angewendet. Dafür wurden verschiedene Preliminarversionen der Korpora benutzt, anhand der Phänomene wie die lexikalische Variation überprüft wurden. Stellte sich heraus, daß ein Text, der vorläufig in das Computerlernerkorpus aufgenommen worden war, im Vergleich zu den ermittelten Mittelwerten z.B. eine extrem hohe lexikalische Variation aufwies, also dementsprechend über einen sehr großen Wortschatz verfügte, wurde er näher untersucht. Dabei stellte sich oft heraus, daß er trotz der Streichung der vom Verfasser selbst als Zitate angegebenen Passagen weitere enthielt, die nicht vom Verfasser des Textes stammten.

Die Profilierung des zu untersuchenden Computerlernerkorpus, d.h. die Bestimmung der relevanten Phänomene, konnte allerdings nicht an sich selber erfolgen. Sie wurde anhand des Vergleiches mit einem Referenzkorpus durchgeführt. Dieses Referenzkorpus oder Computermuttersprachlerkorpus (CMK) wurde möglichst mit den gleichen Kriterien zusammengestellt und ließ ebenfalls die im Computerlernerkorpus ausgeschlossenen Variablen aus. Es dient der quantitativen Bestimmung der zu untersuchenden, potentiell relevanten Phänomene und dem

folgenden Vergleich zwischen ihrer Ausprägung im CLK mit der im CMK.

Damit wird das Referenzkorpus zur Norm (vgl. Leech 1998), denn es stellt die Grundlage der Untersuchungen dar, die Abweichungen im CLK im Vergleich zum CMK festzustellen versuchen. Dennoch muß darauf hingewiesen werden, daß das CMK als Norm nur deskriptiven, aber keinen präskriptiven Charakter hat, denn muttersprachliche Studenten sind nicht unbedingt das zu erreichende Ideal (Leech 1998: xix). Auf eine Wertung ist aber außerdem aufgrund unterschiedlicher Implikationen der gefundenen Werte zu verzichten. So kann der Wert der lexikalischen Variation, der Reichtum des Wortschatzes, nicht vorbehaltlos als Indiz für stilistische Kompetenz gewertet werden, da er eher als Indikator für den Grad von Synthese und Analyse zu werten ist (Tuldava 1996: 133). Die Realisierung von Autor und Leser im Text ihrerseits ist stark von textsortenspezifischen und kulturellen Aspekten geprägt, wie Petch-Tyson (1998: 107) darstellt, was aber ebenfalls keine direkten Rückschlüsse auf die stilistische Kompetenz zuläßt.

Das Computermuttersprachlerkorpus als Norm wurde für die Quantifizierung der Phänomene verwendet, deren Untersuchung im Computerlernerkorpus Abweichungen hinsichtlich des CMK ergeben hatte. Dabei wird Abweichung im Sinne einer quantitativen Darstellung als Unterschied zu einer meßbaren Entität aufgefaßt (vgl. Altmann 1995, Schlobinski 1996). Erst die quantitative Darstellung der Phänomene ermöglichte die

Untersuchung des CMK und des CLK auf anwesende sprachliche Phänomene und somit die Bestimmung der Abweichung in der Anwendung dieser Phänomene. Im Gegensatz zur Analyse von Fehlern wird hier nicht nach der Korrektheit oder Inkorrektheit einer sprachlichen Äußerung gefragt, sondern nach dem Grad, in dem sie in beiden Korpora angewendet wird.

Ferner konnte die Profilierung nicht a priori vorgenommen werden, da es sich um einen Wechselprozeß handelt, in dem ein Phänomen untersucht, mit andern verglichen, näher bestimmt, erneut untersucht und schließlich aufgenommen oder ausgeschlossen wird (vgl. 4.1). Grundlegend für die Profilierung war jedoch die Anwendbarkeit der jeweiligen Methode zur Bestimmung des Phänomens.

Die Untersuchung einiger allgemeiner statistischer Verfahren[7], die auf die Texte des Korpus angewendet werden konnten, hatte bewiesen, daß diese ungenügend sind, um eine Profilierung durchzuführen, die die Ziele dieser Arbeit zu erreichen erlaubte. Aus diesem Grund wurden die einzelnen Texte mit Wortartentags versehen, d.h. maschinell bearbeitbarer Information zu den Wortarten, wobei jedem Wort eine einzige Wortart zugeschrieben wird. Andere Auszeichnungssysteme, wie syntaktische oder semantische, wurden aufgrund der reduzierten technischen Anwendbarkeit ausgeschlossen. Durch diese Vorgehensweise konnte die Profilierung auf der Grundlage

[7] Unabhängig von der jeweiligen Sprache.

zweier Ebenen durchgeführt werden: die des reinen Textes (allgemeine Verfahren) und die der syntaktischen Wortarten (Part of Speech oder POS), die den Texten beigefügt waren.

Alle erkennbaren Phänomene mußten dementsprechend auf Analysen beruhen, die auf diese Information zurückgreifen konnte. In diesem Sinne waren mittels der verwendeten Konkordanzprogramme einzelne Wörter und rekurrente Wortverbindungen und statistische Phänomene hinsichtlich der Wort und Satzlänge direkt erkennbar. Gleichfalls konnten Phänomene der Wortartentags auch direkt untersucht werden.

Indirekt ermöglichten kombinierte Untersuchungen auch die Analyse einiger weiterer Eigenschaften. Dabei handelte es sich beispielsweise um Wortarten, die Indikatoren für syntaktische Strukturen darstellten, wie die Relativpronomina, die auf Relativsätze verweisen. Phänomene, die sich weder auf Wortebene noch auf Wortartenebene widerspiegelten, konnten dementsprechend nicht betrachtet werden. Uneingeleitete Nebensätze, Appositionen und andere syntaktische Strukturen benötigen einer syntaktischen Auszeichnung, die im englischen Sprachraum schon automatisiert möglich ist, im deutschen unseres Wissens aber noch nicht mit zufriedenstellenden Ergebnissen.

Problematisch sind dabei Funktionen, die grundsätzlich von einer Wortart abgeleitet werden, sich aber in mehreren oder gar in syntaktischen Strukturen manifestieren. Konjunktionen z.B. werden aus der traditionellen grammatischen Perspektive

als „Bindemittel" gewertet, als Wörter, die Sätze oder Satzteile miteinander verbinden. Ihre Eigenschaft, abgesehen von syntaktischen ebenfalls semantische Beziehungen herzustellen, teilen sie allerdings mit Elementen anderer Wortarten, wie den Adverbien. Da diese als Gruppe semantisch nicht ausschließlich als Verbindung gewertet werden können, ist die Untersuchung der Funktionen immer eng mit Konkurrenzformen zu verbinden (vgl. auch Altenberg/Tapper 1998).

Alle so erkennbaren Phänomene wurden an Preliminarversionen beider Computerkorpora untersucht, um auf diese Weise potentiell relevante Phänomene festzustellen. Die potentiellen Phänomene dienten zum einen der Untersuchung selbst, um Vorgehensweisen für die spätere Auswertung abzugrenzen, waren aber auch von grundlegender Wichtigkeit für das Korpusdesign, weil sie halfen, Textsorten näher anhand sprachlicher Phänomene zu bestimmen.

Relevante Phänomene wurden anhand des CMK bestimmt und mit dem CLK vergleichend dargestellt. Viele dieser Phänomene sind extrem frequent, wie die Wortlänge oder die Verwendung des Artikels, was das Fehlen des probabilistischen Ansatzes in den Statistiken erklärt; andere hingegen so selten, daß andere Eigenschaften dieser Phänomene als die statistischen in den Vordergrund traten. Eine eingehende mathematisch-statistische probabilistische Darstellung ist demzufolge für spätere Einzeluntersuchungen zu leisten, in denen das

Untersuchungsobjekt anhand der vorliegenden Information näher bestimmt wird.

1.4 Terminologische Aspekte

Zunächst sollen hier noch einige Hinweise auf die verwendete Terminologie gegeben werden. In Zusammenhang mit korpuslinguistischen Fragen ist die Bezeichnung bedeutend, die die verwendeten Texte als Ganzes erhalten. Wir unterscheiden dabei zwischen Textsammlung, Korpus und Computerkorpus. Textsammlung bezeichnet eine Art traditionelles Korpus, das aus Texten gebildet wird, die nicht mit dem Ziel zusammengestellt wurden, eine bestimmte Gruppe aus statistischer Sicht darzustellen. Dementsprechend können die Untersuchungsergebnisse der Forschungen, die anhand von Textsammlungen betrieben werden, nicht verallgemeinert werden, d.h. sie beziehen sich nur auf die verwendeten Texte, nicht aber auf andere Texte, auch wenn diese vergleichbar sind (vgl. S. 142; Ellis 1994: 49). Das bedeutet, daß eine Untersuchung anhand dieser Texte zu einem Ergebnis kommen kann, eine identische Untersuchung jedoch, die eine andere Textsammlung verwendet, zu anderen.

Korpus hingegen bezieht sich auf eine Menge strukturierter Texte, die das Kriterium der Repräsentativität berücksichtigt. Dementsprechend sind die Forschungsergebnisse verallgemeinerbar, d.h. eine Untersuchung zur Lernersprache anhand eines Deutschlernerkorpus von englischen

Hochschulstudenten sollte bei der Verwendung identischer Methoden zu keinen anderen Ergebnissen kommen, selbst wenn alle Texte durch neue ersetzt werden, die in identischer Form ausgewählt wurden.

Computerkorpus ist eine Erweiterung des Begriffs Korpus und fügt diesem die Anforderung der maschinellen Lesbarkeit und der Auszeichnungen hinzu, d.h. ein Computerkorpus muß im Gegensatz zu einem „normalen" Korpus maschinenlesbar sein und über Zusatzinformation verfügen, die vom Namen des Verfassers hin bis zu syntaktischen Auszeichnungen reicht.

Im Bereich der zu untersuchenden Elemente beziehen wir uns mit sprachlichem Phänomen auf jedes linguistisch meßbare Element, das zur quantitativen Darstellung herangezogen werden kann. Dabei ist bei uns ein relevantes sprachliches Phänomen jenes, das sich von seiner Frequenz in einem der Computerkorpora her von dem anderen unterscheidet. Unter Frequenz verstehen wir hier die Zahl der Belege eines sprachlichen Phänomens. Frequenz ist mit Verteilung eng verbunden, die im adjektivischen Gebrauch zu „distributiv" wird (naheliegend war die Verwendung von *Distribution* und *distributionell*, was allerdings zu Überschneidungen mit theoretisch-distributionellen Aspekten des Distributionalismus geführt hätte). Verteilung bezeichnet die Verwendung eines sprachlichen Phänomens an einer bestimmten Stelle eines Textes oder Satzes, aber in den statistischen Darstellungen ebenfalls die Frequenz von Einzelelementen einer Gruppe.

Innerhalb der sprachlichen Phänomene wurden Wortarten anhand eines Vorschlags zur maschinell verwendbaren Auszeichnung von Texten ausgezeichnet (Schiller et al. 1995), die für spezifische Untersuchungen (vgl. 197) leicht abgeändert wurden; auf diese syntaktisch gewonnenen Wortarten beziehen wir uns im allgemeinen sowohl mit dem Begriff syntaktische Wortart als auch einfach nur mit „Wortart". Weitere terminologische Aspekte werden in dem jeweiligen Abschnitt thematisiert.

1.5 Gliederung der Arbeit

In Kapitel 2 werden die Kriterien zur Zusammenstellung eines Computerkorpus dargestellt, denn Computerkorpora, erst 1960 eingeführt, sind jung im Vergleich zu anderen methodischen Vorgehensweisen. Da die Erstellung von Computerlernerkorpora dementsprechend noch jünger ist, bestand die Notwendigkeit, die Kriterien kritisch zu untersuchen.

Kapitel 3 ist der Beschreibung der Methode zur Zusammenstellung unseres Computerlernerkorpus und Computermuttersprachlerkorpus gewidmet, der als Vergleich dient.

In Kapitel 4 wird die Methode zur Ermittlung der zu untersuchenden sprachlichen Phänomene dargestellt und anschließend die Untersuchung der relevanten Phänomene im Computerlernerkorpus und im Computermuttersprachlerkorpus unternommen.

Im darauffolgenden abschließenden Kapitel 5 wird auf der Grundlage des deskriptiven Datenmaterials der Versuch unternommen, die sprachlichen Phänomene auf Wechselbeziehungen zu untersuchen. Daneben werden die Einschränkungen unseres Ansatzes zur Untersuchung von Lernersprache besprochen, die Möglichkeiten zum Ausbau der Untersuchungsmöglichkeiten und die eventuellen Anwendungsgebiete des korpuslinguistischen Ansatzes zur Untersuchung von Lernersprache.

Im Anhang schließlich befindet sich ein Abdruck von Schiller et al. (1995), in dem die Wortarten beschrieben werden, die die Grundlage eines großen Teils unserer quantitativen Untersuchungen darstellen. Er wurde der Deutlichkeit halber vollständig übernommen.

2 Korpusdesign

In dem vorliegenden Kapitel folgt eine kritische Darstellung der Beiträge der modernen Korpuslinguistik zur Bestimmung der Voraussetzungen und Eigenschaften zur Erstellung von Korpora. Schon im vergangenen Jahrhundert wurden korpuslinguistische Untersuchungen durchgeführt, doch erfuhr die Anwendung von Korpora für sprachwissenschaftliche Untersuchungen 1960 eine entscheidende Wende mit dem Brown Korpus, 1964 an der Brown University erstellt (Francis 1979). Allgemein anerkannt ist dies als das erste Korpus, das spezifisch für die Benutzung mit einem Rechnersystem konzipiert wurde und somit als Vorläufer moderner Textkorpora anzusehen ist (Zierl 1997), d.h. der Computerkorpora. In sprachwissenschaftlichen Untersuchungen waren bis zu diesem Augenblick Korpora verwendet worden, die als empirische Datengrundlage fungierten und Anwendung in Forschungsbereichen fanden, die empirisches Textmaterial, geschrieben oder gesprochen, als methodische Grundlage für Untersuchungen verwendeten. Beispiele für Forschungsbereiche, die auf Textkorpora zurückgriffen, waren,

wie bekannt und in Kapitel 1 kurz erwähnt, der Strukturalismus, der Distributionalismus u.a. In diesem Sinne stellt die Anwendung von Korpora eine Methode dar, um das im Forschungsbereich gesetzte Ziel zu erreichen. Mit der Einführung des maschinenlesbaren Formats im Brown Korpus wurde eigentlich eine Methode wiederbelebt, die schon im vorigen Jahrhundert von Lexikologen und Spracherwerbsforschern angewendet wurde, die Korpuslinguistik. Sie wird hier verstanden als Methode zur statistischen Auswertung von textuellem Material, die es ermöglicht, sprachliche Belege quantitativ darzustellen und somit Abstand vom Einzelfall zu nehmen. Wir unterscheiden hier (vgl. McEnery 1996) zwischen früher Korpuslinguistik, deren Erkenntnisse aus der manuellen Auswertung von Korpora entstehen, und der modernen Korpuslinguistik, die für die Bearbeitung auf Computerkorpora, wie das Brown Korpus, zurückgreift.

In der Korpuslinguistik wendete sich das Interesse von der Untersuchung *einzelner Äußerungen* ab und der *statistischen Auswertung* des Sprachgebrauchs zu. Die statistische Komponente der modernen Korpuslinguistik erforderte die Berücksichtigung von Kriterien, die einen Rückschluß auf die durch die statistischen Ergebnisse dargestellte Grundgesamtheit ermöglicht. Die Übersetzung statistischer, der Mathematik entsprungener Methoden für die Sprachwissenschaft, die seit den 60er Jahren von Labov für soziolinguistische Untersuchungen vorgenommen worden war (vgl. z.B. Labov 1983), konnte zum Teil von der modernen Korpuslinguistik übernommen

werden. Im Gegensatz zu theoretisch fundierten Disziplinen wie der Soziolinguistik kann Korpuslinguistik als Methode dagegen keinem eigenen, ausschließlichen Forschungsobjekt und -ziel entsprechen und hat jene zu berücksichtigen, die der betreffenden Disziplin entspringen, für die sie angewendet wird. In dieser Hinsicht ist die Möglichkeit ausgeschlossen, allgemeingültige Kriterien zur Erstellung eines Computerkorpus zu geben. Die der Methode eigene Beschränkung erklärt, daß keine umfassende Darstellung zu in der Korpuslinguistik verwendbaren Kriterien vorliegt. Eine Darstellung der anwendbaren Kriterien ist unserer Ansicht nach jedoch im Falle einer Abgrenzung des Anwendungsbereiches der Korpuslinguistik möglich. Eine solche Abgrenzung stellt der tatsächliche Anwendungsbereich der Korpuslinguistik auf textuelle Untersuchungen von mündlichen und schriftlichen Äußerungen dar, da diese trotz abweichender Forschungsziele zahlreiche Überschneidungen aufweisen. In diesem Kapitel werden Aspekte aus Werken, die korpuslinguistische Fragen aus theoretischer Sicht bearbeiten, und aus Arbeitsberichten strukturiert und besprochen. Die so erhaltenen Aspekte bilden die theoretische Grundlage für die Zusammenstellung der beiden Computerkorpora, die das empirische Material für die vorliegende Untersuchung darstellen, d.h. des Computerlernerkorpus und des Computermuttersprachlerkorpus.

Zu den aus methodischer Sicht in allen Computerkorpora vertretenen Kriterien zählen

- die *Zielsetzung* als allgemeine Grundlage wissenschaftlicher Untersuchungen (vgl. 2.1.1),
- die *Repräsentativität*, die die Gewährleistung methodischer Anforderungen darstellt, der Anwendung statistischer Methoden entspricht und durch Strukturierung erreicht wird (vgl. 2.1.2), und
- die materielle Möglichkeit der *maschinellen Lesbarkeit*, die das Hauptabgrenzungsmerkmal von Computerkorpora darstellt (vgl. 2.1.5).

Zu diesen Grundkriterien sind in Abhängigkeit von ihrer Umsetzung weitere Kriterien zu rechnen, was sich vor allem in der Auswahl der Variationskriterien widerspiegelt (vgl. 2.1.2.1 bis 2.1.2.5), aber auch in praktischen Aspekten wie der Textlänge (2.2.1.2) oder den Auszeichnungen (2.2.2).

Initiativen wie die Empfehlungen der *Expert Advisory Group on Language Engineering Standards* (EAGLES) oder von der *European Language Ressources Association* (ELRA) stellen in letzter Zeit einen Schritt zur verallgemeinernden Darstellung und Anwendungsform der Grundkriterien und der Zusatzkriterien dar, beziehen sich jedoch größtenteils auf statistische Aspekte (Repräsentativität und Strukturierung).

Außerdem sind seit Ende der 80er und Anfang der 90er Jahre mehrere allgemeine Einführungen in Buchform zum Thema

Korpuslinguistik entstanden (z.B. Sinclair 1991a, McEnery 1996, Barnbrook 1996), und noch zahlreicher sind die in Fachzeitschriften wie *Computational Linguistics* oder Sammelbänden wie Aarts/Haan/Oostdijk (1993) oder Oostdijk/Haan (1994) erschienenen spezifischen Artikel zum Thema. Die Beschreibung der theoretischen und praktischen Grundlagen zur Erstellung eines Computerkorpus in diesen Werken ist meistens an das Beispiel eines Computerkorpus gebunden, wie im Falle von Oostdijk/Haan (1994). In nur wenigen Werken, wie McEnery (1996), werden sie unabhängig von der praktischen Anwendung allgemeingültig darzustellen versucht. Weitere Ausnahmen bilden die schon erwähnten Organisationen ELRA und EAGLES. ELRA veröffentlicht Arbeitsberichte und Empfehlungen, die anhand von Umfragen und empirischer Beobachtung zusammengestellt werden, aber fast ausschließlich zu Auszeichnungen und deren Format (wie bei Baker et al. 1997 oder bei McEnery et al. 1998). EAGLES veröffentlicht ebenfalls Empfehlungen zu korpuslinguistischen Themen wie Korpustypologie (EAGLES 1996a) und Texttypologie (EAGLES 1996b). Dabei handelt es sich um Vorversionen, deren endgültige Fassung von dem Ergebnis noch nicht abgeschlossener Untersuchungen und Auswertungen zum Thema abhängt.

Wie schon angemerkt gibt es noch keine umfassende Beschreibung der möglichen Kriterien, die die Erstellung, Bearbeitung und Auswertbarkeit eines Computerkorpus betreffen. Daher ist es notwendig, rekurrente Kriterien, d.h. Kriterien, die wiederholt in Korpusprojekten berücksichtigt werden,

strukturiert darzustellen. Dafür sind die zwei Aspekte zu berücksichtigen, die Leech in einer seiner Definitionen von Computerkorpora nennt, wenn er diese als aus empirischem textuellen Material bestehende Einheiten kennzeichnet (Leech/Fligelstone 1992: 115ff). Die Kriterien zu

- korpusbezogenen Aspekten, die sich auf die Einheit des Korpus beziehen und zu
- textbezogenen Aspekten, die das einzelne textuelle Material betreffen

werden in den folgenden Abschnitten behandelt.

Die korpusbezogenen Kriterien beziehen sich somit auf die formalen Voraussetzungen, die ein Computerkorpus unabhängig von den individuellen Texten erfüllen muß, um als statistisch gültige Repräsentation einer Grundgesamtheit zu gelten, während die textbezogenen Kriterien auf der niedrigeren Ebene des individuellen Textes und des Inhalts anzusiedeln sind und die Aufnahme eines Textes in ein Korpus betreffen.

Bei den korpusbezogenen Kriterien handelt es sich in erster Linie um theoretische Vorüberlegungen, um die Zielsetzung und um die Repräsentativität. Durch allgemeine theoretische Vorüberlegungen werden im Vorfeld alle wichtigen Aspekte geklärt, die Einfluß auf die Erstellung eines Korpus nehmen können. Zusammen mit der Zielsetzung wird der Erstellung des Korpus somit ein Rahmen gegeben, auf den alle Entscheidungen zurückführbar sind (vgl. 2.1.1). Die Repräsentativität

hingegen entspringt der Einbeziehung statistischer, quantitativer Methoden in den Zielen der Forschungsbereiche, die Korpuslinguistik als Methode und Computerkorpora als empirische Grundlage anwenden (in der Lexikographie, der Didaktik, der Interkulturalität, der Fachsprachenforschung, der maschinellen Sprachverarbeitung usw.[8]) (vgl. 2.1.2).

Die statistischen Methoden ergeben sich ihrerseits aufgrund des im allgemeinen großen Umfangs des Datenmaterials, das qualitativ nicht vollständig ausgewertet werden kann, sondern erst nach vorhergehender Delimitierung der zu untersuchenden Phänomene.

Repräsentativität wird durch die Anwendung von zwei weiteren, untergeordneten Kriterien ermöglicht:

- Strukturierung, d.h. die Aufteilung des Korpus in homogene Gruppen und
- Variation, anhand der die oben genannten Gruppen ermittelt werden. Sie ermöglicht die Beschreibung von Faktoren oder Variablen, die Aspekte der Textproduktion einheitlich beeinflussen (wie der Sprachstand, der Einfluß auf den „Wortschatz" eines Lerners hat) und so die Grundlage der Strukturierung bilden.

[8] Ausführlicheres zu den Anwendungen, vgl. Barnbrook 1996: 131-147.

Beide Kriterien sind eng miteinander verbunden und ergeben sich aus der Anwendung statistischer Methoden zur Auswahl der Einzeltexte. Im Normalfall handelt es sich dabei um die Methode der stratifizierten Auswahl. Sie ermöglicht eine Einteilung der Individuen (Texte) in Gruppen (Strukturierung), und die anschließende Auswahl innerhalb dieser Gruppen von Einzelelementen mittels vorbestimmter Kriterien (Variationskriterien) aufgrund der Annahme, daß Unterschiede durch den Einfluß erkennbarer Eigenschaften erklärt werden können.

Die detaillierte Beschreibung und Auswertung der einzelnen korpusbezogenen Aspekte folgt in 2.1. Die textbezogenen Aspekte hingegen bestimmen Kriterien auf der Ebene der Einzeltexte. Hier ist die Festlegung der Auswahlkriterien der einzelnen Texte zu nennen, d.h. die statistische Vorgehensweise, in der potentiell verwendbare Texte einbezogen werden. Dabei handelt es sich um die Textlänge, die Bearbeitung, d.h. die Auszeichnungen und ihr Format u.a., die ausführlicher in 2.2 behandelt werden.

2.1 Korpusbezogene Aspekte

Die Aspekte, die in diesem Abschnitt zusammengefaßt werden, beziehen sich wie gesagt auf ein Korpus als Einheit, die folglich eine Grundgesamtheit darstellt, die ihrerseits nicht vollständig untersucht werden kann. Wir stimmen mit Biber (1990) überein, daß theoretische Vorüberlegungen und

Zielsetzung obligatorische Oberkriterien sind, denn sie bestimmen die Realisierung der weiteren Aspekte; ohne präzise Vorstellungen dazu können andere Aspekte keine konkrete Form annehmen (vgl. 2.1.1).

Die Repräsentativität bezieht sich ebenfalls auf ein Korpus als statistisch bedingte Einheit. Da sie absichert, daß Rückschlüsse von dem Korpus auf die Grundgesamtheit gezogen werden können, die es repräsentieren soll, ist sie eng mit den statistischen Methoden verbunden, die es ermöglichen, dieses Ziel zu erreichen: die Strukturierung und die Variationskriterien (vgl. 2.1.2).

2.1.1 Allgemeine theoretische Vorüberlegungen und Zielsetzung

Biber (1993a: 243) verweist treffend darauf, daß die Planung eines Korpusprojekts auf der Klärung der Aspekte aufbaut, die Einfluß auf die Zusammenstellung des Korpus nehmen können. Die zu berücksichtigenden Aspekte können dabei hierarchisch gegliedert werden, mit der von Leech und Fligelstone (1992: 116) genannten Zielsetzung als übergeordnetes Element, das die Umsetzung der weiteren Kriterien beeinflußt, wie z.B. die Bestimmung der Variationskriterien und der Selektionskriterien für die Auswahl der Texte.

In ihrer modernen Auffassung sind Computerkorpora keine „Anhäufungen" oder Textsammlungen mehr (wie EAGLES 1996a: 5 hervorhebt), deren Wert in der Bereitstellung von beliebig

ausgewählten textuellen Äußerungen liegt. Ihre Form wird bestimmt von dem Ziel, mit dem sie geschaffen wurden, so daß klare Unterschiede bestehen zwischen Korpora z.B. für soziolinguistische oder stilistische Untersuchungen. Unabhängig davon, ob ihre Anwendung für Forschung, Lehre oder andere Ziele bestimmt ist, beeinflußt die Zielsetzung den Realisierungsgrad der weiteren Aspekte, die für die Zusammenstellung eines Computerkorpus erfüllt werden können: von der Auswahl der Variationskriterien bis hin zur Bestimmung der im allgemeinen statistisch bedingten Selektionsmethode der Texte. Dabei ist zu beachten, daß sich ein Computerkorpus durch seine Wiederverwendbarkeit auszeichnet. In diesem Sinne definiert Sampson (1991a: 192) ein Computerkorpus folgendermaßen: „A machine-readable language corpus (of any type) is a general-purpose resource, capable of being put to uses that were never envisaged by its creators".

Die Möglichkeit, ein Computerkorpus für andere Ziele zu nutzen als für die, für die es ursprünglich geschaffen wurde, ergibt sich aus der Überschneidung und Kompatibilität der angewendeten Kriterien. Während bei Untersuchungen anhand spezifisch für eine Untersuchung zusammengestellter Computerkorpora die Anwendbarkeit der Untersuchung auf das entsprechende Computerkorpus vorauszusetzen ist, muß bei einer Wiederverwendung bereits bestehender Computerkorpora die Kompatibilität der Ziele der Untersuchung zu den Kriterien, mit denen es zusammengestellt wurde, überprüft werden.

Die zweifache Anwendungsform von Computerkorpora führt im Bereich ihrer Erstellung zu zwei Methoden, die von den Zielen der Untersuchung bestimmt werden. *Computerkorpora mit spezifischer Zielsetzung* werden für konkrete Untersuchungen zusammengestellt und stehen trotz der allgemeingültigen Wiederverwendbarkeit nur beschränkt für weitere Untersuchungen bereit. *Computerkorpora mit allgemeiner Zielsetzung* werden nicht mit einer konkreten Forschungsabsicht erstellt, sondern sollen einem breiten sprachwissenschaftlichen Untersuchungsbereich empirisches, korpuslinguistisches Datenmaterial zur Verfügung stellen.

Computerkorpora mit allgemeiner Zielsetzung

Computerkorpora mit allgemeinen Zielen entsprechen dem Versuch, anhand der Anwendung möglichst kompatibler oder standardisierter Kriterien (vgl. Ide 1996) in der Erstellung eines Korpus ein breites Spektrum möglicher sprachwissenschaftlicher Forschungen abzudecken. Das ist der Fall des British National Corpus (Burnard 1995), des Brown Korpus (Francis 1979) und seines deutschen Pendants, des Limas Korpus (Institut für Deutsche Sprache 1998d). Mit dem Computerkorpus wird so der Forschungsgemeinschaft empirisches Datenmaterial angeboten, das sich durch einen hohen Grad an Wiederverwendbarkeit auszeichnet und versucht, die aus der Anwendung sehr allgemeiner Kriterien entstehenden Ungenauigkeiten in der Zusammenstellung durch Menge

aufzuheben. Aus diesem Grund bestehen heutige Computerkorpora mit solchem Ziel aus mehreren Millionen Wörtern.

Der Umfang allgemeiner Computerkorpora stellt in vielen Fällen eine Einschränkung für die intensive Bearbeitung und Auszeichnung der Texte dar. Dies ergibt sich aus dem hohen Arbeits- und Kostenaufwand bei dem Editieren und Auszeichnen (Lenders 1993; Leech 1998). So werden die Auszeichnungen oft auf die Bestimmung von Wortarten (*Part of Speech*) anhand meistens traditioneller Kriterien und mehr oder weniger umfangreicher morphologischer Elemente beschränkt. Die Auszeichnung von Wortarten und morphologischer Kategorie kann für eine Vielzahl von Sprachen automatisiert vorgenommen werden (Xerox 1997a und Xerox 1997b). In manchen Fällen werden automatisch ausgezeichnete Texte nicht manuell, sondern halbautomatisch revidiert, so z.B. das BNC (British National Corpus 1997b). Problematisch dabei ist, daß eine ungenügende Revision der Auszeichnungen eine starke Beeinträchtigung der Forschungsergebnisse mit sich ziehen kann. Besondere Auszeichnungsformen, wie die semantischer Strukturen, werden in den meisten Fällen vollkommen beiseite gelassen. Im *Helsinki Corpus of English Texts* (Kytö 1998) z.B. wurde nur der diachronische Teil mit syntaktischen Auszeichnungen versehen, der das *Penn-Helsinki Parsed Corpus of Middle English* (*PPCME*) bildet (PPCME 1998).

Allgemeine Computerkorpora sind häufig Nationalkorpora, die den Standard für eine gegebene Sprache, wie z.B. das schon

erwähnte British National Corpus, und zudem eine Standardreferenz für sprachwissenschaftliche Untersuchungen in der betreffenden Sprache darstellen wollen. Sie werden oft aus schon bestehenden Computerkorpora zusammengestellt oder nutzen das Textmaterial, das andere Korpora bereitstellen, indem sie es umstrukturieren. Die Struktur großer allgemeiner Korpora ermöglicht im allgemeinen die Erstellung von Teilkorpora, die nach eventuell notwendiger Bearbeitung als spezifische Computerkorpora Verwendung finden. Die weiter bearbeiteten Texte eines Teilkorpus können erneut in das allgemeine Korpus eingefügt werden, so daß ein Teil des allgemeinen Korpus mit zusätzlichen Auszeichnungen versehen ist, die für eine spätere Nutzung bereitstehen.

Beispiele für allgemeine Korpora wären im englischen und amerikanischen Sprachraum das schon erwähnte *British National Corpus* (BNC), das *Brown* und das *London-Oxford-Bergen* Korpus (LOB), und im deutschen Sprachraum verschiedene Korpora des IdS Mannheim, insbesondere das erwähnte Limas Korpus, das analog zum Brown Korpus konstruiert wurde (Institut für Deutsche Sprache 1998d).

Die Anwendungsmöglichkeiten eines Korpus mit allgemeinen Zielen sind vielfältig und entstehen oft erst nach der Zusammenstellung; so wird z.B. das Brown Korpus immer noch für Untersuchungen zur sprachlichen Variation zwischen dem amerikanischen und dem britischen Englisch angewendet oder zum diachronischen Vergleich, z.B. mit dem *The Freiburg - LOB*

Corpus of British English (Hundt et al. 1998). Während die Kompatibilität der für die Zusammenstellung angewendeten Kriterien einen hohen Grad an Wiederverwendbarkeit ermöglicht, hat der große Umfang allgemeiner Korpora oft den entgegengesetzten Effekt: die Reduktion der Auszeichnungen und ihrer Qualität, bedingt durch Umfang und Arbeitsaufwand, mindert die Wiederverwendbarkeit, die jedoch weiterhin höher ist als im Falle spezifischer Korpora.

Computerkorpora mit spezifischer Zielsetzung

Korpora mit einer spezifischen Zielsetzung entspringen einem konkreten Forschungsziel und werden im allgemeinen im Rahmen eines Forschungsprojekts angelegt. Sie zeichnen sich aufgrund der bei der Erstellung angewendeten Kriterien durch einen relativ geringen Grad an Wiederverwendbarkeit aus, der sich aus der Einschränkung der Kompatibilität zu anderen Forschungszielen ergibt. Da mit ihrer Hilfe bestimmte, festgelegte Phänomene untersucht werden sollen, wird die Bearbeitung der Texte auf das Wesentliche reduziert[9], was zu einer weiteren Einschränkung der Wiederverwendbarkeit beiträgt. Beispiele solcher Korpora sind die CHILDES Datenbank zur Untersuchung des Lernprozesses der englischen Sprache bei

[9] Die Bearbeitung von Korpora, d.h. ihre Strukturierung und Aufarbeitung für die maschinelle Auswertung, erfolgt aus dem Gesichtspunkt der Realisierbarkeit, insofern nur für die Untersuchung notwendige Kriterien berücksichtigt werden.

Kindern (Childes 1998), oder im deutschen Sprachraum das Wende-Korpus, anhand dessen „die außerordentlichen Veränderungen, die im Herbst 1989 zur ‚Wende' in der DDR und 1990 schließlich zur Wiedervereinigung Deutschlands führten", dokumentiert werden sollen (Institut für deutsche Sprache 1998c).

Wie anfangs angemerkt wurde, bestimmt das Ziel, mit dem ein Korpus zusammengestellt wird, in jedem Fall alle weiteren Kriterien. Dazu gehören vor allem die Repräsentativität des Korpus, durch die Anwendung von statistisch bedingten Variationskriterien garantiert, die Abgeschlossenheit, die für die Unterscheidung zwischen synchronischen und diachronischen Untersuchungen relevant ist, und der Umfang des Korpus, der Einfluß auf die Repräsentativität hat. Diese Kriterien legen die Grundform eines Korpus fest und sind obligatorisch bei der Erstellung zu berücksichtigen (vgl. EAGLES 1996a). Alle weiteren Kriterien sind an das Forschungsziel gebunden und werden in Abhängigkeit von ihm realisiert oder nicht.

2.1.2 Repräsentativität, Strukturierung und Variationskriterien

Wie schon erwähnt führt die Einbeziehung der statistischen Komponente in Computerkorpora dazu, daß Repräsentativität eine der Grundeigenschaften ist[10]. In mathematisch-statistischen

[10] Vgl. Biber (1988); zu Ausnahmen, beispielsweise im Falle literarischer Computerkorpora, vgl. Barnbrook (1996: 24).

Verfahren im allgemeinen und insbesondere im korpuslinguistischen Bereich ist Repräsentativität die Forderung, daß ein Teil (Stichprobe) als reduzierte Darstellung eines größeren Ganzen (Grundgesamtheit) angesehen werden kann, dessen Eigenschaften es in reduzierter Form widerspiegelt (vgl. Schlobinski 1996: 26). Im sprachwissenschaftlichen Bereich werden reduzierte Gruppen für empirische Untersuchungen zusammengestellt, denn „das zu untersuchende linguistische oder literarische Material steht äußerst selten vollständig zur Verfügung bzw. ist sehr schwer zu ermitteln" (Altmann 1995: 3). Repräsentativität stellt dementsprechend die Grundlage empirisch ausgerichteter Forschungen zur *Performance* dar (Francis 1980: 192). Aus diesem Grund ist sie heute unumgänglich für Disziplinen wie die Lexikographie (dazu Meijs 1992: 147). Sie findet Berücksichtigung in den Korpora vieler sprachwissenschaftlicher Forschungsrichtungen, wie des Strukturalismus (Pelz 1996: 87), stellt sich aber für Computerkorpora aufgrund des Umfangs der verwendeten Datenmengen als zentrales Kriterium heraus, das nicht mehr nur optional Verwendung finden kann.

Die statistischen Verfahren der Stichprobentheorie sehen zwei Methoden zum Erlangen von Repräsentativität vor, wobei beide die Auswahl der Individuen (Texte) betreffen, die den Teil (Korpus) bilden werden, der als Abbild des Ganzen (Sprachebene) fungiert: die Zufallsstichprobe und die Quotenauswahl (Schlobinski 1996: 25) (vgl. 2.2.1.3). Von den

beiden Methoden ist die Quotenauswahl zumindest ebenso repräsentativ wie die Zufallsstichprobe und ist dieser dementsprechend vorzuziehen. Der Nachteil besteht in einer größeren Anforderung an die Verfahren zur Zusammenstellung eines Computerkorpus. Die Quotenauswahl setzt eine Strukturierung des Korpus voraus, d.h. die Einteilung der Texte in Gruppen, die hinsichtlich vergleichbarer Eigenschaften geordnet sind. Und die Bestimmung dieser Gruppen erfolgt anhand von Variablen, die Einfluß auf die Realisierung sprachlicher Äußerungen nehmen, den Variationskriterien.

Die aus der mathematischen Statistik entlehnten Verfahren zum Erlangen von Repräsentativität wurden schon in den ersten Computerkorpora der 60er Jahre angewendet, wie dem Brown Korpus, vor allem aber in Disziplinen wie der Soziolinguistik. Die Soziolinguistik stellte tatsächlich einen wichtigen Schritt in der Anpassung der mathematisch-statistischen Verfahren an sprachliche Äußerungen dar; gleichzeitig erleichterte sie die späteren theoretischen Arbeiten der korpuslinguistischen Forscher, die diese Erkenntnisse ihrerseits auf geschriebene Textkorpora anwendeten. Daher werden wir jetzt jene Verfahren erörtern, die aus der Soziolinguistik übernommen wurden.

In der Soziolinguistik wird die zu untersuchende Sprachgemeinschaft, die Grundgesamtheit, nach sogenannten Variationskriterien strukturiert. Variationskriterien stellen die Variablen dar, die Einfluß auf die sprachlichen Phänomene

nehmen. So kann z.B. das Alter ein Variationskriterium bei der Untersuchung der Anwendung von dialektalen Ausdrücken sein, unter der beispielhaften Annahme, daß diese mehr von älteren als von jüngeren Sprechern benutzt werden. Allgemein in der Soziolinguistik angewendete Variationskriterien, d.h. jene, die in einem großen Teil soziolinguistischer Untersuchungen Anwendung finden, wenn auch nicht unbedingt in allen, sind beispielsweise Alter, Geschlecht oder Beruf. In Abhängigkeit vom gesetzten Forschungsziel sind viele andere möglich, wie Ausbildung, Zweisprachigkeit, Muttersprache u.v.a.

In der korpuslinguistischen Forschung stellt sich die Bestimmung von allgemeingültigen Variationskriterien als komplexer als in der Soziolinguistik heraus. Soziolinguistik weist als Disziplin relativ einheitliche Forschungsziele auf. Korpuslinguistik als Methode hingegen ist von den zahlreichen Forschungszielen der Forschungsbereiche, die korpuslinguistische Methoden anwenden, abhängig. In jedem Einzelfall ist eine individuelle Ermittlung der möglichen Einflüsse auf die sprachliche Produktion erforderlich. Schon die Unterscheidung zwischen Korpora von mündlichen und geschriebenen Äußerungen führt zu stark voneinander abweichenden Variationskriterien. Denn während in Korpora der gesprochenen Sprache zahlreiche Variationskriterien der Soziolinguistik hinzugezogen werden, wie z.B. im *Wellington Corpus of Spoken New Zealand English* (Holmes et al. 1998), erscheinen in den Korpora geschriebener Sprache Kriterien, die

zusätzlich Aspekte der geschriebenen Sprache, wie Textsorten, pragmatische Aspekte u.a. betrachten.

Die Anzahl der Variationskriterien schwankt in Abhängigkeit von dem Forschungsziel und dem erwünschten Repräsentativitätsgrad, doch gilt ebenso wie bei anderen statistischen Verfahren, daß eine sehr große Anzahl an Variationskriterien zwar zu einer detaillierteren Strukturierung führt, aber gleichzeitig auch der Arbeitsaufwand für das Zusammentragen des textuellen Materials wächst (Schlobinski 1996). Demnach muß für jedes Variationskriterium eine Anzahl an Texten gefunden werden, die eine spätere Auswahl so ermöglicht, daß sie die Proportionalität in der Grundgesamtheit widerspiegelt. Dies kann besonders bei der Zusammenstellung von historischen Korpora zu Problemen führen, da nicht immer genügend Material vorhanden ist.

Beispiel für einen solchen Fall ist das ZEN-Projekt (*Zurich English Newspaper Corpus;* Nevalainen 1996: 125), ein Computerkorpus englischer Zeitungstexte von 1660 bis Ende des 18. Jahrhunderts. Repräsentativität soll hier durch ein aleatorisches Auswahlverfahren garantiert werden, wobei die Texte in 10-Jahres-Intervallen ausgewählt werden. Dabei ergab sich eine Überrepräsentation von Texten einer einzigen Londoner Zeitung; der Versuch eines Ausgleiches anhand anderer Zeitungstexte stellte sich aber als problematisch dar, insofern keine andere Zeitung kontinuierlich im Laufe des

angegebenen Zeitraumes veröffentlicht wurde. Hier wäre wahrscheinlich eine Einschränkung des Forschungsziels auf die einzige kontinuierlich veröffentlichte Zeitung angebracht gewesen. Für ein ähnliches Problem findet die Forschungsgruppe um Josef Schmied eine pragmatischere Lösung (vgl. Siemund 1997: 61). Für das *Lampeter Corpus of Early Modern English Tracts* (Texte zwischen 1640 und 1730, also ungefähr die gleiche Zeitspanne wie die Texte des ZEN-Projekts) wird die Population so definiert, daß jeder Autor nur einmal vertreten sein darf; auch in diesem Fall stammt die Mehrzahl der Veröffentlichungen aus London, aber bei der Auswahl der Verfasser wird auf deren Herkunft geachtet, so daß durch die Einbeziehung des Variationskriteriums *Verfasser* das Korpus nicht nur für London, sondern auch für den Rest des Landes als repräsentativ gelten kann.

Die Methoden zur Bestimmung der Variationskriterien, die zur Strukturierung eines Korpus führen und dadurch seine Repräsentativität begründen, haben seit der Zusammenstellung des ersten Computerkorpus, des Brown Korpus, eine Entwicklung durchgemacht, die mittels der Einteilung in textexterne und textinterne Variationskriterien dargestellt werden kann (Fleskes 1996: 26ff).

Unter textexternen Variationskriterien werden (situative und kommunikativ-funktionale, vgl. Fleskes 1996) Variablen verstanden, die Einfluß auf die Realisierung eines Textproduktes haben, jedoch nicht aus dem Text selbst zu

erschließen sind. Dazu gehören alle den Verfaser betreffende Eigenschaften (Alter, Geschlecht, Bildung usw.). Unter textinternen Variationskriterien werden (grammatische, vgl. Fleskes 1996) Variablen zusammengefaßt, die Einfluß auf die Realisierungsform eines Textproduktes haben, und aus dem Sprachsystem, d.h. dem Text, abgeleitet werden, wie z.B. statistische Eigenschaften.

Die Bestimmung der Variationskriterien kann nicht theoretisch a priori vorgenommen werden, da sie sich aus dem jeweiligen Forschungsziel ergibt. Aber ebenso wie in der Soziolinguistik, bei der sich rekurrent verwendete Variationskriterien herauskristallisiert haben (Geschlecht, Alter usw.), erscheinen auch bei der Erstellung geschriebener Computerkorpora wiederholt Variationskriterien. Es folgt nun eine Übersicht über die Faktoren, die als Variationskriterien aufgrund ihrer Textbezogenheit wiederholt Anwendung in Computerkorpora finden und in Abhängigkeit von der Zielsetzung in der Korpuslinguistik miteinander kombiniert werden. Dabei erscheinen auch einige der Variationskriterien, die spezifisch für die gesprochene Sprache sind, denn in vielen Computerkorpora werden Texte gesprochener und geschriebener Sprache miteinander kombiniert.

Die Anwendung der Variationskriterien durch den Forscher beruht auf den genannten externen und internen Eigenschaften der Texte; ihre Strukturierung erfolgt jedoch in Abhängigkeit von den Eigenschaften hierarchisch und/oder parallel. So kann

ein Korpus hierarchisch zunächst nach der Obergruppe gesprochene und geschriebene Textproduktion eingeteilt werden, und danach innerhalb der gesprochenen Produktion Kriterien zu den Textsorten anwenden (vgl. Tabelle 1).

Hierarchische Strukturierung eines Korpus								
Ebene 1: Produktionsform	Gesprochene Textproduktion				Geschriebene Textproduktion			
Ebene 2: Textsorte, abhängig von Ebene 1	Interview	Telefongespräch	Private Konversation	usw.	Zeitungsartikel	Literatur	Protokoll	usw.
Ebene 3: Abhängig von Ebene 2	Spontan vs vorbereitet usw.	Öofentlich vs privat usw.	Bekanntschaftsgrad, Vertrauensgrad	usw.	Politik Wirtschaft, Sport usw.	Roman, Theater usw.	Ereignisprotokoll, usw.	usw.

TABELLE 1

Ein Korpus kann aber ebenfalls einerseits in die Obergruppe gesprochene und geschriebene Textproduktion eingeteilt werden, danach aber parallel in geographische Varianten, so daß sowohl der gesprochene als auch der geschriebene Teil geographisch strukturiert werden (vgl. Tabelle 2). Diese Form der Strukturierung, in der eine hierarchisch hochgestellte Einteilung nach Textsorten übergangen wird, erfolgt im allgemeinen nur dann, wenn das Computerkorpus im Vorfeld auf eine einzige Textsorte beschränkt wurde (so z.B. das *Lampeter Corpus of Early Modern English Tracts*, Schmied 1998).

	Parallele Strukturierung eines Korpus					
Ebene 1: Produktionsform	Gesprochene Textproduktion			Geschriebene Textproduktion		
Ebene 2: Geographische Variation	Norden	Mitte	Süden	Norden	Mitte	Süden
Ebene 3: Soziale Variation	Niedrig	Mittel	Gehoben	Niedrig	Mittel	Gehoben

TABELLE 2

Zur Darstellung der Variationskriterien für die schriftliche Textproduktion wurden aus Arbeitsberichten rekurrente Kriterien entnommen und dargestellt. Hier werden neben den Variationskriterien gegebenenfalls auch die Ziele der Untersuchungen angegeben, wenn die Bestimmung der Variationskriterien von der theoretischen Zielsetzung abhängig ist. Für andere, nicht in den als Grundlage dieses Abschnittes dienenden Arbeitsberichten vorgesehene Ziele kann eine weitere, fast unendliche Zahl an Variationskriterien bestimmt werden, so daß es unmöglich ist, alle näher zu beschreiben.

Es folgt die Darstellung der hier aufgelisteten Variationskriterien:

- Produktionsform
  - Mündlich vs. schriftlich
- Texteinteilung
  - Gattung
  - Textsorte
- Textursprung
  - Geographische Aspekte
  - Temporale Aspekte

- Verfasser
  - Population
  - Idiolekte
- Kommunikationsform
  - Privat vs. öffentlich
  - Veröffentlicht vs. nicht veröffentlicht

Diese Variationskriterien werden einzeln in den folgenden Abschnitten behandelt.

2.1.2.1 Produktionsform

Die großen Nationalkorpora wie das *British National Corpus* oder das deutsche Limas-Corpus verbinden oft mündliche und schriftliche Texte. Man spricht dann von gemischten Computerkorpora. Der Versuch, Sprache hinsichtlich der zwei grundlegenden Produktionsformen darzustellen, führt zur Unterscheidung der Variationskriterien mündliche und schriftliche Textproduktion. Die Einteilung gesprochen vs. geschrieben steht auf der obersten Ebene der Variationskriterien, und uns ist kein Computerkorpora mit gesprochenen und geschriebenen Texten bekannt, das nicht dieses Variationskriterium berücksichtigt.

Die Variationskriterien, die auf gesprochene Sprache angewendet werden, sollen hier nicht weiter erläutert werden, denn die vorliegende Arbeit beschränkt sich auf die Untersuchung der Produktion geschriebener Sprache. Die die gesprochene Textproduktion beeinflussenden Faktoren können in Werken zur Soziolinguistik nachgelesen werden, aber auch in

korpuslinguistischen Richtlinien wie die der Spoken Language Working Group EAGLES zur Strukturierung Computerkorpora gesprochener Sprache (Eagles 1996c). Dennoch scheint es interessant, darauf hinzuweisen, daß Computerkorpora gesprochener Sprache im Vergleich zu Computerkorpora geschriebener Sprache nicht sehr zahlreich sind, und daß der Anteil gesprochener Texte in gemischten Computerkorpora ebenfalls sehr reduziert ist (Eagles 1996a). Das *British National Corpus* enthält beispielsweise nur ugf. 10 Prozent transkribiertes gesprochenes Material (British National Corpus 1997a). Diese Beschränkung ergibt sich auf materieller Ebene aus dem extrem hohen Arbeitsaufwand, der für die Bearbeitung eines gesprochenen Korpus anfällt, und auf organisatorischer Ebene aus den noch ungenügenden Richtlinien zur Transkription und technischen Bearbeitung des Materials (vgl. EAGLES 1996a: 7 und 8f). Die gesprochene Äußerung (im Gegensatz zur geschriebenen Äußerung, vgl. Briz Gómez 1998: 19ff und 24) stellt aufgrund der Vielfalt der Variablen, die die Realisierung mündlicher Texte beeinflussen, eine weitere Einschränkung dar. Geschriebene Texte hingegen tendieren zu einer größeren Formalität (Briz Gómez 1998: 26f), besonders bei einer Reduzierung der Textsorten des Computerkorpus, wie z.B. auf akademische Texte, so daß z.B. kolloquiale Varianten größtenteils ausgeschlossen werden können.

Die starke Entwicklung geschriebener Computerkorpora im Vergleich zu den mündlichen ist verbunden mit der paradoxen Tatsache, daß der Mensch zwar für primär die Verarbeitung

gesprochener Sprache geschaffen ist, der Computer aber besser mit geschriebenen Texten umgeht (Leech 1998)[11].

Durch die Besetzung der obersten Ebene der Variationskriterien mit der Einteilung in gesprochen vs. geschrieben (so das LOB-Korpus und das BNC), wird die Einführung weiterer Variationskriterien notwendig, die zu einer präziseren Strukturierung und größeren Repräsentativität des Korpus beitragen. Dabei ist zu beachten, daß, wenn für ein Variationskriterium genügend Texte zur Verfügung stehen, aufgrund des Variationskriteriums ein Subkorpus aus dem Computerkorpus extrahiert werden kann (z.B. das LOB Korpus), oder ein geschriebenes Korpus zu nur einer Textsorte von mehreren (z.B. Biber 1990) und somit für Untersuchungen bereitsteht. Im Folgenden werden die spezifisch für Korpora geschriebener Sprache angewendeten Kriterien besprochen, die teilweise aber auch in Korpora mündlicher Sprache Anwendung finden, wie z.B. geographische Aspekte.

[11] Der gleiche Mechanismus kann aber ebenfalls zur Erklärung herangezogen werden, warum bestimmte Forschungsrichtungen stärker als andere betrieben werden, denn die Aufbereitung der schriftlichen Korpora mit Zusatzinformation, die nicht in automatisierter Weise vorgenommen werden kann, unterliegt ähnlichen Restriktionen wie die mündlichen Korpora: der hohe Kosten- und Arbeitsaufwand wirkt hier abschreckend.

2.1.2.2 Texteinteilung

Die Strukturierung von Computerkorpora hat seit dem Brown Korpus eine Entwicklung durchgemacht, die von subjektiv-traditionellen Ansätzen über Versuche einer objektiv-merkmalbestimmten Einteilung (wie z.B. Biber 1988) bis zur heutigen kompromiß-pragmatischen Lösung der Einbeziehung mehrerer Ansätze reicht (z.B. EAGLES 1996a).

Man darf nicht vergessen, daß in den 60er Jahren den Erstellern von Computerkorpora noch nicht das theoretische Werkzeug der Textlinguistik und der Textsorteneinteilung zur Verfügung stand. Dementsprechend basierte die Strukturierung des Brown Korpus auf externen, formalen Kriterien (Francis/Kucera 1979: 2f), die sich stark an den traditionellen Gattungsbegriff anlehnten. Unterschieden wird im Brown Korpus auf der obersten Ebene zwischen *Informative Prose* und *Imaginative Prose*. Beide Gruppen sind in Subtypen eingeteilt, wie z.B. *Press: Reportage, Press: Editorial* usw. in der ersten Gruppe und *General Fiction, Mystery and Detective Fiction* usw. in der zweiten.

Folgende Computerkorpora berücksichtigten für die Strukturierung der Texte zusätzlich textinterne Kriterien, bei denen allerdings im engen Sinne sprachliche Aspekte unberücksichtigt blieben. Ein erstes Beispiel dafür ist das London-Oslo-Bergen Korpus, das mit dem Ziel der Vergleichbarkeit zum Brown Korpus die gleiche Strukturierung mit den gleichen Textsorten anwendet. Ohne auf spezifisch

linguistische Kriterien zurückzugreifen, finden interne Kriterien Anwendung in der Form einer Strukturierung der einzelnen Texte nach inhaltlichen Elementen. Zusätzlich wurden für das Korpus die Kriterien zur Auswahl der Texte näher bestimmt (Johansson et al. 1978: 5).

Am Ende der 80er und Anfang der 90er Jahre wurden interne sprachliche Kriterien zur Strukturierung eines Korpus in Textsorten entwickelt. Die Notwendigkeit dieser Entwicklung ergibt sich aus der wachsenden zahl der Korpusprojekte und der damit steigenden Spezifizität der Korpora. Während in großen Korpora, wie dem Brown, für die Darstellung eines umfassenden sprachlichen Bereichs (in diesem Fall *American Contemporary Prose*) ebenso umfassende und gleichzeitig wenig definierte Textsorten angewendet werden konnten, da das fehlende Detail in der Definition durch die Menge aufgehoben wurde (die Auswahl wurde demnach mehr nach dem Prinzip der Zufallsstichprobe gefällt als durch Quotenauswahl; vgl. Schlobinski 1996: 26), muß bei der Bearbeitung eines spezifischen Korpus näher bestimmt werden, welche Eigenschaften die Textsorten kennzeichnen.

Die Entwicklung der Kriterien zur Bestimmung dieser Eigenschaften erfolgt zum einen auf rein praktischer Weise im Rahmen der Forschungsprojekte, die an der Zusammenstellung eines Korpus arbeiten. Zahlreiche Beispiele solcher Definitionen befinden sich in den Projekten, die in der Reihe ICAME dargestellt werden.

Aber auf theoretischer Ebene wachsen ebenfalls die Bemühungen um eine nähere, maschinell durchführbare Bestimmung der abgrenzenden Eigenschaften von Textsorten, die stark der Textlinguistik verpflichtet ist. Die Forschungen von Douglas Biber sind hier grundlegend, insofern er versucht, die aufgrund externer Kriterien gewonnenen Phänomene einer Textsorte mit den Besonderheiten zu kombinieren, die sie in linguistischer Hinsicht offenbaren kann, wie wir bereits in den einleitenden Bemerkungen kurz dargestellt haben.

Die Mehrzahl der heutigen Strukturierungsverfahren zur Erstellung eines Computerkorpus berücksichtigen eine Einteilung nach Textsorten. Ausnahmen bilden nur jene Computerkorpora, die sich von Anfang an auf eine konkrete Textsorte beschränken und somit keine weitere Einteilung zur Abgrenzung verschiedener Textsorten benötigen.

In der Textlinguistik gibt es keinen Konsens hinsichtlich der Bestimmung von Textsorten, doch kann zusammenfassend festgehalten werden, daß die meisten Ansätze auf eine Kombination von textexternen und textinternen Kriterien zurückgreifen (Gülich/Raible 1972; Isenberg 1983; Adamzik 1995).

In der Korpuslinguistik hat sich eine Einteilung der Texte nach der Unterscheidung zwischen textinternen und textexternen Eigenschaften durchgesetzt, was sich in der Unterscheidung

zwischen Gattung und Textsorte widerspiegelt. Hauptvertreter dieses Ansatzes ist Douglas Biber[12]. Die Kombination von externen und internen Aspekten erfolgt bei ihm anhand der Unterscheidung zwischen Gattung und Textsorte, die die Grundlage der Strukturierung des Korpus bildet. Gattung ist das oberste Variationskriterium, wobei jeder Gattung in einer zweiten Strukturierungsphase verschiedene Textsorten entsprechen (Biber 1993a: 244f) (vgl. Tabelle 3).

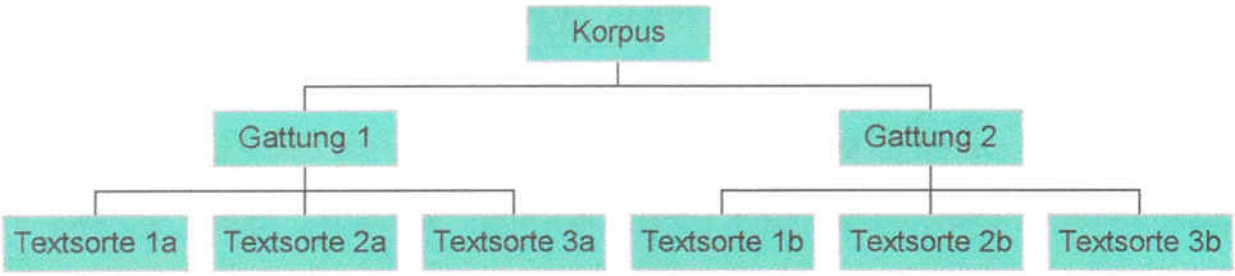


TABELLE 3

Gattung wird in diesem Ansatz anhand externer Kriterien definiert, die sich auf das kommunikative Ziel des Autors oder des Sprechers beziehen (Biber 1988a: 68ff). Demnach kann Gattung präziser als „situationsdefinierte Textkategorie" bezeichnet werden (Biber 1993a: 244).

Die Anwendung externer Kriterien zur Strukturierung wurde schon in den ersten Korpora angewendet und wird aufgrund einer

[12] Vgl. Biber 1988a und Biber 1995a.

fehlenden klaren Bestimmung der Textsorten anhand interner, formaler und quantitativer Aspekte weiterhin in den meisten Korpusprojekten berücksichtigt (Francis/Kucera 1979, Hofland/Johansson 1982, Sinclair 1988 und Atkins/Clear/Ostler 1992). Beispiele für die angewendeten externen Kriterien sind Kommunikationsteilnehmer, Kontext, soziale Komponenten, kommunikative Funktion u.a. (EAGLES 1996a: 5). Es handelt sich dabei um eine größtenteils von der Texttypologie übernommene Einteilung, die in einer stark hierarchischen Strukturierung, wie sie die statistische Erstellung einer Stichprobe darstellt, zu Problemen bei der Zuschreibung der Texte zu bestimmten Gruppen führen kann.

Wright (1993: 25f) stellt die Anwendung der traditionellen Auffassung der Kategorie Gattung als "literarhistorisch fixierbare Dichtungsform" (Hinck 1977) als problematisch dar, wenn es in konkreten Texten Überschneidungen hinsichtlich der Definition einer gegebenen Gattung gibt. So wird die Textsorte Brief anfangs als nicht-literarisch eingestuft. Wenn jedoch von privaten Briefen des 17. und 18 Jahrhunderts in England die Sprache ist, ist darauf zu achten, daß diese in vielen Fällen literarische Eigenschaften hatten, die schließlich zur literarischen Textsorte des Briefromans führten, der nicht mehr privaten, sondern öffentlichen Charakter hatte. In diesem Sinne ist also bei einer Einteilung in Textsorten auf Besonderheiten zu achten, die für die Variation von Bedeutung sein können und sich, von vielfältigen Faktoren bestimmt, kaum vorhersehen lassen. In diesem wie auch im Falle aller anderen

Variationskriterien muß demnach von einer allzu automatischen Umsetzung der allgemeinen Regeln Abstand genommen werden.

Zur Überwindung des Problems, das eine Einteilung nach ausschließlich externen Kriterien aufwirft, wird in Bibers Ansatz zusätzlich eine Bestimmung anhand interner Kriterien vorgenommen, durch die Textsorten definiert werden können. Textsorte erklärt er anhand inhärenter sprachlicher, von der Gattung unabhängiger Eigenschaften, die in den Texten erst entdeckt und überprüft werden müssen (Biber 1988a: 70): "Text types are identified on the basis of shared linguistic co-occurrence patterns, so that the texts within each type are maximally similar in their linguistic characteristics, while the different types are maximally distinct from one another." (Biber 1993a: 245)

Zur Differenzierung der Textsorten werden zwei Analyseformen angewendet, einerseits zur Bestimmung der Parameter, die die Variation zwischen den Textsorten erzeugen (makroskopische Analyse), und andererseits zur Bestimmung der Eigenschaften und Interaktion dieser Parameter (mikroskopische Analyse). Diese Methode soll gewährleisten, daß die Auswahl der Textsorten und der einzelnen Texte eines Korpus zur Repräsentativität beitragen (Biber 1988a: 61ff).

Makroskopische Analysen werden intertextuell angewendet und helfen, die Parameter zu bestimmen, anhand derer eine Differenzierung von Textsorten möglich ist. Der Vergleich zwischen den Texten hilft Parameter zu bestimmen, die einer

Textsorte eigen sind, wie z.B. die Verwendung von Passiv-Strukturen in wissenschaftlichen Texten und aktivischen Verbformen in literarischen Texten. Die Frequenz der Phänomene ist variabel, d.h. ein wissenschaftlicher Text enthält nicht ausschließlich Passiv-Strukturen, wohl aber einen höheren Prozentsatz als die Mehrzahl literarischer Texte. Die Frequenz eines Phänomens stellt demnach einen Hinweis auf eine Textsorte dar. Da es nicht möglich ist, aufgrund eines einzigen Phänomens auf eine Textsorte zu schließen, müssen Phänomenkombinationen bestimmt werden. Dies geschieht anhand mikroskopischer Analysen, aus denen Verbindungen hervorgehen, die eine Textsorte kennzeichnen. So wäre wissenschaftlicher Text beispielsweise als die Verbindung zwischen einer hohen Frequenz von Passiv-Gebrauch und langen Wörtern (d.h. zusammengesetzten Wörtern) bestimmbar.

Anhand mikroskopischer Analysen werden die Parameter intratextuell zur Konkretisierung der Auswahlmethoden für die einzelnen Texte untersucht. Dabei wird die Verteilung eines in der makroskopischen Analyse gefundenen Parameters bestimmt und auf Zusammenwirkung mit anderen Parametern untersucht. So läßt sich in Wechselwirkung zur makroskopischen Analyse festlegen, welche Parameter kombiniert in einer Textsorte auftreten und welche frequentiellen Eigenschaften sie aufweisen. Die Anwendung der Ergebnisse auf neu dem Korpus hinzuzufügende Texte erlaubt eine Selektion jener Texte, die dem Schema nicht entsprechen. Die nähere Untersuchung dieser Texte muß dann bestimmen, ob sie eine Ausnahme bilden und nicht in das Korpus

aufgenommen werden, oder ob die Parameter der mikroskopischen Analyse revidiert werden müssen.

Die Anwendung der beiden Analysen erfolgt nicht zeitlich nacheinander, sondern stellt eine ständige Wechselbeziehung der Ergebnisse dar. Die mikroskopische Analyse führt nicht zu allgemeinen Variationsparametern, gleichzeitig gibt sie aber erste Hinweise auf die in der makroskopischen Analyse zu untersuchenden Parameter: „Linguistic features that are potentially important indicators of variation within the domains must be identified in advance and measured in each of the texts" (Biber 1988a: 65). In diesem Prozeß beeinflussen sich beide Verfahren gegenseitig.

Zusammenfassend muß jedoch angemerkt werden, daß die theoretischen Schlüsse zwar zu einer detaillierteren Bearbeitung der theoretischen Voraussetzungen und Auswahlmethoden beigetragen haben, aber in der Praxis beim augenblicklichen Stand der Programme zur grammatischen Bearbeitung eines Textes nur beschränkt anwendbar sind.

2.1.2.3 Textursprung

In diesem Abschnitt werden Aspekte zum Ursprung der Texte aus unterschiedlichen Perspektiven besprochen. Zusammenfaßbar als geographische und diachronische Variation, prägen sie die linguistische Form eines Textes und sind somit als Variationskriterien aufzufassen, die für die Strukturierung

des Korpus in Abhängigkeit der gesetzten Ziele beachtet werden müssen.

Das Variationskriterium *geographische Variation* dient hauptsächlich zur Untersuchung der sprachlichen Unterschiede hinsichtlich des geographischen Raumes, wie im deutschsprachigen *Wendekorpus* (Institut für deutsche Sprache 1996d) oder dem englischen *Helsinki Corpus of English Texts* (Kytö 1993).

Der vergleichbare Aufbau des *Corpus of Early American English* und des *Helsinki Corpus*, bei dem das Variationskriterium geographische Variation berücksichtigt wurde, ermöglicht beispielsweise die Untersuchung der Unterschiede zwischen amerikanischem und britischem Englisch (Kytö 1993). Da eine geographische Variation mit einer Zeitverschiebung augenblicklich uninteressant für korpuslinguistische Untersuchungen ist, weil dies zu verzerrten Ergebnissen führen würde, setzt die geographische Variation eine synchronische Übereinstimmung voraus, die präzise zu bestimmen ist.

Eng verbunden mit der geographischen Variation ist der Faktor „Dialekt". Aufgrund des im allgemeinen mündlichen Charakters findet Dialekt Anwendung in Korpora gesprochener Sprache (wie im *Corpus of Spoken American English*, vgl. Chafe/Du Bois/Thompson 1991 oder dem Freiburger Korpus, vgl. Institut für Deutsche Sprache 1996a). Das Variationskriterium dient hauptsächlich der Untersuchung umgangssprachlicher

Besonderheiten und der Wechselbeziehungen zwischen Umgangssprache und regionalem Dialekt.

Die *diachronische Variation* bezieht sich auf den Zeitpunkt der Entstehung der Texte und trägt zu einer ausgeglichenen Auswahl über einen festgelegten Zeitraum bei. Sie wird angewendet für die Untersuchung der Entwicklung einer Sprache, eventuell kombiniert mit der geographischen Variation, wie z.B. im Falle des *Corpus of Older Scots* (Kytö 1993) oder des *Cambridge Corpus of Early Modern English* (Wright 1993). Diese Computerkorpora werden im allgemeinen so erstellt, daß ihre Struktur vergleichbar ist mit der schon bestehender Korpora, wie im Falle des *Freiburg - LOB Corpus of British English*, das den strukturellen Kriterien des Brown und des LOB Korpus folgt und so vergleichbar mit ihnen ist (Hundt et al. 1998).

Der historische Zeitraum, dem die Texte entnommen werden, wird im allgemeinen genau festgelegt (Kytö 1993: 4) und hängt im Falle älterer Epochen mehr von der Verfügbarkeit des Textmaterials als von anderen Variationskriterien[13] ab, was des öfteren kritisierbar scheint, da dadurch oft Kriterien wie die Repräsentativität vernachlässigt werden müssen. Das *Cambridge Corpus of Early Modern English* hat hier eine akzeptable Lösung gefunden: aufgrund der Schwierigkeit, die Texte traditionellen Textsorten zuzuteilen (problematisch ist z.B. der oft fließende Übergang von der epistolaren zur literarischen Form,

[13] Wie z.B. Einteilung in Epochen oder nach historischen Ereignissen.

vgl. auch 2.1.2.2), werden die Texte nach Autoren geordnet, doch um die Kompatibilität zu schon bestehenden Korpora abzusichern, werden die Texte jeweils auch mit evtl. komplexer Textsorteninformation versehen, was ihre Auswahl für kontrastive Untersuchungen erleichtert (Wright 1993: 26)[14]. Soll ein diachronisches Computerkorpus zudem noch geographisch mit einem ähnlich aufgebauten diachronischen Computerkorpus vergleichbar sein, erscheint das Problem, daß beide wahrscheinlich unterschiedlichen Variationskriterien unterliegen (literarische Epochen, gesellschaftliche Ereignisse, usw.), so daß die Vergleichbarkeit nicht mehr absolut ist oder zumindest erschwert wird (Kytö 1993: 4).

Zum zeitlichen Aspekt sei noch das Kriterium der Synchronie erwähnt, das einschränkend wirkt und deshalb nicht zu den eigentlichen Variationskriterien gezählt werden kann, da keine Alternative zu dem gewählten Zeitpunkt gegeben wird. Es dient der Limitierung der Texte eines Computerkorpus auf Texte, die zu einem bestimmten Zeitpunkt geschrieben oder veröffentlicht wurden. Dies ist der Fall des Brown Korpus, in dem nur Texte erscheinen, die im Kalenderjahr 1961 veröffentlicht wurden. Ein weiteres Beispiel für diese Einschränkung ist das deutsche Wendekorpus, das aufgrund des Zieles der Dokumentation der Veränderungen während der Wende 1989 und 1990, sowohl geographisch (West- und Ostdeutschland) als auch synchronisch

[14] Vgl. Wright (1993: 26f) hinsichtlich der Diskussion zur Anwendbarkeit von Textsorten auf Computerkorpora.

aufgebaut (1989 und 1990 als zeitliches Limit) ist (Institut für deutsche Sprache 1996d).

2.1.2.4 Verfasser

In der Gruppe Verfasser werden die Aspekte zusammengefaßt, die eine nähere Bestimmung des Textes in Bezug auf seinen Verfasser ermöglichen. Es handelt sich dabei um die Kriterien Population und Idiolekt. Unter Population verstehen wir hier die Individuen und die sie betreffenden textexternen Eigenschaften, die Einfluß auf die Textproduktion haben (vgl. dazu Lorente et al. 1997: 121). Idiolekt hingegen bezieht sich auf die linguistische Eigenschaft des a priori bestimmten Stils des Verfassers (vgl. Tuldava 1995).

Population

In soziolinguistischen Untersuchungen ist dieses Einteilungskriterium laut Labov (1983: 34f) als grundlegend zu betrachten. Population dient in Korpora geschriebener Sprache dazu, sprachliche Unterschiede auf möglicherweise relevante Eigenschaften des Individuums zurückzuführen. Die Anwendung textexterner Subkriterien wie Geschlecht, Alter, soziale und regionale Herkunft u.v.a. führt zu einem weiteren Strukturierungsgrad des Korpus. Insofern das Kriterium Population zur Repräsentativität beitragen soll, muß die Auswahl der Population anhand festgelegter Parameter erfolgen (vgl. Butler 1985).

Idiolekte

Das Variationskriterium des Idiolekts (vgl. Kytö 1993: 3; Wright 1993: 25) ermöglicht die Strukturierung eines Korpus auf der Grundlage der stilistischen Phänomene, die für einen Verfasser als charakteristisch angesehen werden. Die Berücksichtigung der stilistisch-individuellen Komponente des Idiolekts erklärt, daß es sich dabei bis jetzt hauptsächlich um literarische Textkorpora handelt.

Die Strukturierung kann sich einerseits auf einen einzigen Verfasser beziehen, so daß die stilistischen Phänomene dazu dienen, ein repräsentatives Korpus der Texte eines Verfassers zu erstellen. Sie kann aber auch die Darstellung einer Epoche oder literarischen Bewegung anstreben, wobei die individuellen stilistischen Phänomene mehrerer Autoren zu bestimmen und zu berücksichtigen sind.

Eine Einteilung nach Idiolekten ermöglicht aufgrund der Annahme, daß die Abweichung zwischen Autoren stärker oder wichtiger als zwischen Textsorten ist, eine stärkere Hervorhebung der individuellen Unterschiede, während die sprachlich gemeinsamen Eigenschaften, die einer geschichtlichen und textsortenspezifischen Evolution unterliegen, in den Hintergrund treten. Problematisch ist dies allerdings für korpuslinguistische Untersuchungen, die eine anonyme Grundgesamtheit darzustellen versuchen. Das ist der Fall lexikographischer Untersuchungen, da durch diese Einteilung nicht die gemeinsame Bedeutung eines Wortes

hervortritt, sondern seine unterschiedliche Verwendung bei verschiedenen Autoren.

2.1.2.5 Kommunikationsform

Das Kriterium Kommunikationsform bezieht sich auf die Art, in der ein Text seinen Leser ursprünglich erreichen sollte. In dieser Hinsicht kann zwischen privater und öffentlicher Kommunikation einerseits und veröffentlichten und nicht veröffentlichten Textprodukten unterschieden werden.

Die Unterscheidung zwischen privater und öffentlicher Kommunikation ist bedeutend für die Untersuchung stilistischer, argumentativer und anderer pragmatischer Unterschiede (Greenbaum 1991: 90). Greenbaum definiert als privat persönliche Briefe und Notizen, und als öffentlich alles normalerweise gedruckte Material, das sich an eine unbestimmte Leserzahl richtet. Das Variationskriterium kann einen Widerspruch zur Unterscheidung zwischen veröffentlicht und nicht veröffentlicht darstellen, wenn ein ursprünglich privater Brief veröffentlicht wird (z.B. aus dem Nachlaß eines Schriftstellers) oder wenn die epistolare Form schon von Anfang an für die Veröffentlichung gedacht ist und stilistisch mit diesem Ziel angepaßt wird, wie es in den Briefromanen des 18. Jhs. geschieht (Wright 1993; Kytö 1993).

Die Differenzierung veröffentlicht vs. nicht veröffentlicht dient zur Abgrenzung von Gebrauchstextsorten. Häufig bildet sie die Grundlage von sprachgeschichtlich ausgerichteten

Untersuchungen, da die Veröffentlichung eines Textes einer Sprache A in einer sprachlichen Epoche mit dem Einfluß einer zu dieser Zeit vorherrschenden Sprache B verbunden sein kann. In dieser Hinsicht sind die Anmerkungen von Anneli Meureman-Solin (1995: 53f) über den Einfluß der englischen Schriftsprache auf das Schottische bedeutend. In bezug auf die Veröffentlichung sind detailliertere Einteilungen möglich, wie die Greenbaums für das *International Corpus of English*, wo zwischen *scripted* (geschrieben, um gelesen zu werden), *nicht-gedruckt* (handschriftlich oder ausgedruckt) und *gedruckt* unterschieden wird (Greenbaum 1991: 90; im Spanischen Briz Gómez 1998: 19ff).

2.1.2.6 Zusammenfassung

Variationskriterien bilden die Grundlage der Strukturierung eines Korpus und tragen somit zu seiner Repräsentativität bei. Ihre Anwendung erfolgt in Abhängigkeit von den gesetzten Forschungszielen. Der theoretische Rahmen, in dem die Forschung anzusiedeln ist, bedingt dabei die Auswahl der anzuwendenden Variationskriterien. So werden in der Soziolinguistik beispielsweise mehr Variationskriterien zum Verfasser verwendet, wie Alter, Herkunft, Geschlecht, Zugehörigkeit zu bestimmten Kulturgruppen usw. (siehe auch dazu Wheeler 1997: 72).

Durch ihre Anwendung wird ein Korpus hierarchisch strukturiert; zunächst nach den in Abhängigkeit vom

Forschungsziel als grundlegend angesehenen Variationskriterien (z.B. Medialität, d.h. mündlich-schriftlich), anschließend nach untergeordneten Variationskriterien (z.B. nach der Einteilung in mündlich-schriftlich in literarisch, technisch, wissenschaftlich usw. im Falle geschriebener Texte).

Die Strukturierung nach festgelegten Variationskriterien würde die großen Nationalkorpora in Frage stellen, da die vorhandene Strukturierung nicht unbedingt den Forschungszielen einzelner Untersuchungen entspricht. Doch die Zusatzinformation, die in den einzelnen Dateien angegeben wird, erlaubt es dem Forscher, aus dem Nationalkorpus ein eigenes, spezifisches Korpus zusammenzustellen, das die erforderlichen Variationskriterien berücksichtigt.

2.1.3 Korpusumfang

Der Umfang eines Computerkorpus ist ein weiterer Aspekt, der für seine Zusammenstellung in Betracht zu ziehen ist. Der Umfang wird normalerweise in Wörtern berechnet auf der Grundlage der einzelnen Texte, wobei diesen wie im Brown Korpus eine feste Zahl an Wörtern zugewiesen werden kann (Francis/Kucera 1979: 2), oder wie im *International Corpus of Learner English* (ICLE) die Gesamtheit der Wörter eines jeden Textes (Granger 1998: 10), d.h. der Volltext. In letzter Zeit wird die Auswahl von Volltexten für angebracht gehalten, weil ursprüngliche technische Einschränkungen mit der Entwicklung

der Computerhardware nicht mehr vorhanden sind (EAGLES 1996a).

Am Anfang der korpuslinguistischen Untersuchungen wurde versucht, allgemeingültige Kriterien zur Bestimmung des Umfangs eines Korpus zu geben, die mit der Zeit präziseren, mathematisch berechenbaren Vorschlägen gewichen sind. So nannte Francis für ein Computerkorpus mit allgemeinen Forschungszielen, wie dem von ihm beschriebenen Brown Korpus, eine Million Wörter: dies sei genug für viele Forschungen, nicht zu viel für die Rechenleistung der Computer und adäquat für die nur beschränkt zur Verfügung stehenden Mittel (Francis 1980: 193). Da sich seit dem Jahre 1980 oder der Erstellung des Brown Korpus gegen Anfang der 60er Jahre die Rechenleistung der Computer vervielfacht hat und auch die Forschung sich zunehmend Phänomenen mit niedrigen Frequenzen widmet, sind heutzutage größere Computerkorpora gefragt, wie das BNC mit 100 Millionen Wörtern.

Diese Konkretheit der Vorschläge wurde in den letzten Jahren durch theoretische Aspekte ergänzt, die versuchen, den Umfang eines Computerkorpus aufgrund seiner Ziele zu bestimmen. So wurden in den 80er Jahren Berechnungen zur Bestimmung der Länge eines Computerkorpus entwickelt (vgl. z.B. Francis 1982), wie der Einfluß der Frequenz eines Phänomens auf den für seine Untersuchung erforderlichen Korpusumfang. Beispielsweise ist demnach für die Untersuchung des Artikels im Deutschen ein kleines Korpus ausreichend, da es sich dabei

um ein hochfrequentes Phänomen handelt; für die Analyse wenig frequenter sprachlicher Phänomene jedoch wäre ein größeres Korpus notwendig, um über umfangreiches empirisches Datenmaterial zu verfügen.

In diesem Zusammenhang ist das Problem der „Reduzierung der lexikalischen Treffer"[15] zu nennen: je größer das Korpus, desto niedriger der Prozentsatz der Einträge eines bestimmten Wortes, und desto höher die Frequenz der *hapax legomena*, der Einzelerscheinungen eines Wortes. Dies kann nach Meijs (1992: 146) dazu führen, daß viele Wörter von ihrem Kontext her nicht erschlossen werden können, da sie nicht oft genug erscheinen.

Einen anderen Ansatz zur Betrachtung des Problems der Wörter niedriger Frequenzen stellt die lexikalische Resolution dar. Frequenzen werden in diesem lexikalisch ausgerichteten Ansatz als problematisch oder unproblematisch aufgrund ihrer „lexikalischen Resolution" eingestuft, die die Möglichkeit darstellt, die Bedeutung eines Wortes aus dem Kontext zu erschließen (Renouf 1987a: 121). Frequente Wörter stellen demnach aus lexikographischer Sicht kein Problem dar, weil ein gegebenes Korpus über genug Einträge verfügt, um sie aus ihrem Kontext erschließen zu können. Wörter mit sehr niedrigen Frequenzen oder sogar *hapax legomena* sind ebenfalls wenig problematisch, da sie im allgemeinen unbrauchbar für ein Wörterbuch sind. Problematisch sind von diesem Standpunkt aus

[15] „Law of diminishing lexical returns" (Meijs 1992: 146).

nur Wörter mittlerer Frequenzen, also weder hochfrequent noch *hapax legomena*, für die ein großes Korpus benötigt würde, weil die gefundenen Einträge nicht die Möglichkeit anbieten, das Wort aus seinem Kontext zu erschließen.

Zahlreiche weitere theoretische Überlegungen zum Korpusumfang weisen auf die Abhängigkeit von den Forschungszielen hin. In diesem Sinne muß die Länge des Korpus für Barnbrook (1996) im voraus bestimmt werden, obwohl dies seiner Meinung nach oft mit Problemen hinsichtlich der Präzision der ermittelten Werte verbunden ist. Anhand des angesammelten textuellen Materials werden Pretests zur Bestimmung der Frequenzen der Phänomene durchgeführt, die später Einfluß auf den Gesamtumfang des Korpus nehmen. Dabei müssen die Frequenzen der Pretests aber nicht vollständig mit den Frequenzen des endgültigen Computerkorpus übereinstimmen, so daß das Korpus erneut korrigiert werden müßte. Allgemein kann jedoch gesagt werden, daß ein ausgewogen zusammengestelltes kleines Korpus mit sorgfältig ausgesuchten Textsorten nützlich für konkrete Untersuchungen ist, während ein großes zu weit angelegten quantitativen Untersuchungen verführt (Johansson/Stenström 1991: 115; Aston 1998). In diesem Sinne argumentiert Granger (1993: 61) ebenfalls, daß ein kleines, gut ausgewähltes Korpus nützlicher als ein großes heterogenes Korpus sei.

Überlegungen zum Umfang eines Computerkorpus spiegeln sich in der Darstellung konkreter Computerkorpora wider. Beispiele für Computerkorpora reduzierten Umfangs sind zahlreich. So

arbeiten Collot/Belmore (1993) mit einem Korpus elektronischer Sprache, dessen Umfang weniger als 200.000 Wörter aus 9 verschiedenen elektronischen Diskussionsforen beträgt, von denen jedes ein Variationskriterium darstellt. Das Nijmegen-Korpus (Haan 1991) ist mit seinen ugf. 130.000 Wörtern noch reduzierter, ebenso wie das *Lancaster Parsed Corpus*, ein Subkorpus des LOB, das mit grammatischen Auszeichnungen zu den Wortarten (*POS-Tagging*) und zusätzlich syntaktischen Auszeichnungen (*Parsing*) aufbereitet wurde (Icame 1998). Hinsichtlich der Computerlernerkorpora sei hier das *International Corpus of Learner English (ICLE)* verwiesen, das mit ugf. 200.000 Wörtern ebenfalls einen geringen Umfang aufweist (Granger 1998b), jedoch gleichzeitig durch die Einschränkung auf Lernertexte bestimmter Textsorten hochspezifisch ist.

Auch praktische Aspekte haben einen bedeutenden Einfluß auf den Umfang eines Computerkorpus. So nennt Leech (1991a: 10ff) zwei Aspekte, die den Umfang einschränken.

- Das Copyright begrenzt z.B. den Umfang der reproduzierbaren Textextrakte, und verhindert im Normalfall vollkommen, daß ganze literarische Texte, wie Kurzerzählungen, in ein öffentlich verfügbares Computerkorpus aufgenommen werden (Lehr 1996: 99).

- Die nicht parallele Entwicklung der Computerhardware und Software ist als zweiter Aspekt zu nennen. Mit der Zeit sind die Speichermöglichkeiten gewachsen, so daß Computerkorpora

sich von dieser Begrenzung befreit haben, während die Software keine vergleichbare Entwicklung genommen hat. Das führt zur paradoxen Tatsache, daß zwar große Textmengen gesammelt und gespeichert, aber nicht bearbeitet werden können, da die verfügbare Software nicht zuverlässig genug arbeitet. Zu dieser Software gehören augenblicklich hauptsächlich Tagger und Parser. Tagger fügen automatisch in die Texte Information zu den Wortarten (Part of Speech oder POS) ein. Bei diesem Prozeß, dem sogenannten Tagging, wird ein hoher Prozentsatz an korrekten Auszeichnungen erreicht, doch ist eine manuelle Revision immer noch unumgänglich. Parser hingegen zeichnen Texte mit syntaktischer Information aus, die aufgrund eines vorhergehenden Taggings bestimmt wird. Der Prozeß des Parsings geht mit erheblichen Fehlerquoten einher (vgl. Aarts 1996: 105f), so daß ein hoher Aufwand an manuellen Revisionen notwendig ist. Andere automatisiert anwendbare Auszeichnungssysteme (beispielsweise semantische oder diskursive Elemente) befinden sich noch im Anfangsstadium ihrer Entwicklung. All diese Faktoren bremsen die Entwicklung größerer Computerkorpora oder von Computerkorpora mit spezifischen Auszeichnungen, d.h. Auszeichnungen, die nicht nur Wortarten und Syntax berücksichtigen.

2.1.4 Abgeschlossenheit

Eng mit dem Korpusumfang verbunden ist die Abgeschlossenheit. Dadurch wird bestimmt, ob ein Computerkorpus einen festgelegten, unveränderlichen Umfang hat (d.h., es ist abgeschlossen) oder ob dieser Umfang in Abhängigkeit von den Forschungszielen variabel ist (d.h., es ist offen), so daß zu den schon vorhandenen Texten ständig weitere hinzugefügt werden.

Bei offenen Computerkorpora, sogenannten Monitorkorpora (Vgl. Renouf, A. [1987] und EAGLES [1996a: 11]), wird auf der Grundlage eines Urkorpus durch Hinzufügung neuer Texte oder Streichung alter Texte eine ständige Aktualisierung erzeugt. Es ist dieser Aktualisierungsprozeß, der ihre Anwendungsmöglichkeit bedingt. Die Dynamik, der sie unterliegen, ermöglicht die Untersuchung dynamischer diachronischer Prozesse.

Beispiele für diese Forschungsmöglichkeiten sind das Projekt *COBUILD* um Sinclair, in dem an einem Monitorkorpus zur lexikographischen Erschließung der Bedeutungsänderung gearbeitet wird. Ein weiteres Beispiel für die Ausnutzung des dynamischen Charakters des Monitorkorpus ist ein Projekt der Universität Birmingham. Durch ständige Aktualisierung der Texte im Vergleich zu einem bestehenden, nicht aktualisierten Korpus wird untersucht, welche Wörter diachronisch neu in der englischen Sprache gebildet werden, und welche lexiko-grammatischen Tendenzen bestehen (vgl. Renouf 1993).

Gleichzeitig schränkt der dynamische Aktualisierungsprozeß von Monitorkorpora ihre Anwendbarkeit ein. Wie McEnery (1996: 22f) treffend bemerkt, sind sie aufgrund der sich ändernden Wortzahl nur beschränkt für quantitative sprachliche Untersuchungen geeignet. Diese Beschränkung ergibt sich aus der Methode ihrer Erstellung. Offene Computerkorpora erweisen sich problematisch aufgrund der Repräsentativität und der Überprüfbarkeit der Forschungsergebnisse. Bei einem sich ständig erweiternden Korpus müßten die neuen Texte so hinzugefügt werden, daß alle ursprünglich vorgesehenen Variationskriterien zur Strukturierung gleichzeitig berücksichtigt würden. D.h., für jeden neu hinzuzufügenden Text müßten Texte bereitgestellt werden, die den weiteren Variationskriterien entsprechen. Des weiteren wäre die Überprüfung der Ergebnisse einer konkreten Forschung seitens anderer Forscher nur schwer nachvollziehbar, da die Vielzahl an Versionen eines offenen Korpus fast unübersichtlich wäre. Aber nicht nur deshalb, sondern auch aufgrund der aktuellen Forschungsinteressen, die mehrheitlich eine statistische Beschreibung eines gegebenen Zustandes anstreben, werden in der korpuslinguistischen Forschung abgeschlossene Korpora bevorzugt, wie es das Brown, das LOB oder das Limas Korpus sind. Angesichts der methodischen Probleme, die ihre Erstellung mit sich bringt, bevorzugt es Sinclair, einer der leitenden Forscher auf dem Gebiet der Monitorkorpora, diese Art Korpus eine „Textsammlung" zu nennen.

Abgeschlossene Korpora hingegen stellen nach Fertigstellung im Umfang unveränderliche Textsammlungen dar. Ihre Anwendungsmöglichkeiten unterliegen einem ständigen Erweiterungsprozeß, der sich parallel zum Fortschritt der Korpuslinguistik entwickelt. Wurden sie anfangs noch fast ausschließlich für lexikalische Untersuchungen und Analysen zur Wortfrequenz benutzt, finden sie heutzutage mehr Anwendungen im Bereich des NLP (*Natural Language Processing*, vgl. Barnbrook 1996), zur statistischen Disambiguierung von Wortbedeutungen, zur automatischen Erstellung von ein und zweisprachigen Wörterbüchern usw.

Ebenso wie die Korpuslinguistik haben sich auch Computerkorpora entwickelt. Die Korpora der sechziger und siebziger Jahre, damals als groß angesehen, werden von den heutigen Großkorpora übertroffen, die mittlerweile wie das British National Corpus ab ungefähr 100 Millionen Wörter enthalten.

2.1.5 Maschinelle Lesbarkeit

Eine weitere Eigenschaft moderner Computerkorpora besteht in der Möglichkeit einer maschinellen Bearbeitung der Texte (vgl. McEnery 1996, Barnbrook 1996). Durch die Speicherung der Texte in einem Computer und ihre Analyse mit spezifischer Software werden methodisch effiziente Untersuchungen ermöglicht, die mit einer nicht-maschinellen Methode kaum oder sogar gar nicht möglich waren. Maschinelle Lesbarkeit spielt sich auf mehreren

Ebenen ab. Die Ebenen betreffen textexterne und textinterne Eigenschaften. Textextern bedeutet die Möglichkeit, aus einem Computerkorpus nur bestimmte Texte aufgrund der Angaben zu textexternen Eigenschaften (Autor, Datum der Veröffentlichung, geographischer Ursprung usw.) zu extrahieren, d.h. mit einem Subkorpus zu arbeiten.

Die eigentliche Innovation liegt jedoch in der maschinellen Bearbeitung textinterner Eigenschaften. Anhand spezifischer Software können Computerkorpora in ihrer rohen (oder nach der Terminologie von Biber [1998: 257] unkodierten Form, d.h. ohne Aufbereitung mit Wortarten) oder in ausgezeichneter (oder kodierter Form, d.h. versehen mit Zusatzinformation zum Text) bearbeitet werden. Die Bearbeitung roher Computerkorpora wird durch sogenannte Concordancer ermöglicht, die automatisch alphabetische Wortlisten, Wortfrequenzlisten und weitere statistische Berechnungen anbieten[16]. Für die Aufbereitung eines Korpus mit Zusatzinformation ist hingegen ein zusätzlicher Auszeichnungsprozeß notwendig, anhand dessen den jeweiligen Texteinheiten Information zu den Wortarten, syntaktischen Funktionen und anderen Einheiten beigegeben wird. Eine nähere Darstellung der Auszeichnungsmethoden und Auszeichnungssysteme befindet sich in Abschnitt 2.2.2.

[16] Beispiele für Concordancer sind Word Smith, der in unserer Untersuchung in seiner Version 2.0 verwendet wurde (Scott 1998), Monoconc von Michael Barlow (Barlow 1998), und TACT von der Universität Toronto (TACT 1998).

Bei der Aufbereitung von Textkorpora für Computer handelt es sich um einen eng an die Entwicklung der technischen Möglichkeiten gebundenen Prozeß. Obwohl die ersten maschinell lesbaren Korpora der 60er Jahre schon auf Lochkarten gespeichert wurden, wie das Brown Korpus, war es auch später noch keine Selbstverständlichkeit, Korpora in maschinenlesbarem Format zu veröffentlichen. Erst 1991 meinte Johansson, daß in Zukunft niemand auf die Idee kommen würde, ein Korpus anzulegen, ohne es maschinell lesbar zu machen (Johansson 1991a: 305), wodurch angedeutet wird, daß zu dieser Zeit Korpora angelegt wurden, die nicht anhand eines Computers ausgewertet werden sollten. Auch für Leech ist die maschinelle Lesbarkeit Bestandteil der Definition eines modernen Korpus (Leech 1992: 115f), und in seiner Einleitung zum Thema Korpuslinguistik verteidigt McEnery (1996), daß ein Korpus heutzutage „fast immer" maschinell lesbar sei.

Während sich im englischsprachigen Raum die Auffassung durchgesetzt hat, daß Korpora in maschinell lesbarer Form vorliegen sollten, ist die Situation im deutschsprachigen Raum anders gestaltet. Die Auffassung, daß Korpora maschinell lesbar sein sollen und somit mit korpuslinguistischen Methoden bearbeitbar werden, hat sich nur langsam durchgesetzt. Hier folgt ein kurzer Überblick über augenblicklich verfügbare deutschsprachige Computerkorpora, unterscheidend zwischen Computerkorpora gesprochener und geschriebener Sprache.

Die deutschen Korpora gesprochener Sprache, wie das Freiburger Korpus (FKO), das Dialogstrukturenkorpus (DSK) und das Pfeffer-Korpus (PFE)(Institut für deutsche Sprache 1996a; 1996e und 1996b), entstanden teilweise erst gegen Mitte der 90er Jahre oder waren Aufbereitung von Korpora gesprochener Sprache, die nur in gedruckter Form vorlagen. Zu diesem Zeitpunkt lagen schon lange erste Korpora zu gesprochener englischer Sprache vor.

Das Freiburger Korpus behandelt den Zeitraum zwischen 1966 und 1972; die ugf. 700.000 Wörtern sind Texten wie Diskussionen, Interviews, Vorträgen, Berichten, Erzählungen und Reportagen entnommen und wurden an der früheren Außenstelle des Instituts für deutsche Sprache in Freiburg erstellt (Institut für deutsche Sprache 1996a).

Das Pfeffer-Korpus wurde auf der Grundlage der in Buchform im Niemeyer-Verlag veröffentlichten *Texte zur gesprochenen deutschen Gegenwartssprache (Grunddeutsch-Texte)* an der Stanford-University von Prof. J. Alan Pfeffer erarbeitet; mit seinen ugf. 650.000 Wörtern ist es relativ umfangreich, doch leicht veraltet, da die Aufnahmen von Anfang der 60er Jahre stammen. Interessant ist hier, daß das Hauptkriterium der Variation nicht wie im Freiburger Korpus die Textsorte war, sondern die regionalen und nationalen Varietäten der deutschen Sprache; so ist Umgangssprache aus der BRD, der DDR, Österreich und der Schweiz vertreten.

Die Korpora deutscher geschriebener Sprache sind demgegenüber ausführlicher. Über das internetbasierte System COSMAS (Institut für deutsche Sprache 1996c) sind mehrere computerlesbare Korpora zugänglich, wie die Mannheimer Korpora, das Grammatik-Korpus, das Bonner Zeitschriftenkorpus, das Handbuchkorpus, das Wendekorpus, das Limas-Korpus u.a.

Bis auf das Wendekorpus und das LIMAS-Korpus leiden Korpora wie die Mannheimer Korpora, das Handbuchkorpus oder das Grammatik-Korpus darunter, daß sie aufgrund des Arbeitsaufwandes anhand von elektronisch leicht zugänglichem Material zusammengestellt wurden; die Mannheimer Korpora sind Zeitungskorpora, die schon in elektronischer Form vorlagen, ebenso wie das *Korpus Magazin Lufthansa Bordbuch* (Institut für deutsche Sprache 1998d). Diese Faktoren führen, wie Lehr (1996: 65) betont, zu einer Minderung der Repräsentativität und Bearbeitbarkeit der vorhandenen Computerkorpora.

2.2 Textbezogene Aspekte

Während korpusbezogene Aspekte mit dem Korpus als abstraktes Gebilde verbunden waren, beziehen sich die textbezogenen Aspekte auf die Texte, die das Korpus bilden. Die für diesen Abschnitt zu beachtenden Aspekte lassen sich in zwei große Gruppen einteilen: die der Auswahl der Texte und die der Bearbeitung derselben.

2.2.1 Auswahl

Für die Zusammenstellung eines Computerkorpus ist die Auswahl der Texte, d.h. die Methode zur Bestimmung der Texte, die in das Korpus aufgenommen werden, bedeutend für die Repräsentativität. Repräsentativität wird durch die sogenannte Stichprobe erreicht (Schlobinski 1996), die sich aus einzelnen Texten zusammensetzt (vgl. 2.1.2). Erst wenn die Stichprobe und demnach die Einzeltexte die Allgemeinheit darstellen, ist die Voraussetzung für verallgemeinernde Auswertungen gegeben. In diesem Abschnitt werden die Aspekte dargestellt, die Einfluß auf diese Auswahl haben. Folgende Eigenschaften sind dabei zu beachten: die Verallgemeinerbarkeit, durch die aufgrund zu bestimmender Phänomene abgesichert wird, daß ein einzelner Text Teil eines größeren Ganzen ist; die Textlänge und mit ihr verbunden die Auswahlmethode der Texte, d.h. zwei zusätzliche statistisch-methodische Aspekte, die es ebenfalls ermöglichen, von der Stichprobe auf die Grundgesamtheit zurückzuschließen.

2.2.1.1 Verallgemeinerbarkeit

Bei der Auswahl der Einzeltexte muß abgesichert werden, daß diese verallgemeinerbar sind, also Teil der Grundgesamtheit bilden und diese auch darstellen. Damit wird der Auswahlbias vermieden, d.h. die systematische Abweichung einer Statistik vom Parameter (Schlobinski 1996: 26). Das Verfahren besteht in der Ermittlung von Parametern, die charakteristisch für den

Text sind, und anschließend in der Zuweisung von Maximal- und Minimalwerten für diese Parameter. Damit ein Text in ein Computerkorpus aufgenommen werden kann, muß er die Parameter aufweisen und darf die Maximal- und Minimalwerte nicht überschreiten oder unterschreiten.

Die Ermittlung der Parameter erfolgt projektbezogen aufgrund textinterner und textexterner Kriterien, worauf schon in 2.1.2 verwiesen wurde. Als textexterne Kriterien sind Eigenschaften wie Alter, soziale Herkunft u.a. zu nennen; zu den textinternen werden linguistische Eigenschaften wie Textstruktur, Tempusformen, Wortartenfrequenzen u.a. gerechnet.

Hinsichtlich des Kriteriums Verallgemeinerbarkeit sind zwei weitere Aspekte zu beachten: die Vielfalt und die Authentizität.

Unter Vielfalt ist zu verstehen, daß Wiederholungen von sich nicht ausschließenden Kriterien oder von nicht festgelegten Kriterien vermieden werden sollten. Bei der Strukturierung eines Computerkorpus geschriebener Sprache in literarische und wissenschaftliche Texte wäre es ursprünglich kein Widerspruch, wohl aber ein Verstoß gegen die Vielfalt, literarische und gleichzeitig wissenschaftliche Texte des gleichen Autors in das Computerkorpus aufzunehmen; ebenso wäre zu vermeiden, daß literarische und wissenschaftliche Texte alle im gleichen Verlag erschienen, selbst wenn Verlag innerhalb der

vorgegebenen Strukturierung nicht als Variationskriterium angesehen wurde.

Die Authentizität bezieht sich hingegen auf den Ursprung des Materials. Bei einem Korpus sollte es sich bis auf wenige, begründete Ausnahmen um Originalmaterial handeln, das von den Forschern nicht geändert wurde. Stubbs gliedert das Textmaterial in *„attested, actual, authentic data"*, *„modified data"* und *„invented, intuitive, introspective data"*. Akzeptabel sind für ihn nur die beiden ersten, während er einen Großteil des Werkes der Kritik der letzten Gruppe widmet (vgl. Stubbs 1996a: 28). Ein Problem wirft hier im Falle von Computerkorpora gesprochener Sprache die Transkription auf, im Falle der geschriebenen z.B. die Abweichung zum Original, die entsteht, sobald der Einzeltext als Fragment seinem Umfeld entnommen wird und somit seinen Kontext verliert, um Teil eines Korpus zu werden. Für Francis (1980: 197) sind Änderungen des Originals nicht angebracht, so daß im Brown Korpus bei der Kodierung sogar die Worttrennung berücksichtigt wurde.

Nach der Bestimmung der oben genannten Faktoren und der Sammlung jener verfügbaren Texte, die alle Kriterien erfüllen, wird im allgemeinen eine aleatorische Auswahl derselben vorgenommen, um den vorgesehenen Umfang des Korpus zu erreichen. Diese Auswahl kann aus ganzen Texten oder Fragmenten von einer vorher festgelegten Länge bestehen, wie im folgenden Abschnitt dargestellt wird.

2.2.1.2 Textlänge

Ein weiterer Aspekt, der bei der Auswahl der einzelnen Texte zu berücksichtigen ist, ist ihre Länge. Wie schon in 2.1.3 angesprochen, besteht die Möglichkeit, nur Fragmente eines Textes für das Computerkorpus zu verwenden, Volltexte oder Passagen, wobei die Entscheidung für eine dieser Methoden wiederum von dem Forschungsziel abhängig ist. *Fragmente* sind in diesem Sinne Textteile, die eine bestimmte Länge haben und nur mit der Satzmarkierung des Originaltextes übereinstimmen, nicht aber weitere thematisch-strukturelle Einheiten, wie z.B. Absätze oder Kapitel, berücksichtigen. *Volltexte* hingegen stellen abgeschlossene Texteinheiten dar, d.h. den ganzen Text. Und *Passagen* schließlich spiegeln einen strukturellen Auszug eines Volltextes wider, bei dem z.B. eine thematische Einheit (Absatz oder Kapitel) vollständig aufgenommen wurde.

Das deutsche Wendekorpus ist ein Beispiel für Computerkorpora mit Volltexten. Computerkorpora, die Fragmente verwenden, d.h. Texte einer festgelegten Länge, sind das Limas Korpus oder im englischen Sprachraum das Brown Korpus. Heutige Computerkorpora zur repräsentativen Darstellung des Werkes eines Schriftstellers hingegen würden auf Passagen zurückgreifen. Anzumerken ist dabei, daß die Erweiterung der Textspeichermöglichkeiten mittlerweile zur Empfehlung führt, Volltexte oder zumindest Passagen zu verwenden (Eagles 1996a).

Für stärker inhaltsbezogene Untersuchungen ist die Möglichkeit, über ganze Texte zu verfügen, von besonderer

Bedeutung. Stubbs (1996a: 129) verteidigt dies, denn „the restriction to data fragments poses problems of evidence and generalization". Da bestimmte Regelmäßigkeiten nicht direkt betrachtbar und eher probabilistischer Natur sind, also höchst selten erscheinen, verteidigt er die Zusammenstellung von Textkorpora mit langen Texten, die nach diesen Regelmäßigkeiten untersucht werden können: „The analysis of short texts and text fragments must be complemented by the analysis of long texts, since some patterns of repetition and variation only occur across long texts" (Stubbs 1996a: 152).

Einen theoretischen Ansatz zur statistischen Lösung des Problems liefern Biber und Finegan (1991: 211f). Sie vertreten, daß ein Text lang genug sein muß, um die linguistischen Eigenschaften des ganzen Textes darzustellen, aber nicht länger, um das Korpus nicht unnötig zu vergrößern. Die innere Variation, die ein Text z.B. aufgrund der strukturellen Entfaltung aufweisen kann[17], führt zu zwei Auswahlmethoden: die Einbeziehung ausgewählter Passagen des Textes, die eine Zusammenfassung seiner sprachlichen Eigenschaften darstellen; oder die Bestimmung sprachlicher Elemente, deren Untersuchung angestrebt wird, mit anschließender Auswahl nur jener Passagen des Textes, der diese sprachlichen Elemente enthält. Ungelöst bleibt bei

[17] Wir beziehen uns auf Dokumentation, Analyse, Argumentation, Interpretation usw.; vgl. Bünting et al. 1996: 17ff; Rückriem et al. 1983: 100f.

diesem Ansatz die Klassifizierung von Texten nach Textsorte, aufgefaßt als konventionalisierte Klasse von Texten (wie z.B. Kochrezept, Rezension usw.) oder ihre Beschreibung nach Texttypen, aufgefaßt als kombinierbare Liste von Zuordnungen bestimmter Texteigenschaften (wie z.B. Funktionstyp, Sequenzierung, Situationstyp usw.)[18].

2.2.1.3 Textauswahl

Nach den Vorbereitungsphasen der Zuweisung der Texte zu Gruppen, die aufgrund der Variationskriterien bestimmt wurden, nach der anschließenden Bestimmung der textinternen und der textexternen Aspekte eines Textes, die eine Überprüfung der Zugehörigkeit zu den Gruppen ermöglichen, und, anschließend, nach der Entscheidung, ob ganze Texte oder Fragmente in das Korpus aufgenommen werden, erfolgt anhand des zusammengetragenen Textmaterials die Auswahl der Texte, die den Anforderungen entsprechen. Dies geschieht anhand zwei statistischer Vorgehensweisen: der Quotenauswahl und der Zufallsstichprobe (vgl. 2.1.2).

Bei der Stichprobentechnik der Quotenauswahl (auch geschichtetes Sample genannt, vgl. Schlobinski 1996: 25) wird aus den Textgruppen, die in Abhängigkeit von den erarbeiteten Variationskriterien zusammengestellt wurden, aleatorisch aus

[18] Vgl. Adamzik 1995: 7-40

jeder Gruppe eine bestimmte Anzahl von Texten herausgesucht, die dann Teil des Korpus bilden[19].

Im Gegensatz zu dieser Methode steht die Zufallsstichprobe, für die aleatorisch aus all den vorhandenen Texten, ohne Berücksichtigung der Textgruppen, eine bestimmte Zahl von Texten gewählt wird.

Während die Quotenauswahl strukturiert vorgeht, d.h. in einer ersten Phase ein repräsentatives Sample einer Textgruppe zusammenstellt und die verschiedenen Samples in einer zweiten Phase zu einem Korpus zusammenfügt, übergeht die Zufallsstichprobe die Phase der Textgruppen und stellt das Korpus anhand aller verfügbarer Texte zusammen.

Der Vergleich beider Verfahren ergibt, daß die Quotenauswahl immer mindestens so repräsentativ ist wie die Zufallsstichprobe und normalerweise repräsentativer (Biber 1990). Dementsprechend hat sich mit der Zeit in der Korpuslinguistik die Technik der Quotenauswahl durchgesetzt.

2.2.2 Auszeichnungen

Nach der Auswahl der Texte sind sie einer Bearbeitung zu unterwerfen, die ihre maschinelle Bearbeitung ermöglicht und die Möglichkeiten der Auswertung eines Computerkorpus

[19] Vgl. die Technik, die im Brown Korpus angewendet wurde, Francis 1980: 193ff.

beeinflußt. Wichtige Faktoren sind dabei die Prozesse der Säuberung und Vereinheitlichung der Texte und die methodisch geprägten Fragestellungen zu den Auszeichnungen, dem Format der Auszeichnungen und die Auszeichnungsebene.

Durch den Prozeß der Säuberung werden Texte so aufbereitet, daß sie keinen sogenannten Datenmüll mehr enthalten und vereinheitlicht zur Auswertung vorliegen, was die weitere Verarbeitung erleichtert und in manchen Fällen erst möglich macht[20].

Nach der Erstellung "sauberer" Dateien sind methodische Fragen zu den Auszeichnungen, dem Format der Auszeichnungen und der Auszeichnungsebene zu klären.

Auszeichnungen ermöglichen komplexe Untersuchungen auf verschiedenen Ebenen. Ein nicht ausgezeichnetes Computerkorpus (die erwähnten sogenannten rohen Korpora oder *raw corpora*) kann nur auf der Ebene des orthographischen Wortes untersucht werden, nicht aber auf der Ebene theoriebedingt zugewiesener

---

[20] Bei Datenmüll handelt es sich um Zeichen, die nicht mit dem Text auf inhaltlicher Ebene verbunden sind, sondern nur zu seiner graphischen Darstellung beitragen, wie z.B. Formatierungszeichen, Verweise auf Fußnoten usw. Vereinheitlichungen hingegen beziehen sich auf die einheitliche Darstellung identischer Elemente, wie z.B. die Transformation der Konjunktion *dass* in *daß*. Fehlende Vereinheitlichungen können ungeahnte Probleme aufwerfen, z.B. wenn Programme zur Analyse der Wortarten die Konjunktion *dass* nicht als Konjunktion *daß* erkennen, sondern als unbekanntes Wort deuten, womit evtl. die Analyse des folgenden Satzes ungewöhnlich viele Fehler aufweist.

Eigenschaften. Rohe Korpora ermöglichen beispielsweise keine Unterscheidung zwischen der Konjunktion *aber* und dem Adverb *aber*. Auszeichnungen zu den Wortarten (Part of Speech oder POS) hingegen erlauben eine klare Abgrenzung beider grammatischer Funktionen, weil diese in dem Tag, d.h. der Auszeichnung, angegeben sind. Somit werden auch Untersuchungen komplexer Strukturen möglich. Zusammen mit der maschinellen Lesbarkeit stellen Auszeichnungen das differenzierende Element der Computerkorpora im Gegensatz zu traditionellen Korpora dar.

Bei den ersten Computerkorpora der sechziger Jahre handelte es sich um rohe Korpora, die einzig die textuelle Information wiedergaben, d.h. das mit keiner Zusatzinformation versehene orthographische Wort. Die Einschränkung, die dies für die Untersuchbarkeit eines Korpus darstellte, führte zur Empfehlung, die Texte eines Korpus mit Zusatzinformation in Form von Auszeichnungen zu versehen. Daß es sich dabei um eine Empfehlung und keine methodische Forderung handelt, geht klar aus den Anmerkungen verschiedener Autoren hervor. Barnbrook (1996: 107f) sieht Auszeichnungen zwar für viele Untersuchungen als unumgängliche Voraussetzung an, doch nicht für alle, und auch Sampson (1991a: 182) meint, daß ein Korpus ein „more valuable research tool" wird, wenn er zusätzlich ausgezeichnet wird, doch die Auszeichnungen nicht absolut notwendig sind. Leech fordert ebenfalls nicht unbedingt, daß ein Korpus zusätzliche Information enthält, erwähnt diese Möglichkeit aber als wünschenswert (Leech/Fligelstone 1992:

116). Zusätzlich führen Auszeichnungen zu einer Vervielfachung der Untersuchungsmöglichkeiten (Leech 1991a: 19). Auch für Johansson (1991a: 308) enthalten unausgezeichnete Texte viele Untersuchungsmöglichkeiten, doch sei das grammatische Tagging, d.h. die automatische Zuweisung von Auszeichnungen zu den Wortarten, für stilistische und syntaktische Untersuchungen sehr wichtig.

Ein rohes Korpus, das rein textuelles Material enthält, ermöglicht nur Untersuchungen auf der Grundlage des orthographischen Wortes (allgemeinen statistische Untersuchungen wie Wort, Satz und Absatzlänge und lexikalische Variation und Untersuchungen aufgrund der Methode des Pattern-Matching, wie verschiedene lexikalische Analysen bestimmter, abgeschlossener Wortarten, die keine Homonyme aufweisen). Die Frequenzlisten eines rohen Korpus beschränken sich auf die verschiedenen Erscheinungsformen des orthographischen Wortes, da keine Lemmatisierung der Einträge vorliegt, so daß sich wiederholende Lemmas als verschiedene Wörter berechnet werden.

Die Auszeichnung von a priori bestimmten Eigenschaften, wie die Wortarten, erlaubt jedoch die Untersuchung eines Korpus auf mehreren und miteinander kombinierbaren Ebenen. Mit Auszeichnungen kann der Forscher über „einige durch logische Operatoren formulierbare Relationen" hinauskommen und komplexe Suchformen erstellen (Stein 1995: 2f). In dieser Hinsicht können beispielsweise Diskursfunktionen mit grammatischen Phänomenen statistisch in Verbindung gebracht werden oder

Frequenzlisten von starken Verben anhand der Lemmatisierung der entsprechenden Einträge aufgestellt werden.

Bei Auszeichnungen ist sowohl der formale Aspekt des Formats zu beachten, als auch das theoretische System, auf dem sie beruhen (McEnery 1996: 30). Das Format der Auszeichnungen, wie z.B. die heutige *Standard Generalized Markup Language* (SGML), hat Einfluß auf die Kompatibilität zu anderen Korpora und die verwendbaren Computerprogramme. Das Auszeichnungsschema hingegen bestimmt, was in dem Korpus untersucht werden kann; so ermöglicht die Information zu den Wortarten Aussagen über die Anwendung bestimmter Wortarten, die Auszeichnungen zur Syntax über die Anwendung syntaktischer Strukturen.

Während das im folgenden Abschnitt besprochene Format technische Aspekte betrifft, hat die im darauffolgenden dargestellte Auszeichnungstypologie Auswirkungen auf erkenntnistheoretische Aspekte und somit das Wissen, das einem Korpus entnommen werden kann.

2.2.2.1 Auszeichnungsformat

Die Zusatzinformation kann dem Text in verschiedenen Formaten beigegeben werden. In heutigen Computerkorpora wird die Zusatzinformation meistens innerhalb des Textes, direkt bei dem markierten Element mit Sonderzeichen versehen eingetragen, so daß es das Programm zur Bearbeitung der Texte ist, das die Auszeichnungen herausfiltert und dem Anwender die Möglichkeit bietet, entweder den Originaltext ohne Auszeichnungen zu sehen

oder den Text versehen mit den entsprechenden Auszeichnungen. Ältere Korpora verwendeten jedoch die Methode der Querverweise, wobei die Wörter der ursprünglichen Textdatei numeriert wurden, so daß die Zusatzinformation in mit der Originaldatei verbundenen Dateien gespeichert und abgerufen werden konnte.

Die Vielfalt der Auszeichnungen, die einem Text beigefügt werden kann, führt zu einer komplexen und manchmal für den Forscher schwer verständlichen Darstellung des Originaltextes. Die Entscheidung, das Format leicht lesbar (anwenderfreundlich) oder leicht computationell verarbeitbar (programmierfreundlich) zu halten, ist dem einzelnen Forscher überlassen. Mit der Entwicklung des SGML-Formats jedoch, einer HTML-ähnlichen Kodierungssprache, bei der sowohl der Anfang als auch das Ende des auszuzeichnenden Elementes markiert wird, hat sich in den letzten Jahren ein Standard *de fakto* entwickelt (Lehr 1996: 103). Die SGML-Sprache ist relativ unüberschaubar für das menschliche Auge, die Vielfalt an Software aber, die auch dank der Entwicklung der HTML-Sprache im Internet bereitsteht, ermöglicht es, einfache Lösungen für das Problem der Lesbarkeit zu finden. Für beide Ansätze gelten jedoch die allgemeinen Hinweise von Leech (1993: 275), dessen 7 Leitlinien lauten:

- Auszeichnungen müssen entfernbar sein
- Auszeichnungen müssen für sich allein extrahierbar und anderswo speicherbar sein

- Auszeichnungen müssen auf einem System beruhen, das dem Anwender mitgegeben wird
- Es muß klar ausgezeichnet werden, wie die Auszeichnungen ausgeführt wurden und wer diese Arbeit übernommen hat
- Der Anwender muß sich dessen bewußt sein, daß Auszeichnungen Fehler enthalten können
- Auszeichnungen sollten auf allgemein akzeptierten oder neutralen Prinzipien beruhen
- Kein Auszeichnungssystem darf sich selbst als Standard ansehen.

Die meisten Auszeichnungsformate erfüllen diese sieben Leitlinien. Das bei weitem verbreitetste Auszeichnungsformat ist SGML (*Standard Generalized Markup Language*), obwohl daneben und aufgrund der leichteren Lesbarkeit auch andere Formate verwendet werden, wie das Format des Brown Korpus.

Neben dieser „nicht normativen Standardsprache" bestehen weiterhin frei erfundene Formate, die vor allem von der vorherigen Tradition in der Linie des Brown Korpus, den vorhandenen Auszeichnungsprogrammen und den technischen Einschränkungen abhängig sind. Technische Probleme bereiten diese einfacher lesbaren Systeme, wenn es sich um komplexe Auszeichnungen handelt (von diskontinuierlichen Elementen, Mehrwortelementen u.a.).

Im Brown Korpus (Vgl. Francis 1980) wurde in den Texten die POS-Struktur nur für einzelne Wörter ausgezeichnet, die

mittels einer Unterstreichung mit dem entsprechenden Kode verbunden waren:

das_Art Haus_NN

Dieses Format wird heute immer noch verwendet, z.B. für den POS-Tagger des Instituts für deutsche Sprache. Es ermöglicht einzig die Auszeichnung einzelner Wörter, nicht aber von Wortgruppen („Die Katze *auf dem Dach*"). Dies führte zur Einführung des SGML-Formats. Anders als im Brown Korpus verfügt ein Tag bei diesem Format über zwei Teile: einen Anfangstag und einen Endtag. Dadurch können Mehrwortelemente ausgezeichnet werden, wie zum Beispiel eine Präpositionalgruppe (PG):

Die Katze <PG-Anfang>auf dem Dach<PG-Ende>

Im Vergleich zu ähnlichen Systemen mit Anfangs- und Endmarkern in Form von Klammern (oft in der Programmiersprache verwendet und auch für syntaktische Auszeichnungen wie die Penn-Treebank, vgl. Marcus et al. 1993: 320), bietet dieses System einige Vorteile. Die Klammern ermöglichen zwar die Einnistung von Elementen und somit eine parallele Auszeichnung der Texte auf mehreren Ebenen, z.B. POS-Auszeichnungen und syntaktische Auszeichnungen gleichzeitig. Die POS-Auszeichnung sähe folgendermaßen im Klammerformat aus:

[Die ART] [Katze NN] [auf PRÄP] [dem ART] [Dach NN]

Die syntaktische so:

[Die Katze SUBJ] [auf dem Dach PG]

Die Verbindung der beiden ergäbe folgendes Format:

[[Die ART] [Katze NN]SUBJ] [[auf PRÄP] [dem ART] [Dach NN]PG]

Die Klammern bereiten Probleme hinsichtlich der Lesbarkeit, obwohl sich der Leser schnell daran gewöhnt, was an der steigenden Produktivität der Forscher erkennbar ist, die Revisionen an den Texten durchführen. Für den Computer hingegen stellt es ein Ideal dar, da die Maschine hierarchisch die Auszeichnungen ablesen und zuordnen kann. Klammern stoßen jedoch an ihre Grenzen, wenn grundverschiedene, sich überlappende Auszeichnungssysteme angewendet werden, so daß der Computer die Tags nicht mehr korrekt zuordnen kann. Wenn beispielsweise in einem Theaterstück gleichzeitig die *dramatis personae* angegeben werden zusammen mit thematischen Einheiten. Nehmen wir folgendes Beispiel:

Hans: Die Katze ist auf...

Volker: Ich habe es gesehen. Auf dem Dach ist die Katze. Aber nun zum Hund.

Wenn dabei gleichzeitig die verschiedenen Sprecher ausgezeichnet werden sollen und das Thema Katze, wäre das mit dem Klammersystem unerreichbar, weil die Zuordnung der Öffnungsklammern nicht möglich wäre oder für das menschliche Auge ganz unverständlich. SGML bietet hier eine sowohl für den Menschen als auch den Computer überschaubarere Lösung, da

zusätzlich zu dem Sonderzeichen des Tags noch der Auszeichnungstyp erscheint. Dadurch werden die verschiedenen Einheiten überschaubarer:

> Hans: <Sprecher Hans Anfang><Thema Katze Anfang>Die Katze ist auf...<Sprecher Hans Ende>
>
> Volker: <Sprecher Volker Anfang>Ich habe es gesehen. Auf dem Dach ist die Katze.<Thema Katze Ende> <Thema Hund Anfang> Aber nun zum Hund. <Thema Hund Ende><Sprecher Volker Ende>

Abgesehen von der Wahl der konkreten Auszeichnungsmethode gibt es drei Vorgehensweisen zur Anwendung der Auszeichnungen (nach Leech/Fligelstone 1992: 133f.):

- Die automatische Verarbeitung mit Nachbearbeitung, wobei der jeweilige Text von einem Programm automatisch bearbeitet wird und das Ergebnis des Prozesses einer manuellen Revision unterworfen wird;
- die interaktive Verarbeitung, wobei das Programm den Anwender jeweils fragt, welches Element einer vorgegebenen Liste eingegeben werden soll; und schließlich
- die manuelle Eingabe mit maschineller Unterstützung, wobei einige Funktionen mit Unterstützung des Programme vorgenommen werden.

Diese drei Vorgehensweisen werden in Abhängigkeit von den auszuzeichnenden Phänomenen und der Fehlerquote, die das Programm, das den automatisierten Teil der Arbeit übernimmt,

ausgewählt. Eine nur manuelle Vorgehensweise wird im allgemeinen vermieden, weil dies zu einer hohen Anzahl an Fehlern in der Eingabe führen kann. Der Vergleich der Ergebnisse einer ausschließlich manuellen POS-Auszeichnung mit einer automatisierten Auszeichnung und anschließender menschlicher Korrektur zeigt, daß die manuelle Auszeichnung doppelt so lang wie die automatisierte plus menschlicher Korrektur dauert und zu 50% mehr Fehlern in der Zuweisung der Tags führt (Marcus et al. 1993: 318f).

2.2.2.2 Auszeichnungsebenen

Hinsichtlich der Auszeichnungsebenen, d.h. der theoretisch bedingten Elemente, die im Text mit Tags versehen werden, werden in Computerkorpora hauptsächlich Wortarten und Syntax automatisiert gekennzeichnet. Der Prozeß der Zuweisung von Wortarten wird, wie schon erwähnt, Tagging genannt, das Verfahren zur syntaktischen Auszeichnung hingegen Parsing. Andere Auszeichnungen, die dank des SGML-Formats auf einen Text angewendet werden können, der schon getaggt oder geparst ist, wären Auszeichnungen mit lexikalischer, pragmatischer oder anderer Information. Dementsprechend werden Korpora, die keine Auszeichnungen enthalten, dann rohe Korpora genannt. Im folgenden sollen die Eigenschaften der verschiedenen Auszeichnungsebenen dargestellt werden.

Getaggte Korpora (*tagged corpora*) enthalten Auszeichnungen zu den grammatischen Wortarten (*Part of Speech Tagging* oder kurz

POS Tagging). Aufgrund noch bestehender technischer Einschränkungen wird nur das orthographisch vorliegende Wort bearbeitet. Dementsprechend werden Mehrworteinheiten (z.B. *eine Frage stellen*), getrennt erscheinende Einheiten (z.B. *er stellt ihm eine Frage*) oder diskontinuierliche Einheiten (z.B. *er stellt ihm einen Brief zu*) nicht gesondert ausgezeichnet. Der Automatisierungsprozeß für diese Auszeichnungsform erzielt hohe Erfolgsraten, jedoch ist immer noch eine manuelle Revision erforderlich. Bei geparsten Korpora (*parsed corpora*) werden die Texte in syntaktische Strukturen zerlegt und ausgezeichnet. Im Gegensatz zu den Tags der Wortarten (vgl. z.B. Schiller et al. 1995), die auf "allgemein akzeptierten Einheiten" beruhen (Leech 1993: 275), entspringen die Einheiten der syntaktischen Auszeichnung einer grammatischen Theorie (vgl. Souter 1993: 198f.; Souter/Atwell 1994: 143f. und Barnbrook 1996: 127). Zusätzlich zu diesen Auszeichnungsschemata können in einem Korpus noch weitere Phänomene ausgezeichnet werden.

Beim Tagging oder der Auszeichnung der Wortarten wird, wie angegeben, jedem Wort eine grammatische Kategorie zugewiesen. Die Kategorien werden aufgrund von allgemein akzeptierten Kriterien erstellt, also einer Zusammenfassung von grammatischen Regeln verschiedenen Ursprungs (vgl. Leech 1993), für die Bearbeitung durch ein Computerprogramm geeignet sind, nicht aber aufgrund einer bestimmten grammatischen Theorie. Nichtsdestotrotz ist anzumerken, daß das Hauptkriterium zur Bestimmung der Wortarten distributionell

ist (vgl. Schiller et al. 1995), und semantische und morphologische normalerweise erst an zweiter Stelle rangieren. Der Umfang der Kategorien reicht von den Grundwortarten bis hin zu Subklassifizierungen einer jeden Wortart und kann zusätzlich noch morphologische Information enthalten. Die Kategorien werden mittels eines Taggingprogrammes[21] auf die entsprechenden Texte angewendet. Da aus Bearbeitungsgründen jedem Wort im getaggten Text nur eine Kategorie zugewiesen werden darf, es aber mehreren angehören kann (je nach Ausführlichkeit des zusammengestellten Tagsets[22]), und aufgrund der Probleme, die die computationelle Bearbeitung bereitet[23], arbeiten diese Programme mit einer bestimmten Fehlerquote, die von der Arbeitsweise und der Komplexität des Programms abhängt. Die Programme erzielen mittlerweile Trefferquoten von 98,3% (Leech 1997b), aber auf der Grundlage der Wortarten, so daß einige sehr häufige schwer identifizierbare Wortarten die Gesamtfehlerquote stark ansteigen lassen können. Andererseits

---

[21] Wie z.B. der ImS-Tagger (Institut für maschinelle Sprachverarbeitung 1998).

[22] Das Tagset STTS besteht aus 51 Tags (Schiller et al. 1995), das C5 Tagset des BNC aus 61 Tags und C7 Tagset des BNC aus 160 Tags (British National Corpus 1997b). Zu dem in dieser Arbeit verwendeten Tagset STTS vgl. den Anhang.

[23] Taggingprogramme funktionieren im allgemeinen hierarchisch; sie weisen den Wörtern aufgrund des Vergleiches mit generellen Regeln und einem Lexikon mehre Kategorien zu, die dann mit den vorhergehenden und nachfolgenden verglichen werden. Wenn eine bestimmte Verbindung nicht möglich ist, wird der fehlerhaft zugeordnete Tag gestrichen, so daß am Ende jedes Wort nur einen Tag erhält.

bedeutet dies auch, daß selbst bei der Annahme von einer Trefferquote von 90% auf einer standardisierten 300-Wort-Seite immer noch ugf. 30 Fehler erscheinen, die manchmal nur schwer erkennbar sind und eine mehr oder weniger computerunterstützte Nachbearbeitung erfordern.

Im Gegensatz zum Tagging ist das Parsing computationell nicht so weit entwickelt, da es eine wesentlich größere technische Komplexität aufweist. Das syntaktische Auszeichnungsschema beruht nicht mehr wie beim Parsing auf einer pragmatischen Zusammenstellung von Kategorien, sondern auf einer spezifischen syntaktischen Theorie, wie z.B. im *English Constraint Grammar Parser* (vgl. Kytö/Voutilainen 1995). Es werden nicht mehr Einzelwörter ausgezeichnet, sondern die Abhängigkeitsbeziehungen der Wörter oder Wortgruppen untereinander dargestellt, die auch getrennt erscheinen können, wie im Falle der mehrteiligen Konjunktionen. Ein Beispiel für diesen grundlegenden Unterschied ist im Englischen die Struktur *the one*. Ein POS-Tagger zeichnet sie als *CD* (Cardinal Number) aus. Im Plural, *the ones* wird *ones* dann aber trotz paralleler Funktion als NNS (Plural Common Noun) angegeben. In der syntaktischen Auszeichnung würde das *one* hingegen als N (Singular Common Noun) anzugeben sein, und *ones* als NNS (Plural Common Noun) (Marcus et al. 1993: 315f).

Die technischen Probleme zur Darstellung dieser Strukturen werden mit syntaktischen Bäumen dargestellt und als Klammern in Maschinensprache umgewandelt. Die Fehlerquote der

Programme, die diese Arbeit übernehmen, ist sehr groß, so daß die Programm heute fast nur zur Unterstützung des Auszeichnungsprozesses eingesetzt werden.

Abgesehen von den beiden häufigsten Auszeichnungsebenen, dem POS-Tagging und dem syntaktischen Parsing, werden weitere Auszeichnungsebenen projektspezifisch angewendet.

Dazu gehört beispielsweise die semantische Auszeichnung zu untersuchender Phänomene; dieser Prozeß befindet sich im Vergleich zum Tagging und Parsing noch in einem Anfangsstadium (Meunier 1998: 25). Die Wörter eines Textes werden bestimmten vorgegebenen Kategorien zugeteilt (vgl. Wilson/Rayson 1993: 216), deren Anzahl variieren kann; das System von Wilson und Rayson z.B. sieht 21 Hauptkategorien vor, wobei Antonyme mit den Zeichen + oder - versehen werden.

Die Fehleranalyse hat ebenfalls seit kurzem die Anwendbarkeit eines Auszeichnungsschemas erkannt; Granger verwendet für die Erstellung des *International Corpus of Learner English* (ICLE) ein hierarchisches Fehlerauszeichnungsschema, das die Untersuchung der Textproduktion von Englischlernern ermöglicht (Granger 1996). Problematisch bei diesen beiden Ansätzen ist, daß sie nur bis zu einem geringen Grad automatisierbar sind (vgl. Meunier 1998: 26); im Falle der Fehleranalyse kommt hinzu, daß Ansätze zu einer Automatisierung sich oft auf einer oberflächlichen und sehr reduzierten Ebene bewegen, die keinen Ansatz zu neuen Entwicklungen anbieten. So schlägt beispielsweise Meunier vor, zur Auszeichnung von

orthographischen Fehlern auf die orthographischen Korrekturfunktionen der Textverarbeitungen zurückzugreifen, wobei aber die teilweise sehr komplexen Fehlleistungen von Lernern nicht berücksichtigt werden.

3 Computerlernerkorpus (CLK) und Computermuttersprachlerkorpus (CMK)

Computerlernerkorpora gehören der Gruppe spezifischer Computerkorpora an (Granger 1998b), denn sie werden mit einem konkreten Forschungsziel erstellt, das zusammen mit den textinternen und textexternen Eigenschaften der Textprodukte der Informanten die Form des Korpus bedingt. Dementsprechend eingeschränkt ist, zumindest im Deutschen, die Möglichkeit zur Wiederverwendung bestehender Korpora. Viele deutschsprachige Korpora zeichnen sich augenblicklich noch durch die beschränkte Auswahl der Textsorten aus. Für die angestrebte textsortenabhängige Untersuchung stand kein geeignetes Computerkorpus zur Verfügung, und Computerkorpora zu Deutschlernern im Sinne der im vorhergehenden Kapitel besprochenen Eigenschaften sind unseres Wissens völlig unexistent. Daraus ergab sich die Notwendigkeit, die notwendigen Computerkorpora für die vorliegende Arbeit zu erstellen. Einerseits wurde ein zu untersuchendes Computerkorpus von Lernertexten benötigt, andererseits ein muttersprachliches Kontrollkorpus, mit dem das Computerlernerkorpus verglichen werden konnte.

Zur Bestimmung der Kriterien, die für die Erstellung beider Computerkorpora berücksichtigt werden mußten, wurden folgende Aspekte in Betracht gezogen:

- Erkenntnisse angrenzender Forschungsgebiete, wie der Soziolinguistik, der Fehleranalyse und der Lernersprachenanalyse.
- Die Kriterien zur Erstellung des *International Corpus of Learner English* (ICLE), des einzigen uns bekannten umfangreichen Lernerkorpus.
- Theoretische Aspekte der nicht spezifisch lernerorientierten korpuslinguistischen Methode.

Als Erkenntnisse angrenzender Forschungsgebiete fassen wir methodische Vorgehensweisen der Fehleranalyse und der Lernersprachenanalyse auf, als auch Methoden der Soziolinguistik, die die Lernersprachenanalyse beeinflußt haben, die jetzt kurz besprochen werden sollen.

Im Rahmen der Fehleranalyse wurde gegen Ende der 60er und in den 70er Jahren der Versuch unternommen, Fehler als Hinweis auf die Phasen des Spracherwerbs zu nutzen. Für die Untersuchungen wurden authentische Texte gesammelt und ausgewertet (vgl. Nickel 1972). Die Auswertung ist als qualitativ zu bezeichnen[24] und statistische-korpusbedingte Aspekte fanden kaum Beachtung.

[24] Vgl. die Arbeiten von Gnutzmann (1972) und Drubig (1972).

Die Untersuchungen der Fehleranalyse wurden weiterentwickelt durch die Lernersprachenanalyse. Sie hatte zum Ziel, Aspekte des Fremdsprachenerwerbs mutidisziplinär zu erklären. Anfangs wurden methodische Elemente der Soziolinguistik angewendet, was zur Erforschung der Variablen führte, die bestimmte Wiederholungen oder Regelmäßigkeiten bedingten. In Anlehnung an die Soziolinguistik wurde dabei der gesprochenen Lernersprache Vorrang gegeben. Empirisches Material konnte als repräsentatives Material für eine abgegrenzte Lernergruppe aufgefaßt werden, die im Extremfall —im Kontrast zu Computerlernerkorpora— aus einem einzigen Individuum bestehen konnte. Dieses empirische Textmaterial wurde zu Korpora zusammengefaßt, die allerdings nicht unbedingt den Anforderungen von Computerlernerkorpora entsprachen, wie Biber (1998: 172ff) betont.

Fehleranalyse und Lernersprachenanalyse unterscheiden sich von Untersuchungen zu Computerlernerkorpora aufgrund ihres theoretischen Ranges (vgl. S. 57). Bei Fehleranalyse und Lernersprachenanalyse handelt es sich um mehr oder weniger unabhängige Forschungsbereiche, d.h. Disziplinen, die als solche eigene Forschungsziele aufweisen. Computerlernerkorpora hingegen stellen empirisches Datenmaterial bereit. Sie können in vielen Disziplinen Anwendung finden, doch aufgrund ihrer konkreten Form (maschinelle Lesbarkeit, Auszeichnungen usw.) können sie einer Methode zugeschrieben werden, die ihrerseits den Anwendungsbereich in eigenständigen Disziplinen bedingt, wie der Lexikographie, der Didaktik, der interkulturellen

Forschung usw. Dementsprechend können Computerkorpora auch für Lernersprachenanalysen herangezogen werden, während die Korpora der Lernersprachenanalyse nicht unbedingt den Ansprüchen eines Computerlernerkorpus entsprechen. Dennoch enthalten Fehleranalyse und Lernersprachenanalyse verwendbare Ansätze, die später besprochen werden.

Aus der unterschiedlichen Konzeption von Korpora in Fehleranalyse und Lernersprachenanalyse einerseits und Korpuslinguistik andererseits (vgl. S. 17ff) entsteht die Notwendigkeit, die Anwendungsmöglichkeiten von nicht maschinell lesbaren Lernerkorpora (z.B. der Lernersprachenanalyse) von denen der Computerlernerkorpora abzugrenzen. Bis heute verfügbare nicht maschinell lesbare Lernerkorpora zeichnen sich durch zwei Eigenschaften aus: einerseits handelt es sich um *Korpora*, die anhand festgelegter Methoden zusammengestellt werden (vgl. Kapitel 2); andererseits dienen sie dem *Zweck* der Untersuchung der Textproduktion von Lernern.

Computerlernerkorpora reihen sich in die produktorientierten Untersuchungen zur Lernersprache ein; anhand des Textmaterials soll laut Leech (1998) dargestellt werden können,

- welche sprachlichen Phänomene ein Lerner öfter oder seltener anwendet als ein Muttersprachler;
- inwieweit die Lernersprache von der Muttersprache beeinflußt wird;
- in welchen Bereichen Vermeidungsstrategien zum Ausdruck kommen;

- in welchen Bereichen nahezu muttersprachliche Kompetenz erreicht wird;
- welche sprachlich verallgemeinerbaren Phänomene einer Gruppe am stärksten als nicht muttersprachlich einzustufen sind, so daß an ihnen besonders intensiv gearbeitet werden kann.

Diese Ziele weichen von der mentalistischen Tendenz der Lernersprachenanalyse ab, die die prozeßorientierte Erklärung der Lernprozesse anstrebt. Beide Ansätze entsprechen somit unterschiedlichen Auffassungen des Forschungsgegenstands (Lehr 1996: 10).

Für die Zusammenstellung des Computerlernerkorpus und des Computermuttersprachlerkorpus wurde die Einteilung in korpusbezogene Eigenschaften, d.h. Kriterien, die sich auf das Korpus als repräsentative Darstellung einer gegebenen Grundgesamtheit beziehen (vgl. Lehr 1996: 120 zu Fragen der angewandten Repräsentativität), und textbezogenen Eigenschaften, die eine nähere Charakterisierung der Einzeltexte ermöglichen, berücksichtigt. Dabei führten die Abweichungen der Besonderheiten zwischen Computerlerner- und Computermuttersprachlerkorpus zu einer nicht immer identischen Anwendung aller korpusbezogenen Eigenschaften, wie Zielsetzung, Strukturierung, Repräsentativität, Korpusumfang, Abgeschlossenheit und maschinelle Lesbarkeit, da z.B. im Computermuttersprachlerkorpus keine Sprachstandebenen vorhanden waren, d.h. nur im Computerlernerkorpus konnten Lerner nach erreichten Sprachkenntnissen eingeteilt werden. Auch von den textbezogenen Aspekten konnten aufgrund der

Spezifizität beider Korpora weder alle berücksichtigt noch identisch angewendet werden.

3.1 Korpusbezogene Voraussetzungen

Die Anwendung der korpusbezogenen Voraussetzungen (vgl. 2.1) auf das Computerlerner- und das Computermuttersprachlerkorpus erfolgte unter Berücksichtigung der Spezifizität der Korpora und der vorzunehmenden Untersuchungen. Die Umsetzung der Voraussetzungen wurde besonders durch die abweichende Anwendung der beiden Korpora geprägt, als Lernerkorpus und als Referenzkorpus. Auch durch die Forderung der Kompatibilität und der Vergleichbarkeit, die beide Korpora trotz differierenden Forschungszwecks zu erfüllen hatten, hatte Einfluß auf die Umsetzung der Voraussetzungen.

3.1.1 Allgemeine theoretische Vorüberlegungen und Zielsetzung

Die Zusammenstellung der beiden Computerkorpora mußte so erfolgen, daß die Abweichungen im Sprachgebrauch zwischen Lernern und Muttersprachlern darstellbar würden. Zudem sollten beide Korpora sowohl zusammen als auch getrennt für weiterführende Untersuchungen bereitstehen. Beide Ziele konnten durch die Berücksichtigung von zwei Aspekten erreicht werden: die allgemeinen Kriterien zur Zusammenstellung eines Computerkorpus und die spezifisch von den Computerlernerkorpora bedingten Kriterien, u.a. den Kriterien, die auch nicht maschinell lesbare Lernerkorpora betreffen.

Die Besonderheiten des Computerlerner- und des Computermuttersprachlerkorpus, wie z.B. die Fehlerkorrektur, der die Texte des Lernerkorpus unterworfen wurden, die differierenden Zwecke der beiden Korpora, die Abweichungen in der Textlänge usw. erforderten zusätzlich spezifische Lösungen, die in den entsprechenden Abschnitten besprochen werden.

Angesichts der Möglichkeit weiterführender Untersuchungen wurde ebenfalls darauf geachtet, daß beide Korpora ohne Minderung ihrer Anwendbarkeit für die gesetzten Ziele auch in Zukunft für Untersuchungen über Lernersprache und eventuell für andere Forschungen wiederverwendbar seien. Dies bedingt konkrete Entscheidungen vor allem hinsichtlich des Formats der Texte und zu den Auszeichnungen, die es ermöglichen, daß das vorliegende Lernerkorpus teil eines größeren Deutschlernerkorpus bilden könnte, das u.a. auch für die Untersuchung kulturell bedingter Lernunterschiede herangezogen werden könnte (vgl. Leech/Fallon 1992a).

3.1.2 Repräsentativität, Strukturierung und Variationskriterien

Beide Korpora, das Computerlerner- und das Computermuttersprachlerkorpus (wie schon angegeben CLK und CMK), unterscheiden sich hinsichtlich ihrer Anwendung in dieser Untersuchung. Die Verwendung des CMK als Referenz- oder Vergleichskorpus und des CLK als zu untersuchendes Korpus führt zu einer grundlegenden Abweichung der Repräsentativität

beider Computerkorpora. Im Falle des CLK soll eine reduzierte Population untersucht werden, weil die Zusammenstellung eines internationalen CLK den Rahmen dieser Arbeit gesprengt und zu einem viel größeren CLK geführt hätte. Im CMK hingegen wird eine unvergleichlich große Population zusammengefaßt, deren sprachliches Verhalten als Vergleich dient.

Gleichzeitig war zu beachten, daß sich die Verfasser der Texte beider Computerkorpora aufgrund ihres Sprachstandes unterscheiden. Bei den Autoren des CMK war der maximal mögliche Sprachstand vorauszusetzen, d.h. daß sie die Sprache auf muttersprachlicher Ebene beherrschen. Ebenfalls war anzunehmen, daß die Autoren beim Verfassen der Texte sprachliche Strukturen ohne die Begrenzung nur ungenügender sprachlicher Kenntnisse anwendeten, d.h. jene auswählten, die ihnen jeweils angebracht schienen. Die Verfasser der Texte des CLK hingegen befanden sich im Prozeß der Aneignung sprachlicher Kenntnisse, so daß die Anwendung sprachlicher Strukturen sich nicht ausschließlich aus einer freien Wahl zwischen anwendbaren Strukturen ergab, sondern der Einschränkung des bis zum Augenblick der Niederschrift erreichten Sprachstands unterlag.

Die Differenzierung der Repräsentativität und des Sprachstands beider Populationen führt zu einer Unterscheidung der anwendbaren Methoden zur Erstellung der Korpora. Während das CMK aufgrund der in Kapitel 2 dargestellten Kriterien aufgebaut werden konnte und als normales, spezifisches

Computerkorpus anzusehen war, schien es im Falle des CLK angebracht, besondere Kriterien hinsichtlich des Lernercharakters zu berücksichtigen, d.h. CLK-spezifische Kriterien für die Umsetzung von Repräsentativität, Strukturierung und Variation. Einen ersten Ansatz zur Ermittlung CLK-spezifischer Kriterien bot die Fehleranalyse der 60er und 70er Jahre. Dabei war jedoch zu beachten, daß methodische Unstimmigkeiten der Fehleranalyse der 60er und 70 Jahre[25] vermieden wurden, die die Repräsentativität der angelegten Textsammlungen beeinträchtigt und somit auch die Validität der Untersuchungen.

Für die ersten Untersuchungen zur Produktion von Lernertexten wurden ebenfalls Textsammlungen von Lernertexten verwendet. Anhand solcher Textsammlungen konnten Ansätze zur Bestimmung von Variablen ermittelt werden, die Einfluß auf die Textproduktion von Lernern hatten, wie z.B. der Einfluß der Muttersprache, die Häufigkeit von fehlerhaften Realisierungen in Abhängigkeit des Sprachstands usw. Die Ergebnisse der Untersuchungen sollten verallgemeinerbar sein, d.h. zur allgemeinen Beschreibung des Lernprozesses einer Fremdsprache beitragen. An dieser Vorgehensweise erwies sich hauptsächlich die methodische Vorgehensweise als problematisch. Granger (1998b) faßt folgende Aspekte zusammen, die auf der

[25] Vgl. z.B. Nickel 1972a: fehlende Angaben zum Umfang der Korpora und der Einzeltexte, zu den Eigenschaften der Texte und Verfasser u.a.

Unterscheidung zwischen Textsammlung, Korpora und Computerkorpora beruhen:

- Der Umfang der Textsammlungen war oft sehr reduziert. In einigen Fällen handelte es sich um wenig mehr als 2000 Wörter.
- Selbst größere Korpora machen oft keine Angaben zum exakten Umfang.
- Die Auswahl der Texte berücksichtigte nicht die erforderlichen Variationskriterien.

Diese drei Aspekte erschweren die Verallgemeinerung der Forschungsergebnisse, wie Ellis (1994: 49) hervorhebt: „Many EA studies have not paid sufficient attention to these factors, with the result that they are difficult to interpret and almost impossible to replicate".

Odlin kritisiert ebenfalls die methodischen Grundlagen dieser Untersuchungen zur Fehleranalyse; er bemerkt eine „considerable variation in the number of subjects, in the backgrounds of the subjects, and in the empirical data, which come from tape-recorded samples of speech, from student writing, from various types of tests, and from other sources" (Odlin 1989: 151), so daß eine Verbesserung der Methode zur Erstellung des Materials für ihn erstrebenswert erscheint.

Ähnliche Kritiken können an der deutschsprachigen Forschung zur Fehleranalyse geäußert werden. Gnutzmann (1972: 67) äußert sich in seiner *Analyse lexikalischer Fehler* weder zum Gesamtumfang seiner Textsammlung, noch scheint er eine

Einteilung in Textsorten bei seiner Zusammenstellung zu beachten; die Angaben zu den lexikalischen Fehlern erfolgen in Prozent, aber ohne daß dabei darauf hingewiesen würde, ob der Umfang der benutzten Textsorten vergleichbar wäre. Ebenso Drubig (1972: 79); er beschränkt sich zwar auf die Textsorte „Nacherzählung", gibt aber auch keine Werte zur Textsammlung an.

Die Anwendung der Methoden für die Zusammenstellung von Computerkorpora tragen zu einer Revision der Aufbereitung des empirischen Datenmaterials bei, das in der Fehleranalyse verwendet wurde. Eines der Probleme der Textsammlungen der Fehleranalyse war die Variation.

Bei den Texten der Lerner und der Muttersprachler handelt es sich um variable, nicht uniforme und heterogene Sprachrealisationen. Die Heterogeneität der Texte ist aber nicht ausschließlich willkürlich, wie Untersuchungen ergaben[26].

[26] Zur Überprüfung wurden Lernertexte zunächst ohne Strukturierung auf sprachliche Phänomene wie lexikalische Variation oder Anwendung von Wortarten untersucht. Die Ergebnisse waren stark heterogen, unterschieden sich jedoch deutlich von denen der Texte der Muttersprachler. Nach einer vorläufigen Strukturierung der Lernertexte nach Variationskriterien wie Sprachstand und Textsorte und der erneuten Untersuchung einiger sprachlicher Phänomene (lexikalische Variation oder Anwendung von Wortarten) ergab sich, daß die Heterogeneität stark zurückging, sich aber die ermittelten Werte weiterhin deutlich von den vergleichbar strukturierten Texten der Muttersprachler unterschieden.

Der klare Hinweis auf eine strukturierte und nicht freie Variation innerhalb der Texte (vgl. Turell 1995b: 20f) führte dementsprechend zur Notwendigkeit, die Texte nach Variationskriterien zu strukturieren, d.h. Variablen, die die Textproduktion beeinflussen.

Zur Ermittlung der Variationskriterien wurden in anderen Bereichen berücksichtigte Variablen auf ihre Anwendbarkeit untersucht. Zu diesen Bereichen gehören die Soziolinguistik, das *International Corpus of Learner English* (*ICLE*) (Granger 1998a), und allgemeinen korpuslinguistische Variationskriterien (vgl. 2.1.2).

Die Variationskriterien der Soziolinguistik und des ICLE sollten bestimmen helfen, inwiefern textexterne Faktoren wie Alter, Bildungsgrad u.a. Einfluß auf die Textproduktion der gewählten Population hatten. Aus der Überprüfung allgemeiner Variationskriterien, wie die in 2.1.2 besprochenen, erhofften wir uns zusätzliche Hinweise auf Variablen, die auf unsere Untersuchung anwendbar waren.

Die Variationskriterien der Soziolinguistik beziehen sich auf textexterne Faktoren, die hauptsächlich einer näheren Definition des Informanten dienen, wie z.B. Alter, Bildungsgrad oder Beruf. Ihre Anwendung setzt eine strukturell-heterogene Population voraus, die im Fall unserer Informanten nicht gegeben war. Somit waren Bildung und Beruf entweder irrelevant oder nicht anwendbar: bei den Informanten handelte es sich um Studenten, die sich noch in der

Bildungsphase befinden und im allgemeinen noch keinen festen Beruf ausüben, der ihr sprachliches Verhalten hätte grundlegend prägen können. Auch das Geschlecht der Informanten, ein in der Soziolinguistik weitverbreitetes Variationskriterium, scheint aufgrund der hohen Spezifizität der Textsorten wenig relevant zu sein, was anhand entsprechender Voruntersuchungen überprüft wurde.

Die im ICLE verwendeten Variationskriterien konnten aufgrund der Eigenschaften der Population nur auf das CLK angewendet werden, bekräftigten jedoch die ursprüngliche Einteilung unseres CLK nach Sprachstand und Textsorte. Weitere Variationskriterien des ICLE hingegen, die nicht mit Sprachstand und Textsorte verbunden sind, erwiesen sich als nicht ergiebig, vor allem aufgrund der Zielsetzung des ICLE, die stark von unserem CLK abweicht.

Das ICLE ist ein internationales Englischlernerkorpus, für das alle Lerner der englischen Sprache in allen Lernkontexten berücksichtigt werden. Zur Darstellung dieser Zielgruppe werden im ICLE zwei Hauptebenen berücksichtigt, Sprache und Lernerinformation (Granger 1998b: 8ff), denen jeweils Variationskriterien entsprechen (zur Hauptebene Sprache gehören die Variationskriterien Gattung, Thema, Technizität und Aufgabe, zur Hauptebene Lernerinformation die Variationskriterien Alter, Geschlecht, Muttersprache u.a.).

Für unser Computerlernerkorpus erwies sich nach entsprechender Voruntersuchung nur das Variationskriterium Gattung, das wir

als Textsorte übernommen haben, als ergiebig, da hier große Unterschiede festgestellt werden konnten. Im ICLE wird Gattung (*Genre*) als sprachübungsspezifische Textsorte aufgefaßt, d.h. es handelt sich um im Sprachunterricht produzierte Übungstexte der Art Aufsatz, Brief u.a. Diese Typologie konnte nicht auf unsere Lernertexte angewendet werden, da diese nicht ausschließlich mit dem Ziel einer Sprachübung produziert wurden, sondern auch der Umsetzung akademischer Vorgehensweisen dienten. Dementsprechend wurde bei der Strukturierung unserer Texte der Begriff Gattung des ICLE als Textsorteneinteilung nach akademischen Kriterien berücksichtigt, die zur Bestimmung von Textsorten wie Exposé, Einleitung und Bericht führten.

Die Unterscheidung hinsichtlich der Textsorten führte ebenfalls dazu, daß die weiteren Variationskriterien der Hauptebene Sprache des ICLE, d.h. Thema, Technizität und Aufgabe, sich als unanwendbar oder überflüssig erwiesen. In den Voruntersuchungen waren keine Differenzen aufgrund des Themas oder der Technizität zu erkennen, was auf die Spezifizität der betrachteten akademischen Textsorten zurückzuführen ist. Zusätzlich ist zu beachten, daß bei den quantitativen Untersuchungen keine lexikalischen Analysen vorgesehen waren, die Verbindung zum Thema oder der Technizität aufweisen konnten, so daß die Berücksichtigung dieser Variationskriterien unnötig war.

Die Aufgabenstellung hingegen, die zum Text führte, erübrigte sich, weil sie als Sprachübung in den Texten unserer Informanten nicht vorhanden war, d.h. die Informanten wurden nicht dazu angehalten, z.B. Relativsätze oder Partizipialkonstruktionen anzuwenden[27].

Hinsichtlich des Lerners wurde in verschiedenen Voruntersuchungen überprüft, ob Alter oder Geschlecht Einfluß auf die Textproduktion hatten. Dabei stellte sich heraus, daß nur hinsichtlich des Alters Unterschiede festzustellen waren[28]. Das Geschlecht hingegen führte zu keinen relevanten Schwankungen der Ergebnisse der Voruntersuchungen.

Zusätzlich zu den Variationskriterien wurde der Vermeidung häufig auftretender Fehler im Korpusdesign besondere Beachtung geschenkt. Schlobinski weist auf zwei Aspekte hin, die für die statistische Repräsentativität zu beachten sind:

> Bei der Beurteilung der Stichprobe sind zwei Gütekriterien zu beachten, nämlich der **Auswahlbias** (systematischer Fehler oder kurz: Bias) und der **Auswahlfehler** (Zufallsfehler). Der

---

[27] Weitere Untersuchungen sollten jedoch hinsichtlich didaktischer Anwendungen unternommen werden, die die Aufgabenstellung der Textproduktion einbeziehen, nicht als Schreibübung, sondern als authentisches kommunikatives Produkt.

[28] Bei den nach Alter eingeteilten Texten konnten Abweichungen lexikalischer Natur entdeckt werden: ältere Studenten erzielten einen extrem hohen Grad an lexikalischer Variation. Es handelte sich um sehr wenige Studenten, die alle älter als 40 waren. Aufgrund des stark reduzierten Textmaterials, das von diesen Studenten zur Verfügung stand, wurden ihre Texte nicht in das CLK einbezogen.

> Auswahlfehler liegt dann vor, wenn die Präzision der Stichprobe mangelhaft ist. Je kleiner die Stichprobe, desto wahrscheinlicher und größer ist der Auswahlfehler. Unter Bias versteht man die „systematische Abweichung einer Statistik vom Parameter" (Knieper 1993: 60). Ein Bias liegt dann vor, wenn die Stichprobe nach anderen Parametern gezogen wird, als sie in der Grundgesamtheit vorkommen. Die Stichprobe ist dann ‚verzerrt' bzw. ‚verfälscht'.[29] (Schlobinski 1996: 26)

Beide Fehlerquellen waren für unsere Computerkorpora insofern wichtig, als daß die Gesamtlänge der Korpora aufgrund schon erwähnter Faktoren reduziert gehalten werden mußte: je größer die Auswahl, desto niedriger die Gefahr, daß ein Auswahlfehler auftritt. Um das Auftreten eines Auswahlfehlers zu begrenzen, wurden die Texte nur auf hochfrequente Phänomene untersucht, so daß ihre Repräsentativität hinsichtlich dieser Phänomene stieg. Außerdem half hierbei die Textsortenbestimmung, die durch ihre starke Einschränkung auf nur Berichte, Rezensionen und Einleitungen zu einer vergleichbar hohen Anzahl an Texten pro Kategorie führte.

Zur Vermeidung eines Bias hingegen wurden die Computerkorpora parallel zu ihrem Aufbau immer wieder anhand schon in der Anfangsphase erkennbarer Kriterien untersucht; Variationskriterien wie Satzlänge, lexikalische Variation, Frequenz von Passiva u.a. wurden auf die einzelnen Texte, auf die Textsorten, den Sprachstand und die jeweiligen

---

[29] Das Zitat von Knieper stammt aus Knieper, Thomas (Hrsg.) (1993): *Statistik. Eine Einführung für die Kommunikationsberufe*. München.

Momentanversionen der Computerkorpora angewendet, um größere Abweichungen zu erkennen, zu dokumentieren und evtl. zu korrigieren. Bei abweichenden Ergebnissen wurde untersucht, ob andere Variationskriterien dafür hätten verantwortlich sein können.

Ebenfalls wurde darauf geachtet, daß die Texte das Kriterium der Vielfalt erfüllten. Dies war leichter im CMK zu erreichen, da hier aufgrund der Vielfalt der Quellen Texte von verschiedenen Autoren für das Korpus herangezogen werden konnte. Im CLK hingegen und aufgrund der Besonderheiten der Population[30] ergaben sich Wiederholungen, deren Auswirkung allerdings durch die Einschränkung auf einen Text pro Textsorte von jedem Verfasser aufgehoben wurden. In diesem Sinne kann nur die Sammlung weiterer Texte im Laufe der Jahre zu einer größeren Vielfalt führen.

Hinsichtlich der Authentizität der Texte wurde berücksichtigt, daß die Texte in keiner besonderen Weise für z.B. eine Veröffentlichung bearbeitet worden waren. Dabei ist jedoch zu beachten, daß die Texte des CLK einer Revision unterworfen wurden.

Im folgenden werden die Variationskriterien besprochen, die für das CLK und das CMK von Bedeutung waren.

[30] Die begrenzte Zahl der Studenten in den Seminaren führte dazu, daß ein Teil mehrmals im Lernerkorpus repräsentiert ist.

3.1.2.1 Produktionsform

Unter Produktionsform verstehen wir hier die Einteilung in mündliche und schriftliche Textprodukte, ohne Berücksichtigung weiter Gliederungen, die z.B. auf den Kontext der Produktion verweisen könnten. Als Produktionsform wurde nur die schriftliche Textproduktion in Betracht gezogen. Die allgemeinen Probleme des Designs Korpora gesprochener Sprache wurden schon auf S. 40 besprochen. Auf die Möglichkeiten Korpora gesprochener Sprache in Bezug auf die vorliegende Untersuchung hingegen wird in dem letzten Kapitel eingegangen.

3.1.2.2 Texteinteilung

Die Einteilung der Texte nach Textsorten war das Hauptvariationskriterium beider Korpora. Voruntersuchungen anhand der Lernertexte und der Texte der Muttersprachler hatten zwar ergeben, daß auch dem Sprachstand besonderes Gewicht beigemessen werden muß, doch war dies nicht auf unser Computermuttersprachlerkorpus anwendbar und konnte so nicht als primäres Einteilungskriterium verwendet werden.

Dabei unterliegt die Bestimmung der Textsorten einer traditionellen Auffassung, die anhand textexterner Eigenschaften wie z.B. Produktionskontext (akademischer Bereich) oder Kommunikationsziel (Hausarbeit, Bericht usw.) beschrieben werden. Nicht beachtet wurden u.E. präzisere Ansätze, wie z.B. die Berücksichtigung diskursiver Funktionen, da theoretische Erkenntnisse auf diesem Gebiet nur schwer auf

den korpuslinguistischen Bereich anzuwenden sind. In diesem Sinne sprechen wir hier von Text*sorten* und nicht von Text*typen* (vgl. Adamzik 1995, der Differenzierung Isenbergs folgend). Weitere, frühe Ansätze der Textlinguistik wie die Einteilung von Texten nach Wissensgebieten oder nach kommunikativer Funktion (vgl. Fernández-Villanueva 1989) konnten hier ebenfalls nicht mit einbezogen werden.

Die Einteilung nach Textsorten ergab bei den ersten Voruntersuchungen zu unserer Arbeit deutliche Unterschiede bei der Verwendung der Wortarten wie Personalpronomina und anderen Einheiten. Die Ergebnisse stimmen mit Untersuchungen in anderen Sprachen überein: verschiedenen Textsorten liegen allgemeine, wenn auch nicht in jeder Sprache identische Eigenschaften zugrunde[31].

Die Bestimmung und die Auswahl der in die Korpora aufgenommenen Textsorten erfolgte aufgrund der Kriterien Verfügbarkeit[32], Relevanz für die Untersuchungen und Vergleichbarkeit zwischen dem CMK und dem CLK.

---

[31] Vgl. dazu die Untersuchung Tuldavas (1995a: 73-92) zu statistisch-stilistischen Merkmalen Estonischer zeitgenössischer Schriftsteller, aus der klar hervorgeht, daß bedeutende statistische Unterschiede in den Textsorten hinsichtlich verschiedener Sprachen auftreten können, obwohl dazu angemerkt werden muß, daß diese Unterschiede ebenso auf individuellem Textniveau erscheinen. Qualitativ gesehen ist dies höchstwahrscheinlich anders, denn kulturelle und kommunikative Strategien prägen die Struktur der Textsorten.

[32] Vgl. Granger (1998b: 10) zu den Problemen, die die Sammlung und Bearbeitung von Lernertexten aufwirft und 3.1.3 für die Anwendung auf

Verfügbarkeit bezieht sich auf die Möglichkeit, Texte zusammenzutragen, die den festgelegten Kriterien entsprechen; nicht immer sind solche Texte vorhanden oder stehen dem Forscher leicht zur Verfügung.

Relevanz hingegen betrifft die Bedeutung, die eine Textsorte für die Untersuchungen und die Ergebnisse haben kann.

Vergleichbarkeit schließlich ist das Kriterium, anhand dessen garantiert wird, daß zwischen CLK und CMK eine Übereinstimmung der Grundmerkmale der Texte besteht und der Vergleich zwischen nicht analogen Textsorten vermieden wird.

Die Verfügbarkeit stellte eine Einschränkung bezüglich des Kontextes dar, in dem der jeweilige Text verfaßt worden war. Er sollte einem reellen produktiven Kontext entsprechen, d.h. seine textuelle Funktion sollte der Textsorte angepaßt sein, da dies gleichzeitig zur Authentizität des Datenmaterials beitragen würde. Ein produktiver Kontext mit dem Ziel, einzig und allein einen grammatischen Punkt in dem Fremdsprachenseminar zu üben, hätte zu einer Verzerrung der Ergebnisse im CLK geführt oder zumindest wie im ICLE als Variationskriterium berücksichtigt werden müssen. Zudem hätte ein produktiver Kontext die Vergleichbarkeit mit dem CMK eingeschränkt, da diesem keine grammatischen Aufgaben zugrunde lagen. Aus diesem Grund wurden nur Texte zusammengetragen, die

---

unsere Korpora.

den Prozeß der Aneignung von (stilistischer, methodischer und kommunikativer) Kompetenzen in der entsprechenden Textsorte widerspiegelten (Aufgabenorientiertheit).

Ein weiteres Problem der Verfügbarkeit bestand darin, daß die Textsorten des CLK denen der Texte der Muttersprachler entsprechen mußten (siehe unten), gleichzeitig aber auch den Verfassern der Lernertexte bekannt sein sollten. Obwohl für das CMK auch Textsorten wie z.B. Protokolle und verschiedene Arten der Mitschrift zur Verfügung standen[33], schienen diese Textsorten in ihrer deutschsprachigen und deutschkulturellen Form der akademischen Praxis der Autoren der Lernertexte zu fremd; ihre Einbeziehung hätte dazu geführt, daß die Informanten erst hätten lernen müssen, diese Texte zu produzieren, was wiederum zu Verfremdungen geführt hätte, die mehr mit dem Lernprozeß einer neuen Textsorte als mit dem Lernprozeß der Sprache zu tun gehabt hätten.

Die oben genannten Einschränkungen führten dazu, daß für das CLK und das CMK im akademischen Bereich die Textsorten Hausarbeit oder Seminararbeit, Bericht, Rezension, Aufsatz, Erzählung und Nacherzählung in Frage kamen. Diese Textsorten werden jetzt kurz erörtert.

Hausarbeiten werden im Verlaufe des ganzen Studiums geschrieben und entsprechen einer produktiven Situation, die

[33] Vgl. Bünting et. al (1996) hinsichtlich der Definition von Protokoll und Mitschrift.

mit der der Muttersprachler vergleichbar ist: es handelt sich dabei um die Untersuchung eines vorgegebenen Themas mit dem Ziel, Kompetenz auf inhaltlicher und formaler Ebene unter Beweis zu stellen und zudem eine bestimmte Bewertung von seiten des Dozenten zu erhalten.

Berichte, oder auch Erlebnisberichte, können im Rahmen eines Seminars verfaßt werden, v.a. als Übung in Fremdsprachenseminaren, aber auch mit dem nicht strikt akademischen Ziel, Kommilitonen über bestimmte Themen zu informieren.

Rezensionen werden ebenfalls in verschiedenen Seminaren verfaßt, einerseits mit Zielen, die vergleichbar sind mit denen der Hausarbeit, andererseits mit dem Ziel, andere Studenten über bestimmte Aspekte eines literarischen oder wissenschaftlichen Werkes zu informieren.

Aufsätze werden hauptsächlich als Übung zur Textproduktion geschrieben oder erscheinen in Prüfungen als Textproduktion. Es handelt sich dabei um eine Textsorte, die aufgrund einer bestimmten Übungstypologie verfaßt wird und von uns nicht als akademische Textsorte aufgefaßt wird. Sie weist keine strukturelle Einheit auf. Strukturell können Aufsätze an vorgegebene Elemente gebunden sein, d.h. einer strukturellen thematischen Entfaltung folgen, oder nicht.

Erzählungen dienen, wie die Aufsätze, in Sprachseminaren dem mehr oder weniger bestimmten Ziel der Textproduktion. Ohne die

Komponente des Sprachlernens sind sie stilistisch frei und sollen im Lernkontext dem Studenten helfen, eventuelle Schreibschwierigkeiten zu überwinden.

Nacherzählungen werden ebenfalls hauptsächlich in Sprachseminaren produziert und bilden eine Untergruppe der Erzählung. Sie sind stilistisch markiert durch die Vorgabe eines Originals und haben zum Ziel, vorhandene Information zusammenzufassen. Sie bilden die (normalerweise reduzierte) Version des Originals eines muttersprachlichen Verfassers und reproduzieren deshalb viele der sprachlichen Eigenschaften des Vorbilds.

Weitere Textsorten, wie evtl. die *Übersetzung*[34], häufig in der Fehleranalyse verwendet (vgl. Nickel 1972a), wurden nicht hinzugezogen, weil sie mittlerweile entweder nur selten oder gar nicht in Fremdsprachenseminaren erscheinen.

Der zweite wichtige Punkt für die Bestimmung und Auswahl der Textsorten war die *Relevanz*, die dieselben für die quantitativen Untersuchungen haben konnte. Die gewählten Textsorten mußten zu den allgemeinen Zielen der Untersuchung beitragen können, und zwar aufgrund der Auswertbarkeit und Untersuchbarkeit hinsichtlich der quantitativen Analysen. Dies setzte sprachliche Eigenschaften in den Texten voraus, die

[34] Des weiteren wäre die Textsorte *Übersetzung* wenig relevant für die vorliegende Untersuchung, da sie wie die Nacherzählung bei Lernern dazu tendiert, sich stark an das Original anzulehnen.

computationell zu bearbeiten waren, und diese sind nach Biber (1988) leichter in formal stark markierten Textsorten zu finden (wie z.B. in der Rezension) als in denen, die dem Verfasser einen großen stilistischen Spielraum anbieten, wie die Erzählung u.a. Das führte zu einer groben Einteilung der Texte in wissenschaftliche, stilistisch stark markierte Texte und kreative, stilistisch schwach markierte Texte.

Das Korpus sollte jedoch nicht ausschließlich aus wissenschaftlichen Texten bestehen, denn die Einbeziehung von kreativen Texten erlaubte es eben aufgrund ihres stilistischen Spielraums, über eine Kontrollgruppe innerhalb des Korpus zu verfügen: anhand dieser Texte konnten sprachliche, stark abweichende Eigenschaften sowohl korpusintern (d.h. innerhalb eines der Computerkorpora) als auch korporaintern (d.h. zwischen den verschiedenen Computerkorpora) erneut überprüft werden.

Das dritte Kriterium für die Bestimmung der Textsorten ist die *Vergleichbarkeit* mit den Texten der Muttersprachler. Aufgrund der Einschränkung des Kriteriums Verfügbarkeit kamen die Textsorten Rezension, Bericht und Haus- oder Seminararbeit, Erzählung und Aufsatz in Betracht.

Verfügbarkeit	Relevanz	Vergleichbarkeit
Aufsatz		
Bericht		Bericht
Erzählung		Erzählung
Hausarbeit	Hausarbeit	Hausarbeit
Nacherzählung		
Rezension	Rezension	Rezension

TABELLE 4

Aus den verfügbaren Textsorten (vgl. Tabelle 4) wurden die beiden akademischen Textsorten gewählt, die den drei Kriterien entsprachen, d.h. Hausarbeit und Rezension. Diesen beiden sollte eine nicht akademische Textsorte hinzugefügt werden, als Kontrollgruppe. Die Wahl fiel dabei auf den Bericht in Form eines Erlebnisberichtes, der eine Abgrenzung zu den akademischen Textsorten ermöglichte, und eventuell in den beiden anderen Textsorten nicht erscheinende Aspekte abdeckt. Der Bericht ermöglicht im Vergleich zum Aufsatz und zur Erzählung zumindest eine thematische Abgrenzung (siehe weiter unten). Aufsatz, Erzählung und Nacherzählung weisen ihm gegenüber eine hohe stilistische und strukturelle Vielfalt auf, die eine Systematisierung erschwert.

Zusätzlich ist anzumerken, daß, obwohl als Textsorte auch die Hausarbeit aufgenommen wurde, diese nicht in ihrer vollständigen Form zum Teil des Korpus wurde. Die normale Länge einer Hausarbeit hätte zu einer Überrepräsentation sehr langer Texte (bis zu 30 Seiten) im Korpus geführt. Zusätzlich erwies sich die Tatsache als problematisch, daß innerhalb des Hauptteils einer Hausarbeit zahlreiche Zitate erscheinen, die

mit den vorhandenen computationellen Mitteln schwer zu bearbeiten waren und im Falle einer Nicht-Berücksichtigung zu verfälschenden Ergebnissen geführt hätten. Die Möglichkeit, daß lange Texte eine starke innere Variation hinsichtlich von Aspekten wie Argumentation, Exposition usw. aufzeigen können, trug ebenfalls zu dieser Entscheidung bei.

Von der vorläufigen Auswahl der Textsorten ausgehend, wurden Bericht, Einleitungen (zu Hausarbeiten) und Rezension anschließend hinsichtlich ihrer thematischen und stilistischen Eigenschaften näher bestimmt.

Als *Bericht* werden jene Texte definiert, die Auskunft über ein langes Erlebnis eines Studenten geben. In dem CLK und dem CMK wurden sie spezifisch definiert als Berichte entweder über einen Auslandsaufenthalt (Erasmus-Berichte, Berichte über Au-Pair-Tätigkeit) oder über die Situation der eigenen Stadt und/oder Universität. Sie hatten alle informativen Charakter und waren an Studenten gerichtet, die eventuell einen gleichen oder ähnlichen Aufenthalt planten oder sich über eine Stadt und/oder eine Universität informieren wollten[35]. Berichte enthalten nicht alle Eigenschaften von akademischen Texten und dienen hier als Kontrollgruppe, in dem eine formal nicht so stark vorgegebene Sprache angewendet werden kann, die bis hin zur Umgangssprache reicht. Die Länge der Berichte schwankt

[35] Die Authentizität des Materials sowie der kommunikativen Aufgabe wurde im Rahmen von Studentenaustauschprogrammen abgesichert.

zwischen 210 und 2.859 Wörtern, mit einem mittleren Wert von 972 Wörtern.

Einleitungen (zu Hausarbeiten) sind Einführungen zu wissenschaftlichen Arbeiten, die im Rahmen eines Seminars geschrieben wurden. Im Gegensatz zum Bericht wurden sie vom jeweiligen Dozenten bewertet und bilden Teil einer größeren Arbeit, der Seminar- oder Hausarbeit, der sie entnommen wurden. Da lange Texte verschiedene strukturelle Eigenschaften aufweisen können und diese nicht konstant innerhalb des Textes verteilt sind, haben wir aus den uns zur Verfügung stehenden Seminar- und Hausarbeiten die restlichen Teile entfernt. Einleitungen haben in diesem Sinne die Funktion, das im Hauptteil der Arbeit dargestellte Thema einzuleiten und eine kurze problemorientierte Zusammenfassung über den Stand der Forschung und (aus der Sicht des Studenten, der eine Bewertung erwartet) den Kenntnisstand des Verfassers zu geben. In einigen Arbeiten erschien überhaupt keine Einleitung, die höchstens in den Text eingearbeitet war; diese Fälle konnten aber nicht berücksichtigt werden und wurden dementsprechend nicht in das Korpus aufgenommen. Die Länge der Einleitungen schwankt zwischen 151 und 1669 Wörtern, mit einem Mittelwert von 375 Wörtern.

Bei der Textsorte *Rezension* handelt es sich um Zusammenfassungen von literarischen oder wissenschaftlichen Werken; der Verfasser einer Rezension will mit seinem Text dem Leser eine Übersicht über das rezensierte Werk geben, in der

je nach Eigenschaften (literarisch oder wissenschaftlich) verschiedene Punkte berücksichtigt werden, wie Inhalt, Struktur, Thema, Methode usw.

3.1.2.3 Der Sprachstand

Hinsichtlich des CLK erwies sich zusätzlich das Variationskriterium des Sprachstand als bedeutend. Wie in der späteren Auswertung beider Korpora zu sehen sein wird, folgen viele Abweichungen einer quantitativ steigenden Tendenz von niedrigem Lernstadium bis hin zum Text eines Muttersprachlers (vgl. 5.1.4). Im CLK wird das Kriterium anhand der Unterscheidung zwischen zwei Sprachstandebenen aufgenommen, Fortgeschrittene und stark Fortgeschrittene[36].

Weitere Sprachstandebenen wurden nicht berücksichtigt. Die Seminare des Niveaus Mittelstufe I oder darunter erzeugten zu wenige Texte, um statistisch analysierbar zu sein, und aufgrund der noch fehlenden sprachlichen Kenntnisse lag die Vermutung nahe, daß die Auswertung zu stark von Faktoren der Reduzierung (auf lexikalisch, grammatisch oder semantisch einfache Strukturen) betroffen gewesen wäre. Seminare zwischen den gewählten Sprachstandebenen Mittelstufe II und Oberstufe wurden anfangs in einige Voruntersuchungen mit einbezogen, wiesen aber konstant Mittelwerte zwischen den beiden anderen

[36] Fortgeschrittene entsprechen dem Niveau Mittelstufe II, stark Fortgeschrittene dem Niveau Oberstufe.

Sprachstandebenen auf, so daß zur Darstellung der Tendenzen hinsichtlich der Sprachstandebene die zwei Extreme genügten. Dadurch konnte eine Vergrößerung des Lernerkorpus vermieden werden, die sich negativ auf die weitere Bearbeitung ausgewirkt hätte.

3.1.2.4 Textursprung

Als Textursprung fassen wir hier den geographischen Ort auf, in dem der Informant seinen Text produziert hat, ohne jedoch weiter auf den geographischen Ursprung des Verfasser selbst einzugehen. CLK und CMK weisen bedeutende Unterschiede hinsichtlich ihres Ursprungs auf. Gemeinsam haben sie den Zeitraum der Entstehung der Texte, zwischen 1995 und 1997. Geographisch aber liegt dem CLK eine stark eingeschränkte Auswahl zugrunde. Alle Verfasser lebten zum Zeitpunkt der Textproduktion in Katalonien, mehrheitlich in Barcelona, und waren an der Universität Barcelona im Fachbereich Germanistik eingeschrieben. Die Verfasser der Texte des CMK hingegen wurden so ausgewählt, daß das CMK eine minimale geographisch Repräsentativität für Deutschland darstellen kann, allerdings mit der Einschränkung, daß es sich dabei nicht um ein primäres Kriterium handelte. Verfasser anderer Nationalitäten wurden nicht aufgenommen.

Für die Auswahl des CMK wurden Texte von verschiedenen Universitäten zusammengetragen: Berlin, München, Düsseldorf und Hamburg und eine reduzierter Auswahl von geographisch

weiter verteilten Universitäten. Eine größere Repräsentativität erfordert mehr geographische Punkte und die Überprüfung, ob der Verfasser auch einheimisch ist an der Universität, für die er den Text produziert hat.

3.1.2.5 Verfasser

Als mögliche Variationskriterien zum Verfasser wurden in 2.1.2.4 Population und Idiolekt dargestellt. Mit dem Ziel einer verfasserunabhängigen Darstellung des Materials wurde das Idiolekt aus dem Design ausgeschlossen und eine zu große Repräsentation einzelner Verfasser gezielt vermieden. Das Kriterium Population hingegen, die besonderen Eigenschaften der Verfasser, wurde definiert und beim Design aufgrund seines Einflusses auf die Textproduktion berücksichtigt.

In das CMK wurde jeweils nur ein Text von einem Verfasser aufgenommen. Dadurch konnte die Überrepräsentation einzelner Verfasser vermieden werden, die als Idiolektale Eigenschaften Einfluß auf die Auswertung des Datenmaterials hätte haben können. Aufgrund der reduzierten Auswahlmöglichkeiten bei den Lernern hingegen war dieses Verfahren nicht anzuwenden, weshalb bestimmt wurde, daß jeder Verfasser nur einmal pro Textsorte vertreten sein durfte, aber mehrmals pro Sprachstandebene.

In bezug auf die Population wurden Alter, Studiengang und Muttersprache als zu berücksichtigende Eigenschaften bestimmt. Zudem wurden andere angegeben, die für spätere Untersuchungen

von Nutzen sein können, wie Geschlecht und Staatsangehörigkeit/Herkunft, und zudem Einschränkungen festgelegt, die Einfluß auf die oben genannten Eigenschaften haben konnten.

Das Alter der Verfasser beider Korpora wurde auf zwischen 18 und 35 Jahren festgelegt. Dadurch sollte z.B. vor allem im CLK die oben genannte Abweichung vermieden werden (vgl. S. 40), daß ältere Studenten, die möglicherweise über mehr lexikalische, aber weniger grammatische Kenntnisse verfügen, die Auswertung in diesem Sinne beeinflussen konnten[37].

Als Studiengang für die Textsorten Einleitung und Rezension wurde germanistische Sprach- und Literaturwissenschaft bestimmt. Auf diese Weise konnten die Textsorten leichter abgegrenzt werden und größere stilistische fachsprachlich bedingte Unterschiede, wie sie z.B. im Vergleich zur Textproduktion aus dem Jurabereich entstehen können, vermieden. Diese Bestimmung wurde für Berichte aufgehoben, da dieser stilistische Einfluß hier nicht zu vermuten war. Im Falle des CLK war eine weitere Einschränkung, daß es sich bei den Verfassern um Studenten der Universität Barcelona handeln

[37] Die Untersuchung der Textproduktion älterer Studenten in der Designphase hatte ergeben, daß mehrere von ihnen komplexe und gehobene lexikalische Elemente anwendeten, wobei das überdurchschnittlich gehobene lexikalische Niveau allerdings mit unterdurchschnittlichen grammatischen Kenntnissen einher ging. Dies war anscheinend mit vorherigen Studien dieser Studenten verbunden und wäre in Zukunft zur Absicherung der Repräsentativität näher zu untersuchen.

mußte, die in den entsprechenden Seminaren des Fachbereiches Germanistik eingeschrieben waren.

Die Muttersprache der Verfasser des CMK mußte Deutsch sein[38], die der Verfasser des CLK Spanisch oder Katalanisch. Hier wurde eine Einschränkung auferlegt: Texte von Halbmuttersprachlern, ehemaligen Schülern einer sich in Spanien befindlichen deutschsprachigen Schule oder Studenten, die mehr als ein Jahr in letzter Zeit oder mehrere Jahre in der Vergangenheit in einem deutschsprachigen Land verbracht hatten, wurden nicht in das CLK aufgenommen. Auf diese Weise sollten die Einflüsse eines nicht mit den anderen Verfassern vergleichbaren Spracherwerbs vermieden werden, die in diesen Ausnahmefällen zu besonderen Eigenschaften in der Textproduktion führen können.

Wenn verfügbar, wurde weitere Information zu den Verfassern angegeben, jedoch nicht für die Auswahl berücksichtigt; diese Angaben hatten in den Voruntersuchungen keine relevanten Eigenschaften aufgewiesen oder waren statistisch nicht anwendbar. Hinsichtlich des Geschlechts z.B. waren im CLK einerseits nur wenige männliche Verfasser vertreten; eine nähere Untersuchung in dieser Richtung hatte jedoch keine relevanten Abweichungen ergeben. Die Information wurde dem

[38] Dies wurde durch die entsprechende Nachfrage beim Informanten oder dem betreuenden Dozenten überprüft.

jeweiligen Text jedoch beigegeben und kann für spätere Untersuchungen herangezogen werden.

3.1.2.6 Kommunikationsform

Das Kommunikationsziel der Texte beider Korpora liegt zwischen dem Zweck akademischer Texte, Wissen mitzuteilen, und dem Zweck von Lernertexten, eine Leistungsmessung von Seiten des Dozenten zu erhalten. Schreiben wird in diesem Sinne als Produktion sprachlicher Äußerungen aufgefaßt, die in die „allgemeine Handlungsdynamik des Menschen eingegliedert ist; Menschen schreiben, weil sie bestimmte Ziele in bestimmten Situationen auf eine bestimmte Weise erreichen wollen" (Grabowski 1995: 12). Die Texte beider Korpora erfüllten dementsprechend nicht nur den Zweck, Wissen mitzuteilen, sondern sie sind gleichzeitig für eine Bewertung bestimmt. Dadurch unterscheiden sie sich von rein akademischen Texten, die ausschließlich Wissen mitteilen, und deshalb sind unsere beiden Korpora nicht mit veröffentlichten akademischen Texten vergleichbar, wo Vermeidungsstrategien, wenn überhaupt angewandt, einer ganz anderen Natur sind.

3.1.3 Korpusumfang

Als Umfang jedes der effektiv behandelten Computerkorpora wurden ugf. 40.000 Wörter festgelegt. Eine präzise mathematische Bestimmung ist nicht zu leisten für ein Korpus, das auf frequentiell stark abweichende Phänomene untersucht

werden soll (Präpositionen erreichen eine Frequenz von ugf. 10‰, Adverbien jedoch nur 0,5‰). Voruntersuchungen anhand des CLK und des CMK zu verschiedenen Wortarten hatten ergeben, daß mit 40.000 Wörtern auch relativ wenig frequente Phänomene in die Untersuchungen einbezogen werden konnten.

Aufgrund statistischer Verfahren besteht die Notwendigkeit, für jedes angewendete Variationskriterium eine bestimmte Zahl von Texten bereitzustellen (vgl. Biber 1988), die proportional dem Gewicht des Variationskriteriums hinsichtlich des ganzen Korpus entspricht. Bei gleicher Proportion aller Variationskriterien wäre die Zahl der benötigten Texte zu berechnen, indem die Summe der Variationskriterien mit einer gleichbleibenden Zahl an Texten multipliziert wird; bei ungleicher Proportion wäre für jedes Variationskriterium zunächst die Zahl der Texte jedes einzelnen Variationskriteriums zu bestimmen und anschließend die Gesamtzahl der Texte zu addieren.

Im Falle des CLK sind 5 Variationskriterien vorhanden die drei genannten Textsorten (Bericht, Einleitung und Rezension; 3.1.2.2) und die ebenfalls schon besprochenen zwei Sprachstandebenen (Mittelstufe und Oberstufe, vgl. 3.1.2.3). Das Kriterium Sprachstand ist nicht auf das CMK anzuwenden, weshalb dieses nur über die drei Variationskriterien der Textsorte verfügt. Die Zahl der Variationskriterien wurde bewußt so niedrig wie möglich gehalten, ohne größere

Beeinträchtigung der Repräsentativität, um den Umfang des Korpus nicht wesentlich zu steigern.

Die nähere Bestimmung des notwendigen Korpusumfangs ist bei abweichenden Frequenzen der Phänomene extrem schwierig und hängt zudem von weiteren Faktoren ab. Statistisch kann der Umfang annähernd durch die Bestimmung des für ein Phänomen notwendigen Umfangs erfolgen (Granger 1998b: 11).

Ein Phänomen mit einer Frequenz von beispielsweise 10‰ erscheint statistisch gesehen einmal alle hundert Wörter; damit dieses Phänomen in jeder Textsorte des CLK erscheint, das nach 5 Variationskriterien eingeteilt wurde, würden dementsprechend 5 Texte à mindestens 100 Wörter benötigt. Für eine Auswertung wäre die Zahl der empirischen Belege des Phänomens zu niedrig (vgl. Meijs 1992: 146f), so daß in Abhängigkeit von der Frequenz des Phänomens die Zahl der verfügbaren Texte erhöht werden muß und dementsprechend mehr Wörter zur Auswahl stehen. Bei 1000 Wörtern in jedem der 5 Subkorpora (die nach Variationskriterien definiert werden) dürften schon 10 Treffer in jedem Subkorpus und 50 im ganzen Korpus erhalten werden; je nach Komplexität des Phänomens, also ob es verschiedene Anwendungen findet (homonymische Präpositionen mit semantischer Differenzierung beispielsweise oder die Untersuchung der Stellung einer Wortart im Satz) oder nicht, könnte dies als ausreichend angesehen werden.

Die Mehrzahl der untersuchten Phänomene lag jedoch bei einer Frequenz unter 10‰, mit einem Großteil leicht über 1‰. Aus

diesem Grund wurde bestimmt, daß für Frequenzen über 1‰ und mit variierender Komplexität der Phänomene ein Korpus von 40.000 Wörtern ausreichen müßte, um die Phänomene darzustellen und ihre Verteilung zu untersuchen. Phänomene mit einer Frequenz von 1‰ bei einem Korpusumfang von 40.000 Wörtern würden statistisch gesehen 40 Belege erzielen. Eine eingehende Untersuchung jedes Phänomens müßte zunächst auf der Grundlage der provisorisch ermittelten Frequenzen und in Abhängigkeit von der an einem Phänomen zu untersuchenden Eigenschaften den notwendigen Korpusumfang erneut festlegen, so daß dieser gegebenenfalls erweitert würde. Dies bedeutet, daß ein für die Untersuchung eines wenig frequenten Phänomens ein größeres Korpus benötigen würde.

3.1.4 Abgeschlossenheit

Beide Korpora waren in der für diese Untersuchung verwendeten Version abgeschlossen, d.h. während der Darstellung der Phänomene wurden keine weiteren Texte hinzugefügt. Demzufolge erfolgen alle Angaben und Berechnungen auf der gleichen Grundlage.

Ein Korpus muß für die Untersuchung eines Phänomens immer einen festen, unveränderlichen Umfang haben. Ansonsten könnten keine Ergebnisse hinsichtlich eines Phänomens erzielt werden. Auch bei der Untersuchung mehrerer Phänomene ist ein abgeschlossenes Korpus angebracht, da dadurch die quantitative Beschreibung verschiedener Phänomene miteinander vergleichbar

wird, und Interferenzen vermieden werden, die aus der Benutzung verschiedener Versionen entstehen.

Die Texte beider Korpora wurden über drei Jahre (1995-1997) gesammelt und anschließend anhand vorläufiger Designkriterien strukturiert. Aus diesem Prozeß entstand eine erste Vorversion der Korpora, an der Voruntersuchung vorgenommen wurden. Die Ergebnisse führten zu einer Revision der Kriterien zur Zusammenstellung der endgültigen Korpora, die aufbereitet und mit Auszeichnungen versehen wurden.

3.1.5 Maschinelle Lesbarkeit

Wie angegeben (vgl. 2.1.5) ist die maschinelle Lesbarkeit Voraussetzung für ein Computerkorpus, da dadurch Teile der Bearbeitung und der Auswertung der Texte mit einem Computer durchgeführt werden können. Der größere Arbeitsaufwand, der für die Erstellung eines Computerkorpus im Vergleich zu einem traditionellen Korpus oder einer Textsammlung notwendig ist, wird darüber hinaus durch schnellere und präzisere Auswertungsverfahren aufgehoben. Ein weiterer Grund für die Verwendung maschinell lesbarer Korpora ist die Möglichkeit, die Texte in automatisierter Form mit Zusatzinformation, wie POS-Tags, zu versehen (vgl. 3.2.2.).

Nach der Eingabe der Texte in den Computer[39] wurden sie aufbereitet, d.h. einer Vereinheitlichung unterworfen, die der

[39] Die Texte des Lernerkorpus wurden von den Informanten größtenteils auf

Kompatibilität der Texte diente und sogenannten Datenmüll entfernen sollte. Dabei wurden drei verschiedene Prozesse durchgeführt: Vereinheitlichung, Streichung und Markierung.

Die *Vereinheitlichung* diente der maschinellen Bearbeitbarkeit und sollte gleichzeitig Berechnungsfehler verhindern. Folgende Elemente wurden vereinheitlicht:

- Unterschiedliche Schrifttypen und kursiv- oder fettgedruckte Formate wurden gestrichen oder in Anführungszeichen gesetzt.
- Die Silbentrennung wurde abgeschaltet und schon vorhandene Trennungen eliminiert.
- Überflüssige Leerzeilen und Tabulatoren wurden gestrichen.
- Abweichende Schreibweisen wurden vereinheitlicht: *ss* wurde nach den Regeln der alten Rechtschreibung zu *ß*[40], Abkürzungen wurden graphisch gleich gestaltet[41].

Streichungen erfolgten mit dem Zweck der Reduzierung der Texte auf das tatsächlich vom Autor verfaßte Textmaterial und der

---

Diskette abgegeben. Proportional zu den handschriftlichen Texten wurden einige Texte manuell eingegeben, um evtl. auf das Medium zurückführbare Abweichungen ebenfalls zu berücksichtigen. Die Texte des Muttersprachlerkorpus hingegen lagen allesamt in elektronischer Form vor.

[40] Das *ss* ist vielleicht eines der klarsten Beispiele für die Notwendigkeit von Vereinheitlichungen. Unterschiedliche Schreibweisen identischer Formen (*dass* oder *muss* anstatt *daß* und *muß*) hätten bei der statistischen Auswertung zu Fehlergebnissen geführt (was bei ersten Voruntersuchungen auch geschah, weil *dass* nicht als Konjunktion erkannt wurde) und zu Fehlern beim automatisierten POS-Tagging.

[41] So wurden Formen wie *"z.b."*, *"Z.B."*, *"z. B."*, *"z B"* usw. zu *"z.B."*.

Minimierung von Fehlerquellen. Gestrichen wurden folgende Elemente:

- Fußnoten
- Verweise auf Fußnoten
- Graphiken und Legenden
- Weitere Elemente, die nicht dem laufenden Text angehören, wie Titelblatt, Inhaltsverzeichnis, bibliographische Angaben, Bibliographie usw.

Durch die Markierung schließlich sollten Elemente gekennzeichnet werden, die bei der Auswertung übergangen werden sollten, bei einer manuellen Revision aber für das Textverständnis notwendig waren. Dazu gehörten:

- Titel und Überschriften, die aufgrund ihres unvollständigen Satzcharakters nicht in die Berechnung einbezogen wurden[42].
- Zitate, die, als Fremdmaterial, nicht vom Verfasser des Textes stammten, aber für sein Verständnis notwendig sind[43].

---

[42] Sie hätten negativen Einfluß auf die allgemeine Frequenz der Verben in den Korpora gehabt, da in Überschriften im allgemeinen keine Verben erscheinen.

[43] Ihre Streichung bereitete jedoch Probleme, die zur Ausschließung der Texte führte, in denen sie stark vertreten waren. Zitate erschienen in den Texten in zwei Formen: unabhängig und in den Text eingebunden. In Ihrer unabhängigen Form konnten die Eliminierung ohne Einfluß auf die Untersuchungen erfolgen, weil keine argumentativen Strukturen berücksichtigt wurden. Die in den Text eingebundene Form führte jedoch bei einer Streichung zu syntaktisch unvollständigen Sätzen. Aus diesem Grund wurde nicht nur das Zitat, sondern auch der einbettend Satz gestrichen, mit der Einschränkung, daß alle Texte aus den Korpora ausgeschlossen wurden,

3.2 Textbezogene Aspekte

Hinsichtlich der textbezogenen Aspekte waren für das Design der beiden Korpora die Auswahl und die Aufbereitung anhand von Auszeichnungen entscheidend. Dabei weichen die Auswahlverfahren zum Teil zwischen beiden Korpora voneinander ab, da die repräsentierte Grundgesamtheit nicht direkt auf der gleichen Ebene liegen. Hinsichtlich der Auszeichnungen waren aus methodischer Sicht die Ziele der Untersuchung zu berücksichtigen, doch war auch dem technischen Aspekt der vorhandenen Mittel zur Durchführung der Auszeichnung Beachtung zu schenken.

3.2.1 Auswahl

Durch statistische Verfahren wie die Verallgemeinerbarkeit, die Bestimmung der Textlänge und die Methode, die zur Auswahl der Texte angewendet wird, wird die Repräsentativität erreicht. Die Repräsentativität spielt sich bei Textkorpora auf zwei Ebenen ab: der des Korpus insgesamt, das eine Widerspiegelung der Grundgesamtheit anstrebt, und auf der Ebene der Texte, die Teil dieser Grundgesamtheit sein müssen. D.h., ein Text, der nicht die Aspekte erfüllt, die das Korpus darstellt, wie in unserem Fall die akademische Textproduktion, sondern einer außerhalb dieser Gruppe angesiedelten Textsorte angehört, wie dem literarischen Text, kann nicht in das Korpus

---

bei denen Zitate mehr als 20% des Gesamtumfangs ausmachten.

aufgenommen werden. In den folgenden Abschnitten sollen die Kriterien Verallgemeinerbarkeit, Textlänge und Auswahlmethode dargestellt werden, die eine Abgrenzung der verwendbaren Texte ermöglichen.

3.2.1.1 Verallgemeinerbarkeit

Wie in 2.2.1.1 dargestellt, können Einzeltexte nicht als Stichprobe aufgefaßt werden, denn sie müssen Eigenschaften aufweisen, anhand der sie verallgemeinerbar sind, also als *Teil* der Stichprobe angesehen werden können. Dies wurde durch die Analyse der in den Voruntersuchungen als relevant bestimmten Phänomene überprüft. Dabei handelt es sich um ein mit der Profilierung vergleichbaren Prozeß, anhand dessen auf zunächst subjektiver Basis sprachliche Phänomene bestimmt werden, deren Frequenzen als Hinweis gedeutet wird, daß ein Text einer bestimmten Gruppe angehört (wie z.B. Nominalisierungen in technischen Texten), um dann anschließend aufgrund der ermittelten Frequenzen empirisch wechselseitige Beeinflussungen festzustellen, und diese einer Gruppe zuzuschreiben (Crystal 1991: 227).

Diese Überprüfung ist unabdingbar, um die Einbeziehung von textuellem Material zu vermeiden, das nicht der Grundgesamtheit entspricht, die das Korpus darstellen soll. Eine ungenügende Kontrolle kann zu erheblichen Verzerrungen führen. So wurde beispielsweise bei Voruntersuchungen der Preliminarversionen der Korpora die Hypothese untersucht, daß

die lexikalische Variation ein relevantes Phänomen ist. Aufgrund der Verteilung der Werte des CMK wurden im CLK näher jene Texte untersucht, deren Werte durchschnittlich über denen des CMK lagen. Als Ergebnis konnte festgehalten werden, daß in der betreffenden Preliminarversion des CLK Texte mit zahlreichen Zitaten vertreten waren, die vom Verfasser nicht als solche angegeben worden waren. Dies führte zur Aufnahme der lexikalischen Variation als relevantes Phänomen und zur Ausschließung der betreffenden Texte, die dementsprechend als nicht verallgemeinerbar oder authentisch im Sinne der Ziele des Korpus einzustufen waren.

3.2.1.2 Textlänge

Textlänge stellt einen weiteren Faktor bei der Erstellung eines Korpus dar. Aus statistischen, mittlerweile nicht weiter angewendeten Gründen (vgl. EAGLES 1996b: 20), verwendete das Brown Korpus Fragmente einer vorbestimmten Länge (Francis/Kucera 1979), wobei durch die Dekontextualisierung strukturelle Eigenschaften der Texte verloren gingen. Um die in Lernertexten vertretenen sprachlichen Phänomene zu berücksichtigen, die in Abhängigkeit der strukturellen Eigenschaften der Texte an einer bestimmten Position derselben erscheinen, wurden in unsere Korpora nur ganze Textpassagen laufenden Texts aufgenommen, mit den in 3.1.5 genannten Einschränkungen (Vereinheitlichung, Streichung und Markierung); dieses Kriterium stimmt mit dem des *International*

Corpus of Learner English (ICLE) überein (vgl. Granger 1998b: 10).

Aufgrund ihrer hohen Frequenz erscheinen viele Phänomene von der Verteilung her mit einer großen Regelmäßigkeit in einem Text, wie z.B. die Artikel; andere hingegen lassen nur schwer solche Regelmäßigkeiten erkennen. Die in Tabelle 5 dargestellte Verteilung von Wörtern à 15 Zeichen in den Berichten des CMK mit mehr als 1400 Wörtern folgt keinem einheitlichen, klar erkennbaren Schema[44]. Dabei stellt jede durchnumerierte Zeile in ihrer Länge eine Datei dar; jeder Strich steht für ein 15-Zeichen-Wort, so daß anhand des Abstands und der Position die Verteilung abgelesen werden kann (*Plot*).

Verteilung von 15-Zeichen-Wörtern in Berichten mit mehr als 1400 Wörtern des CMK

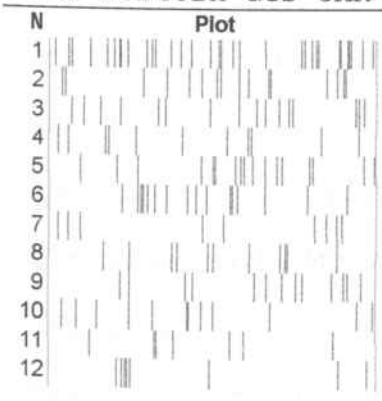


TABELLE 5

[44] Die Begrenzung auf Dateien über 1400 Wörtern erfolgte, um nur Texte darzustellen, die das Phänomen in angebrachter Menge enthielten.

Ebenso unsystematisierbar scheint das Negationselement *nicht* in Berichten des CMK mit mehr als 2000 Wörtern zu sein (vgl. Tabelle 6)[45].

Verteilung der Negation *nicht* in Berichten mit mehr als 2000 Wörtern des CMK

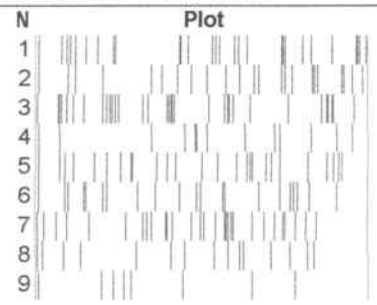


TABELLE 6

Andere Phänomene weisen komplexere Verteilungsschemata auf. Sie können unsystematisch in einer Textsorte erscheinen, aber mit einer hohen Systematizität in einer anderen. Das Personalpronomen *ich* ist ein Beispiel dafür. In den formal relativ freien Berichten des CMK ist keine klare Systematizität zu erkennen, und nur aufgrund von zwei der ausgewählten Dateien (unter den Berichten des CMK mit mehr als 1400 Wörtern) könnte eine Tendenz zur Anwendung gegen Ende des Textes erkannt werden (vgl. Tabelle 7, Zeilen 11 und 12), was wohl auf eine persönliche Schlußfolgerung hindeutet.

[45] Idem.

Verteilung des Personalpronomens *ich* in Berichten mit mehr als 1400 Wörtern des CMK

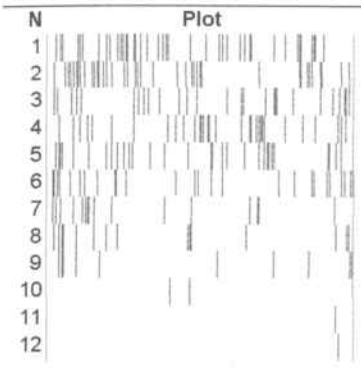


TABELLE 7

Das gleiche Phänomen in der Rezension ergibt ein sehr systematisches distributives Schema: das Personalpronomen *ich* erscheint, wenn es verwendet wird, gegen Anfang oder Ende des Textes, was klar aus dem Plot der Tabelle 8 hervorgeht.

Verteilung des Personalpronomens *ich* in allen Rezensionen des Muttersprachlerkorpus

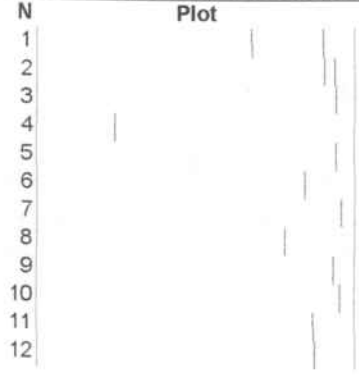


TABELLE 8

In diesem Sinne könnte das Personalpronomen *ich* als textsortenabhängiges Phänomen gedeutet werden, das stilistisch

in einigen von ihnen, in diesem Fall in der Rezension, zur Anwendung gegen Anfang und gegen Ende des Textes tendiert; als deutende Hypothese könnte hierfür angegeben werden, daß sein Erscheinen am Anfang einer Rezension auf die Begleitumstände verweist, in denen der Verfasser der Rezension das besprochene Buch gelesen hat, und daß seine Verwendung gegen Ende des Textes Hinweise auf eine persönliche Meinung des Verfassers ist.

Die Verbindung zwischen strukturellen Eigenschaften, wie die Verwendung von Personalpronomina gegen Ende der Rezension, im Gegensatz zur distributiv nicht klaren Verwendung in Berichten, schlägt sich auch in anderen Phänomenen nieder.

Die oft regelmäßige Verteilung von Phänomenen war dementsprechend zu berücksichtigen, indem nur ganze Texte in die Korpora aufgenommen wurden. Doch ein weiterer Grund, um die Texte ungekürzt aufzunehmen, war die Wiederverwendbarkeit des Korpus, dessen Untersuchung sich in dieser Arbeit zwar auf hauptsächlich grammatische Phänomene bezieht, aber in Zukunft auch für andere Analysen, wie die Kohäsion und Kohärenz oder thematische Progression herangezogen werden soll[46].

[46] Vgl. auch Granger 1998b: 10 und die Struktur des International *Corpus of Learner English* ICLE.

3.2.1.3 Auswahlmethode

Der dritte Faktor, der die Auswahl der Texte bedingt, ist die Auswahlmethode. Während Verallgemeinerbarkeit bestimmte, ob ein Text einer vorgegebenen Grundgesamtheit angehört und somit nicht alle konformen Texte auszusondern hilft, und Textlänge bestimmte, ob die verbleibenden Texte auszugsweise (fragmentarisch) oder vollständig aufgenommen werden, dient die Auswahlmethode dazu, den Mechanismus zu bestimmen, anhand dessen aus den vorhandenen Texten jene gewählt werden, die Teil des Korpus bilden werden. Für den Prozeß der Auswahl, das *Sampling*, d.h. die Zusammenstellung repräsentativer Texte, wurde die Stichprobentechnik der Quotenauswahl[47] gewählt, die zu einer repräsentativen Teilerhebung der Grundgesamtheit führt. Da die Unterschiede hinsichtlich der Textsorten in den Voruntersuchungen klar zum Vorschein getreten waren, wurde die Zufallsstichprobe verworfen, weil damit alle Texte, sei es nun des Niveaus Mittelstufe oder Oberstufe, eine einzige Gruppe gebildet hätten. Die Quotenauswahl hingegen bot hier die Möglichkeit, die schon erkannten Unterschiede der Sprachstandebene besser zu berücksichtigen (vgl. Butler 1985: 2f.). Für jede Textsorte wurde eine Anzahl an Texten vorgegeben (vgl. 3.1.3); die Auswahl der Einzeltexte erfolgte aleatorisch unter Berücksichtigung der Variationskriterien.

[47] Die Quotenauswahl oder stratifizierte Auswahl entspricht der strukturierten Variation: eine Grundgesamtheit wird in Gruppen eingeteilt (Quoten), und für jede Gruppe eine bestimmte Zahl von Texten ausgewählt (vgl. Schlobinski 1996: 26).

Dabei ist anzumerken, daß im Lernerkorpus die Textsorte *Einleitung* für die Sprachstandebene Mittelstufe dieses Minimum nicht erreichte, da auf dieser Ebene der sprachlichen Ausbildung während der betreffenden Jahre nur wenige Informanten eine Hausarbeit geschrieben hatten. Diese Einschränkung wurde in der Auswertung des Datenmaterials in den betreffenden Fällen angemerkt.

3.2.2 Auszeichnungen

Für die Entscheidung zur Aufbereitung der Korpora mit Auszeichnungen waren verschiedene Voruntersuchungen[48] entscheidend, die ergeben hatten, daß die mit den Zielen dieser Untersuchungen verbundenen Phänomene ohne Auszeichnungen nicht zu entdecken gewesen wären. Die Voruntersuchungen, durchgeführt mit dem Konkordanzprogramm WordSmith (Scott 1998), dienten zur Überprüfung der Möglichkeit, ob anhand allgemeiner statistischer Verfahren (Wort-, Satz und Absatzlänge, lexikalische Variation usw.) und der Analyse der Frequenzen auf rein orthographischer Wortebene genug Phänomene gefunden werden könnten, die den Zielen dieser Untersuchung entsprachen.

Die Voruntersuchungen waren jedoch mit verschiedenen Problemen verbunden, die zur Notwendigkeit einer zusätzlichen Auszeichnung führten. Auf der Ebene der allgemeinen

[46] Oder *Pretest*, vgl. Schlobinski (1996: 27).

statistischen Verfahren erwiesen sich Satzlänge, Absatzlänge und lexikalische Variation als problematisch.

Konkordanzprogramme wie WordSmith bearbeiten Satzlänge und Absatzlänge aufgrund der Markierung der entsprechenden Satzendmarker (wie Punkt, Fragezeichen oder Ausrufezeichen) und Absatzmarker (Zeilenumbruch). Diese Markierungsmerkmale können in einigen Fällen auch anderen Eigenschaften entsprechen, zum Beispiel einer Abkürzung („z.B.", „ugf."). Die notwendige Disambiguierung der Satzendmarker und Absatzmarker erfolgte mit Hilfe eines in den schließlich ausgewählten POS-Tagger eingebetteten Unix-Scripts.

Bei der Bestimmung der Absätze sollten zusätzlich die Titel aus der Auswertung ausgeschlossen werden. Sie sind nicht Teil der rein argumentativen Struktur eines Textes und weisen wie bereits erwähnt deutliche Unterschiede zu den syntaktischen Eigenschaften des laufenden Textes auf, wie z.B. keine Präsenz von Verben, weshalb sie zu Verzerrungen der Ergebnisse geführt hätten.

Die lexikalische Variation konnte aufgrund des unbearbeiteten oder rohen Computerkorpus erfolgen, d.h. eines Korpus ohne Auszeichnungen, doch hätten Phänomene wie die Genitivattribution, die im CMK stärker vertreten ist, oder differierende Frequenzen von anderen Flexionsformen zu Abweichungen in den Ergebnissen geführt. Die Anwendung eines lemmatisierten Computerkorpus, d.h. eines Korpus, das statt des normalen Textes alle Wörter auf Grundeinheiten reduziert

darstellt (Bußmann 1990: 445), konnte dazu beitragen, eine präzisere Vorstellung der lexikalischen Kenntnisse der Informanten zu erhalten, unabhängig von ihren syntaktischen Kenntnissen. Eine lemmatisierte Version der Computerkorpora konnte jedoch aus technischen Gründen nur zusammen mit einer ausgezeichneten Version der Korpora erzeugt werden, da zur Bestimmung des Lemmas die Kenntnis der Wortart notwendig ist.

Die lemmatisierte Version ermöglichte jedoch aus technischen Gründen keine Auflösung der Homographen. Dazu wäre eine semantische Disambiguierung notwendig gewesen, und eine weitere Auszeichnungsebene, die eine Unterscheidung von Homographen ermöglichte. Da Konkordanzprogramme die Listen aufgrund des Vergleichs von Zeichenketten und nicht des semantischen Inhalts eines Wortes erstellen, unterscheiden sie auch nicht zwischen den verschiedenen Bedeutungen, die ein und die gleiche Zeichenkette haben kann (z.B. „Bank" zum Sitzen und Kreditinstitut).

Eine weitere Möglichkeit zur Untersuchung eventuell relevanter Phänomene anhand nicht aufbereiteter Korpora liegt in ihrer orthographische Erkennbarkeit (Leech 1997a: 4). Dabei werden den Phänomenen, hauptsächlich Wortarten, orthographische Formen zugeschrieben, die in dem Korpus gesucht werden. Dies ist jedoch nur bei Wortarten möglich, deren Elemente eine finite und invariable Liste bilden. So kann eine Liste der Konjunktionen erstellt werden, für die eine Konkordanzliste erstellt wird, aus der dann manuell alle falschen

Konjunktionen (gleichlautende Adverbien, Partikeln usw.) gestrichen werden. Bei offenen Wortarten wie den Verben und Substantiven erweist sich diese relativ umständliche Vorgehensweise als unangebracht.

Die Untersuchungen und Vergleiche von Phänomenen, die mit Methoden durchzuführen sind, die nicht auf Auszeichnungen zurückgreifen, sind unpräzise und vom Arbeitsaufwand wenig ökonomisch. Dementsprechend wurden die Texte einer Auszeichnung unterworfen, was gleichzeitig positive Auswirkungen auf die Wiederverwendbarkeit des Korpus hatte (Leech 1997a: 5). Trotzdem sei darauf verwiesen, daß jede Art der Auszeichnung mit einem theoretischen Auszeichnungssystem verbunden ist, das als solches nicht perfekt sein kann. So bauen POS-Tagging-Systeme auf einer Wortartenklassifizierung auf, die wie schon angemerkt auf theoretisch nicht einheitlichen Aspekten beruht (vgl. 2.2.2.2).

3.2.2.1 Auszeichnungsformat

Das Format der Auszeichnungen oder Zusatzinformationen, d.h. die physische Form, in der sie den Texten beigegeben werden, wird von methodischen Grundlagen einerseits und der Bearbeitbarkeit des eigentlichen Formats andererseits bestimmt. Wir beachten in dem Computerlerner- und dem Computermuttersprachlerkorpus die in Leechs 7 Leitlinien

zusammengefaßten methodischen Grundlagen[49] und übersetzen sie in ein Format, das sich mittlerweile zu einem *de facto* Standard entwickelt hat, dem SGML-Auszeichnungssystem. Dies trägt zu einer größeren Wiederverwendbarkeit der Korpora bei, die mit allgemein vorhandenen technischen Mitteln wie Textverarbeitung und Konkordanzprogrammen bearbeitet werden können, so daß nur in Ausnahmefällen die Entwicklung spezifischer Software notwendig war, z.B. zur Erzeugung eines lemmatisierten Korpus.

Das SGML-Format (vgl. 2.2.2.1) respektiert implizit die erste und die zweite Leitlinie Leechs und ermöglicht die Beachtung der dritten und vierten. Aufgrund des Formats, das SGML seinen Tags gibt, die mit spitzen Klammern eingeführt werden, können diese Tags aus dem Korpus entfernt oder separat gespeichert werden. Der *Header*, der textexterne Zusatzinformation enthält, kann seinerseits das Auszeichnungssystem wiedergeben oder darauf verweisen, und zugleich Angaben zur Person machen, die den Text bearbeitet hat.

Die Leitlinien 5, 6 und 7 von Leech beziehen sich auf die Auszeichnungen. Eine Zusammenfassung des verwendeten Tagsets befindet sich in Abschnitt 3.2.2.2, ebenso wie die Darstellung seiner Adäquatheit für unsere Untersuchung. Die ausführliche

[49] Entfernbarkeit, Extrahier- und Speicherbarkeit, Systematizität, Auszeichnungsmethode, Fehlerhaftigkeit, Allgemeinheit, Standard, vgl. 2.2.2.1.

Beschreibung des Tagsets ist bei Schiller et al. 1995 nachzulesen (vgl. Anhang, S. 409).

Hier ist jedoch vorwegzunehmen, daß in der Theorie theorieneutrale Tagsets gefordert werden, was zur Allgemeingültigkeit der Auszeichnungen beitragen soll (vgl. Leech 1993: 275). In der Praxis erweist sich die Vorstellung einer theorieneutralen Wortarteneinteilung jedoch als extrem problematisch. Aufgrund der Berücksichtigung teilweise unvereinbarer methodischer Ansätze entstehen Überschneidungen in den zugewiesenen Kategorien, die zu widersprüchlichen Ergebnissen führen können. So werden beispielsweise lexikalisierte Partizipien in STTS als Partizip (VVPP) getaggt, wobei jedoch Passivpartizipien „je nach Kontext auch eine adjektivische Lesart zulassen", wie im Falle „der Tisch wird verrückt" und „der alte Mann wird verrückt" (Schiller et al. 1995: 23). Ferner ist zu berücksichtigen, daß Tagsets mit distributionellen Kriterien erstellt werden, was sich aus computationellen Gründen ergibt. Die Anwendung von POS-Taggern, wie der Tagger des Instituts für maschinelle Sprachverarbeitung (ImS), mit dem unsere Korpora bearbeitet wurden, erzeugt zwei Arten von Problemfällen: einerseits typische Programmfehler (das Programm verfügt nicht über genügend Information zur korrekten Bestimmung eines Wortes und weist ihm ein inkorrektes Tag zu), andererseits aber auch theoriebedingte Problemfälle, d.h. problematische Zuweisungen von Tags, die dem zugrunde liegenden distributionellen Modell inhärent sind. Ein Beispiel dieser zweiten Kategorie ist die

Unterscheidung zwischen Adverb und adverbial gebrauchtem Adjektiv, die bei Schiller et al. (1995: 56) listenbasiert geleistet wird. Während Programmfehler durch eine manuelle Revision berichtigt werden können, werden theoriebedingte Problemfälle erst gelöst werden können, wenn die zugrunde liegende Theorie alle sprachlichen Phänomene erklären kann.

Hinsichtlich der Auszeichnungsprinzipien ist auf die Arbeit von Schiller et al. (1995) hinzuweisen, die in der Einleitung auf die Kombination von Prinzipien hinweisen: „Das STTS resultiert aus einer gegenseitigen Abstimmung zweier Part-of-Speech-Tagsets [...]. Als wichtigste Gliederungsaspekte bei der Einteilung der Wortarten wurden distributionelle Kriterien, aber auch traditionell-linguistische Kriterien (z.B. semantische und morphologische) zugrundegelegt." (Schiller et al. 1995: 3).

Die konkrete Anwendung dieser Prinzipien auf unsere Arbeit wird auf Seite 198 besprochen (vgl. dazu den Anhang).

Wie oben angemerkt, wurde SGML als Format für die Auszeichnungen übernommen. Hauptsächlich drei Gründe sprachen für die Anwendung eines schon bekannten Formats: erstens konnten dank der vorhandenen Bibliographie zu dem Format technische Probleme bei der Programmierung schneller und einfacher gelöst werden; zweitens konnten wir so sicher sein, daß für jedes auftretende Auszeichnungsproblem auch eine Lösung vorhanden war oder daß zumindest ein Ansatz zur Überwindung des Problems zu finden wäre, z.B. hinsichtlich der

Auszeichnung von Mehrworteinheiten; und drittens schließlich arbeitete die Mehrzahl der verwendeten Programme mit einem solchen Format, und es war zeit- und kostengünstiger, spezifische Fragen diesem allgemeinen Schema anzupassen, als erneut eigene Programme zur Verarbeitung herzustellen.

SGML bot eine Lösung für die meisten Probleme, die bei dem Auszeichnungsprozeß auftreten konnten. Die hierarchische Strukturierung der textuellen Eigenschaften ließ genug Freiraum für die Einbindung von Zusatzinformation (sowohl im *Header* als auch im Text innerhalb der Auszeichnungen), und die Deutlichkeit des Systems erlaubte es, mit relativ wenig Aufwand kleinere Programme zur Auswertung der Texte zu schreiben[50].

Für die Strukturierung der Dateien erübrigte sich im CLK und im CMK die Anwendung einiger technischer Elemente von SGML, wie die öffentlichen Deklarationen mit Information zum verwendeten Zeichensatz und der internen Struktur des Dokuments[51].

3.2.2.2 Auszeichnungsschema

Unter Auszeichnungsschema verstehen wir hier die systematisierte Zusatzinformation, die dem Text beigegeben

---

[50] Zum Aufbau eines SGML-Dokuments, vgl. 2.2.2.

[51] Diese Information kann für die zukünftige Bereitstellung im Institut für Maschinelle Sprachverarbeitung hinzugefügt werden.

wird. Die Zusatzinformation erscheint auf zwei Ebenen, der textexternen und der textinternen. In physischer Form wird textexterne Information im *Header* wiedergegeben, einer Art Inhaltsverzeichnis, das dem eigentlichen Text vorangestellt ist und kontextuelle Information zum Text, zum Verfasser oder zum Verarbeiter enthält. Textinterne Information hingegen erscheint im *Body*, der den Text enthält. Sowohl Header als auch Body sind SGML-Konform, d.h. sie entsprechen den Strukturierungsnormen eines SGML-Dokuments (vgl. Tabelle 9). Im folgenden werden die Auszeichnungsschemata zu textexterner und textinterner Information dargestellt.

Struktur eines Dokuments

```
<Doc>
    <Header>
        ...
    </header>
    <Body>
        <Titel> ... </titel>
        ...
        <Zitat> ... </zitat>
    </body>
</doc>
```

TABELLE 9

Der Header

Der *Header* ist die dem eigentlichen, zu untersuchenden Text vorangestellte Inhaltsangabe. Die Information im Header ist streng geordnet, um die Verarbeitung und Auswahl der Texte nach bestimmten Kriterien einfach und sicher zu gestalten; er enthält für die Ziele unserer Untersuchung im engen Sinne

nicht strikt notwendige Elemente, jedoch wurde diese Information bei Vorhandensein eingegeben, um die Wiederverwendbarkeit des Korpus für spätere Untersuchungen abzusichern.

Die Header der Texte des CMK und des CLK enthalten, abgesehen von den obligatorisch vertretenen Elementen, noch weitere, sprachspezifische Information, die CMK von CLK unterscheidet. Das in Tabelle 10 angegebene kumulative Headermodell ist jedoch vereinheitlicht, so daß manche Felder nur im CLK, andere nur im CMK ausgefüllt sind.

Struktur eines Headers

```
<Header>
    <Verfasser></verfasser>
    <Alter></alter>
    <Geschlecht></geschlecht>
    <Studium></studium>
    <Semester></semester>
    <Seminar></seminar>
    <Dozent></dozent>
    <Universität></universität>
    <Datum></datum>
    <Herkunft></herkunft>
    <Textsorte></textsorte>
    <Version></version>
    <Fehlerkorrektur></fehlerkorrektur>
    <Vereinheitlichungen></vereinheitlichungen>
    <Streichungen></streichungen>
    <Korrektor></korrektor>
    <Bearbeiter></bearbeiter>
    <Auszeichnungen></auszeichnungen>
    <Weiteres></weiteres>
</header>
```

TABELLE 10

Die im Header enthaltene Information erscheint linear in der textuellen Version, doch kann sie in zwei große Gruppen

eingeteilt werden: in Information zum Text und Information zur korpuslinguistischen Bearbeitung des Textes.

Die Information zum Text bezieht sich auf alle Parameter, die im Zusammenhang mit der Form und dem Inhalt des Textes stehen. An erster Stelle handelt es sich um Information zu dem Verfasser, die obligatorisch ist und der Überprüfung der ausgeglichenen Auswahl der Texte des Korpus dient. Es waren zwar Wiederholungen der Verfasser im Computerlernerkorpus erlaubt, und teilweise war dies sogar wünschenswert, um die Entwicklung der Evolution des Lernprozesses besser verfolgen zu können, doch sollte auch im CLK kein Student im Korpus überrepräsentiert sein. Die Anzahl der anonymen Texte liegt in keinem der Subkorpora über 10 Prozent. Zum Verfasser wurde ebenfalls das Alter angegeben; im CMK war das Alter nicht immer bekannt, doch wurde überprüft, daß es sich im Rahmen der für die Population angegebenen Bereiche bewegte (vgl. 3.1.2).

Ebenfalls angegeben wurde der Titel der Lehrveranstaltung, der der Text entnommen wurde, denn zusammen mit der Textsorte können thematische Unterschiede, die sich in dem Thema der Lehrveranstaltung widerspiegeln, Einfluß auf einige Phänomene und die quantitative Auswertungen eines Korpus haben.

Weitere Punkte waren das Geschlecht des Verfassers, gekennzeichnet als maskulin (*masc*) oder femenin (*fem*); der Studiengang, das Semester, in dem sich der Student befand, der Dozent dieses Seminars, im Falle des Computermuttersprachlerkorpus die Universität, wodurch

eventuelle regionale Varianten berücksichtigt werden konnten, das Datum des Textes, die Herkunft, also Information zum Ursprung des Textes und die Textsorte, um so eine präzisere Auswahl der Texte zu ermöglichen.

Die Information zur Bearbeitung bezieht sich auf die Aufbereitung des jeweiligen Textes für seine Einbindung in das Korpus; hier werden die Angaben zusammengefaßt, die formelle Aspekte betreffen und die in den Auszeichnungen enthaltene Zusatzinformation, aber auch jene, die die Herkunft des Textes und die Angemessenheit seiner Einbeziehung in das Korpus zu kontrollieren erlauben.

Zunächst erscheint Information zur Version, womit der jeweilige Bearbeitungsstand der Datei kontrolliert werden kann (vgl. Marcos Marín 1994: 100). Zusätzlich wird angegeben, ob die Datei einer Fehlerkorrektur unterworfen wurde. Dies ist nur der Fall der Texte des CLK, da an den Texten des CMK nur Änderungen hinsichtlich der Tippfehler oder auf orthographischer Ebene vorgenommen wurden; diese Änderungen wurden nicht im Header, sondern in der entsprechenden Logfile vermerkt, eine Datei, in der alle Hinweise zur Bearbeitung des Korpus vermerkt werden. Außerdem wurde im Header ebenfalls angegeben, welchen Vereinheitlichungen der Text unterzogen wurde und ob dadurch eventuell wichtige Elemente verloren gingen (z.B. Typographie, *ss* zu *ß*, Abkürzungen usw.). Der Eintrag Streichungen gibt an, was aus dem Original entfernt wurde: Titel, Inhaltsangabe, Zitate, Bibliographie, Register,

usw. Auch wird hier vermerkt, wer im Falle der Texte des CLK der Korrektor war. Danach folgt die Information zum Bearbeiter, also jener Person, die den Text vereinheitlicht hat, das POS-Tagging und die Revision der Tags durchführte. Das Feld *Auszeichnungen* gibt an, welches Auszeichnungssystem angewendet wurde, und zusätzlich welche Version des Systems. Die genauere Beschreibung befindet sich in der Logfile. Das letzte Feld, *weiteres*, ermöglichte die Eingabe von anfangs nicht berücksichtigter Information.

Der Body

Der Body ist der Abschnitt einer Datei, in dem der eigentliche Text wiedergegeben wird. Er enthält einerseits die eigentlich textuelle Information, d.h. den Text an sich; andererseits ist dieser Text mit Information angereichert, anhand der zusätzliche Untersuchungen an dem Korpus vorgenommen werden können.

Die Auszeichnung von Zusatzinformation bezieht sich auf die Wortartenklassifizierung STTS (POS-Tags), die in Schiller et al. dargestellt werden. Zusätzlich zu den POS-Tags enthalten die Texte noch Information zu Titeln von Abschnitten und Zitaten, die jedoch für den analytischen Teil dieser Arbeit nicht berücksichtigt wurden. Andere Auszeichnungsschemata, wie das syntaktische Tagging, wurden bei der Aufbereitung der Korpora nicht angewendet. In Tabelle 11 ist ein Beispiel für

die interne Struktur eines Bodys zu sehen; im folgenden Abschnitt wird das angewendete Taggingsystem besprochen.

Beispiel für den Body eines mit POS-Tagging versehenen Textes

```
<p><S>Es<PPER[es]> gibt<VVFIN[geben]> ein<ART[ein]>
Problem<NN[Problem]> mit<APPR[mit]> der<ART[d]>
Bildung<NN[Bildung]> der<ART[d]> Studenten<NN[Student]>
.</Korrektur>s<k> <S>Es<PPER[es]> kommt<VVFIN[kommen]>
aus<APPR[aus]> verschiedenen<ADJA[verschieden]>
Ursachen<NN[Ursache]> :<$.[:]> das<ART[d]> Geld<NN[Geld]>
,<$,[,]> die<PRELS[d]> Qualität<NN[Qualität]> und<KON[und]>
die<ART[d]> Bildung<NN[Bildung]> .<$.[.]> </s> <S>Die<ART[d]>
Zusammenfassung<NN[Zusammenfassung]> könnte<VMFIN[können]>
sein<VAINF[sein]> ,<$,[,]> daß<KOUS[daß]> die<ART[d]>
Verwaltung<NN[Verwaltung]> nicht<PTKNEG[nicht]> eine<ART[ein]>
ganze<ADJA[ganz]> öffentliche<ADJA[öffentlich]>
Universität<NN[Universität]> möchte<VMFIN[mögen]> .<$.[.]> </s>
<S>Deshalb<PAV[deshalb]> hat<VAFIN[haben]> die<ART[d]>
Verwaltung<NN[Verwaltung]> das<ART[d]> Etat<NN[Etat]>
gekürzt<VVPP[kürzen]> ,<$,[,]> und<KON[und]> jetzt<ADV[jetzt]>
ist<VAFIN[sein]> es<PPER[es]> sehr<ADV[sehr]>
schwer<ADJD[schwer]> ,<$,[,]> es<PPER[es]> zu<PTKZU[zu]>
wechseln<VVINF[wechseln]> ,<$,[,]> weil<KOUS[weil]> die<ART[d]>
Zeit<NN[Zeit]> ,<$,[,]> um<KOUI[um]> das<PDS[d]> zu<PTKZU[zu]>
machen<VVINF[machen]> ,<$,[,]> vorbei<ADV[vorbei]>
ist<VAFIN[sein]> .<$.[.]> </s> </p>
```

TABELLE 11

POS-Tagging

Die Entscheidung für die Aufbereitung der Texte mit Information zu den Wortarten (POS) fiel in einer relativ frühen Phase der Untersuchung. Unter den Auszeichnungsschemata, die in 2.2.2.2 genannt werden, ist das POS-Tagging augenblicklich das einzige maschinell anwendbare System. Andere Auszeichnungssysteme bringen einen hohen manuellen Arbeitsaufwand mit sich, entweder, weil die Auszeichnungen von Hand eingefügt werden müssen (wenn auch mit

der Unterstützung eines Programmes), oder weil die Fehlerquote so hoch liegt, daß die Bearbeitung größerer Textmengen nur mit erheblichen finanziellen Mitteln erfolgen kann. Ferner stellt das POS-Tagging die Grundlage für in Zukunft auszuführende Auszeichnungen dar, da die Bestimmung syntaktischer Strukturen notwendigerweise auf diese Information zurückgreift.

Die Verfügbarkeit eines Wortartensystems wie das von STTS, das in deutscher Sprache das detaillierteste, uns bekannte System zur Auszeichnung von Wortarten in Computerkorpora ist, verbunden mit dem Vorhandensein eines entsprechenden Taggers, der diese Wortarten automatisiert auf die Texte anwenden konnte, war entscheidend für die Wahl dieses Systems. Gleichzeitig bedingte STTS jedoch teilweise die Untersuchbarkeit sprachlicher Phänomene. Die distributionelle Unterscheidung von Konjunktionen und Adverbien in STTS ermöglichte so beispielsweise keine automatisierte Untersuchung zur vergleichenden Anwendung von Konjunktionen und Konjunktionaladverbien.

Für einige Voruntersuchungen wurde das Programm Morphy (Lezius et al. 1996) von Wolfgang Lezius verwendet, das gegenüber dem Tagger des Instituts für maschinelle Sprachverarbeitung die Vorteile bietet, unter DOS zu arbeiten und trotz einer stark reduzierten Anzahl an Wortartentags Information zu flektierten Elementen anzugeben. Die Anwendung dieses Programms erleichterte die Zusammenarbeit mit der vorhandenen Software

unter Windows, die schon teilweise für die Verarbeitung der Korpora angepaßt worden war.

Das Programm Morphy, zuerst in seiner DOS-Version 1.3, später in seiner Windows-Beta-Version 3.0 und dann der Endversion 3.01, arbeitete mit einer Trefferquote von ugf. 90 Prozent (Lezius et. al 1996). Das Programm lieferte Information zu einem reduzierten Wortartentagset und zudem morphologische Angaben (Lezius 1996). Nach eingehender Revision der Ausgabe von Morphy wurde jedoch klar, daß die Fehlerquote zu einer sehr aufwendigen Revision geführt hätte und daß das System sehr anfällig für nicht vorhergesehene Sonderzeichen und orthographische Abweichungen war. Zudem hätten die Wortartentags nicht zu einer differenzierten Untersuchung der sprachlichen Phänomene führen können. Die Tags von Morphy beschränken sich auf die 6 Hauptwortarten *Substantive, Adjektive, schwache* und *nicht-schwache Verben, Eigennamen* und *Sonstige*. Der Vergleich mit der Ausgabe des ImS-Taggers zeigte, daß dieser eine höhere Trefferquote erzielte, was die Revisionsarbeit erleichterte, und daß, trotz fehlender morphologischer Auszeichnung, die Differenziertheit der Wortartentags von STTS entscheidend für die Bestimmung und differenzierte Darstellung der sprachlich relevanten Phänomene war.

Der Tagger des Instituts für maschinelle Sprachverarbeitung konnte nach der Bestätigung der Lizenzbestimmungen frei angewendet werden. Seine Einschränkung ist, daß er nur unter

UNIX läuft, doch gibt es eine spezifische Version für Linux, so daß er auch auf einem normalen Rechner installiert werden konnte.

Das Format der Auszeichnungen wurde grundlegend von der automatischen Ausgabe des ImS-Tagger bestimmt. Nach der Bearbeitung zeigt der ImS-Tagger die Information in Tabellenform an, wobei an erster Stelle das Originalwort erscheint, an zweiter der vom Programm zugeordnete Tag und an dritter die Lemmatisierung des betreffenden Wortes, soweit dieses im Lexikon enthalten ist; ansonsten wird als Lemmatisierung das Originalwort ausgegeben:

kleinen	ADJA	klein
Golfballes	NN	Golfballes

Diese „falsche" Lemmatisierung wurde während des Revisionsprozesses korrigiert, um mögliche Verzerrungen der Ergebnisse in der Untersuchung zur Type-Token-Ratio zu vermeiden, die aufgrund der lemmatisierten Einträge berechnet wurde.

In STTS wird für die Angabe der lexikalischen und morphologischen Kategorien ein System benutzt, das die Bearbeitung mit dem Computer erleichtert, da die verschiedenen Angaben von Sonderzeichen getrennt werden. Auf das Originalwort folgt der eigentliche POS-Tag mit Querstrich, danach die lexikalische Information mit spitzer Klammer und

schließlich die morphologische Information mit Doppelpunkt, wobei die jeweiligen morphologischen Kategorien wie Genus, Kasus, Numerus usw. durch Punkte voneinander getrennt sind, jedoch immer in der gleichen Reihenfolge, was die Identifizierung der Elemente erleichtert:

```
Wort/POS<LexikalischeInformation:MorphologischeInformation
Lemmatisierung
```

Da unser Bearbeitungssystem auf anderen Zeichenfolgen und Routinen beruhte, die für die Korrektur der Lernertexte entwickelt worden waren, entschieden wir, nicht die Programmierroutinen zu ändern, was einen hohen Arbeitsaufwand mit sich gebracht hätte, sondern STTS so zu übersetzen, daß es sich diesen Routinen anpaßte. Hierbei war eine Übersetzung der entsprechenden Sonderzeichen notwendig, die wenn nötig wieder in STTS übersetzbar sind:

```
Wort*<POS*:LexikalischeInformation/MorphologischeInformation
[Lemmatisierung*]
```

Ebenfalls war zu beachten, daß bei der späteren Hinzufügung von morphologischen und lexikalischen Kategorien nicht alle berücksichtigt wurden, die in STTS vorgesehen sind, sondern nur jene ausgewählt wurden, die für die quantitativen Untersuchungen wichtig erschienen. Aus diesem Grund waren nur die mit * gekennzeichneten Elemente obligatorisch, also das Originalwort, der POS-Tag und die Lemmatisierung.

Das POS-Tagset von STTS

Das Tagset für die POS-Auszeichnungen beruht auf den „Vorläufigen Guidelines für das Tagging deutscher Textcorpora mit STTS" (Schiller et. al 1995; vgl. Anhang), an dem allerdings zu einigen Kategorien kleinere Änderungen eingefügt wurden.

Das STTS-Tagset besteht in seiner aktuellen Form aus folgenden Kategorien:

POS	Beschreibung	Beispiele
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	Vergleichspartikel, ohne Satz	als, wie
NN	normales Nomen	Tisch, Herr, [das] Reisen
NE	Eigenname	Hans, Hamburg, HSV
PDS	substituierendes Demonstrativpronomen	dieser, jener
PDAT	attribuierendes Demonstrativpronomen	jener [Mensch]
PIS	substituierendes Indefinitpronomen	keiner, viele, man, niemand
PIAT	attribuierendes Indefinitpronomen ohne Determiner	kein [Mensch], irgendein [Glas]
PIDAT	attribuierendes Indefinitpronomen mit Determiner	[ein] wenig [Wasser], [die] beiden [Brüder]
PPER	irreflexives Personalpronomen	ich, er ihm, mich, dir
PPOSS	substituierendes Possessivpronomen	meins, deiner
PPOSAT	attribuierendes Possessivpronomen	mein [Buch], deine [Mutter]
	Relativpronomina	
PRELS	substituierend	[der Hund,] der
PRELAT	attribuierend	[der Mann,] dessen [Hund]
PRF	reflexives Personalpronomen	sich, einander, dich, mir
PWS	substituierendes Interrogativpronomen	wer, was
PWAT	attribuierendes	welche [Farbe], wessen [Hut]
PWAV		warum, wo, wann,

	Interrogativpronomen adverbiales Interrogativ- oder Relativpronomen	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",	anzukommen, loszulassen
VVPP	voll	gegangen, angekommen
VAFIN	Partizip Perfekt, voll	[du] bist, [wir] werden
VAIMP	finites Verb, aux	sei [ruhig!]
VAINF	Imperativ, aux	werden, sein
VAPP	Infinitiv, aux	gewesen
VMFIN	Partizip Perfekt, aux	dürfen
VMINF	finites Verb, modal	wollen
VMPP	Infinitiv, modal	[er hat] gekonnt
	Partizip Perfekt, modal	
XY	Nichtwort, Sonderzeichen enthaltend	D2XW3
$,	Komma	,
$.	Satzbeendende	.?!;:
$(	Interpunktion	- [ ] ( )
	sonstige Satzzeichen; satzintern	

TABELLE 12

Bei diesem Tagset wird jeder Wortform ein einziger Tag zugewiesen; unter dem Begriff Wortformen werden auch Ziffern, Satzzeichen und weitere Sonderzeichen zusammengefaßt (Schiller et al. 1995: 4).

Das Problem der POS-Auszeichnungen besteht darin, daß im Unterschied zum *Parsing* (vgl. 2.2.2.2) aufgrund der Auszeichnungsmethode aus technischen Gründen keine Mehrwortlexeme angegeben werden können (Schiller et al. 1995:

4), diese also manuell untersucht werden mußten und sich nicht als Auszeichnungen in dem Korpus niederschlugen. Funktionsverbgefüge, Redewendungen usw. wurden aus diesem Grund anhand von einigen allgemeinen Eigenschaften untersucht und waren deshalb nicht als einheitliche Klasse zusammenfaßbar, sondern nur als statistischer Hinweis auf das Verhalten der jeweiligen Gesamtgruppe.

Das führt ebenfalls dazu, daß Mehrwortlexeme im analytischen Teil als genau die Anzahl von Wörtern berechnet werden, die sie getrennt darstellen; vgl. dazu in 4.2 die Fragen, die dies für Wort- und Satzlänge und im besonderen für die Type-Token-Ratio (d.h. die lexikalische Variation) aufwirft. Dies bedeutet ebenso, daß Mehrwortlexeme nicht direkt als Einheiten, die sie ja tatsächlich darstellen (vgl. Fleischer 1982: 35 und Wotjak 1992: 3), berechnet werden können; zur Lösung des Problems vgl. die Kommentare im analytischen Teil zu rekurrenten Wortkombinationen in 4.2.1.3.

Eine weitere Einschränkung des Taggers ist die fehlende Information zu morphologischen und lexikalischen Eigenschaften. Diese wurden durch ein problemorientiertes Tagging eingegeben. Im Gegensatz zum Ganztagging, das automatisiert alle vorhandene Information eingibt, wird beim problemorientierten Tagging nur jene Information berücksichtigt, die zur Lösung eines bestimmten Problems notwendig ist (vgl. Haan 1991: 52). Ein Beispiel dafür sind

die Adjektive partizipialen Ursprungs oder die Genitivattribution.

4 Auswertung des empirischen Materials

In diesem Kapitel wird das empirische Textmaterial der in Kapitel 3 beschriebenen Korpora untersucht, ausgewertet und dargestellt. Der erste Abschnitt ist der Beschreibung der Methode gewidmet, die zur Bestimmung und endgültigen Auswahl der quantifizierbaren Phänomene angewendet wurde. Der zweite Teil dieses Kapitels, der aus der Untersuchung der ermittelten Phänomene in den Korpora besteht, ist als empirische Grundlage für die qualitativen Schlußfolgerungen des folgenden Kapitels aufzufassen.

4.1 Bestimmung und Auswahl relevanter Phänomene

Anhand der Auszeichnungen und der Texte selbst konnte eine große Anzahl an Untersuchungen an den Korpora vorgenommen werden, von denen jedoch nicht alle dem gesetzten Ziel zu nutzen gewesen wären. Eine Bestimmung der Phänomene (*Profilierung*), die eine Antwort auf die gestellten Fragen ermöglicht, ist jedoch fast unmöglich, wenn sie a priori vorgenommen werden soll. Erste Untersuchungen anhand der Preliminarversionen der Korpora hatten ergeben, daß nur wenige

der aleatorisch ausgewählten Phänomene, wie z.B. die Satzlänge oder die Frequenz der Adjektive, signifikante Unterschiede ergaben. Das Problem der Bestimmung der Phänomene wird oft erwähnt, aber nie gelöst. Crystal (1991: 227) z.B. schlägt eine eher empirisch geprägte Lösung vor: „This is the paradox of profiling: one needs to divise a profiling procedure in order to discover whether a profiling procedure is possible. Profiling procedures grow as they are used. They drive on experience and application." Biber (1995a:35) fügt zu dem Problem der Profilierung noch das der *co-occurrence* hinzu, da bei seiner Untersuchung zur Bestimmung von Textsorten anhand von sprachlichen Phänomenen nicht ein einzelnes Phänomen wichtig ist, sondern seine Verbindung mit anderen: „few empirical investigations are based on the analysis of co-occurring linguistic features. [...] In part, this shortcoming is due to the fact that the empirical identification of co-occurrence patterns has proven to be quite difficult."

Für die Profilierung wurden verschiedene Quellen herangezogen: Grammatiken, korpuslinguistische Untersuchungen über relevante Phänomene[52] und weitere Phänomene, die anhand der Untersuchung der Preliminarversionen bestimmt wurden. Die Phänomene wurden danach hierarchisch an den Vorversion der Computerkorpora überprüft, mit dem Ziel, zu klären, ob ihre Untersuchung relevant und möglich war.

[52] Vgl. Crystal 1991: 226.

Es muß jedoch darauf hingewiesen werden, daß die Überprüfung dieser Phänomene anhand der Preliminarversion der Korpora erneut an der Endversion vorgenommen wurde, denn es gab Phänomene, die keine Unterschiede in der Preliminarversion aufwiesen, wohl aber in der vorläufigen Endversion, was dann zwangsweise zur Korrektur des Korpusdesigns führte[53] und schließlich zu einer endgültigen Endversion.

Als wichtigste Quelle zur Profilierung erwies sich das Modell der Registervariation (Biber 1995a: 29f), in dem jene Elemente angegeben werden, die in den häufigsten Fällen zur Beschreibung einer bestimmten sprachlichen Variante hinsichtlich ihrer Textsorte beitragen. Das Modell der Registervariation bezieht sich auf die englische Sprache und hat primär das Ziel, Register (Kombinationen von sprachlichen Phänomenen) zu erstellen, anhand der sich eine Textsorte von einer anderen unterscheidet. Obwohl dieser Ansatz nur in bezug auf die Strukturierung nach Textsorten mit dem dieser Arbeit verbunden scheint, handelt es sich bei Biber um Phänomene, die oft erst dank der POS-Auszeichnungen beschreibbar werden. Zu den Phänomenen der Registervariation wurden weitere hinzugefügt, die verschiedenen korpuslinguistischen

[53] Beispiele für Phänomene, die in der Vorversion des Korpus keine Unterschiede ergaben, aber in der Endversion ja, waren die Konjunktionen. Durch die ermittelten Frequenzen wurde entdeckt, daß es sich bei den Texten des Lernerkorpus in hohem Maße an nicht angegebenen oder ausgezeichnetem fremdtextuellen Material handelte.

Untersuchungen entstammen. Die Phänomene der Registervariation werden in der folgenden Darstellung mit *RV* gekennzeichnet und sind in Biber (1995a) nachzulesen; alle weiteren Phänomene, inklusive die der Grammatiken, werden mit dem Namen des Autors gekennzeichnet.

Die Gliederung der Phänomene erfolgt hier mit Rücksicht auf ihre Erkennbarkeit in den Korpora. Dabei handelt es sich um

- Sprachunabhängige statistische Phänomene, d.h. Phänomene, die anhand statistischer Methoden quantifiziert werden können und nicht von der konkreten Sprache abhängig sind;
- lexikalische Phänomene, die mit Hilfe von Wortlisten in den Korpora gesucht werden;
- morphologische Phänomene, auf die erst nach der morphologischen Auszeichnung der Texte zurückgegriffen werden kann;
- grammatische oder POS-Phänomene, für deren Bearbeitung eine POS-Auszeichnung benötigt wird;
- syntaktische Phänomene, die ebenfalls einer speziellen Auszeichnung bedürfen;
- argumentative, semantische und weitere theoriegebundene Phänomene, für die die Texte entsprechend auszuzeichnen sind, einem Auszeichnungsschema folgend, das, ebenso wie die syntaktische Auszeichnung, an eine bestimmte Theorie gebunden ist;
- nicht erkennbare Elemente oder manuell zu bestimmende Phänomene, die keinem der vorhergehenden Schemata klar zuzuschreiben sind. Es kann sich hier um komplexe lexikalische Einheiten handeln oder um Elemente, die eine grammatische Form haben, aber einer anderen zuzurechnen sind, wie z.B. die Konkurrenzform des Passivs; und nicht

anwendbare Elemente, die in anderen Sprachen relevant sein können, aber im Deutschen keine Anwendung finden.

Sprachunabhängige statistische Phänomene

Zu diesen Phänomenen können die Wortlänge oder lange Wörter der Registervariation (RV) und die Satzlänge[54] gerechnet werden und die lexikalische Variation aufgrund der in den Texten erscheinenden orthographischen Wörter (also nicht aufgrund der Lemma), aber auch die Cluster, d.h. rekurrent erscheinende Wortkombinationen, die oft nur statistisch oder zufallsmäßig bedingt sind (*in die*), oft aber auch lexikalischen Ursprung haben (*Einführung in*).

Lexikalische Phänomene

Zu den lexikalischen Phänomenen können folgende Elemente gerechnet werden: Negation anhand von Präfixen (siehe Kwon 1997: 21ff), maskuline oder feminine Berufsbezeichnungen (siehe Holmes 1994: 31ff), Modalverben (siehe Collins 1991), die allerdings leichter zu erkennen sind, wenn die POS-Auszeichnungen sie gesondert angeben, Proverben (RV), Sprechaktverben (RV), Verben mentaler Prozesse (RV), und Nominalisierungen (RV), die nur anhand von Suffixen bestimmt werden.

[54] Vgl. Haan 1993, Haan 1996: 23

Morphologische Phänomene

Es handelt sich bei den morphologischen Phänomenen um Tempus und Aspekt (RV), Futur und Ausdruck für Zukunft, wobei hier morphologisch nur das Futur zu erkennen wäre, während andere Realisierungen der Zukunftsform spezifisch ausgezeichnet werden müßten (siehe Berglund 1997: 7ff).

Grammatische oder POS-Phänomene[55]

Zu dieser Gruppe gehören fast alle Wortarten, besonders aber Nomen (RV), Pronomina (RV), Relativpronomina (RV), Adjektive (RV), Adverbien (RV), Modalverben (RV), Negationselemente (RV), Koordination (RV) (nur anhand von koordinierenden Konjunktionen) und Kollokationen hochfrequenter grammatischer Wörter[56].

Syntaktische Phänomene

Zu den syntaktischen Phänomenen gehören die adverbiale Subordination (RV), die Präpositionalphrasen (RV), die Apposition, die Relativsätze (RV) und die Infinitivkonstruktionen. Die Liste der möglichen syntaktischen

[55] Man darf grammatische POS-Phänomene nicht mit syntaktischen Phänomenen verwechseln. In dem ersten Fall wird „grammatisch" im Sinne von syntaktischem Wort verwendet; vgl. S. 46.

[56] Siehe den Begriff Collocational Framework, Renouf/Sinclair 1991

Untersuchungen wäre lang, doch gibt es erst wenige syntaktisch ausgezeichnete Korpora.

Argumentative, semantische und weitere theoriegebundene Phänomene

Es handelt sich dabei um die Strukturierung von Information (RV), den Ausdruck der Möglichkeit (siehe Løken 1997: 43ff), den Ausdruck für Zukunft (siehe Berglund 1997: 7ff), verschiedene Grammatische Informationsstrukturierungsmittel (RV) u.a.

Nicht erkennbare Elemente oder manuell zu bestimmende Phänomene

In dieser Gruppe erscheinen die Konkurrenzformen des Passivs (RV) und Funktionsverbgefüge.

Die augenblicklichen Auszeichnungen des CMK und des CLK stellten eine Einschränkung der Erkennbarkeit dar. Von den Gruppen konnten, ausgehend von dem vorliegenden Material, grundsätzlich nur die allgemeinen statistischen Phänomene bearbeitet werden und die verschiedenen syntaktischen Wortarten, die vom POS-Tagger ausgezeichnet wurden. Eine Beschränkung auf z.B. syntaktische Phänomene anstatt der syntaktischen Wortarten war nicht möglich, weil das eine weitere Auszeichnung verlangen würde, die den Rahmen einer

solchen Untersuchung überschreitet. Die syntaktischen Wortarten stellen immerhin das erste Stadium der Erkennbarkeit und Auszeichnung der syntaktischen Phänomene dar. Dennoch werden einige syntaktische Phänomene berücksichtigt, die anhand der Auszeichnung der Wortarten identifizierbar sind, wie z.B. Relativsätze anhand von Relativpronomina.

Von den restlichen Phänomenen wurden jene gestrichen, die hinsichtlich ihrer Relevanz für die vorliegende Untersuchung, ihrer Analysierbarkeit, Vertrauenswürdigkeit der Daten und Relevanz der gefundenen Abweichungen nicht berücksichtigt werden konnten.

Phänomene, die nach dieser Vorgehensweise für eine Untersuchung schließlich in Betracht kamen, waren nach Gruppen:

- Sprachunabhängige statistische Phänomene: Wort- und Satzlänge, lexikalische Variation, Cluster.
- Lexikalische Phänomene: Negation, die einerseits in den POS-Auszeichnungen als Negationspartikel berücksichtigt wird, andererseits in der Untersuchung zur Wortbildung.
- Grammatische oder POS-Phänomene: Alle syntaktischen Wortarten von STTS.
- Syntaktische Phänomene: Relativsätze, Infinitivkonstruktionen.

Die Phänomene wurden in zwei große Gruppen eingeteilt, die sich aufgrund der zur Untersuchung angewendeten Methode

ergeben: zuerst die allgemeinen statistischen Untersuchungen, in denen jene Phänomene zusammengefaßt werden, die ohne Auszeichnungen untersucht werden; zweitens die grammatischen Untersuchungen, für deren Bearbeitung ausschließlich die getaggten Korpora benutzt wurden, da die Untersuchung nicht mehr anhand des orthographischen Wortes vorgenommen wird, sondern anhand der jedem Wort zugewiesenen POS-Kategorien.

Die einzelnen Phänomene dieser Liste werden im Abschnitt 4.2 näher untersucht; die qualitative Analyse, in der die multivarianten Verbindungen zwischen den Phänomenen verdeutlicht werden, wird in Kapitel 5 dargestellt.

4.2 Auswertung sprachlicher Phänomene

Die Auswertungen der sprachlichen Phänomene dieses Abschnitts sind aufgrund der zugrundeliegenden Methode in *allgemeine statistische und lexikalische Phänomene* und *Darstellung der STTS-Wortarten* eingeteilt. Während in der ersten Gruppe jene Phänomene untersucht werden, die unabhängig von konkreten sprachlichen Phänomenen durchgeführt werden können, wie die Satzlänge, und aufgrund von lexikalischen Phänomenen (lexikalische Variation, Affigierung u.a.), werden in der zweiten Gruppe die syntaktischen Wortarten von STTS untersucht.

Diese Einteilung wurde zur gesonderten Bestimmung und Bewertung jener Phänomene eingeführt, die vollkommen oder

teilweise unabhängig von einer Bearbeitung oder Auszeichnung der Texte mit einem POS-Tagging-Programm erhalten werden können. Der Vorteil vollkommen sprachunabhängiger Phänomene liegt darin, daß ihre Anwendung auf Einzeltexte oder Textgruppen keine Bearbeitung voraussetzt, also kein zusätzlicher Arbeitsaufwand entsteht, bevor erste Werte erarbeitet werden können. Die Untersuchung der Texte anhand dieser Phänomene erlaubt so erste Ansätze zur Bestimmung einiger textinterner Eigenschaften, wie die schon erwähnte Satzlänge.

Wenngleich die endgültige Zuordnung zu einer Textsorte oder einem sprachlichen Niveau immer in Verbindung mit weiteren, aus der POS-Auszeichnung gewonnenen Phänomenen erfolgen muß[57], ermöglichen es diese ersten Untersuchungen anhand sprachunabhängiger Phänomene, Grundeigenschaften von den Texten zu überprüfen und einzelne Texte, die sich von den anderen durch extrem hohe Werte abheben, näher zu untersuchen und eventuell auszusondern, so daß sie schließlich nicht mehr dem aufwendigen Prozeß des POS-Taggings und seiner Revision unterworfen werden müssen.

Zwischen sprachunabhängigen Untersuchungen und POS-Untersuchungen gibt es eine Gruppe, die wir in dieser Arbeit zusammen mit den sprachunabhängigen darstellen. Es handelt

[57] Vgl. hierzu die Diskussion über interne und externe Kriterien hinsichtlich des Korpusdesigns in EAGLES (1996b: 4).

sich tatsächlich um statistische und lexikalische Phänomene, die aber unter Berücksichtigung der POS-Tags leichter und schneller maschinell untersucht werden konnten.

Die maschinelle Unterscheidung zwischen Wörtern, Satzzeichen, Leerzeichen u.a. stellt eine Segmentierung der Zeichenkette, die eine Computerdatei enthält, in erste analysierbare Elemente dar. Auf der Grundlage des sogenannten *scanning*, die Einteilung der Texte in klar abgegrenzte Einheiten (Meya/Huber 1986: 41), können mathematisch-statistische Verfahren angewendet werden, die Information zur Beschaffenheit, Verteilung und Interaktion dieser Einheiten geben. Diese Verfahren ergeben sich aus den Kombinationsmöglichkeiten der einzelnen Segmente oder der Bedeutung, die denselben zugeschrieben wird[58].

Für die vorliegende Untersuchung wurden jedoch nicht alle möglichen Kombinationen zwischen diesen Segmenten als relevant angesehen. Als statistisch wertvoll haben sich Wort- und Satzlänge herausgestellt, aber auch die rekurrenten Wortkombinationen oder *Cluster*, d.h. die statistische Frequenz, mit der zwei oder mehr Wörter überdurchschnittlich oft zusammen erscheinen (wie es z.B. der Fall des Konnektors

[58] So stellt eine alphabetische Zeichenkette begrenzt durch Leerzeichen oder Satzzeichen ein Wort dar, während ein Punkt und ein Fragezeichen ein Satzende bedeuten. Die Kombination von Wort und Satzende ergibt den Wert der Satzlänge in Wörtern, die Kombination von Wort mit seiner strukturellen Eigenschaft „alphabetische Zeichen" die Wortlänge in Zeichen.

so daß wäre), sowie die lexikalische Variation (Type-Token Ratio oder TTR) und die Wortbildung anhand der Affigierung. Nicht relevant schien hingegen die Untersuchung anderer Phänomene, wie die Absatzlänge, die Frequenz von Klammern u.a.

Die zweite Gruppe bilden die grammatischen Untersuchungen, für deren Durchführung die vom POS-Tagger zugewiesenen Tags notwendig waren, d.h. eine Analyse auf der Basis syntaktischer Wortarten. Ein unbearbeitetes Korpus läßt z.B. keine Untersuchung zur Frequenz von Substantiven zu, weil diese nur anhand eines Taggers von anderen Wortarten abgegrenzt werden können. Ebenso wäre es nicht möglich zu analysieren, wie oft zwei Adjektive zwischen Artikel und Substantiv erscheinen, ohne über die entsprechenden Tags zu verfügen. So wurden in dieser Gruppe all jene Untersuchungen zusammengefaßt, die auf diese Information zurückgreifen und erst nach dem Tagging und der Revision der Tags durchgeführt werden können.

Die einzelnen Untersuchungen wurden auf der Grundlage von den zwei Versionen der Korpora durchgeführt, die in 3.2 beschrieben sind, die rohe Textversion und die POS-Version.

Die Textversion wurde für die allgemeinen statistischen Untersuchungen verwendet, mußte aber in manchen Fällen auch mit der POS-Version kombiniert werden, wie z.B. zur Bestimmung der Satzlänge. Dennoch bildet die Textversion die Grundlage der mathematischen Berechnungen, da sie auch zur Bestimmung des Umfangs der Korpora herangezogen wurde; so wird z.B. die

Satzlänge berechnet, indem der Wert der Wortzahl des rohen Textkorpus durch die Anzahl der Satzendetags des POS-Korpus dividiert wird.

Die POS-Version wurde für die grammatischen Untersuchungen herangezogen, weil sie kombinierte Suchen zwischen grammatischen Tags und lexikalischen Einheiten ermöglichte. Sie diente des weiteren als Grundlage für eine lemmatisierte Textversion, die nur aus den lemmatisierten Texten bestand und zur Berechnung der lexikalischen Variation benutzt wurde.

Schließlich sei darauf hingewiesen, daß die Untersuchung der Phänomene rein deskriptiv ist und nur Unterschiede und Übereinstimmungen festhält, es aber in diesem Kapitel keine Erklärungsansätze gegeben werden. Die Auswahl der Phänomene selbst erfolgte jedoch aufgrund qualitativer Kriterien, die dem Ziel entsprachen, jene Elemente darzustellen, die zur Erläuterung des Sprachstands des Lerners beitragen konnten.

Ein weiterer qualitativer Ansatz in diesem Kapitel besteht in den untersuchten Aspekten eines Phänomens. Diese Aspekte wurden so ausgewählt, daß sie als empirische Grundlage für die qualitativen Untersuchungen als Schlußfolgerungen dienten. Zusammenfassend werden in den folgenden Untersuchungen jene Phänomene dargestellt, die Abweichungen aufweisen, und innerhalb der Phänomene jene Aspekte quantitativ untersucht, die zu einer qualitativen Erklärung und Beschreibung übergeordneter Phänomene beitragen. Ausgelassen wurden jene

Phänomene, die keine Abweichung aufwiesen und Aspekte, die sich als nicht relevant für Differenzierung eines Phänomens erwiesen.

4.2.1 Allgemeine statistische Untersuchungen

Die in diesem Abschnitt dargestellten Untersuchungen beziehen sich auf fünf Phänomene, die in verschiedenen Voruntersuchungen statistisch hohe Abweichungen ergeben hatten.

- Die großen Unterschiede im hohen Bereich der Wortlänge führte zur Überlegung, daß dafür Wortbildungsmechanismen verantwortlich seien.

- Die stark abweichenden Werte der Satzlänge ihrerseits deuteten auf eine Verwendung von entweder komplexeren syntaktischen Satzbauplänen hin oder von mehr freien Elementen, wie Adjektiven, Adverbien oder Präpositionalgruppen.

- Die rekurrenten Wortkombinationen waren ein Hinweis auf stärker lexikalisierte Formeln oder phraseologische Komponenten.

- Die lexikalische Variation bildete eine Hinweis auf den aktiven Wortschatz, der in beiden Computerkorpora angewendet wurde. In vielen Untersuchungen wird die lexikalische Variation aufgrund eines rohen Textkorpus berechnet; wir

ziehen hier die Berechnung anhand der lemmatisierten Einträge vor. Dies hat den Vorteil, daß die größere Variation von Flexionsformen im CMK (wie die Verwendung des Genitivattributs, vgl. 4.2.2.1) keinen Einfluß auf die lexikalischen Werte nimmt.

- Das letzte Phänomen der ersten Gruppe ist die Affigierung, ein Indiz, zusammen mit der Wortlänge, für die Anwendung und Variation von Wortbildungsmechanismen.

4.2.1.1 Wortlänge

Zur Definition von Wortlänge ist zunächst computationell der Terminus Wort zu definieren. Die traditionelle Einteilung der Worteinheit in Form und Inhalt (vgl. Clément 1996: 19) kann dabei nicht berücksichtigt werden, da dem Computer dafür erst die lexikalisch notwendige Information zur Bedeutung eines Wortes im Kontext einzugeben wäre. Syntaktische Definitionen leisten ebenfalls keine Hilfe, da computationell die Mechanismen Verschiebbarkeit und Ersetzbarkeit (vgl. Bußmann 1990: 849) noch nicht anwendbar sind.

Grundlage für die computationelle Definition ist die Trennung auf orthographisch-phonemischer Ebene durch Leerzeichen. So werden als Worteinheit alle nicht-leeren Zeichenfolgen zwischen Leerraum, Satzzeichen gefolgt von Leerzeichen oder Sonderzeichen gefolgt von Leerzeichen aufgefaßt (vgl. Leidner 1997); zum Wort gehören alle alphabetischen Zeichen, aber auch

Sonderzeichen (wie der Bindestrich und das Apostroph: *Baden-Württemberg, geht's*), wenn ihnen kein Leerzeichen folgte.

Als Wortlänge definieren wir hier die Zahl der Zeichen eines Wortes. Andere Maßeinheiten für die Bestimmung der Länge eines Wortes, wie Silben, Morpheme und andere, werden aufgrund der angewendeten Computerprogramme und der allgemein in der Korpuslinguistik angewendeten Methoden nicht in Betracht gezogen (zur Diskussion dieser Faktoren und ihres Einflusses, siehe Grotjahn/Altman 1993: 141ff).

Diese Definition von Wort und Wortlänge bereitet einige computationelle und methodische Probleme. Computationell sind die Dateien von sogenanntem Datenmüll zu reinigen (vgl. Blackwell 1992: 97ff), was durch eine intensivere manuelle Bearbeitung der Texte erreicht wurde[59]. Methodisch gab es keine Möglichkeit, eine einheitliche computationell ohne Einzelauszeichnung anwendbare Definition für Wort zu finden.

---

[59] Aufgrund ihres Ursprungsformats enthielten viele Texte Sonderzeichen und Formatierungselemente, die die Berechnungen verzerrt hätten. So erschienen in vielen Texten z.B. Bindestriche am Ende einer Zeile zur Markierung der Worttrennung, die für den Korpus nicht mehr notwendig waren und zu einer falschen Berechnung der Wortlänge geführt hätten. In diesen Fällen wäre der Bindestrich ein weiteres Zeichen in der Wortlänge gewesen; vgl. Grefenstette / Tapanainen (1994: 3f.). Des weiteren erschienen Referenzen zu den Fußnoten in Klammern und numeriert innerhalb des Textes, die ebenfalls eliminiert wurden, weil auch die entsprechende Fußnote nicht in den Korpus aufgenommen wurde. Der Rest des Datenmülls wurde ebenfalls gestrichen (HTML-Codes, nach der Streichung von Tabellen und Grafiken noch verbleibende Legenden usw.).

Getrennt erscheinende trennbare Elemente (*er fing an*), mehrteilige Konjunktionen (*wenn ... dann*), Wörter mit Bindestrich (*Baden-Württemberg*) und andere seltenere Phänomene wurden entsprechend der vorhergehenden Definition und den Hinweisen von Grotjahn/Altman (1993: 143) als mehrere Wörter (*er fing an*: 3, *wenn ... dann*: 2) oder ein Wort (*Baden-Württemberg*: 1) berechnet[60].

Die Untersuchbarkeit des Phänomens ergibt sich aus den von dem Konkordanzprogramm erstellten Wortlisten, in denen die Anzahl der Wörter einer gegebenen Länge *n* in jedem Text des Korpus angegeben werden. Da das Wort die konstitutive Grundeinheit

---

[60] Elemente wie Baden-Württemberg und auch die E-Mail-Adresse strunk@lingua.fil.ub.es wurden als ein Wort berechnet, denn nach den beiden Sonderzeichen Bindestrich und Klammeraffe und den Satzzeichen in der E-Mail-Adresse erscheinen keine Leerzeichen, die die Einleitung für ein neues Wort wären. Doch nach dieser Definition werden auch extreme Beispiele wie der Flug London-Berlin mit 3 Wörtern berechnet, wobei es in Wirklichkeit 4 sind. Wenn wir aber das Beispiel der Flug London-New York nehmen, erhalten wir wieder 4, jedoch ebenfalls fälschlicherweise. Eine weitere Einschränkung stellten die uneinheitlichen Darstellungsformen von Zahlen, insbesondere der Datumsangaben, dar. Schrieb ein Student 4.1.98, wurde dies als ein Wort berechnet; schrieb er 14. Januar 1998, als drei. Computationell wäre für die Lösung all dieser Probleme eine Einzelbestimmung von Worteinheiten notwendig gewesen, was aber mit den augenblicklich allgemein verfügbaren Computerprogrammen noch nicht korrekt und einheitlich zu bewältigen ist. Nach einer Stichprobenuntersuchung, aus der hervorging, daß es sich im Falle der Sonderzeichen um wenig frequente Phänomene handelt, wurde der Entschluß getroffen, auf diese mögliche aber minimale Verzerrung hinzuweisen, ohne sie jedoch gesondert zu beheben.

eines Textes bildet, ist es obligatorisch in jedem Text vertreten.

Zunächst war zu untersuchen, ob die mittlere Wortlänge Abweichungen hinsichtlich der Textsorten mit sich zieht und wie stark diese Abweichung sein kann; dafür wurde das Computermuttersprachlerkorpus als Grundkorpus herangezogen und der Mittelwert aller Wörter jeder Textsorte mit den Werten der anderen Textsorten verglichen.

Dabei stellte sich heraus, daß bis zu ugf. 6 Zeichen die Textsorte *Bericht* höhere Werte als die Textsorten Einleitung und Rezension aufwiesen (vgl. Tabelle 13), sich die Relation aber ab 9 Zeichen klar zugunsten der Textsorten *Einleitung* und *Rezension* umkehrte (vgl. Tabelle 13 und Tabelle 14), wobei ab 12 Zeichen der Bericht proportional viel weniger Wörter dieser Länge enthält und die Werte der *Einleitung* und der *Rezension* vergleichbar sind. In beiden Tabellen sind die Ergebnisse gruppiert in 4-Zeichen-Einheiten, so daß kleinere Schwankungen ausgeglichen werden und die allgemeine Tendenz klarer dargestellt werden kann.

Histogramm. Vergleichende Wortlänge nach Textsorten im Muttersprachlerkorpus (Werte in ‰)

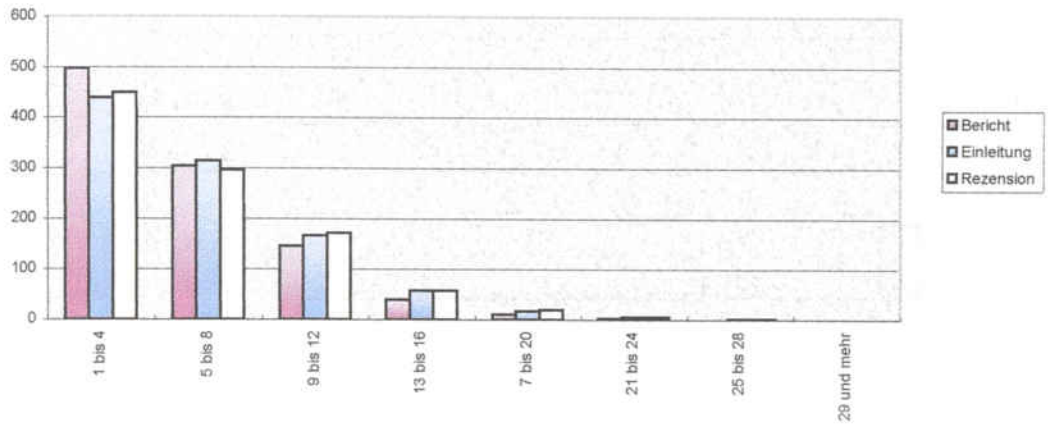


TABELLE 13

Diese Verteilung der Frequenzen scheint die Vermutung zu bestätigen, daß formal stärker definierte Textsorten, wie die Einleitung und die Rezension, auch eine höhere Proportion an längeren Wörtern aufweisen als Textsorten, die stilistisch relativ frei sind wie der Bericht.

Histogramm Vergleichende Wortlänge nach Textsorten im Muttersprachlerkorpus. Wörter ab 13 Zeichen

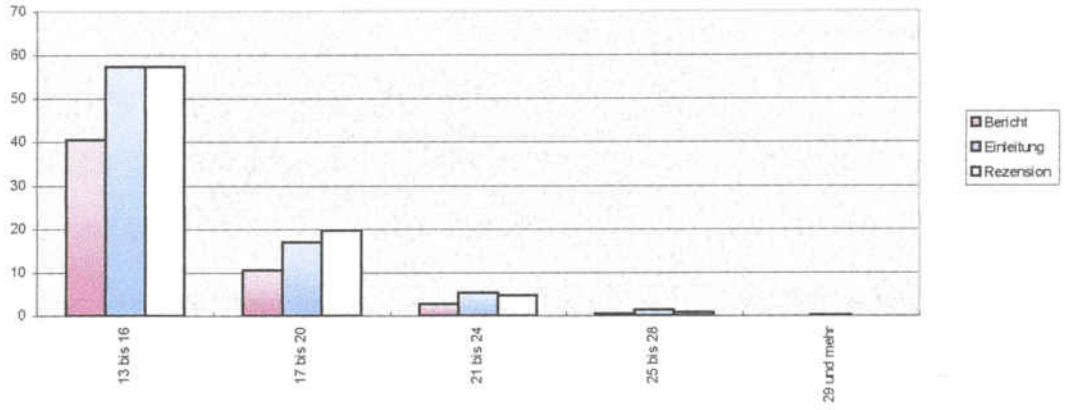


TABELLE 14

Zum Vergleich der Wortlänge zwischen dem CLK und dem CMK wurden die Werte nun so verglichen, daß auch die Subkorpora des CLK berücksichtigt wurden (Mittelstufe, MS und Oberstufe, OS). Wie schon in dem Textsortenvergleich des CMK ist auch hier zu erkennen, daß sich eine Gruppe in den niedrigen Werten der Wortlänge über die andere auszeichnet, dafür aber bei den höheren Werten zurückbleibt: das CLK weist höhere Werte im Bereich bis zu 6 Zeichen auf, mit einem überproportional hohen Wert in Subkorpus MS (vgl. Tabelle 15).

Histogramm Wortlänge CLK und CMK: Wörter zwischen 1 und 30 oder mehr Zeichen (in 0/00)

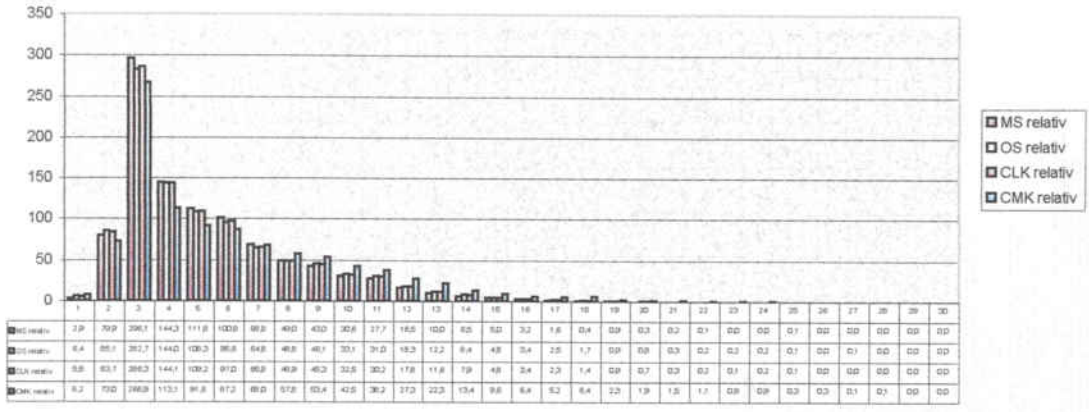


TABELLE 15

Die Untersuchung der längeren Wörter (ab 10 Zeichen; vgl. Tabelle 16) ergibt, daß die Werte des CMK einen immer größeren Abstand zu dem CLK aufweisen; dabei weist zwar auch das Subkorpus OS größere Werte auf, befindet sich aber in jedem Fall dem Subkorpus MS näher als dem CMK.

Histogramm Wortlänge CLK und CMK: Wörter zwischen 10 und 30 oder mehr Zeichen (in ‰)

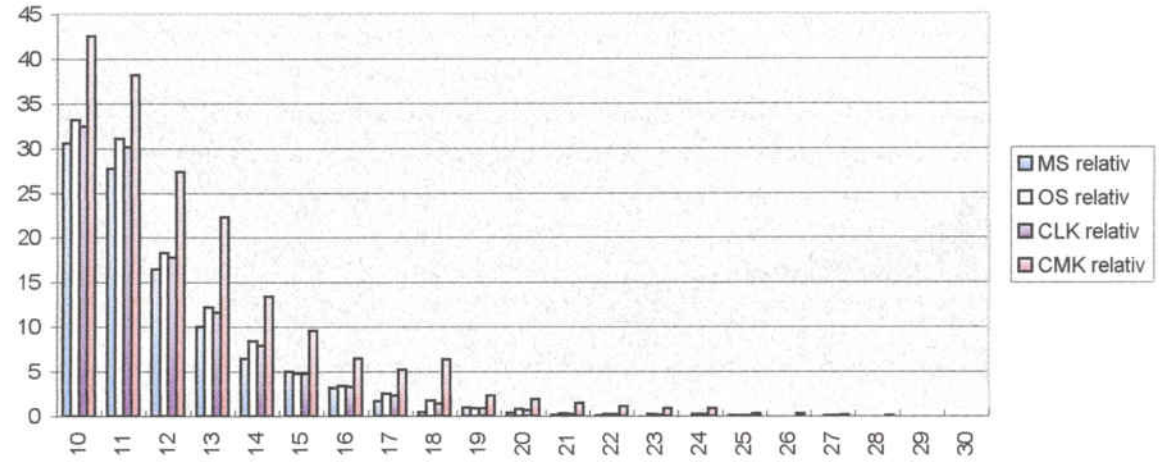


TABELLE 16

Die generelle Tendenz wird noch klarer, wenn man ausschließlich das CLK mit dem CMK vergleicht, ohne Einteilung des CLK in die Subkorpora MS und OS. Anhand der Verteilung beider Korpora auf einer 100%-Basis wird erkenntlich, daß das CLK bis zu 6 Zeichen pro Wort höhere Werte als das CMK aufweist, bei 7 Zeichen gleiche Werte besitzt und ab 8 Zeichen stetig sinkt, wobei die Schwankungen im höheren Bereich (hier 27 Zeichen) auf die extrem niedrigen Frequenzen zurückzuführen sind, die sich im Bereich zwischen 0 und 5 Wörtern mit mehr als 25 Zeichen pro Tausend bewegen.

Histogramm Wortlänge CLK und CMK: Verteilung der Wortlänge in Zeichen auf die Summe der Korpora

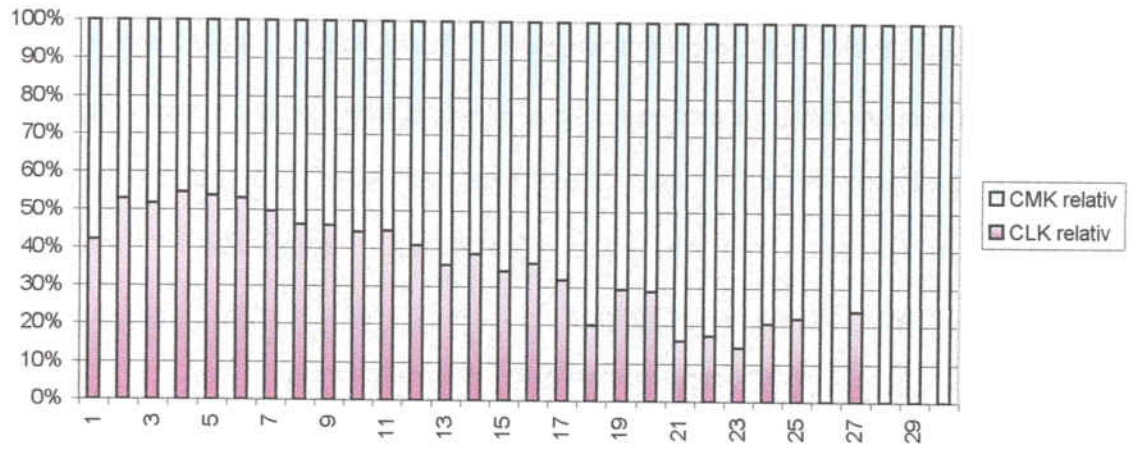


TABELLE 17

4.2.1.2 Satzlänge

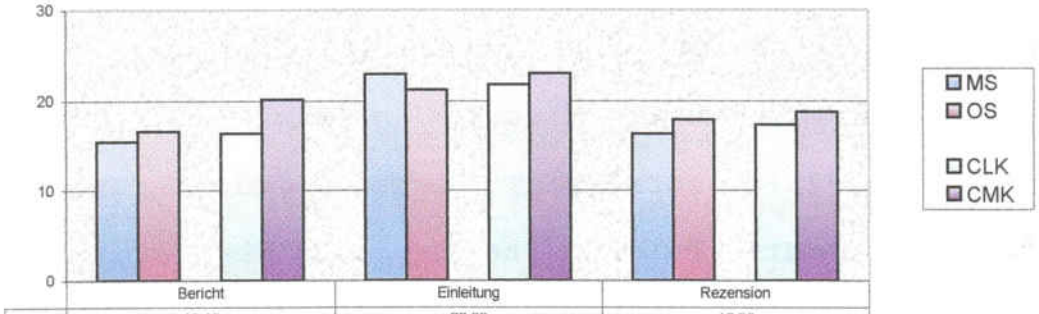
Satz-Einheiten werden definiert als ausschließlich von den Satzzeichen Punkt, Ausrufezeichen und Fragezeichen abgregrenzte Zeichenketten (vgl. Leidner 1997); in diesem Sinne handelt es sich um eine orthographische und keine syntaktische Unterscheidung. Manuell mußten bei dieser Perspektive jene Fälle korrigiert werden, bei denen die Satzzeichen eine andere Funktion hatten als die Abgrenzung eines Satzes, wie zum Beispiel die Anwendung des Punktes für Abkürzungen[61].

[61] Zur automatischen Bestimmung von Sätzen, vgl. Grefenstette/Tapanainen (1994: 4). Zur Abgrenzung von Punkt als Satzzeichen, vgl. Oppenheim (1988), Burrows (1987a: 213-216).

Die Berechnungen der Satzlänge basieren auf der Kombination der Textversion und der POS-Version der Korpora. Die Textversion diente als Grundlage für die Bestimmung der Wortzahl eines jeden Textes, das POS-Tag </s>, mit dem das Satzende markiert wird, zur Bestimmung der Satzzahl. Danach wurde die Wortzahl durch die Satzzahl dividiert und das Ergebnis als Satzlänge angegeben.

Zur Bestimmung der Eigenschaften der Satzlänge wurde zunächst die mittlere Satzlänge zwischen den Textsorten des CMK berechnet, was die Grundlage für den darauffolgenden Vergleich mit den Werten des CLK bildet (vgl. Tabelle 18). Auch in folgenden Vergleichen wird das CMK als Kontrollkorpus herangezogen (vgl. S. 147).

Vergleichende Satzlänge in den Textsorten: Mittlere Gesamtwerte und Standardabweichung



	Bericht	Einleitung	Rezension
MS	15,42	22,93	16,28
OS	16,57	21,19	17,89
CLK	16,34	21,78	17,33
CMK	20,09	22,98	18,74

TABELLE 18

Die Berechnung der Werte nach Textsorten im CMK ergab, daß die mittlere Satzlänge der einzelnen Texte starken Schwankungen unterliegt (vgl. Tabelle 19); im Falle der Einleitungen geht sie von 12 bis 35, der Berichte von 14 bis 25 und der Rezensionen von 12 bis 31.

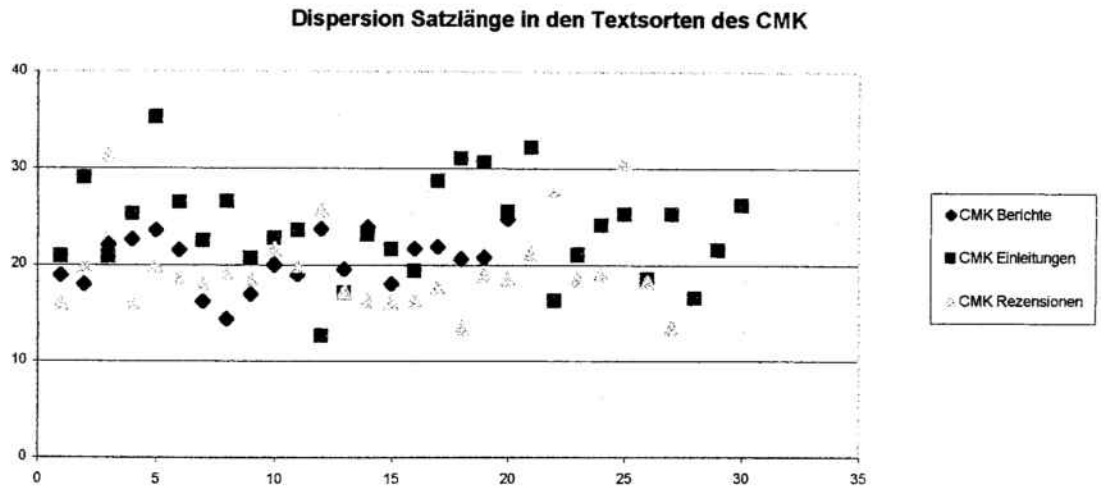


TABELLE 19

Die stärksten Schwankungen sind in den Einleitungen zu erkennen. Die mittleren Werte für die Textsorten jedoch spiegeln Differenzen wider, die auf Konstanten in den unterschiedlichen stilistischen Phänomenen zurückzuführen sind; auffallend sind dabei die niedrigen Werte der Rezension, die mit 19,52 Wörtern pro Satz deutlich unter dem Bericht (22,60 Wörter/Satz) und der Einleitung (23,61 Wörter/Satz) liegt, die vergleichbare Werte erzielen.

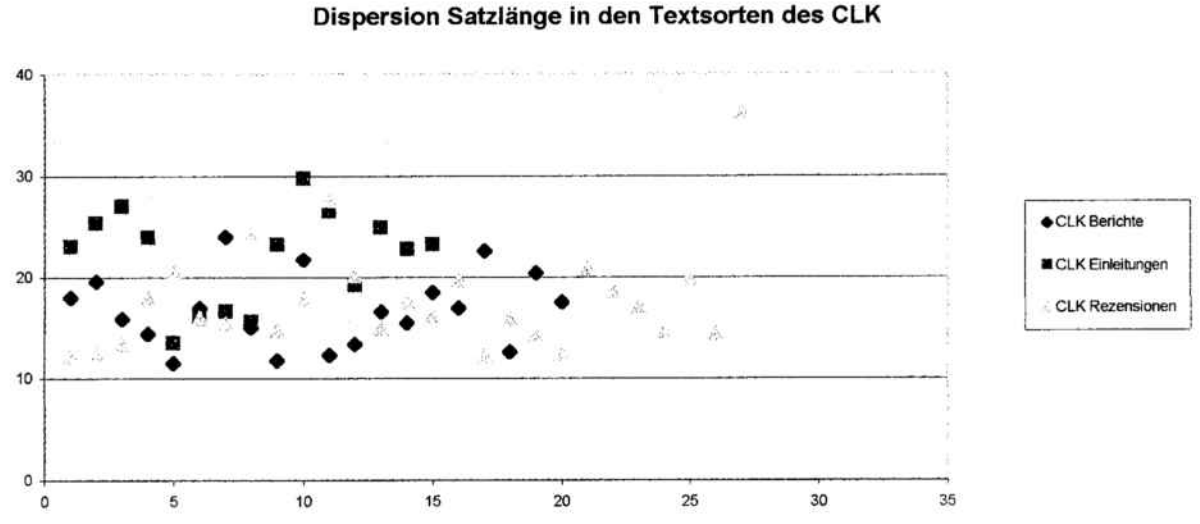


TABELLE 20

4.2.1.3 Rekurrente Wortkombinationen (Cluster)

Rekurrente Wortkombinationen, auch *Cluster* genannt, sind „kontinuierliche Wortketten, die mehr als einmal identisch erscheinen" (Altenberg 1993: 227). Maschinell werden sie wie normale Wortlisten erstellt, die nicht aufgrund der Frequenz eines orthographischen Wortes berechnet werden, sondern anhand der Verbindung von beliebig vielen Wörtern. So entstehen Frequenzlisten aus Mehrworteinheiten, die zur Überprüfung des „idiom principle" Sinclairs (1991) dienen. Nach diesem Prinzip verwendet ein Sprecher in seiner mündlichen Produktion eher idiomatische Formeln, anstatt immer wieder neues Material zu schaffen (vgl. auch Cock et al. 1998: 67), was sich in

Kjellmers für französische Englischlerner zugeschnittene Hypothese widerspiegelt, nach der der fremdsprachliche Charakter der Ausdrucksweise eines Lerners auf die Verwendung von individuellen Bausteinen und nicht von vorgefertigten Elementen zurückzuführen ist (vgl. Kjellmer 1991: 124). Nach Altenberg (1993: 227), der diese idiomatischen Formeln als „phraseologische Komponenten" bezeichnet, ist dies auch ein Phänomen des geschriebenen Diskurses.

Für die vorliegende Untersuchung haben wir Frequenzlisten für Zweiwortcluster und Dreiwortcluster zusammengestellt und sie auf feste Verbindungen untersucht. Größere Cluster ergeben ähnliche Muster wie die Zweiwortcluster und die Dreiwortcluster, so daß sie hier nicht dargestellt werden. Die Zahl der Fehltreffer, d.h. Verbindungen, die z.B. nur aus Funktionswörtern (wie Hilfsverb und Artikel in *hat einen*) bestehen, erschweren zusammen mit der Vielfalt der so gefundenen Phänomene die systematische Untersuchung der Listen. Hinzu kommt, daß hier aufgrund der im Vergleich zur mündlichen Produktion niedrigen Anzahl an Treffern alle maschinell gefundenen Cluster dargestellt werden, und keine semantisch bedingte Auswahl wie die Reduzierung auf Ausdrücke des Zögerns vorgenommen wird (vgl. Cock et al. 1998: 74).

Auch ist zu beachten, daß Kontraktionen, wie die Verbindung einer Präposition mit einem Artikel (z.B. *ins* oder *im*), nicht als Cluster bearbeitet werden, obwohl die gleiche Struktur mit einem anderen Genus als Cluster berechnet würde (z.b. *in der*).

Zweiwortcluster

Die Untersuchung der Frequenzliste der Zweiworteinheiten des CMK ergibt, daß an den ersten Stellen Verbindungen von Präposition plus Artikel erscheinen (*in der* mit einer relativen Frequenz von 0,34‰, gefolgt von *in den, für die, auf die* usw.), mit der Ausnahme allerdings der zweiten Stelle, die eine Verbindung von der hochfrequenten Konjunktion *und* mit einem Artikel darstellt (*und die*). Erst an achter Stelle erscheint die Verbindung von einem Modalverb mit einem Pronomen in invertierter Stellung (*kann man*), an neunter ein Substantiv mit Artikel (*das Buch*), ein starkes Indiz für die Einheitlichkeit der Themen der Texte. Erst an 20. Stelle erscheinen klar erkennbare argumentative Konnektoren, wie *wenn man* und an 32. Stelle *so daß*. Die danach erscheinenden Einträge wiederholen die schon genannten Schemata ohne größere Variationen, wobei noch an der 71. Stelle die reflexive Formel *ich mich* anzumerken ist. In Tabelle 21 sind die hier besprochenen Phänomene zusammengefaßt.

CMK Zweiwortcluster	CMK Absolut	CMK Relativ
IN DER	137	3,34
UND DIE	71	1,73
IN DEN	67	1,63
KANN MAN	52	1,26
DAS BUCH	51	1,24
GIBT ES	40	0,97
WENN MAN	34	0,83
SO DAß	29	0,70
ICH MICH	17	0,41

TABELLE 21

Die ersten beiden Stellen des CLK stimmen mit denen des CMK überein: *in der* und *und die* sind auch im CLK die häufigsten Cluster. Während jedoch im CMK die anderen Spitzenstellungen von einer Präposition mit Artikel besetzt waren, erscheint im CLK an dritter Stelle die verbale Form *gibt es* in invertierter Stellung, direkt gefolgt von der normalen Stellung *es gibt*; im CMK erscheint die invertierte Form erst an 16. Stelle, die normale sogar erst an 59. Stelle. Nur ein weiterer Eintrag des CLK stimmt ebenfalls mit dem CMK überein, *kann man*. Alle anderen weichen ab: das Verb *sein* als Vollverb erscheint im CLK an 5. Stelle, im CMK an 15. Stelle. Das Substantiv im CLK erscheint auf Platz 15, doch mit ähnlichen Frequenzen wie *das Buch* im CMK, was als Indiz für eine Abweichung hinsichtlich der Themen der Texte interpretiert werden kann. Als präzisierendes Element erscheint im CLK *nicht nur* auf Platz 36, mit vergleichbaren Frequenzen zu dem Konnektor des CMK (*wenn man*). Und während die erste Infinitivkonstruktion des CLK an 40. Stelle erscheint, tut sie es im CMK, *zu können*, erst an 71. Stelle mit annähernd der Hälfte an Treffern, 17. Vergleichbares geschieht mit der ersten reflexiven Konstruktion im CLK: mit einer Frequenz von 0,70‰ liegt sie deutlich über der ersten vergleichbaren Konstruktion im CMK, die einen Wert von 0,41‰ aufweist.

CLK Zweiwortcluster	CLK Absolut	CLK Relativ
IN DER	151	3,44
UND DIE	85	1,32
GIBT ES	56	1,27
KANN MAN	55	1,25
IST EIN	50	1,14
DIE STUDENTEN	48	1,09
NICHT NUR	33	0,75
ZU MACHEN	32	0,73
MAN SICH	31	0,70

TABELLE 22

Die Untersuchung der Zweiwortcluster auf vertikaler Ebene zwischen den beiden Korpora zeigt, daß einige Elemente konstant in beiden Korpora erscheinen, diese aber ihre Spitzenstellung nicht aufgrund der Einheit belegen, die sie darstellen, sondern weil beide getrennt ebenfalls hochfrequente Wörter der Frequenzlisten sind. Zweiwortcluster, die eine semantische Einheit bilden, sind in den ersten Positionen kaum vertreten. Die einzigen Ausnahmen bilden das Substantiv *Studenten* mit seinem Artikel und ein Teil einer zweiteiligen Konjunktion, *nicht nur [sondern auch]*.

Dreiwortcluster

Im CMK werden die ersten Stellen von Verbindungen zwischen Substantiven und Präposition besetzt, sei es ein Substantiv gefolgt von der von ihm valenzbedingten Präposition (*Einführung in die, Überblick über die, die Frage nach*) oder eine Präpositionalgruppe allein (*An der Uni*). Danach erscheinen schon unter den ersten 25 Einträgen mehrere

lexikalisierte Wendungen, wie *im Rahmen des, im Anschluß daran,* oder *in der Regel, in diesem Zusammenhang, auf jeden Fall* oder *in bezug auf,* und ebenfalls einige Verben mit den von ihnen valenzbedingten Präpositionen (*beschäftigt sich mit*).

Weitere Wiederholungen ergeben sich aus den schon in den Zweiwortclustern erwähnten Phänomenen, zu denen dann noch ein weiteres Element in der Liste hinzugefügt wird, was die Grundstruktur aber nicht ändert.

Im CLK wird die erste Stelle von dem Cluster *in der Nähe* besetzt, gefolgt von einem ersten formelhaften Ausdruck, *meiner Meinung nach.* Darauf folgen wie im CMK Präpositionalgruppen (*an der Universität*), aber keine Substantive mit der von ihnen regierten Präposition bis zur 32. Stelle (*Bezug auf die*), viel später als im CMK. Auch Verben mit ihren valenzbedingten Präpositionen erscheinen nicht unter den ersten 50 Treffern, und nur ein argumentativer Konnektor (*aus diesem Grund*), der hinsichtlich seiner strukturierenden Funktion mit den formelhaften Ausdrücken des CMK zu vergleichen ist.

Hier kann also deutlicher noch als bei den Zweiwortclustern erkannt werden, daß die Einheit, die Lerner für ihre Textproduktion vorziehen, das einzelne Wort ist, und daß in statistisch weniger Fällen als im CMK Mehrwortverbindungen Anwendung finden.

Während die Zweiwortcluster kaum lexikalische Einheiten erkennen ließen, erscheint bei den Dreiwortclustern eine hohe Zahl an formelhaften Ausdrücken und komplexeren lexikalischen Einheiten, wie z.B. Konnektoren. Hier zeigt sich ebenfalls, daß die Texte des CLK weniger formelhafte Verbindungen verwenden und in diesen weniger Substantive oder Verben mit entsprechender präpositionaler Rektion erscheinen.

4.2.1.4 Type-Token-Ratio: lexikalischer Variationsindex

Unter *Token* wird hier die Gesamtanzahl der in einem Text vorkommenden Wörter[62] verstanden, unter *Types* die Anzahl der sich wiederholenden Wörter[63]. Der Berechnung der Types liegen den vom ImS-Tagger lemmatisierten Einträgen zugrunde, die einer manuellen Revision unterworfen wurden. Die Bearbeitung und Berechnung des Datenmaterials erfolgte maschinell, was aufgrund verschiedener Phänomene zu leichten Verzerrungen der Ergebnisse führen kann. Die Einschränkungen dieser Vorgehensweise sind folgende:

- Ohne manuelle Disambiguierung des semantischen Gehalts eines Wortes und evtl. seiner Etymologie (vgl. Clément 1996: 23f) ist der Computer nicht in der Lage, Homonymie und Polysemie

[62] Vgl. Definition von Wort in 4.2.1.1.

[63] Vgl. McEnery (1996a: 67) und die Hilfefunktion von WordSmith 2.0 unter dem Eintrag „Type-token ratio"

zu unterscheiden und die Realisierung eines Wortes als verschiedene Grundformen zu bearbeiten (Vgl. Barnbrook 1996: 58 und 60). Homonymische (homographische) und polysemische Wörter werden somit immer als ein Type berechnet. Clément fügt zu dem Problem noch das der Synonymie hinzu, weil bei unterschiedlicher orthographischer Form und gleicher Bedeutung immer verschiedene Wörter bei der Zählung angegeben werden (Vgl. Clément 1996: 24).

- Wortbildungen werden immer als ein Type angegeben, und nicht als zwei oder evtl. mehr.
- Mit Bindestrich getrennte Types konnten entweder nur als Einzelwörter oder nur als getrennte Wörter interpretiert werden. Im Einklang zur Definition von Wort wurden sie ebenfalls immer als ein Type berechnet.
- Zahlen wurden als ein Type berechnet, weil die einheitlicher Ansicht vertreten wurde, daß sie sowohl ausgeschrieben als auch in Ziffern den gleichen semantischen Wert darstellen.
- Kontraktionen (*im, zum* usw.) werden im Einklang mit der Definition von Wort als Type behandelt, auch wenn sie sich aus zwei Types zusammensetzen, die ebenfalls im Text vorkommen können und dann getrennt berechnet werden.
- Eigennamen wie *New York* oder *Universität Heidelberg* werden als jeweils zwei Types berechnet.
- Mehrwortlexeme werden immer als verschiedene Types behandelt. So z.B. idiomatische Redewendungen, die als stabile oder relativ stabile Verbindungen von Wörtern aufgefaßt werden können[64].

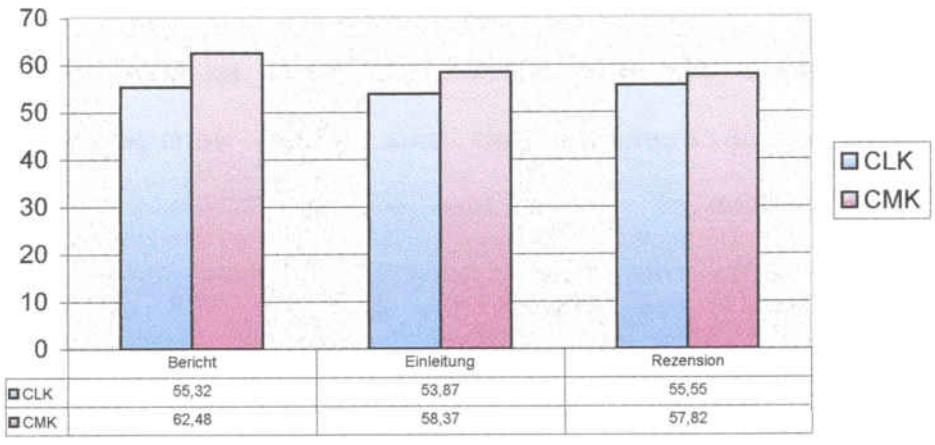
[64] Vgl. Bußmann (1990: 320f) und Wotjak (1992: 3f).

Zur Berechnung der Type-Token-Ratio (TTR) wurden die Texte in Segmente à 200 Wörter eingeteilt und dann der Mittelwert für den jeweiligen Text berechnet und graphisch als Streudiagramm dargestellt; beide Werte werden in eigenen Diagrammen dargestellt. Diese Segmentierung in 200-Wort-Einheiten wurde wegen der starken Abhängigkeit der TTR von dem Textumfang vorgenommen. Tuldava (1995: 133) weist darauf hin, daß ein direkter Vergleich nur zwischen gleich langen Texten möglich ist, denn „the relation between the size of text and the size of vocabulary does not remain unchanged at any value of text size. The size of the text (*N*) increases faster than the size of vocabulary (*L* or *V*)". Die Textlänge beeinflußt also stark die TTR, wie den Beispielen Martin Volks (Volk 1998) entnommen werden kann: in Smiths *Computers and Human language*, Kapitel 3, kommen 2.427 Types auf 13.000 Token (TTR = 5,35), im Scanworx-Manual für das Deutsche 3.700 Types auf 48.000 Token (TTR = 12,97), im englischen Brown Korpus ugf. 50.000 Types auf 1 Millionen Token (TTR = 20), im deutschsprachigen Zeitungskorpus *Die Welt* 166.484 Types auf 2,5 Millionen Token (TTR = 15,01).

Die uneinheitliche Länge der Texte der Computerkorpora CMK und CLK hätte so bei einer auf der Gesamtlänge basierten Berechnung der Ratio zu Werten geführt, die nicht untereinander vergleichbar gewesen wäre. Die Segmentierung jedes Textes erlaubt jedoch einen Vergleich auf der Grundlage von vergleichbaren Einheiten. Nach der Ermittlung der TTR für

jedes Fragment wird der mittlere Wert berechnet, der dann als TTR für die spezifizierte Segmentlänge des bearbeiteten Textes angegeben wird. Der Wert 200 wurde gewählt, weil er die untere Grenze der Textlänge darstellte und nur sehr wenige Texte diesen Wert unterschritten.

Der horizontale Vergleich des CMK zeigt, daß die TTR größeren Schwankungen hinsichtlich der Textsorten unterliegt (vgl. Tabelle 23), mit hohen Werten im Bericht und sich stark ähnelnden, niedrigeren Werten in der Einleitung und in der Rezension. Dies spiegelt sich aber nicht im horizontalen Vergleich des CLK wider: die TTR liegt hier relativ einheitlich zwischen 55,55 und 53,87, allerdings mit größeren intertextuellen Schwankungen (vgl. Tabelle 25). Der vertikale Vergleich zwischen den beiden Korpora weist außerdem stetig eine größere lexikalische Variation im CMK als im CLK auf.

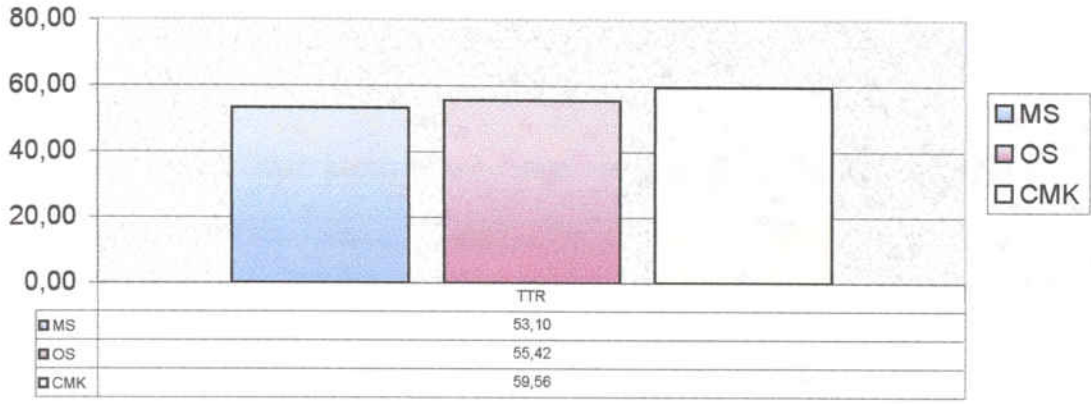


	Bericht	Einleitung	Rezension
CLK	55,32	53,87	55,55
CMK	62,48	58,37	57,82

TABELLE 23

Wird die TTR der Subkorpora mit einbezogen, tritt eine Steigerung des Wertes von MS (53,10) über OS (55,42) bis hin zu CMK (59,56) auf: je fortgeschrittener der Sprachstand, desto größer ist der Wert der TTR (vgl. Tabelle 24). Diese Tendenz wird nur von den Einleitungen des Subkorpus MS unterbrochen, die aufgrund ihres niedrigen Repräsentationsgrades als Ausnahme zu gelten haben.

TTR der Textsorten in CMK und CLK. Mittelwertberechnung aufgrund von Segmenten à 200 Wörtern



	TTR
MS	53,10
OS	55,42
CMK	59,56

TABELLE 24

Anhand des Streudiagramms (vgl. Tabelle 25) ist klar die große Variation der Werte zu erkennen. Die Texte des Subkorpus MS zeichnen sich durch die Werte im unteren Bereich aus, die des CMK durch Werte im höheren. Die individuellen Abweichungen lassen sich nur qualitativ erklären, da der Inhalt des jeweiligen Textes, seine thematische Entfaltung und andere Faktoren Einfluß auf die lexikalische Variation nehmen.

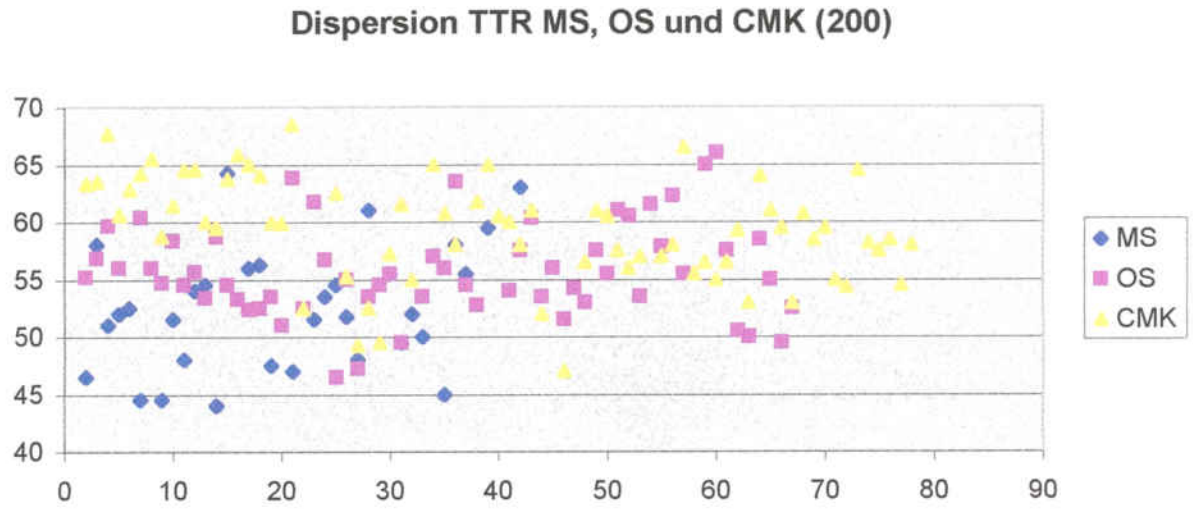


TABELLE 25

Die TTR ist durchschnittlich ein steigender Wert, vom Subkorpus MS über OS bis hin zu CMK. Die Werte zeigen nur im CMK eine deutliche Differenz auf, während sie relativ stabil im CLK sind. Diese allgemeinen Tendenzen würden sich bei größeren Segmenten als Grundlage zur Berechnung stabilisieren oder sogar noch deutlicher hervortreten.

4.2.1.5 Affixe

Die Affigierung ist eine der wichtigsten Methoden der Wortbildung, zusammen mit dem Ablaut, der Reduplikation, der Zusammensetzung; weniger wichtig sind Abkürzungen und Kontamination (Hentschel/Weydt 1990: 21ff). Hier soll die Anwendung von Präfixen und Suffixen hinsichtlich der Frequenz ihrer Anwendung untersucht werden, wobei Präfixe und Suffixe

je nach Wortart, bei der sie angewendet werden, eingeteilt wurden, so daß die Suche nach der Kombination von Präfix oder Suffix und Wortartentag erfolgen konnte. Dabei wurden nur jene Affixe untersucht, die nicht getrennt erscheinen; im Falle der verbalen Präfixe kann auf ein eigenes Wortartentag verwiesen werden, PTKVZ, mit dem getrennt erscheinende Präfixe ausgezeichnet werden (vgl. 4.2.2.10).

Die Untersuchung des Phänomens ist von Interesse zur Bestimmung der produktiven Fähigkeiten und Gewohnheiten der Lerner, dient aber auch dem Ziel, Regelmäßigkeiten und Frequenzen ihrer Benutzung im CMK darzustellen, denn Voruntersuchungen an den Preliminarversionen der Computerkorpora führten zur Annahme, daß die in Lehrwerken eingeführten Elemente mehr didaktische Ziele verfolgten als die Anwendung der Elemente im reellen Sprachgebrauch wiederzugeben.

Die hohe Zahl an produktiven Präfixen und Suffixen, vor allem im verbalen Bereich, führte zur Entscheidung, für diese Untersuchung nur die häufigsten zu untersuchen und die weniger frequenten außer acht zu lassen.

Als Grundlage für die Einteilung wurde eine Verteilung nach assoziierten Wortarten unternommen, was eine detailliertere Differenzierung zuließ. Aufgrund der so erhaltenen quantitativen Beschreibung ihrer Anwendung konnten eventuell

weitere Auswertungen vorgenommen werden, wie z.B. eine semantische Analyse ihrer Anwendung.

Für die Präfixe wurde eine Liste zusammengestellt, in der die in Grammatiken aufgezeichneten Präfixe aufgenommen wurden.

Auf der Grundlage des lemmatisierten CMK wurde zunächst anhand der alphabetisch geordneten Wortliste untersucht, welche Präfixe wiederholt in dem Korpus erschienen, so daß sie dann in eine provisorische Liste aufgenommen wurden. Danach wurden die Suffixe auf eine ähnliche Weise ermittelt, allerdings mit einer alphabetischen revers geordneten Liste, d.h. die nicht nach dem Anfang des Wortes geordnet war, sondern nach dem Ende.

Für all die so ermittelten Präfixe und Suffixe wurde dann eine Konkordanzliste im POS-getaggten CMK (s.o.) erstellt; als Kriterien für die Einbeziehung in die schließlich zu untersuchende Liste galt, daß das Präfix oder Suffix mindestens 4 verschiedene Wörter betreffen mußte und eine Mindestfrequenz von 0,20‰ im CMK aufzuweisen hatte. Danach wurden die Präfixe für ihre endgültige Untersuchung den mit ihnen auftretenden Wortarten zugeschrieben, so daß ein Präfix mehrmals in der ganzen Liste erscheinen kann.

Die endgültige Liste der untersuchten Elemente bestand aus folgenden Präfixen und Suffixen:

Substantive		Adjektive		Verben	
Präfix	Suffixe	Präfixe	Suffixe	Präfixe	
Um-	-heit	un-	-lich	an-	hin-
	-keit	über-	-isch	auf-	mit-
	-ismus		-bar	aus-	nach-
	-logie		-ell	be-	um-
	-nahme		-haft	durch-	unter-
	-nis		-ig	ein-	über-
	-tion		-los	ent-	ver-
	-ung		-sam	er-	vor-
			-tiv	ge-	weiter-
				heraus-	zu-
				hervor-	zusammen-

TABELLE 26

Allgemeines

Nach der Erstellung der Konkordanzlisten wurden zunächst aus diesen alle Einträge gestrichen, bei denen das Präfix oder Suffix nicht dem gesuchten entsprach[65]. Im Falle eines zusammengesetzten Wortes wurde es nur berücksichtigt, wenn das zusammengesetzte Wort mit dem entsprechenden Präfix oder Suffix (oder seiner Flexionsform) endete. Die danach in der Liste verbleibenden Einträge bilden dann den absoluten Wert, der als Grundlage für die Berechnung der Frequenzen in ‰ diente. Aufgrund der unterschiedlichen Länge der Korpora erscheint nicht der absolute Wert in den Tabellen, sondern nur der relative.

Zur Berechnung der Variation wurden alle sich wiederholenden Einträge gestrichen, wobei unter Wiederholung auch

[65] Die Suche nach dem Präfix *un-* z.B. brachte auch alle Präfixe *unter-* hervor, die dann manuell gelöscht werden mußten.

Flexionsformen verstanden wurden. Beibehalten wurden auch alle Wortzusammensetzungen, obwohl in manchen Fällen das gesuchte Präfix oder Suffix schon in einer anderen Wortzusammensetzung erschien[66]. Der Variationswert gibt so an, wie viele verschiedene Formen des untersuchten Präfixes oder Suffixes pro tausend Wörter im Text erscheinen und ist somit ein Indikator für den Grad der lexikalischen Variation und bildet, im besonderen Falle der Präfixe und Suffixe, die Grundlage für spätere qualitative Untersuchungen zur textuellen Kreativität hinsichtlich der untersuchten Elemente.

Zur Darstellung des Vergleiches zwischen CLK und CMK wurde den relativen Werten und den Werten der Variation noch ein vergleichender Wert auf prozentueller Basis beigegeben, die Abweichung. Sie wird berechnet auf der Grundlage des Richtwertes CMK, der als Referenzkorpus betrachtet wird und den Wert 0 darstellt, im Vergleich zu den Werten, die im CLK erzielt werden. Daraus ergibt sich eine prozentuelle Abweichung, die positiv oder negativ sein kann (positiv im Falle der Überrepräsentation des Phänomens im CLK, negativ im Falle der Unterrepräsentation im CLK). Die Berechnung erfolgt anhand der Formel [CLK * 100 / CMK - 100].

Als Beispiel nehmen wir einen Wert 10 im CMK. Wenn das CLK einen Wert 20 erreicht, wäre folgende Berechnung

---

[66] So erscheint in der Liste des Suffixes -lich *wirklich*, aber auch *unwirklich*.

durchzuführen: [20 * 100 / 10 - 100 = 100]. Der Wert stellt eine positive Abweichung dar, weil die Werte des CLK höher als die des CMK sind.

Bei einem Wert von 5 im CLK wäre folgende Berechnung anzuwenden: [5 * 100 / 10 - 100 = -50]. Hier ist der Wert negativ, weil der Wert des CLK unter dem des CMK liegt.

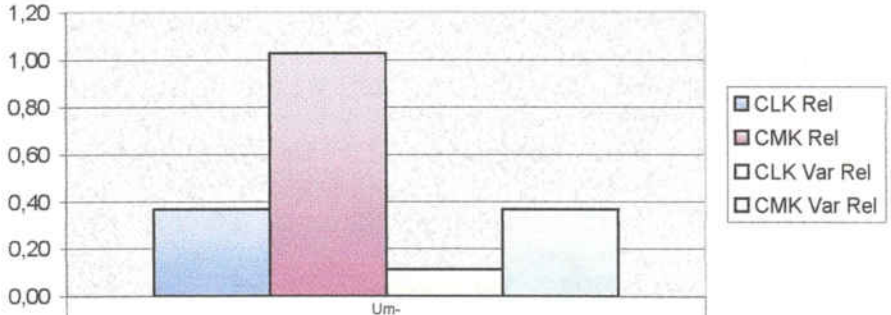
In dieser Berechnung wird am Ende nur die Abweichung (AB) des CLK dargestellt. Die Abweichung des CMK ist dabei implizit in der von der Darstellung von CLK enthalten, weil das CMK die Grundlage der Berechnung bildet. Bei der Darstellung ist zu beachten, daß die Deutung positiver und negativer Werte nicht vergleichbar ist. Positive Abweichungen sind offen, d.h. das CLK kann ein Phänomen *n* Mal verwenden, wobei *n* irgendeine positive Zahl darstellen kann. Negative Abweichungen hingegen sind auf einer Skala von 0 bis 99,99 Periode zu deuten. Bei einem Wert von 10 im CMK und 10 im CLK wird eine Nullabweichung von 0% erreicht, es gibt also keine Abweichung. Bei einem Wert von 10 im CMK und 5 im CLK wird eine Abweichung von -50% berechnet, bei einem Wert von 10 im CMK und 2 im CLK eine Abweichung von -80%, bei einem Wert von 10 im CMK und 1 im CLK eine Abweichung von 90%, bei einem Wert von 10 im CMK und 0,1 im CLK eine Abweichung von -99%.

Substantive: Präfixe

Das einzige substantivische Präfix, das anhand der Konkordanzlisten gefunden wurde und die Einschränkung von einer Frequenz von mindestens 0,20‰ überschritt und mindestens in vier verschiedenen Kombinationen auftrat, war *Um-*. Anhand der Konkordanzliste wurden zwar andere frequente Präfixe gefunden (*Ab-* mit einer relativen Frequenz von 1,65‰, z.B.), doch waren sie ausschließlich verbalen Ursprungs und wurden hier nicht betrachtet.

Tabelle 27 ist zu entnehmen, daß im CMK das Präfix *Um-* mehr als dreimal so oft angewendet wird wie im CLK, und die Variationswerte des CMK ebenso deutlich über denen des CLK liegen.

Substantivische Präfixe: Um-



	Um-
CLK Rel	0,37
CMK Rel	1,03
CLK Var Rel	0,11
CMK Var Rel	0,37

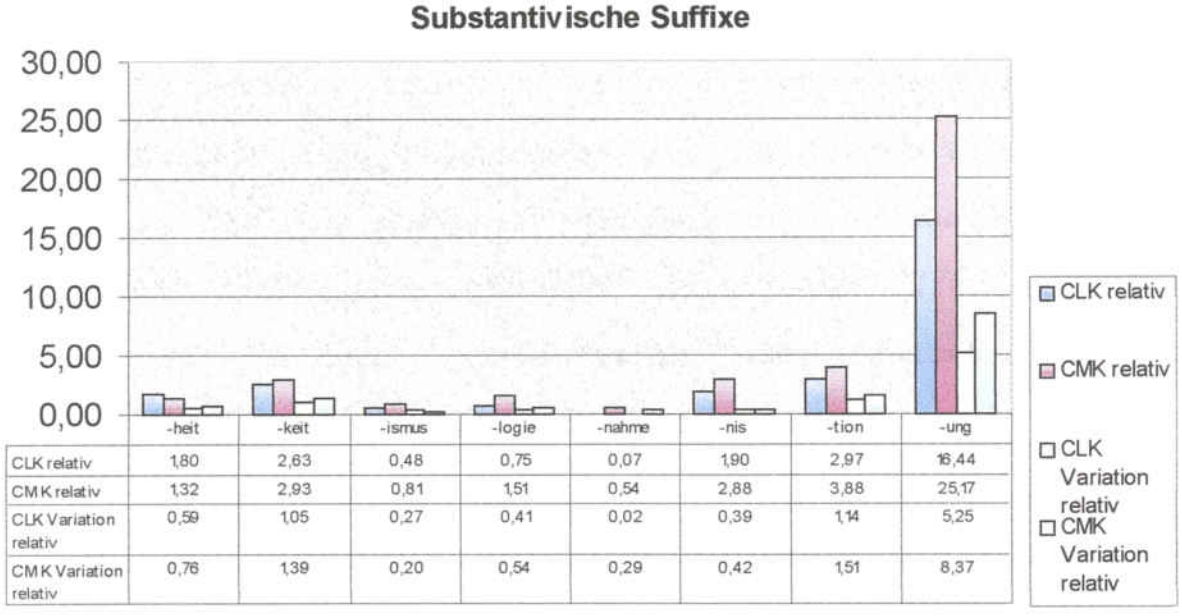
TABELLE 27

Aufgrund der reduzierten Anzahl an sich wiederholenden substantivischen Präfixen für Substantive sind die in Tabelle 27 angegebenen Werte nicht quantitativ systematisierbar und müssen später in Zusammenhang mit den anderen Wortarten analysiert werden. Für eine spätere qualitative Untersuchung ist hier der Vergleich zu dem substantivischen Präfix *Selbst-* zu erwähnen. Es erscheint nicht in der Untersuchung, weil es die genannten Mindestwerte nicht erreichte, befand sich aber an der Grenze derselben. Die Anwendung dieses Präfixes zeigt nur noch einen leicht höheren Gebrauch im CMK an; eine Hypothese ist, daß sich dieser Unterschied aus der Identifizierbarkeit des semantischen Inhalts der beiden Präfixe ergeben könnte, wobei die präzisere Definition von *Selbst-* durch den Lerner zu vergleichbaren Werten führen würde, während die Ungenauigkeit und Bedeutungsvariation von *Um-* inhibitorisch auf den Lerner wirken könnte.

Substantive: Suffixe

Substantivische Suffixe zeigen starke Schwankungen hinsichtlich ihrer Anwendung in beiden Korpora auf (vgl. Tabelle 28); die Werte sind in fast allen Fällen höher im CMK als im CLK. Die allgemeine Tendenz ist, daß das CMK höhere Werte sowohl im relativen Wert als auch in der Variation aufzeigt. Es werden demnach mehr Suffixe angewendet und gleichzeitig ist die Variation derselben größer.

Ausnahmen bilden nur der relative Wert des Suffixes *-heit*, das öfter im CLK als im CMK verwendet wird, und der Variationswert des Suffixes *-ismus*, der ebenfalls höher im CLK als im CMK ist.



	-heit	-keit	-ismus	-logie	-nahme	-nis	-tion	-ung
CLK relativ	1,80	2,63	0,48	0,75	0,07	1,90	2,97	16,44
CMK relativ	1,32	2,93	0,81	1,51	0,54	2,88	3,88	25,17
CLK Variation relativ	0,59	1,05	0,27	0,41	0,02	0,39	1,14	5,25
CMK Variation relativ	0,76	1,39	0,20	0,54	0,29	0,42	1,51	8,37

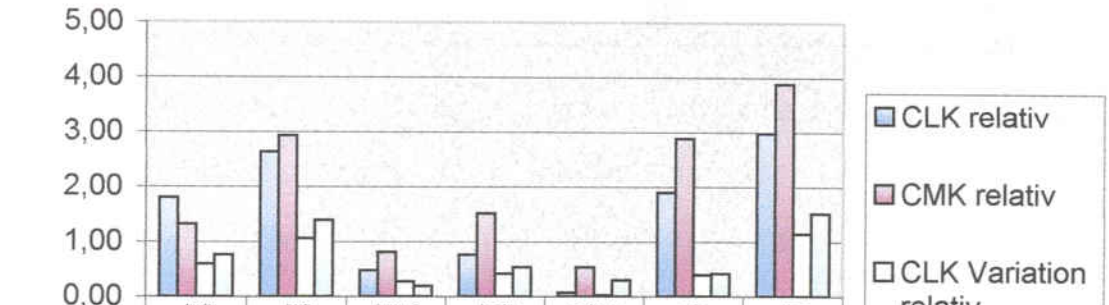
TABELLE 28

Dabei liegen jedoch die Werte der meisten untersuchten Suffixe unter 5‰; zur anschaulicheren Darstellung werden sie erneut in Tabelle 29 zusammengefaßt. Hier wird einerseits klar erkennbar, daß sich beide Ausnahmen zur Regel -daß im CMK mehr dieser Suffixe verwendet werden- im unteren Frequenzbereich befinden, also stark von den statistisch niedrigen Frequenzen des Phänomens beeinflußt sein können, und daß dazu noch im

Falle des höheren Variationswertes des Suffixes *-ismus* dieser eine sehr niedrige Differenz aufweist.

Ein zweites Merkmal dieser Ausnahmen ist, daß sie nicht, wie die allgemeine Tendenz des CMK anzeigt, beide Werte (Relativer Wert und Variationswert) betreffen, sondern nur einen von beiden, also wenig konsistent hinsichtlich ihrer Tendenz sind.

Substantivische Suffixe: Frequenzen unter 5 ‰

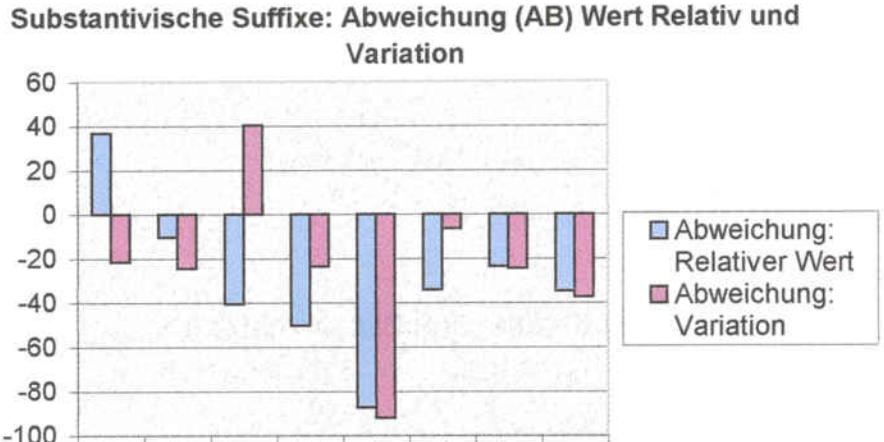


	-heit	-keit	-ismus	-logie	-nahme	-nis	-tion
CLK relativ	1,80	2,63	0,48	0,75	0,07	1,90	2,97
CMK relativ	1,32	2,93	0,81	1,51	0,54	2,88	3,88
CLK Variation relativ	0,59	1,05	0,27	0,41	0,02	0,39	1,14
CMK Variation relativ	0,76	1,39	0,20	0,54	0,29	0,42	1,51

TABELLE 29

Der Vergleich der Abweichung läßt dies deutlicher erkennen (vgl. Tabelle 30). Bis auf den schon erwähnten Relativen Wert von *-heit* und die Variation von *-ismus* liegen alle Werte des CLK unter denen des CMK. Die abweichenden Werte des CLK erreichen gerade 40% der Werte des CMK, während die des CMK im Falle von *-nahme* fast 90% erreichen, obwohl hier erneut

anzumerken ist, daß es sich dabei um ein niedrigfrequentes Element handelt.



	-heit	-keit	-ismus	-logie	-nahme	-nis	-tion	-ung
Abweichung: Relativer Wert	36,86	-10,35	-40,47	-50,21	-87,24	-34,20	-23,51	-34,67
Abweichung: Variation	-21,54	-24,50	40,33	-23,46	-92,20	-6,45	-24,55	-37,27

TABELLE 30

Zusammenfassend ist festzuhalten, daß relativer Wert und Variation der untersuchten substantivischen Suffixe im CMK erheblich über den Werten des CLK liegen (vgl. Tabelle 31), wobei die beiden Ausnahmen auf die statistisch sehr niedrigen Werte zurückzuführen sind und ihnen nur relative Bedeutung beigemessen werden kann.

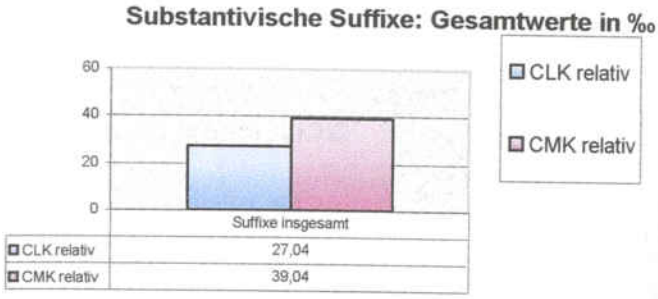
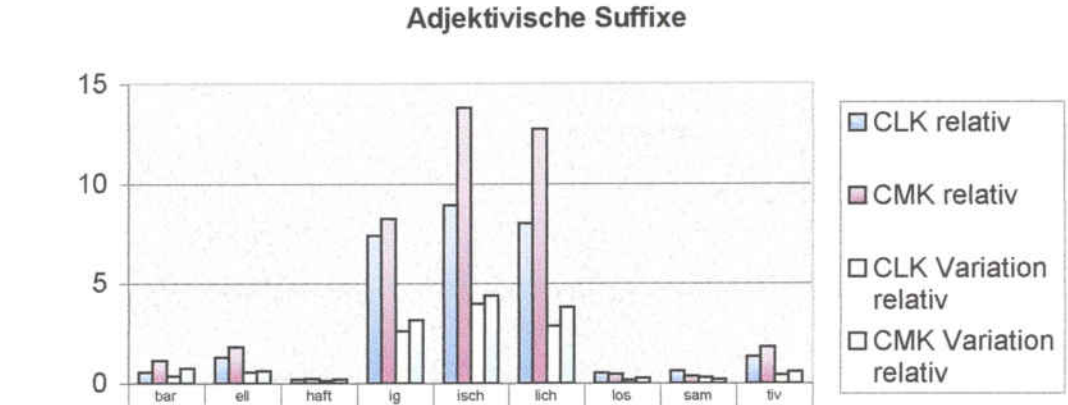


TABELLE 31

Adjektivische Suffixe

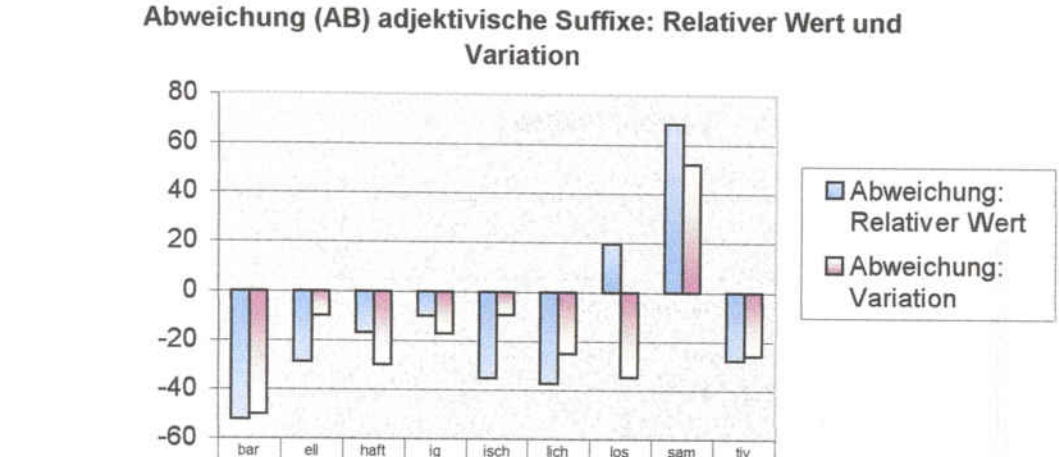
Die adjektivischen Suffixe zeigen ähnliche Tendenzen wie die substantivischen Suffixe auf. An erster Stelle ist festzuhalten, daß fast alle Werte des CMK, sowohl der relative Wert als auch die Variation, höher als die des CLK sind, mit den Ausnahmen der Suffixe *-los* und *-sam* (vgl. Tabelle 32).



	bar	ell	haft	ig	isch	lich	los	sam	tiv
CLK relativ	0,55	1,30	0,18	7,40	8,93	7,99	0,53	0,62	1,32
CMK relativ	1,15	1,83	0,22	8,23	13,79	12,72	0,44	0,37	1,83
CLK Variation relativ	0,37	0,55	0,14	2,60	3,97	2,85	0,16	0,30	0,43
CMK Variation relativ	0,73	0,61	0,20	3,15	4,39	3,81	0,24	0,20	0,59

TABELLE 32

Allerdings handelt es sich dabei einerseits wieder um sehr niedrigfrequente Suffixe, andererseits um sehr geringe Abweichungen der Werte. Es ist jedoch anzumerken, daß das Suffix *-sam* kongruent mit der Tendenz ist und beide Werte, sowohl der relative Wert als auch die Variation, höher als die des CMK sind, während im Falle von *-los* nur der relative Wert leicht über dem des CMK liegt, die Variation des CMK aber fast doppelt so groß wie im CLK ist. Diese Abweichungen sind besser anhand der Darstellung in Tabelle 33 erkennbar: die Ausnahme *-los* zeigt nur eine geringfügig höhere Variation als das CMK (20%) an, während *-sam* eine Variation von 68,39% und 52,02% hinsichtlich des Mittelwertes erreicht, Werte, die sich im CMK in nicht ganz so stark ausgeprägter Form in allen anderen Suffixen widerspiegeln.



	bar	ell	haft	ig	isch	lich	los	sam	tiv
Abweichung: Relativer Wert	-52,23	-28,90	-16,84	-10,06	-35,26	-37,15	19,54	68,39	-27,65
Abweichung: Variation	-50,11	-10,19	-29,84	-17,33	-9,57	-25,04	-34,51	52,02	-25,94

TABELLE 33

Verbale Präfixe

Hinsichtlich der verbalen Präfixe ist zunächst anzumerken, daß für ihre Bestimmung nicht getrennt erscheinende Präfixe in Betracht gezogen wurden. Aus diesem Grund ist es nicht möglich, die Frequenzen der einzelnen Präfixe direkt untereinander zu vergleichen, da zwar alle untrennbar verwendeten Präfixe erscheinen, von den trennbaren aber nur jene, die zusammen mit ihrem Verb auftraten. Des weiteren erscheinen aufgrund der Beschränkung auf Elemente, die eine höhere Frequenz als 0,20‰ aufweisen, einige Präfixe wie *zer-* nicht, obwohl es im normalen Sprachunterricht eine wichtige Stellung einnimmt; hier sei anzumerken, daß es im CMK nur ein einziges mal erscheint (*zerstreuen*), und im CLK dreimal, aber mit nur zwei Formen (*zerstören* und *zerbrechen*).

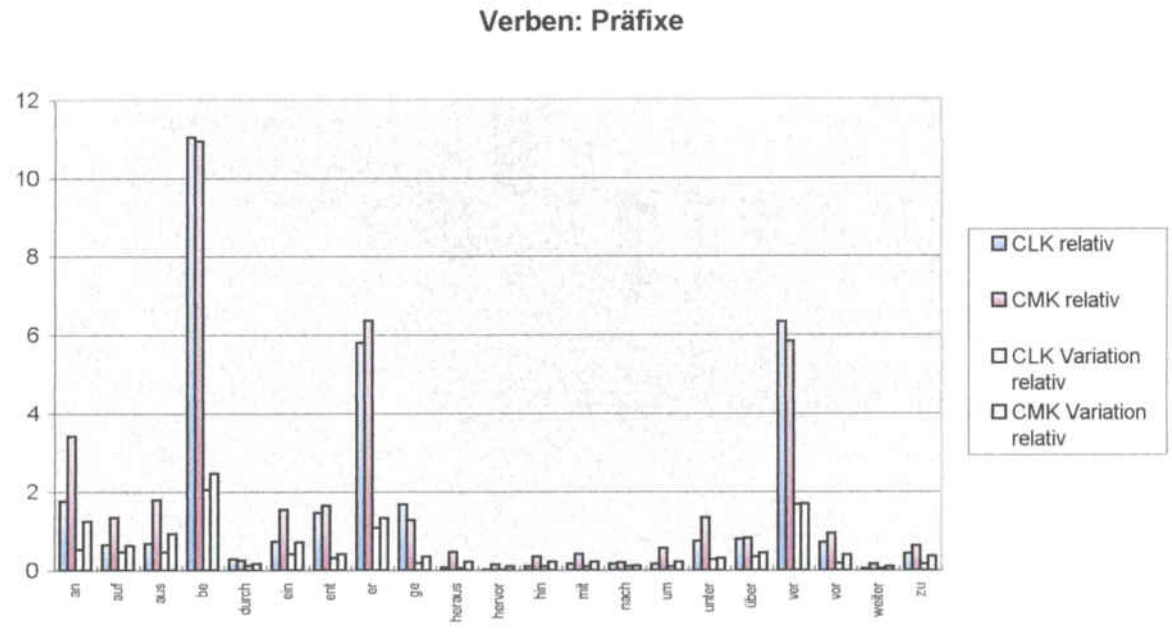


TABELLE 34

Verbale Präfixe zeigen ebenfalls die allgemeine Tendenz der Suffixe zu höheren Werten im CMK auf (vgl. Tabelle 34). Die Tendenz zu hochfrequenten Präfixen ist vergleichbar im CLK und im CMK; *er-* und *ver-* mit ugf. 6‰ und *be-* mit 11‰ haben annähernd gleiche Werte. Aufgrund der stark abweichenden Frequenzen der verschiedenen Suffixe, die eine klare Darstellung erschweren, werden die niedrigfrequenten erneut in Tabelle 35 zusammengefaßt.

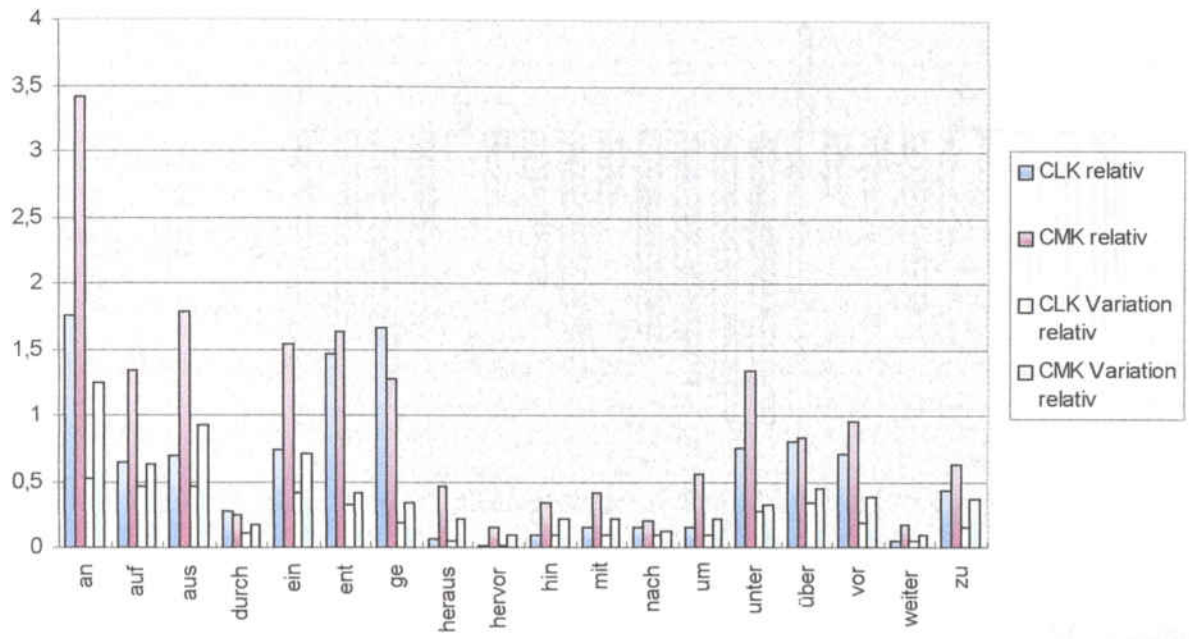


TABELLE 35

Große Abweichungen zugunsten des CMK weisen die Präfixe *an-*, *auf-*, *aus-*, *ein-*, *heraus-*, *um-* und *unter-* auf. Die Variation ist in diesen Fällen höher bei den Muttersprachlern und erreicht bei *heraus* und *hervor* -80% im CLK (vgl. Tabelle 36).

Abweichung (AB) verbale Präfixe: Relativer Wert und Variation

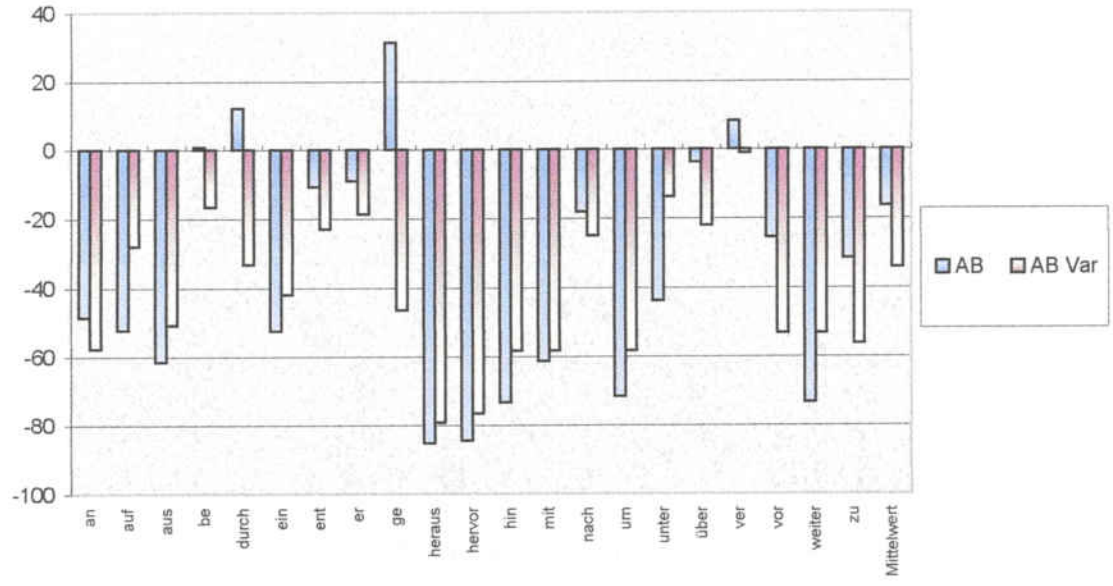


TABELLE 36

Im CLK sind einzig das wenig frequente Präfix *durch-* und das frequente *ge-* öfter als im CMK vertreten. Doch selbst bei höheren relativen Frequenzen ist die Variation in beiden Fällen niedriger im CLK als im CMK.

4.2.2 Wortarten von STTS

Die Einheit der Gruppe der grammatischen Untersuchungen ergibt sich aus der für die Untersuchungen angewendete Information der Dateien. Während den allgemeinen statistischen Untersuchungen das Korpus in seiner rohen Textversion zugrunde lag, werden für die Analysen dieses Abschnittes die grammatischen Auszeichnungen des POS-Systems benötigt, was

gleichzeitig eine nähere Bestimmung der zusätzlich notwendigen Auszeichnungen ermöglichte.

Dem vorhergehenden Abschnitt kann entnommen werden, daß die allgemeinen statistischen Verfahren eine starke Aussagekraft haben. Die Werte weisen starke Abweichungen auf und sind konstant in der vertikalen Dimension, hinsichtlich des Vergleiches beider Korpora, aber auch in der horizontalen Dimension, beim Vergleich der Textsorten eines der Korpora. Ferner zeigen die Werte der Wort- und der Satzlänge aber auch eine Progression in der anzunehmenden Phase der Beherrschung der Sprache auf: sind sie niedrig im Subkorpus MS, steigen sie im Subkorpus OS und erreichen den höchsten Wert im CMK.

Die im folgenden dargestellten Phänomene sind im Vergleich dazu uneinheitlicher. Es handelt sich um stark differenzierte Phänomene, bei denen es nicht mehr ausreicht, eine Gesamtübersicht zu geben, da diese nicht die zugrundeliegenden Tendenzen erkennen läßt. Die Bestimmung und Abgrenzung dieser Faktoren erfolgt hier einerseits empirisch anhand der Überprüfung der von dem Konkordanzprogramm erstellten Listen, andererseits ist sie aber auch an die Hypothesen gebunden, die sich im Laufe der Untersuchungen herauskristallisieren und im folgenden Kapitel qualitativ dargestellt werden. Die Untersuchung der kumulativen Werte einer einzigen syntaktischen Wortart von STTS kann ungenügend für die adäquate Beschreibung dieser Wortart sein, weil andere Faktoren zu einer Über- oder Unterrepräsentation beitragen

können, wie z.B. attributiv verwendete Partizipien und "normale" Adjektive (vgl. 4.2.2.1).

Vergleichende Untersuchungen innerhalb der Textsorten zu einigen syntaktischen Wortarten des CMK zeigten, daß die Anzahl der jeweiligen Wortart je nach Textsorte stark schwanken konnte, weswegen in dem ganzen Abschnitt nicht nur die Gesamtwerte des CLK und des CMK verglichen werden, sondern auch die Werte der verschiedenen Textsorten untereinander. Die hier angegebenen Wortarten entsprechen den syntaktischen Wortarten von STTS, in dem aufgrund distributioneller Kriterien Wortarten verschiedener Grammatiken zusammengestellt werden (vgl. S. 126).

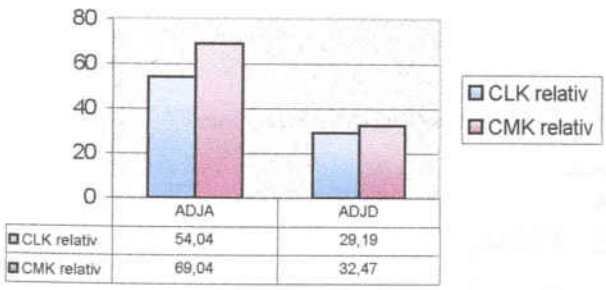
Dabei ist erneut anzumerken, daß nicht alle Tags in der Untersuchung berücksichtigt wurden, weil sie in den Voruntersuchungen auch unterschiedliche Werte oder Ansätze zu einer Erklärung liefern mußten (vgl. 4.1). Ausgeschlossen wurden z.B. Kardinalzahlen (CARD), fremdsprachliches Material (FM), Interjektionen (ITJ), Antwortpartikel (PTKANT), Nichtwörter (XY) und Interpunktionselemente ($).

4.2.2.1 Adjektive (ADJA und ADJD)

STTS teilt Adjektive in attributive Adjektive (ADJA) und prädikativ oder adverbial verwendete Adjektive (ADJD) ein. Adjektive erzielen höhere Werte im CMK und scheinen zusammen mit anderen Wortarten einer der Gründe für längere Sätze im

CMK zu sein. Zum einen handelt es sich um eine sehr häufige Wortart (insgesamt, d.h. ADJA und ADJD zusammengerechnet, mehr als 10% des Gesamtumfangs), so daß Abweichungen in ihrer Verteilung starke Schwankungen erzeugen können. Andererseits ist die Abweichung v.a. der attributiven Adjektive sehr groß. Und schließlich handelt es sich im Falle attributiven Adjektivs ADJA um ein stilistisch freies grammatisches Phänomen.

Adjektive: Verteilung



	ADJA	ADJD
CLK relativ	54,04	29,19
CMK relativ	69,04	32,47

TABELLE 37

Attributive Adjektive stellen im CLK 54,04‰ aller Wörter dar, im CMK hingegen erreichen sie einen Wert von 69,04‰, d.h. exakt 15‰ über dem Wert des CLK. Die prädikativen Adjektive weisen ebenfalls eine Unterrepräsentation im CLK auf mit einem Wert von 29,19‰, der 3,28‰ unter dem Wert des CMK liegt (32,47‰). Kumulativ gesehen sind Adjektive um 18,28‰ seltener im CLK als im CMK.

Doch mehr als die unterschiedlichen Frequenzen der Adjektive scheint der lexikalische Ursprung derselben von Bedeutung zu sein. Dafür haben wir in den beiden Korpora die Adjektive auf ihren verbalen Ursprung untersucht, also unter den Adjektiven jene ausgesondert, die aus einer partizipialen Form abgeleitet sind. Zur Bestimmung dieser attributiven Partizipien wurden verschiedene Suchkriterien kombiniert. Aus den so erhaltenen Listen wurden die falschen Partizipien gestrichen und die restlichen in die endgültige Liste aufgenommen.

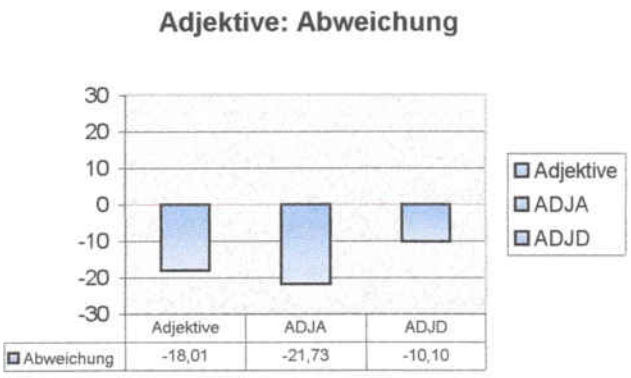


TABELLE 38

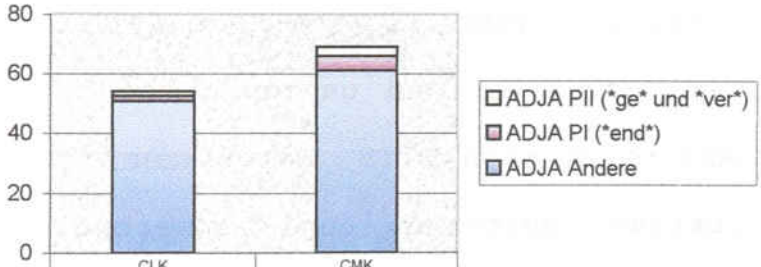
Im Falle des Partizip I wurde ein morphologisches Suchkriterium angewendet, das die PI-Formen der schwachen Verben abdeckt (Adjektive, die die Buchstabenfolge *end* enthalten: ADJA + *end*).

Im Falle des Partizip II wurden folgende morphologische Suchkriterien kombiniert: Adjektive, die mit *ge-* anfingen und

innerhalb des Wortes oder an seinem Ende ein *t* enthielten (*gespielte, gerettete* usw.), und Adjektive, die mit *ver-* beginnen.

Aufgrund fehlender Auszeichnungen zu morphologischen Phänomenen kann mit dieser Methode nur eine Teilgruppe der Adjektive partizipialen Ursprungs gekennzeichnet werden. Nicht erkennbar sind starke Verben, gemischte Verben und präfigierte Verben ohne trennbares Präfix. Doch diese Methode ermöglicht zumindest eine Annäherung an erste quantitative Ergebnisse und Tendenzen.

Verteilung der Adjektive: lexikalischer Ursprung



	CLK	CMK
ADJA PII (*ge* und *ver*)	1,58	3,15
ADJA PI (*end*)	1,60	4,88
ADJA Andere	50,86	61,01

TABELLE 39

Ein weiterer Faktor, der die höhere Frequenz von Adjektiven im CMK erklären kann, ist die Anwendung von zwei oder mehr Adjektiven vor einem Substantiv.

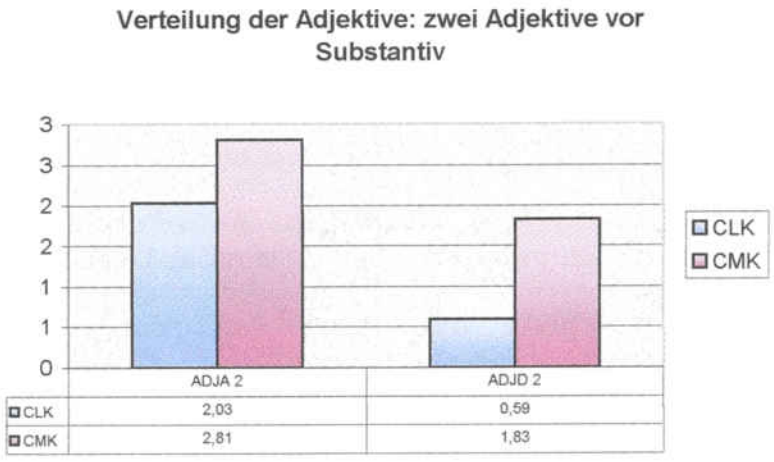


TABELLE 40

4.2.2.2 Adverbien (ADV)

STTS (1995: 55) faßt Adverbien nur als „reine, nicht von Adjektiven abgeleitete, nicht flektierbare Modifizierer von Verben, Adjektiven, Adverbien und ganzen Sätzen" auf. Ebenso gehören zu den Adverbien nach STTS „Wortformen, die auch als attributive Adjektive auftreten und adverbial verwendet werden, die aber semantisch nichts (mehr) mit dem Adjektiv verbindet, und die meistens auch *nicht* prädikativ verwendet werden können".

In diesem Abschnitt wird die kumulative Verteilung der Adverbien dargestellt und auf eine weitere Einteilung verzichtet. Diese Gruppe wird aufgrund distributioneller und syntaktischer Kriterien erstellt, was dazu führt, daß einige semantische Markierungen von Adverbien eine Konkurrenz zu anderen Wortarten herstellen (wie die kausalen Adverbien, die

als Konnektor aufzufassen sind). STTS sieht dabei eine Ausnahme vor, die Pronominaladverbien, die in Abschnitt 4.2.2.8 dargestellt werden und durch die Verbindung distributioneller und lexikalischer Kriterien getaggt werden können. Für die nähere Untersuchung der Adverbien als Gruppe wäre zusätzlich eine weitere Auszeichnungsebene notwendig gewesen, die semantische Information zusammenfaßt.

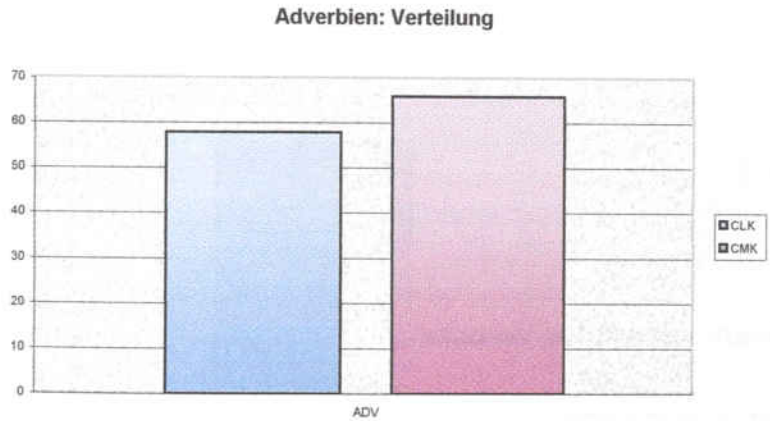


TABELLE 41

Festzuhalten sei hier eine Abweichung von ugf. 10% zugunsten des CMK.

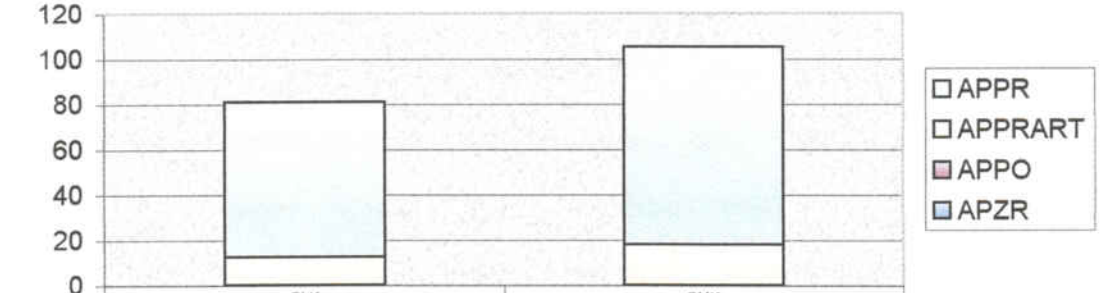
4.2.2.3 Adpositionen

STTS unterscheidet in der Gruppe der Adpositionen zwischen Präpositionen (*auf*), Postpositionen (*entlang*) und Zirkumpositionen (*von ... an*); bei letzteren wird der erste

Teil als Präposition getaggt, während der zweite Teil das Tag APZR erhält.

Die vergleichende kumulative Verteilung zeigt, daß dieses Phänomen eine negative Abweichung im CLK darstellt, daß also höhere kumulative Werte im CMK erzielt werden. Der Hauptteil der Gruppe Adpositionen entspricht den Präpositionen (APPR), gefolgt von den Präpositionen mit Artikel (APPRART); die beiden anderen Gruppen sind mit absoluten Werten zwischen 3 und 13 so selten vertreten, daß sie statistisch kaum Wert haben und in diesem Abschnitt nicht weiter besprochen werden (vgl. Tabelle 42).

Adpositionen: kumulative Verteilung



	CLK	CMK
APPR	68,29	87,18
APPRART	12,31	17,89
APPO	0,30	0,07
APZR	0,11	0,07

TABELLE 42

Präpositionen (APPR)

Die vergleichende Verteilung der Präpositionen ergibt eine negative Abweichung des CLK, die im Falle der Präpositionen, die mit einer Frequenz von mehr als 100‰ ugf. 10% des ganzen Wortschatzes des CMK ausmachen (vgl. Tabelle 42), bei etwas mehr als 20% liegt.

Präpositionen und APPRART: Abweichung (AB)

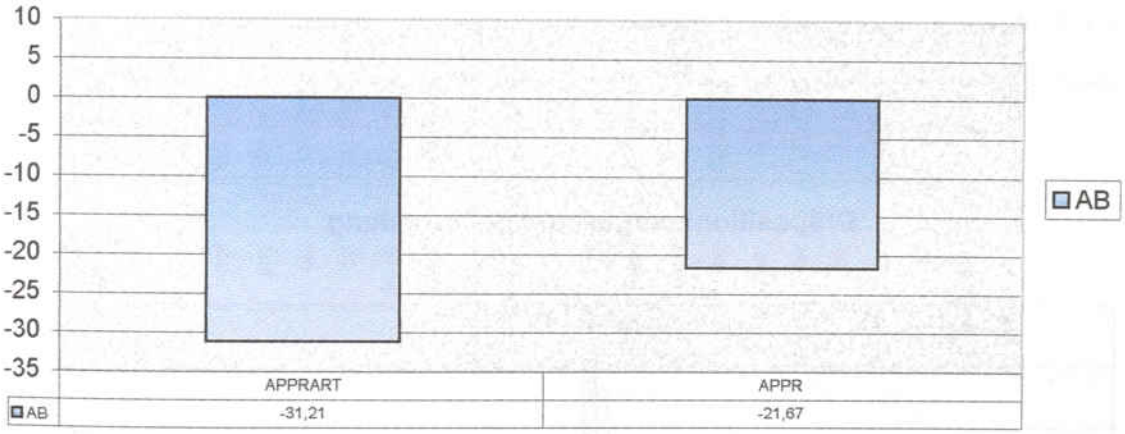


TABELLE 43

Werden die Präpositionen einzeln betrachtet, sind fast alle Werte höher im CMK. Als Tendenz ist festzuhalten, daß häufig erscheinende Präpositionen proportional öfter im CMK verwendet werden als im CLK, während Präpositionen mit sehr niedrigen Frequenzen, wie *gegen, ohne* oder *zwischen*, leicht höhere Werte im CLK erzielen können (vgl. Tabelle 44). Im allgemeinen folgt das CLK den distributiven Tendenzen des CMK: wenn eine

Präposition sehr oft im CMK benutzt wird, erscheint sie auch oft im CLK, wenn auch mit leicht niedrigeren Werten. Die einzige Ausnahme dabei bildet die Präposition *von*, die vergleichbare Werte im CMK und im CLK erzielt, was teilweise auf eine geringere Verwendung von Genitivattributen im CLK und ihre Ersetzung durch Konstruktionen der Art *von + Substantiv* zurückzuführen ist.

Unter *Andere* sind in Tabelle 44 alle Präpositionen zusammengefaßt, die eine Frequenz von weniger als 0,5‰ in *beiden* Korpora aufwiesen; dazu gehört z.B. auch *wegen*.

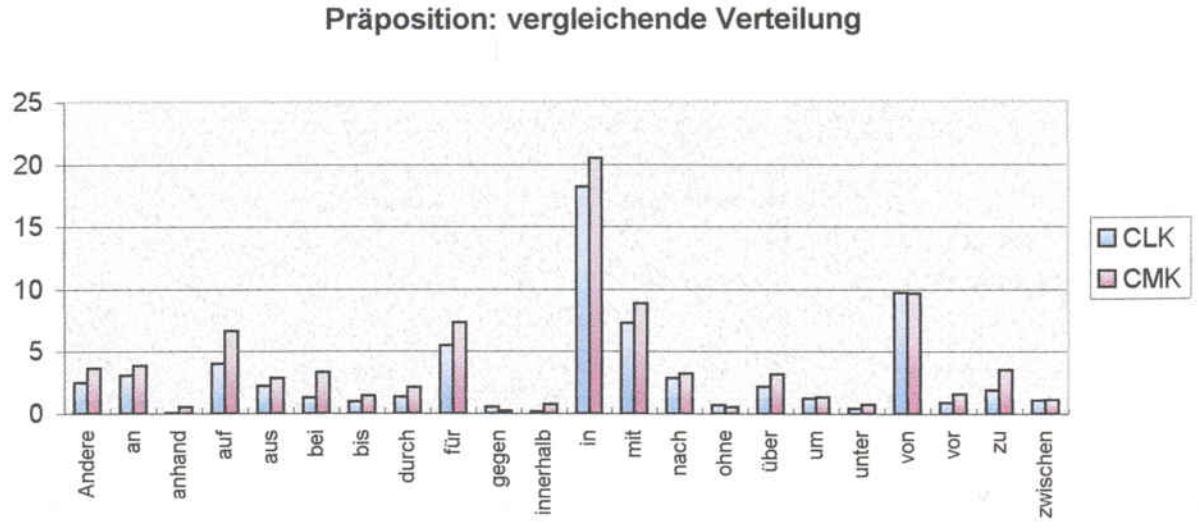


TABELLE 44

Die Untersuchung der Abweichung der einzelnen Präpositionen zeigt, daß nur seltene Präpositionen hohe Abweichungen

aufweisen, während die gebräuchlichsten sich bei einem Wert um die 20% bewegen (vgl. Tabelle 45). Dabei gehören die meisten vertretenen Präpositionen der Gruppe der primären Präpositionen an, während die der sekundären kaum vertreten sind (vgl. Weinrich 1993: 615f).

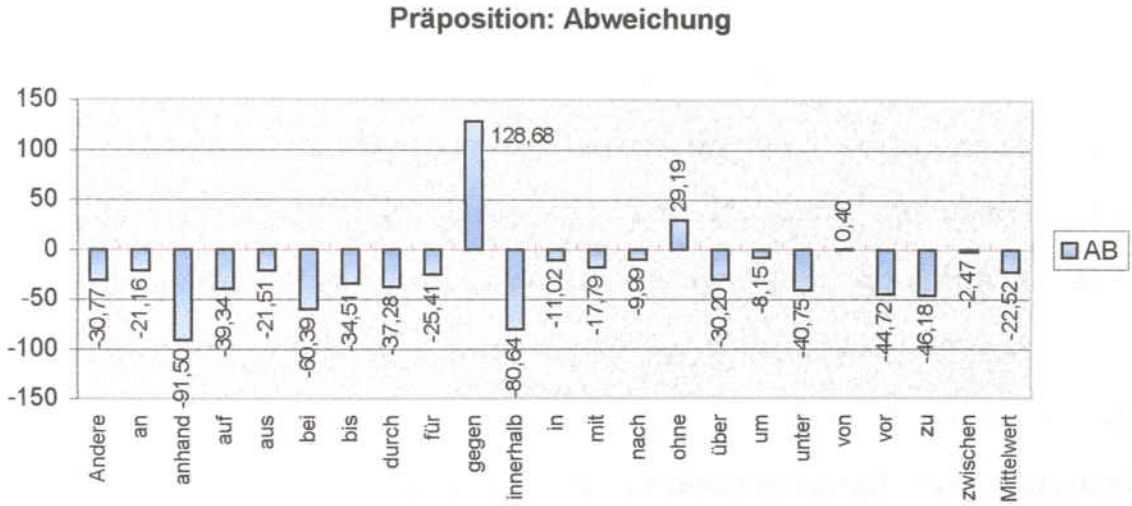


TABELLE 45

Zusammenfassend weist die als APPR getaggte Wortart der Präpositionen im CLK niedrigere Frequenzen auf, wobei allerdings die allgemeine Verteilung des CMK widergespiegelt wird. Dieses Verhalten gibt sich erneut in den Kontraktionen von Präposition und Artikel, was ein Indiz für die Konsistenz des Phänomens ist.

APPRART

Zusammen mit den normalen Präpositionen bilden die Verbindungen von Artikel und Präposition den größten Teil der Gruppe Adpositionen. Ihre Frequenz liegt im CMK bei ugf. 17‰ und bei 12‰ im CLK (vgl. Tabelle 42). Dies stellt eine Abweichung von -30% dar, die damit größer als die Abweichung der normalen Präpositionen (APPR) ist (vgl. Tabelle 43).

Die Übereinstimmung der Tendenz zur Unterrepräsentation der Präpositionen und kontrahierten Präpositionen im CLK würde darauf hindeuten, daß dieses Phänomen der Eigenschaften der Wortart Präpositionen, mehr als dem Artikel, zuzuschreiben ist. In diesem Sinne sei zu beachten, daß STTS Präpositionen nicht innerhalb ihrer syntaktischen Funktion definiert, sondern nur als Einzelelement, so daß keine klare Trennung von Partizipialattributen und nicht regierten Präpositionen vorgenommen werden kann. Die größere Abweichung von APRRART andererseits, für dessen Bildung Wortbildungsmechanismen angewendet werden müssen, ist möglicherweise ein Indiz dafür, daß zusätzlich für diese Abweichung eine andere, von den Präpositionen unabhängige Erklärung gefunden werden kann.

4.2.2.4 Artikel

Artikel werden in STTS ausschließlich als bestimmter oder unbestimmter Artikel definiert, wobei allerdings keine Unterscheidung zwischen ihnen gemacht wird, mit der

Begründung, daß sie sich distributiv gleich verhalten (Schiller et al. 1995: 32). Andere Artikelwörter als *der* oder *ein* und die ebenfalls die substantivische Gruppe eröffnen (vgl. Helbig/Buscha 1991: 355ff) sind in diesem Tag nicht mit inbegriffen, was eine präzise Analyse des Phänomens erschwert. Diese anderen Artikelwörter (oder Determinativa), wie *jener, welcher, manch ein, derselbe* u.a. werden in STTS anderen Wortarten zugeordnet, vor allem den Pronomina (*mancher* wird als attribuierendes Indefinitpronomen ohne *Determiner* ausgezeichnet, *solch ein* als substituierendes Indefinitpronomen usw.). Aus diesem Grund wird hier nur die weitere Einteilung in bestimmter und unbestimmter Artikel eingeführt, die anhand der Bestimmung des bestimmten oder unbestimmten Artikels (*der* oder *ein*) voneinander getrennt werden konnten. Zusätzlich ist zu beachten, daß in einer POS-Auszeichnung ebenfalls keine Nullartikel berücksichtigt werden, was sich als weiteres Problem hinsichtlich der Analyse der Artikel erweist.

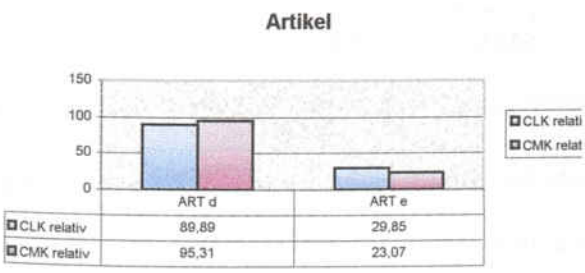


TABELLE 46

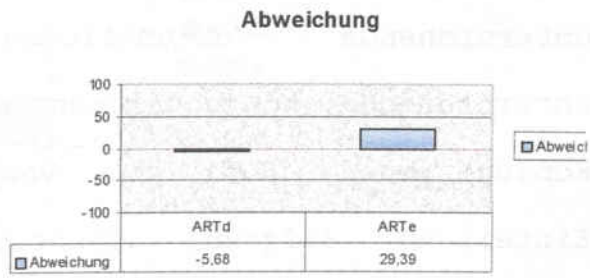


TABELLE 47

Die distributive Analyse zeigt, daß im CLK der bestimmte Artikel leicht unterrepräsentiert ist, während der unbestimmte Artikel eine positive Abweichung von fast 30% aufzuweisen hat (vgl. Tabelle 46 und Tabelle 47). Die kumulative Verteilung des Artikels jedoch ist in beiden Korpora vergleichbar; sie liegt bei 119,74‰ im CLK und bei 118,38‰ im CMK. Auffallend dabei sind jedoch diese vergleichbaren Werte mit den höheren Werten der Substantive (NN) im CMK. Die Differenz wäre erklärbar, wenn als Hypothese davon ausgegangen würde, daß im CLK andere Artikelwörter unterrepräsentiert sind (*solche, manche* usw.) und daß Nullartikel und artikelloser Gebrauch (vgl. Weinrich 1993: 410) ebenfalls niedrigere Frequenzen aufweisen. Da weder Nullartikel noch Artikelwörter in STTS gesondert ausgezeichnet werden, ist diese Überprüfung hier nicht möglich.

4.2.2.5 Konjunktionen

In STTS werden Konjunktionen in vier Gruppen eingeteilt: unterordnende Konjunktionen mit Infinitiv (KOUI), unterordnende Konjunktionen mit Satz (KOUS), nebenordnende Konjunktionen (KON) und Vergleichspartikel (KOKOM). Diese Einteilung aufgrund distributioneller und lexikalischer Kriterien, wobei die Stellung der Elemente der Wortliste „Konjunktionen" innerhalb des Satzes untersucht wird, ermöglicht die Untersuchung des Phänomens in Verbindung mit

besonderen Satzstrukturen, nicht aber die Analyse anderer Konnektoren, die ähnliche Funktionen haben können wie die Konjunktionen, wie z.B. Konjunktionaladverbien (vgl. Ruipérez 1992: 181f).

Die kumulativen Frequenzen aller Konjunktionen liegen im CLK deutlich über denen des CMK. Dies spiegelt sich in den Frequenzen der unterordnenden (KOUS und KOUI) und nebenordnenden Konjunktionen (KON) wider, die im CLK höher als im CMK sind. Die Ausnahme bilden die Vergleichspartikel (KOKOM), die zwar im CMK viel höhere Frequenzen aufweisen, aufgrund der niedrigen relativen Werte dieser Gruppe jedoch nicht weiter betrachtet werden, da sie keinen größeren Einfluß auf die kumulative Verteilung aller Konjunktionen haben (vgl. Tabelle 48).

Konjunktionen: kumulative Verteilung

	CLK	CMK
KOKOM	5,94	8,64
KON	41,34	36,20
KOUS	18,80	13,21
KOUI	4,04	1,39

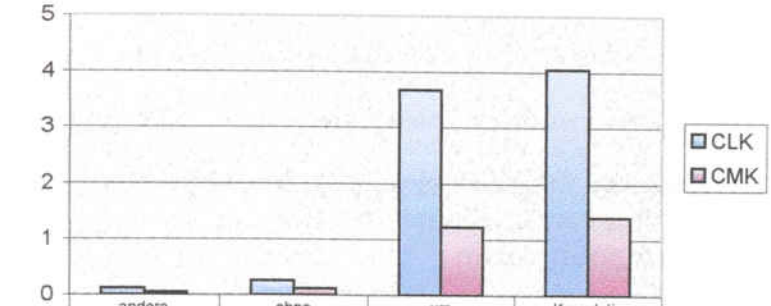
TABELLE 48

Unterordnende Konjunktionen mit Infinitiv (KOUI), stärker im CLK vertreten als im CMK, erreichen ebenfalls nur geringe Werte im Vergleich zu den beiden großen Konjunktionalgruppen, KON und KOUS. Während jedoch KON und KOUS zahlreiche Elemente enthalten, wird die Frequenz der Gruppe KOUI von nur einem innerhalb der Gruppe sehr frequenten Element bestimmt.

*Unterordnende Konjunktionen mit Infinitiv (*KOUI*)*

In STTS werden Konstruktionen wie *anstatt zu, ohne zu* und *um zu* als unterordnende Konjunktionen mit Infinitiv (KOUI) ausgezeichnet. Nur wenige andere Konjunktionen gehören dieser Gruppe an, wie *statt zu*. Aufgrund der niedrigen Frequenzen der meisten Elemente der Gruppe KOUI wurden sie in drei Gruppen eingeteilt, *um zu, ohne zu* und *andere*.

Unterordnende Konjunktionen mit Infinitiv (KOUI): individuelle und kumulative Verteilung



	andere	ohne	um	Kumulativ
CLK	0,11	0,25	3,68	4,04
CMK	0,05	0,12	1,22	1,39

TABELLE 49

Um zu ist mit einer Frequenz von fast 3,68‰ im CLK eine sehr frequente Konjunktion, die innerhalb der Gesamtgruppe der Konjunktionen im CLK und im CMK nur von *daß* übertroffen wird. Dabei ist zu beachten, daß die Konjunktionen dieser Gruppe zusätzlich eine Infinitivkonstruktion beinhalten, also innerhalb der Subordination ebenfalls das Tag PTKZU erscheint.

Unterordnende Konjunktionen mit Satz (KOUS)

Die vergleichende Darstellung der Frequenzen der unterordnenden Konjunktionen mit Satz zeigt große Unterschiede in der Anwendung der einzelnen Elemente auf. Die Überrepräsentation im CLK geht von sehr niedrig im Falle der Konjunktionen *bevor* oder *ob* bis hin zu extremen Abweichungen im Falle von *als, obwohl, weil* oder *wenn*. In nur einigen Ausnahmen wird die allgemeine Tendenz zur Überrepräsentation

im CLK gebrochen, bei den niedrigfrequenten *indem* und *nachdem* und viel akzentuierter im Falle von *da*. Anzumerken ist ebenfalls, daß im CMK mehr verschiedene als KOUS ausgezeichnete Konjunktionen verwendet werden (24) als im CLK (19), also von einer größeren lexikalischen Variation des Phänomens im CMK gesprochen werden kann.

Unterordnende Konjunktionen mit Satz (KOUS): individuelle Verteilung

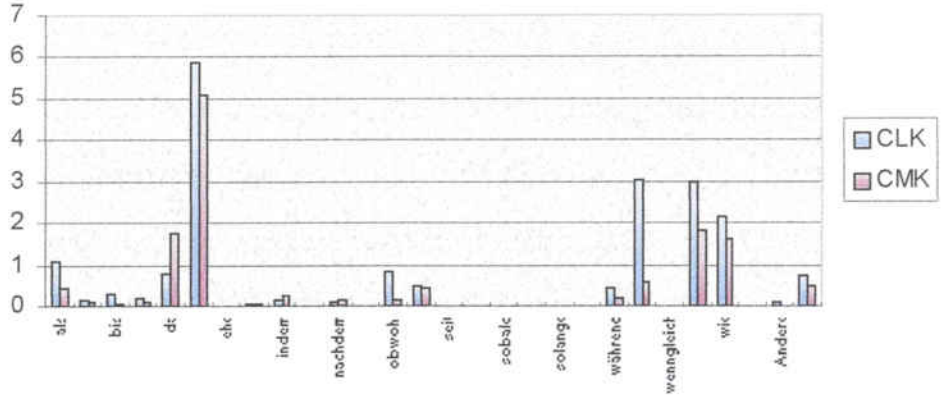


TABELLE 50

Die kumulative Verteilung von KOUS im CLK liegt bei 18,80‰, im CMK bei 13,21‰. Es sind deutlich die hochfrequenten Elemente wie *daß*, *weil* und *wenn* überrepräsentiert. Die getrennte Darstellung der weniger frequenten Elemente (unter 1‰) zeigt jedoch andere distributive Eigenschaften auf, wie Tabelle 51 zu entnehmen ist. Hier ist die Abweichungen der Frequenzen

nicht mehr so stark ausgeprägt, während die Variation der benutzten Elemente klar zugunsten des CMK tendiert.

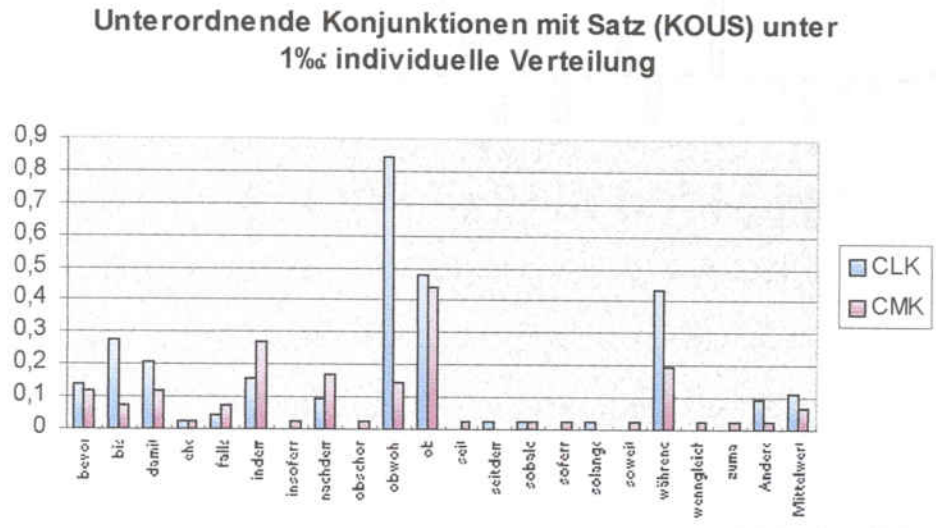


TABELLE 51

Die Abweichungswerte aller Elemente der Gruppe sind in Tabelle 52 zusammengefaßt; dabei können große Schwankungen im CLK erkannt werden.

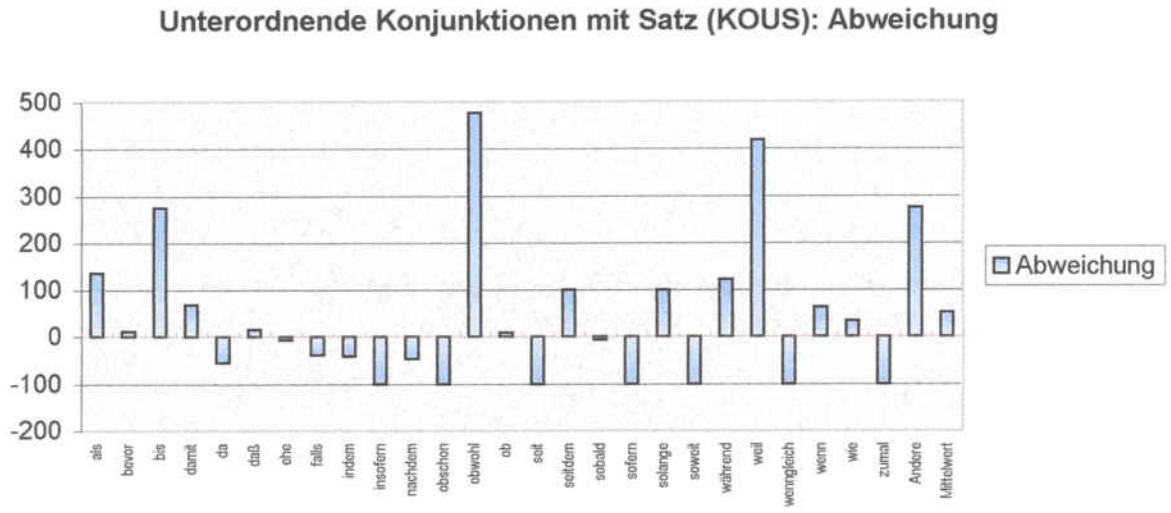


TABELLE 52

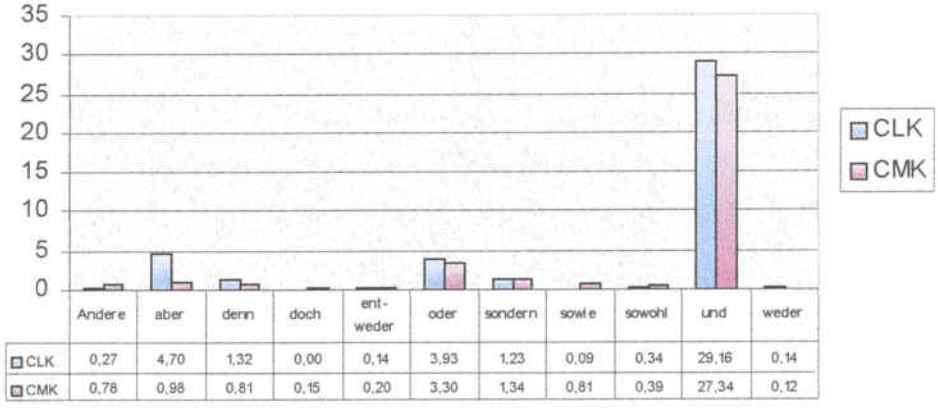
Während in vielen anderen Wortarten die Tendenz besteht, daß zwar nicht unbedingt die relativen Frequenzen der Elemente vergleichbar sind, wohl aber die Tendenz zu höheren oder niedrigeren Werten, stellen die Konjunktionen als uneinheitlicher Abweichungsmuster dar. Die Frequenzen entsprechen einander nicht und Abweichungen erreichen positive Werte im CLK von fast 500% im Falle von *obwohl*, das im CMK sehr zurückhaltend gebraucht wird. Aber auch *weil* stellt einen besonderen Fall dar, mit einer fast viermal so hohen Frequenz im CLK. Obwohl man dazu tendieren würde, *weil* subjektiv als frequente Konjunktion einzustufen, erscheint sie eher selten im CMK. Gründe dafür sind wohl in der höheren Frequenz von dem semantisch vergleichbaren *da* im CMK zu suchen, in der Verwendung von Konjunktionaladverbien und Pronominaladverbien,

mit entsprechendem Satzbau, und eventuell auch in kulturspezifischen Argumentationsformen.

Nebenordnende Konjunktionen (KON)

Einfache, mehrteilige und satzeinleitende Konjunktionen, die eine V2-Stellung zulassen, werden innerhalb von STTS als KON getaggt. Einfache Konjunktionen sind dementsprechend *und* und *oder*, mehrteilige Konjunktionen *entweder ... oder*, *weder ... noch* u.a., satzeinleitende Konjunktionen *denn*, *aber*, *doch* oder *jedoch*. Während die einfachen Konjunktionen eindeutig ausgezeichnet werden, erscheint bei den mehrteiligen das Problem, daß beide Elemente das Tag KON tragen. Die satzeinleitenden Konjunktionen ihrerseits werden aufgrund distributioneller Kriterien bestimmt. So wird *aber* in nebensatzeinleitender Position als KON getaggt, ansonsten aber als Adverb (ADV). In den Tabellen werden unter dem Eintrag *Andere* als KON getaggte Elemente ausgezeichnet, die nicht klar einer bestimmten Konjunktion zugeschrieben werden konnten, wie z.B. *und/oder* in *sogar die Erklärung des Beispiels und/oder auch weitere Möglichkeiten*.

Nebenordnende Konjunktionen (KON): individuelle Verteilung (in ‰)



	Andere	aber	denn	doch	ent-weder	oder	sondern	sowie	sowohl	und	weder
CLK	0,27	4,70	1,32	0,00	0,14	3,93	1,23	0,09	0,34	29,16	0,14
CMK	0,78	0,98	0,81	0,15	0,20	3,30	1,34	0,81	0,39	27,34	0,12

TABELLE 53

Die distributive Untersuchung der nebenordnenden Konjunktionen zeigt, daß sich starke Abweichungen hauptsächlich im Falle von *aber* ergeben, mit einer Frequenz von 4,7‰ im CLK und 0,98‰ im CMK. In Beziehung zur kumulativen Tendenz zur Überrepräsentation von Konjunktionen im CLK ist hier bemerkenswert, daß sowohl die mehrteiligen Konjunktionen als auch selten benutzte Konjunktionen eine Tendenz zu ähnlichen Werten im CLK und im CMK aufweisen, und daß diese Werte sogar leicht höher im CMK sind. Im Falle der mehrteiligen Konjunktionen sind das *entweder, sowohl* und *weder,* und im Falle der niedrigfrequenten Konjunktionen vor allem *sowie,* die anscheinend im CLK vermieden wird. Zur besseren Darstellung dieser niedrigfrequenten Phänomene befindet sich in Tabelle 54 eine Zusammenfassung der nebenordnenden Konjunktionen ohne die hochfrequente *und.*

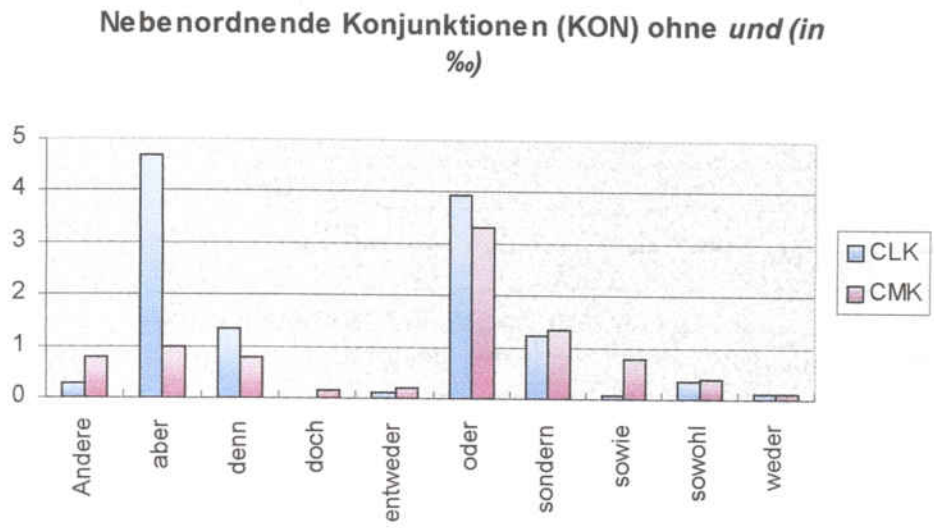
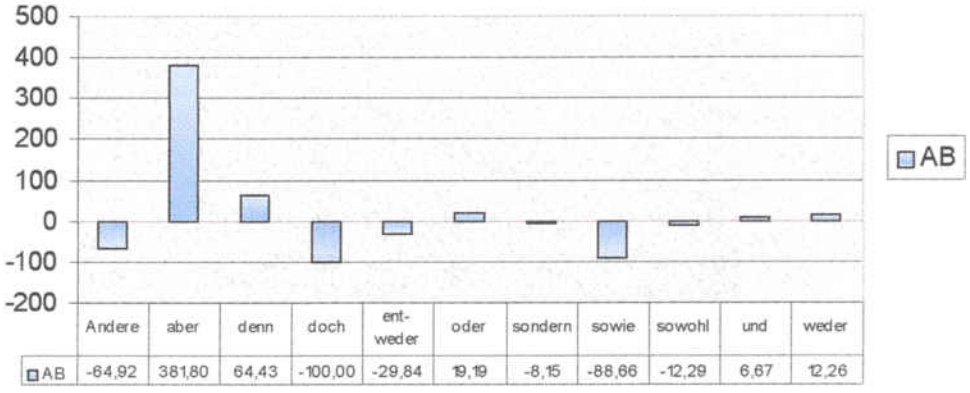


TABELLE 54

Die Überrepräsentation der Konjunktionen im CLK ergibt sich hier statistisch aus dem extrem hohen Wert der Konjunktion *aber*, zusammen mit leicht höheren Werten von *denn* und *oder* (vgl. Tabelle 54). Dabei ist anzumerken, daß der ImS-Tagger *aber* immer dann als Konjunktion taggt, wenn es satzeinleitend erscheint.

Nebenordnende Konjunktionen (KON): Abweichung (AB)



	Andere	aber	denn	doch	ent-weder	oder	sondern	sowie	sowohl	und	weder
AB	-64,92	381,80	64,43	-100,00	-29,84	19,19	-8,15	-88,66	-12,29	6,67	12,26

TABELLE 55

Eine weitere Abweichung ist die von *denn*, das mit einer Frequenz von fast 1‰ im CMK vergleichbar mit den relativen Werten von *aber*, *sondern* und *sowie* im CMK ist. *Denn* weist eine positive Abweichung von 64,43% im CLK auf und bildet somit zusammen mit dem schon erwähnten *aber* die einzige nennenswerte positive Abweichung. Negative Abweichungen hingegen stellen die Konjunktionen *doch* und *sowie* dar. *Doch* ist wenig frequent in den akademischen Texten des CMK, *sowie* hingegen erstaunlich oft repräsentiert, wahrscheinlich aufgrund textsortenspezifischer Argumentationsstrukturen. Die negative Abweichung im CLK beträgt -88.68%.

Vergleichspartikel (KOKOM)

Innerhalb der Konjunktionen sind die Vergleichspartikel die einzigen Elemente, die als Gruppe eine negative Abweichung im

CLK darstellen, also höhere Werte im CMK erzielen. Vergleichspartikel sind für STTS ausschließlich *als* und *wie*. Diese beiden Elemente werden nur als KOKOM getaggt, wenn sie nicht satzeinleitend angewendet werden.

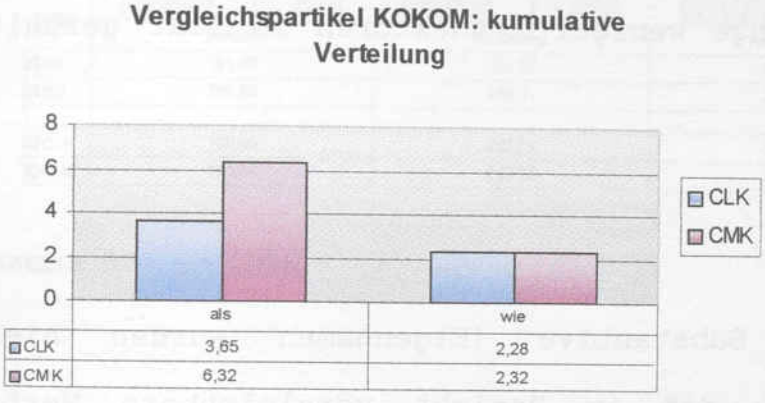


TABELLE 56

Die distributive Analyse der beiden Elemente dieser Gruppe ergibt, daß die unterschiedlichen Frequenzen der kumulativen Werte von KOKOM aus der Abweichung von *als* entspringen, nicht aber von *wie*, das nahezu identische Werte in beiden Korpora aufweist. Die große Abweichung von dem Vergleichspartikel *als* ergibt sich aus den syntaktischen und semantischen Strukturen, für die es in seiner modalen komparativen Bedeutung angewendet werden kann (vgl. Helbig/Buscha 1991: 454f).

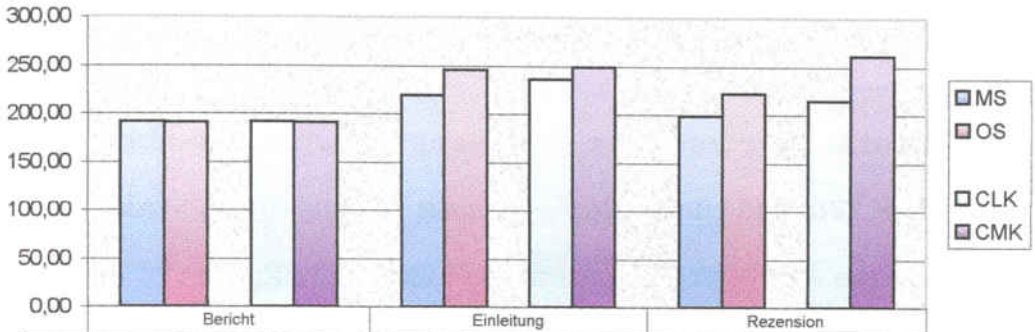
Die Untersuchung der Frequenzen der Konjunktionen sind aufgrund des semantischen Gehalts dieser Wortart und ihrer Abhängigkeit von der Argumentationsstruktur anhand anderer

Textsorten zu relativieren. Außerdem sind Konjunktionen textuell als Konnektoren aufzufassen, weshalb weitere Untersuchungen dazu diese mit einbeziehen müßten, wie die Untersuchung zu konjunktionsartigen Ausdrücken von Hoey (1993: 67). Des weiteren ist der Einfluß anderer Phänomene zu beachten, wie z.B. der Hinweis von Crystal (1991: 234), daß mit zunehmender Satzlänge weniger Konnektoren im Text gezählt werden.

4.2.2.6 Substantive (NN)

Die Verteilung der Substantive (Eigennamen wurden hier ausgeschlossen) zeigt, daß im Bericht vergleichbare Werte erzielt werden (um 190‰ in beiden Korpora), während die Abweichung in der Einleitung zugunsten des CMK steigt (236,24‰ vs. 248,97‰) und in der Rezension beachtlich ist (214,47‰ vs. 261,31‰) (vgl. Tabelle 57).

Substantive: Vergleichende Distribution im CLK und im CMK (in ‰)



	Bericht	Einleitung	Rezension
MS	191,49	219,37	198,43
OS	190,82	245,51	222,30
CLK	190,95	236,24	214,47
CMK	190,84	248,97	261,31

TABELLE 57

Die distributive Untersuchung dieses Phänomens anhand morphologischer Phänomene schien vielversprechend für eine Erklärung der Abweichung; Voruntersuchungen mit den von Morphy morphologisch getaggten Texten hatten ergeben, daß z.B. die Werte der Genitivattribute im CMK bedeutend über denen des CLK lagen.

Diese morphologische Analyse, die aufgrund der fehlenden morphologischen Auszeichnungen der Texte nur einen Ansatz darstellt, wurde anhand der Konkordanzlisten der Genitivformen des bestimmten (*des*) und unbestimmten (*eines*) Artikels der Genus Maskulin und Neutrum vorgenommen. Beide Formen werden ausschließlich für diesen Kasus angewendet und konkurrieren so nicht mit anderen Formen, im Gegensatz zu „der", was zur manuellen Bearbeitungen sehr langer Konkordanzlisten geführt

hätte. Die femininen Formen der Artikel mußten dementsprechend ausgelassen werden, da sie stark mit anderen Artikelformen konkurrieren.

Nicht betrachtet werden in dieser hypothesenbezogenen Untersuchung die Verteilung der Formen Nominativ, Akkusativ und Dativ, da die entsprechenden Artikelformen mit anderen Kasus konkurrieren. Ebenfalls wurden die mit Nullartikel erscheinenden Genitivformen nicht betrachtet, da diese aufgrund der fehlenden morphologischen Auszeichnung ebenfalls nicht erkennbar waren. Als Grundlage für eine eingehende distributive Analyse der Verteilung von NN wäre eine morphologische Auszeichnung der Korpora notwendig gewesen. Diese Art Auszeichnung ist zwar von STTS vorgesehen (Schiller et al. [1995: 11] nennen Genus, Kasus, Numerus und Flexion als Hauptkategorien für die morphologische Auszeichnung), wird aber vom entsprechenden Tagger des Instituts für maschinelle Sprachverarbeitung nicht durchgeführt und hätte so manuell eingefügt werden müssen.

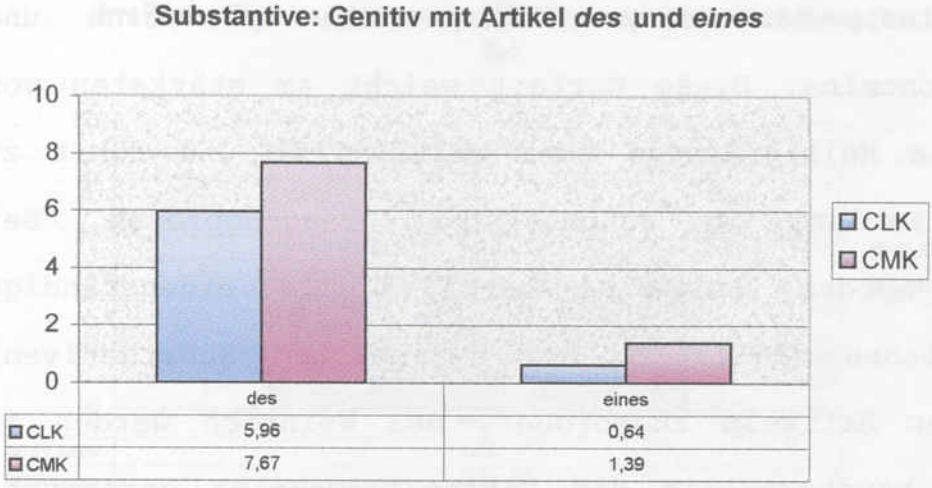


TABELLE 58

Die Abweichung zwischen den Korpora summiert 3‰; bei dieser Berechnung wurden, wie beschrieben, nur Genitivformen mit Artikel des Maskulin und Neutrum dargestellt. Dies führt zur Überlegung, daß eine eingehende Auszeichnung der morphologischen und syntaktischen Phänomene zu präziseren Untersuchungen der Anwendung nominaler Phrasen und Strukturen führen kann, die die in Tabelle 57 zusammengefaßten Abweichungen näher erklären könnten.

4.2.2.7 Pronomina

STTS teilt Pronomina in attribuierende und substituierende ein, wobei die ersten innerhalb der NP auftreten und die zweiten diese ersetzen. Weiterhin unterscheidet das System zwischen Pronominaladverbien, Demonstrativpronomina, Indefinitpronomina, irreflexiven Personalpronomina, reflexiven

Personalpronomina, Possessivpronomina, Relativpronomina und Interrogativpronomina. Diese Wortart weicht am stärksten von Grammatiken wie Helbig/Buscha oder Weinrich ab und führt zu den meisten Fehlern im automatischen Taggingprozeß. Bei Helbig/Buscha werden Pronomina gänzlich als eigenständige Wortart gestrichen (1991: 21) und erscheinen Substantiven, Adjektiven oder Artikeln zugeordnet. Bei Weinrich werden zu den Pronomina hauptsächlich die Personalpronomina gerechnet, die in STTS ein gesondertes Tag erhalten; Reflexivpronomina werden bei Weinrich gesondert behandelt.

Nicht reflexive Personalpronomina (PPER)

Die in STTS vorgesehenen nicht reflexiven Personalpronomina bilden eine hinsichtlich der reflexiven Personalpronomina abgegrenzte Gruppe, gebildet aus folgenden Elementen: *ich, meiner, du, deiner, er, sie, es, seiner, ihrer, ihm, ihn, ihr, wir, unser, ihr, euer, sie, ihrer, ihnen, mich, dich, dir* und *mir*. Sie überschneiden sich mit den attribuierenden Possessivpronomina (ihr Kleid, euer Auto), die das Tag PPOSAT erhalten.

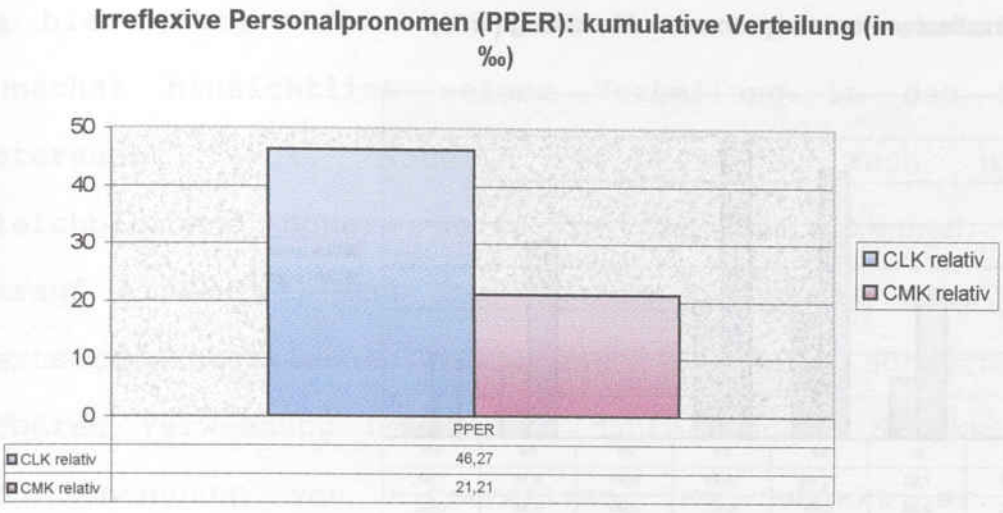
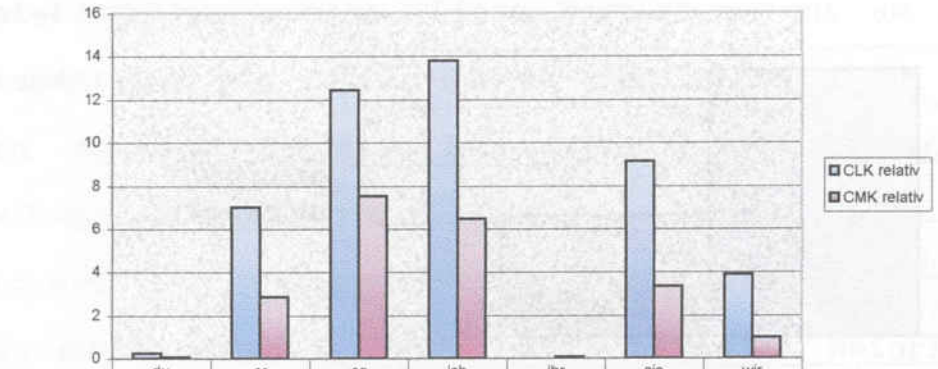


TABELLE 59

Die kumulative Verteilung zeigt eine große positive Abweichung des CLK in der Anwendung von PPER (vgl. Tabelle 59). Zur differenzierten Erklärung dieses Unterschiedes wurden die nicht reflexiven Personalpronomina einzeln alphabetisch geordnet anhand der Grundform dargestellt (vgl. Tabelle 60). Die Überrepräsentation im CLK ist nicht auf ein einziges Pronomen zurückführbar, sondern sie spiegelt sich in allen wider.

Irreflexive Personalpronomen: Verteilung (in ‰)



	du	er	es	ich	ihr	sie	wir
CLK relativ	0,23	7,01	12,45	13,79	0,00	9,14	3,88
CMK relativ	0,05	2,86	7,52	6,47	0,05	3,34	0,95

TABELLE 60

Im Falle des Personalpronomens *du* handelt es sich um ein niedrigfrequentes Phänomen, das sowohl im CLK (Absoluter Wert: 9) als auch im CMK (Absoluter Wert: 2) ausschließlich in den Berichten erscheint. Im CMK wird es nur in der direkten Rede verwendet, als Zitat, und zwar in einem einzigen Text des Korpus (*mkb0018*). Im CLK wird *du* zweimal in der direkten Rede verwendet, ansonsten als Substitut für Strukturen, wo normalerweise ein unpersönliches *man* zu erwarten wäre (*Dann brauchst du keine Aufenthaltserlaubnis* oder *entweder sprichst du deutsch, oder du ißt nicht*). Die 9 Belege für *du* im CLK verteilen sich auf 6 Dateien, so daß hier ebenfalls nicht von einem allgemeinen Gebrauch des Pronomens gesprochen werden kann.

Mit absoluten Werten von 307 Belegen im CLK und 117 im CMK ist das Personalpronomen *er* hingegen ein sehr frequentes Phänomen.

Da hier genug statistische Daten vorhanden waren, wurde es zunächst hinsichtlich seiner Verteilung in den Textsorten untersucht (vgl. Tabelle 61), doch auch hier sind gleichbleibend höhere Werte im CLK zu erkennen, was also darauf hindeutet, daß die höheren Frequenzen sich nicht aus textsortenspezifischen Phänomenen ergeben, sondern aus der höheren Verwendung bestimmter syntaktischer Strukturen, z.B. der Anwendung von Alternativen zum Subjekt *er*. Für die Bestimmung und Auswertung solcher Strukturen wäre jedoch ein syntaktisches Tagging der Korpora notwendig gewesen. Dieser Prozeß zur Erkennung syntaktischer Strukturen, *parsing* (Meya/Huber 1986: 41), anhand dessen z.B. das Subjekt der Sätze statistisch bearbeitet werden konnte, ist aufgrund der technischen Beschränkungen der heutigen Parser noch mit einem sehr hohen manuellen Arbeitsaufwand verbunden.

Irreflexive Personalpronomen: Verteilung von *er* in den Textsorten (in ‰)

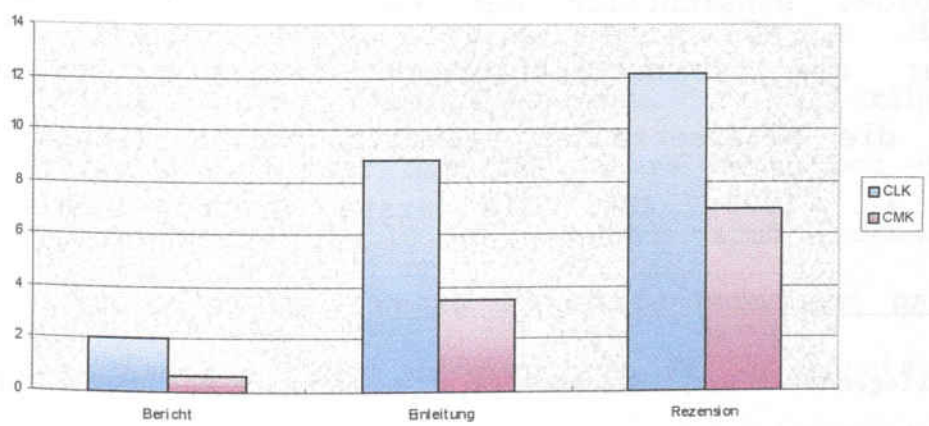


TABELLE 61

Auch die Verteilung nach Kasus ergibt, zum großen Teil aufgrund der niedrigen Frequenzen, keine eindeutigen Abweichungsschemata:

PPER	er	ihn	ihm
CLK	0,59	0,05	0,06
CMK	0,22	0,03	0,03

Auf dieses Phänomen wird in den qualitativen Untersuchungen eingegangen, mit der Hypothese, daß im CLK mehr PPER als Substantiv ersetzende Elemente verwendet werden. Diese Untersuchung kann aufgrund der fehlenden syntaktischen Auszeichnung nicht in den quantitativen Darstellungen vorgenommen werden.

Relativpronomina (PRELAT und PRELS)

STTS unterscheidet hinsichtlich der Relativpronomina zwei große Gruppen: die nomenattribuierenden Relativpronomina (PRELAT)[67] und die NP-ersetzenden Relativpronomina (PRELS)[68] (Schiller et al. 1995: 48). Die erste Gruppe besteht

[67] Z.B. "[...] liegende Stadt, die<PRELS[d]> durch den [...]". Datei lkbel014.

[68] Z.B. „[...] Textarbeit, welche<PRELS[welch]> hier angewandt [...]". Datei mkel013.

ausschließlich aus den Relativpronomina *deren* und *dessen*, die zweite aus den Relativpronomina *der, die, das (etc.), welch-* und *was*.

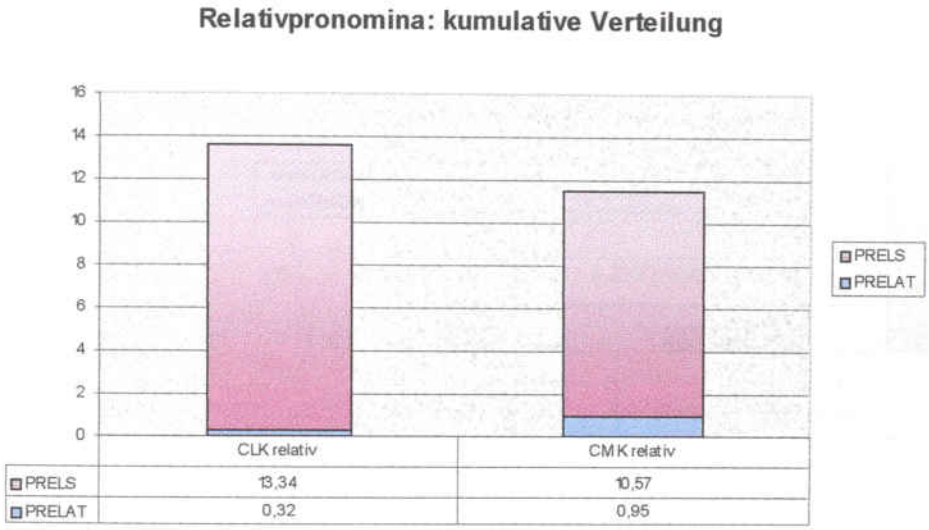
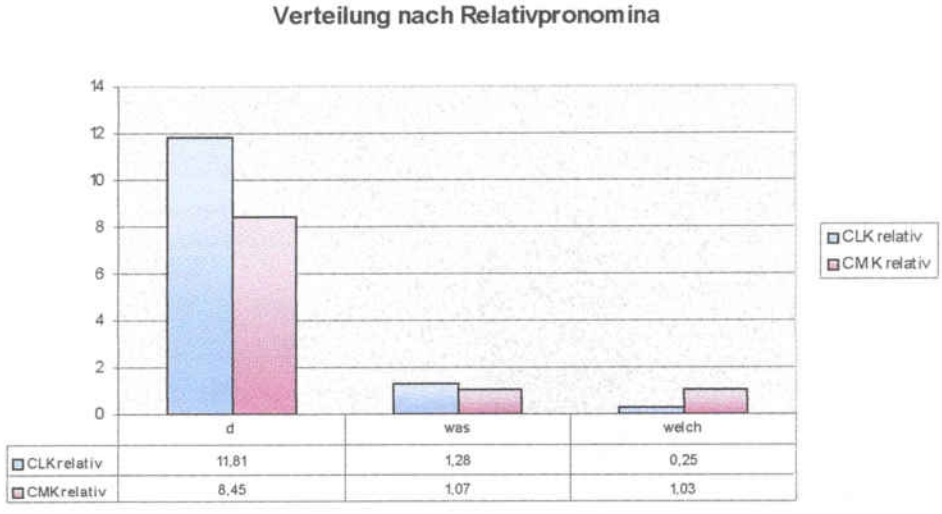


TABELLE 62

Die Untersuchung in bezug auf die kumulativen relativen Werte beider Korpora ergab, daß die des CLK leicht über denen des CMK liegen (13.66‰ im CLK vs. 11.52‰ im CMK), nach der Einteilung der Relativpronomina in attribuierende und substituierende aber das CMK höhere Werte bei diesen letzten aufweist (vgl. Tabelle 62), obwohl dabei anzumerken ist, daß es sich um relativ wenig frequente Phänomene handelt. Wie in vielen anderen schon untersuchten Fällen könnte dies auf die Vermeidung komplexerer Strukturen hindeuten, was durch die

Verteilung der Verwendung der einzelnen Relativpronomina (*der, was, welch-*) klarer dargestellt werden kann.



	d	was	welch
CLK relativ	11,81	1,28	0,25
CMK relativ	8,45	1,07	1,03

TABELLE 63

Die Verteilung der einzelnen Relativpronomina zeigt nur eine große Abweichung hinsichtlich der allgemeinen Tendenz auf (vgl. Tabelle 63); während das Relativpronomen *der* eine Abweichung AB von fast 40% mehr im CLK aufweist, sinkt dieser Wert im Falle von *was* auf weniger als 20% und stellt im Falle von *welch-* sogar eine Unterrepräsentation von 24% im CLK dar (vgl. Tabelle 64).

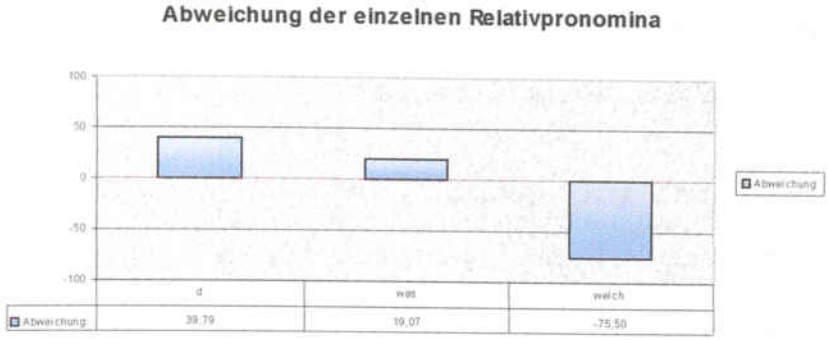


TABELLE 64

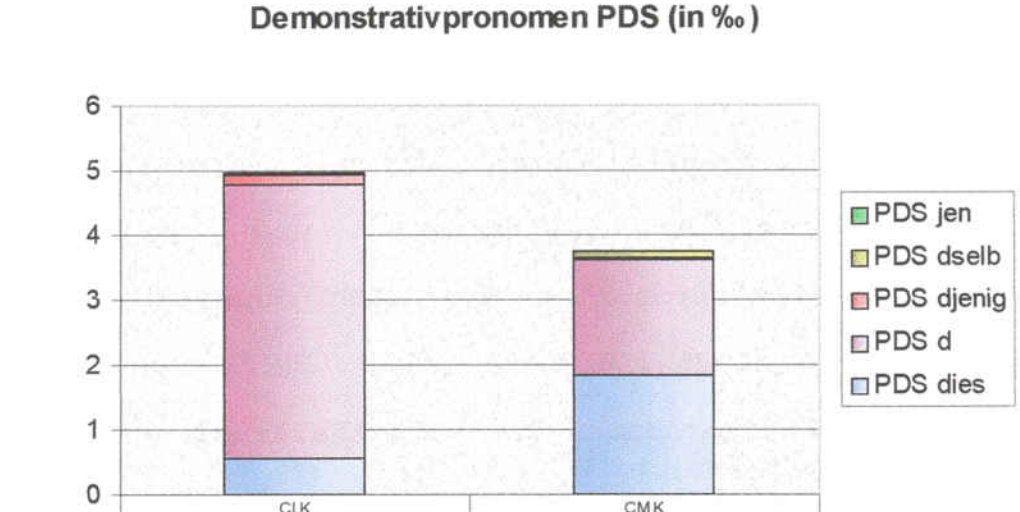
Die Tendenz ist hier eine starke Überrepräsentation von Relativsätzen im CLK, die aber hauptsächlich im Relativpronomen *d-* zum Ausdruck kommt. Das Relativpronomen *was* ist in seiner Anwendung restringiert und wird nach Demonstrativen, Indefinitpronomina und nach Superlativen im Neutrum benutzt (Hentschel/Weydt 1990: 225), was die proportional zum Relativpronomen *d-* zurückhaltende Anwendung im CLK erklären könnte. *Welch-* ist „gegenüber *der/die/das* das seltenere Relativpronomen" (Hentschel/Weydt 1990: 225) und wird als stilistische Alternative verwendet, wenn mehrere gleichlautende Formen vermieden werden sollen. In diesem Sinne läge die Unterrepräsentation von *welch-* im CLK in einer niedrigeren stilistischen Kompetenz begründet.

Demonstrativpronomina (PD)

Demonstrativpronomina werden in STTS in substituierende (PDS) und attribuierende Demonstrativpronomina (PDAT) eingeteilt.

PDAT sind beispielsweise *dieses [Buch]* oder *jene [Frage]*; PDS sind *dies [ist ein Buch]* oder *jenes [ist schwierig]*.

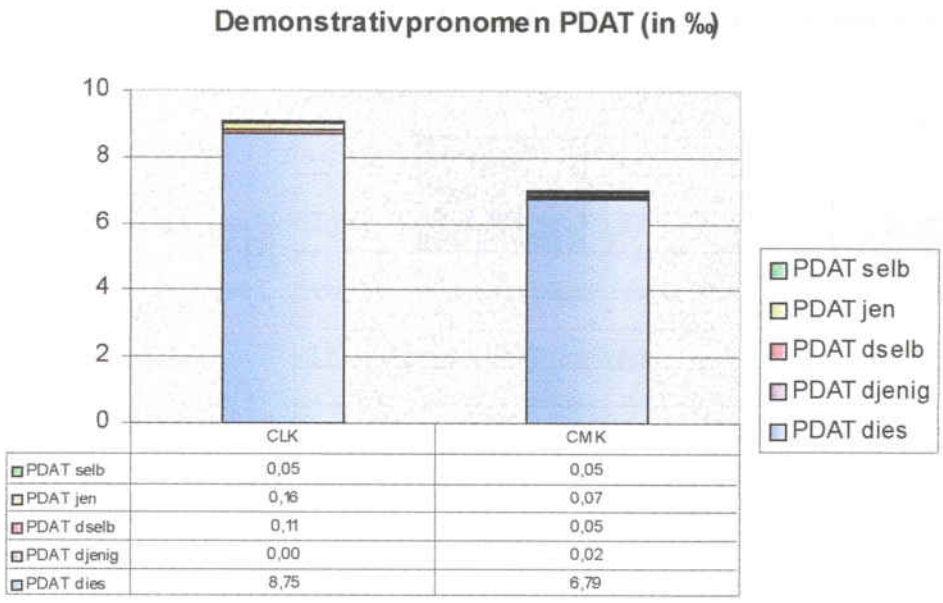
Die Frequenzen des substituierenden Demonstrativpronomens liegen im CLK deutlich über denen des CMK, wobei allerdings auf die sehr starke Überrepräsentation der Form *d-* hinzuweisen ist.



	CLK	CMK
PDS jen	0,02	0,00
PDS dselb	0,02	0,10
PDS djenig	0,14	0,05
PDS d	4,23	1,78
PDS dies	0,57	1,83

TABELLE 65

Eine proportional vergleichbare Überrepräsentation weist das attribuierende Demonstrativpronomina PDAT im CLK auf. Der größte Teil dieser Gruppe besteht aus dem Demonstrativpronomen *dies-*; alle anderen erscheinen zu selten, um sie in die Auswertung mit einzubeziehen.



	CLK	CMK
PDAT selb	0,05	0,05
PDAT jen	0,16	0,07
PDAT dselb	0,11	0,05
PDAT djenig	0,00	0,02
PDAT dies	8,75	6,79

TABELLE 66

Die kumulative Abweichung der Demonstrativpronomina wird im folgenden nach PDS und PDAT zusammengefaßt, ohne weitere Unterscheidung der Einzelelemente. Dabei wird eine positive Abweichung im CLK von fast 35% für PDS und von fast 30% für PDAT errechnet.

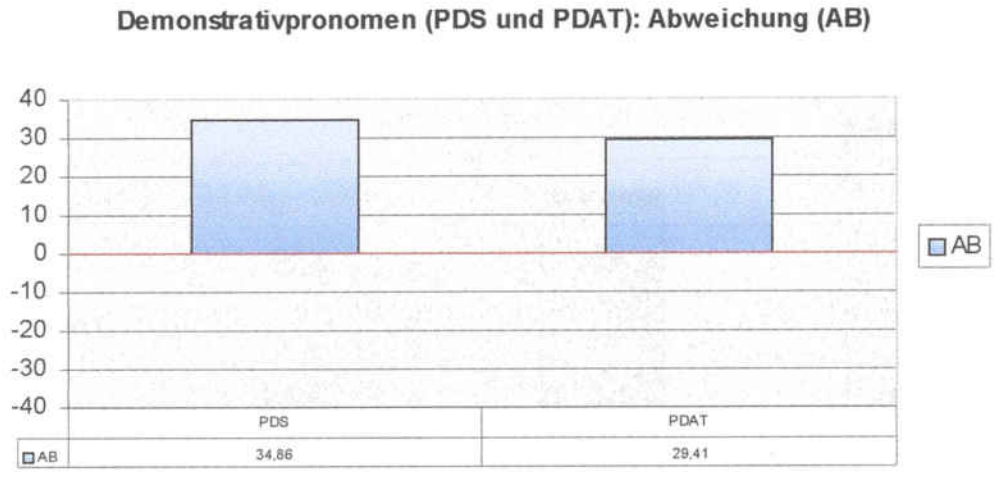


TABELLE 67

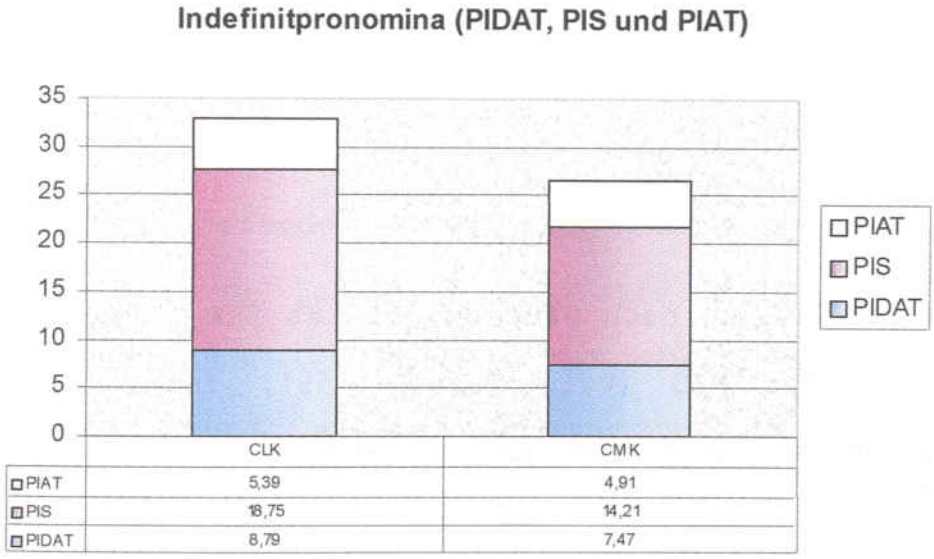
Indefinitpronomina (PI)

STTS teilt Indefinitpronomina in substituierende (PIS) und attribuierende (PIAT und PIDAT) ein. Bei den attribuierenden wird die Unterscheidung gemacht, ob sie mit (PIDAT) oder ohne (PIAT) Determiner auftreten können, wobei unter Determiner der unbestimmte oder bestimmte Artikel verstanden wird oder andere Pronomina davor oder dahinter.

PIS sind z.B. *etwas, nichts, irgendwas, (irgend)wer* und *man*; PIAT sind *etliche [Dinge], zuviele [Fragen]* oder *etwas [Schokolade]*; PIDAT sind *all [die Bücher], solche [eine Frage], beide [Fragen]* oder *viele [Leute]*.

Hentschel/Weydt (1990: 227ff) geben eine semantische Definition für Indefinitpronomina, die „das Merkmal gemeinsam haben, eine unbestimmte Menge, Art, Eigenschaft, unbestimmte

Umstände u.ä. auszudrücken". Helbig/Buscha (1991: 234) vertreten mit Hentschel/Weydt die Ansicht, daß sich die Gruppen je nach semantischem Kriterium ändern. Die Unvereinbarkeit der beiden Ansätze, distributionell im STTS, semantisch bei den genannten Grammatiken, ließ es ratsam erscheinen, Indefinitpronomina zunächst kumulativ darzustellen, d.h. alle drei Gruppen zusammengefaßt.



	CLK	CMK
PIAT	5,39	4,91
PIS	18,75	14,21
PIDAT	8,79	7,47

TABELLE 68

Indefinitpronomina sind mit insgesamt 32.93‰ im CLK überrepräsentiert; das CMK erreicht einen kumulativen Wert von 26,59‰ (vgl. Tabelle 68). Diese Tendenz zu höheren Frequenzen teilen die Indefinitpronomina mit fast all den Pronominalgruppen.

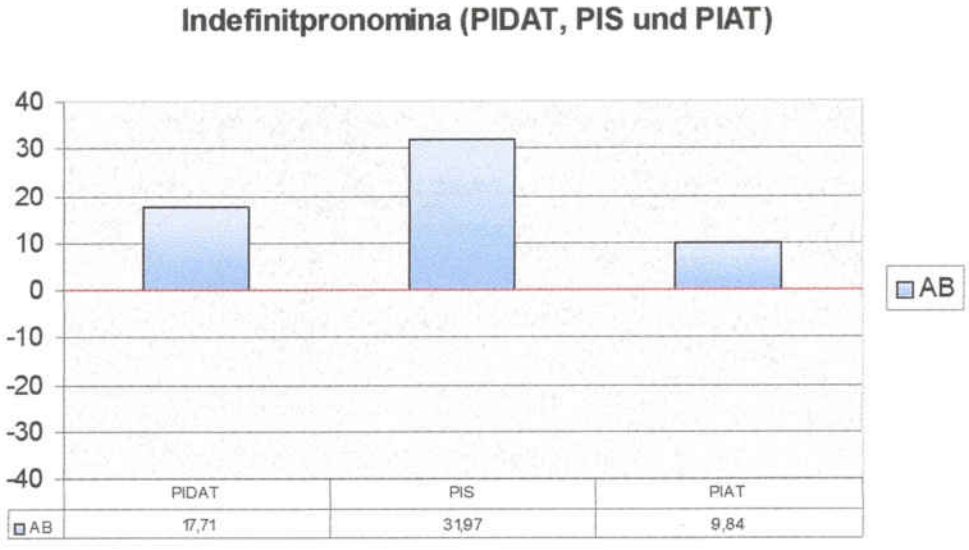


TABELLE 69

Dabei liegt die Abweichung nach Gruppen bei 18% für PIDAT, 10% für PIAT und 32% für PIS (vgl. Tabelle 69). Individuelle Unterschiede können anhand der Einzelelemente der Gruppen erkannt werden.

So weisen die substituierenden Indefinitpronomina die allgemeine Tendenz zur Überrepräsentation im CLK auf, wobei der Großteil der Abweichung auf das hochfrequente unpersönliche *man* zurückzuführen ist (vgl. Tabelle 70). Im unteren Bereich der Frequenzen ist die Tendenz zu mehr Variation im CMK erkennbar, wenn auch nicht zu allgemein höheren Werten.

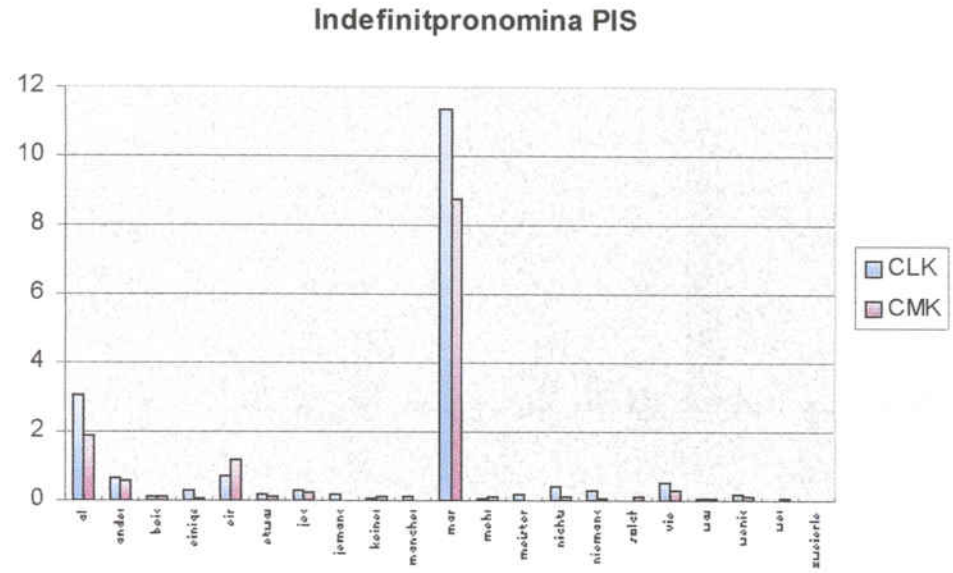
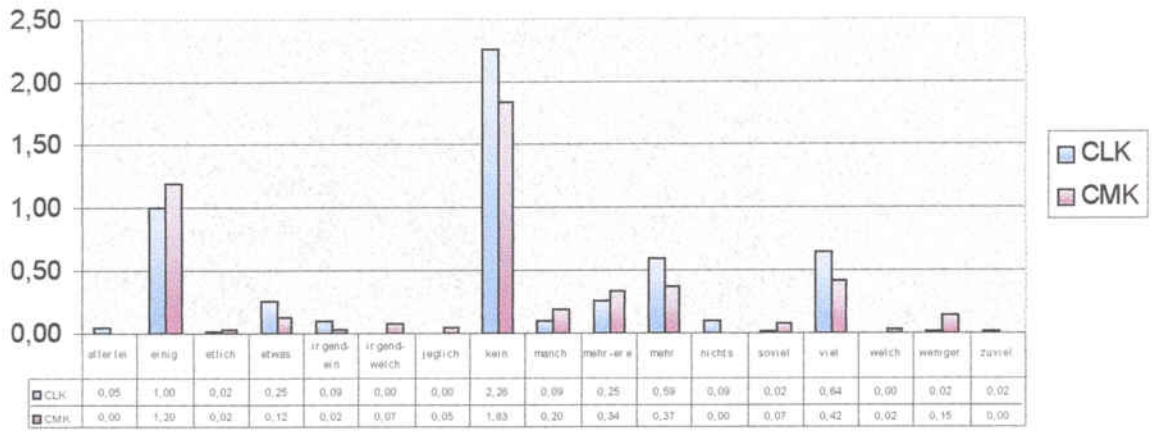


TABELLE 70

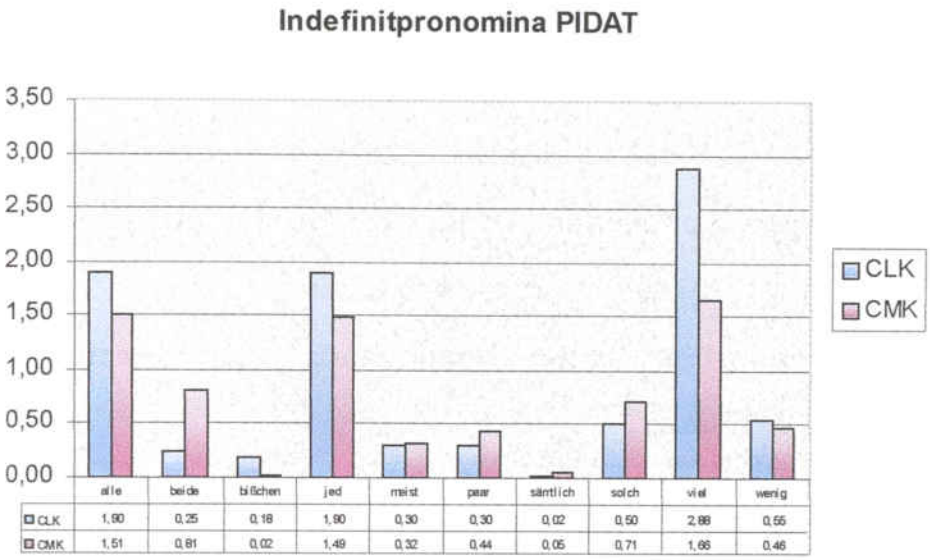
Die attribuierenden Indefinitpronomina ohne Determiner, PIAT, spiegeln diese Tendenzen ebenfalls wider. Im oberen Bereich der Frequenzen ist PIAT im CLK überrepräsentiert (*kein*); im mittleren und unteren Frequenzbereich erscheinen jedoch Inversionen dieser Tendenz (*einige, manche, mehrere*) (vgl. Tabelle 71).



	allerlei	einig	etlich	etwas	irgendein	irgendwelch	jeglich	kein	manch	mehr-ere	mehr	nichts	soviel	viel	welch	weniger	zuviel
CLK	0,05	1,00	0,02	0,25	0,09	0,00	0,00	2,26	0,09	0,25	0,59	0,09	0,02	0,64	0,00	0,02	0,02
CMK	0,00	1,20	0,02	0,12	0,02	0,07	0,05	1,83	0,20	0,34	0,37	0,00	0,07	0,42	0,02	0,15	0,00

TABELLE 71

Das attribuierende Indefinitpronomen mit Determiner, PIDAT, stellt keine Ausnahme zur allgemeinen Tendenz dar. Im oberen Frequenzbereich ist es im CLK überrepräsentiert (*all-, jed-, viel-*), im unteren kann diese Tendenz sich umkehren (*beid-, meist, paar, solch*).



	alle	beide	bißchen	jed	meist	paar	sämtlich	solch	viel	wenig
CLK	1,90	0,25	0,18	1,90	0,30	0,30	0,02	0,50	2,88	0,55
CMK	1,51	0,81	0,02	1,49	0,32	0,44	0,05	0,71	1,66	0,46

TABELLE 72

Dabei gibt es keine Unterschiede in der Variation, da alle Einzelelemente in beiden Korpora vertreten sind.

Indefinitpronomina stellen keine Ausnahme zur allgemeinen Tendenz der Pronomina dar: sie sind als Gruppe im CLK überrepräsentiert, zeigen allerdings weniger Variation als im CMK auf und tendieren oft bei niedrigen und demnach statistisch weniger repräsentativen Frequenzen zur Unterrepräsentation im CLK.

Reflexive Personalpronomina (PRF)

In STTS werden reflexive Personalpronomina im weitesten Sinne, denn sie umfassen *sich, einander, mich* u.a., als PRF

ausgezeichnet. Diese Formen stimmen mit denen von Helbig/Buscha (1991: 64ff) überein, die sie auch Reflexivpronomina nennen und zwischen Verben unterscheiden, bei denen das Reflexivpronomen obligatorisch oder nicht obligatorisch steht. Diese Unterscheidung, wenn auch wünschenswert, ist in STTS nicht vorgesehen und demzufolge nicht gesondert ausgezeichnet.

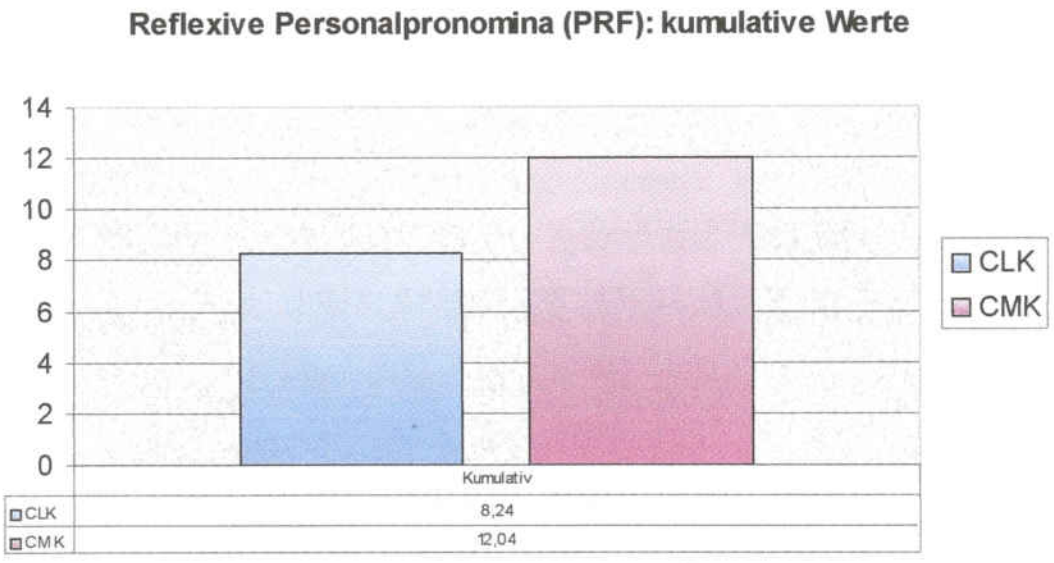
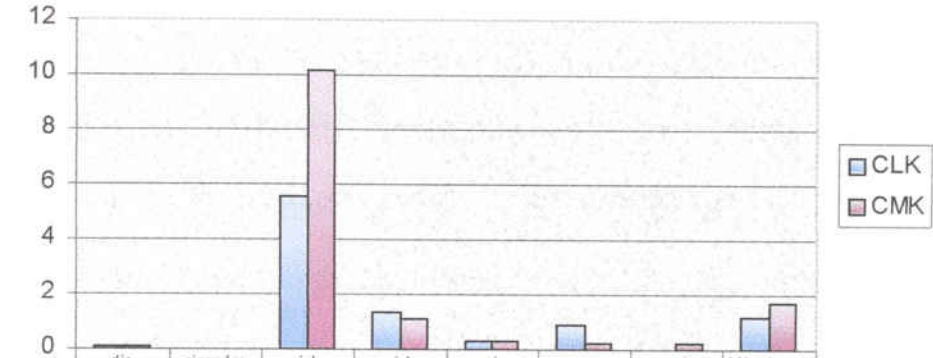


TABELLE 73

Die kumulativen Werte der reflexiven Personalpronomina zeigen eine starke Unterrepräsentation im CLK an, die sich aber hauptsächlich aus dem Pronomen *sich* ergibt (vgl. Tabelle 74).

Reflexive Personalpronomina (PRF): individuelle Verteilung



	dir	einander	sich	mich	mir	uns	euch	Mittelwert
CLK	0,09	0,02	5,53	1,37	0,32	0,91	0,00	1,18
CMK	0,05	0,02	10,13	1,07	0,27	0,24	0,24	1,72

TABELLE 74

Die starke Unterrepräsentation von *sich* im CLK könnte auf die Unterscheidung zwischen obligatorischem und nicht obligatorischem Auftreten des Pronomens zurückzuführen sein. Demnach wäre *sich* prozentuell stärker mit den reflexiven Verben im engeren Sinne verbunden (*reflexiv*), während die anderen reflexiven Formen als Hinweis auf die Anwendung reflexiver (*reziproker*) Konstruktionen zu gelten hätten, bei denen das Pronomen dekliniert erscheint (vgl. Corcoll/Corcoll 1994: 55ff). Im unteren Bereich der Frequenzen sind die reflexiven Personalpronomina im CLK überrepräsentiert, was aber hauptsächlich auf die Frequenz der Pronomina der 1. Person zurückzuführen ist, die im allgemeinen im CLK überrepräsentiert ist (*mich, mir, uns*).

Possessivpronomina (PPOS)

Possessivpronomina werden in STTS in attribuierende (PPOSAT: *seine [Meinung]*) und substituierende (PPOSS: *[das ist] meins*) eingeteilt, weichen aber als Gesamtgruppe gesehen nicht von der Einteilung der Grammatiken von Helbig/Buscha und Hentschel/Weydt ab.

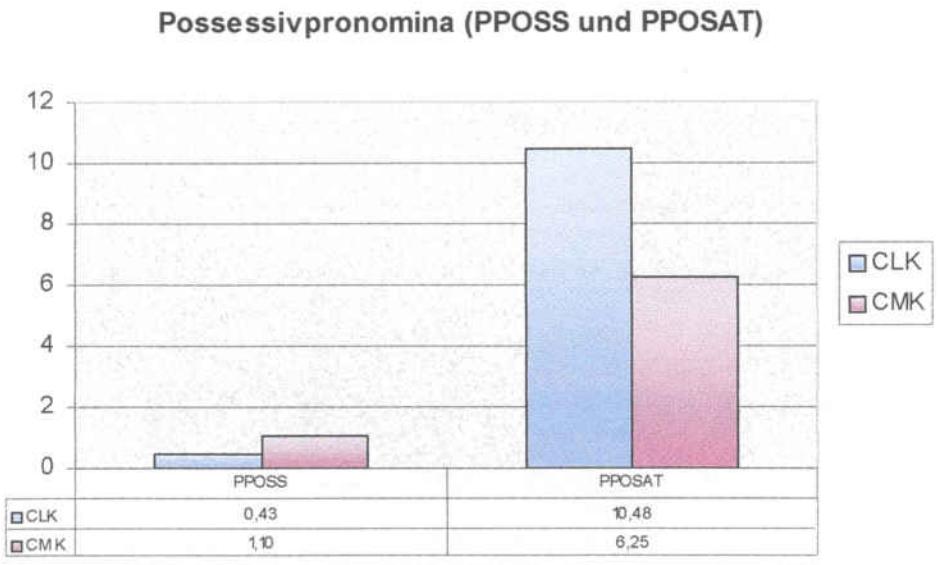


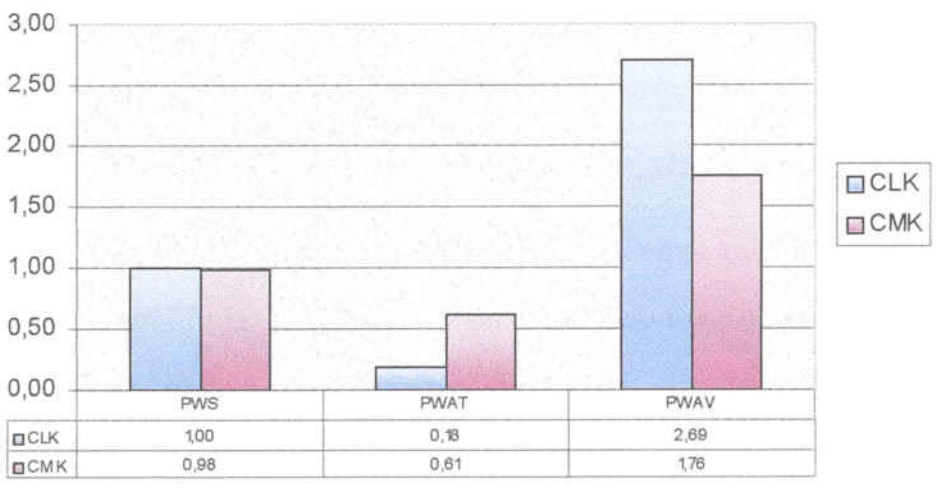
TABELLE 75

Die Untersuchung ihrer Verteilung zeigt, daß im CLK attribuierende Possessivpronomina stark überrepräsentiert sind, mit einer Abweichung, die 67,68% im Vergleich zum CMK beträgt. Substituierende Possessivpronomina hingegen sind unterrepräsentiert im CLK, mit einer negativen Abweichung von -39,09%. Unter- und Überrepräsentation scheinen hier demnach

mit der syntaktischen Funktion des Possessivpronomens verbunden zu sein.

Interrogativpronomina (PW)

Zur Gruppe der Interrogativpronomina gehören nach STTS substituierende Interrogativpronomina (PWS), attribuierende Interrogativpronomina (PWAT) und adverbiale Interrogativ- oder Relativpronomina (PWAV). Die letzte Gruppe, in Schiller et al. (1995) gesondert dargestellt, besteht aus mit *w-* beginnenden Adverbien (*wann, wo, woher, wohin, wieso, weshalb, warum* und die Adverbien gebildet aus *wo(r)-* und Präposition, wie *worüber, wobei, womit usw.*). Diese distributionelle Einteilung weicht von der gängigen semantischen Einteilung nach erfragter Sachverhaltskomponente ab. Des weiteren werden in dieser Gruppe auch die von Helbig/Buscha (1991: 266) als Pronominaladverbien bezeichneten Verbindungen von *wo(r) + Präposition* eingebunden, die nicht mehr in der Gruppe der Pronominaladverbien (PAV) von STTS erscheinen (vgl. 4.2.2.8).



	PWS	PWAT	PWAV
CLK	1,00	0,18	2,69
CMK	0,98	0,61	1,76

TABELLE 76

Kumulativ gesehen sind Interrogativpronomina leicht überrepräsentiert im CLK, mit einem Wert von 3,87‰, im Vergleich zu 3,35‰ im CMK. Abgesehen jedoch von den substituierenden Interrogativpronomina, die vergleichbare Frequenzen in beiden Korpora aufweisen, erscheinen Unterschiede bei den attribuierenden Interrogativpronomina und bei den adverbialen Interrogativ- oder Relativpronomina. Aufgrund der geringen Frequenzen werden hier weder PWS noch PWAT besprochen. Anstatt dessen wird näher auf die individuelle Verteilung von PWAV eingegangen.

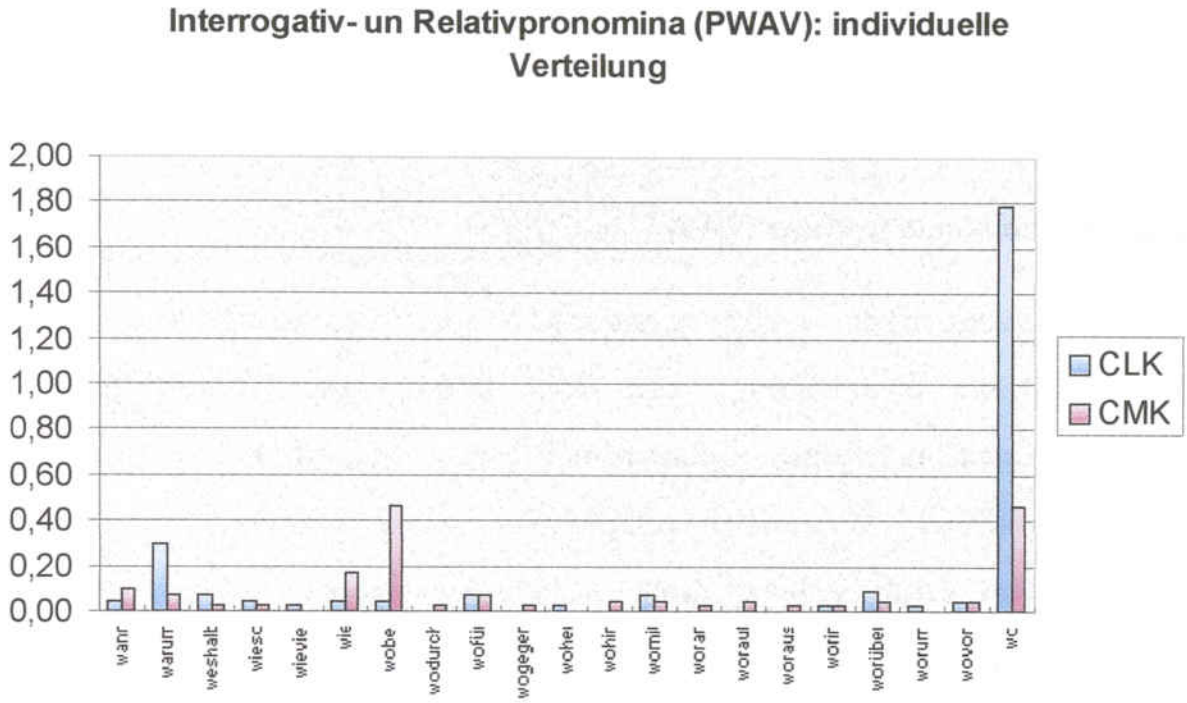


TABELLE 77

Die große positive Abweichung von PWAV im CLK, die kumulativ bei 52,84% liegt, ergibt sich aus der starken Überrepräsentation von *wo*, mit einer Abweichung von 286,96%. In dem Bereich der mittleren Frequenzen ist ein weiteres Element überrepräsentiert (*warum*), ein anderes unterrepräsentiert (*wobei*). Im unteren Bereich der Frequenzen, der unter 0,1‰ liegt, also statistisch wenig präzise ist, kann nur eine größere Variation im CMK vermerkt werden, in dem *wodurch, wogegen, wohin, woran, worauf* und *woraus* vertreten sind, die nicht im CLK erscheinen, während in diesem nur *wieviel, woher* und *worum* allein erscheinen. Die Elemente, die ausschließlich im CMK vertreten sind, haben größtenteils die

Eigenschaft, daß sie eine Präposition enthalten, die von einem Substantiv oder Verb regiert wird.

4.2.2.8 Pronominaladverbien (PAV)

Pronominaladverbien werden innerhalb von STTS als „eine Klasse von Adverbien bezeichnet, die sich aus einer Präposition und einem Pronominalstamm zusammensetzen. Sie treten im Satz anstelle einer Präpositionalphrase als Adverbialbestimmung oder Präpositionalobjekt auf" (Schiller et al. 1995: 53). Zu dieser Gruppe gehören Elemente wie *darauf, daher, hierzu, trotzdem, deswegen* oder *außerdem*. Sie umfaßt folglich auch Konjunktionaladverbien wie *deshalb, danach* oder *trotzdem*.

Die allgemeine Tendenz der Abweichung, die im Falle der Pronominaladverbien beobachtet werden kann, verweist auf höhere Werte des Phänomens im CMK als im CLK. Dabei ist der kumulative Wert im CMK um ugf. 30% höher, wie man aus Tabelle 78 entnehmen kann.

Pronominaladverbien (PAV): kumulative Verteilung

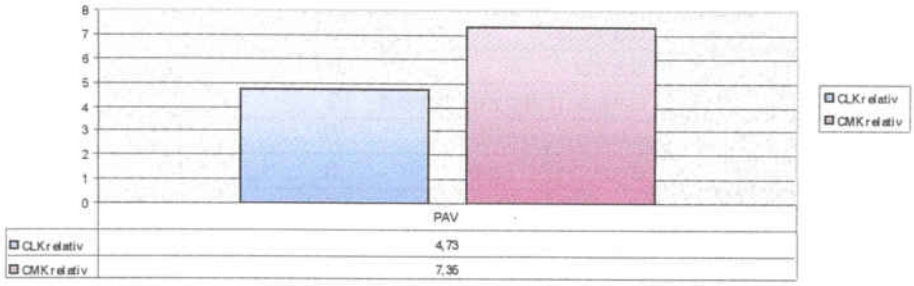


TABELLE 78

Die Untersuchung der einzelnen Elemente dieser Gruppe zeigt jedoch deutliche Abweichungen in der Frequenz der Pronominaladverbien. Zunächst ist darauf hinzuweisen, daß nicht nur die Frequenz im CMK höher ist, sondern daß dort auch mehr Pronominaladverbien angewendet werden, die im CLK nicht erscheinen; vgl. dazu Tabelle 79. Es handelt sich dabei um absolute Werte, da die Frequenzen sehr gering sind.

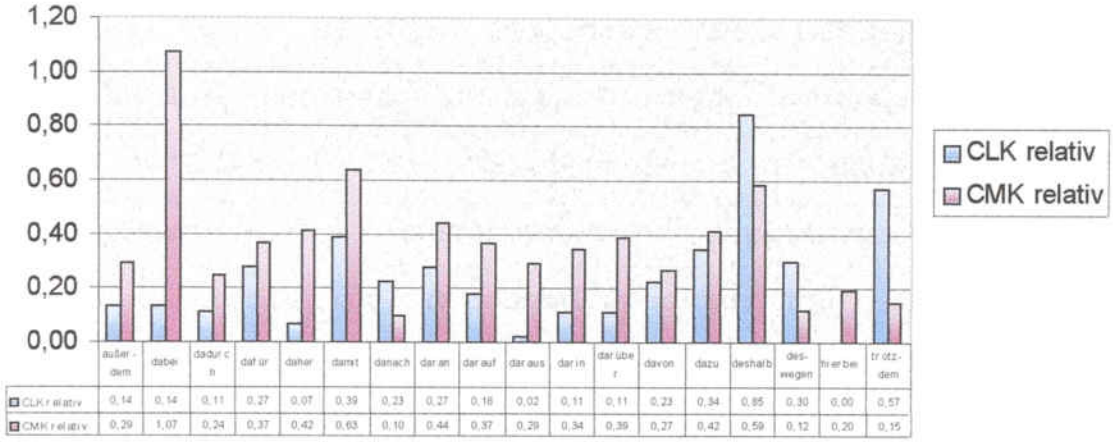
PAV	CLK absolut	CMK absolut
außerdem	6	12
dabei	6	44
dadurch	5	10
dafür	12	15
dagegen	6	4
daher	3	17
dahin	1	0
dahingegen	0	1
damit	17	26
danach	10	4
daran	12	18
daraufhin	1	1
darauf	8	15
daraus	1	12
darin	5	14
darüber	5	16
davon	10	11
dazu	15	17
dementsprechend	1	2
demgegenüber	0	2
demzufolge	0	1
deshalb	37	24
deswegen	13	5
hieran	0	1
hieraus	0	1
hierbei	0	8
hiervon	0	1
hierzu	0	3
seitdem	1	0
trotzdem	25	6
währenddessen	0	1

TABELLE 79

Auffallend ist dabei, daß im CLK kein einziges Pronominaladverb mit *hier-* erscheint (*hieran, hieraus, hierbei, hiervon* und *hierzu* erscheinen nur im CMK). Sie sind zwar relativ wenig frequent als Einzelelement, mit Frequenzen im CMK zwischen 1 und 8 Belegen, doch zusammengefaßt sind sie mit 14 Belegen bedeutend.

In der folgenden Darstellung der Pronominaladverbien wurden nur jene berücksichtigt, die in einem der beiden Korpora mindestens eine Frequenz von 0,20‰ aufwiesen.

Pronominaladverbien (PAV): individuelle Verteilung



	außer-dem	dabei	dadurch	dafür	daher	damit	danach	daran	darauf	daraus	darin	darüber	davon	dazu	deshalb	deswegen	hierbei	trotzdem
CLK relativ	0,14	0,14	0,11	0,27	0,07	0,39	0,23	0,27	0,18	0,02	0,11	0,11	0,23	0,34	0,85	0,30	0,00	0,57
CMK relativ	0,29	1,07	0,24	0,37	0,42	0,63	0,10	0,44	0,37	0,29	0,34	0,39	0,27	0,42	0,59	0,12	0,20	0,15

TABELLE 80

Eine mögliche Erklärung für dieses distributive Schema ist die semantische Klarheit der angewendeten Elemente und ihre Didaktisierung als prototypische Konnektoren, da im CLK höhere Werte hauptsächlich von Homonymen der Konjunktionaladverbien erzielt werden (*danach, deshalb, trotzdem...*). Im CLK würden deshalb Pronominaladverbien wie *deshalb, danach, deswegen* und *trotzdem* häufig benutzt, da sie als Konnektoren klar identifizierbar sind und von keiner verbalen oder substantivischen Rektion abhängen. Aus diesem Grund würden dann andere wie *dabei, darüber* usw. vermieden, da für ihre Anwendung die Rektion des entsprechenden Verbs oder Substantivs beachtet werden muß. Zwischen diesen beiden Polen existieren dann klarere und weniger klare Elemente, deren Rektion gut bekannt ist (dazu, davor lokal, usw.); *außerdem* ist eine Ausnahme.

4.2.2.9 Verben

Aufgrund der vielfältigen verbalen Tags in STTS ist eine detaillierte Analyse der Distributionseigenschaften der verbalen Komponenten in den Korpora möglich. In STTS werden 12 individuelle Verbklassen unterschieden, die aus drei Hauptgruppen und 5 Untergruppen bestehen (vgl. Schiller et al. 1995: 28). Die drei Hauptgruppen bestehen aus den Modalverben (VM) *können, müssen, wollen, dürfen, mögen* und *sollen*; den potentiellen Auxiliaren (VA) *haben, sein* und *werden*, auch wenn sie innerhalb des Satzes als Vollverb benutzt werden; und schließlich den restlichen Verben, die als Vollverben (VV) klassifiziert werden. Weitere Unterschiede werden in STTS nicht angegeben, so daß modalverbähnlich zu verwendende Verben („verbos de modalidad", vgl. Castell 1997: 172ff) wie *drohen, lassen, sehen* u.a. nicht gesondert ausgezeichnet werden.

Die fünf Untergruppen erlauben eine Einteilung der drei Hauptgruppen in *Imperativformen* (IMP), *Finite Formen* (FIN), *Partizipien* (PP), *Infinitive* (INF) und *Infinitive mit zu* (IZU). Imperativformen erhalten in diesem System eine eigene Klasse, da sie sich distributionell von anderen finiten Verbformen unterscheiden (Schiller et al. 1995: 28). Die Kombination dieser Gruppen führt zu der folgenden endgültigen Einteilung der Verben in STTS:

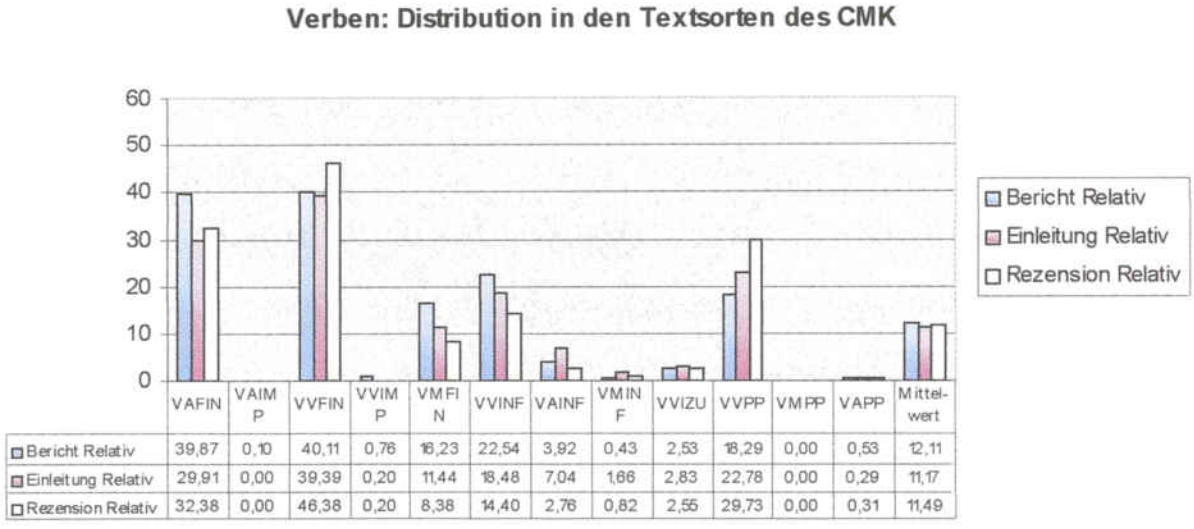
	Modalverb VM	Potentieller Auxiliar VA	Vollverb VV
Imperativ IMP		VAIMP	VVIMP
Finite Formen FIN	VMFIN	VAFIN	VVFIN
Partizip PP	VMPP	VAPP	VVPP
Infinitiv INF	VMINF	VAINF	VVINF
Infinitiv mit zu IZU			VVIZU

Anzumerken ist im Falle der Partizipien, daß weder zwischen aktivischem, passivischem oder prädikativem Gebrauch unterschieden wird, und daß aufgrund distributioneller Kriterien adverbial benutzte Partizipien als prädikatives Adjektiv (ADJD) getaggt werden (Schiller et al. 1995: 31).

Verben: Allgemeine Darstellung

Die vergleichende Darstellung der Textsorten des CMK zeigte starke Schwankungen im Bereich einiger verbaler Tags auf. In den Berichten liegt der Anteil an finiten Hilfsverben (VAFIN) und finiten Modalverben (VMFIN) höher als in den anderen Textsorten. Rezensionen zeigen höhere Frequenzen im Bereich der finiten Vollverben (VVFIN) und der partizipialen Vollverben (VVPP) an, während sich die Werte der Einleitungen bis auf die Ausnahme der wenig frequenten auxiliaren Infinitive in einem mittleren Bereich bewegen. Eine positive Merkmalsausprägung (Überrepräsentation) ist hier aufgrund des Datenmaterials hinsichtlich der finiten Auxiliare in den Berichten und der finiten Vollverben in der Rezension und den partizipialen Vollverben zu erkennen. Eine negative

Merkmalsausprägung (Unterrepräsentation) stellt die reduzierte Anwendung von partizipialen Vollverbformen im Bericht und von finiten Modalverben in der Rezension dar (vgl. Tabelle 81).



	VAFIN	VAIMP	VVFIN	VVIMP	VMFIN	VVINF	VAINF	VMINF	VVIZU	VVPP	VMPP	VAPP	Mittelwert
Bericht Relativ	39,87	0,10	40,11	0,76	16,23	22,54	3,92	0,43	2,53	18,29	0,00	0,53	12,11
Einleitung Relativ	29,91	0,00	39,39	0,20	11,44	18,48	7,04	1,66	2,83	22,78	0,00	0,29	11,17
Rezension Relativ	32,38	0,00	46,38	0,20	8,38	14,40	2,76	0,82	2,55	29,73	0,00	0,31	11,49

TABELLE 81

Der mittlere Wert der gesamten verbalen Formen jedoch, ebenfalls in Tabelle 81 dargestellt, ist mit 12,11% im Bericht, 11,17% in der Einleitung und 11,49% in der Rezension relativ stabil, wobei die größere Differenz des Berichts zu der Einleitung und der Rezension Indiz für einen Stil ist, den wir als "stark verbal geprägt" bezeichnen können. Die größeren Differenzen zwischen den Textsorten werfen jedoch einige Fragen der Kongruenz auf, die im folgenden besprochen werden.

Zunächst sollte ein höherer Wert an auxiliaren finiten Verbformen in den Berichten einem höheren Wert an Partizipien entsprechen, doch die ermittelten Werte spiegeln dies nicht wider. Da STTS die Verben *sein, haben* und *werden* unabhängig von ihrer auxiliaren oder finiten Verwendung als potentielle Auxiliare taggt (VA), wurde eine manuelle Differenzierung dieser Tags durchgeführt. Der Grund für die fehlende Entsprechung zwischen VAFIN und VVPP im Bericht liegt in der hohen Verwendung von *sein* nicht in seiner auxiliaren Form, sondern als Vollverb:

	Bericht	Einleitung	Rezension
sein	24,06	12,61	14,20
haben	7,59	5,08	3,68
werden	8,40	12,22	14,61

TABELLE 82: VAFIN IM BERICHT DES CMK

Ebenfalls höher im Bericht sind die Werte der infinitiven Formen der Vollverben (VVINF), was mit der höheren Frequenz von finiten Modalverben (VMFIN) in Verbindung gebracht werden kann.

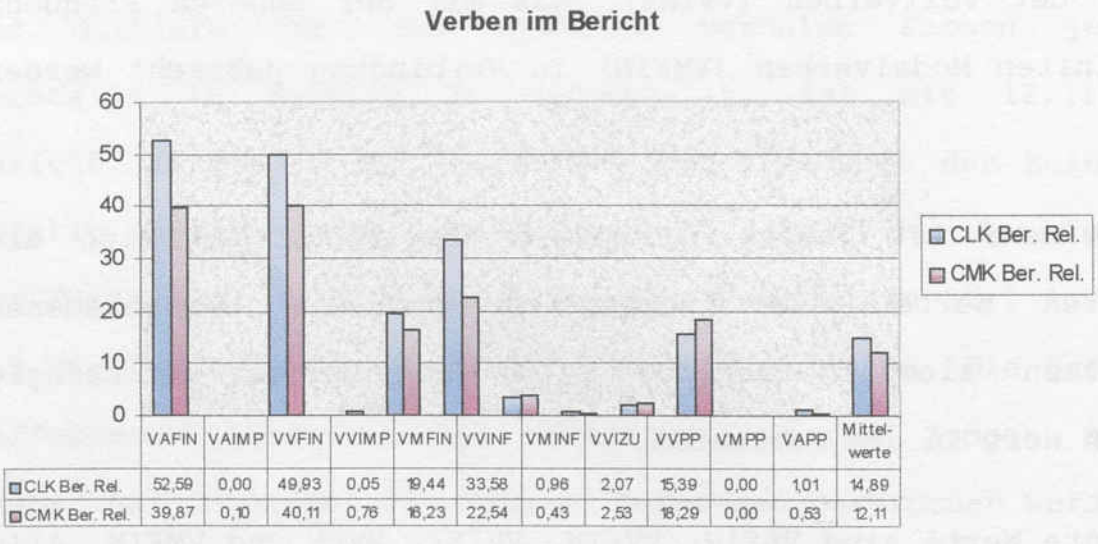
Entscheidend ist jedoch in diesem Vergleich, daß sich die mittleren Werte aller Textsorten für die verschiedenen Verbformen ziemlich ähnlich sind, mit einem geringfügig höheren Wert in den Berichten.

Relevante Werte sind VAFIN, VVFIN, VVINF, VVPP und VMFIN. Alle anderen weisen sehr geringe Frequenzen auf und haben somit

keinen Einfluß auf die Differenzen im CMK und die Abweichung zum CLK.

Betrachten wir die Textsorte Bericht im Vergleich zwischen CLK und CMK, sehen wir, daß die Werte des CLK generell der Tendenz von denen des CMK folgen, doch größtenteils mit Abweichungen verbunden. So sind ebenfalls die finiten Formen der Hilfsverben (VAFIN) die häufigsten, zusammen mit den finiten Formen der Vollverben (VVFIN) (vgl. Tabelle 83). Die Werte dieser beiden Verbklassen liegen jedoch deutlich über denen des CMK, mit einer Abweichung, die bis zu 20% beträgt.

Im CLK ist eine Tendenz zu einer proportional höheren Verwendung von VAFIN und VVFIN zu erkennen, zusammen mit einer Überrepräsentation von VVINF. Nur im Falle von VVPP erscheinen deutlich höhere Werte im CMK.

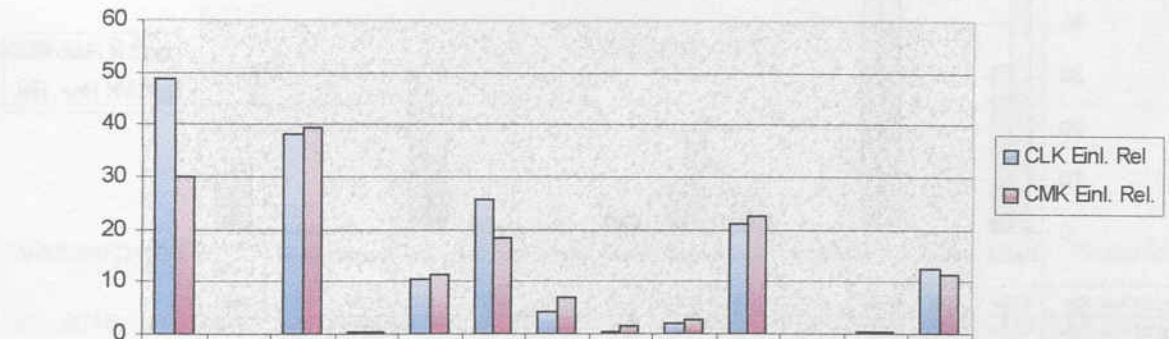


	VAFIN	VAIMP	VVFIN	VVIMP	VMFIN	VVINF	VMINF	VVIZU	VVPP	VMPP	VAPP	Mittel-werte
CLK Ber. Rel.	52,59	0,00	49,93	0,05	19,44	33,58	0,96	2,07	15,39	0,00	1,01	14,89
CMK Ber. Rel.	39,87	0,10	40,11	0,76	16,23	22,54	0,43	2,53	18,29	0,00	0,53	12,11

TABELLE 83

In der Einleitung im CMK ändern sich die Werte von VAFIN und VVPP; im Gegensatz zum Bericht findet hier ein Ausgleich statt (vgl. Tabelle 84), d.h. während VAFIN sinkt, steigt gleichzeitig VVPP, was auf eine engere Verbindung zwischen potentiellen Auxiliaren und Partizipien deutet. Mit reduzierteren Werten erscheint diese Tendenz auch im CLK. Die Werte für VVFIN im CMK bleiben mit denen des Berichts vergleichbar, während die von VMFIN im CMK sinken. Im CLK hingegen fallen die Werte für VVFIN, ebenso wie die von VMFIN. Parallel in beiden Korpora sinken zudem die Werte für VVINF.

Verben in der Einleitung

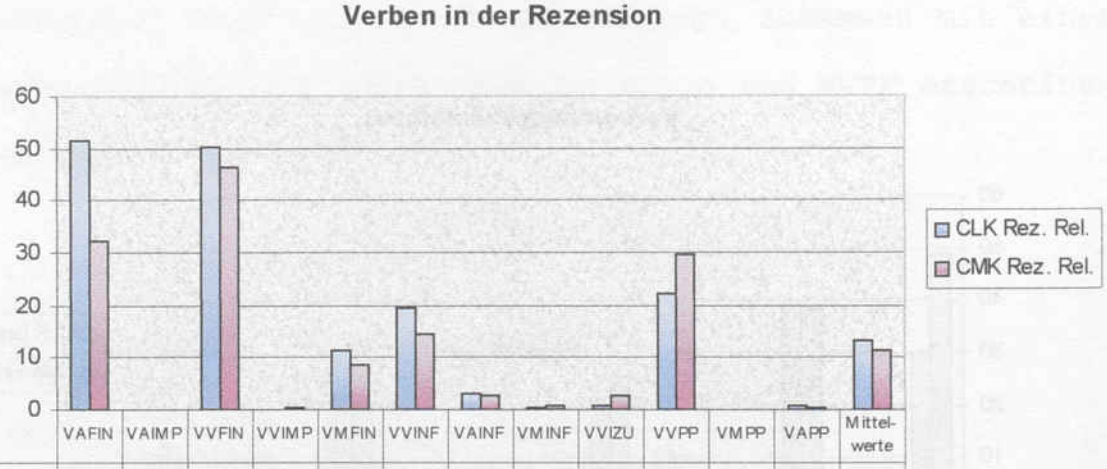


	VAFIN	VAIMP	VVFIN	VVIMP	VMFIN	VVINF	VAINF	VMINF	VVIZU	VVPP	VMPP	VAPP	Mittel-werte
CLK Einl. Rel	48,89	0,00	38,20	0,45	10,69	25,69	4,32	0,45	2,05	20,92	0,00	0,23	12,66
CMK Einl. Rel.	29,91	0,00	39,39	0,20	11,44	18,48	7,04	1,66	2,83	22,78	0,00	0,29	11,17

TABELLE 84

Die allgemeinen Tendenzen für das CMK erscheinen akzentuiert in der Rezension des CMK. VAFIN geht eine engere Verbindung

mit VVPP ein, da proportionell mehr VAFIN mit VVPP verbunden scheinen. Diese Tendenz spiegelt sich aber nicht im CLK wider, bei dem keine Veränderung der Relation VAFIN - VVPP im Vergleich zur Einleitung zu erkennen ist. Und während im CMK VMFIN weiterhin sinkt, bleibt der Wert von VMFIN im CLK konstant im Vergleich zur Einleitung. Andererseits steigen im CLK erneut die Werte für VVFIN, was auch im CMK zu verzeichnen ist. Schließlich ist zu beachten, daß die Werte für VVINF parallel in beiden Korpora hinsichtlich der Einleitung und des Berichts fallen (vgl. Tabelle 85).



	VAFIN	VAIMP	VVFIN	VVIMP	VMFIN	VVINF	VAINF	VMINF	VVIZU	VVPP	VMPP	VAPP	Mittel-werte
CLK Rez. Rel.	51,29	0,00	50,38	0,11	11,15	19,35	3,16	0,43	0,91	22,08	0,00	0,96	13,32
CMK Rez. Rel.	32,38	0,00	46,38	0,20	8,38	14,40	2,76	0,82	2,55	29,73	0,00	0,31	11,49

TABELLE 85

Die allgemeine Tendenz der Verteilung der Verben im CMK ist also eine progressive Annäherung von Auxiliaren und Partizipien von Bericht über Einleitung bis hin zu Rezension,

wobei parallel dazu der Wert der Modalverben und Infinitive fällt. Diese Tendenz des CMK spiegelt sich nicht ganz im CLK wider: die Annäherung von Auxiliaren und Partizipien geht nur bis zur Einleitung, da sie in der Rezension fast mit den Werten dieser Textsorte vergleichbar sind. Die Modalverben zeigen ebenfalls eine fallende Entwicklung auf, die jedoch auch bei der Einleitung anhält und sich nicht weiter in der Rezension auswirkt.

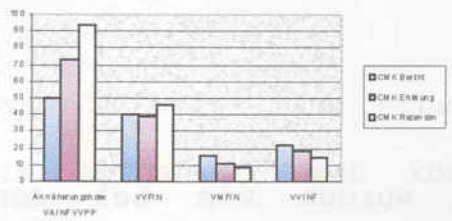


TABELLE 86

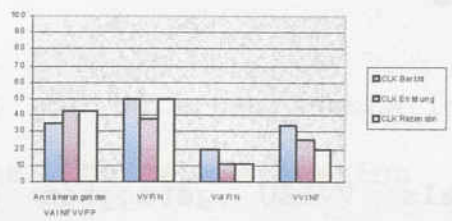


TABELLE 87

Zusammengefaßt sind diese Tendenzen in Tabelle 86 und Tabelle 87, aus denen hervorgeht, daß im CLK nur die Infinitive der Vollverben eine mit dem CMK vergleichbare Tendenz aufweisen, aber in allen Fällen abweichende Werte erscheinen.

Verben: Infinitive mit zu *(VVIZU)*

Eng mit den getrennt erscheinenden Verbzusätzen verbunden sind die Infinitivkonstruktionen mit *zu*, das zwischen dem verbalen

Präfix und dem Verb erscheint. Hier ist wie auch schon im Falle der Verbzusätze eine starke Abweichung zwischen CLK und CMK zu erkennen.

TABELLE 88

Diese als VVIZU getaggten Elemente wurden wie bei den Verbzusätzen PTKVZ auf die einzelnen Zusätze untersucht, um deren individuelle Frequenzen zu bestimmen (vgl. Tabelle 89).

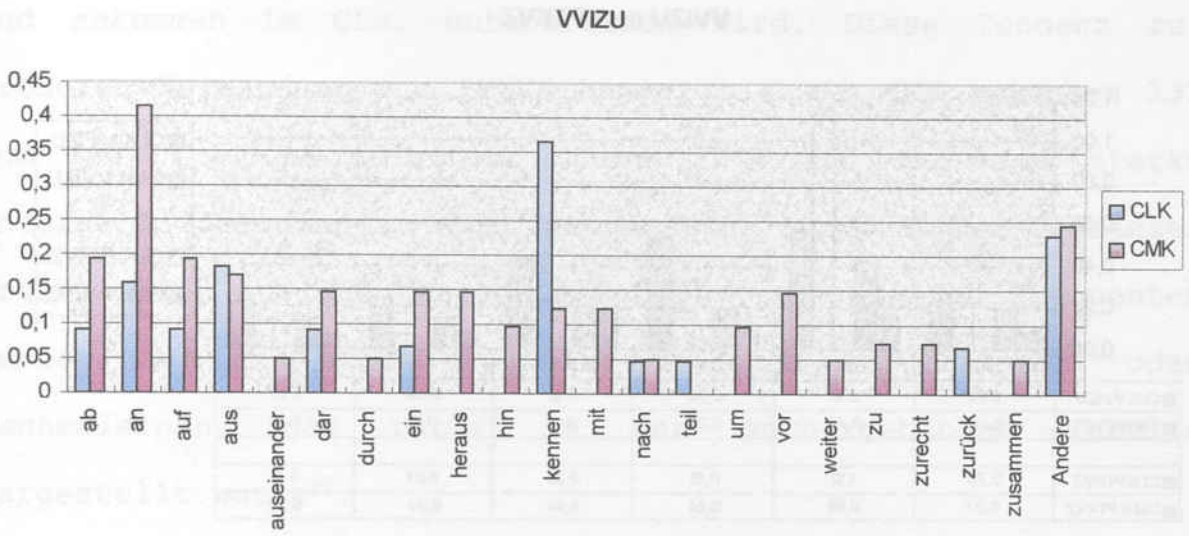
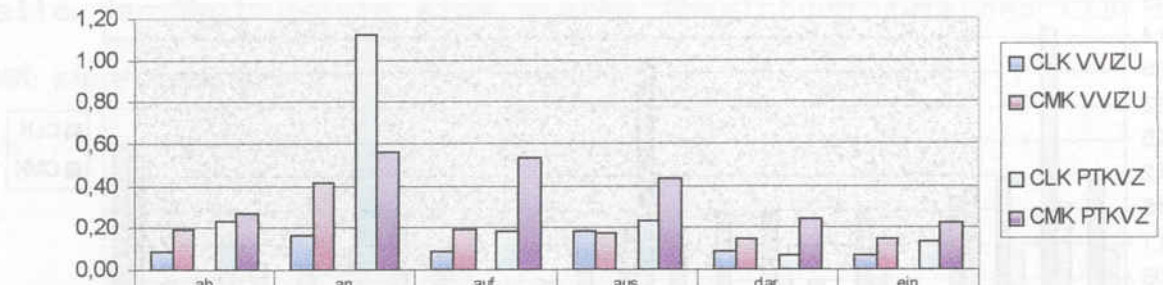


TABELLE 89

Das CLK weist hier eine Tendenz zur Unterrepräsentation auf. Die Ausnahmen dazu können eine Überrepräsentation im CLK darstellen (*kennen-*, *teil-*, und *zurück-*), oder vergleichbare Werte (*aus-* und *nach-*). Ebenso wie bei den Verbzusätzen PTKVZ ist hier eine größere Variation in den angewendeten Elementen zu verzeichnen. Im CMK erscheinen 19 sich wiederholende VVIZU, von denen 11 nicht im CLK auftreten; im CLK werden 11 VVIZU angewendet, von denen nur 2 nicht im CMK vertreten sind.



	ab	an	auf	aus	dar	ein
CLK VVIZU	0,09	0,16	0,09	0,18	0,09	0,07
CMK VVIZU	0,20	0,42	0,20	0,17	0,15	0,15
CLK PTKVZ	0,23	1,12	0,18	0,23	0,07	0,14
CMK PTKVZ	0,27	0,56	0,54	0,44	0,24	0,22

TABELLE 90

Tabelle 90 stellt einen Vergleich zwischen den Verbzusätzen VVIZU und PTKVZ dar, der dazu dient, Übereinstimmungen zwischen den Tendenzen zu verdeutlichen. Berücksichtigt wurden nur die Elemente, die Werte erzielten, d.h., wenn eines der Elemente (VVIZU und PTKVZ) in einem der Korpora (CLK oder CMK) keinen Wert hatte, wurde es nicht in den Vergleich einbezogen.

Die Übereinstimmung in beiden Korpora hinsichtlich der allgemeinen Tendenz der Unterrepräsentation im CLK erscheint bei den Zusätzen *ab-*, *auf-*, *aus-*, *dar-* und *ein*, obwohl dabei größere Frequenzunterschiede hervortreten.

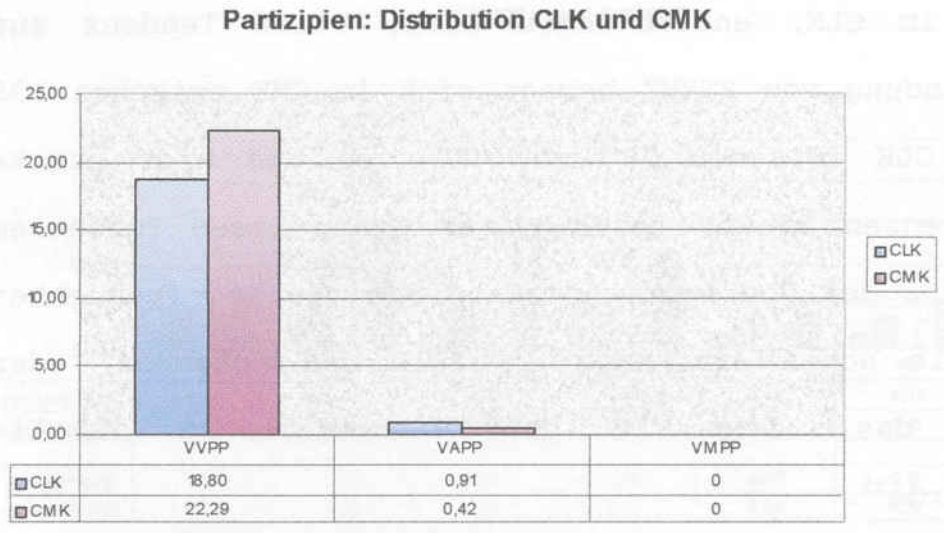
CMK und CLK verwenden erstens als allgemeine Tendenz mehr PTKVZ als VVIZU; das CLK zeichnet sich dabei durch eine große Unterrepräsentation dieser Elemente aus, die nur im Falle von *an-*, aufgrund der frequenten Verwendung der Verben *anfangen*

und *ankommen* im CLK, unterbrochen wird. Diese Tendenz zur größeren Verwendung von PTKVZ bewegt sich im CMK zwischen 135 und 280%, im CLK zwischen 78 und 700%, so daß hier starke interne Differenzen im CLK festzuhalten sind. Diese Tendenzen ergeben sich aus der Überrepräsentation von einigen frequenten Verben, wie die schon erwähnten *anfangen* und *ankommen*, oder *kennenlernen*, das nicht in der vorhergehenden Tabelle dargestellt wurde[69].

Verben: Partizipiale Formen (VVPP, VMPP, VAPP)

Die Verteilung der Partizipien im CMK und im CLK weist eine Abweichung von ugf. 16% zugunsten des CMK auf; die Werte basieren fast ausschließlich auf den Partizipien der Vollverben (VVPP), da die Partizipien der Modalverben und der Auxiliare sehr niedrigfrequente Phänomene darstellen (vgl. Tabelle 91).

[69] Vgl. Kapitel 5 zur Unterscheidung solcher Tendenzen im CLK unter Berücksichtigung des Sprachstands Mittel-Oberstufe

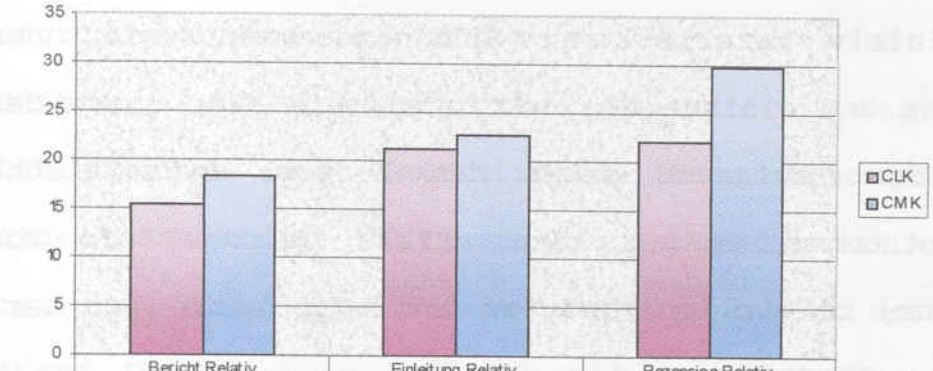


	VVPP	VAPP	VMPP
CLK	18,80	0,91	0
CMK	22,29	0,42	0

TABELLE 91

Die Verteilung zwischen den Textsorten zeigt, daß beide Korpora je nach Textsorte verschiedene Werte aufweisen. Während sich im CMK alle drei Textsorten hinsichtlich ihrer Werte unterscheiden, sind Differenzen beim CLK nur zwischen zwei Gruppen zu erkennen: zwischen dem Bericht einerseits und der Einleitung und der Rezension andererseits (vgl. Tabelle 92).

Partizipien: Verteilung nach Textsorten



	Bericht Relativ	Einleitung Relativ	Rezension Relativ
CLK	15,39	21,15	22,08
CMK	18,29	22,78	29,73

TABELLE 92

So erreicht der Bericht im CLK 15,39‰ VVPP, stabilisiert sich dann aber bei der Einleitung und der Rezension bei einem Wert von ugf. 22‰. Im CMK hingegen steigt der Wert konstant vom Bericht (18,29‰) über die Einleitung (22,78‰) bis hin zur Rezension (29,73‰).

4.2.2.10 Weitere Wortartentags

In diesem Abschnitt werden weitere Wortartentags dargestellt, die STTS auszeichnet, und die bei einer Analyse des Satzes, die jedem Wort nur ein Tag zuweist, als Restmenge nicht mit den morphologisch und syntaktisch verbundenen Wörtern in Beziehung gebracht werden können. Dabei handelt es sich um *zu* vor Infinitiv, die Negationspartikel *nicht*, der abgetrennte Verbzusatz, die Partikel bei Adjektiv oder Adverb und das Kompositions-Erstglied.

Zu vor Infinitiv (PTKZU)

Zu vor Infinitiv stellt in STTS zusammen mit den Relativpronomina und einigen Konjunktionen, wie *daß*, einen der wenigen Tags dar, aufgrund denen den Korpora syntaktische Information entnommen werden kann. PTKZU zeichnet *zu* vor Infinitiv (*[ohne] zu [wollen]*)und vor Partizip Futur (*[in der] zu [zerstörenden Stadt]*) aus.

Infinitivkonstruktionen (PTKZU): Verteilung

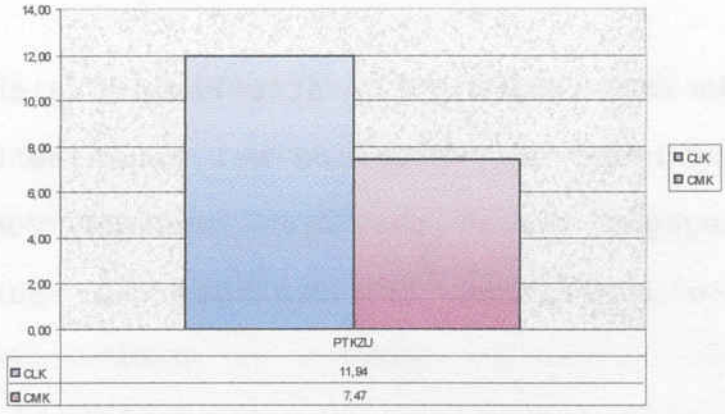


TABELLE 93

Die Verteilung von PTKVZ zeigt eine starke Überrepräsentation im CLK, das mit 11,94‰ 4,47‰ mehr Konstruktionen dieser Art erreicht als das CMK. Die positive Abweichung liegt dabei bei 59,83%. Diese Überrepräsentation ist mit der syntaktischen Struktur der Infinitivkonstruktion in Zusammenhang zu bringen, kann aber wegen noch nicht eingefügten syntaktischen Auszeichnung nicht weiter untersucht werden.

Negationspartikel (PTKNEG)

Aufgrund der Einschränkung von PTKNEG auf *nicht* könnte dieses Phänomen gleichfalls im Abschnitt der allgemeinen statistischen und lexikalischen Untersuchungen erscheinen. Seine Aussagekraft für unsere Untersuchung ist gering, so daß hier nur die allgemeinen Werte dargestellt werden. Für eine nähere Untersuchung wäre es mit anderen Negationselementen, wie die Adverbien *nie, nirgends* u.a. und mit negativen Präfixen zu verbinden.

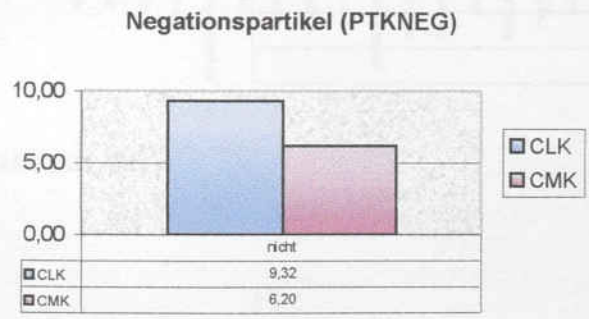


TABELLE 94

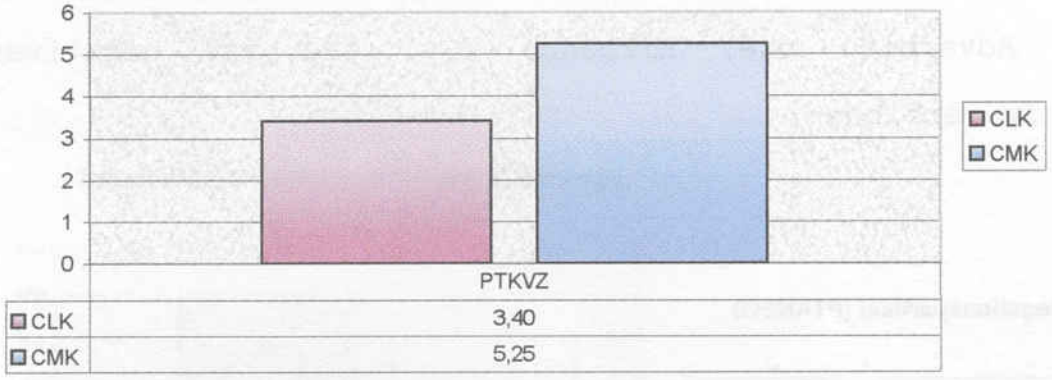
Die allgemeine Verteilung zeigt erneut eine starke Überrepräsentation im CLK, mit einer positiven Abweichung von 50,32%.

Abgetrennter Verbzusatz (PTKVZ)

Abgetrennte Verbzusätze sind in Verbindung zu bringen mit den verbalen Präfixen. Es handelt sich dabei um trennbare verbale

Präfixe, die getrennt vom Verb erscheinen, aber aufgrund der technischen Einschränkungen von STTS nicht mit dem Verb verbunden werden, also als selbständiges Wort zählen.

Abgetrennter Verbzusatz (PTKVZ): kumulative Verteilung



	PTKVZ
CLK	3,40
CMK	5,25

TABELLE 95

Die höheren Frequenzen im CMK stimmen mit den allgemeinen Werten der verbalen Präfixe überein. Da diese jedoch ebenfalls untrennbare Präfixe enthielten, wurden für die detaillierte Darstellung die einzelnen Verbzusätze miteinander verglichen. In Tabelle 96 sind alle Verbzusätze aufgelistet, die in einem der beiden Korpora mindestens eine Frequenz von 0,04‰ aufweisen (was 2 Treffern entspricht). Alle anderen, die nur ein einziges mal erschienen, wurden in der abschließenden Gruppe *Andere* zusammengefaßt, wodurch ebenso die Variation dieser Elemente dargestellt werden kann. Das CMK weist nicht nur in fast allen Fällen höhere Werte auf, sondern auch eine

größere Variation dieser Elemente, denn viele, die im CMK erscheinen, sind im CLK nicht zu finden. Dabei handelt es sich keineswegs um komplexe Elemente (vgl. Tabelle 96).

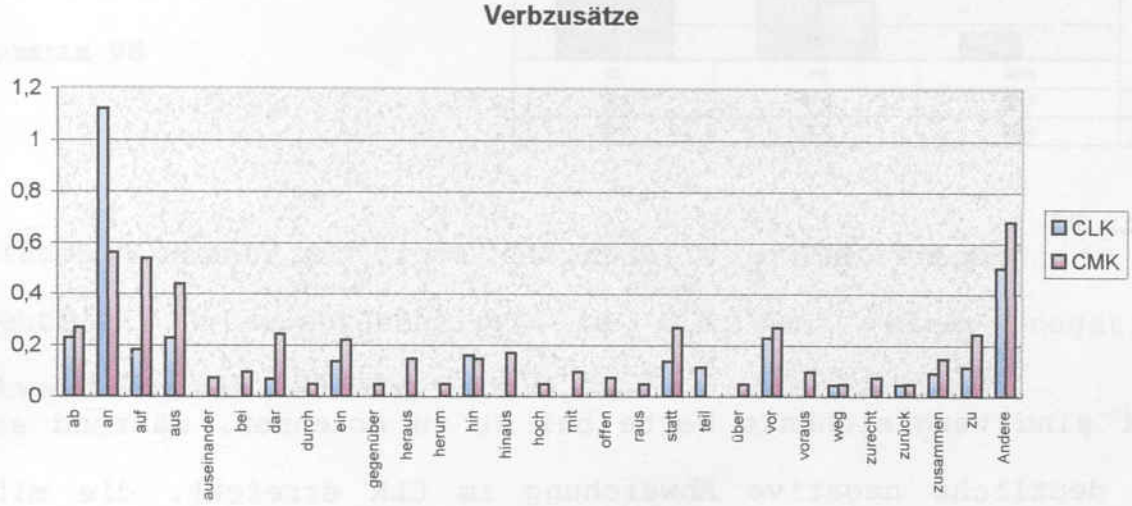
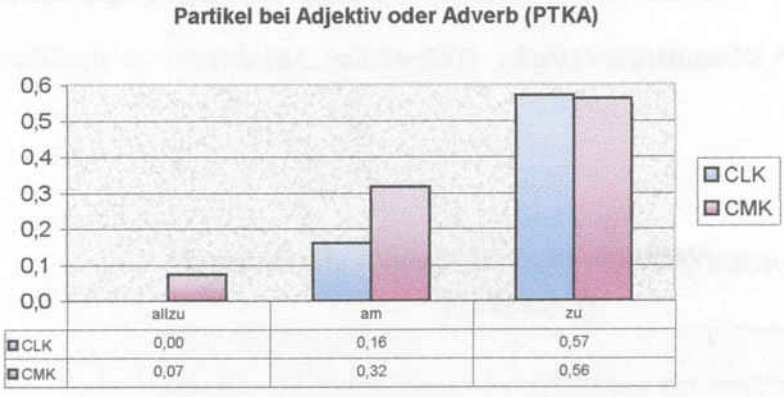


TABELLE 96

Bis auf zwei Ausnahmen weist das CMK höhere Werte auf. Die Ausnahmen betreffen *an-*, mit einer extrem hohen positiven Abweichung, und *hin-*, mit einer sehr negativen. Auffallend aber ist an den Frequenzen, daß viele Verbzusätze gar nicht im CLK erscheinen, die Variation im CMK also dementsprechend höher ist.

Partikel bei Adjektiv oder Adverb (PTKA)

PTKA umfaßt in STTS *am* vor Superlativ und *zu* oder *allzu* vor Adjektiv oder Adverb.



	allzu	am	zu
CLK	0,00	0,16	0,57
CMK	0,07	0,32	0,56

TABELLE 97

Dabei sind vergleichbare Werte bei *zu* zu erkennen, während *am* eine deutliche negative Abweichung im CLK erreicht, die mit der Verwendung von den in STTS nicht morphologisch ausgezeichneten Superlativen in Verbindung zu bringen ist. *Allzu* ist nicht repräsentiert im CLK, was allerdings auch auf die sehr niedrigen Frequenzen im Referenzkorpus zurückgeführt werden kann.

Kompositions-Erstglied (TRUNC)

Als Kompositions-Erstglieder bezeichnet STTS Wortteile, die mit einem Bindestrich enden, der einen Teil des nachfolgenden, mit *und* oder *oder* verknüpften Wortes ersetzt.

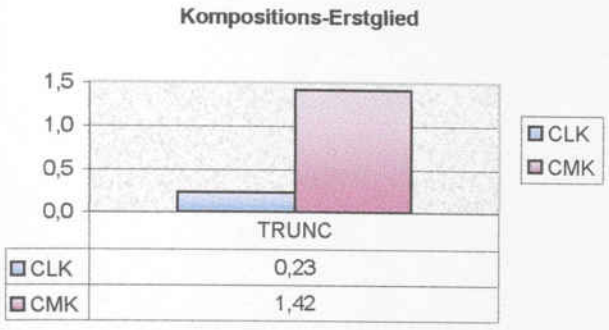


TABELLE 98

Dieses Element ist trotz der relativ großen Frequenz im CMK deutlich unterrepräsentiert im CLK, mit einer negativen Abweichung von -83,80%.

5 Interpretation des empirischen Materials

Im vorhergehenden Kapitel konnte mittels der Darstellung von Frequenzen sprachlicher Phänomene belegt werden, daß verschiedene Sprachstandebenen zu abweichenden Realisierungen sprachlicher Produktion führen. Die Abweichungen, die dabei beobachtet wurden, sind in hohem Maße regelmäßig, worauf bei der Darstellung der einzelnen Phänomene in Kapitel 4 hingewiesen wurde.

Die empirische Darstellung der einzelnen sprachlichen Phänomene, wie die Frequenzen der Wortarten von STTS, wirft jedoch vier Fragen auf:

- Inwieweit einzelne Phänomene miteinander verbunden sind, d.h., welche Wechselbeziehungen zwischen ihnen zu erkennen sind und in welchen Gruppen Phänomene zusammengefaßt werden können.
- Welche Aspekte zu berücksichtigen sind, um weitere Ergebnisse mit dem vorgeschlagenen Modell oder einer erweiterten Version zu erzielen.
- Welchen Einschränkungen ein korpuslinguistisch ausgerichtetes Modell zur Untersuchung von Textprodukten der

Lernersprache unterliegt, d.h., was damit nicht erreicht werden kann.

- Für welche Forschungsbereiche das vorgeschlagene methodische Modell Anwendung finden und somit zu einer tieferen Kenntnis des untersuchten Objektes, der Lernersprache, führen kann.

Das folgende Kapitel ist der Behandlung dieser vier Punkte gewidmet, die dank der durchgeführten Untersuchung zu klären sind und die sich daraus ergeben. Jeder dieser Punkte führt dabei zu Ansätzen, die jedoch keinen Anspruch auf Vollständigkeit erheben können, sondern vielmehr als Vorschläge für weitere ergänzende Untersuchungen in dieser Richtung verstanden werden sollten. Denn sowohl die Wechselbeziehungen der sprachlichen Phänomene als auch Ansätze zur Verbesserung, die Einschränkungen und die Anwendungsbereiche der Methode bilden offene Gruppen, die nur in Abhängigkeit vom jeweiligen Forschungsziel näher zu bestimmen sind.

5.1 Wechselbeziehungen

Die frequentielle Darstellung einzelner sprachlicher Phänomene bestätigt sich, wie der vorhergehenden Untersuchung zu entnehmen ist, als ein nützliches Werkzeug, um im Bereich der Lernersprache Abweichungen aufzudecken (Leech 1998), die eventuell als „Schwachpunkte" gedeutet werden können. Doch aus einer umfangreicheren Perspektive ermöglicht erst die

Bestimmung von Gruppen eine präzisere Erklärung der möglicherweise mit der Abweichung verbundenen Aspekte.

Gruppen von sprachlichen Phänomenen, für deren Abweichung eine einheitliche Erklärung gefunden werden kann, sind dementsprechend unumgänglich für jeden weiteren Erklärungsversuch. Die Methode zur Bestimmung der sprachlichen Phänomene, die einer gleichen Gruppe angehören, soll im folgenden dargestellt werden.

Aufgrund der Menge der sprachlichen Phänomene, die sich potentiell gegenseitig bedingen können, steht für die Bestimmung der Phänomengruppen, in denen sich gegenseitig beeinflussende Phänomene zusammengefaßt werden, kein automatisiertes Verfahren zur Verfügung. In diesem Sinne ist die Bestimmung der Wechselbeziehungen sprachlicher Phänomene mit dem Prozeß der Profilierung zu vergleichen (vgl. S. 45; S. 173; Crystal 1991: 227), in dem aufgrund subjektiver Basis zusammengestellte Phänomene erst in einer zweiten Phase anhand der erhobenen Daten auf ihre Relevanz überprüft werden sollten. Dementsprechend werden die Gruppen sich beeinflussender Phänomene zunächst subjektiv bestimmt, und erst nach der provisorischen Festlegung der miteinander verbundenen sprachlichen Phänomene können statistische Verfahren zur Überprüfung der Intensität der Wechselbeziehung Anwendung gefunden werden.

Aus diesem Grund werden in diesem Abschnitt die erhobenen sprachlichen Phänomene der quantitativen Auswertung hinsichtlich ihrer Relevanz und ihrer Wechselbeziehungen dargestellt werden, in bezug auf die Besonderheiten, die die Lernersprache der betreffenden Population auszeichnen. Es handelt sich dabei um eine erste Phase in der Bestimmung der Phänomengruppen, für die provisorisch Phänomene gruppiert und anhand erster quantitativer Daten miteinander in Verbindung gebracht wurden. Spezifische statistische Verfahren zur Untersuchung des Grades des Wechselbeziehungen wurden hier nicht angewendet, da sie in hohem Maße zielgerichtet sind, d.h. von dem angegebenen Forschungsziel beeinflußt werden.

Zuvor scheint es uns aber notwendig, die Vorgehensweise zu beschreiben, die für die Bestimmung relevanter sprachlicher Phänomene angewendet wurde zusammen mit den Einschränkungen, die aus der besonderen Beschaffenheit der benutzten Korpora entstehen.

Relevante Phänomene konnten in drei verschiedenen Bereichen gesucht werden: in der Wortbildung aus traditionell morphologischer Sicht, in der Untersuchung syntaktischer Wörter im Sinne der neuesten Lexikologie und in den Satzstrukturen (Satzsyntax).

Den drei Bereichen wurden (direkt oder indirekt) sprachliche Phänomene zugeordnet, deren Daten im vorhergehenden Kapitel erhoben wurden.

Die in Kapitel 4 dargestellten Daten sind quantitativer Natur und beruhen, wie gesehen, auf der Einteilung einerseits in eine allgemeine statistische Beschreibungen und andererseits eine grammatische (im weitesten Sinne) Beschreibungen. Der allgemeine statistische Teil gab Information über Wort- und Satzlänge und Cluster, verstanden als sich statistisch überdurchschnittlich oft wiederholende Wortverbindungen. Anhand der Untersuchung der Wortlänge (siehe 5.1.1) konnten Phänomene zur Wortbildung, hauptsächlich der Zusammensetzung, festgestellt werden. Die Satzlänge gab in Verbindung mit der Untersuchung der Wortarten Hinweise auf die Verwendung von stilistisch oft frei hinzufügbaren Elementen, wie attributive Adjektive. Die Cluster ihrerseits führten zur Erkenntnis, daß formelhafte Ausdrücke, die der Lexik zuzurechnen sind, nicht den gleichen Stellenwert in beiden Korpora haben.

Die grammatische Beschreibung der einzelnen syntaktischen Wortarten ließ grundlegende Eigenschaften der Unter- und Überrepräsentation anhand der Tags des ImS erkennen. Einige dieser Wortarten vermitteln rein grammatische (vor allem morphologische) Information, wie der bestimmte und der unbestimmte Artikel; andere hingegen ermöglichen über ihre rein distributiven Qualitäten hinaus auch die Einbeziehung der lexikalisch-syntaktischen Komponente (Verwendung von Konjunktionen, als geschlossene Wortart), der Wortbildungs-Komponente (adjektivische Suffixe oder verbale Präfixe) oder

einen Hinweis auf die Verwendung syntaktischer Strukturen, wie Relativsätze oder Infinitivkonstruktionen.

Anzumerken ist dabei jedoch, daß die Wortartentags von STTS direkt nur Aussagen über die Verwendung der entsprechenden syntaktischen Wortarten zulassen (vgl. Anhang, S. 409). Weitere Information zu morphologischen oder syntaktischen Eigenschaften konnte nur indirekt gesammelt werden anhand der Ausprägung dieser Eigenschaften im lemmatisierten Eintrag zu jedem Wort. So zeichnet STTS zwar nicht die ungetrennt erscheinenden verbalen Präfixe gesondert aus (dies tut es nur, wenn die Präfixe getrennt erscheinen), doch stellt die manuelle Untersuchung der Verblisten auf erscheinende Präfixe eine Lösung zu diesem Problem dar.

In dieser Hinsicht gründen die folgenden Aussagen über relevante sprachliche Phänomene und über ihre Wechselbeziehungen hauptsächlich auf den durch STTS bestimmten Wortarten und zusätzlich auf einigen lexikalischen, morphologischen und syntaktischen Eigenschaften, die indirekt untersucht werden konnten.

Dabei konnten jene Phänomene nicht mit einbezogen werden, die nicht anhand der Tags oder ihrer Verbindung mit orthographischen Ausprägungen untersucht werden konnten.

Diese Einschränkung bezieht sich beispielsweise auf die Konnektoren. Zu den Konnektoren können grundsätzlich alle Konjunktionen von STTS gerechnet werden, einige Adverbien

(Konjunktionaladverbien) und zusätzlich konjunktionsartig verwendete Elemente, bei denen es sich um Verben und andere Wortarten, aber auch komplexe lexikalische Einheiten, handeln kann. Während die Konjunktionen als solche ausgezeichnet sind, müssen die Konjunktionaladverbien von den einfachen Adverbien ausgesondert werden, was anhand von lexikalischen Einträgen mit geringen Arbeitsaufwand erfolgen kann. Diese Vorgehensweise ist aber im Fall der anderen Wortarten und komplexen lexikalischen Einheiten nicht mehr möglich. Sie benötigen einer gesonderten, textuellen Auszeichnung, die die Leistung von STTS überfordert und untersuchungsspezifische Parameter berücksichtigen müßte, wie z.B. eine Einteilung der Konnektoren nach semantischen oder textstrukturellen Kriterien, die Information zu ihrer Wortartenzugehörigkeit und ihrem anaphorischen oder kataphorischen Beziehungen, etc.

Aus diesen Gründen werden hier die relevanten Phänomene von Kapitel 4 hervorgehoben und es wird der Versuch gemacht, die Abhängigkeit und die distributiven Regelmäßigkeiten dieser Abhängigkeit darzustellen. Damit wird explizit auf den Versuch einer Erklärung der Ursachen verzichtet, der zum einen nicht das Ziel dieser Arbeit war, und zum anderen von den betroffenen Disziplinen geleistet werden muß, wie die Didaktik, die Textlinguistik oder die interkulturelle Sprachforschung.

Im folgenden werden die Phänomene in die Gruppen Wortbildung, Lexik und Syntax gegliedert. Alle relevanten Phänomene des

vorhergehenden Kapitels können mindestens einer dieser Gruppen zugeordnet werden, doch ist eine eindeutige Zuweisung wegen der vielfältigen Wechselbeziehung nicht durchführbar. Zur Klärung besonderer Aspekte, wie der Wortzusammensetzung, werden weitere problemspezifische Daten angegeben, die die des Kapitels 4 ergänzen.

5.1.1 Tendenzen in der Wortbildung

Im CLK erscheinen einige relevante morphologische Phänomene, die als der Wortbildung zugehörig angesehen werden können. Innerhalb der Wortbildung werden hier nur Phänomene dargestellt, die den Wortbildungsmechanismen Komposition und Derivation entsprechen. Ausgelassen werden die Mechanismen der Konversion und periphere Prozesse der Wortbildung wie Kürzungen, Abkürzungen oder Kontaminationen (Bußmann 1990: 852f). Flexionsformen werden hier nicht zu den Wortbildungsmechanismen gerechnet und in der Gruppe Syntax dargestellt.

Die erwähnten Wortbildungsmechanismen können im vorhergehenden Kapitel verschiedenen quantitativen Darstellungen zugewiesen werden. Die Prozesse der Derivation (Affigierung) werden in Abschnitt 4.2.1.5 (vgl. S. 242) anhand substantivischer und verbaler Präfixe und substantivischer und adjektivischer Suffixe untersucht. Die Prozesse der Komposition sind indirekt aus der Untersuchung zur Wortlänge (vgl. 4.2.1.1, S. 219)

abzuleiten und werden hier ausführlich mit zusätzlichen problemspezifischen Untersuchungen dargestellt.

Als allgemeine Tendenz hinsichtlich der Wortbildung ist anzumerken, daß die untersuchten Wortbildungsmechanismen des CLK durch allgemein niedrigere Frequenzen und geringere Variationswerte ausgezeichnet sind. Diese Tendenz ist auf allen Ebenen anwendbar. Die Mittelwerte beider Werte respektieren diese Tendenz auf vertikaler Ebene (beim Vergleich zwischen CLK und CMK) und auch innerhalb jeder Gruppe (präfigierte Substantive, präfigierte Verben, suffigierte Substantive, suffigierte Adjektive). Ausnahmen bilden nur einige wenige Elemente der Gruppen, die oft nicht konsistent sind im Sinne, daß eventuell ein Element höhere Werte hinsichtlich der Frequenz aufweist, aber nicht in der Variation. Für die mit der Wortlänge verbundenen Wortbildungsmechanismen werden keine Variationswerte angegeben, doch liegen die Frequenzen des CMK auch in diesem Fall über denen des CLK.

Die quantitative Darstellung des Phänomens Wortlänge hatte ergeben, daß ab ugf. 8 Zeichen im CMK eine höhere Proportion an längeren Wörtern erscheint. Zur Untersuchung der Faktoren, die dieses Phänomen beeinflussen, wurde eine Liste der Wörter ab 16 Zeichen zusammengestellt. Dieser Wert wurde gewählt, weil die Trefferzahl insgesamt 1074 Wörter ausmacht; bei einer reduzierteren Wortlänge (15 oder 14 Zeichen) wäre die Liste extrem schwierig zu bearbeiten gewesen.

Aus den Listen ist zu entnehmen, daß es sich bei den Wörtern ab 16 Zeichen hauptsächlich um Wortzusammensetzungen handelt, und in wenigen Fällen um Flexionsformen oder Ableitungen. Im CLK erscheinen insgesamt 7,72‰ lange Wörter, während dieser Wert im CMK 17,97‰ beträgt, mehr als das zweifache. Ebenfalls ist eine Entwicklung innerhalb der Subkorpora zu erkennen, die von niedrigen Werten bei niedrigem Sprachniveau zu höheren Werten geht. So werden im Subkorpus MS 6,22‰ lange Wörter verwendet, im Subkorpus OS 8,26‰ und im CMK 17,97‰. Die Entwicklung zwischen MS und OS ist jedoch nicht sehr stark ausgeprägt, was mit einer allgemeinen Tendenz zur Vermeidung der Wortbildung erklärt werden kann (vgl. Tabelle 99).

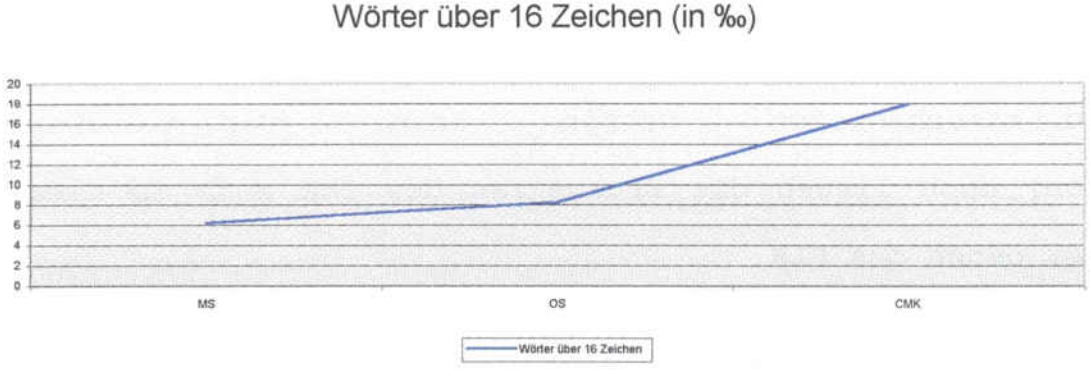


TABELLE 99

Diese Tendenz spiegelt sich auch in den verwendeten Wortarten wider. Durch Wortzusammensetzung werden an erster Stelle Substantive gebildet und danach mit großen Abstand Adjektive und Verben. Im Subkorpus MS führt die Zusammensetzung fast

ausschließlich zu Substantiven, im Subkorpus OS hingegen auch zu einigen Verben. Hinsichtlich der beteiligten Wortarten aber erscheinen sowohl MS als auch im OS fast ausschließlich Substantive und Adjektive und nur im OS zwei Verben. Im CMK hingegen sind auch Partizipien, Adverbien, Präpositionen und Verben an der Zusammensetzung beteiligt, obwohl auch hier Substantive den wichtigsten Platz einnehmen. Die Variation der angewendeten Mechanismen ist damit im CMK ebenfalls höher als im CLK.

Ein weiteres Indiz für die Tendenz zu einer reduzierten Anwendung der Wortzusammensetzung ist das extremere Phänomen der Wortzusammensetzung mit drei oder mehr Elementen (z.B. *Gastvortragsreihe*). Im CMK erscheinen 1,10‰ dieser Wörter, im Gegensatz zum Subkorpus OS mit 0,34‰ und MS mit keinem einzigen dieser Wörter.

Die Komplexität der Zusammensetzungen und der dafür zu beachtenden Regeln (verstanden als grammatische, lexikalische und pragmatische Kenntnisse) führt ebenfalls zur Unterrepräsentation des Phänomens. Das Fugenelement *s* ist ein Beispiel dafür, denn es wird im CLK prozentuell weniger als im CMK verwendet. Während im CMK 36% aller Wörter über 16 Zeichen ein Fugenelement *s* enthalten, sind es 26% im Subkorpus OS und nur noch 18% im MS[70].

[70] Wobei noch die evtl. von der Lehrkraft eingefügten Korrekturen zu berücksichtigen sind.

Die Wortzusammensetzung, erkennbar an der Wortlänge, ist somit ein klares Indiz für Lernersprache. Eine hohe Proportion an kurzen Wörtern impliziert eine Unterrepräsentation an Wortzusammensetzungen, was sich in einer allgemein niedrigeren Frequenz von Zusammensetzungen ausdrückt, und in Zusammensetzungen mit unterschiedlichen Wortarten und Fugenelementen noch stärker ausgeprägt ist.

Die zurückhaltende Anwendung von Wortbildungsmechanismen im CLK spiegelt sich erneut in der Affigierung wider. Mit den Suffixen und den Präfixen werden hinsichtlich der Wortlänge eine Mehrzahl an Wörtern betrachtet, die mindestens 8 Zeichen aufweisen. Diese Wortlänge bedeutete den Anfang der Inversion der Tendenz zu einer höheren Wortlänge im CMK. Mit im allgemeinen zwischen 2 und 4 Zeichen Länge benötigen Präfixe und Suffixe Wörter zwischen 4 und 6 Zeichen, um zu diesen 8-Zeichen-Wörtern gerechnet zu werden, was im allgemeinen der Fall ist, so daß allein der höhere Wert an Wörtern ab 8 Zeichen im CMK Hinweis auf eine stärke Ausprägung des Phänomens Affigierung war.

Wird das relativ wenig repräsentative substantivische Präfix *Um-* beiseite gelassen, ist der quantitativen Darstellung zu entnehmen, daß die relativen Werte der Präfixgruppe Verben (VP) und der Suffixgruppen Substantive (SS), Adjektive (AS) und Verben (VS) im CMK über denen des CLK liegen. Des weiteren ist bei diesen Werten zu beachten, daß auch die Werte der Variation höher sind im CMK. Die allgemeine Tendenz in jeder

Gruppe ist also, daß höhere relative Werte mit einer größeren Variation verbunden sind.

Ausnahmen dazu bildeten die in Tabelle 100 zusammengefaßten Einzelelemente, was aber keinen Einfluß auf die Tendenz der Gruppen als Ganzes nimmt. Die Mehrzahl dieser Ausnahmen lassen sich auf zwei Faktoren zurückführen, die eine Erklärung dafür liefern können: die niedrigen Frequenzen und die Trennbarkeit des Elements im Falle der verbalen Präfixe.

Präfix- oder Suffixklasse	Präfix oder Suffix	Ausnahme relativer Wert	Ausnahme Variationswert	Frequent
SS	*-heit*	Ja	Nein	Nein
SS	*-ismus*	Nein	Ja	Nein
AS	*-los*	Ja	Nein	Nein
AS	*-sam*	Ja	Ja	Nein
VP	*be-*	Ja*	Nein	Ja
VP	*durch-*	Ja	Nein	Nein
VP	*ge-*	Ja	Nein	Nein
VP	*ver-*	Ja*	Nein	Ja

* Abweichung geringfügig

TABELLE 100

Die niedrigen Frequenzen einiger Elemente, wie die adjektivischen Suffixe *-los* und *-sam*, die substantivischen Suffixe *-heit* und *-ismus* und das verbale Präfix *durch-* können zu Verzerrungen der Ergebnisse hinsichtlich der Grundgesamtheit führen und eine Einschränkung der Repräsentativität der Stichprobe darstellen. Niedrige Frequenzen, die fast im Zufallsbereich liegen, führen bei nur wenigen Treffern zu großen prozentuellen Abweichungen und gleichen sich statistisch nur in sehr großen Korpora aus.

Dieses Problem wurde minimiert, indem ein Mindestwert von 0,20‰ für jedes Element in einem der beiden Korpora gefordert wurde (ugf. 9 Einträge im CLK oder im CMK), doch selbst so ist im Grenzbereich dieses Wertes immer noch eine mögliche Verzerrung zu beachten.

Diese Verzerrung wird klar erkennbar, wenn ein Element im Grenzbereich nur einen höheren relativen Wert aufweist, aber keine größere Variation, oder umgekehrt, also der allgemeinen Tendenz zu höheren Werten auf beiden Ebenen widerspricht. Beispiel für Elemente, die nur in einer Ebene höhere Werte erzielen und zudem wenig frequent sind, wären *-heit, -ismus, -los,* und *durch-* und *ge-*. Doch erklärt dies nicht die Abweichung des AS -sam und der VS *be-* und *ver-*, die entweder eine Ausnahme sowohl beim relativen Wert als auch bei der Variation darstellen (*-sam*) oder als frequent eingestuft werden können (*be-* und *ver-*).

Die Erklärung der Ausnahmen, die die verbalen Suffixe betreffen, liegt in der Trennbarkeit der Elemente. Im Falle von *be-* und *ver-* handelt es sich um sehr frequente Elemente mit einer geringen Abweichung in der relativen Frequenz und nur leicht höheren Werten in der Variation zugunsten des CMK. Die Werte beider Elemente sind somit sehr ähnlich. Die Einteilung der VP nach Trennbarkeit zeigt, daß all die untrennbaren Elemente zu vergleichbaren Werten im CMK und im CLK tendieren, während die trennbaren starke Abweichungen aufweisen. So erreichen die untrennbaren VP *be-, ent-* und *ver-*

niedrige Abweichungen (relativer Wert und Variation) von ugf. 20%, im Falle von *ge-* liegt der relative Wert sogar um mehr als 30% höher im CLK, während die Variation mit mehr als 40% viel größer im CMK ist, was als Ausnahme aufzufassen ist.

Die trennbaren Elemente jedoch sind stärker im CMK vertreten; *an-, auf-, aus-, ein-, heraus-, hervor-, hin-, mit-, nach-, um-, unter-, über-, vor-, weiter-* und *zu-* weisen höhere relative Werte auf und haben eine größere Variation; die einzige Ausnahme bildet *durch-*, was allerdings, da es sich um ein sehr niedrigfrequentes Element handelt, aufgrund seiner wenig repräsentativen Frequenz.

Sowohl die Wortzusammensetzung als auch die Affigierung beeinflussen die Wortlänge, vor allem die Proportion an Wörtern ab 8 Zeichen. Ein höhere Wert der Wortlänge kann demnach als Indiz dafür angesehen werden, daß die beiden hier dargestellten sprachlichen Phänomene weniger oft angewendet werden, was wiederum Rückschlüsse auf die Phase des Lernstadiums ermöglicht.

5.1.2 Tendenzen auf Wortebene (syntaktische Wortarten)

Die Phänomene, die der Ebene des syntaktischen Wortes zugeschrieben werden können, haben Einfluß auf die Frequenz der Wortarten von STTS und ihre Variation. Die Ausprägung der damit verbundenen Phänomene folgt drei Tendenzen: Erstens die

Überrepräsentation einer Wortart verbunden mit der Unterrepräsentation einer anderen, die semantisch vergleichbare Elemente enthält, wie im Falle der Konjunktionen, Pronominal- und Konjunktionaladverbien und Adverbien; zweitens die Unterrepräsentation von Wortarten, die in einem hohen Grad frei hinzufügbar sind, z.B. Adjektive; und drittens, die Unterrepräsentation von komplexen lexikalischen Einheiten, wie formelhafte Ausdrücke und valenzbedingte Präpositionen. Diese drei Tendenzen werden in diesem Abschnitt einzeln besprochen; dem folgt als Schluß die Darstellung der Type-Token-Ratio, Indikator für die lexikalische Variation, der stark von dem Umfang der Texte beeinflußt wird.

Überrepräsentation und Unterrepräsentation

Die erste Tendenz impliziert die Überrepräsentation einer Wortart, für die eine entsprechende Unterrepräsentation bei einer anderen Wortart gefunden werden kann, die als semantischer Ersatz für die erste gelten kann. Beispielhaft ist in diesem Sinne der Fall der Konjunktionen und der Pronominaladverbien[71], die sich in ihren absoluten Frequenzen zwar nicht ausgleichen, jedoch eine Annäherung darstellen.

[71] Schiller et al. (1995) unterscheiden in der Gruppe der Pronominaladverbien nicht die Konjunktionaladverbien, was eine Untersuchung des Phänomens erschwert.

Konjunktionen weisen eine Überrepräsentation im CLK auf, die kumulativ ugf. 18% beträgt und im Falle der subordinierenden Konjunktionen KOUS sogar 42%.

Die einzige Ausnahme in der Überrepräsentation des CLK bilden die wenig frequenten Vergleichspartikel KOKOM (ausschließlich *als* und *wie*), mit 8,64‰ im CMK und 5,94‰ im CLK. Sie erreichen eine kumulative Unterrepräsentation von 31% im CLK; die Unterrepräsentation ergibt sich jedoch innerhalb der Gruppe aus nur einem der beiden Elemente, *als*, da *wie* vergleichbare Werte ergibt; *als* steht in diesem Sinne für eine individuelle Unterrepräsentation von 42%.

Alle anderen Konjunktionalgruppen, also nebenordnende Konjunktionen (KON), unterordnende Konjunktionen mit Satz (KOUS) und unterordnende Konjunktionen mit Infinitiv (KOUI) sind im CLK in verschiedenen Graden überrepräsentiert.

Die nebenordnenden Konjunktionen KON weisen eine kumulative Abweichung von etwas mehr als 14% auf, wobei die individuelle Verteilung (vgl. Tabelle 53) jedoch zeigt, daß die Abweichung hauptsächlich auf die Überrepräsentation von zwei Konjunktionen zurückzuführen ist: *und* und *aber*. Dabei beträgt die Abweichung von *und* 6,65%, bei einem Frequenzwert von 27,34‰ (ausgehend vom CMK), und die von *aber* 381%, bei einem Frequenzwert von 0,98‰. Die Betrachtung des unteren Frequenzbereiches (unter 5‰) zeigt, daß die Konjunktionen dort zu einer Unterrepräsentation im CLK tendieren (vgl. Tabelle

54); diese Tendenz wird deutlicher im Bereich unter 1‰: *denn, doch, entweder, sondern, sowie* und *sowohl* weisen höhere Werte im CMK auf, im Falle von *doch* ist die Konjunktion sogar nur im CMK präsent. In diesem Bereich unter 1‰ stellen *denn* und *weder* die einzigen Ausnahmen dar. Zusammenfassend kann festgehalten werden, daß das CLK zur Überrepräsentation weniger hochfrequenter Konjunktionen neigt, besonders von *aber*, und andere, seltener verwendete Konjunktionen leicht unterrepräsentiert sind.

Auffallend ist hier die Überrepräsentation im CLK der Konjunktion *aber*, die 3,81 mal öfter im CLK erscheint. Im CMK ist ihre Verwendungsfrequenz mit der von *denn, sondern* oder *sowie* vergleichbar, die Verteilung im CLK hingegen mit keiner anderen Konjunktion. Dies erklärt sich zumindest teilweise aus der Untersuchung der entsprechenden Einträge in der Konkordanzliste und der Vergleich der Konjunktion *aber* mit dem Adverb *aber*, der im Anschluß an die Beschreibung der weiteren Konjunktionalgruppen dargestellt wird.

Bei den unterordnenden Konjunktionen mit Satz wiederholen sich erneut die distributiven Eigenschaften der nebenordnenden Konjunktionen. Die unterordnenden Konjunktionen des hohen Frequenzbereiches, über 2‰ in einem der beiden Korpora (*daß, weil, wenn* und *wie*), sind stark überrepräsentiert; im mittleren Frequenzbereich liegende Konjunktionen (zwischen 0,4‰ und 2‰: *als, da, obwohl, ob* und *während*) sind ebenfalls deutlich überrepräsentiert, während im unteren Bereich (unter

0,4‰) diese allgemeine Tendenz unterbrochen wird und kein deutliches Schema mehr erkennbar ist, wenn auch festgehalten werden kann, daß eine größere Variation im CMK erscheint. Einige sind im CLK stark überrepräsentiert, wie *bis* und *damit*, andere unterrepräsentiert, wie *indem* und *nachdem*, die meisten jedoch weisen vergleichbare Werte auf, so zum Beispiel *bevor*, *ehe* und *sobald*.

In den unterordnenden Konjunktionen mit Satz KOUS tritt klarer als bei den nebenordnenden KON die größere Variation hervor, die bezeichnend für sie ist. In den Konkordanzlisten erscheinen insgesamt 26 verschiedene KOUS, von denen 19 im CLK vertreten sind und 24 im CMK. Im CLK erscheinen nicht die KOUS *insofern, obschon, seit, sofern, soweit, wenngleich* und *zumal*; dafür enthält das CLK *seitdem* und *solange*, die nicht im CMK erscheinen. Dies Phänomen kann auf die niedrigen Frequenzen dieser Konjunktionen zurückgeführt werden, denn alle erscheinen nur aufgrund eines einzigen Treffers im ganzen Korpus in den Listen (vgl. Tabelle 51), so daß die Bedeutung der Variation hier in der höheren Anzahl an Konjunktionen innerhalb der Gruppe liegt. Dies wird besonders deutlich durch einen semantischen Vergleich der Konjunktionen. Helbig/Buscha erwähnen *obschon* und *wenngleich* als „seltene, der gehobenen Stilschicht angehörige Konjunktionen", die „vollständig durch andere Konjunktionen vertreten werden" (Helbig/Buscha 1991: 445), in diesem Fall von *obwohl*. Während *obwohl* im CLK überrepräsentiert ist, erscheinen im CMK zusätzlich

synonymisch aufzufassende Konjunktionen für *obwohl*, wie *obschon* und *wenngleich*.

Für die unterordnenden Konjunktionen mit Satz sei so festzuhalten, daß sie kumulativ im CLK überrepräsentiert sind, vor allem im Bereich der frequenten Konjunktionen, daß aber höhere Variationswerte im CMK zu verzeichnen sind.

Die unterordnenden Konjunktionen mit Infinitiv bilden keine Ausnahme innerhalb der Konjunktionen und weisen höhere kumulative Frequenzwerte im CLK auf. Dabei ist die größte Überrepräsentation bei *um zu* zu verzeichnen, das im CLK bei einer Frequenz von 1,39‰, ausgehend vom CMK, 2,91 mal so oft wie im CMK verwendet wird; ihm folgt bei einer sehr niedrigen Frequenz von 0,12‰ im CMK *ohne zu* mit einer 2,08 mal größeren Verwendung im CLK. Die weiteren Konjunktionen dieser Gruppe sind eher selten (0,05‰ im CMK) und werden hier nicht weiter betrachtet.

In allen drei Konjunktionalgruppen ist die Tendenz erkennbar, daß einige sehr frequente Konjunktionen sehr stark überrepräsentiert sind im CLK, daß dieser aber in den niedrigeren Frequenzbereichen weniger Variation aufzuweisen hat. Da Konjunktionen nun abgesehen von der syntaktischen Verknüpfung zwischen Elementen auch die Funktion haben können, semantische Beziehungen zwischen diesen Elementen herzustellen (Hentschel/Weydt 1990: 257), stellt sich die Frage, anhand welcher sprachlicher Mittel diese Beziehungen, sofern sie

nicht abwesend oder seltener sind, im CMK realisiert werden, wenn dieses Korpus weniger Konjunktionen aufzuweisen hat.

Dabei ist es üblich, die Zahl der Konjunktionen in relativen Werten auf Gesamtwortbasis anzugeben, also auf der Gesamtlänge der Texte des Korpus (so z.B. auch für einen Lernerkorpus in Altenberg/Tapper 1998), was zu Verzerrungen führen kann, wenn große Abweichungen hinsichtlich von fakultativen Elementen erkennbar sind, wie z.B. attributive Adjektive. Wünschenswert wäre hier eine semantische Auszeichnung, die die Ideen, die Beziehungen zwischen ihnen und die sie verbindenden Konnektoren angäbe, was jedoch einen augenblicklich noch nicht automatisierbaren Prozeß darstellt. Nichtsdestotrotz geben die Werte auf Gesamtwortbasis Hinweise auf Tendenzen, vor allem, wenn mehrere Elemente in einer Gruppe miteinander verglichen werden, und die Gruppe nicht nur aus einem Element besteht.

Einer dieser Hinweise ist, daß im CMK die semantischen Beziehungen anhand anderer Wortarten realisiert werden. Konjunktionaladverbien oder Pronominaladverbien können teilweise Konjunktionen ersetzen (Helbig/Buscha 1991: 445) und konkurrieren mit ihnen auf semantischer Ebene, wobei sie distributionell von den Konjunktionen abgegrenzt werden. Eine weitere konkurrierende Wortart wären die Verben, die jedoch aufgrund der starken Variation als offene Wortarten hier nicht betrachtet werden können.

Konjunktionaladverbien und Pronominaladverbien bilden in der Literatur keine einheitliche Klasse. Helbig bestimmt Konjunktionaladverbien distributiv: sie können die Stelle vor dem finiten Verb allein einnehmen und auch innerhalb des Satzes stehen, gleichzeitig aber vielfach auch am Satzanfang die Rolle einer koordinierenden Konjunktion übernehmen (Helbig Buscha 1990: 341). Pronominaladverbien hingegen werden dort morphologisch nach ihrer Wortbildung bestimmt: es handele sich um Wortverbindungen zwischen bestimmten Präpositionen mit den Adverbien *da-* und *wo-* (Helbig/Buscha 1990: 264)[72]. Hentschel/Weydt definieren Konjunktionaladverbien als Synonym für Pronominaladverbien, und beide also als Wörter, die wie Konjunktionen Sätze miteinander verknüpfen, sich syntaktisch aber wie Adverbien verhalten (Hentschel/Weydt 1990: 276).

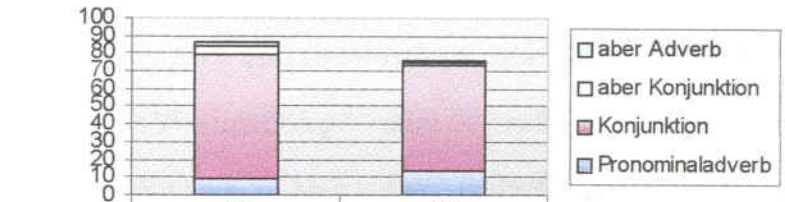
STTS faßt unter Pronominaladverbien (PAV) Pronominaladverbien und die Konjunktionaladverbien zusammen. Dazu muß aber noch *aber* in adverbialer Stellung gerechnet werden, das von Helbig nicht erwähnt wird und von Hentschel/Weydt (1990: 277f) sowohl den Konjunktionen als auch den Konjunktionaladverbien angehören kann.

Bis auf wenige Ausnahmen weisen die Pronominaladverbien (im Sinne von STTS) eine starke Unterrepräsentation im CLK auf,

[72] Hier darf man nicht verkennen, daß Pronomen keine selbständige Wortart bei Helbig/Buscha bilden, sondern anderen verteilt zugeschrieben werden und daß innerhalb der Wortart unterschiedliche Kriterien zur Differenzierung benutzt werden.

allerdings mit Frequenzen im CMK, die nur in einem Fall 1‰ überschreiten. So ist vorwegzunehmen, daß die Pronominaladverbien nicht für sich alleine eine Erklärung zu der starken Überrepräsentation von Konjunktionen im CLK sein können, da die Differenz zwischen Konjunktionen bei ugf. 10‰ zugunsten des CLK liegt, während die Differenz der Pronominaladverbien ugf. 5‰ beträgt; es ist also deutlich eine Annäherung der Werte zu erkennen (vgl. Tabelle 101).

Pronominaladverbien: Verteilung einzelner Elemente



	CLK	CMK
aber Adverb	2,19	1,68
aber Konjunktion	4,7	0,98
Konjunktion	70,12	58,44
Pronominaladverb	9,07	14,04

TABELLE 101

Das von STTS als Adverb getaggte *aber* kennzeichnet sich ebenfalls durch eine unterschiedliche Verteilung im CLK und CMK aus, die der Tendenz der Angleichung teilweise widerspricht. Während die Lerner *aber* vorzugsweise als Konjunktion verwenden (Erststellung), mit 4,70%, und seltener als Konjunktionaladverb, mit 2,19‰, verwenden es

Muttersprachler lieber in seiner adverbialen Funktion mit 1,68‰, als als Konjunktion, mit 0,98‰ (vgl. Tabelle 101).

Die semantische Funktion, die Konjunktionen haben können, scheinen also im CMK durch andere sprachliche Mittel realisiert zu werden als im CLK. Eines dieser sprachlichen Mittel sind die Konjunktionaladverbien, deren Werte aber allein nicht den Unterschied erklären können. Weitere Wortarten, die für eine Erklärung herangezogen werden könnten, wie Verben oder präpositionale Ausdrücke, werden im Rahmen dieser Arbeit nicht dargestellt, da für ihre Untersuchung eine semantische Auszeichnung notwendig wäre; dennoch sei auf den möglichen Einfluß der Präpositionen hingewiesen, die im CLK unterrepräsentiert sind, was aber zum großen Teil auf syntaktische Phänomene zurückzuführen ist, die weiter unten besprochen werden.

Zusammenfassend ist festzuhalten, daß das CLK zur Überrepräsentation gut bekannter Konjunktionen tendiert, und die restlichen sehr wenig variiert. Das CMK hingegen zeigt eine größere Variation auf, auch in stilistisch besonders markierten Konjunktionen (gehobener Sprache), und gleicht zumindest teilweise die Werte der Konjunktionen durch eine größere Anwendung semantisch kompatibler Wortarten aus, wie die Pronominaladverbien.

Unterrepräsentation fakultativer Elemente

Die zweite Tendenz auf lexikalischer Ebene ist an der Unterrepräsentation der frei hinzufügbaren Elemente erkennbar, d.h. der Wortarten oder Verbindung von Wortarten, die ohne Veränderung der Satzstruktur und mit normalerweise reduziertem Bedeutungsunterschied eingebunden oder weggelassen werden können. Elemente, die diese Eigenschaften erfüllen, sind nach Helbig/Buscha (1991: 621) die sekundären Satzglieder und die Attribute. Es handelt sich dabei nicht um Wortarten direkt, da auch ganze Satzglieder (oder Phrasen) betroffen sein können. Sie sind aber innerhalb der Auszeichnungen von STTS hauptsächlich anhand attributiver Adjektive (ADJA), Präpositionen und Genitivattribute (vgl. 4.2.2.1) erkennbar.

In STTS wird zwischen attributiven und prädikativen Adjektiven unterschieden, die die Tags ADJA und ADJD erhalten. Die kumulative Abweichung der Adjektive liegt im CLK bei -18,01%, wobei allerdings die relativ wenig frequenten prädikativen Adjektive eine Abweichung von -10,10% aufweisen und die viel frequenteren attributiven Adjektive einen Wert von -21,73% erreichen (vgl. Tabelle 38). Als fakultative Wortart werden die attributiven Adjektive also proportional weniger im CLK als die nicht immer fakultativen prädikativen Attribute verwendet, und dies in einem viel höheren Maß als im CMK. Einen gewissen Einfluß mag hier der morphologische Ursprung des Adjektivs haben. Die Darstellung der attributiven Partizipien (vgl. Tabelle 39) zeigt, daß im CLK die Adjektive

verbalen Ursprungs unterrepräsentiert sind, was der in 5.1.3 zu besprechenden Hypothese zuzuschreiben ist, daß im CLK Elemente desto stärker unterrepräsentiert sind, je mehr Regeln für ihre Anwendung kombiniert werden müssen. Hier würde es sich um die Verbindung zur Deklination des attributiven Adjektivs mit der Bildung eines Partizips handeln. Ein weiterer Faktor ist die Unterrepräsentation von verdoppelten Adjektiven (vgl. Tabelle 40), die im Falle des attributiven Adjektivs ADJA bei -28% und im Falle der prädikativen Adjektive bei -68% liegt.

Sowohl die Unterrepräsentation von ADJA als auch die von ADJD kann teilweise durch die Ersetzbarkeit dieser Elemente durch andere sprachliche Mittel erklärt werden, wie z.B. die Frequenzen von Relativsätzen, und findet sich in den niedrigeren Frequenzen der adjektivischen Suffixe bestätigt (vgl. Tabelle 32). Da es sich dabei um syntaktische Phänomene handelt, werden beide adjektivischen Formen in dieser Hinsicht erneut in 5.1.3 besprochen. Weil aber attributive Adjektive frei hinzufügbar sind und prädikative nicht unbedingt, kann vermutet werden, daß die attributiven ersatzlos gestrichen werden, was sich in zwei Indikatoren widerspiegelt, der Satzlänge, die im CLK reduzierter als im CMK ist, und der Type-Token-Ratio, die ebenfalls niedrigere Werte im CLK aufweist.

Satzlänge wurde anhand der Zahl der Wörter eines Satzes gemessen (vgl. 4.2.1.2). Die ersatzlose Streichung eines

Wortes bedeutet also automatisch die Reduzierung des erhaltenen Satzlängenwertes. Mit einer Frequenz von ugf. 7% (ugf. ein Adjektiv alle 14 Wörter) und einer Satzlänge von ugf. 20 Wörtern im CMK, kommen 1,42 attributive Adjektive auf jeden Satz. Im CLK beträgt dieser Wert bei einer Frequenz von 5,4% und einer Satzlänge von ugf. 16 Wörtern nur noch 0,86 attributive Adjektive pro Satz, was Einfluß auf die gesamte Satzlänge hat.

Ferner tragen attributive Adjektive ebenfalls zur lexikalischen Variation bei, der Type-Token-Ratio. Diese wurde aufgrund von 200-Wort-Segmenten berechnet; in jedem dieser Segmente erscheinen also im CMK 14 attributive Adjektive, im CLK nur noch 10,8, was eine Differenz von 3,2 ausmacht. Angenommen, die verwendeten attributiven Adjektive würden nicht mit den prädikativen übereinstimmen (täten sie dies, wären sie als identische Wörter berechnet worden, vgl. 4.2.1.4) und sich nicht wiederholen, so ständen sie bei einer Abweichung der Type-Token-Ratio von ugf. 5 (~60 im CMK, ~55 im CLK) für den größten Teil dieser Abweichung. Da sich aber Adjektive wiederholen und die Sättigung durch Funktionswörter in einem 200-Wort-Segment noch nicht abgeschlossen ist (vgl. Tuldava 1995: 131ff), darf diese Berechnung nur als Indiz angesehen werden und müßte anhand der individuellen Untersuchung aller Einträge kontrastiert werden.

Die Abweichung spiegelt sich aber nicht nur im Wert der standardisierten Type-Token-Ratio wider, sondern auch in der

Relation zwischen Types und Token in der Gesamtheit der Texte eines jeden Korpus. So weist das CLK 7199 Types bei 43786 Token auf, während das CMK 8389 Types bei nur 40963 Token enthält. Dies kann teilweise auch durch die Überrepräsentation von Personalpronomina im CLK erklärt werden (PPER, vgl. 4.2.2.7), die eine Differenz von -25.06‰ aufweist (ausgehend von einer Frequenz von 21.21‰ im CMK und 46.27‰ im CLK) und die gleichzeitige Unterrepräsentation von Substantiven (NN, vgl. 4.2.2.6), mit einer Differenz von 19,82 (NN: 213,89 im CLK, 233,71 im CMK). Demnach würde das CLK in einem semantischen Reduktionsprozeß häufiger Pronomina anstelle von Substantiven verwenden, was einen weiteren Teil der unterschiedlichen Type-Token-Ratio erklären kann.

Dabei ist jedoch darauf hinzuweisen, daß die Type-Token-Ratio nach Tuldava (1995: 132) einen Hinweis auf die lexikalische Variation eines Textes gibt, und daß dieser, je größer der Wert ist, desto mehr lexikalische Variation aufweist, sie aber gleichzeitig keine Möglichkeit zur Bewertung der ästhetischen Eigenschaften eines Stils gibt.

Unterrepräsentation komplexer lexikalischer Einheiten

Innerhalb der Gruppe rein lexikalischer Elemente ist die Bearbeitung der Type-Token-Ratio anhand von 200-Wort-Segmenten durch die fehlende Sättigung an Funktionswörtern eingeschränkt. Erst nach dieser Sättigung an Funktionswörtern kann präziser der aktive Wortschatz eines Textes gemessen

werden, doch sind die für diese Untersuchungen verwendeten Texte zu kurz, um eine solche Analyse zu ermöglichen.

Ein Indiz jedoch für die Anwendung des Wortschatzes sind die Cluster, in Form von Zwei- oder Drei-Wort-Einheiten, die sich statistisch überdurchschnittlich oft innerhalb der Texte wiederholen. In 4.2.1.3 sind Beispiele für solche rekurrenten Wortkombinationen dargestellt. Der Unterschied zwischen CLK und CMK liegt v.a. in der Verwendung dieser Elemente, die stärker in den Dreiwortclustern hervortreten.

Zu unterscheiden sind dabei einerseits die rekurrenten Kombinationen zwischen Substantiven oder Verben und den von ihnen regierten Präpositionen und die formelhaften Ausdrücke.

Im ersten Fall ist im CLK eine Unterrepräsentation dieser regierten Präpositionen zu erkennen, was sich später in der reduzierten Verwendung der Pronominaladverbien und der allgemeinen Frequenzen der Präpositionen widerspiegelt. Das CLK scheint Verben und Substantive zu vermeiden, die eine Präposition erfordern, und weist so eine Unterrepräsentation all der Tags auf, die mit diesen Präpositionen verbunden sind: die Präpositionen selbst, die Pronominaladverbien, worüber allerdings noch im Abschnitt zur Syntax zu sprechen sein wird, und die Artikel mit Präposition (APPRART, *zum*). Beispiele dafür sind *Einführung in* u.a.

Im zweiten Fall handelt es sich um formelhafte Ausdrücke, die als komplexe lexikalische Einheiten aufzufassen sind (vgl.

Wotjak 1992)[73] und mit relativ wenig Änderungen wiedergegeben werden, wie *in der Regel* oder *im Anschluß an*. Das CLK hebt sich durch eine starke Unterrepräsentation dieser Mehrworteinheiten ab, was ebenfalls einen gewissen Einfluß auf die Satzlänge haben kann, da die Verwendung dieser Einheiten, die aufgrund der fehlenden syntaktischen Auszeichnungen nur indirekt anhand der Cluster erkannt werden konnten und so im Rahmen dieser Untersuchung nicht quantifiziert wurden, mindestens 2 Wörter betreffen.

5.1.3 Tendenzen auf Satzebene (Satzstrukturen)

Zur Syntax werden hier Phänomene gerechnet, bei denen die Frequenzen einer Wortart auf die Über- oder Unterrepräsentation einer syntaktischen Struktur hinweist und eventuell mit charakteristischen Frequenzen einer anderen Wortart verbunden werden kann. Bei diesen syntaktischen Strukturen handelt es sich um Phänomene wie Deklination, regierte Wortarten und Satzstrukturen. Dabei ist bei fehlender syntaktischer Auszeichnung nur indirekt ein Rückschluß auf die zugrundeliegenden syntaktischen Strukturen erkennbar. Attributive Adjektive geben so Aufschluß über die Komplexität von Satzgliedern und sind als Attribute mit Relativsätzen verbindbar, die ihrerseits anhand des Relativpronomens

[73] Vgl. auch Schröder (1986: 252) hinsichtlich der Präpositionalphrasen, die wie Präpositionen verwendet werden.

erkennbar sind. Syntaktische Phänomene, die nicht auf der Oberflächenstruktur anhand eines POS-bestimmbaren sprachlichen Phänomens erscheinen, wie z.B. die Apposition, können so nicht mit einbezogen werden.

Die Tendenz dieser Gruppe besteht in der Unterrepräsentation von flektierten und regierten Elementen und ihre Ersetzung, wenn möglich, durch die Überrepräsentation semantisch entsprechender Strukturen.

Zu den Elementen, die Einfluß auf diese Phänomene haben, zählen adverbiale Interrogativ- und Relativpronomina, Cluster, die Verbindung von Präposition und Artikel, Relativpronomina, Personalpronomina und Adjektive.

Adverbiale Interrogativ- und Relativpronomina (PWAV) weisen ebenso wie die Konjunktionen eine Reduktion auf ein Grundelement auf (*wo*), das überrepräsentiert ist, während andere Elemente der Gruppe die Tendenz entweder zur Unterrepräsentation oder zur reduzierteren Variation aufweisen (vgl. Tabelle 77). Von den 21 Elementen der Gruppe (beide Korpora zusammengezählt) sind im CLK 14 dieser Elemente vertreten, im CMK 18. Dabei handelt es sich um Elemente mit sehr geringen Frequenzen (von den 21 Elementen liegen 18 in beiden Korpora unter einer Frequenz von 0,20‰), weshalb Einzelelemente nicht beweiskräftig sind. Der Unterschied zu den Konjunktionen liegt in der Rektion einiger Elemente, die von einem Substantiv oder Verb bestimmt wird. Es handelt sich

dabei um die Elemente *wodurch, wofür, wogegen, wohin, womit, woran, worauf, woraus, worin, worüber, worum* und *wovon*. Von diesen 12 Elementen erscheinen 11 im CMK und 7 im CLK, in dem sie, bis auf *womit, worüber* und *worum*, alle leicht unterrepräsentiert sind. Aufgrund der niedrigen Frequenzen sei hier jedoch nur auf die größere Variation im CMK hingewiesen, also auf die leicht reduziertere Anwendung der Elemente der Subgruppe der regierten Elemente der PWAV im CLK.

Das gleiche Phänomen der Unterrepräsentation von regierten Elementen findet sich in der Verteilung der Cluster bestätigt. Wie schon beschrieben, sind im CMK, vor allem bei den Dreiwortclustern, mehr Substantive mit Präpositionen zu verzeichnen (vgl. 4.2.1.3), wie *Einführung in, Überblick über* oder *Frage nach* als im CLK. Da Cluster nebeneinanderstehende Wörter aufdecken, können sie die präpositionale Rektion von Verben und Substantiven darstellen, nicht aber Satzrektionen, wie die Infinitivkonstruktion.

Infinitive mit *zu* werden im STTS als VVIZU getaggt (vgl. Tabelle 88); dabei ist dieses Tag allerdings kein Hinweis auf die unterschiedlichen syntaktischen Strukturen, die ein *zu* + Infinitiv verlangen, sondern gibt nur an, daß ein Infinitiv mit *zu* erscheint. Im CLK erscheinen 1,5‰ VVIZU, von denen ugf. ein Drittel der zweiteiligen Konjunktion *um zu* entsprechen; im CLK erscheinen ugf. 2,6‰ VVIZU, von denen nur 10% der Konjunktion entsprechen. In den restlichen Fällen handelt es sich um valenzbedingte Konstruktionen wie *was ihn befähigt,*

die anderen von dieser Kraft anzustecken oder attributive Konstruktionen wie *diese Kraft, das Göttliche durch ihr Herz wahrzunehmen.* Die unterschiedlichen Frequenzen weisen also auch hier auf eine reduzierte Verwendung von regierten Strukturen hin.

Die Verteilung der Verbindung von Präposition und Artikel, wie *zum* (vgl. APPRART in 4.2.2.4) kann auch mit dem Gebrauch von Vermeidungsstrategien in Verbindung gebracht werden. Das CLK weist eine Abweichung von -32,21% in der Wortart APPRART auf. Der Vergleich der zwei beteiligten Wortarten, Artikel und Präpositionen, zeigt, daß die Unterrepräsentation hier mit letzteren verbunden ist. Während die Frequenz des Artikels (bestimmter und unbestimmter zusammengefaßt) im CMK 118,38‰ und im CLK 119,74‰ beträgt, was im CLK eine Abweichung von 1,15% darstellt (vgl. Tabelle 46), liegt die Abweichung der Präpositionen (APPR, vgl. 4.2.2.3) bei -31,21%, ein mit der negativen Abweichung der syntaktischen Wortart APPRART (-32,21%) vergleichbarer Wert. Die Unterrepräsentation der Präpositionen im allgemeinen, zuzüglich der Präpositionen mit Artikel, ist Hinweis auf die Tendenz zur Vermeidung einerseits von Verben, Substantiven oder Adjektiven, die eine Präpositionalphrase fordern, und andererseits zur Vermeidung von frei hinzufügbaren Elementen, was sie mit den attributiven Adjektiven verbindet (vgl. 4.2.2.1).

Die Unterrepräsentation der Präpositionalgruppen und Adjektive kann sich teilweise durch eine Überrepräsentation anderer

Phänomene ausgleichen, die auf die Verwendung von größeren syntaktischen Einheiten hinweisen und diese Präpositionalgruppen semantisch ersetzen können. Als fakultative Attribute sind präpositionale Attribute und attributive Adjektive semantisch beispielsweise durch Relativsätze ersetzbar, und die Frequenzen der Relativpronomina könnten andeuten, daß hier ein Ersatz in den Relativsätzen zu finden sein könnte.

Mit einer Frequenz von 11,81% im CLK ist *d-* das häufigste Relativpronomen (vgl. Tabelle 63). Die Abweichung zum CMK beträgt dabei fast 40% (vgl. Tabelle 64). Im CLK erscheinen also 40% mehr Relativsätze als im CMK, wobei allerdings erneut auf die Unterrepräsentation von Relativsätzen mit stilistisch gehobenen und dementsprechend seltenen Relativpronomina hingewiesen werden muß, da *welch* mit -75,50% im CLK unterrepräsentiert ist.

Ob die semantische Funktion der Relativsätze, sowohl durch ein Relativpronomen als auch ein adverbiales Relativpronomen eingeleitet, in den Korpora anhand mehr attributiver Adjektive und präpositionaler Attribute realisiert wird oder durch andere sprachliche Mittel, ist anhand der POS-Auszeichnungen nicht mehr erkennbar. Folglich kann erst nach einer syntaktischen und semantischen Auszeichnung beantwortet werden, in welchem Grad semantisch-funktionell vergleichbare syntaktische Strukturen komplementär angewendet werden und bis zu welchem Punkt die potentiell beinhaltete Information

einfach nicht realisiert wird. Eine reduzierte Untersuchung der Einträge beider Korpora legt zumindest die Vermutung nahe, daß eine direkte Umformung der Relativsätze in semantisch entsprechende attributive adjektivische Strukturen nur in wenigen Fällen wahrscheinlich ist; aleatorisch ausgewählte Beispiele für Transformationen des Typs *„die Uni, die groß ist"* → *„die große Uni"* wären *„mit denen die Uni die neuen Studenten begrüßt"*, *„der mir helfen mußte"*, *„das normalerweise für Partys benutzt wird"* u.a. Hier stoßen die Wortartentags an ihre Grenzen, da sie keine Erklärung mehr für die Über- oder Unterrepräsentation von weiteren syntaktischen und semantischen Phänomenen und den Abhängigkeiten dieser Eigenschaften liefern können.

Hinsichtlich der syntaktischen Eigenschaften kann gesagt werden, daß das CLK zu einer Unterrepräsentation der Elemente neigt, die Flexionsmerkmale aufweisen oder von einem Valenzträger abhängig sind, was sich in Adjektiven, Präpositionen und Infinitivkonstruktionen widerspiegelt, während bei syntaktisch komplexeren, aber aus der Sicht der Valenz, der Rektion und der Deklination einfacheren Strukturen eine Überrepräsentation zu beobachten ist, wie die der Relativpronomina und Relativsätze.

5.1.4 Weitere Tendenzen

Die Unterschiede zwischen CLK und CMK weisen weitere Tendenzen auf, die nicht in die vorhergehenden Gruppen eingebunden

werden konnten. Es handelt sich dabei um die Tendenz zur Unterrepräsentation von Elementen oder Strukturen, die die gleichzeitige Aktivierung und Berücksichtigung mehrerer grammatischer, lexikalischer und pragmatischer Kenntnisse in Form von prozeduralen Regeln erfordern und um die Evolution einiger Phänomene hinsichtlich des Sprachstands, ausgehend von einer Einteilung in die Subkorpora MS und OS und das Korpus CMK.

Je mehr Regeln für die Anwendung eines Wortes zu beachten sind, desto seltener wird es im CLK verwendet. Ein Beispiel dafür ist das attributive Adjektiv, das für die Unterrepräsentation fakultativer deklinierter Elemente steht; die Tendenz zur Unterrepräsentation verstärkt sich weiter im partizipialen attributiven Adjektiv. Denn während für die Anwendung des Adjektivs nur die Deklination zu beachten ist, erscheint bei dem Adjektiv partizipialen Ursprungs zusätzlich die morphologische Komponente der Bildung des Partizips und zusätzlich von ihm abhängige Satzglieder. So liegt die mittlere Abweichung von ADJA (alle attributiven Adjektive) bei -21,73%, die des ADJA aus einem Partizip I (was aufgrund der fehlenden morphologischen Phänomene nur einen Teil der partizipialen Adjektive darstellte, zu der Darstellung vgl. Tabelle 39) bei 32,79% und die aus einem Partizip II sogar bei 45,14%.

Ein weiteres Beispiel für die Unterrepräsentation der Elemente, die die Anwendung mehrerer Regeln erfordern, sind

die verbalen Präfixe. Sie erscheinen in der Untersuchung zur Affigierung (vgl. 4.2.1.5), sind aber auch anhand von zwei Tags erkennbar, wie die abgetrennten Verbzusätze (PTKVZ) und die Infinitive mit *zu* (VVIZU).

Der Vergleich der drei Klassen verdeutlicht, daß mit zunehmender Anzahl von grammatischen Regeln das Phänomen im CLK stärker unterrepräsentiert ist. Während die Abweichung der verbalen Präfixe von ungetrennt erscheinenden Verben trotz der hohen individuellen Abweichungen bei insgesamt -16,45% liegt, steigt dieser Wert bei VVIZU auf -59,85% und bei PTKVZ auf -64,76% (vgl. Tabelle 36).

Eine weitere Tendenz der Korpora ist die Evolution des Sprachstands. Wird dieses anhand der Subkorpora in MS, OS und CMK eingeteilt, oder Fortgeschrittene, weit Fortgeschrittene und Muttersprachler, kann das jeweils erreichte Stadium anhand verschiedener sprachlicher Phänomene überprüft werden. Phänomene, die im Rahmen dieser Untersuchung ihre Relevanz in dieser Hinsicht bewiesen haben, sind die Wortlänge, die Satzlänge, die lexikalische Variation und die Substantive.

Für die Evolution nach Lernstadium in der Wortlänge ist die allgemeine Tendenz zwischen CLK und CMK zu beachten, d.h. hohe Werte im CLK im Bereich der kurzen Wörter (unter 8 Zeichen) und niedrige Werte im Bereich der langen Wörter (ab 8 Zeichen). Im Bereich der kurzen Wörter weist das Subkorpus MS die höchsten Werte auf, gefolgt von OS und CMK; im Bereich der

langen Wörter wendet sich die Tendenz zu niedrigsten Werten im MS, gefolgt von OS und CMK. Der Vergleich des untersten Bereiches der Wortlänge, von 1 bis 4 Zeichen, ist exemplarisch für diese Tendenz (vgl. Tabelle 102).

TABELLE 102

Die Nebeneinanderstellung der Gruppierung von 9 bis 15 Zeichen hingegen zeigt, daß im Bereich der langen Wörter eine Umkehrung stattfindet (vgl. Tabelle 103).

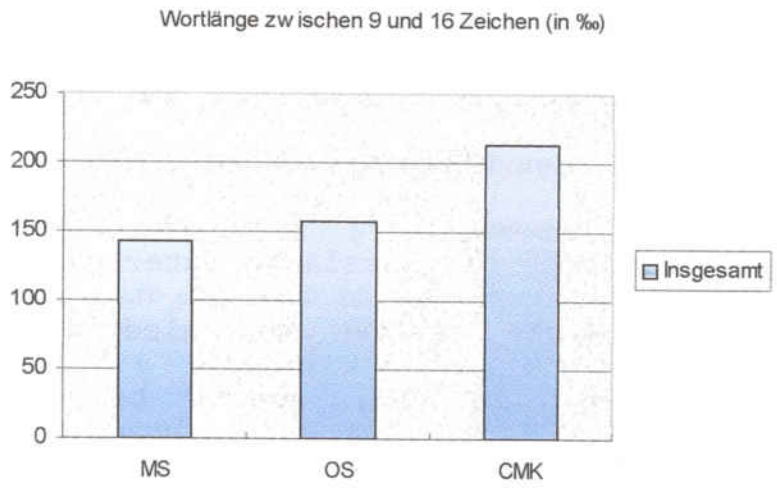


TABELLE 103

Dabei ergibt das CLK Werte in beiden Subkorpora, MS und OS, die zwar nicht mit denen des ganzen CMK vergleichbar sind, wohl aber mit denen der Textsorte „Bericht" des CMK. Während die Werte des CMK bei der Textsorte „Einleitung" und der Textsorte „Rezension" in der Gruppierung 1 bis 4 Zeichen um die 420‰ liegen, weist die Textsorte „Bericht" des CMK deutlich höhere Werte um die 500‰ auf. Diese Tendenz spiegelt sich, wenn auch nicht so stark akzentuiert, im Bereich zwischen 9 und 16 Zeichen wider; das CLK erreicht Frequenzen um die 150‰ in diesem Bereich, das CMK Mittelwerte über 210‰, die Textsorte „Bericht" des CMK aber nur um die 190‰. Zusammenfassend bedeutet dies, daß im untersten Bereich der Wortlänge das CLK sich der Textsorte „Bericht" des CMK nähert, und dieser deutlich von den Textsorten „Einleitung" und „Rezension" getrennt ist, die Wortlänge also ein texsortenspezifisches Phänomen darstellt, das zur

Unterscheidung herangezogen werden kann. Demnach wären die Mittelwerte des CLK in dieser Hinsicht stilistisch dem formal relativ freien Bericht näher als den Mittelwerten des CMK.

Ebenso wie die Wortlänge zeigt auch die Satzlänge Unterschiede in den Subkorpora auf. Die Werte reichen von niedrig im Subkorpus MS bis hin zu hoch in CMK, sowohl bei der Zusammenfassung der Werte der verschiedenen Textsorten, als auch, wenn sie getrennt betrachtet werden (vgl. Tabelle 18). Dabei muß allerdings auf die Ausnahme der Textsorte „Einleitung" des MS hingewiesen werden, die aufgrund der beschränkten Wortzahl des Subkorpus *MS Einleitung* anhand einer Erweiterung dieses Subkorpus überprüft werden müßte.

Die lexikalische Variation ist ein weiterer Indikator für die Evolution des Sprachstands. Anhand der Type-Token-Ratio ist die Evolution des Wortschatzes meßbar, die wie zu erwarten von niedrigen Werten im Subkorpus MS über mittleren Werten im OS hin zu höchsten Werten im CMK geht. MS erreicht in den gemessenen 200-Wort-Segmenten eine Type-Token-Ratio von 53,10, OS 55,42 und CMK 59,56. Dabei unterliegt die Type-Token-Ratio textsortenspezifischen Schwankungen. In der Textsorte „Bericht" des CMK werden die höchsten Werte erzielt, was wahrscheinlich auf inhaltsbezogene Faktoren zurückzuführen ist; die Werte der Textsorten „Einleitung" und „Rezension" hingegen liegen stabil bei ugf. 58. Im CLK hingegen sind diese größeren Schwankungen nicht zu verzeichnen (vgl. Tabelle 23).

Substantive und Partizipien sind weitere Indikatoren für den Sprachstand. Substantive werden normalerweise zusammen mit anderen sprachlichen Phänomenen als Indikatoren für einen nominal geprägten Stil gewertet, gehen im CMK von niedrigen Werten in der Textsorte „Bericht" über mittleren Werten in der Textsorte „Einleitung" zu höchsten Werten in der Textsorte „Rezension". Dies kongruiert mit der formalen Vorgabe dieser Textsorten, wobei der Bericht relativ frei ist, die Einleitung den wissenschaftlichen Texten angehört und die Rezension die höchste Informationskonzentration erfordert (vgl. Bußmann 1990: 530).

Im CLK ist diese Entwicklung nicht in gleichem Maße präsent. Die Berichte aller Textsorten und aller Subkorpora weisen vergleichbare Werte auf und auch die Einleitung erreicht steigende Werte sowohl im Subkorpus MS als auch im Subkorpus OS dar, doch keiner der Subkorpora des CLK steigt weiter in der Rezension, wie es der Fall des CMK ist. Die Übereinstimmung des Subkorpus MS und des Subkorpus OS in diesem Sinne weist auf textsortenspezifische Unterschiede hin, die im CMK realisiert werden, im CLK aber nicht. Die dargestellte Entwicklung der Substantive in den Textsorten muß aber nicht nur lernbedingt sein, denn auch kulturelle Faktoren können hier einen Einfluß haben, im Sinne, daß das CLK die Textsorten stilistisch anders realisiert als das CMK, wobei aber nicht gesagt werden kann, daß die eine oder andere Realisierung besser oder schlechter sei. Aus diesem Grund

scheint es uns nicht ratsam, die Frequenz der Substantive als allgemein lernbedingtes Phänomen anzugeben, wohl aber als stilistisches Indiz.

Substantive sind aber nicht das einzige sprachliche Phänomen, das die Entwicklung des CMK von einem stilistisch freien Bericht bis hin zu einer stilistisch stark markierten Rezension durchmachen. Auch die Partizipien folgen einer textsortenspezifischen Entwicklung. Im CMK steigt ihre Frequenz von der Textsorte „Bericht" über die Textsorte „Einleitung" bis zur Textsorte „Rezension", von 18 über 23 bis 30% (vgl. Tabelle 92). Im CLK jedoch steigt die Frequenz nur von der Textsorte „Bericht" zur Textsorte „Einleitung", stagniert dann aber mit Werten von 15, 21 und 22%. Substantive sind jedoch ein Hinweis dafür, daß nicht alle abweichenden sprachlichen Phänomene für die Bestimmung und Bewertung des Sprachstands herangezogen werden können.

5.2 Erweiterungen, Einschränkungen und mögliche Forschungsbereiche

Aus den vorhergehenden Darstellungen geht hervor, daß die Anwendung konkreter sprachlicher Phänomene quantitativ dargestellt werden kann und die dadurch bereitgestellten Daten einer Interpretation zur Verfügung stehen. Dies ist sogar auf zwei relativ beschränkten Ebenen möglich, einerseits der der sprachlichen Phänomene, die aus allgemeinen statistischen Verfahren hervorgehen, andererseits der der Wortarten, die

hier durch die syntaktischen Wortarten von Schiller et al. (1995) bestimmt wurden.

Die quantitative Beschreibung eröffnet so die Möglichkeit der Feststellung von Abweichungen in der Verwendung sprachlicher Phänomene. Im Gegensatz zur Auszeichnung von Fehlern, die nur fehlerhafte Äußerungen des Lerners behandelt (vgl. Nickel 1972), ermöglicht die Darstellung der angewendeten Phänomene Untersuchungen zum Grad, in dem bestimmte sprachliche Phänomene verwendet werden (Leech 1998).

Die Abweichung einzelner Phänomene steht somit für Interpretationen bereit, wobei allerdings im Hinblick auf Ursachenforschungen die Gruppierung der Phänomene nach sich gegenseitig beeinflussenden Aspekten produktiver scheint.

Der in unserer Arbeit vorgestellte Ansatz kann durch verschiedene methodische Zusätze zu präziseren Auswertungen des vorhandenen textuellen Materials führen. Zusätze wären in dieser Hinsicht einerseits morphologische und syntaktische Auszeichnungen, und je nach Forschungsbereich, in dem das empirische Material verwendet werden soll, weitere zu bestimmende Auszeichnungstypen, wie z.B. zu textstrukturierenden Elementen.

Morphologische Auszeichnungen tragen zu einer präziseren Bestimmung der sprachlichen Phänomene bei. So helfen sie beispielsweise, adjektivisch verwendete Partizipien von nicht deverbativen Adjektiven zu unterscheiden, oder eine nähere

Untersuchung zur Verwendung von Deklinationsformen der Substantive durchzuführen, was besonders interessant für Genitiv- und Dativobjekte wäre.

Zusätzlich zu den POS-Auszeichnungen können syntaktische Auszeichnungen helfen, die Frequenz und die Verteilung syntaktischer Elemente näher zu bestimmen und so Frequenzen der POS-Auszeichnung zu präzisieren. Auf diese Weise könnte beispielsweise die Verwendung von regierten und nicht regierten Präpositionen voneinander abgegrenzt werden.

Zugleich unterliegt der hier dargestellte Ansatz jedoch auch methodischen Einschränkungen, die bei jeder Interpretation zu berücksichtigen sind. So weist jede automatisierte Wortartenklassifizierung Probleme auf, wobei die von STTS keine Ausnahme bildet. Die Anwendung zwar komplementärer, aber nicht immer vereinbarer theoretischer Ansätze für die Bestimmung der Wortarten, die zudem noch von computationellen Faktoren bedingt wird, führt zu den schon dargestellten Überschneidungen, die bei der Beschreibung der einzelnen Wortklassen angegeben und berücksichtigt wurden (vgl. 4.2.2). Hier sei erneut die problematische Zuweisung von Adjektiven und Partizipien als Beispiel genannt. Obwohl solche Überschneidungen im Allgemeinfall sehr wenige Wortarten betreffen und zudem selten erscheinen, ist diese Einschränkung für entsprechende Untersuchungen zu beachten.

Als problematischer in diesem Ansatz erweist sich die Entscheidung für die Durchführung der Fehlerkorrektur, bevor die Texte in das Computerkorpus aufgenommen werden. Sie beruht auf technischen und theoretischen Überlegungen. Aus technischem Gesichtspunkt führt eine automatisierte Auszeichnung von stark fehlerhaften Texten zu so zahlreichen falschen Zuweisungen seitens des Taggers, daß die Ergebnisse nicht mehr interpretierbar wären. Aus theoretischer Sicht hingegen wäre mit nicht korrigierten Texten das Ziel dieser Arbeit, Abweichungen im Sprachgebrauch aufzudecken, verfälscht worden. Ein fälschlicherweise klein geschriebenes, deverbatives Substantiv wäre vom Tagger beispielsweise als Verb ausgezeichnet worden, und nicht als Substantiv. In diesem Sinne mußten die Texte konservativ korrigiert werden, und ergänzend könnten die Ergebnisse der vorliegenden Untersuchung in Zusammenhang mit einer Untersuchung zu der Frequenz von Fehlerkategorien interpretiert werden.

Zudem muß berücksichtigt werden, daß es sich bei diesem Ansatz um eine beschreibende Methode handelt, die nicht direkt zu Interpretationen führen kann und auch nicht Ziel in sich selbst ist. Erklärungen können nur innerhalb theoretisch abgegrenzter Forschungsbereiche gegeben werden, von denen hier nur einige genannt werden sollen.

Die Verwendungsmöglichkeiten eines korpuslinguistischen Ansatzes zur Untersuchung von Lernersprache sind zahlreich und vielseitig. Im englischen Sprachraum hat sich in den letzten

Jahren ein eigener Forschungsbereich durchgesetzt, der die didaktischen Einsatz von Computerkorpora untersucht[74]. Anwendbar sind sie zur Erstellung von Vokabellisten (z.B. unregelmäßiger Verben, vgl. Grabowski/Mindt 1995), zum Selbstlernen und zur Selbstevaluation oder zur Leistungsmessung seitens des Dozenten bzw. Lehrers.

In der Lexikologie tragen sie zur Erstellung von Lernerwörterbüchern oder Glossaren bei, sowohl, was die Auswahl der Einträge betrifft, als auch in bezug auf die erscheinenden Erklärungen.

Für die Lernersprachenanalyse bedeuten korpuslinguistische Methoden einen methodisch präziseren Ansatz zur Untersuchung von Lernersprache, der Phasen, die Lerner durchmachen, und kann auch bedeutend zur Erklärung zahlreicher Äspekte hinzugezogen werden.

In der Textsortenforschung hingegen können korpuslinguistische Vorgehensweisen zu einer klareren Trennung der Textsorten hinsichtlich der sie bedingenden kulturellen Faktoren führen, in der Linie der interkulturellen korpuslinguistischen Untersuchungen Leechs aus lexikalischer Ebene (Leech/Fallon 1992) oder der einsprachigen Untersuchungen von Haan (1996) zu idiosinkratischen Abweichung hinsichtlich der Verwendung der direkten Redewiedergabe.

---

[74] Eine umfangreiche Übersicht befindet sich in Johns (1998).

Somit erweist sich Korpuslinguistik für die Untersuchung der schriftlichen Textproduktion von Lernern einerseits als nützliches methodisches Werkzeug, andererseits als zukunftsträchtiges Mittel, das in zahlreichen Forschungsbereichen Anwendung finden kann.

6 Bibliographie

Aarts, Jan (1996): Grammatical annotation. In: *ICAME Journal. Computers in English Linguistics* Vol. 20, 1996, S. 104-107.

Aarts, Jan / Haan, Pieter de / Oostdijk, Nelleke (Hrsg.) (1993): *English Language Corpora: Design, Analysis and Exploitation. Papers from the thirteenth International Conference on English Language Research on Computerized Corpora, Nijmegen 1992.* Amsterdam: Rodopi 1993.

Adamzik, Kirsten (1995): Textsorten - Texttypologie. Eine kommentierte Bibliographie. Münster: Nodus Publikationen (=Studium Sprachwissenschaft 12).

Aijmer, Karin /Altenberg, Bengt (Hrsg.) (1991): *English corpus linguistics: studies in honour of Jan Svartvik.* London: Longman

Altenberg, Bengt (1993): Recurrent verb-complemental constructions in the London-Lund Corpus. In: Aarts / Haan / Oostdijk 1993, S: 227-245.

Altenberg, Bengt / Tapper, Marie (1998): The use of adverbial connectors in advanced Swedish learners' written English. In: Granger 1998a: 80-93.

Altman, Gabriel (1993): Science and Linguistics. In: Köhler / Rieger 1993: S. 3-10.

Alvar Ezquerra, Manuel / León Hurtado, Luís (1994): Las industrias de la lengua y las aplicaciones de los córpora. In: *Scripta Philologica In Memoriam Manuel Taboada Cid.* Tomo I. Coruña: Servicio de Publicaciones Universidad de Coruña, S. 32-46

ANNO (1996): State of the Art Report. Centre for Computational Linguistics. Faculty of Arts. Katholieke Universiteit Leuven: http://www.ccl.kuleuven.ac.be/about/ANNO/TEKST/begin.html

Aston, Guy (1998): Learning English with the British National Corpus. Paper presented at 6th Jornada de Corpus, UPF, Barcelona, May 1998. http://www.sslmit.unibo.it/guy/barc.htm 10.12.98

Atkins, S. / Clear, J / Ostler, N. (1992): Corpus design criteria. *Literary and Linguistic Computing* 7: 1-16.

Baker, Paul / Burnard, Lou / McEnery, Anthony / Wilson, Andrew (1997): An analytic framework for the validation of language corpora. Workpackage 2. Report for the ELRA Corpus Validation Group. 1. Dec 1997. http://www.icp.grenet.fr/ELRA/valid/wman/index.htm

Barkema, Henk (1993): Idiomaticity in English NPs. In: Aarts / Haan / Oostdijk 1993, S: 257-278.

Barlow, Michael (1998): MonoConc Concordance Programs for Text Analysis. http://www.ruf.rice.edu/~barlow/mono.html 28.12.98

Barnbrook, Geoff (1996): *Language and Computers. A practical introduction to the computer analysis of language.* Edinburgh: Edinburgh University Press (=Edinburgh Textbooks in Empirical Linguistics).

Bauer, Laurie (1993): Progress with a Corpus of New Zealand English and some early results. In: Souter / Atwell 1993, S. 1-10.

Bausch, Karl-Richard / Christ, Herbert / Krumm, Hans-Jürgen (Hrsg.) (1995): Handbuch Fremdsprachenunterricht. 3., überarb. und erw. Aufl. Tübingen, Basel: Francke

Berglund, Ylva (1997): Future in Present-Day English: Corpus-based evidence on the rivalry of expressions. In: *ICAME Journal. Computers in English Linguistics* Vol. 21, 1997, S. 7-19

Bernárdez, Enrique (Ed.)(1987): Lingüística del texto. Madrid: Cátedra.

Biber, Douglas (1988): *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, Douglas (1990a): Methodological issues regarding corpus-based analyses of linguistic variation. In: Literary and Linguistic Computing, 5: 257-269.

Biber, Douglas (1993a): *Representativeness in Corpus Design.* In: Literary and Linguistic Computing, Vol. 8, No. 4, 1993. Oxford: Oxford University Press.

Biber, Douglas (1993b): Co-ocurrence Patterns among Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition. In: Computational Linguistics, September 1993, Volume 19, Number 3, S. 531-538

Biber, Douglas (1993c): Using Register-Diversified Corpora for General Language Studies, S. 231. In: Computational Linguistics.

Biber, Douglas (1995a): *Dimensions of register variation. A cross-linguistic comparision.* Cambridge: Cambridge University Press.

Biber, Douglas / Finegan, Edward (1991): On the exploitation of computerized corpora in variation studies. In: Aijmer /Altenberg 1991: S. 204-220.

Biber, Douglas / Finegan, Edward (1994): Intra-textual variation within medical research-articles. In: Oostdijk / Haan 1994, S. 201-221.

Blackwell, Susan (1993): From dirty data to clean language. In: Aarts / Haan / Oostdijk 1993 S. 97-105.

Briscoe, Ted / Waegner, Nick (1993): Undergeneration and robust parsing. In: Aarts / Haan / Oostdijk 1993, S: 181-196.

British National Corpus (1997a): What ist the BNC? http://info.ox.ac.uk/bnc/what/index.html 09.11.97

British National Corpus (1997b): Linguistic annotation ("tagging") http://info.ox.ac.uk/bnc/what/ucrel.html

Briz Gómez, Antonio (1998): El español coloquial en la conversación. Barcelona: Ariel.

Bünting, Karl-Dieter; Axel Bitterlich; Ulrike Pospiech (1996): Schreiben im Studium: ein Trainingsprogramm. Berlin: Cornelsen Scriptor

Burnage, Gavin / Dunlop, Dominic (1993): Encoding the Britsh National Corpus. In: Aarts / Haan / Oostdijk 1993, S: 79-95.

Burnard, Lou (1992): Tools and Techniques for Computer-assisted Text Processing. In: Butler 1992, S. 1-28.

Burnard, Lou (Hrsg.) (1995): British National Corpus. Users Reference Guide. British National Corpus Version 1.0. ftp://ota.ox.ac.uk/pub/ota/BNC/urg.pdf 04.10.98

Burrows, John F. (1992): Computers and the Study of Literature. In: Butler 1992, S. 167-204.

Bußmann, Hadumod (1990): Lexikon der Sprachwissenschaft. Zweite, völlig neu bearbeitete Auflage. Unter Mithilfe und mit

Beiträgen von Fachkolleginnen und -kollegen. Stuttgart: Alfred Kröner (=Kröners Taschenausgabe 452).

Butler, Christopher S. (1985): Statistics in Linguistics. Oxford, New York: Basil Blackwell

Butler, Christopher S. (Hrsg.) (1992): Computers and Written Texts. Oxford (UK), Cambridge (USA), Blackwell (=Applied Language Studies. Edited by David Crystal and Keith Johnson)

Castell, Anderu (1997): Gramática de la lengua alemana. Editorial Idiomas.

CCL Centre for Computational Linguistics (1996): Anno State of the Art Report. Faculty of Arts. Katholieke Universiteit Leuven. http://www.ccl.kuleuven.ac.be/about/ANNO/TEKST/begin.html

Chafe, Wallace L. / du Bois, John W. / Thompson, Sandra A. (1991): Towards a new corpus of spoken American English. In Aijmer / Altenberg (1991), S. 64-82.

Cherubim, Dieter (Hrsg.) (1980): Fehlerlinguistik. Beiträge zum Problem der sprachlichen Abweichung. Tübingen: Niemeyer (Reihe Germanistische Linguistik 24).

Childes (1998): CHILDES The Child Language Data Exchange System. http://atila-www.uia.ac.be/childes/ 05.10.98

Clément, Danièle (1996): Linguistisches Grundwissen. Opladen: Westdeutscher Verlag (=WV-Studium Bd. 173: Linguistik)

Cock, Sylvie de / Granger, Sylviane / Leech, Geoffrey / McEnery, Tony: An automated approach to the phrasicon of EFL learners. In: Granger 1998a: 67-79.

Collins, Peter (1991): The modals of obligation and necessity in Australian English. In: Aijmer /Altenberg 1991: S. 145-165.

Collins, Peter (1991b): *Will* and *shall* in Australian English. In: Johansson / Stenström 1991: S. 182-199.

Collot, Milena / Belmore, Nancy (1993): Electronic language: A new variety of English. In: Aarts / Haan / Oostdijk 1993, S: 42-55.

Corcoll, Brigitte / Corcoll, Roberto (1994): Programm. Alemán para hispanohablantes. Barcelona: Herder.

Crystal, David (1991): Stylistic profiling. In: Aijmer /Altenberg 1991: S. 221-238.

Crystal, David (1995): Die Cambridge Enzyklopädie der Sprache. Frankfurt / New York: Campus.

Doughty, Catherine (1991): Computer Applications in Second Language Acquisition. Research: Design, Description, and Discovery. In: Pennington / Stevens 1991, S. 127-154.

Drubig, Bernhard (1972): Zur Analyse syntaktischer Fehlleistungen. In: Nickel 1972, 78-91.

Dulay, Heidi / Burt, Marina / Krashen, Stephen (1982): Language Two. New York, Oxford: Oxford University Press.

Dürr, Michael / Schlobinski, Peter (^{2}1994): Einführung in die deskriptive Linguistik. Opladen: Westdeutscher Verlag (=WV Studium 163)

EAGLES (1996a): Preliminary recommendations on Corpus Typology. EAGLES document EAG-TCWG-CTYP/P. Version of May, 1996.

EAGLES (1996b): Preliminary recommendations on Text Typology. EAGLES document EAG-TCWG-TTYP/P. Version of Jun, 1996.

EAGLES (1996c): Preliminary recommendations on spoken text corpora. ftp://ftp.ilc.pi.cnr.it/pub/eagles/corpora/spokentx.ps.gz 10.12.98

Ellis, Rod (1994): The Study of Second Language Acquisition. Oxford: Oxford University Press.

Engel, Ulrich (1988): Deutsche Grammatik. 2., verbesserte Auflage. Heidelberg: Julius Groos.

Fernández-Villanueva Jané, Marta (1989): La lingüística del texto. Tesis de licenciatura dirigida por el Dr. Javier Orduña. Universidad de Barcelona. Septiembre de 1989.

Fleischer, Wolfgang (1982): Phraseologie der deutschen Gegenwartssprache. Leipzig: Bibliographisches Institut.

Fleskes, Gabriele (1996): Untersuchungen zur Textsortengeschichte im 19. Jahrhundert. Tübingen: Niemeyer (=Reihe Germnaistische Linguistik 176)

Francis, W. Nelson (1979a): A Tagged Corpus - Problems and Prospects. In: Greenbaum, S.; Leech, G.; Svartvik, J. (Hrsg.)(1979): Studies in English Linguistics for Randolph Quirk. London: Longman.

Francis, W. Nelson (1980): A tagged corpus - problems and prospects. In Greenbaum / Leech / Svartvik 1980, 192-209

Francis, W.N. (1982): Problems of assembling and computerizing large corpora. In: *ICAME Journal. Computers in English Linguistics* Vol. 16, 1992, S. 83

Francis, W.N. / Kucera, H. (1979): *Brown Corpus Manual*. Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Revised 1971. Revised and amplified 1979. http://www.hd.uib.nora.no/icame/brown/bcm.html 09/12/98

Garside, Roger / Leech, Geoffrey / McEnery, Anthony (Hrsg.) (1997): *Corpus Annotation. Linguistic Information from Computer Text Corpora*. London, New York: Longman.

Gibbon, D. (Hrsg.) (1996): Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference, Bielefeld, October 1996. Berlinn: Mouton de Gruyter

Gnutzmann, Claus (1972): Zur Analyse lexikalischer Fehler. In: Nickel 1972, 67-72.

Grabowski, Eva / Mindt, Dieter (1995): A corpus-based learning list of irregular verbs in English. In: *ICAME Journal. Computers in English Linguistics* Vol. 19, 1995, S. 5-13

Grabowski, Joachim (1995): Schreiben als Systemregulation. Ansätze einer psychologischen Theorie der schriftlichen Sprachproduktion. In: Jakobs / Knorr / Molitor-Lübbert (1995), 11-34.

Granger, Sylviane (1993): International Corpus of Learner English. In: Aarts / Haan / Oostdijk 1993, S: 57-71.

Granger, Sylviane (1996): Exploiting Learner Corpus Data in the Classroom: Form Focused

Instruction and Data Driven Learning. Vortrag gehalten am 9. August 1996. Talc96 Teaching and Language Corpora. Lancaster University, UK, 9th-12th August, 1996.

Granger, Sylviane (1998b): The computer learner corpus: a versatile new source of data for SLA research. In: Granger 1998a: 3-18.

Granger, Sylviane (Hrsg.) (1998a): Learner English on Computer. Longman: London, New York (=Studies in Language and Linguistics).

Greenbaum, Sidney (1991): The development of the International Corpus of English. In Aijmer / Altenberg (1991), S. 83-91.

Greenbaum, Sidney / Leech, Geoffrey / Svartvik, Jan (Hrsg.) (1980): *Studies in English linguistics*. London, New York: Longman.

Greenbaum, Sidney / Nelson, Gerald / Weitzman, Michael (1996): Complement Clauses in English. In: Thomas / Short 1996, S. 76-91.

Grefenstette, Gregory / Tapanainen, Pasi (1984): What is a word, What is a sentence? Problems of Tokenization. Rank Xerox Research Centre. Grenoble Laboratory. Internet.

Grotjahn, R. / Altman, G. (1993): Modelling the Distribution of Word Length: Some Methodological Problems. In: Köhler / Rieger 1993: S. 141-153.

Gülich, Elisabeth / Raible, W. (1972): Textsorten. Differenzierungskriterien aus linguistischer Sicht. Frankfurt a.M.

Haan, Pieter de (1991): On the exploration of corpus data by means of problem-oriented tagging: Postmodifying clauses in the English noun phrase. In: Johansson / Stenström 1991: S. 51-65.

Haan, Pieter de (1992a): The optimum corpus sample size? In: Leitner (1992a)

Haan, Pieter de (1996): More on the language of dialogue in fiction. In: *ICAME Journal. Computers in English Linguistics* Vol. 20, 1996, S. 27-40

Halliday, M.A.K. (1985a): An Introduction to Functional Grammar. London: Edward Arnold Publishers.

Hausser, R. (Hrsg.) (1996): Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994. Tübingen: Niemeyer.

Heidelberger Forschungsprojekt "Pidgin-Deutsch" (1978) *The Acquisition of German Syntax by Foreign Migrant Workers*, in Sankoff 1978, S. 1-21

Helbig, Gerhard (1988): Entwicklung der Sprachwissenschaft seit 1970. Leipzig

Helbig, Gerhard / Buscha, Joachim (1991): Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. 14. durchgesehene Auflage. Berlin, München, Leipzig usw.: Langenscheidt.

Hentschel, Elke / Weydt, Harald (1990): Handbuch der deutschen Grammatik. Berlin, New York: de Gruyter.

Hinck, Walter (Hrsg.)(1977): Textsortenlehre - Gattungsgeschichte. Heidelberg: Quelle und Meyer (=Medium Literatur 4).

Hoey, Michael (1993): How the word reason is used in texts. In: Sinclair / Hoey / Fox 1993, S. 67-82.

Hofland, K. / Johansson, S. (1982): Word frequencies in British and American English. London: Longman.

Holmes, Janet (1994): Inferring language change from computer corpora: Some methodological problems. In: *ICAME Journal. Computers in English Linguistics* Vol. 18, 1994, S. 27-40

Holmes, Janet; Vine, Bernadette; Johnson, Gary (1998): Guide to The Wellington Corpus of Spoken New Zealand English. http://www.hit.uib.no/icame/wsc/index.htm 01.10.98

Hundt, Marianne; Sand, Andrea; Siemund, Rainer (1998): Manual of Information accompany The Freiburg - LOB Corpus of British English (‚FLOB') http://www.hit.uib.no/icame/flob/index.htm

ICAME (1998): Icame Corpus Collection - Information: Lancaster Parsed Corpus. http://www.hit.uib.no/icame/lanpeks.html 10.12.98

ICAME Journal. Computers in English Linguistics, Num. 16, 1992

ICAME Journal. Computers in English Linguistics, Num. 17, 1993

ICAME Journal. Computers in English Linguistics, Num. 18, 1994

ICAME Journal. Computers in English Linguistics, Num. 19, 1995

ICAME Journal. Computers in English Linguistics, Num. 20, 1996

ICAME Journal. Computers in English Linguistics, Num. 21, 1997

Ide, Nancy (1996): Corpus Encoding Standard. http://www.cs.vassar.edu/CES/ 05.11.98

Institut für Deutsche Sprache (1996a): Freiburger Korpus (fko). http://www.ids-mannheim.de/ldv/cosmas/corpora-fko.html (12.06.98)

Institut für Deutsche Sprache (1996b): PFEFFER-Korpus (pfe). http://www.ids-mannheim.de/ldv/cosmas/corpora-pfe.html (12.06.98)

Institut für Deutsche Sprache (1996c): Textkorpora des IDS. http://www.ids-mannheim.de/ldv/cosmas/corpora.html (12.06.98)

Institut für Deutsche Sprache (1996d): Wendekorpus (wk). http://www.ids-mannheim.de/kt/corpora-wk.html (06.01.99)

Institut für Deutsche Sprache (1996e): Dialogstrukturenkorpus (dsk). http://www.ids-mannheim.de/ldv/cosmas/corpora-dsk.html (12.06.98)

Institut für Deutsche Sprache (1998a): Cosmas. http://www.ids-mannheim.de/ldv/cosmas/cosmas.html (12.06.98)

Institut für Deutsche Sprache (1998b): Cosmas II. http://www.ids-mannheim.de/ldv/cosmas2/ cosmas2.html (12.06.98)

Institut für Deutsche Sprache (1998c): Textkorpora des IDS. http://www.ids-mannheim.de/ldv/cosmas/corpora-ges.html (12.06.98)

Institut für Deutsche Sprache (1998d): Limas-Korpus (lim). http://www.ids-mannheim.de/kt/corpora-lim.html (06.01.99)

Institut für Deutsche Sprache (1998e): Korpus Magazin Lufthansa Bordbuch / DEUTSCH (mld). http://www.ids-mannheim.de/kt/corpora-mld.html (06.01.99)

Institut für maschinelle Sprachverarbeitung (1998): TreeTagger - ein sprachunabhängiger Wortart-Tagger. http://www.ims.uni-stuttgart.de/www/Tools/DecisionTreeTagger-de.html (10.06.98)

Isenberg, Horst (1983): Cuestiones fundamentales de tipología textual. In: Bernárdez 1987.

Jakobs, Eva-Maria / Knorr, Dagmar / Molitor-Lübbert, Sylvie (Hrsg.) (1995): Wissenschaftliche Textproduktion. Mit und ohne Computer. Frankfurt a.M., Berlin u.a.: Peter Lang. S. 179-192.

Johansson, Stig (1991a): Times change, and so do corpora. In: Aijmer /Altenberg 1991: S. 305-314.

Johansson, Stig (1991b): Computer Corpora in English language research. In: Johansson / Stenström 1991: S. 3-6.

Johansson, Stig / Leech, Geoffrey N. / Goodluck, Helen (1978); Manual of Information to accompany The Lancaster-Oslo/Bergen Corpus of British English, for use with Digital Computers. http://www.hd.uib.nora.no/icame/lob/lob-dir.html

Johansson, Stig / Stenström, Anna-Brita (Hrsg.) (1991): *English Computer Corpora. Selected Papers and Research Guide*. Berlin, New York: Mouton de Gruyter.

Johns, Tim (1998): Tim Johns Data-driven Learning Page. http://sun1.bham.ac.uk/johnstf/timconc.htm (10.12.98)

Kammer, Manfred (1993): Korpora geschriebener Sprache. In: Lenders 1993a, S. 49-62

Kasper, Gabriele (1995): Funktionen und Formen der Lernersprachenanalyse. In: Bausch / Christ / Krumm (1995), S. 263-267.

Kjellmer, Göran (1991): A mint of phrases. In: Aijmer /Altenberg 1991: S. 111-127.

Köhler, Reinhard / Rieger, Burghard B. (Hrsg.) (1993): Contributions to Quantitative Linguistics. Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier, 1991. Dordrecht, Boston, London: Kluwer Academic Publishers.

Königs, Frank G. (1995a): Fehlerkorrektur, in Bausch / Christ / Krumm (1995) S. 268-272.

Kuhs, Katharina (1989): Sozialpsychologische Faktoren im Zweitsprachenerwerb. Eine Untersuchung bei griechischen Migrantenkindern in der Bundesrepublik Deutschland. Eine Pilotstudie auf der Grundlage von schriftlichen Texten. Tübingen: Narr (=Tübinger Beiträge zur Linguistik. Ser. A, Language Development 10)

Kwon, Heok-Seung (1997): Negative Prefixation from 1300 to 1800: A case study in *in-/un-* variation. In: *ICAME Journal. Computers in English Linguistics* Vol. 21, 1997, S. 21-42

Kytö, Merja (1989): Progress report on the diachronic part of the Helsinki Corpus. In: *ICAME Journal Computers in English Linguistics* Vol. 13, 1989, S. 12-15

Kytö, Merja (1993): A supplement to the Helsinki Corpus of English Texts: The Corpus of Early American English. In: Aarts / Haan / Oostdijk 1993, S. 3-10.

Kytö, Merja / Rissanen, Matti (1996): English historical corpora: Report on developments in 1995. In: *ICAME Journal. Computers in English Linguistics* Vol. 20, 1996, S. 117-133.

Kytö, Merja / Voutilainen, Atro (1995): Applying the Constraint Grammar Parser of English to the Helsinki Corpus. In: *ICAME Journal. Computers in English Linguistics* Vol. 19, 1995, S. 23-48

Labov, William (1983): Modelos sociolingüísticos. Übersetzung von José Miguel Marinas Herreras. Madrid: Ediciones Cátedra.

Lancashire, Ian (1993): The Early Modern English Renaissance Dictionaries Corpus. In: Aarts / Haan / Oostdijk 1993, S: 11-24.

Leech, Geoffrey (1991a): The state of the art in corpus linguistics. In: Aijmer /Altenberg 1991: S. 8-29

Leech, Geoffrey (1993): Corpus Annotation Schemes. In: Literary and Linguistic Computing, Vol. 8, No. 4, 1993, S. 275-281.

Leech, Geoffrey (1997a): Teaching and Language Corpora: a Convergence. In: Wichmann / Fligelstone / McEnery / Knowles 1997: S. 1-23

Leech, Geoffrey (1997b): A Brief User's Guide To The Grammatical Tagging Of The British National Corpus. http://info.ox.ac.uk/bnc/what/gramtag.html (10/12/98)

Leech, Geoffrey (1998): *Preface*. In: Granger 1998a: xiv-xxii.

Leech, Geoffrey / Fallon, Roger (1992): Computer Corpora - What do they tell us about culture? In: *ICAME Journal. Computers in English Linguistics*, Num. 16, April 1992

Leech, Geoffrey / Fligelstone, Steven (1992): Computers and Corpus Analysis. In: Butler 1992, S. 115-140.

Lehr, Andrea (1996): Kollokationen und maschinenlesbare Korpora. Ein operationales Analysemodell zum Aufbau lexikalischer Netze. Tübingen: Niemeyer (=Reihe Germanistische Linguistik 168)

Leidner, Jochen (1997): Korpusstatistiken. http://www.linguistik.uni-erlangen.de/~leidner/

Leitner, Gerhard (1992a): New Directions in English Language Corpora. Methodology, Results, Software Developments. Berlin, New York: Mouton de Gruyter, 1992

Lenders, Winfried (Hrsg.)(1993): Computereinsatz in der angewandten Linguistik. Frankfurt am Main: Lang

Lewis, C.I. (1966): Philosophy. In: Enciclopedia Americana. International Edition. Vol. 21, 779-771.

Lezius, Wolfgang (1996): Morphologiesystem MORPHY. In: Hausser 1996.

Lezius, Wolfgang / Rapp, Reinhard / Wettler, Manfred (1996): A Morphology-System and Part-of-Speech Tagger for German. In: Gibbon 1996.

Lloret, Maria Rosa / Boix, Emili / Lorente, Mercè / Payrató, Luís / Perea, M. Pilar (Hrsg.) (1997): Anàlisi de la variació lingüística. Actes de la 2ª jornada sobre variació lingüística i del 3r col.loqui lingüístic de la Universitat de Barcelona (Club-3). Barcelona: Promociones y Publicaciones Universitarias.

Løken, Berit (1997): Expressing possibility in English and Norwegian. In: *ICAME Journal. Computers in English Linguistics* Vol. 21, 1997, S. 43-59.

Lorente, Mercè / Cabré, M. Teresa / Yzaguirre, Lluís de (1997): Lèxic i variació. In: Lloret et al. S. 121-147.

Mair, Christian (1991): Quantitative or qualitative corpus analysis? Infinitival complement clauses in the Survey of English Usage corpus. In: Johansson / Stenström 1991: S. 67-80.

Marcos Marín, Francisco / Sánchez Lobato, Jesús (1988): Lingüística Aplicada. Madrid: Editorial Síntesis.

Marcos Marín, Francisco A. (1994): Informática y Humanidades. Madrid: Gredos

Marcus, Mitchell P. / Santorini, Beatrice / Marcinkiewicz, Mary Ann (1993): Building a large annotated corpus of English: the Penn Treebank", In: Computational Linguistics, September 1993, Volume 19, Number 3, S. 313-330

Maritxalar, M. / Díaz de Ilarraza, A. / Alegría, I. / Ezeiza, N. (1996): Modelización de la competencia gramatical en la interlengua basada en el análisis de corpus. Donostia: Lengoaia eta Sistema Informatikoak saila Euskal Herriko Unibertsitatea (UPV/EHU)

McEnery, Tony / Burnard, Lou / Wilson, Andrew / Baker, Paul (1998): Validation of Linguistic Corpora. http://www.icp.grenet.fr/ELRA/valid/wp3/index.htm

McEnery, Tony / Wilson, Andrew (1996a): Corpus Linguistics. Edinburgh: Edinburgh University Press.

Meijs, Willem (1992): Computers and Dictionaries. In: Butler 1992, S. 141-165.

Menge, Heinz H. (1993): Korpora gesprochener Sprache. In: Lenders 1993a

Meunier, Fanny (1998): Computer Tools for the analysis of learner corpora. In: Granger 1998a, S. 19-37.

Meurman-Solin, Anneli (1995): A New Tool: The Helsinki Corpus of Older Scots (1450-1700). In: *ICAME Journal. Computers in English Linguistics 1995 N° 19*

Meya, Montserrat / Huber, Wolfgang (1986): Lingüística computacional. Barcelona: Teide.

Meyer, Charles F. (1991): A corpus-based study of apposition in English. In: Aijmer /Altenberg 1991: S. 166-181.

Nevalainen, Terttu und Raumolin-Brunberg, Helena (1996): ZEN - The Zurich English Newspaper Corpus. In: *ICAME Journal. Computers in English Linguistics N° 20*

Nickel, Gerhard (1972b): Grundsätzliches zur Fehleranalyse und Fehlerbewertung. In: Nickel 1972a, S. 8-24.

Nickel, Gerhard (Hrsg.) (1972a): Fehlerkunde. Beiträge zur Fehleranalyse, Fehlerbewertung und Fehlertherapie. Berlin: Cornelsen-Velhagen & Clasing.

Odlin, Terence (1989): Language transfer. Cross-linguistic influence in language learning. Cambridge: Cambridge University Press.

Oostdijk, Nelle; Haan, Pieter de (1994): Clause patterns in Modern British English: A corpus-based (quantitative) study. In: *Icame Journal. Computers in English Linguistics*, Num. 18, 1994, S. 41-79

Oostdijk, Nelleke / Haan, Pieter de (Hrsg.) (1994): Corpus-based research into language. In honour of Jan Aarts. Amsterdam: Rodopi

Pelz, Heidrun (1996): Linguistik. Eine Einführung. Hamburg: Hoffmann und Campe (=Campe-Paperback)

Pennington, Martha C. / Stevens, Vance (Hrsg) (1991): Computers in Applied Linguistics: An International Perspective. Clevedon, Bristol, Adelaide: Multilingual Matters

Petch-Tyson, Stephanie (1998): Writer/reader visibility in EFL written discourse. In: Granger 1998a: 107-118.

Peters, Pam (1993): Corpus evidence on some points of usage. In: Aarts / Haan / Oostdijk 1993, S: 247-255.

Pienemann, Manfred / Jansen, Louise (1991): Computational Analysis of Language Acquisition Data. In: Pennington/Stevens 1991, S. 201-243.

PPCME (1998): The Penn-Helsinki Corpus of Middle English. http://ling.upenn.edu/mideng/ 05.01.99

Raabe, Horst (1980): Der Fehler beim Fremdsprachenerwerb und Fremdsprachengebrauch. In: Cherubim 1980, S. 61-93.

Renouf, A. (1987): Corpus development, in Sinclair 1987: 1-40

Renouf, Antoinette (1993): A word in time: First findings from the investigation of dynamic text. In: Aarts / Haan / Oostdijk 1993, S: 279-288.

Renouf, Antoinette / Sinclair, John M. (1991): Collocational frameworks in English. In: Aijmer /Altenberg 1991: S. 128-143.

Rothkegel, Annely (1995): Konzept für eine Werkbank zum Schreiben. In:

Rückriem, Georg; Stary, Joachim; Franck, Norbert (1983): Die Technik wissenschaftlichen Arbeitens: praktische Anleitung zum Erlernen wissenschaftlicher Techniken am Beispiel der Pädagogik, unter besonderer Berücksichtigung gesellschaftlicher und psychischer Aspekte des Lernens. München, Wien, Zürich: Schöningh

Ruipérez, Germán (1992): Gramática alemana. Madrid: Cátedra.

Sampson, Geoffrey (1991a): Analysed Corpora of English: A Consumer Guide. In: Pennington/Stevens 1991, S. 181-200.

Sankoff, David (Hrsg.) (1978): Linguistic Variation. Models and Methods. New York: Academic Press

Schafer, John C. (1981): The linguistic analysis of spoken and written Texts. In Barry M. Kroll und Roberta J. Vann (Hrsg.): *Exploring speaking-writing relationships: connection and contrasts*. Urbana, IL: National Council of Teachers of English. S. 1-31.

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1995): Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Draft vom 14. November 1995.

Schlieben-Lange, Brigitte (1991): Soziolinguistik. Dritte, überarbeitete und erweiterte Auflage. Stuttgart, Berlin, Köln: Kohlhammer (=Urban Taschenbücher 176).

Schlobinski, Peter (1996): Empirische Sprachwissenschaft. Opladen: Westdeutscher Verlag (=WV Studium 174)

Schmied, Josef; Hertel, Eva (1991): Lampeter Corpus of Early Modern English Tracts. http://www1.tu-chemnitz.de/~ehe/real/lampr.htm 10.12.98

Schmitz, Ulrich; Gerhardt, Tom C.: *Sprache und Datenverarbeitung. International Journal for Language data Processing.* Heft 1-2. 17. Jahrgang 1993

Schröder, Bernhard (1993): Sprachkorpora als linguistische Belegsammlungen. In: Schmitz 1993, S. 103-130

Schröder, Jochen (1986): Lexikon deutscher Präpositionen. Leipzig: Verlag Enzyklopädie.

Scott, Mike (1998): WordSmith Tools. http://www.liv.ac.uk/~ms2928/index.htm 01.07.98

Siemund, Rainer / Claridge, Claudia (1997): The Lampeter Corpus of Early Modern English Tracts. In: *ICAME Journal. Computers in English Linguistics* Vol. 21, 1997, S. 61-70.

Sinclair, John M. (1987): Looking up. Collins

Sinclair, John M. (1991a): Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sinclair, John M. / Hoey, Michael / Fox, Gwyneth (Hrsg.) (1993): Techniques of Description. Spoken and written discourse. A festschrift for Malcolm Coulthard. London, New York: Routledge.

Souter, Clive (1993): Towards a standard format for parsed corpora. In: Aarts / Haan / Oostdijk 1993, S: 197-212.

Souter, Clive / Atwell, Eric (1994): Using parsed corpora: A review of current practice. In: Oostdijk / Haan 1994, S. 143-158.

Souter, Clive / Atwell, Eric (Hrsg.) (1993): Corpus-based Computational Linguistics. Amsterdam: Rodopi.

Stary, Joachim; Horst Kretschmer (1994): Umgang mit wissenschaftlicher Literatur. Eine Arbeitshilfe für das sozial- und geisteswissenschaftliche Studium. Frankfurt am Main: Cornelsen Scriptor

Stein, Achim (1995): Maschinenlesbare Textkorpora für das Französische. In: *Zeitschrift für französische Sprache und Literatur*. Vol. 105(1995) Stuttgart: Franz Steiner Verlag, S. 1-25

Stein, Gabriele / Quirk, Randolph (1991): On having a look in a corpus. In: Aijmer /Altenberg 1991: S. 197-203.

Stubbs, Michael (1996a): Text and Corpus Analysis. Computer-assisted Studies of Language and Culture. Oxford: Blackwell.

Svartvik, Jan (1990b): Tagging and parsing on the TESS project.

Svartvik, Jan (Hrsg.) (1990a): The London-Lund Corpus of Spoken English. Description and Research. Lund: Lund University Press.

TACT (1998): Text Analysis Computing Tools (TACT) http://www.chass.utoronto.ca:8080/ cch/tact.html 28.12.98

Thomas, Jenny / Short, Mick (Hrsg.) (1996): Using Corpora for Language Research. London, New York: Longman.

Tottie, Gunnel (1992a): Grammar and corpus linguistics. In: *ICAME Journal. Computers in English Linguistics 1992*

Tuldava, Juhan (1995): Methods in Quantitative Linguistics. Trier: Wissenschaftlicher Verlag Trier (=Quantitative Linguistics 54).

Turell Julià, M. Teresa (1995b): La base teòrica i metodològica de la variació lingüística. In: Turell Julià 1995a, S. 17-49

Turell Julià, M. Teresa (Hrsg.) (1995a): *La sociolingüística de la variació*. Barcelona: Promociones y Publicaciones Universitarias.

Volk, Martin (1998): Token, Types, Häufigkeiten und automatische Wortartenerkennung (Statistik-basiertes Tagging). Morphologieanalyse und Lexikonaufbau. http://www.ifi.unizh.ch/CL/volk/LexMorphVorl/Lexikon06.Freq.html

Weinrich, Harald (1993): Textgrammatik der deutschen Sprache. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Wichmann, Anne / Fligelstone, Steven / McEnery, Tony / Knowles, Gerry (1997): Teaching and Language Corpora. London, New York: Longman (=Applied Linguistics and Language Study).

Wilson, Andrew / Rayson, P. (1993): The automatic content analysis of spoken discourse. In: Souter / Atwell 1993, S. 215-226.

Woods, Anthony / Fletcher, Paul / Hughes, Arthur (1986): Statistics in language studies. Cambridge: Cambridge University Press (=Cambridge Textbooks in Linguistics).

Wotjak, Barbara (1992): Verbale Phraseolexeme in System und Text. Tübingen: Niemeyer

Wright, Susan (1993): In search of history: English language in the eighteenth century. In: Aarts/Haan/Oostdijk 1993, S: 25-39.

Xerox (1997a): Morphology. http://www.xrce.xerox.com/research/mltt/fsNLP/tagger.html (19.11.98)

Xerox (1997b): Part of Speech Tagging. http://www.xrce.xerox.com/research/mltt/fsNLP/tagger.html (19.11.98)

Zierl, Marco (1997): Entwicklung und Implementierung eines Datenbanksystems zur Speicherung und Verarbeitung von Textkorpora. Magisterarbeit in der Philosophischen Fakultät II (Sprach- und Literaturwissenschaften) der Friedrich-Alexander-Universität Erlangen-Nürnberg. http://www. linguistik.uni-erlangen.de/tree/html/corsica/zierl97/ zierl97.html (01.08.98)

7 Anhang

Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS

Anne Schiller, Simone Teufel, Christine Stöckert
Universität Stuttgart
Institut für maschinelle Sprachverarbeitung

Christine Thielen
Universität Tübingen
Seminar für Sprachwissenschaft

Draft

14. November 1995

Inhaltsverzeichnis

1 Einleitung **3**

1.1 Zuweisung von Tags 4
1.2 Mehrwortlexeme 4
1.3 Behandlung von Abkürzungen 4
1.4 Behandlung von Fehlern im Text 5
1.5 STTS – Übersicht 6

2 Beschreibung der einzelnen Tags **9**

2.1 Nomina 9
2.1.1 NN: normale Nomina 9
2.1.2 NE: Eigennamen 13
2.2 Adjektive 16
2.2.1 ADJA: attributive Adjektive 17
2.2.2 ADJD: prädikativ oder adverbial gebrauchte Adjektive 22
2.2.3 ADJD oder VVPP? 23
2.3 Zahlen 26
2.3.1 CARD: Kardinalzahlen 26
2.4 Verben 28
2.4.1 VAFIN, VAIMP, VVFIN, VVIMP, VMFIN: finite Formen 28
2.4.2 VVINF, VAINF, VMINF, VVIZU: Inifinitiv 30
2.4.3 VVPP, VMPP, VAPP: Partizip Perfekt 31
2.5 Artikel 32
2.5.1 ART: bestimmter und unbestimmter Artikel 32
2.6 Pronomina 34
2.6.1 PPER, PRF: Personal– und Reflexivpronomina 34
2.6.2 PPOSAT, PPOSS: Possessivpronomina 37
2.6.3 PDAT, PDS: Demonstrativpronomina 38
2.6.4 PIDAT, PIS, PIAT: Indefinitpronomina 40
2.6.5 PRELAT, PRELS: Relativpronomina 48
2.6.6 PWAT, PWS: Interrogativpronomina 50
2.6.7 PWAV: adverbiale Interrogativ– oder Relativpronomina 52
2.6.8 PAV: Pronominaladverbien 53
2.7 Adverbien 55
2.7.1 ADV: "echte" Adverbien 55
2.7.2 ADJD oder ADV? 56

2.8 Konjunktionen 58
2.8.1 KOUI: unterordnende Konjunktion mit Infinitiv 58
2.8.2 KOUS: unterordnende Konjunktion mit Satz 58
2.8.3 KON: nebenordnende Konjunktion 59
2.8.4 KOKOM: Vergleichspartikel 60
2.9 Adpositionen 62
2.9.1 APPR: Präposition 62
2.9.2 APPRART: Präposition mit Artikel 63
2.9.3 APPO: Postposition 64
2.9.4 APZR: Zirkumposition rechts 65
2.10 Partikel 66
2.10.1 PTKZU: "zu" vor Infinitiv 66
2.10.2 PTKNEG: Negationspartikel 66
2.10.3 PTKVZ: abgetrennter Verbzusatz 67
2.10.4 PTKA: Partikel bei Adjektiv oder Adverb 69
2.10.5 PTKANT: Antwortpartikel 69
2.11 Interpunktionen 69
2.11.1 $, $(, $. 69
2.12 Sonstige 70
2.12.1 ITJ: Interjektionen 70
2.12.2 TRUNC: Kompositions–Erstglied 70
2.12.3 XY: Nichtwörter 71
2.12.4 FM: Fremdsprachliches Material 72

Kapitel 1

Einleitung

Das vorliegende Papier ist ein Anleitung für die manuelle Annotierung von deutschen Textkorpora mit STTS (Stuttgart–Tübingen Tagset).
Das STTS resultiert aus einer gegenseitigen Abstimmung zweier Part–of–Speech–Tagsets, die an der Universität Stuttgart (IMS) und an der Universität Tübingen (SfS) entwickelt wurden. Damit soll eine Übereinstimmung bei der Korpus–Annotation erreicht werden, die die gegenseitige Nutzung bereits durchgeführter Korpusarbeit ohne umständliche Anpassung unterschiedlicher Tagsets ermöglicht.
Als wichtigste Gliederungsaspekte bei der Einteilung der Wortarten wurden distributionelle Kriterien, aber auch traditionell–linguistische Kriterien (z.B. semantische und morphologische) zugrundegelegt.
In Stuttgart wurde dieses POS–Tagset noch hinsichtlich lexikalischer und morphologischer Eigenschaften von Wortformen erweitert. Bei der Spezifikation der konkreten Tagsets können je nach Anwendung nur einzelne Blöcke verwendet oder höhere Ebenen der Hierarchie ausgewählt werden.
Der augenblickliche Stand wurde nach wiederholter Diskussion am 18.08.1995 in Tübingen festgelegt.

1.1 Zuweisung von Tags

Als allgemeine Regel gilt, daß jede Wortform genau ein Tag erhält. Der Begriff Wortform umfaßt neben "echten" Wortformen auch Zahlen in Ziffern, Satzzeichen, Sonderzeichen (wie z.B. §, $), abgetrennte Wortteile oder Kompositions–Erstglieder (wie z.B. **Ein–** und Ausgang) etc. Es wird davon ausgegangen, daß für das manuelle Taggen die Texte so aufbereitet sind, daß jede Zeile genau eine Wortform enthält.

1.2 Mehrwortlexeme

Damit ist es also (aus technischen Gründen) nicht möglich, Mehrwortlexeme als Ganzes zu taggen, oder kontraktive Formen mit einer Kombination aus mehreren Tags zu versehen. Idealerweise sollten feststehende Ausdrücke wie *vor kurzem, vor allem* als Mehrwortlexeme (**multi word items**) aufgefaßt werden und von Tokenizer und Tagger so behandelt werden. Solange dies technisch noch nicht möglich ist, werden als Kompromiß die einzelnen Teile annähernd so behandelt, als wenn die Teile einzeln stehen würden:

Beispiele:

- **New/NE York/NE** nicht: **New York/NE**
- **so/ADV daß/KOUS** nicht: **so daß/KOUS**
- **zum/APPRART** nicht: **zum/APPR ART**

Bei aus 2 Teilen bestehenden Konjunktionen (*entweder – oder, weder – noch*) werden **beide** Teile als KON getaggt. In den folgenden guidelines werden Mehrwortlexeme durch das Zeichen **ml:** gekennzeichnet, was besagt, daß diese Wortform idealerweise ein gemeinsames Tag bekommen sollte (welches hinter den Zeichen **ml:** angegeben wird), als Kompromißlösung aber wie angegeben getaggt wird.

1.3 Behandlung von Abkürzungen

Es gibt kein eigenes Tag für Abkürzungen. Abgekürzte Wortformen werden generell so getaggt wie die ausgeschriebene Form. Abkürzungen für mehrere Worte, die nicht durch Leerzeichen getrennt sind, werden entsprechend ihrer syntaktischen Funktion klassifiziert.

Beispiele:

- **Herr/NN Dr./NN Maier/NE**
- *die* **gem./ADJA** *Verhandlungen*
- *mit Haus* **u./KON** *Garten*
- **z./APPRART B./NN**
- **z.B./ADV**
- **d./PDS h./VVFIN**
- **d.h./KON**
- **sondern/KON**

- **aber**/KON *es klang nicht so, als ob...*
- **USA**/NE
- **LB**/NE
- **AG**/NN
- *die* **Zwei**/NN
- *die Zahl* **Zwei**/NN

1.4 Behandlung von Fehlern im Text

- Schreibfehlertolerantes Vorgehen: Wenn der Sinn erkennbar ist, wird die WF verbessert, und es wird so getaggt, wie die richtige Wortform ausgesehen hätte:
  - Hautür ⇒ **Haustür**/NN
  - neuhlich ⇒ **neulich**/ADV
- Auch syntaxverfälschende Fehler sollen so behandelt werden:
  - Er hat im das gesagt ⇒ **ihm**/PPERS
  - Sie hat das Haus, daß sie gestern sah, gekauft ⇒ **das**/PRELS
- Dokumentation all dieser Veränderungen in einem Administrationsfile.
- Vollkommen unverständliche Sätze, fehlende Satzteile, doppelte Satzteile: Wenn die Struktur des Satzes nicht mehr zu erkennen ist, wird der ganze Artikel nicht mehr verwendet.

1. Nomina (N)	7. Adverbien (ADV)
2. Verben (V)	8. Konjunktionen (KO)
3. Artikel (ART)	9. Adpositionen (AP)
4. Adjektive (ADJ)	10. Interjektionen (ITJ)
5. Pronomina (P)	11. Partikeln (PTK)
6. Kardinalzahlen (CARD)	

Tabelle 1.1: Die Hauptwortarten und ihre *tags*

1.5 STTS – Übersicht

Das Tagset ist hierarchisch strukturiert. Die aus unseren Überlegungen resultierenden Hauptwortarten und ihre Unterwortarten spiegeln sich in den *tags* wider. Die *tags* bestehen aus möglichst selbsterklärenden Buchstabensequenzen, die von links nach rechts gelesen zuerst die Hauptwortart und dann die Unterwortart kodieren, also von der allgemeinen Information zur spezifischeren hinführen.[1]
Damit wird eine gewisse Flexibilität erreicht, die dem Benutzer erlaubt, je nach Anspruch, nur auf die Hauptwortarten oder auf wortartenspezifische Informationen zuzugreifen.
Das Tagset umfaßt 11 Hauptwortarten (Tabelle 1.1), die weitgehend nach allgemein anerkannter linguistischer Terminologie in den *tags* kodiert sind. Sie orientieren sich am "TEI Starter Set Of Grammatical–Annotation Tags"[2] mit Ausnahme der Kardinalzahlen, die durch den Wert *cardinal* beim Merkmal *numeral* der Adjektive abgedeckt werden und der Konjunktionen, die dort von den 2 Kategorien *subordinators* und *coordinators* repräsentiert werden.

Diese Hauptwortarten sind unterschiedlich stark subklassifiziert. So werden z.B. die Pronomina in weitere 8 Untergruppen unterschieden, wobei die Untergruppen wieder unterteilt sein können, je nachdem ob sie NP–ersetzende (substituierend, *tag*: S), nomenbegleitende (attribuierend, *tag*: AT) oder adverbiale (*tag*: AV) Funktion innehaben.[3]

Insgesamt enthält STTS 54 tags, von denen 48 reine POS–tags sind und 6 zusätzliche tags für fremdsprachliches Material (FM), Kompositions–Erstglieder (TRUNC), Nichtwörter (XY) und Satzzeichen ($,, $., $() verwendet werden. In Tabelle 2 werden alle Tags kurz beschrieben.

[1] Dabei sollte die Buchstabensequenz möglichst kurz sein, damit die Leserlichkeit eines so getaggten Korpus nicht zu sehr beeinträchtigt wird.
[2] beschrieben in [TEI 91]
[3] siehe Abbildung 2.1, Seite 35.

POS =	Beschreibung	Beispiele
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	Vergleichspartikel, ohne Satz	*als, wie*
NN	normales Nomen	*Tisch, Herr, [das] Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativ-pronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativ-pronomen	*jener [Mensch]*
PIS	substituierendes Indefinit-pronomen	*keiner, viele, man, niemand*
PIAT	attribuierendes Indefinit-pronomen ohne Determiner	*kein [Mensch],* *irgendein [Glas]*
PIDAT	attribuierendes Indefinit-pronomen mit Determiner	*[ein] wenig [Wasser],* *[die] beiden [Brüder]*
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessiv-pronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein [Buch], deine [Mutter]*
	Relativpronomen	
PRELS	substituierend	*[der Hund,] der*
PRELAT	attribuierend	*[der Mann ,] dessen [Hund]*

POS =	Beschreibung	Beispiele
	Relativpronomen	
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	*[er hat] gekonnt*
XY	Nichtwort, Sonderzeichen enthaltend	*D2XW3*
$,	Komma	,
$.	Satzbeendende Interpunktion	. ? ! ; :
$(	sonstige Satzzeichen; satzintern	- [](}

Kapitel 2

Beschreibung der einzelnen Tags

2.1 Nomina

Bei den Nomina wird zwischen "normalen" Nomina und Eigennamen unterschieden.
Beispiele:

- *der Vater/NN von Klaus/NE*
- *die Schiffahrt/NN auf Rhein/NE und Mosel/NE*

2.1.1 NN: normale Nomina

Klassifikation von NN	POS

POS =	Beschreibung	Beispiele
NN	konkrete und abstrakte Substantive	*Haus, Anwendung*
	Maßangaben	*Liter, Meter, Kilo*
	Titel oder Anreden	*Herr, Professor, Graf, Bundeskanzler*
	Produkte	*[ein] Porsche, [ein] Dinkelacker*
	Herkunftsbezeichnung	*[ein] Frankfurter, [ein] Schweizer*
	substantiviertes Adjektiv	*[der] Blinde, [das] Junge*
	substantivierte Partizipien	*[das] Gewünschte, [der] Schlafende*
	substantivierte Infinitive	*[das] Reisen, [des] Laufens [müde]*
	Determinativkomposita (NE+NN)	*[in der] Mozartstraße, Bachkantate, Gretchenfrage*
	Monate	*[im] Juli*
	Wochentage	*[am] Sonntag*
	Sprachen	*[er spricht] Esperanto/Englisch*
Aber:		
ADJA	adjektivischer Gebrauch von Herkunftsbezeichnungen	**Schweizer**/ADJA *Käse,* *der* **Frankfurter**/ADJA *Flughafen*
NE	Firmennamen	*der Vorstand von* **Porsche**/NE

Beispiele:

- *das* **Haus**/NN *von* **Herrn**/NN **Dr.**/NN *Maier*
- *der* **Arme**/NN *ging 10* **km**/NN *weit*
- *ein* **Viertel**/NN *der* **Angestellten**/NN *liebt das* **Reisen**/NN
- *er wurde* **Dritter**/NN
- *der* **Alte**/NN
- *der* **Abgeordnete**/NN
- *ein* **Liebender**/NN
- *das* **Spielen**/NN
- *ich lerne* **Deutsch**/NN
- **Hunderte**/NN
- *ein* **Vierter**/NN
- *ein* **Viertel**/NN
- **Abk.**/NN
- **AG**/NN
- *der* **Spieler**/NN
- *die* **Anwendung**/NN
- *ich trage meistens eine* **Jeans**/NN
- *ich entspanne mich durch* **Yoga**/NN
- aber: *ich entspanne mich durch* **Tai**/FM **Chi**/FM
- *ich trinke gerne* **Kerner**/NN *und* **Trollinger**/NN
- aber: *ich trinke gerne* **Kerner**/ADJA *und* **Trollinger**/ADJA *Wein*
- *der Film "*Ein/ART **Fisch**/NN **namens**/APPR **Wanda**/NE*"*
- *ich gehe ins Gasthaus "*Ewige/ADJA **Lampe**/NN*"*
- aber: *ich gehe ins Gasthaus* **Lampe**/NE
- **Deutsch**/NN *ist leichter als* **Russisch**/NN

Kriterien zur Abgrenzung NN/NE:

- Komplexe Namen: jedes Teil wird getaggt wie im prototypischen Kontext.
- Einzelwortformen: semantisches Kriterium. Namenssemantik.
- Indefiniter Artikel kann verwendet werden → Anzeichen für NN.

Kriterien zur Abgrenzung NN/FM:[1]

- Deutsche Flexion → NN
- Großgeschrieben, wenn das entsprechende Wort in Originalsprache kleingeschrieben wurde → NN, z.B. die **Contras**/NN

[1] siehe dazu auch Abschnitt 2.12.4

Lexikalische Kategorien für NN LEX

LEX =	Erläuterung	Beispiele
ABK	Abkürzung	*Abk.*/NN<**ABK**, *AG*/NN<**ABK**
ADJ	substantivierte Adjektive	*der Alte*/NN<**ADJ**
CARD	Kardinalzahl	*Hunderte*/NN<**CARD** *von Tonnen*
FRAKT	Bruchzahl	*ein Viertel*/NN<**FRAKT**
ORD	Ordinalzahl	*ein Vierter*/NN<**ORD**
VINF	substantivierter Infinitiv	*das Spielen*/NN<**VINF**
VPART	substantivierte Form des Partizip Perfekts	*der Abgeordnete*/NN<**VPART**
	substantivierte Form des Partizip Präsens	*ein Liebender*/NN<**VPART**
Aber:		
–	Derivationsformen	*der Spieler*/NN, *die Anwendung*/NN

Morphologische Merkmale von NN MOR

Attribut	MOR =	Beispiele
Genus	**Masc**	*der Tisch*/NN:**Masc**.Nom.Sg.*
	Fem	*den Städten*/NN:**Fem**.Dat.Pl.*
	Neut	*das Reisen*/NN<VINF:**Neut**.Nom.Sg.*
	*	*die Kosten*/NN:*.Nom.Pl.* , *den Deutschen*/NN<ADJ:*.Dat.Pl.Sw
Kasus	**Nom**	*ein Tisch*/NN:Masc.**Nom**.Sg.*
	Gen	*der Frau*/NN:Fem.**Gen**.Sg.*
	Dat	*der Frau*/NN:Fem.**Dat**.Sg.*
	Akk	*den Grünen*/NN<ADJ:Masc.**Akk**.Sg.Sw
Numerus	**Sg**	*das Haus*/NN:Neut.Nom.**Sg**.*
	Pl	*die Häuser*/NN:Neut.Nom.**Pl**.*
Flexion	**Sw**	*der Beamte*/NN:Masc.Nom.Sg.**Sw**
	St	*ich Armer*/NN<ADJ:Masc.Nom.Sg.**St**
	Mix	*eine Rote*/NN<ADJ:Fem.Nom.Sg.**Mix**
	*	*ein Haus*/NN:Neut.Akk.Sg.*

Beachte:

Genus: Für Genus ist der Wert * zugelassen:

- bei Substantiven, die nur im Plural vorkommen (z.B. *die Kosten*) und
- bei nominalisierten Adjektiven im Plural (z.B. *die Alten*).

Kasus:
- Der Kasus wird bei Nomina immer angegeben.
- Bei engen Appositionen, wie z.B. *im Hotel Beckmann* wird grundsätzlich nur das Kopfnomen flektiert, das Appositiv trägt die Nominativmorphologie.

Beispiele:

- *im Hotel*/NN:Neut.Dat.Sg.* *Beckmann*/NE:*.Nom.Sg
- *Peter*/NE:Masc.Nom.Sg *Müllers*/NE:*.Gen.Sg *Haus*/NN:Neut.Nom.Sg.*

Numerus: Der Numerus muß immer angegeben werden.

Flexion: Die Flexion wird in folgenden Fällen angegeben:

- bei substantivierten Adjektiven,
- Partizipien,
- Ordinalzahlen und
- bei adjektivisch deklinierten Nomina wie z.B. *[der] Beamte*.

In allen anderen Fällen wird ein * gesetzt.

Beispiele:

- *das Haus*/NN:Neut.Nom.Sg.* *auf dem Lande*/NN:Neut.Dat.Sg.*
- *mit Herrn*/NN:Masc.Dat.Sg.* *Dr.*/NN<ABK:Masc.Nom.Sg.* *Maier*/NE:*.Dat.Sg
- *der Arme*/NN<ADJ:Masc.Nom.Sg.Sw *ging 10 km*/NN<ABK:Masc.Akk.Pl.* *weit*
- *ein Viertel*/NN<FRACT:Neut.Nom.Sg.* *der Angestellten*/NN<VPART:*.Gen.Pl.Sw *liebt das Reisen*/NN<VINF:Neut.Akk.Sg.*
- *er wurde Dritter*/NN<ORD:Masc.Nom.Sg.St

2.1.2 NE: Eigennamen

Klassifikation von NE POS

POS =	Beschreibung	Beispiele
NE	Vornamen	*Hans, Uli*
	Familiennamen	*Maier, Krafft*
	Tiernamen	*Fifi, Hansi, Betzi*
	Firmennamen	*Mercedes, LB*
	Ortsnamen	*Stuttgart, Moskau, Heslach*
	Ländernamen und Gebietsnamen	*England, Schweiz, USA, Baden-Württemberg, Pfalz*
	Gewässernamen	*Rhein, Bodensee, Pazifik*
	Bergnamen	*Zugspitze, Lemberg*
	Gebirgsnamen	*Alpen, Alb, Hunsrück*
	Planetennamen	*Venus, Mars, Jupiter*
	Namen von Stadtvierteln	*Ostend, Stuttgart-West*
	fremdspr. Namensteile	*Vincent van Gogh, New York*
Aber:		
NN	Produktnamen	*ein* **Mercedes**/NN, *eine* **Cola**/NN
NN	aus NN abgeleitete Eigennamen	*die* **Grünen**/NN
NN	Determinativkomposita (NE+NN)	**Mozartstaße**/NN, **Bachkantate**/NN, **Gretchenfrage**/NN
NN	Monate, Wochentage	**Januar**/NN, **Montag**/NN
NN	Stadtviertel nach Richtungen	*Im Stuttgarter* **Westen**/NN

Generelle Regel für komplexe Namen:

- deutsch: Teile werden entsprechend ihrer Distribution getaggt (z.B. **Freie**/ADJA **Universität**/NN **Berlin**/NE)
- fremdsprachliche Teile werden als Eigennamen getaggt (z.B. **New**/NE **York**/NE)

Beispiele:

- **Wernher**/NE **von**/APPR **Braun**/NE
- **Weil**/NE **am**/APPRART **Rhein**/NE
- *die* **Freie**/ADJA **Universität**/NN **Berlin**/NE
- *die* **Bundesrepublik**/NN **Deutschland**/NE
- *die* **Deutsche**/ADJA **Angestellen-Gewerkschaft**/NN
- *die* **DAG**/NE
- *der* **VfB**/NE *spielt gegen den* **HSV**/NE
- *ich gehe ins "*Holiday/NE Inn/NE*"*

- **Frankf./NE**
- *die Strecke* **Hamburg–Berlin/NE**
- *ich habe in* **Berlin–Ost/NE** *gewohnt*
- *die* **Treuhand/NE**
- **amnesty/NE international/NE**

Lexikalische Kategorien für NE LEX

LEX =	Erläuterung	Beispiele
ABK	Abkürzungen	*Frankf.*/NE<**ABK**, *DAG*/NE<**ABK**

Morphologische Merkmale von NE MOR

Attribut	MOR =	Beispiele
Genus	**Masc**	*der HSV*/NE<ABK:**Masc**.Nom.Sg
	Fem	*Maria*/NE:**Fem**.Akk.Sg
	Neut	*Englands*/NE:**Neut**.Gen.Sg
	*	*Familie Maier*/NE:*.Dat.Sg, *Uli*/NE:*.Nom.Sg
Kasus	**Nom**	*Hans*/NE:Masc.**Nom**.Sg *geht*
	Gen	*Frau Maiers*/NN:*.**Gen**.* *Hut*
	Dat	*an der Donau*/NE:Fem.**Dat**.Sg
	Akk	*in die USA*/NE<ABK:*.**Akk**.Pl
	*	*van*/NE:*.*.* *Gogh*
Numerus	**Sg**	*Paris*/NE:Neut.Nom.**Sg**
	Pl	*Maiers*/NE:*.Nom.**Pl** *kommen*
	*	*New*/NE:*.*.* *York*

Beachte:

Genus: Der Genus ist immer undefiniert bei:

- geschlechtsneutralen Vornamen (z.B. *Uli*) und
- Familiennamen (z.B. *Müller*).

Kasus und Numerus: Müssen immer angegeben werden.

Sonstiges: Bei fremdsprachigen Namensteilen bleiben Genus, Kasus und Numerus undefiniert.

Beispiele:

- *Wernher*/NE:Masc.Nom.Sg *von*/APPR *Braun*/NE:*.Nom.Sg
- *Weil*/NE:Neut.Nom.Sg *am*/APPRART:Masc.Dat.Sg *Rhein*/NE:Masc.Dat.Sg
- *die Freie*/ADJA:Pos.Fem.Nom.Sg.Sw *Universität*/NN:Fem.Nom.Sg.* *Berlin*/NE:Neut.Nom.Sg
- *die Bundesrepublik*/NN:Fem.Nom.Sg.* *Deutschland*/NE:Neut.Nom.Sg
- *die Deutsche*/ADJA:Pos.Fem.Nom.Sg.Sw *Angestellen-Gewerkschaft*/NN:Fem.Nom.Sg.*
- *die DAG*/NE<ABK:Fem.Nom.Sg
- *der VfB*/NE<ABK:Masc.Nom.Sg *spielt gegen den HSV*/NE<ABK:Masc.Akk.Sg
- *Vincent*/NE:Masc.Nom.Sg *van*/NE:*.*.* *Gogh*/NE:*.Nom.Sg
- *New*/NE:*.*.* *Yorks*/NE:Neut.Gen.Sg *Bürgermeister*/NN:Masc.Nom.Sg.*

2.2 Adjektive

Bei den Adjektiven wird zwischen attributivem Gebrauch und nicht–attributivem Gebrauch unterschieden. Zur Klasse **ADJA** zählen alle flektierten Adjektive, sowie nicht–flektierte Formen, die vor einem Nomen stehen, auch vor einem "leeren" Nomen (Ellipsen). Mit **ADJD** werden prädikativ und adverbial (auch wenn andere Adjektive modifiziert werden) gebrauchte, sowie nachgestellte, nicht flektierte Adjektive bezeichnet.

Beispiele:

- *das* **rote**/ADJA *Kleid*
- *das* **lila**/ADJA *Kleid*
- *vor* **kurzem**/ADJA *(ml:ADV)*
- *seit* **langem**/ADJA *(ml:ADV)*
- *im* **übrigen**/ADJA *(ml:ADV)*
- **aber:** *vor* **allem**/PIS *(ml:ADV)*
- **aber:** *unter* **anderem**/PIS *(ml:ADV)*
- *in* **ganz**/ADJA *Deutschland*
- *ein* **freundlich**/ADJA *Wort*
- *ein* **lustig**/ADJA *Liedchen*
- *das Auto ist* **schnell**/ADJD
- *das Auto fährt* **schnell**/ADJD
- *ein* **schnell**/ADJD **fahrendes**/ADJA *Auto*
- *Hänschen* **klein**/ADJD
- *die* **50er**/ADJA *Jahre*
- *das* **320–seitige**/ADJA *Werk*
- *die Verfolgung* **politisch**/ADJD **Andersdenkender**/NN

Folgende Wortformen gehören zu den Adjektiven:

- mannigfach -> nur ein Vorkommen in 100 Mio WF taz, adverbial. In duden grammatik kein Hinweis: ADJA?
- mehrfach, vielfach,
- vielfältig

2.2.1 ADJA: attributive Adjektive

Klassifikation von ADJA POS

POS =	Beschreibung	Beispiele
ADJA	"echte" Adjektive	*[die] große [Stadt],*
	(Positiv)	*[das] lila [Kleid]*
	(Komparativ, Superlativ)	*[das] kleinere/kleinste [Übel]*
	attributiv gebrauchtes Partizip Perfekt	*[der] gesuchte [Dieb]*
	attributiv gebrauchtes Partizip Präsens	*[das] lachende [Kind]*
	attributiver Gebrauch von Herkunftsbezeichnungen und Orte in Straßennamen	*Schweizer [Käse], [der] Frankfurter [Flughafen] Rottweiler [Straße]*
	Ordinalzahlen	*[die] zweite [Besetzung]*
	Multiplikativzahlen	*[der] zweifache/zweimalige [Sieger]*
	Bruchzahlen	*[ein] dreiviertel [Liter Milch]*
Aber:		
NN	substantivisch gebrauchte Adjektive oder Partizipien[2]	*ein* **Großer**/NN, *der* **Gesuchte**/NN
CARD	Kardinalzahlen	*die* **drei**/CARD *Männer*
PIDAT	Indefinitpronomen "all–", "beid–", "viel–", "wenig–"	*die* **vielen**/PIDAT *Leute,* **alle**/PIDAT *diese Leute* *ein* **wenig**/PIDAT *Wasser*
PIAT	Indefinitpronomen "viel"	**viel**/PIAT *Gutes,* **viel**/PIAT *Wasser*

Beispiele:

- *der* **große**/ADJA *und der* **kleine**/ADJA *Klaus*
- *mit einem* **lachenden**/ADJA *und einem* **weinenden**/ADJA *Auge*
- *das* **vermißte**/ADJA *Kind*
- *das* **schnellere**/ADJA *Auto*
- *der* **vordere**/ADJA *Wagen*
- *die* **Schweizer**/ADJA *Schokolade in* **lila**/ADJA *Verpackung*
- *den* **ganzen**/ADJA *Tag*
- *in* **ganz**/ADJA *Deutschland*[3]

[2] Wenn das Adjektiv klein geschrieben (d.h. das zugehörige Nomen ausgelassen) ist, bleibt die Klassifizierung als ADJA.

[3] Die Stellung und nicht die Flexion entscheidet hier!

- **aber:** *die Vase ist* **ganz**/ADJD
- *sie werden als* **letzte**/ADJA *geheuert*
- *der* **größte**/ADJA *Zwerg*
- **aber:** *er ist der* **Größte**/NN
- *die beiden* **ersten**/ADJA *Sieger*
- *der* **27.**/ADJA *Februar*
- *der* **dreimalige**/ADJA *Sieger*
- *der* **vielfache**/ADJA *Weltmeister*
- *das* **vielfältige**/ADJA *Angebot*
- *der* **dritte**/ADJA *Sieger*
- **aber:** *die* **drei**/CARD *Sieger*
- *ein* **halbes**/ADJA *Pfund*
- *ein* $\frac{3}{4}$/ADJA *Liter Milch*
- *die* **zahlreichen**/ADJA *Besucher*
- **aber:** *die* **vielen**/PIDAT *Besucher*
- *vor* **kurzem**/ADJA *(ml:ADV)*[4] *war er da*
- **aber:** *vor* **allem**/PIS*(ml:ADV)*
- *der* **gefeierte**/ADJA *Star*
- *das* **sinkende**/ADJA *Schiff*
- *die* **anzuwendende**/ADJA *Regel*
- *die* **gem.**/ADJA *Wohnung*
- *die* **anwendbare**/ADJA *Regel*
- *die* **50er**/ADJA *Jahre*

Lexikalische Kategorien für ADJA LEX

LEX =	Erläuterung	Beispiele
ABK	Abkürzungen	*die gem.*/ADJA<**ABK** *Wohnung*
FRAKT	Bruchzahlen	*ein halbes*/ADJA<**FRAKT** *Pfund*
ORD	Ordinalzahl	*der dritte*/ADJA<**ORD** *Mann*
VPART	Partizip Perfekt	*der gefeierte*/ADJA<**VPART** *Star*
VPART	Partizip Präsens	*das sinkende*/ADJA<**VPART** *Schiff,* *die anzuwendende*/ADJA<**VPART** *Regel*
Aber:		
–	Derivationsformen	*die anwendbare*/**ADJA** *Regel*

Morphologische Merkmale von ADJA MOR

Attribut	MOR =	Beispiele
Grad	**Pos**	*das kleine*/ADJA:**Pos**.Neut.Nom.Sg.Sw *Haus*
	Comp	*das kleinere*/ADJA:**Comp**.Neut.Nom.Sg.Sw *Haus*
	Sup	*das kleinste*/ADJA:**Sup**.Neut.Nom.Sg.Sw *Haus*
	*	*das dritte*/ADJA<ORD:*.Neut.Nom.Sg.Sw *Haus*
Genus	**Masc**	*ein schneller*/ADJA:Pos.**Masc**.Nom.Sg.Mix *Wagen*
	Fem	*eine schnelle*/ADJA:Pos.**Fem**.Nom.Sg.Mix *Fahrt*
	Neut	*ein schnelles*/ADJA:Pos.**Neut**.Nom.Sg.Mix *Auto*
	*	*die schnellen*/ADJA:Pos.*.Nom.Pl.Sw *Autos*, *ein lila*/ADJA:Pos.*.*.*.* *Kleid*
Kasus	**Nom**	*der rote*/ADJA:Pos.Masc.**Nom**.Sg.Sw *Hut*
	Gen	*des roten*/ADJA:Pos.Masc.**Gen**.Sg.Sw *Hutes*
	Dat	*mit rotem*/ADJA:Pos.Masc.**Dat**.Sg.St *Hut*
	Akk	*ohne roten*/ADJA:Pos.Masc.**Akk**.Sg.St *Hut*
	*	*im lila*/ADJA:Pos.*.*.*.* *Kleid*
Numerus	**Sg**	*eine halbe*/ADJA:Pos.Fem.Nom.**Sg**.Mix *Sache*
	Pl	*keine halben*/ADJA:Pos.*.Nom.**Pl**.Mix *Sachen*
	*	*die Schweizer*/ADJA:*.*.*.*.* *Banken*
	*	*ein viertel*/ADJA<FRACT:*.*.*.*.* *Pfund*
Flexion	**St**	*mit ganzem*/ADJA:Pos.Masc.Dat.Sg.**St** *Einsatz*
	Sw	*mit dem ganzen*/ADJA:Pos.Masc.Dat.Sg.**Sw** *Hausrat*
	Mix	*mit einem ganzen*/ADJA:Pos.Masc.Dat.Sg.**Mix** *Apfel*
	*	*in ganz*/ADJA:Pos.*.*.*.* *Europa*

[4] Mehrwortlexem, s. Abschnitt 1.2.

Beachte:

Grad: Der Steigerungsgrad ist nicht definiert für:

- Ordinalzahlen (z.B. *das zweite*),
- Bruchzahlen (z.B. *ein viertel*) und
- Herkunftsbezeichnungen (z.B. *Frankfurter Würstchen*).

Ansonsten muß er immer angegeben werden. Er richtet sich nach der Form, nicht nach der Semantik des Adjektivs.

Beispiele:

- *der bestmögliche*/ADJA:**Pos**.Masc.Nom.Sg.Sw *Weg*
- *ein optimaler*/ADJA:**Pos**.Masc.Nom.Sg.Mix *Ansatz*

Genus: Der Genus bleibt bei Adjektiven im Plural immer undefiniert.

Kasus und Numerus: Müssen immer angegeben werden.

Flexion: Läßt sich aus dem vorausgehenden Determiner ableiten:

- **schwache Flexion:**
  - nach bestimmtem Artikel *der, die, das* (auch nach Präposition mit inkorporiertem Artikel wie *im, zur, etc.*),
  - nach Demonstrativpronomen *dies-, jen-, derselb-, derjenig-*
  - nach *jed-, jeglich-, jedwed-, all-, beid-, sämtlich-*
  - nach *manch-, solch-, welch-*
- **starke Flexion**
  - ohne Artikel
  - nach *manch, solch, welch, viel, wenig, etwas, mehr*
- **gemischte Flexion**
  - nach unbestimmtem Artikel *ein-*
  - nach *kein-*
  - nach Possessivpronomen *mein-, dein-, sein-, ...*

Sonstiges:
- Bei nichtflektierenden Adjektiven wie *lila, rosa, ganz* wird nur der Steigerungsgrad **Pos** angegeben. Alle anderen Attribute bleiben undefiniert.
- Bei Herkunftsbezeichnungen (z.B. *Schweizer Schokolade*) bleiben alle Attribute undefiniert.

Beispiele:

- *der große*/ADJA:Pos.Masc.Nom.Sg.Sw *und der kleine*/ADJA:Pos.Masc.Nom.Sg.Sw *Klaus*

- *mit einem lachenden*/ADJA<VPART:Pos.Neut.Dat.Sg.Mix *und einem weinenden*/ADJA<VPART:Pos.Neut.Dat.Sg.Mix *Auge*
- *das schnellere*/ADJA:Comp.Neut.Nom.Sg.Sw *Auto*
- *der vordere*/ADJA:Pos.Masc.Nom.Sg.Sw *Wagen*
- *die Schweizer*/ADJA:*.*.*.*.* *Schokolade in lila*/ADJA:Pos.*.*.*.* *Verpackung*
- *den ganzen*/ADJA:Pos.Masc.Akk.Sg.Sw *Tag*
- *in ganz*/ADJA:Pos.*.*.*.* *Deutschland*
- *sie werden als letzte*/ADJA:Pos.*.Nom.Pl.St *geheuert*
- *er ist der Größte*/NN<ADJ:Masc.Nom.Sg.Sw
- *der 27.*/ADJA<ORD:*.Masc.Nom.Sg.Sw *Februar*
- *der siebenundzwanzigste*/ADJA<ORD:*.Masc.Nom.Sg.Sw *Platz*
- *vor kurzem*/ADJA:Pos.Neut.Dat.Sg.St
- *ein viertel*/ADJA<FRACT:*.*.*.*.* *Pfund*

2.2.2 ADJD: prädikativ oder adverbial gebrauchte Adjektive

Klassifikation von ADJD POS

POS =	Beschreibung	Beispiele
ADJD	"echte" Adjektive	
	(Positiv)	*[sie ist] groß, [es ist] lila*
	(Komparativ, Superlativ)	*[er läuft] schneller/am schnellsten*
	usprüngliche Nomina	*[es ist] recht*
	adverbial gebrauchtes	
	Partizip Präsens	*[er kam] lachend [herein]*
	adverbial gebrauchtes	*gekonnt [gespielt]*
	Partizip Perfekt	*[schneller als] geplant*
	Ordinalzahlen	*[sie sind zu] zweit*
Aber:		
ADV	nur adverbial	*er kommt* **nämlich**/ADV *morgen*
	gebrauchte Form	
CARD	Kardinalzahlen	*sie waren* **zwei**/CARD
VVPP	nicht flektiertes	*er wird* **gesucht**/VVPP,
	Partizip Perfekt	*es ist* **geplant**/VVPP
PTKVZ	adjektivische abgetrennte	*[er hält]* **geheim**/PTKVZ
	Verbzusätze	

Beispiele:

- *er liegt* **krank**/ADJD *im Bett*
- *er kam* **völlig**/ADJD **durchnäßt**/ADJD *an*
- *er kommt wie* **geplant**/ADJD
- aber: *er kommt, wie er es* **geplant**/VVPP *hat*
- *er ist* **länger**/ADJD *als* **breit**/ADJD
- *er ist am* **schnellsten**/ADJD
- *er ist* **schuld**/ADJD
- *mir ist* **angst**/ADJD
- *er kommt zu* **spät**/ADJD
- *sie kamen* **zu**/PTKA **dritt**/ADJD
- *mir ist* **angst**/ADJD *und* **bange**/ADJD
- **gebraucht**/ADJD *kaufen*
- **rasend**/ADJD *werden*
- *zu* **dritt**/ADJD
- **halb**/ADJD *voll*
- *Die Regel ist* **anwendbar**/ADJD
- *eine* **zugegeben**/ADJD **frei**/ADJD **erfundene**/ADJA *Geschichte*

2.2.3 ADJD oder VVPP?

- Partizipien in adverbialer Stellung: ADJD.

  Beispiele:

  - *er spielt* **gekonnt**/ADJD
  - *er kommt* **geflogen**/ADJD
  - *die Mittel wurden* **gezielt**/ADJD **eingesetzt**/VVPP

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

  Beispiele:

  - *er macht es wie* **geplant**/ADJD
  - *sie lügt wie* **gedruckt**/ADJD
  - *die* **geplante**/ADJA *Sache*
  - *das* **gewollt**/ADJD **verlorene**/ADJA *Spiel*

- 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).

  Beispiele:

  - *der Tisch wird* **verrückt**/VVPP
  - aber: *der alte Mann wird* **verrückt**/ADJD

  Kriterien für Disambiguierung Kopulakonstruktionen mit ADJD vs. Verlaufspassiv mit VVPP:

  - Verdacht auf VVPP: kann der Satz ins Aktiv gesetzt werden mit gleicher Semantik? Ja → VVPP
  - von–PP oder ähnliche PP, die auf Verbsemantik hinweist → VVPP
  - Ersetzung durch semantisch nahes Adjektiv möglich → ADJD

  Beispiele:

  - *wo Menschen selbst* **betroffen**/VVPP *seien oder sich* **betroffen**/ADJD *fühlten.*

  Meist muß der weitere Kontext (satzübergreifend) herangezogen werden, um Partizipien, wie in dem folgenden Beispiel zu desambiguieren:

  - er hat die Haare kurz **geschnitten**/ADJD [er = der Friseur]
    vs. er hat die Haare kurz **geschnitten**/VVPP [= er hat kurze Haare]

In Abhängigkeit von Perfekt-*haben* werden dieselben Formen eindeutig als Partizip identifiziert:

- er ist **verrückt**/ADJD [= irre]
  vs. er hat den Schrank **verrückt**/VVPP
- sie ist **geladen**/ADJD [= zornig]
  vs. sie hat ihr Gewehr **geladen**/VVPP
  vs. sie ist zum Fest **geladen**/VVPP
- er ist sehr **bewegt**/ADJD [= gerührt]
  vs. er hat den Kopf **bewegt**/VVPP
- sie ist ziemlich **geschafft**/ADJD [= müde]
  vs. sie hat es endlich **geschafft**/VVPP
- er ist **gelehrt**/ADJD
  vs. er hat ihn Astrologie **gelehrt**/VVPP
- eine Frage ist **angebracht**/ADJD
  vs. eine Frage wird von Peter **angebracht**/VVPP
- er ist in der Stadt hoch **angesehen**/ADJD
  vs. er wird von Peter hoch **angesehen**/VVPP

Liste dieser lexikalisierten Partizipien (die dann je nach Kontext als ADJD getaggt werden):

abgebrüht/ADJD	**abgedreht**/ADJD	**abgeklärt**/ADJD
abgerissen/ADJD	**abgeschieden**/ADJD	**abgespannt**/ADJD
angegriffen/ADJD	**angemessen**/ADJD	**angeschlagen**/ADJD
angeschmiert/ADJD	**angespannt**/ADJD	**aufgeblasen**/ADJD
aufgedreht/ADJD	**aufgekratzt**/ADJD	**aufgelöst**/ADJD
aufgeräumt/ADJD	**aufgeschlossen**/ADJD	**aufgeschmissen**/ADJD
ausgefallen/ADJD	**ausgekocht**/ADJD	**ausgelassen**/ADJD
ausgeschlossen/ADJD	**ausgewogen**/ADJD	**begabt**/ADJD
begehrt/ADJD	**begeistert**/ADJD	**bekannt**/ADJD
beherrscht/ADJD	**beliebt**/ADJD	**benommen**/ADJD
betroffen/ADJD	**bewährt**/ADJD	**eingebildet**/ADJD
eingeschnappt/ADJD	**erschlagen**/ADJD	**gefaßt**/ADJD
gefragt/ADJD	**gehemmt**/ADJD	**geknickt**/ADJD
gekonnt/ADJD	**geladen**/ADJD	**gelassen**/ADJD
gelöst/ADJD	**geplättet**/ADJD	**gerädert**/ADJD
gerecht/ADJD	**gereizt**/ADJD	**gerissen**/ADJD
geritzt/ADJD	**geschickt**/ADJD	**geschwollen**/ADJD
gesetzt/ADJD	**gespannt**/ADJD	**getragen**/ADJD
gewagt/ADJD	**gewandt**/ADJD	**gewollt**/ADJD
hingerissen/ADJD	**niedergeschlagen**/ADJD	**verbissen**/ADJD
überwältigt/ADJD	**überzeugt**/ADJD	**verkehrt**/ADJD
verkannt/ADJD	**verloren**/ADJD	**vermessen**/ADJD
verschlagen/ADJD	**verschwiegen**/ADJD	

TO BE CONTINUED

Bemerkung: zu manchen dieser Partizipien exisitiert das entsprechende Verb nicht mehr, Beispiel *beliebt*. In diesem Fall ist nur noch die ADJD–Lesart zugelassen.

- Partizipien in festen Wendungen → ADJD:
  - von jdm/etw **angetan**/ADJD sein
  - jdm/einer Sache **zugetan**/ADJD sein
  - von sich **eingenommen**/ADJD sein
  - vor Schreck **gebannt**/ADJD sein
  - jdm für etw **verbunden**/ADJD sein
  - um etw **verdient**/ADJD sein
  - auf etw **versessen**/ADJD sein
  - mit jdm/etw **verwandt**/ADJD sein
  - gut/schlecht **aufgelegt**/ADJD sein
  - geistig/körperlich **zurückgeblieben**/ADJD

Lexikalische Kategorien für ADJD LEX

LEX =	Erläuterung	Beispiele
FRACT	Bruchzahl	*halb*/ADJD<**FRACT** *voll*
NN	Ursprüngliche Nomina	*mir ist angst*/ADJD<**NN**
ORD	Ordinalzahl	*zu dritt*/ADJD<**ORD**
VPART	Partizip Perfekt	*gebraucht*/ADJD<**VPART** *kaufen*
VPART	Partizip Präsens	*rasend*/ADJD<**VPART** *werden*
Aber:		
–	Derivationsformen	*Die Regel ist anwendbar*/ADJD

Morphologische Merkmale von ADJD MOR

Attribut	MOR =	Beispiele
Grad	**Pos**	*er fährt zu schnell*/ADJD:**Pos**
	Comp	*er ist schneller*/ADJD:**Comp** *als du*
	Sup	*er springt am höchsten*/ADJD:**Sup**
	*	*sie sind zu*/PTKA *zweit*/ADJD<ORD:*

Beachte:

Grad: Der Steigerungsgrad ist nicht definiert bei:

- Ordinalzahlen,
- Bruchzahlen und
- Adjektiven, die durch Konversion aus Nomina gebildet sind (z.B. *angst*)

Beispiele:

- *er liegt krank*/ADJD:Pos *im Bett*
- *er kam völlig*/ADJD:Pos *durchnäßt*/ADJD<VPART:Pos *an*
- *er ist länger*/ADJD:Comp *als breit*/ADJD:Pos
- *er ist am schnellsten*/ADJD:Sup
- *er ist schuld*/ADJD<NN:*
- *er kommt zu spät*/ADJD:Pos
- *sie kommen zu dritt*/ADJD<ORD:*
- *das Glas ist dreiviertel*/ADJD<FRACT:* *voll*

2.3 Zahlen

Nur für Kardinalzahlen wird unter den Numeralia eine eigene Wortklasse definiert. Ordinal–, Multiplikativ– und Fraktalzahlen werden entsprechend ihrer Distribution zu den Adjektiven oder Nomina gezählt.

Beispiele:

- *der* **Vierte**/NN
- *der* **vierte**/ADJA *Mann*
- *ein* **Viertel**/NN
- **dreiviertel**/ADJD *voll*

2.3.1 CARD: Kardinalzahlen

Klassifikation von CARD	POS

POS =	Beschreibung	Beispiele
CARD	geschriebene ganze Zahlen	*drei [Männer]*
	ganze Zahlen in Ziffern	*3 [Männer]*
	Jahreszahlen	*[im Juni] 1993*
	Dezimalzahlen in Ziffern	*7,5 [Prozent]*
	Römische Zahlen	*[Kapitel] IV*
	Sportergebnisse	*[der VfB verliert] 0:6*
	Postleitzahlen	*72074 [Tübingen]*
Aber:		
ART	"ein–" in Artikelposition	**eine**/ART *Million*
PIS	NP–substituierendes "ein–"	**einer**/PIS, *der zuhört*
NN	substantivische Zahlwörter	*drei* **Millionen**/NN
NN	Nominalisierungen	*[die] Zwei [gewinnt]*
ADJD	Bruchzahlen	**dreiviertel**/ADJD *voll*
ADJA	Bruchzahlen	*ein* **3/4**/ADJA *Liter*
XY	Postleitzahlen mit Länderkennung	**D–72074**/XY *Tübingen*
XY	Modellkennungen	*das Modell* **DX3E**/XY

Beispiele:

- **eins**/CARD *und* **eins**/CARD *zusammengezählt und die Zahl* **Zwei**/NN *herausbekommen*
- **zwei**/CARD *Häuser weiter*
- **anderthalb**/CARD *Pfund Mehl*
- aber: **ein**/ART **viertel**/ADJA *Pfund Mehl*
- **15**/CARD **Millionen**/NN *Menschen*
- **hundert**/CARD *Prozent*
- aber: *ein* **halbes**/ADJA **Hundert**/NN
- *Schlag* **zwölf**/CARD
- *im Jahre* **2000**/CARD
- *am* **3.2.1994**/CARD
- aber: *am* **3.**/ADJA **2.**/ADJA **1994**/CARD
- *er zählt von* **eins**/CARD *bis* **zehn**/CARD
- aber: *ich habe* **eins**/PIS *gesehen*
- **ein**/CARD *bis* **zwei**/CARD **Millionen**/NN
- aber: **eine**/ART **Million**/NN

2.4 Verben

Im STTS werden drei Typen von Verben unterschieden:

- Die Klasse der mit **VM** getaggten Modalverben umfaßt *können, müssen, wollen, dürfen, mögen* (und auch die Konjunktiv–Form von *mögen, möchten*) und *sollen*.
- Mit **VA** werden die potentiellen Auxiliare *haben, sein* und *werden* gekennzeichnet, unabhängig davon, ob sie im Satz tatsächlich als Voll– oder Hilfsverben gebraucht sind.
- Alle anderen Verben werden als **VV** klassifiziert.

Beispiele:

- *er* **muß/VMFIN einkaufen/VVINF**
- *er* **läßt/VVFIN einkaufen/VVINF**
- *er* **ist/VAFIN** *gegangen*
- *er* **ist/VAFIN** *groß* *(nicht:* **ist/VVFIN***)*

2.4.1 VAFIN, VAIMP, VVFIN, VVIMP, VMFIN: finite Formen

Imperativformen erhalten eine eigene Klasse (VAIMP, VVIMP), da sie sich distributionell von allen anderen finiten Verbformen (VFIN) unterscheiden (V1–Stellung, fehlendes Personalpronomen).

Klassifikation von VAFIN, VAIMP, VVFIN, VVIMP, VMFIN | POS

POS =	Beschreibung	Beispiele
VVFIN	Finite Verbform	*[du] gehst*
VAFIN	(außer Imperativ)	*[sie] wären*
VMFIN	(außer Imperativ)	*[wir] wollten*
VAIMP	Imperativ	*sei [leise !], habt [Geduld !]*
VVIMP		*geh [!], geht [!]*

Beispiele:

- *ich* **würde/VAFIN gehen/VVINF**
- *er* **sagt/VVFIN** , *daß sie* **gehen/VVINF sollen/VMFIN**
- *er* **hat/VAFIN** *ein Auto*
- *er* **hat/VAFIN gehen/VVINF wollen/VMINF**
- *er* **wird/VAFIN geschlagen/VVPP**
- *er* **wird/VAFIN** *ihn* **schlagen/VVINF**
- *er* **wird/VAFIN** *langsam* **wütend/ADJD**

Lexikalische Kategorien für VFIN,VIMP LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von VFIN MOR

Attribut	MOR =	Beispiele
Person	**1**	*ich gehe*/VVFIN:**1**.Sg.Pres.Ind
	2	*du gingst*/VVFIN:**2**.Sg.Past.Ind
	3	*er ist*/VAFIN:**3**.Sg.Pres.Ind
Numerus	**Sg**	*sie habe*/VAFIN:3.**Sg**.Pres.Konj
	Pl	*sie sind*/VAFIN:3.**Pl**.Pres.Ind
Tempus	**Pres**	*du kannst*/VMFIN:2.Sg.**Pres**.Ind
	Past	*du konntest*/VMFIN:2.Sg.**Past**.Ind
Modus	**Ind**	*er hilft*/VVFIN:3.Sg.Pres.**Ind**
	Konj	*er helfe*/VVFIN:3.Sg.Pres.**Konj**

Beachte:

Alle Attribute müssen angegeben werden.

Beispiele:

- *er wird*/VAFIN:3.Sg.Pres.Ind *rot*
- *er werde*/VAFIN:3.Sg.Pres.Konj *sehen*/VVINF
- *er wurde*/VAFIN:3.Sg.Past.Ind *geschlagen*/VVPPF
- *er würde*/VAFIN:3.Sg.Past.Konj *gehen*/VVINF
- *er sagt*/VVFIN:3.Sg.Pres.Ind , *daß sie gehen*/VVINF *sollen*/VMINF
- *wir möchten*/VMFIN:1.Pl.Past.Konj *gehen*/VVINF
- *sie müßten*/VMFIN:3.Pl.Past.Konj *da sein*/VAINF

Morphologische Merkmale von VIMP MOR

Attribut	MOR =	Beispiele
Numerus	**Sg**	*geh*/VVIMP:**Sg** *!*
	Pl	*geht*/VVIMP:**Pl** *!*

Beachte:

Numerus: Muß immer angegeben werden

Sonstiges: Da es im Deutschen nur Imperativformen für die 2. Person gibt, wird auf das Attribut *Person* verzichtet.

Beispiele:

- *gib*/VVIMP:Sg *mir das Buch !*
- *laßt*/VVIMP:Pl *ihn gehen !*
- *werde*/VAIMP:Sg *bloß nicht gleich sauer !*

2.4.2 VVINF, VAINF, VMINF, VVIZU: Inifinitiv

Klassifikation von VVINF, VAINF, VMINF, VVIZU | POS

POS =	Beschreibung	Beispiele
VVINF	reiner Infinitiv, voll	*ankommen, loswerden*
VAINF	reiner Infinitiv, aux	*haben, sein, werden*
VMINF	reiner Infinitiv, modal	*können, müssen*
	Ersatzinfinitiv	*[er hat kommen] wollen*
VVIZU	Infinitiv mit "zu"	*anzukommen, dazusein, loszuwerden*
Aber:		
NN	substantivierter Infinitiv	*das* **Reisen**/NN *macht ihm Spaß*

- In Verbindung mit Infinitiven wird bei manchen Verben (z.B. Modalverben) das Partizip durch den Infinitiv ersetzt. Diese *Ersatzinfinitive* werden auch als Infinitiv (**VMINF**) getaggt.
- Zusammensetzungen aus adverbialen, adjektivischen oder sonstigen Präfixen und *haben, sein, werden* bzw. Modalverben werden als VV..., nicht als VA... bzw als VM... getaggt!!!
  - **bekanntgeworden**/VVPP
  - **dabeisein**/VVINF ist alles
  - **dafürkönnen**/VVINF

Beispiele:

- *er* **will**/VMFIN **kommen**/VVINF
- *er* **verspricht**/VVFIN **zu**/PTKZU **kommen**/VVINF
- *er* **muß**/VMFIN **weggehen**/VVINF
- *er* **ist**/VAFIN **gezwungen**/VVPP **wegzugehen**/VVIZU
- *er* **hat**/VAFIN **gehen**/VVINF **wollen**/VMINF
- *er* **hat**/VAFIN *ihn* **spielen**/VVINF **sehen**/VVINF
- *er* **wird**/VAFIN *ihn* **verraten**/VVINF
- aber: *er* **wird**/VAFIN *von ihm* **verraten**/VVPP
- **dafürzukönnen**/VVIZU
- **dafürkönnen**/VVINF
- **dabeisein**/VVINF
- **kann**/VMFIN **nichts**/PIS **dafür**/PTKVZ

2.4.3 VVPP, VMPP, VAPP: Partizip Perfekt

Klassifikation von VVPP, VMPP, VAPP POS

POS =	Beschreibung	Beispiele
VVPP **VMPP** **VAPP**	nicht–flektiertes Partizip Perfekt	*[er wird] gesucht* *[er hat] gewollt* *[er ist] geworden.*
Aber:		
ADJD	modifizierendes Partizip	**gezielt**/ADJD *eingesetzte Mittel*
ADJD	adverbiales Partizip	*er sucht* **gezielt**/ADJD
ADJA	attributives Partizip	*der* **gesuchte**/ADJA *Verbrecher*

- Es wird nicht unterschieden zwischen aktivischem, passivischem oder prädikativem Gebrauch des Partizips:

  Beispiele:

  - *er* **hat**/VAFIN **gehen**/VINF **wollen**/VMINF
  - *er* **hat**/VAFIN *das Buch* **gewollt**/VMPP
  - *er* **ist**/VAFIN **geschlagen**/VVPP **worden**/VAPP
  - *er hat die Sache* **geplant**/VVPP
  - *die Sache wurde* **geplant**/VVPP
  - *die Sache ist* **geplant**/VVPP

- Partizipien, die adverbial gebraucht werden, werden als ADJD getaggt (distributives Kriterium)!! vgl. dazu Abschnitt 2.2.3

  Beispiele:

  - *er* **hat**/VAFIN *ihn* **verraten**/VVPP
  - *er* **wird**/VAFIN *von ihm* **verraten**/VVPP
  - **aber:** *er* **wird**/VAFIN *ihn* **verraten**/VVINF
  - *er* **ist**/VAFIN **verraten**/VVPP **worden**/VAPP
  - *er* **muß**/VAFIN **verraten**/VVPP **worden**/VAPP **sein**/VAINF
  - *er* **hat**/VAFIN *ihn* **reiten**/VVINF **gelehrt**/VVPP
  - *er* **hat**/VAFIN *ins Kino* **gewollt**/VMPP
  - *er ist als* **vermißt**/ADJD **gemeldet**/VVPP
  - *er kommt früher als* **erwartet**/ADJD
  - **abgesehen**/VVPP **davon**/PAV

2.5 Artikel

2.5.1 ART: bestimmter und unbestimmter Artikel

Bei den Artikeln wird nicht zwischen unbestimmten und bestimmten Artikel unterschieden, da sie sich distributionell betrachtet gleich verhalten.

Klassifikation von ART POS

POS =	Beschreibung	Beispiele
ART	bestimmter oder unbestimmter Artikel	*der, die, das* *ein, eine*
Aber:		
PIS	Indefinitpronomen	**einer**/PIS, *der kommt*
PDS	Demonstrativpronomen	**das**/PDS, *was er sagt*
CARD	Kardinalzahl	**ein**/CARD *bis zwei Millionen*
ADJA	attributives Adjektiv	*der* **eine**/ADJA *und* **andere**/ADJA *Mensch*
PTKVZ	Verbzusatz	*ich lade* **ein**/PTKVZ

Ambiguitäten:

- ART/PDS/PDAT/PRELS/PRELAT:
  - **der**/ART **das**/ART Haus streichende Mann
  - **diese**/PDAT Meinung weicht von **der**/PDS **der**/ART meisten Menschen ab.
  - ist **das**/PDS **die**/ART Frau, **die**/PRELS **die**/ART Tasche verloren hat?
  - der Junge, **dessen**/PRELAT Vater Polizist ist
  - ist das musikalische Äquivalent **dessen**/PDS, was Truman Capote ...
  - Regine und **deren**/PDAT Mann
  - dreiviertel **dessen**/PDS, was hier geredet wird
  - **dessen**/PDS **ungeachtet**/APPO
- ART/PIS/CARD/ADJA:
  - **eine**/ART Tat **eines**/ART guten Mannes
  - **einer**/PIS von insgesamt 16 Abgeordneten
  - **ein**/CARD bis **zwei**/CARD Millionen
  - **der**/ART **eine**/ADJA Arm

Lexikalische Kategorien für ART LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von ART MOR

Attribut	MOR =	Beispiele
Definitheit	**Def**	*der*/ART:**Def**.Masc.Nom.Sg *Hund*
	Indef	*eine*/ART:**Indef**.Fem.Nom.Sg *Katze*
Genus	**Masc**	*ein*/ART:Indef.**Masc**.Nom.Sg *Vogel*
	Fem	*einer*/ART:Indef.**Fem**.Dat.Sg *Giraffe*
	Neut	*ein*/ART:Indef.**Neut**.Nom.Sg *Pferd*
	*	*die*/ART:Def.*.Nom.Pl *Tiere*
Kasus	**Nom**	*der*/ART:Def.Masc.**Nom**.Sg *Elefant*
	Gen	*eines*/ART:Indef.Masc.**Gen**.Sg *Pinguins*
	Dat	*dem*/ART:Def.Neut.**Dat**.Sg *Kamel*
	Akk	*einen*/ART:Indef.Masc.**Akk**.Sg *Frosch*
Numerus	**Sg**	*eine*/ART:Indef.Fem.Nom.**Sg** *Fliege*
	Pl	*die*/ART:Def.*.Akk.**Pl** *Fische*

Beachte:

Genus: Bleibt bei Pluralformen undefiniert.

Kasus und Numerus: Müssen immer angegeben werden.

2.6 Pronomina

Possessiv-, Demonstrativ-, Indefinit-, Interrogativ- und Relativpronomina werden nach ihrer Distribution unterschieden. Als *attribuierend, -AT*, werden Pronomina bezeichnet, die innerhalb einer NP auftreten, *substituierend, -S*, sind Pronomina, die anstelle einer NP stehen. Die jeweils letzten beiden (bzw. der letzte) Buchstaben geben diese Unterscheidung an. Abb. 2.1 zeigt den Aufbau der Tags für Pronomina.

Beispiele:

- **ich**/PPER *wasche* **mich**/PRF
- **meine**/PPOSAT *Bücher*
- **diese**/PDAT *Bücher*
- *das ist* **alles**/PIS
- *der Mann,* **dessen**/PRELAT *Frau hier war*
- **Wohin**/PWAV *gehst du?*
- **das**/PDS *hast du* **davon**/PAV

2.6.1 PPER, PRF: Personal- und Reflexivpronomina

Bei Personalpronomina wird unterschieden zwischen reflexiven Formen

- mich, dich, sich, uns, euch, mir, dir, einander

und sonstigen Personalpronomina:

- ich, du, er, sie, es, wir, ihr (Nom)
- mich, dich, ihn, sie, es, uns, euch (Akk)
- mir, dir, ihm, ihr, ihnen (Dat)
- meiner, deiner, ihrer, seiner, unser(er), eurer (Gen)

Achtung: Es gibt Überschneidungen bei *mir, dir, dich, mich, euch, uns*, die sowohl reflexiv als auch irreflexiv sein können.
In der nachfolgenden Tabelle sind alle möglichen Formen von Personalpronomina aufgeführt.

Klassifikation von PPER, PRF POS

POS =	Beschreibung	Beispiele
PPER	Personalpronomen	*ich, meiner, du, deiner, er, sie, es, seiner, ihrer, ihm, ihn, ihr, wir, unser, ihr, euer, sie, ihrer, ihnen, mich, dich, dir, mir*
PRF	reflexives Personalpronomen	*sich, einander, mich, dich, uns, euch, mir, dir,*
Aber:		
PPOSAT	attribuierendes Possessivpronomen	**ihr**/PPOSAT *Kleid* **euer**/PPOSAT *Auto*

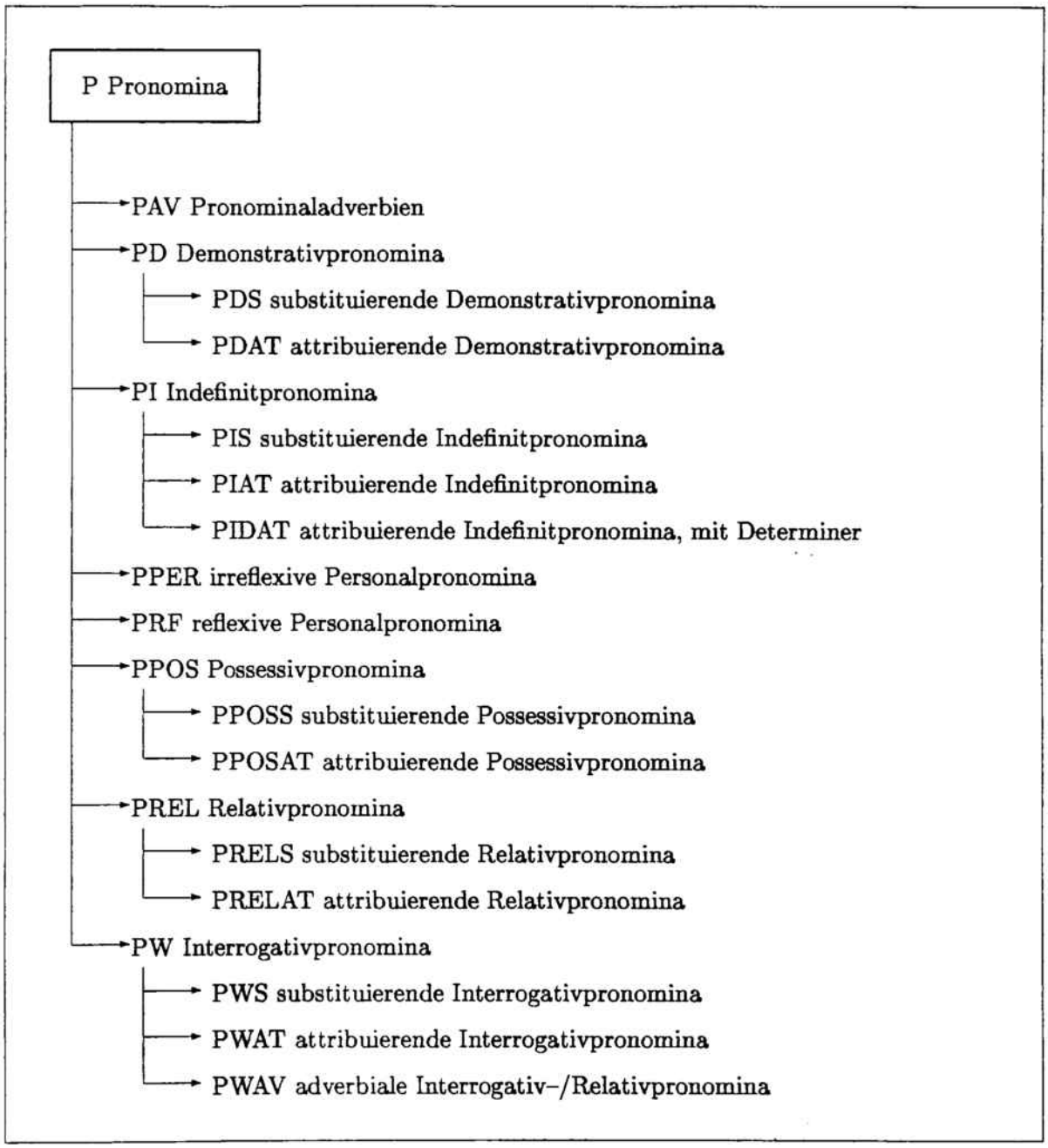


Abbildung 2.1: Pronomina

Beispiele:

- er/PPER *begibt* **sich**/PRF *mit* **dir**/PPER *zu* **ihr**/PPER
- **sie**/PPER *geben* **einander**/PRF *die Hand*
- **aber:** **sie**/PPER *spielen* **miteinander**/ADV, **durcheinander**/ADV, **füreinander**/ADV
- **er**/PPER *ist* **sich**/PRF **ihrer**/PPER *sicher*
- *das ist* **ihr**/PPOSAT *Mann*
- *das ist* **ihrer**/PPOSS

Lexikalische Kategorien für PPER, PRF LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von PPER MOR

Attribut	MOR =	Beispiele
Person	**1**	*wir*/PPER:**1**.Pl.*.Nom
	2	*deiner*/PPER:**2**.Sg.*.Gen
	3	*ihnen*/PPER:**3**.Pl.*.Dat
Numerus	**Sg**	*es*/PPER:3.**Sg**.Neut.Nom
	Pl	*ihr*/PPER:2.**Pl**.*.Nom
Genus	**Masc**	*ihn*/PPER:3.Sg.**Masc**.Akk
	Fem	*sie*/PPER:3.Sg.**Fem**.Nom
	Neut	*es*/PPER:3.Sg.**Neut**.Nom
	*	*du*/PPER:2.Sg.*.Nom
Kasus	**Nom**	*sie*/PPER:3.Pl.*.**Nom**
	Gen	*unser*/PPER:2.Pl.*.**Gen**
	Dat	*ihm*/PPER:3.Sg.Neut.**Dat**
	Akk	*ihn*/PPER:3.Sg.Masc.**Akk**

Beachte:

Genus: Ist nur für die 3. Person Singular definiert.

Person, Numerus und Kasus: Müssen immer angegeben werden.

Beispiele:

- *ich*/PPER:1.Sg.*.Nom *gehe ohne ihn*/PPER:3.Sg.Masc.Akk *zu ihr*/PPER:3.Sg.Fem.Dat
- *du*/PPER:2.Sg.*.Nom *gibst es*/PPER:3.Sg.Neut.Akk *mir*/PPER:1.Sg.*.Dat
- *es*/PPER:3.Sg.Neut.Nom *regnet*

Morphologische Merkmale von PRF MOR

Attribut	MOR =	Beispiele
Person	**1**	*mich*/PRF:**1**.Sg.Akk
	2	*dir*/PRF:**2**.Sg.Dat
Numerus	**Sg**	*dich*/PRF:2.**Sg**.Akk
	Pl	*uns*/PRF:1.**Pl**.Dat
Kasus	**Dat**	*mir*/PRF:1.Sg.**Dat**
	Akk	*uns*/PRF:1.Pl.**Akk**

Beachte:

Person: Für dieses Attribut gibt es nur die Werte *1* und *2*

Numerus: Muß immer angegeben werden.

Kasus: Als Werte gibt es nur Akkusativ und Dativ

Sonstiges: **sich** und **einander** bekommen bei allen Attributen den Wert *.

Beispiele:

- *ich*/PPER:1.Sg.*.Nom *wasche mich*/PRF:1.Sg.Akk
- *ihr*/PPER:2.Pl.*.Nom *gebt ihn*/PPER:3.Sg.Masc.Akk *uns*/PPER:1.Pl.*.Dat
- *sie*/PPER:3.Pl.*.Nom *geben einander*/PRF:*.*.* *die Hände*
- *er begibt sich*/PRF:*.*.* *zu ihr*/PPER:3.Sg.Fem.Dat

2.6.2 PPOSAT, PPOSS: Possessivpronomina

Klassifikation von PPOSAT, PPOSS POS

POS =	Beschreibung	Beispiele
PPOSAT	attribuierendes Possessivpronomen	*seine [Meinung]*
PPOSS	substituierendes Possessivpronomen	*[das ist] meins*

- Die Formen *meinig–, deinig–, seinig–, (etc.)* werden als attribuierende Possessivpronomina getaggt, da sie zwar ohne Nomen, aber nicht anstelle einer vollständigen NP stehen (vgl. attributive Adjektive in Abschnitt 2.2.1).

Beispiele:

- *Das ist* **mein**/PPOSAT *Buch* .
- *Das ist* **meines**/PPOSS .
- *Das ist* **meines**/PPOSAT *Vaters Buch* .
- *Das ist das* **meinige**/PPOSS.

- **mein**/PPOSAT *Vater gibt dem* **deinigen**/PPOSAT **eines**/PIS **seiner**/PPOSAT *Bücher*

Lexikalische Kategorien für PPOSAT, PPOSS LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von PPOSAT, PPOSS MOR

Attribut	MOR =	Beispiele
Genus	**Masc**	*dein*/PPOSAT:**Masc**.Nom.Sg *Bruder*
	Fem	*meine*/PPOSAT:**Fem**.Nom.Sg *Freundin*
	Neut	*das ist ihres*/PPOSS:**Neut**.Nom.Sg
	*	*unsere*/PPOSAT:*.Nom.Pl *Freunde*
Kasus	**Nom**	*seine*/PPOSAT:Fem.**Nom**.Sg *Frau*
	Gen	*meines*/PPOSAT:Masc.**Gen**.Sg *Bruders*
	Dat	*mit ihrem*/PPOSAT:Neut.**Dat**.Sg *neuen Kleid*
	Akk	*ohne euer*/PPOSAT:Neut.**Akk**.Sg *Zutun*
Numerus	**Sg**	*das ist meins*/PPOSS:Neut.Nom.**Sg**
	Pl	*mit deinen*/PPOSAT:*.Dat.**Pl** *Sachen*

Beachte:

Genus: Ist bei Pluralformen nicht definiert.

Kasus und Numerus: Müssen immer angegeben werden. Sie richten sich nicht nach dem Besitzer, sondern nach dem (nachfolgenden) Nomen.

Beispiele:

- *seine*/PPOSAT:**Fem**.Nom.Sg *Mutter*
- **nicht:** *seine*/PPOSAT:**Masc**.Nom.Sg *Mutter*
- *seine*/PPOSAT:*.Nom.**Pl** *Kinder*

2.6.3 PDAT, PDS: Demonstrativpronomina

Klassifikation von PDAT, PDS POS

POS =	Beschreibung	Beispiele
PDAT	attribuierendes Demonstrativ-pronomen	*dieses [Buch]* *jene [Frage]*
PDS	substituierendes Demonstrativ-pronomen	*dies [ist ein Buch],* *jenes [ist schwierig]*
Aber:		
PIDAT	manch, solch, welch	**manch**/PIDAT *einer*
PIAT	mancher	**mancher**/PIAT *sagt*

- **nur substituierend** vorkommende Demonstrativpronomina (**nur** /PDS) sind
  - *der, die, das*
- **nur attributiv** vorkommende Demonstrativpronomina gibt es nicht.
- **substituierend** oder **attribuierend** (/PDS oder /PDAT) verwendet werden:
  - *selb-*
  - *dies-, jen-, {der, die, das}jenig-, {der, die, das}selb-*
  - *ebenjen-, ebendies-, etwelch-*

Beispiele:

- **das**/PDS *weiß ich nicht*
- **diejenige**/PDAT *Person, die* **dasselbe**/PDAT *Kleid trägt*
- **derjenige**/PDS, *der* **dasselbe**/PDS *sagt*
- **das**/PDS *ist* **einer**/PDS, *der ihr gefällt*
- *im* **selben**/PDAT *Monat*

Lexikalische Kategorien für PDAT, PDS

LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von PDAT, PDS

MOR

Attribut	MOR =	Beispiele
Genus	**Masc**	*dieser*/PDAT:**Masc**.Nom.Sg *Tag*
	Fem	*jene*/PDS:**Fem**.Nom.Sg *gefällt ihm*
	Neut	*dieses*/PDAT:**Neut**.Akk.Sg *Mal*
	*	*die*/PDS:*.Nom.Pl *gefallen ihm nicht*
Kasus	**Nom**	*derjenige*/PDS:Masc.**Nom**.Sg , *welcher*
	Gen	*trotz dieses*/PDAT:Masc.**Gen**.Sg *Einwands*
	Dat	*mit diesem*/PDAT:Masc.**Dat**.Sg *Hut*
	Akk	*ohne denjenigen*/PDS:Masc.**Akk**.Sg *zu fragen*
Numerus	**Sg**	*dieser*/PDAT:Masc.Nom.**Sg** *Punkt*
	Pl	*dieselben*/PDAT:*.Nom.**Pl** *Leute*

Beachte:

Genus: Ist bei Pluralformen nicht definiert.

Kasus und Numerus: Müssen immer angegeben werden.

Beispiele:

- *das*/PDS:Neut.Akk.Sg *weiß ich nicht*
- *diejenige*/PDAT:Fem.Nom.Sg *Person, die dasselbe*/PDAT:Neut.Akk.Sg *Kleid trägt*
- *derjenige*/PDS:Masc.Nom.Sg , *der dasselbe*/PDS:Neut.Akk.Sg *sagt*

2.6.4 PIDAT, PIS, PIAT: Indefinitpronomina

Die Indefinitpronomina werden in substituierende (PIS) und attribuierende (PIAT, PIDAT) unterschieden. Bei den attribuierenden gilt das Unterscheidungskriterium, ob das Indefinitpronomen mit Determiner (unbestimmter/bestimmter Artikel, andere Pronomina davor oder dahinter) auftreten kann oder nicht.
Zu den Adjektiven werden nur solche Wortformen gezählt, die sowohl nach definitem als auch nach indefinitem Artikel stehen können (z.b. *ander-*).

Klassifikation von PIDAT, PIS, PIAT POS

POS =	Beschreibung	Beispiele
PIAT	attribuierendes Indefinit–pronomen ohne Determiner vorkommend	*etliche [Dinge], zuviele [Fragen], etwas [Schokolade]*
PIDAT	attribuierendes Indefinit–pronomen, mit Determiner vorkommend	*all [die Bücher] solch [eine Frage] beide [Fragen], viele [Leute]*
PIS	substituierendes Indefinit–pronomen	*etwas, nichts, irgendwas (irgend)wer, man*

PIS:		
all-	allerlei	alles
ander-	anderlei	beid-
beides	beiderlei	bißchen
deinesgleichen	dergleichen	derlei
dreierlei	ebensoviel-	ebensowenig-
ein-	einerlei	einig-
erster-	etlich-	etwas
etwelch-	euresgleichen	ihresgleichen
irgendein-	(irgend)etwas	(irgend)jemand
(irgend)was	(irgend)welch-	(irgend)wem
(irgend)wen	(irgend)wer	(irgend)wessen
jed-	jedermann	jedermanns
jedwed-	jeglich-	jemand
kein-	letzter-	man
manch-	mancherlei	mehr
mehrer-	mehrerlei	meinesgleichen
meist-	nichts	niemand
nix	(ein) paar	reichlich
sämtlich-	seinesgleichen	solch-
solcherlei	sonstjemand	sonstwas
sonstwem	sonstwen	sonstwer
soviel	soviel-	sowas
unsereinem	unsereinen	unsereiner
unsereines	unsereins	unseresgleichen
viel	viel-	vielerlei
vieles	wenig	wenig-
weniger	wenigst-	zuviel
zuviel-	zuwenig	zuwenig-
zweierlei		

PIAT:		
allerlei	anderlei	beiderlei
derlei	dreierlei	ebensoviel
ebensowenig	einig-	etlich-
etwas	etwelch-	euresgleichen
ihresgleichen	irgendein-	jedwed-
jedermanns	kein	kein-
keinerlei	lauter	manch-
mancherlei	mehr	mehrer-
mehrerlei	nichts	reichlich
solcherlei	sovielsoviel	soviel-
sowas	unseresgleichen	vielviel
vielerlei	weniger	zuvielzuviel
zuviel-	zuwenig	zuwenig-
zweierlei		

PIDAT:		
all	all-	beid-
bißchen	erster-	jed-
jeglich-	letzter-	manch
meist-	(ein) paar	sämtlich-
solch	solch-	viel-
welch	wenig	wenig-
wenigst-		

Ambiguitäten zwischen PIS, PIDAT, PIAT

- nur **PIS**:
  - *jemand, niemand, man, jedermann*
  - *ein-, (irgend)was, (irgend)wer, sonstwer, sonstwas*
  - *meinesgleichen, deinesgleichen, ...*
  - *unsereiner, unsereins*
  - *beides, vieles, alles*
  -
- nur **PIAT**:
  - *irgendein [Buch]*
  - *kein [Mensch]*
  - *lauter [Verrückte]*
  - *reichlich [Alkohol]*
  - *keinerlei [Verständnis]*
  - *solcherlei [Unsinn]*

- Nur **PIDAT**:
  - *all [die Leute]*
  - *manch [ein Mensch]*
  - *solch [eine Sache]*
  - *welch [ein Unsinn]*
- **PIAT** oder **PIS**:
  - *kein–, irgendein–* (keine Blumen / keiner kam)
  - *etwas, nichts* (etwas Wasser / etwas ist geschehen; nichts aufregendes / nichts hat sich zugetragen)
  - *viel* (viel Zucker / viel ist geschehen)
  - *mehr, weniger* (attribuierend: **mehr**/PIAT Post; als Head einer NP: **mehr**/PIS kann nicht passieren, **mehr**/ADV als 200 Leute; adverbial: Das war **mehr**/ADV als gut,)
  - *zuviel, zuwenig* (zuviel Zucker / zuviel ist schon geschehen)
  - *(eben)soviel, ebensowenig* (soviel Zucker / soviel ist vorgefallen)
  - *soviel–* (soviele Leute / sovielen ist schlecht geworden)
  - *zuviel–, zuwenig–* (zuviele Gäste / zuviele gingen früh)
  - *etlich–* (etliche Kilometer / etliche kamen im Auto)
  - *jedwed–* (jedweder Fehler / jedweden gefiel es)
  - *manch–* (mancher Mensch / mancher)
  - *mehrer–, einig–* (mehrere Prozentpunkte / mehrere gingen zu Fuß)
  - *mancherlei, vielerlei, allerlei, ...* (allerlei Nonsens / mancherlei ist inzwischen geschehen)
  - *einerlei, zweierlei, dreierlei,...* (zweierlei Kuchen / zweierlei ist inzwischen geschehen)
  - *dergleichen, derlei*(dergleichen/dergleichen Unsinn)
- **PIDAT** oder **PIS**:
  - *meist–, wenigst–* (die meisten Frauen/ die meisten)
  - *all–, sämtlich–* (alle Frauen/ alle)
  - *beid–* (beide Männer/ beide sind gekommen)
  - *jed–, jedwed–, jeglich–* (ein jeder Mensch / ein jeder)
  - *solch–* (ein solcher Mensch/ ein solcher)
  - *erster–, letzter–* (ersterer Bruder/ ersterer)
  - *viel–, wenig–* (viele Menschen/viele)
  - *wenig*(**ein**/ART **wenig**/PIDAT Schokolade/ **ein**/ART **wenig**/PIS war genug) (ml: PIS/PIDAT)

- *bißchen* (**ein**/ART **bißchen**/PIDAT Wein/ **ein**/ART **bißchen**/PIS) (ml: PIS/PIDAT)
- *paar* (ein paar Brote/ ein paar sind schon gegangen) (ml: PIS/PIDAT)

- **ADJA** oder **PIS**:
  - *ander-* (die anderen Leute/ die anderen)
  - *erst-* (ein erster Kontakt/ die ersten)

- **ADV** oder **PIDAT, PIAT oder PIS**
  - etwas
  - reichlich
  - wenig
  - viel
  - mehr
  - zuviel
  - bißchen
  - soviel

Test zur Desambiguierung von PIS/ADV:

- Ersetzung des Wortes durch *nichts*, oder besser noch mögliche Ergänzung zu einer NP ⇒ PIS
- Ersetzung durch *nicht* ⇒ ADV

z. B.

- er hat **wenig**/PIS gegessen
- er hat **wenig Gemüse** gegessen
- er hat **nichts** gegessen (unmarkiert)
- er hat **nicht** gegessen
- er hat **reichlich**/ADV gelacht
- er hat **nicht** gelacht
- *er hat **nichts** gelacht
- *er hat **reichlich Lachen** gelacht

Beispiele:

- **mehr**/ADV *als 20 Mio*
- **nur**/ADV **mehr**/ADV *600 Leute*
- *das weiß ich nicht* **mehr**/ADV

- **viel**/PIAT **mehr**/ADV *als du*
- *die* **einen**/PIS *und die* **anderen**/PIS *sind gegangen*
- *der* **eine**/ADJA *und der* **andere**/ADJA *Arm*
- *wir haben* **andere**/ADJA *Torten gegessen*
- **etwas**/PIAT *Schokolade*
- *unter* **anderem**/PIS *ist das hier der Fall*
- *ein* **anderer**/ADJA *Fall*
- *viele Länder: kein* **anderes**/ADJA *hat so viele Probleme*
- **andere**/PIS *mögen das* **anders**/ADV *sehen*
- aber: **etwas**/ADV *gequält*
- **etwas**/PIS *geschieht*
- **solche**/PIDAT *Farben*
- **solch**/PIDAT *ein Theater*
- **manches**/PIAT **andere**/ADJA *Thema*
- **manch**/PIDAT **anderes**/ADJA *Thema*
- **manch**/PIDAT **schöne**/ADJA *Stunde*
- **manche**/PIAT **schöne**/ADJA *Stunde*
- **kein**/PIAT *Mensch*
- **keiner**/PIS *war da*
- *in* **keiner**/PIAT *Form*
- *er hat* **viele**/PIDAT *Bücher*
- *er trinkt* **viel**/PIAT *Wein*
- aber: *er trinkt* **viel**/PIS
- *er ißt* **zuviel**/PIAT *Fleisch*
- **viel**/ADV *ferngesehen*
- **viel**/ADV *gelacht*
- **viel**/PIS *gegessen*
- **viel**/PIS *gesehen*
- **viel**/ADV **zuviel**/PIS *gemacht*
- **viel**/ADV **zu**/PTKA **viel**/PIS *gemacht*
- aber: *er ißt* **zuviel**/PIS
- *er sieht* **vieles**/PIS *ein*
- **alles**/PIS , *was recht ist*
- **all**/PIAT **diese**/PDAT **vielen**/PIDAT *Leute*
- *die* **beiden**/PIS *kamen gleichzeitig*

- **beide**/PIS *waren da*
- **beide**/PIDAT *Läufer waren gleich schnell*
- *die Läufer waren* **beide**/PIS *gleich schnell*
- *wir waren* **beide**/PIS *sofort zur Stelle*
- *wir tanzten* **alle**/PIS *bis um vier Uhr*
- *wir* **alle**/PIS *waren damals ABBA–Fans*
- *er ißt* **viel**/PIS
- *er ißt* **viel**/PIAT *Schokolade*
- *er lacht* **wenig**/PIS
- *er lacht ein* **wenig**/PIDAT
- **mehr**/ADV *als 200 Leute*
- **mehr**/ADV *als verdoppeln*
- *er weiß* **viel**/ADV **mehr**/PIS *als du*
- *er weiß* **nichts**/PIS **mehr**/ADV
- *er fährt jetzt* **viel**/ADV *schneller*
- **alle**/PIDAT *Kinder* **all**/PIDAT **meiner**/PPOSAT *Freunde*
- **alles**/PIS , *was recht ist*
- **all**/PIDAT **diese**/PDAT **vielen**/PIDAT *Leute*
- *er ist mein* **ein**/PIS *und* **alles**/PIS
- *die* **beiden**/PIS *kamen gleichzeitig*
- **beide**/PIS *waren da*
- **beide**/PIDAT *Läufer waren gleich schnell*
- **alle**/PIDAT *diese Laster*
- **viele**/PIS **dieser**/PDAT *Laster*
- *vor* **allem**/PIS *(ml: ADV)*
- *wir stehen* **alle**/PIS *auch auf schwarze Musik*
- *zufrieden waren denn auch* **alle**/PIS
- *die 8 Betreuer, die* **alle**/PIDAT *in den Gemeinden mitarbeiten*
- *deshalb existieren* **alle**/PIDAT *ihre Institutionen hier*
- *Sie* **alle**/PIDAT *konnten sich im Riesenslalom behaupten*
- *Hinterher sind wir* **alle**/PIDAT *schlauer*

Lexikalische Kategorien für PIDAT, PIS, PIAT

LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von PIDAT, PIS, PIAT

MOR

Attribut	MOR =	Beispiele
Genus	**Masc**	*keiner*/PIS:**Masc**.Nom.Sg *war da*
	Fem	*erstere*/PIDAT:**Fem**.Nom.Sg *Königin*
	Neut	*manches*/PIAT:**Neut**.Nom.Sg *Thema*
	*	*viele*/PIS:*.Nom.Pl *sagen das*
Kasus	**Nom**	*kein*/PIAT:Masc.**Nom**.Sg *Mensch*
	Gen	*keines*/PIAT:Masc.**Gen**.Sg *Menschen*
	Dat	*keinem*/PIAT:Masc.**Dat**.Sg *Menschen*
	Akk	*keinen*/PIAT:Masc.**Akk**.Sg *Menschen*
	*	*man*/PIS:*.*.* *sagt*
Numerus	**Sg**	*einer*/PIS:Masc.Nom.**Sg** *wird eingestellt*
	Pl	*viele*/PIDAT:*.Nom.**Pl** *Leute*

Beachte:

Genus: Ist bei Pluralformen nicht definiert.

Kasus und Numerus: Müssen immer angegeben werden.

Ausnahmen: Bei nicht-flektierten Formen wird für alle Attribute der Wert * vergeben.

Beispiele:

- *solche*/PIDAT:*.Nom.Pl *Sachen*
- **aber:** *solch*/PIDAT:*.*.* *ein Wetter*
- *etwas*/PIS:*.*.* *geht vor sich*
- *viele*/PIDAT:*.Nom.Pl *Kinder waren auf dem Fest*
- **aber:** *er trinkt viel*/PIDAT:*.*.* *Mineralwasser*
- *vor allem*/PIS:Neut.Dat.Sg
- *all*/PIDAT:*.*.* *die vielen Jahre*
- *alle*/PIDAT:Fem.Nom.Sg *Mühe war umsonst*

2.6.5 PRELAT, PRELS: Relativpronomina

Es werden nur nomenattribuierende (PRELAT) und NP–ersetzende (PRELS) Relativpronomina unterschieden, die adverbialen Relativpronomina werden als PWAV getaggt.

Klassifikation von PRELAT, PRELS POS

POS =	Beschreibung	Beispiele
PRELAT	attribuierendes Relativpronomen	*[der Mann ,] dessen [Hut]*
PRELS	substituierendes Relativpronomen	*[derjenige ,] welcher, [das ,] was*
Aber:		
PWAT	attributives Interrogativpronomen	*er weiß ,* **welcher**/PWAT *Zug fährt,* **wessen**/PWAT *Frau er sah*
PWS	substituierendes Interrogativpronomen	*er fragt ,* **was**/PWS *es gibt*
PWAV	adverbiales Interrogativpronomen	*der Grund,* **warum**/PWAV *ich gehe*

- attribuierende Relativpronomen sind nur *deren* und *dessen*.
- substituierende Relativpronomina sind *der, die, das, (etc.), welch–* und *was*.

Beispiele:

- **das**/PDS, **was**/PRELS *er gesagt hat*
- *das Kind,* **das**/PRELS *er kennt*
- *der Mann,* **der**/PRELS **das**/PDS *gesagt hat*
- **aber:** **wer**/PWS *so fragt, ist ein Esel*
- *die Dinge,* **deren**/PRELAT *Nutzen wir erkennen*
- *die Dinge,* **deren**/PRELS *wir uns bedienen*
- *die Dinge,* **derer**/PRELS *wir uns bedienen*
- *die Frage,* **welche**/PRELS *gestellt wurde*
- **aber:** *die Frage,* **welche**/PWAT *Aufgaben gestellt wurden*
- **aber:** *das Buch,* **worüber**/PWAV *wir gesprochen haben*

Lexikalische Kategorien für PRELAT, PRELS LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von PRELAT MOR

(keine morphologischen Merkmale)

Morphologische Merkmale von PRELS MOR

Attribut	MOR =	Beispiele
Genus	**Masc**	*der Mann, der*/PRELS:**Masc**.Nom.Sg *singt*
	Fem	*die Frau, welche*/PRELS:**Fem**.Nom.Sg *spricht*
	Neut	*das Kind, das*/PRELS:**Neut**.Nom.Sg *spielt*
	*	*die Leute, die*/PRELS:*.Akk.Pl *wir trafen*
Kasus	**Nom**	*die Sache, die*/PRELS:Fem.**Nom**.Sg *uns beschäftigt*
	Gen	*die Sache, aufgrund derer*/PRELS:Fem.**Gen**.Sg *wir beschlossen ...*
	Dat	*die Sache, mit der*/PRELS:Fem.**Dat**.Sg *wir uns beschäftigen*
	Akk	*die Sache, ohne die*/PRELS:Fem.**Akk**.Sg *wir nicht auskommen*
Numerus	**Sg**	*das, was*/PRELS:Neut.Nom.**Sg** *uns fehlt*
	Pl	*die Dinge, die*/PRELS:*.Nom.Pl *uns fehlen*

Beachte:

Genus: Ist bei Pluralformen nicht definiert.

Kasus und Numerus: Müssen immer angegeben werden.

Beispiele:

- *der Mann, der*/PRELS:Masc.Nom.Sg *das gesagt hat*
- *das, was*/PRELS:Neut.Akk.Sg *er gesagt hat*
- *die Dinge, deren*/PRELS:*.Gen.Pl *wir uns bedienen*
- *die Dinge, derer*/PRELS:*.Gen.Pl *wir uns bedienen*
- **aber:** *die Dinge, deren*/PRELAT *Nutzen wir erkennen*
- *die Frage, welche*/PRELS:Fem.Akk.Sg *gestellt wurde*

2.6.6 PWAT, PWS: Interrogativpronomina

Interrogativpronomina sind *wer, was, welch–, ...* Sie kommen in direkten oder indirekten Fragesätzen vor (nach *fragen, erkundigen, ...*), aber auch nach *wissen, erklären, ...*

Beispiele:

- **Wer**/PWS *kommt?*
- *er fragt,* **wer**/PWS *kommt.*
- *er weiß,* **wer**/PWS *kommt.*

Klassifikation von PWAT, PWS POS

POS =	Beschreibung	Beispiele
PWAT	attribuierendes Interrogativpronomen	*wessen [Mantel], welche [Farbe]*
PWS	substituierendes Interrogativpronomen	*was [ist los ?], wer [ist da ?]*
Aber:		
PRELS	Relativpronomen	*das ,* **was**/PRELS *er sagt derjenige,* **welcher**/PRELS *meint*

- **PWAT** sind nur Formen von *welch–*, sowie das attributivgebrauchte *wessen*
- **PWS** sind *wer, wessen, wem, wen, was* und allein stehendes *welch–*

Beispiele:

- **welchen**/PWAT *Hut hast du ausgesucht?*
- **welchen**/PWS *von beiden hast du gesehen ?*
- *er will wissen,* **wer**/PWS **wann**/PWAV *mit* **welchem**/PWAT *Zug kommt*
- **wer**/PWS *das sagt, weiß nicht,* **was**/PWS *los ist*
- **Wieviele**/PWAT *Autos du hast!*
- **Wieviele**/PWAT *Autos hast du ?*
- **Was**/PWS **für**/APPR **welche**/PWS *hast du? ('was für welche' → ml: PWS)*
- aber: **Wie**/KOKOM *grosse Autos du hast!*
- aber: **der**/PRELS *das sagt, weiß nicht,* **was**/PWS *los ist*

Lexikalische Kategorien für PWAT, PWS LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von PWAT MOR

Attribut	MOR =	Beispiele
Genus	**Masc**	*welcher*/PWAT:**Masc**.Nom.Sg *Tag ist heute ?*
	Fem	*welche*/PWAT:**Fem**.Nom.Sg *Lage ist besser ?*
	Neut	*welches*/PWAT:**Neut**.Nom.Sg *Los gewinnt ?*
	*	*welche*/PWAT:*.Akk.Pl *Fragen haben Sie ?*
Kasus	**Nom**	*welches*/PWAT:Neut.**Nom**.Sg *Haus ist es ?*
	Gen	*aufgrund welcher*/PWAT:Fem.**Gen**.Sg *Sache ... ?*
	Dat	*mit welchem*/PWAT:Neut.**Dat**.Sg *Recht ... ?*
	Akk	*welchen*/PWAT:Masc.**Akk**.Sg *Wagen fährst du ?*
	*	*wessen*/PWAT:*.*.* *Hut ist das?*
Numerus	**Sg**	*welches*/PWAT:Neut.Nom.**Sg** *Ergebnis ... ?*
	Pl	*welche*/PWAT:*.Nom.Pl *Ergebnisse ... ?*
	*	*wessen*/PWAT:*.*.* *Eltern ... ?*

Beachte:

Genus: Ist bei Pluralformen nicht definiert.

Sonstiges:
- Bei den Formen von *welch-* müssen Kasus und Numerus immer angegeben werden.
- Bei *wessen* sind **alle** Attributwerte undefiniert.

Beispiele:

- *er will wissen, mit welchem*/PWAT:Masc.Dat.Sg *Zug sie kommt*
- *es ist ihm egal, von wessen*/PWAT:*.*.* *Geld er lebt*

Morphologische Merkmale von PWS

MOR

Attribut	MOR =	Beispiele
Genus	**Masc**	*welcher*/PWS:**Masc**.Nom.Sg *ist es ?*
	Fem	*welche*/PWS:**Fem**.Nom.Sg *ist gemeint ?*
	Neut	*welches*/PWS:**Neut**.Akk.Sg *nimmst du ?*
	*	*welche*/PWS:*.Nom.Pl *sind schöner ?*
Kasus	**Nom**	*wer*/PWS:*.**Nom**.Sg *ist da ?*
	Gen	*wessen*/PWS:*.**Gen**.Sg *wird er beschuldigt ?*
	Dat	*mit wem*/PWS:*.**Dat**.Sg *ist er fort ?*
	Akk	*wen*/PWS:*.**Akk**.Sg *hast du gesehen ?*
Numerus	**Sg**	*welchen*/PWS:Masc.Akk.**Sg** *will er ?*
	Pl	*welche*/PWS:*.Akk.Pl *meinst du ?*

Beachte:

Genus: Ist nicht definiert bei:

- Formen von *wer* und bei
- Pluralformen.

Kasus und Numerus: Müssen immer angegeben werden.

Sonstiges: Das Interrogativpronomen *was* erhält als Genus *Neut* und als Numerus it Sg.

- Die morphologischen Merkmale der substituierenden *welch*-Formen stimmen mit denen der attribuierenden überein.
- Bei Formen von *wer* ist Genus nicht definiert und Numerus immer *Sg*.
- Das Interogativpronomen *was* erhält als Genus *Neut* und Numerus *Sg*.

Beispiele:

- *er will wissen, wer*/PWS:*.Nom.Sg *mit wem*/PWS:*.Dat.Sg *kommt*
- *er erklärt, was*/PWS:Neut.Nom.Sg *passiert ist*
- *er weiß, was*/PWS:Neut.Akk.Sg *er gesagt hat*
- *welcher*/PWS:Masc.Nom.Sg *der beiden ist schöner*

2.6.7 PWAV: adverbiale Interrogativ– oder Relativpronomina

Die mit *w*– beginnenden Adverbien (*wann, wo, worüber, ...*) können sowohl als Interrogativ– als auch Relativpronomina verwendet werden. Da die Distribution in indirekten Fragesätzen und Relativsätzen übereinstimmt, werden beide Klassen zusammengefaßt.

Beispiele:

- **Wo**/PWAV *wohnt er?*
- *er fragt ,* **wo**/PWAV *er wohnt*
- *der Ort ,* **wo**/PWAV *er wohnt*

Klassifikation von PWAV POS

POS =	Beschreibung	Beispiele
PWAV	adverbiales Interrogativpronomen adverbiales	*wann [verreist du ?], wo [bist du ?], wann [kommt sie ?]* *[der Grund ,] warum*
Aber:		
KOKOM	Vergleichspartikel	*so schnell* **wie**/KOKOM *er*

- **PWAV** sind
  - *wo, woher, wohin, wann*
  - *wieso, weshalb, warum*
  - *wo* + Präposition: *worüber, wobei, womit, ...*
- In Nebensätzen nach *so* mit Adjektiv oder Adverb ist *wie* Vergleichspartikel.
- *wie* ist PWAV nur in V2–Sätzen, also zum Beispiel in direkten Fragen.

Beispiele:

- **wann**/PWAV *kommst du?*
- *er will wissen,* **wann**/PWAV *du kommst*
- **Wie**/PWAV *geht es dir?*
- *er will wissen,* **wie**/KOUS *es ihr geht*
- **aber:** *er erklärt,* **wie**/KOUS *ein Auto fährt*
- **aber:** *er sieht,* **wie**/KOUS *das Auto um die Ecke fährt*
- **aber:** *er kommt so schnell,* **wie**/KOKOM *er kann*
- *er weiß,* **worüber**/PWAV *er spricht*
- **aber:** **wie**/KOUS *auch immer*
- **aber:** *er will wissen,* **ob**/KOUS *du kommst*

2.6.8 PAV: Pronominaladverbien

Als Pronominaladverbien wird ein Klasse von Adverbien bezeichnet, die sich aus einer Präposition und einem Pronominalstamm zusammensetzen. Sie treten im Satz anstelle einer Präpositionalphrase als Adverbialbestimmung oder Präpositionalobjekt auf.

Klassifikation von PAV POS

POS =	Beschreibung	Beispiele
PAV	"da(r)" + Präposition	*darauf, daneben, daher*
	"hier" + Präposition	*hierauf, hierzu, hiermit*
	der + Präposition	*trotzdem, deswegen, außerdem*
Aber:		
PWAV	"wo(r)" + Präposition	*worüber, womit, wogegen*
PWAV	*wer* + Präposition	*weswegen, weshalb*

Zu den Pronominaladverbien zählen

- *dabei, dadurch, dafür, dagegen, daher, damit, danach, darüber, daran, darauf, daraus, darin, darum, darunter, davon, davor, dazu, dazwischen*
- auch verkürzte Formen *drüber, dran, drauf, drum, drunter*

- *hierbei, hierdurch, hierfür, hierher, hiermit, hierüber, hieran, hierauf, hieraus, hierin, hierum, hierunter, hiervon, hiervor, hierzu*
- *deswegen, deshalb*
- *demzufolge, dementsprechend, demgemäß*
- *seitdem, trotzdem, außerdem*
- *außerdem* ist immer PAV! *daher, dagegen* immer PAV! nie ADV!

Beispiele:

- *er wehrt sich* **dagegen**/PAV
- *er beruft sich* **hierauf**/PAV
- *er hat sich* **dementsprechend**/PAV *verhalten*
- *er hat sich* **seitdem**/PAV *ruhig verhalten*
- **aber:** *er hat sich ruhig verhalten,* **seitdem**/KOUS *er die Strafe kannte*
- *er hat sich* **damit**/PAV *gut ausgekannt*
- **aber:** *er hat sich ruhig verhalten,* **damit**/KOUS *er nicht erwischt wurde.*

2.7 Adverbien

2.7.1 ADV: "echte" Adverbien

Als Adverbien werden nur reine, nicht von Adjektiven abgeleitete, nicht flektierbare Modifizierer von Verben, Adjektiven, Adverbien und ganzen Sätzen verstanden. Wortformen, die auch als attributive Adjektive auftreten und adverbial verwendet werden, die aber semantisch nichts (mehr) mit dem Adjektiv verbindet, und die meistens auch nicht prädikativ verwendet werden können, werden zu den Adverbien gezählt (z.B. *nämlich*).

Klassifikation von ADV | POS

POS =	Beschreibung	Beispiele
ADV	lokale Adverbien	*dort, da, fort*
	temporale Adverbien	*heute, dann, oft*
	modale Adverbien	*gerne, sehr*
	kausale Adverbien	*darum, sonst*
	Abtönungspartikel	*ja, aber, denn, doch, zwar*
	Präp. + "einander"	*miteinander, nebeneinander*
	Ordinalzahlen	*erstens, zweitens, drittens*
	Multiplikativzahlen	*einmal, zweimal, dreimal*
	abgekürzte Formen	*bzw., u.a., z.B.*
Aber:		
ADJD	adverbial gebrauchtes Adjektiv	*er fährt* **schnell**/ADJD, *ein* **schlecht**/ADJD *gespieltes Stück*
	adverbial gebrauchtes Partizip Perfekt	*er fährt* **gekonnt**/ADJD
PAV	Pronominaladverb	*er steht* **daneben**/PAV
PWAV	Interrogativpronomen	**wo**/PWAV *bist du ?*
KON	satzeinleitende Konjunktion	*er will,* **aber**/KON *er kann nicht*
PTKNEG	"nicht"	*er kommt* **nicht**/PTKNEG
PTKVZ	adverbiale Verbpartikel	*er kommt* **vorbei**/PTKVZ

Beispiele:

- *er kommt* **sehr**/ADV **bald**/ADV dort/ADV *an*
- *das hat sich* **immer**/ADV *noch nicht geändert*
- *er wird schon* **irgendwo**/ADV *sein*
- *er kommt* **dann**/ADV **ja**/ADV **wohl**/ADV **doch**/ADV *nicht*
- *er geht* **nur**/ADV **einmal**/ADV *um den Block* **herum**/PTKVZ
- *das hat er* **so**/ADV *gewollt*

Weitere Adverbien:

- *bislang*
- *andermal, jedesmal, manchmal, mehrmals, vielemal, vielmals, einmal*
- *meistens, wenigstens, erstens*

Auch satzinitial:

- **auch**/ADV die Entscheidungsphase fiel schwer.

2.7.2 ADJD oder ADV?

Die Entscheidung, ob ein Adverb oder ein adverbial verwendetes Adjektiv vorliegt, ist in einigen Fällen problematisch, nämlich in den Grenzfällen, bei denen adverbiale und prädikative Lesarten zwar homonym sind, ihre Semantik aber verschiedene Lexikoneinträge rechtfertigt. Das Kriterium in STTS ist listenbasiert: Wortformen, die auf der ADV–Liste stehen, können, wenn ihre Bedeutung satzadverbial ist oder eine der anderen adverbialen Bedeutungen hat (z.B. Modifikation eines Adjektives oder Adverbs).

ADV oder **ADJA** diese Wortformen sind niemals **ADJD**, weil sie keine Kopulakonstruktion bilden können:

- nämlich: *die Frage ist* **nämlich**/ADV, *ob ...* vs. *die* **nämliche**/ADJA *Frage*
- äußerst: *sie waren* **äußerst**/ADV **gewitzt**/ADJD vs. *zur* **äußersten**/ADJA *Not*
- eigentlich: *die Sache ist* **eigentlich**/ADV *die* vs. **die**/ADJA *Frage ist, ...*
- längst: *alles ist* **längst**/ADV *vergessen* vs. *die* **längste**/ADJA *Strecke*
- kürzlich: *ich habe ihn* **kürzlich**/ADV *gesehen* vs. *der Anlaß meines* **kürzlichen**/ADJA *Besuches.*

Nur **ADV** möglich:

- **schließlich**/ADV
- **lediglich**/ADV

ADV oder **ADJD** → Semantisches Kriterium:

- früher: *er ist heute* **früher**/ADJD *gekommen* vs. **früher**/ADV *ist er nie so spät gekommen.*
- eben: *die Straße ist* **eben**/ADJD vs. *das ist* **eben**/ADV *die Frage.*
- gerade: *die Linie ist* **gerade**/ADJD vs. *es ist* **gerade**/ADV *14 Uhr*
- natürlich: *der Baum ist ökologisch und total* **natürlich**/ADJD *großgeworden* vs.**Natürlich**/ADV *haben wir ihn nicht chemisch gedüngt!*
- endlich: *die Menge ist* **endlich**/ADJD vs. *er kommt* **endlich**/ADV

- rund: *der Ball ist* **rund**/ADJD vs. *es waren* **rund**/ADV *100 Gäste da*
- weit: *das Ziel ist* **weit**/ADJD vs. *es sind* **weit**/ADV **mehr**/ADV *als 100 Gäste*
- weiter: *er hat* **weiter**/ADV *nichts zu sagen* vs. *Die Straße ist seit gestern* **weiter**/ADJD.
- ganz: *es war* **ganz**/ADV *dunkel* vs. *das Ei blieb* **ganz**/ADJD.
- sicher: *er geht sehr* **sicher**/ADJD *über den Baumstamm* vs. *Er hat das* **sicher**/ADV *nicht mit Absicht getan.*

Nur **ADJD** möglich

- *die Frage ist* **häufig**/ADJD, *die* **häufige**/ADJA *Frage*
- *die Frage wird* **häufig**/ADJD *gestellt*
- **früh**/ADJD *erkannt*
- **gleich**/ADJD *gemacht!*

ADV oder PIS[5]

- *ein* **etwas**/ADV *farbloser Technokrat*
- *wir haben* **viel**/ADV *gelacht*
- aber: *wir haben* **viel**/PIS *gegessen*
- *wir haben* **reichlich**/ADV *gelacht*
- aber: *wir haben* **reichlich**/PIS *gegessen*

Liste der ADV:

allesamt/ADV	**allzuviel**/ADV	**ausgerechnet**/ADV
ausschließlich/ADV	**beispielsweise**/ADV	**bloß**/ADV
etwas/ADV	**früher**/ADV	**ganz**/ADV
ganz/ADV und **gar**/ADV	**gleich**/ADV	**gut**/ADV (10 Kilo)
halt/ADV	**knapp**/ADV (10 Kilo)	**kürzlich**/ADV
lange/ADV	**längst**/ADV	**letztendlich**/ADV
möglichst/ADV	**natürlich**/ADV	**reichlich**/ADV
rund/ADV (10 Kilo)	**schier**/ADV	**sicher**/ADV
unbedingt/ADV	**vermutlich**/ADV	**wahrlich**/ADV
weitaus/ADV	**ziemlich**/ADV	**zukünftig**/ADV

*** Anmerkung: diese Liste kann man aus Morphologie holen: Wortformen mit ambiger Analyse ADJD ADV ***
Adverbien in prädikativer Stellung:

- Er ist **soweit**/ADV
- sie ist jetzt endgültig **fort**/ADV

[5] siehe dazu auch Abschnitt 2.6.4, Seite 44

2.8 Konjunktionen

2.8.1 KOUI: unterordnende Konjunktion mit Infinitiv

Klassifikation von KOUI	POS

POS =	Beschreibung	Beispiele
KOUI	unterordnende Konjunktion mit Infinitiv	*{um [zu], ohne [zu], anstatt [zu], statt [zu] }*
Aber:		
APPR	Präposition	**ohne**/APPR *daß er es weiß*

Beispiele:

- *er kam,* **um**/KOUI *ihn danach* **zu**/PTKZU *fragen*
- *sie tun alles* **um**/KOUI **zu**/PTKZU *überleben*
- *er trat ein* **ohne**/KOUI *anzuklopfen*
- **aber:** *er trat ein,* **ohne**/APPR *daß es ihm jemand erlaubt hätte*
- **anstatt**/KOUI *sich stur* **zu**/PTKZU *stellen, hätte er verhandeln sollen*

2.8.2 KOUS: unterordnende Konjunktion mit Satz

Die Konjunktionen dieser Klasse leiten einen finiten Nebensatz ein, in der Regel mit Verb–Letzt–Stellung.

Klassifikation von KOUS	POS

POS =	Beschreibung	Beispiele
KOUS	unterordnende Konjunktion mit Satz	*daß, weil, wenn, obwohl, als, damit*
Aber:		
KOKOM	Vergleichspartikel	*besser* **als**/KOKOM *er*
		so gut **wie**/KOKOM *er*
PWAV	Interrogativpronomen	*er weiß,* **weswegen**/PWAV *sie kam*

Beispiele:

- *er weiß ,* **daß**/KOUS *du kommst*
- *er will wissen ,* **ob**/KOUS *du kommst,* **damit**/KOUS *er planen kann*
- **aber:** *er will wissen ,* **wann**/PWAV *du kommst*
- **wenn**/KOUS *du kommen könntest, würde er sich freuen*
- **obwohl**/KOUS *es dunkel war, sah er,* **wie**/KOUS *die Tür aufging*

- Ausnahme: *weil* läßt auch einen V2–Satz zu, wird aber trotzdem zu den unterordneten Konjunktionen gezählt:

  Beispiele:

  - *ich frage ihn gar nicht erst,* **weil**/KOUS *er ja doch nichts weiß*
  - *ich frage ihn gar nicht erst,* **weil**/KOUS *er weiß ja doch nichts*
  - *ich sage nichts,* **zumal**/KOUS *du ja ohnehin nicht antworten wirst.*

2.8.3 KON: nebenordnende Konjunktion

Die Konjunktionen dieser Klasse erlauben V2–Stellung.

Klassifikation von KON | POS

POS =	Beschreibung	Beispiele
KON	einfache Konjunktion	*und, oder*
	mehrteilige Konjunktion	*entweder ... oder, werde ... noch*
	satzeinleitende Konjunktion	*denn, aber, doch, jedoch*
Aber:		
ADV	eingeschobenes "aber", "doch"	*er war* **doch**/ADV *gar nicht da, er ging* **aber**/ADV *gleich wieder*

- Einfache nebenordnende Konjunktionen sind nur *und, oder, sowie*
- Mehrteiligen nebenordnenden Konjunktionen sind *entweder ... oder; sowohl ... als (auch); weder ... noch.* Dabei werden alle Teile als **KON** getaggt.
- Abgekürzte mehrteilige Konjunktionen (ohne Leerzeichen: *d.h., z.B., bzw.*) werden gesamt als **KON** getaggt.
- Satzeinleitend sind *aber, doch, denn, jedoch.* Sie werden nur dann als **KON** getaggt, wenn sie am Anfang des nebengeordneten Satzes stehen, sonst als **ADV**.

Beispiele:

- **je**/KOUS *schöner die Spatzen singen,* **desto**/KON *später ist es.*[6]
- **je**/KOUS *später der Abend,* **um**/APPR **so**/ADV *schöner die Gäste.*
- **je**/KOUS *später der Abend,* **umso**/KON *schöner die Gäste.*
- *Waren sie auch hungrig,* **so**/ADV *aßen sie doch noch nicht.*
- **So**/ADV *gingen sie denn.*
- **So**/ADV *gut war das auch nicht!*
- *es wird immer später,* **je**/KOUS *öfter ich auf die Uhr sehe.*
- **je**/ADV *nach Familienstand*

[6] *je* regiert einen VL–Satz, *desto* oder *umso* einen V2–Satz.

- **je**/ADV *mehr sich die Familien anpassen*
- **aber:** *diese Wohnung kostet 1000 DM* **je**/APPR *Quadratmeter.*
- *Männer* **und**/KON *Frauen*
- *Männer* **wie**/KOKOM *Frauen*
- **sowohl**/KON *Männer* **als**/KON **auch**/ADV *Frauen*
- *Männer* **sowie**/KON *Frauen*
- **aber:** **Sowie**/KOUI *er sie sah, kam er angelaufen*
- **entweder**/KON *er* **oder**/KON *ich*
- **weder**/KON *er* **noch**/KON *ich*
- **sowohl**/KON *Kinder* **als**/KON **auch**/ADV *Eltern*
- *er sah sie,* **aber**/KON *er erkannte sie nicht wieder*
- *er sah sie,* **aber**/KON *erkannte sie nicht wieder*
- **aber:** *er sah sie, er erkannte sie* **aber**/ADV *nicht wieder*
- **aber:** *er sah sie, erkannte sie* **aber**/ADV *nicht wieder*
- *er sah sie,* **jedoch**/KON *er erkannte sie nicht wieder*
- *er sah sie,* **jedoch**/KON *erkannte sie nicht wieder*
- *er sah sie,* **jedoch**/KON *erkannte er sie nicht wieder*
- *Peter,* **d.h.**/KON *mein Bruder, ...*
- *ein Mittelklassewagen,* **z.B.**/KON *ein Golf, ...*
- **aber:** *Frauen* **wie**/KOKOM *Männer*

2.8.4 KOKOM: Vergleichspartikel

POS =	Beschreibung	Beispiele
KOKOM	Vergleichspartikel ohne Satz! Auch ohne Vergleichssemantik	*{als, wie }* *als [Taxifahrer]*
Aber:		
KOUS	Satzeinleitendes *wie* oder *als*	*als [er schwamm],*
	Relativpronomen	*[die Art ,] wie er es macht*
	Interrogativpronomen	*[er weiß,] wie [es geht]*
PWAV	Direkte Fragen mit *wie*	*wie [geht es dir?]*

- Vergleichspartikel sind nur *als, wie.*
- KOKOM bezeichnet alle *als, wie*, die NICHT satzeinleitend verwendet werden, z.B. NP folgt, AP folgt...

- Als Kompromiß werden die konjunktionsartigen *wie, als* (KOUS) von den übrigen *wie, als* (KOKOM) getrennt. Letztere Klasse könnte man in solche mit vergleichender Semantik und solche ohne Vergleichssemantik einteilen; erstere in solche, wo *wie auf diese Art und Weise* bedeutet und in alle anderen. Da diese semantischen Unterscheidungen vage sind, treffen wir im jetzigen Tagset nur die syntaktischen.
- *wie* in direkten Fragesätzen ist immer PWAV!

Beispiele:

- *er kommt öfter* **als**/KOKOM *geplant*
- **aber:** *er fährt schneller ,* **als**/KOUS *die Polizei erlaubt*
- **aber:** *ich lachte,* **als**/KOUS *er ins Zimmer kam*
- *er arbeitet* **als**/KOKOM *Taxifahrer*
- *er gilt* **als**/KOKOM *fleißig*
- **wie**/PWAV *soll das weitergehen?*
- *er weiß,* **wie**/KOUS *gut sie kocht*
- *er kommt nicht so oft* **wie**/KOKOM *du*
- *er benimmt sich* **wie**/KOKOM *ein Idiot*
- *entpuppte sich* **als**/KOKOM *stimmenträchtiges Zugpferd.*
- *Einrichtungen* **wie**/KOKOM *Krankenhäuser*
- **aber:** *einen Betrieb,* **wie**/KOUS *ihn die Gewerkschaft definiert*

Damit haben *wie* und *als* je 3 Analysen:

- er arbeitet **als**/KOKOM Bauer
- **als**/KOUS er hereinkam, ...
- sowohl Kinder **als**/KON auch Frauen
- **Wie**/PWAV geht es dir?
- **wie**/KOUS aus dem Innenministerium verlautete, ...
- **Wie**/KOKOM schnell du bist!
- so schnell **wie**/KOKOM Brigitte
- Männer **wie**/KOKOM Frauen
- arbeitet **wie**/KOKOM ein Wilder

2.9 Adpositionen

Es wird grundsätzlich zwischen Präpositionen, Postpositionen und Zirkumpositionen unterschieden. Allerdings wird bei einer Zirkumposition (z.B. *von ... an*) der erste Teil immer als Präposition getaggt und nur der zweite Teil durch **APZR** gekennzeichnet. Im Deutschen kann eine Reihe von Präpositionen auch als Postposition auftreten.

Beispiele:

- **entlang**/APPR *der Straße*
- *die Straße* **entlang**/APPO

Es wird nicht nach lokalen, temporalen, kausalen oder modalen Präpositionen unterschieden oder danach, welchen Kasus sie beim Bezugswort fordern.

2.9.1 APPR: Präposition

Klassifikation von APPR | POS

POS =	Beschreibung	Beispiele
APPR	Präposition	*mit, ohne, bis, mittels, trotz,*
	lokal	*auf, unter, über,*
	temporal	*während, zwischen,*
	kausal	*infolge, unbeschadet,*
	modal	*einschließlich, gemäß*
	linker Teil einer	*um [der Sache willen],*
	Zirkumposition	*von [heute an]*
Aber:		
APPRART	Präposition mit Artikel	*er geht* **zum**/APPRART *Arzt*
PTKA	"zu" vor Adjektiv	*er geht* **zu**/PTKA *schnell*
PTKZU	"zu" vor Infinitiv	*er braucht nicht* **zu**/PTKZU *kommen*
PTKVZ	abgetrennter Verbzusatz	*er kommt* **an**/PTKVZ
KOKOM	"wie", "als"	*er arbeitet* **als**/KOKOM *Lehrer*
		sie arbeitet **wie**/KOKOM *eine Irre*

Beispiele:

- *er steht* **mit**/APPR *dem Hund* **auf**/APPR *der Straße*
- *er denkt* **an**/APPR *seinen Urlaub* **in**/APPR *Spanien*
- *er arbeitet* **von**/APPR *sieben* **bis**/APPR *vier*
- *er arbeitet* **von**/APPR *morgen* **an**/APZR
- **dank**/APPR *Susanne*
- **zeit**/APPR *seines Lebens*

- **mittels**/APPR *Susannes Fleckenlöser*
- **hinsichtlich**/APPR *unseres Zeitplans*
- **bis**/APPR **zu**/APPR *20 Mark (ml: APPR)*
- **bis**/APPR **zur**/APPR *Haustür (ml?)*
- **rund**/ADV **um**/APPR *die Uhr*
- **um**/APPR **so**/ADV *schöner sang sie (ml: KON)*
- *je größer die Torte,* **umso**/KON *größer die Freude*
- **östlich**/APPR *der Elbe*

Lexikalische Kategorien für APPR LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von APPR MOR

Attribut	MOR =	Beispiele
Kasus	**Nom**	*Behandlungsdauer je*/APPR:**Nom** *angemeldeter Patient*
	Gen	*hinsichtlich*/APPR:**Gen** *dieses Anklagepunktes*
	Dat	*aus*/APPR:**Dat** *sozialer Verantwortung*
	Akk	*durch*/APPR:**Akk** *diese hohle Gasse*

Beispiele:

- *er steht mit*/APPR:Dat *dem Hund auf*/APPR:Dat *der Straße*
- *er denkt an*/APPR:Akk *seinen Urlaub in*/APPR:Dat *Spanien*
- *er arbeitet von*/APPR:Dat *morgen an*/APZR
- *dank*/APPR:Gen *seines Wissens*
- *zeit*/APPR:Gen *seines Lebens*
- *mittels*/APPR:Gen *Susannes Fleckenlöser*
- *bis*/APPR:Akk *zur*/APPR:Dat *Haustür*
- *rund*/ADV *um*/APPR:Akk *die Uhr*

2.9.2 APPRART: Präposition mit Artikel

Klassifikation von APPRART POS

POS =	Beschreibung	Beispiele
APPRART	Präposition mit inkorporiertem Artikel	*am, ans, zur, zum*
Aber:		
PTKA	"am" vor Superlativ	*es ist* **am**/PTKA *besten*

Beispiele:

- *er geht* **am**/APPRART *Montag wieder* **zur**/APPRART *Arbeit*
- *er denkt* **beim**/APPRART *Arbeiten immer* **ans**/APPRART *Schlafen*

Lexikalische Kategorien für APPRART LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von APPRART MOR

Attribut	MOR =	Beispiele
Genus	**Masc**	*am*/APPRART:**Masc**.Dat *Montag*
	Fem	*zur*/APPRART:**Fem**.Dat *Sache*
	Neut	*im*/APPRART:**Neut**.Dat *Haus*
Kasus	**Dat**	*beim*/APPRART:Neut.**Dat** *Essen*
	Akk	*ins*/APPRART:Neut.**Akk** *Theater*

Beachte:

Genus und Kasus: Müssen immer angegeben werden.

Sonstiges: Verschmelzungen kommen nur mit defintitem Artikeln im Singular vor. Deshalb wird auf die Attribute *Definitheit* und *Numerus* verzichtet.

Beispiele:

- *er geht am*/APPRART:Masc.Dat *Montag wieder zur*/APPRART:Fem.Dat *Arbeit*
- *er denkt beim*/APPRART:Neut.Dat *Arbeiten immer ans*/APPRART:Neut.Akk *Schlafen*

2.9.3 APPO: Postposition

Klassifikation von APPO POS

POS =	Beschreibung	Beispiele
APPO	Postposition	*zuwider, wegen, entlang, halber*
Aber:		
APPR	Präposition	**wegen**/APPR *der Liebe*
APZR	rechter Teil einer Zirkumposition	**von**/APPR *Rechts* **wegen**/APZR
PTKVZ	abgetrennter Verbzusatz	*er fährt die Straße* **entlang**/PTKVZ

Beispiele:

- *der Liebe* **wegen**/APPO
- *seiner Mutter* **zuliebe**/APPO *kommt er heim*
- *den ganzen Weg* **entlang**/APPO *stehen Bäume*
- *den Tag* **über**/APPO *arbeitete er fleißig*

APPO:		
entgegen/APPO	**entlang**/APPO	**gegenüber**/APPO
gemäß/APPO	**halber**/APPO	**nach**/APPO
über/APPO	**ungeachtet**/APPO	**weg**/APPO
wegen/APPO	**zufolge**/APPO	**zuliebe**/APPO

Lexikalische Kategorien für APPO LEX

(keine lexikalischen Kategorien)

Morphologische Merkmale von APPO MOR

Attribut	MOR =	Beispiele
Kasus	**Gen**	*aller Ermahnungen ungeachtet*/APPO:**Gen**
	Dat	*der nächsten Generation zuliebe*/APPO:**Dat**
	Akk	*die Heizkosten mitgerechnet*/APPO:**Akk**

Beispiele:

- *der Liebe wegen*/APPO:Dat
- *seiner Mutter zuliebe*/APPO:Dat *kommt er heim*
- *den ganzen Weg entlang*/APPO:Akk *stehen Bäume*
- *den Tag über*/APPO:Akk *arbeitete er fleißig*

2.9.4 APZR: Zirkumposition rechts

Klassifikation von APZR POS

POS =	Beschreibung	Beispiele
APZR	rechter Teil einer Zirkumposition	*[um ...] willen, [von ...] an*
Aber:		
ADV	"her" + Präposition	*um Ulm* **herum**/ADV
ADV	"hin" + Präposition	*auf den Berg* **hinauf**/ADV

Beispiele:

- **von**/APPR *morgen* **an**/APZR *wird alles anders*

- **aber:** **von**/APPR *heute* **bis**/APPR *morgen*
- **um**/APPR *des lieben Friedens* **willen**/APZR

APZR:		
ab/APZR	**an**/APZR	**aus**/APZR
wegen/APZR	**willen**/APZR	

2.10 Partikel

2.10.1 PTKZU: "zu" vor Infinitiv

Klassifikation von PTKZU POS

POS =	Beschreibung	Beispiele
PTKZU	"zu" vor Infinitiv	*[ohne] zu [wollen]*
	"zu" vor Partizipien Futur	*[in der] zu [zerstörenden Stadt]*
Aber:		
PTKA	"zu" vor Adjektiv	*er ist* **zu**/PTKA *groß,*
	"zu" vor Adverb	*er fährt* **zu**/PTKA *schnell*
APPR	Präposition "zu"	*er geht* **zu**/APPR *ihr*
PTKVZ	abgetrennter Verbzusatz "zu"	*er stimmt* **zu**/PTKVZ

- Die Klasse **PTKZU** enthält als einzige Wortform *zu*, das unmittelbar vor einem Infinitiv steht.

Beispiele:

- *er bittet ihn* **zu**/PTKZU **kommen**/VVINF
- *er redete ohne* **zu**/PTKZU *überlegen drauflos*

2.10.2 PTKNEG: Negationspartikel

Klassifikation von PTKNEG POS

POS =	Beschreibung	Beispiele
PTKNEG	"nicht"	*[er kommt] nicht*
Aber:		
ADV	negative Adverbien	*er kommt* **nie**/ADV
PIS	Indefinitpronomen 'kein–'	**keiner**/PIS *kam*

- Die Wortart **PTKNEG** umfaßt nur die Wortform *nicht*. Andere Formen wie *nie, niemals, nirgends, ...* werden als Adverbien getaggt.

Beispiele:

- *er kommt heute* **nicht**/PTKNEG
- *er kommt heute* **gar**/ADV **nicht**/PTKNEG
- *ist das* **nicht**/PTKNEG *schön*
- *was die Kinder* **nicht**/PTKNEG *alles wissen*

2.10.3 PTKVZ: abgetrennter Verbzusatz

Das Tag PTKVZ umfaßt sowohl "echte" trennbare Verbpräfixe wie *an-[kommen], ein--[kaufen], um-[formen]* als auch nominale (oder ähnliche) Verbzusätze wie *statt[finden], teil[nehmen]* oder *überhand[nehmen], fehl[schlagen]*.
Zu den Verbzusätzen werden auch solche Formen, die als Adverb, Adjektiv oder Postposition auftreten können, gerechnet!!!!
Ein Verbzusatz tritt nur mit finiten Verben in Sätzen mit Hauptsatzstellung (V2 oder V1) frei auf. In Infinitiv, Partizip oder Nebensätzen (VL) sind Verbzusätze mit dem Verb verbunden und werden nicht getrennt getaggt.

Beispiele:

- *er* **hört**/VVFIN **auf**/PTKVZ
- **hör**/VVIMP **auf**/PTKVZ *!*
- **aber:** *er will* **aufhören**/VVINF
- **aber:** *er hat* **aufgehört**/VVPP
- *er kommt* **herbei**/PTKVZ
- *er gehört* **dazu**/PTKVZ

Klassifikation von PTKVZ | POS

POS =	Beschreibung	Beispiele
PTKVZ	trennbare Verbpräfixe	*[er kommt] an*
	nominale Verbzusätze	*[er nimmt] teil, [er läuft] eis*
	andere Verbzusätze	*[es schlägt] fehl, [er setzt] instand*
	adverbiale Verbzusätze	*[er kommt] herum*
	adjektivische Verbzusätze	*[er hält] geheim*
	Postpositionen	*[er geht die Straße] entlang*

- Die trennbaren Verbzusätze umfassen *ab, an, auf, aus, bei, dar, durch, ein, mit, nach, um, vor, zu*
- Weitere Verbzusätze sind Formen, die aus Nomen oder Präposition + Nomen abgeleitet sind:
  - *rad[fahren], eis[laufen]*
  - *statt[finden], teil[nehmen]*
  - *zustande[kommen], zunichte[machen]*

- Andere Verbzusätze, die in Form und Distribution mit einem Adverb, Adjektiv oder Postposition übereinstimmen. Beispielsweise die folgenden Formen:
  - *her*(+ Präposition)
  - *hin*(+ Präposition)
  - Präposition + *einander*
  - *fort, wohl, ...*

PTKVZ oder ADV bei mehreren Partikeln (Beispiel 'mit')

- Wenn 'mit' + das entsprechende Verb ein Präfixverb ergibt ⇒ PTKVZ
- Wenn 'mit' + das entsprechende Verb nicht lexikalisiert ist ⇒ ADV
- sonst Test: Topikalisierung der Partikel moeglich? ⇒ topikalisierter Partikel → ADV, anderer Partikel → PTKVZ, z.B. *kommst Du* **mit**/ADV **runter**/PTKVZ *in den Keller?*

  mit in den Keller runterkommen

  **runter in den Keller mitkommen*

Beispiele:

- *er werkelte* **mit**/PTKVZ
- *steigst Du* **mit**/ADV **auf**/PTKVZ *den Berg? (aufsteigen)*
- *steigst Du* **mit**/ADV **auf**/PTKVZ*?*
- *kommst Du* **mit**/PTKVZ *schwimmen? (mitkommen)*
- *kommst Du* **mit**/ADV **runter**/PTKVZ *in den Keller?*
- *er kam* **an**/PTKVZ*, packte seine Sachen* **aus**/PTKVZ *und fuhr wieder* **weg**/PTKVZ
- aber: *er ist* **angekommen**/VVPP*, hat seine Sachen* **ausgepackt**/VVPP *und ist wieder* **weggefahren**/VVPP
- *der Senat stimmt* **ab**/PTKVZ *und der Präsident* **zu**/PTKVZ
- *er fährt* **rad**/PTKVZ
- aber: *er fährt* **Auto**/NN
- *er steht* **kopf**/PTKVZL
- aber: *er steht* **Schlange**/NN
- *er geht* **aus**/PTKVZ
- *er geht* **zugrunde**/PTKVZ
- *es geht der Sonne* **entgegen**/PTKVZ
- *er geht* **hinein**/PTKVZ
- *er geht* **verloren**/PTKVZ
- *er geht* **spazieren**/PTKVZ
- aber: *er geht* **langsam**/ADJD
- aber: *er geht* **waschen**/VVINF

2.10.4 PTKA: Partikel bei Adjektiv oder Adverb

Klassifikation von PTKA	POS

POS =	Beschreibung	Beispiele
PTVA	"am" vor Superlativ "zu", "allzu" vor Adjektiv oder Adverb	*am [besten]* *[er ist] zu [groß]* *[er fährt] zu [schnell]*
Aber:		
ADV	Adverb	*er fährt* **sehr**/ADV *schnell* *er fährt* **viel**/ADV *schneller*

Beispiele:

- *er war nicht* **allzu**/PTKA *begeistert*
- *sie kamen* **zu**/PTKA *dritt* **zu**/PTKA *spät* **zu**/APPR *der Party*
- *er war* **am**/PTKA *schnellsten* **am**/APPRART *Ziel*

2.10.5 PTKANT: Antwortpartikel

Als Antwortpartikel werden die Wortformen *ja, nein, danke, bitte* bezeichnet, die im allgemeinen nur in direker Rede vorkommen und dann alleine einen Satz bilden oder in einem Antwortsatz als Bejahung, Verneinung oder Verstärkung verwendet werden.

Klassifikation von PTKANT	POS

POS =	Beschreibung	Beispiele
PTVANT	Antwortpartikel	{*ja, nein, danke, bitte, doch* }
Aber:		
ADV	Abtönungspartikel	*er ist* **ja**/ADV *schon da*

Beispiele:

- *er sagte :* " **Nein**/PTKANT , **danke**/PTKANT " , *und ging*
- **aber:** *sein* **Nein**/NN *zur EG*
- *Kommst du nicht?* **Doch**/PTKANT, *ich komme.*

2.11 Interpunktionen

2.11.1 $, $(, $.

Klassifikation von $, $(, $.	POS

POS =	Beschreibung	Beispiele
$,	nur Komma	,
$(	satzintern, nicht Komma	([{ “
$.	satzfinale Satzzeichen	. ! ? : ;

Beispiele:

- *in* **Glass**/NE *'/$(Besitz*

2.12 Sonstige

2.12.1 ITJ: Interjektionen

Interjektionen sind Wörter,

> die zum Ausdruck von Empfindungen, Flüchen und Verwünschungen sowie zur Kontaktaufnahme dienen. ... sie sind formal unveränderlich, stehen syntaktisch außerhalb des Satzzusammenhanges und haben (im strengen Sinn) keine lexikalische Bedeutung. ([Bußmann 1990])

Klassifikation von ITJ	POS

POS =	Beschreibung	Beispiele
ITJ	Interjektion	*ach, äh, mhm, tja, hoppla, bravo, ...*
Aber:		
ADV	Abtönungspartikel	*er ist* **ja**/ADV *schon da*

2.12.2 TRUNC: Kompositions–Erstglied

Mit **TRUNC** werden Wortteile bezeichnet, die mit einem Bindestrich enden, der einen Teil des nachfolgenden, mit *und, oder* verknüpften Wortes ersetzt.

Klassifikation von TRUNC	POS

POS =	Beschreibung	Beispiele
TRUNC	Präfix	*be– [und entladen],*
		Ein– [und Ausgang],
	Kompositionsglied	*Damen– [und Herrenbekleidung]*
Aber:		
PTKVZ	abgetrenntes Verbpräfix	*er packt* **ein**/PTKVZ

Beispiele:

- *der* **Obst–**/TRUNC *und Gartenbauverein*

- **Ein–**/TRUNC *und Ausgang*
- *er wird es* **ein–**/TRUNC *und auspacken.*
- **aber:** *er packt es* **ein**/PTKVZ *und wieder* **aus**/PTKVZ

2.12.3 XY: Nichtwörter

Nicht–alphabetische Zeichen (§, ©, *$ etc.*), römische Zahlzeichen etc sind so zu taggen, wie das ausgeschriebene Wort getaggt würde, in Analogie zu Abkürzungen.

Beispiele:

- *Er wurde nach* §/NN **301**/CARD *verurteilt.*
- *Sie hat* **$**/NN **200**/CARD *verloren.*

Ist dies nicht möglich (vor allem bei größeren Symbolgruppen, Nichtwörtern sowie Kombinationen aus Ziffern und Zeichen, die sich nicht als CARD oder ADJA einordnen lassen), so wird das Tag XY vergeben.

Beispiele:

- *Das Modell* **DX3E**/XY *gehorcht all Ihren Wünschen.*
- **aber:** *Das Match ging* **4:3**/CARD *aus.*
- *Schicken Sie es in die Blumenstraße 2,* **D–70186**/XY *Stuttgart.*
- *um* **16.03**/CARD *Uhr*

Klassifikation von XY POS

POS =	Beschreibung	Beispiele
XY	Nichtwort	*D–70174 [Stuttgart]* *08/15*
Aber:		
NE	Eigennamen	**C&A**/NE
NN	Währungen, Paragraph ...	**$**/NN, §/NN
CARD	Kardinalzahl	**17,5**/CARD **70174**/CARD *Stuttgart*
ADJA	Ordinalzahl	**23.**/ADJA *Mai*

Beispiele:

- *laut* §/NN **234b**/XY *muß er* **35**/CARD **$**/NN *zahlen*
- *in Kapitel* **II**/CARD *und* **IV**/CARD

2.12.4 FM: Fremdsprachliches Material

Größere Textstücke, die einer fremden Sprache angehören, und nicht als Eigennamen klassifiziert werden können, werden als fremdsprachliches Material getaggt.

Beispiele:

- *Er hat das mit "***but/FM this/FM was/FM not/FM so/FM***" übersetzt.*
- *der spanische Film "***mujer/FM de/FM Benjamin/NE***"*
- *Sie hat ihn dann einfach "***lazy/FM***" genannt.*
- *Diese Sache kann auch in anderen europäischen Sprachen zu Problemen führen: "***je/FM ne/FM sais/FM pas/FM***" ist äquivalent zu "***j'ai/FM pas/FM***", und somit ...*

Auf keinen Fall ist das fremdsprachliche Material auf die deutsche Syntax zu übertragen!! Was als Eigennamen erkannt wird, ist **mit** /NE zu taggen.

Beispiele:

- *Der Film "***A/FM fish/FM called/FM Wanda/NE***" lief nicht in jedem Theater.*
- **New/NE York/NE**
- **University/NE of/NE Michigan/NE**

Lexikalisierte Lehnwörter sind als entsprechende Kategorie zu taggen:

Beispiele:

- *Er macht viel* **Yoga/NN** *in* **Jeans/NN**, *und er* **joggt/VVFIN** *auch häufig.*
- *sie besitzt einen* **Cadillac/NN**

Komplexe fremdsprachliche Ausdrücke, die eine syntaktische Funktion im Satz erfüllen, sind vom Tokenizer zu bündeln. Sie sollten wie entsprechende deutsche Ausdrücke getaggt werden. Problem i.A.: Tokenizer

Beispiele:

- **last–but–not–least/ADV**
- **persona–non–grata/NN**
- **per–se/ADV**

Als Notlösung können die entsprechenden Einzelteile mit **FM** getaggt werden.

Literaturverzeichnis

[Bußmann 1990] Hadumod Bußmann: *Lexikon der Sprachwissenschaft.* Alfred Kröner Verlag, Stuttgart, 1990.

[Duden 1984] Günther Drodowski et al. (Hrsg): *Duden Bd. 4, Grammatik der deutschen Gegenwartssprache.* Dudenverlag, Mannheim, Wien, Zürich, 1994.

[Helbig, Buscha 1991] Gerhard Helbig und Joachim Buscha: *Deutsch Grammatik. Ein Handbuch für den Ausländerunterricht.* Langenscheidt - Verlag Enziklopädie, Leipzig, Berlin, München, Wien, Zürich, New York, 1991.

[TEI 91] TEI AI1W2 (1991): *List of Common Morphological Features For Inclusion in TEI Starter Set Of Grammatical-Annotation Tags.*