Dissertationen aus der Philosophischen Fakultät II
der Universität des Saarlandes

Automatische Extraktion von bilingualen Valenzwörterbüchern aus deutsch-englischen Parallelkorpora

Eine Pilotstudie

Oliver Čulo

universaar
Universitätsverlag des Saarlandes
Saarland University Press
Presses Universitaires de la Sarre

Oliver Čulo

Automatische Extraktion von bilingualen Valenzwörterbüchern aus deutsch-englischen Parallelkorpora

Eine Pilotstudie

universaar
Universitätsverlag des Saarlandes
Saarland University Press
Presses Universitaires de la Sarre

D 291

Universitätsverlag des Saarlandes
Saarland University Press
Presses Universitaires de la Sarre

Postfach 151150, 66041 Saarbrücken

ISBN 978-3-86223-034-1 gedruckte Ausgabe
ISBN 978-3-86223-035-8 Online-Ausgabe
URN urn:nbn:de:bsz:291-universaar-292

zugl.: Dissertation, Universität des Saarlandes, 2010

Projektbetreuung *universaar*: Isolde Teufel

Satz: Oliver Čulo
Umschlaggestaltung: Julian Wichert

Gedruckt auf säurefreiem Papier von Monsenstein & Vannerdat

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Meinen Eltern

Danksagung

Dass eine Doktorarbeit ein zeitaufwendiges und arbeitsreiches Unterfangen ist, dürfte eine allgemein anerkannte Wahrheit sein. Ebenso dürfte auch anerkannt sein, dass sie nie nur das Werk eines Einzelnen ist. Direkt und indirekt haben viele am Entstehen dieser Arbeit mitgewirkt, und diesen Personen sei an dieser Stelle Dank gesagt.

Zuallererst gilt mein Dank meinem Doktorvater, Prof. Dr. Johann Haller, der mich überhaupt erst dazu gebracht hat, zu promovieren. Auf diesem Weg hat er mich mit Rat und Tat begleitet, und während meiner Zeit am Lehrstuhl und am IAI habe ich sehr viel für meine Doktorarbeit und meine berufliche Zukunft gelernt! Genauso gilt mein Dank auch meiner Zweitbetreuerin, Jun.-Prof. Dr. Silvia Hansen-Schirra, für die starke inhaltliche Betreuung und die vertrauensvolle Zusammenarbeit während CroCo und darüber hinaus! Ich hatte das große Glück, gleich zwei sehr engagierte Betreuer zu haben, die immer, wirklich immer, wenn ich darum gebeten habe, „mal ein halbes Stündchen“ für mich hatten!

Den praktischen Teil der Arbeit konnte ich im Rahmen des CroCo-Projekts durchführen. Dabei hatte ich viel Unterstützung durch ein großartiges Saarbrücker Team bestehend aus Peggy Daut, Kerstin Kunz, Karin Maksymski, Mary Mondt, Stella Neumann, Prof. Dr. Erich Steiner, Marc Summkeller und einem quirligen Heer von HiWis. Danke für das schöne Gefühl, morgens (oder auch mal mittags) mit Freude ins Büro reinzukommen.

Auch wenn ich viel in Saarbrücken war, geschah die Hauptarbeit doch in meiner Zeit in Germersheim. Hier habe ich ebenfalls wunderbaren Kollegen zu danken, für eine tolle Bürogemeinschaft und entspannende Mittagessen, die die Launen des Arbeitstages vertreiben, und organisatorische sowie inhaltliche Unterstützung: Melanie Arnold, Prof. Dr. Dieter Huber, Doris Kinne, Christoph Rösener und Johanna Wismeth.

Nicht vergessen sei die Zeit als Mitarbeiter am IAI und an der Universität des Saarlandes, die meinen Berufsstart mit einem Einblick in ein breites Anwendungsfeld prägte. Hier seien für ihre Hilfsbereitschaft insbesondere erwähnt: Gernot Hensberg und Paul Schmidt vom IAI sowie Hendrik Zender vom DFKI.

Während der ganzen Zeit der Doktorarbeit hatte ich, wie in allen anderen Lebenslagen auch, meine Familie im Rücken. Mein Bruder hat immer aufmunternde Worte für mich gefunden. Meine Eltern haben sich immer für meine bestmögliche Ausbildung – und nicht nur dafür – eingesetzt, ohne Wenn und Aber. Ihnen ist daher diese Arbeit gewidmet, zum Dank für ihre Unterstützung und Liebe.

Und nicht zuletzt haben mich während der letzten Jahre viele Menschen begleitet, die ebenfalls einen Platz in meinem Herzen haben – meine Freunde innerhalb wie außerhalb von Universität und Arbeit. Alle mir wichtigen Menschen hier aufzuzählen würde den Rahmen sprengen, aber im Zusammenhang mit der Doktorarbeit seien ganz besonders die Mitglieder des L.V. an der Universität des Saarlandes sowie die Mannheimer Vereinigung des ‚ménage À 3' erwähnt.

Mannheim, im März 2010

Inhaltsverzeichnis

1 Einleitung..... 1

1.1 Einfluss der Valenztheorie..... 4
1.2 Valenz und Korpora: Einordnung in das Forschungsumfeld.....5
1.3 Struktur der Arbeit..... 6

2 Valenz.....9

2.1 Valenz in der Theorie.....9
2.1.1 Lucien Tesnière.....10
2.1.2 Die Prager Schule..... 12
2.1.3 Die Leipziger Valenzschule..... 13
2.1.4 Frame-Semantik.....15
2.1.5 Übersicht.....17
2.2 Valenz in computerlinguistischen Grammatikformalismen.....19
2.2.1 HPSG..... 20
2.2.2 LFG.....22
2.2.3 Übersicht.....24
2.3 Valenz in der Lexikografie.....25
2.3.1 Monolinguale Valenzlexikografie.....25
2.3.2 Bilinguale Valenzlexikografie..... 31
2.3.3 Übersicht.....33
2.4 Valenzextraktion aus Korpora.....35
2.4.1 Extraktionsexperimente mit monolingualen englischen Texten.....36
2.4.2 Parallele Valenzrahmenextraktion aus der Prague Czech-English Dependency Treebank..... 38
2.4.3 Kollokationen und Verbentsprechungen aus EUROPARL.....40
2.4.4 Übersicht.....41
2.5 Valenz in der maschinellen Übersetzung.....42
2.5.1 Valenz in SUSY..... 43
2.5.2 Das METAL-Leuven-Framework..... 44
2.5.3 Das rollenbasierte Modell von EUROTRA und die Interlingua von Dorr.....46
2.5.4 LFG-Modelle..... 49

2.5.5 HPSG-Modelle 50
2.5.6 Valenz-MÜ im Rahmen der PCEDT 52
2.5.7 Übersicht 54

3 Das deutsch-englische CroCo-Korpus 57

3.1 Aufbau 58
3.2 Die CroCoAPI 62
3.3 Vor- und Nachteile der Nutzung von CroCo für die vorliegende Studie 66

4 Studien zur parallelen Valenzextraktion 69

4.1 Hypothesen zur Parallelitätsannahme 69
4.2 Versuchsaufbauten 71
4.2.1 Auswertung von Alignierungslücken und -divergenzen 72
4.2.2 Extraktion und Nutzbarmachung bilingualer Valenzwörterbucheinträge 76
4.3 Ergebnisse 81
4.3.1 Ergebnisse der Vorstudien geordnet nach valenzrelevanten Kategorien 81
4.3.1.1 Empty links bei Sätzen 81
4.3.1.2 Empty links bei Einzelsätzen 82
4.3.1.3 Empty links bei grammatischen Funktionen 88
4.3.1.4 Crossing lines zwischen Wörtern und grammatischen Funktionen 96
4.3.1.5 Übersicht 104
4.3.2 Divergenzen bei Prädikatsausdrücken 106
4.3.3 Übersicht 117
4.4 Anwendung der Ergebnisse 118
4.4.1 Erweiterung der Regeln und Transferlexika von MÜ-Systemen 119
4.4.2 Ein webbasiertes Wörterbuch für die HÜ 131

5 Diskussion 137

5.1 Einordnung der Ergebnisse in das Forschungsumfeld 137
5.2 Beitrag zur theoretischen Diskussion 140
5.3 Desiderata an die zukünftige Korpusannotation 145

6 Fazit und Ausblick 147

Literaturverzeichnis..... 149

Abbildungsverzeichnis

Abbildung 2.1: Eine vereinfachte Beispiel-Merkmalstruktur für den Lexikoneintrag sehen..... 20
Abbildung 2.2: Das Head-Feature Principle der HPSG (übernommen von Oliva 2003, S.661)..... 21
Abbildung 2.3: Das Subcategorization Principle der HPSG (übernommen von Oliva 2003, S.663)..... 21
Abbildung 2.4: Die Stellung der a-structure zwischen lexikalischer Semantik und syntaktischer Struktur (adaptiert von Bresnan 2001, S.306)..... 23
Abbildung 2.5: Abbildung von a-structure auf f-structure und Korrespondenzen zwischen c-structure und f-structure (adaptiert von Bresnan 2001, S.303) 23
Abbildung 2.6: Vereinfachte LCS für das Ereignis gehen (adaptiert von Dorr 1994: 601)..... 49
Abbildung 3.1: Der Aufbau des CroCo-Korpus..... 57
Abbildung 3.2: Annotations- und Alignierungsebenen des CroCo-Korpus…59
Abbildung 3.3: Das CroCoXML stand-off Format..... 62
Abbildung 3.4: Die drei Ebenen der CroCoAPI..... 64
Abbildung 3.5: Die hierarchischen Ebenen des CoReTool..... 65
Abbildung 4.1: Empty links bei der Wortalignierung..... 73
Abbildung 4.2: Crossing lines zwischen Funktionen: von FIN nach PRED..74
Abbildung 4.3: Pseudo-Code für die Abfrage nach crossing lines zwischen Einzelsätzen und Sätzen..... 75
Abbildung 4.4: Pseudo-Code für die Extraktion und Ausgabe von grammatikalisch-funktionalen Valenzmustern..... 78
Abbildung 4.5: Ein extrahiertes Satzpaar, mit Hauptverben und VRM, aus dem Register G2E_SPEECH..... 79
Abbildung 4.6: Absolute Zahlen für Einzelsätze..... 83
Abbildung 4.7: Alignierte und nicht alignierte Einzelsätze in Prozent..... 83
Abbildung 4.8: Anteil der Funktionen, die in einem Satzpaar nur in der AS vorhanden sind, Übersetzungsrichtung englisch-deutsch..... 88
Abbildung 4.9: Anteil der Funktionen, die in einem Satzpaar nur in der AS enthalten sind, Übersetzungsrichtung deutsch-englisch..... 89
Abbildung 4.10: Anteil der Funktionen, die in einem Satzpaar nur in der ZS vorhanden sind, Übersetzungsrichtung englisch-deutsch..... 90

Abbildung 4.11: Anteil an Funktionen, die in einem Satzpaar nur in der ZS vorhanden sind, Übersetzungsrichtung deutsch-englisch.....91
Abbildung 4.12: Crossing lines zwischen Wörtern und grammatischen Funktionen..... 97
Abbildung 4.13: Häufige Kollokate von essen in Netzwerkdarstellung auf der Webseite des DWDS (Quelle: http://www.dwds.de, 24.02.2010).....134

Tabellenverzeichnis

Tabelle 2.1: Übersicht über die Bezeichnungen für Aktanten und Modifikatoren..... 18
Tabelle 2.2: Übersicht über Motivation und Inhalt verschiedener monolingualer und bilingualer Valenzwörterbücher.....34
Tabelle 2.3: Übersicht der Ansätze zur Extraktion von Valenzwörterbüchern aus Korpora.....41
Tabelle 3.1: Grammatische Funktionen in CroCo..... 60
Tabelle 3.2: Die verschiedenen Tokenisierungen von It's von MPro und TNT. 63
Tabelle 4.1: Tendenzen für typische Wanderungsbewegungen von Funktionen, von oben nach unten sinkt die Häufigkeit. Der Ausdruck dobj→subj steht z.B. dafür, dass ein direktes Objekt aus der AS in die Subjektfunktion in der ZS wechselt..... 99
Tabelle 4.2: Übersicht über Divergenzen, Ursachen und den Bezug zur Valenz mit Berücksichtigung des Englischen (EN) und des Deutschen (DE)..106
Tabelle 4.3: Auszählungsergebnis für Satzpaarkategorien mit Bezug auf die Vergleichbarkeit ihrer Prädikatsausdrücke..... 114
Tabelle 4.4: Anteil an direkten Verb-zu-Verb-Entsprechungen ohne Perspektivwechsel..... 115

Abkürzungsverzeichnis

AS	Ausgangssprache	PH	Phraseologismus
ebd.	ebenda	PKA	Prädikatsausdruck
FVG	Funktionsverbgefüge	PP	Präpositionalphrase
ggf.	gegebenenfalls	u.a.	unter anderem
HÜ	Humanübersetzung	u. a.	und andere
KP	Kopulakonstruktion	v.a.	vor allem
MÜ	maschinelle Übersetzung	z.B.	zum Beispiel
NP	Nominalphrase	ZS	Zielsprache

Notationskonventionen

kursiv, Times New Roman	Korpus- und Sprachbeispiele; bei der ersten Erwähnung: fremdsprachliche Ausdrücke, Namen (wenn nicht in Großbuchstaben)
„“	Zitate
`kursiv, Courier New`	Code-Beispiele
fett, Times New Roman	bei der ersten Erwähnung: zentrale Fachbegriffe
GROSSBUCHSTABEN	Namen (wenn nicht kursiv), Abkürzungen

1 Einleitung

Die fortschreitende Europäisierung und Globalisierung erhöht derzeit stetig den Bedarf an Übersetzungen. Neben den gedruckten Wörterbüchern wird das Internet zu einer immer wichtigeren Quelle für Übersetzungen, wie die Nutzerstatistiken beispielsweise des Online-Wörterbuchs *LEO*[1] belegen. Während Online-Wörterbücher, wie die Erfahrung zeigt, eine ähnlich hohe Qualität wie gedruckte Werke erreichen können sind Werkzeuge zur Übersetzung ganzer Texte oder Sätze höchstens für eine Rohübersetzung geeignet, die aber nicht an die Qualität einer menschlichen Übersetzung heranreicht. Ein Experiment mit drei Testsätzen aus dem Deutschen ins Englische soll diese Aussage belegen. Übersetzt wurden die Sätze mit einem der mit am häufigsten verwendeten, weil frei verfügbaren Übersetzungswerkzeuge, *Google Translate*[2]. Google verwendet eine recht aktuelle Übersetzungsstrategie in der maschinellen Übersetzung, bei der nicht Wörter, sondern Phrasen übersetzt werden. Um das Übersetzungslexikon zu erstellen, werden in Ausgangs- und Zieltexten Phrasen markiert und miteinander aligniert, aber abgesehen von der Phrasenerkennung keine weitere linguistische Information verwendet (Franz u. a. 2003; Hoang u. a. 2009).

Die Testsätze sind allesamt dem CroCo-Korpus (Neumann und Hansen-Schirra 2005), ein Textkorpus, das u.a. deutsche und englische Originale und Übersetzungen enthält, entnommen. Im Folgenden sind die deutschen Sätze und dann die *Google Translate*-Übersetzungen aufgeführt. Beispiel (1) stammt aus einer politischen Rede, Beispiel (2) aus einem Aktionärsbrief, und Beispiel (3) aus einem fiktionalen Text. Alle drei sind Extrakte aus in CroCo enthaltenen Sätzen, und die jeweiligen Musterübersetzungen sind in der jeweils dritten Zeile angegeben:

(1) *Die Unterstützung der Türkei straft diesen Mythos Lügen.*
**The support of Turkey punishes this myth lies.*
The support of Turkey lays this myth to rest.

(2) *Unter Einbeziehung von Europa belaufen sich unsere weltweiten Umsätze insgesamt auf über 3 Milliarden US-Dollar .*
**With the inclusion of Europe amount to our worldwide sales total of more than 3 billion U.S. dollars.*

1 http://dict.leo.org, Angaben zur Nutzerentwicklung unter http://dict.leo.org/pages.ende/about_de.html?lp=ende&lang=de
2 http://translate.google.de

When European sales are included, our global coverage will be more than $ 3 billion.

(3) *Mich beunruhigt schon die vom Himmel fallende Asche.*
**I have worried from the sky falling ash.*
I am already disturbed by the ash falling from the sky.

Die ersten beiden Übersetzungen erscheinen zumindest verständlich; die dritte ist ohne Kenntnis des Originals kaum interpretierbar. Allerdings ist keine der drei Übersetzungen kritiklos annehmbar. In der ersten Übersetzung wurde nicht erkannt, dass das Prädikat nicht nur *strafen*, sondern auch *Lügen* umfasst (*etw. Lügen strafen*), und der gesamte Ausdruck eher mit der englischen Wendung *lay ... to rest* zu übersetzen wäre. Bei der zweiten Übersetzung ist nicht deutlich zu erkennen, welches syntaktische Element das Subjekt und welches das Objekt ist. Das Subjekt *unsere weltweiten Umsätze* steht im Deutschen hinter dem finiten Verb. Das englische Subjekt *our worldwide sales* müsste vor dem finiten Verb platziert werden. Man kann davon ausgehen, dass die Übersetzungsprozedur das Subjekt gar nicht erkannt hat. Die dritte Übersetzung ist als nicht erfolgreich zu werten. Die Rolle des Objekts *mich* als erfahrendes Individuum der Beunruhigung wurde erkannt und dies als Subjekt *I* in *I have worried* übersetzt; allerdings ist der Anschluss des deutschen syntaktischen Subjekts *die vom Himmel fallende Asche* als Objekt im Englischen nicht korrekt mit *about* (*I have worried about...*) realisiert. Auch wurde die eingebettete Adjektivphrase *vom Himmel fallende* nicht korrekt als abhängig vom Nomen *ash* übersetzt.

In allen drei Fällen spielt das Verb eine zentrale Rolle für die korrekte Übersetzung. Im ersten Beispiel wurde das Prädikat, ein Phraseologismus, nicht in seinem vollen Umfang erkannt und daher nicht korrekt übersetzt. Im zweiten und dritten Beispiel dagegen wurden syntaktische Vorgaben des Verbs – das Subjekt (meist) vor dem finiten Verb, bzw. Anschluss des Objekts mit *about* – nicht korrekt umgesetzt.

Dass das Verb syntaktisch wie semantisch eine zentrale Stellung im Satz einnimmt, ist in der Linguistik allgemein anerkannt. Ágel (2000:7) formuliert diese zentrale Stellung wie folgt:

„Wörter – vor allem Verben – prädeterminieren die Satzstruktur."

Diese Eigenschaft wird als **Valenz** des Verbs bezeichnet.

Alle eingangs verwendeten Beispielsätze sind im CroCo-Korpus mit ihrer jeweiligen Übersetzung bzw. dem entsprechenden Originalsatz enthalten. Es liegt nahe, solche Übersetzungskorpora dafür zu verwenden, um Valenzeigenschaften von Verben sowie Valenzeigenschaften im Kontext von Übersetzungen zu studieren. Für das Gebiet der statistischen maschinellen Übersetzung, für das große Datenmengen nötig sind, um Übersetzungsregeln bzw. -modelle zu erlernen, ist es wünschenswert, die Valenzeigenschaften von Verben im Kontext von Original und Übersetzung möglichst automatisch zu extrahieren und in Form von elektronischen Valenzwörterbüchern festzuhalten. Für den Humanübersetzer oder Fremdsprachenlerner können Valenzwörterbücher ebenfalls eine große Hilfe sein. Weiß man einmal, was man sagen möchte, ist oft das genaue Wie schwierig. Ein Valenzwörterbuch liefert zu einem Verbeintrag idealerweise Informationen über die korrekten Fälle wie auch über den semantischen Typ der jeweiligen Verbbegleiter.

In der Erstellung aktueller Wörterbücher, insbesondere von im Internet verfügbaren Ressourcen wie z.B. dem *Digitalen Wörterbuch der deutschen Sprache* (DWDS)[3] oder dem Leipziger Wortschatzprojekt[4], aber auch von gedruckten Werken, z.B. dem englischen Valenzwörterbuch von Herbst et al. (Herbst u. a. 2004a), haben sich zuletzt Korpusmethoden etabliert. Diese korpusgestützt erstellten Wörterbücher warten neben sehr aktuellen Daten – dank schneller maschineller Auswertung auch neuester Texte – mit einer hohen Anzahl von Verwendungsbeispielen sowie zusätzlichen Informationen wie etwa typischen Kollokationen auf, wie sie in der Anzahl in traditionellen Wörterbüchern kaum vorkommen.

Die genannten Ressourcen basieren auf einsprachigen Korpora und sind daher einsprachige Wörterbücher. Die Verfügbarkeit von immer mehr mehrsprachigen Korpora, die sowohl im Rahmen der maschinellen Übersetzung (MÜ) als auch der Humanübersetzung (HÜ) eingesetzt werden, sollte es aber bald ermöglichen, für diverse Sprachenpaare auf automatischem Wege (Valenz-)Wörterbücher zu erstellen.

Die hier vorliegende Arbeit untersucht daher auf Basis des CroCo-Korpus die Frage, inwiefern Valenzwörterbücher aus einem Korpus von Originalen und Übersetzungen extrahiert werden können. Dazu werden Fälle untersucht, in denen es zwischen Originalen und Übersetzungen zu syntaktischen Divergenzen zwischen Ausgangssprache (AS) und Zielsprache (ZS) kommt. Die Divergenzen werden durch moderne Methoden der computergestützten Korpuslinguistik aufgefunden und mit Blick auf valenzrelevante Unterschiede analy-

3 http://www.dwds.de
4 http://wortschatz.uni-leipzig.de

siert und diskutiert. Dabei wird darauf eingegangen, inwiefern sich die Korpusdaten zur Extraktion von Valenzwörterbüchern eignen, welche Erkenntnisse aus den Korpusdaten zu gewinnen sind und wie die Korpusaufbereitung sowie die Extraktion aussehen müssen, um valenzrelevante Divergenzen adäquat beschreiben zu können.

1.1 Einfluss der Valenztheorie

Die Valenz ist, seit ihrer Einführung als abgeschlossene Theorie durch Lucien Tesnières posthum erschienenes Werk „Éléments de syntaxe structurale“ (1959), eines der zentralen Themen der linguistischen Forschung. Dies ist belegt durch eine Vielzahl von Arbeiten zur Valenztheorie (vgl. Abschnitt 2.1) sowie durch die Adaption des Valenzkonzepts durch verschiedene linguistische Teildisziplinen: Grammatik, kontrastive Linguistik, (Fremd-)Spracherwerb, Lexikografie, computerlinguistische Grammatiktheorien und maschinelle Übersetzung, um die wichtigsten zu nennen. Diese Bandbreite und Vielseitigkeit legt es nahe, mittels neuerer korpuslinguistischer Methoden das Phänomen der Valenz zu untersuchen.

Valenz wurde in ihrer ursprünglichen Konzeption v.a. als Eigenschaft des Verbs verstanden; diese Arbeit beschäftigt sich denn auch ausschließlich mit der Verbvalenz. Die Faszination der Valenz geht davon aus, dass sie eine Eigenschaft ist, die Wechselwirkungen mit verschiedenen linguistischen Ebenen hat, sei es Syntax, Semantik – oder sogar Pragmatik.

Die Tatsache, dass die Valenz eines Verbs quasi den Satzbauplan vorbestimmt, und zudem noch die Wörter, die in dem Satz vorkommen können, einschränkt, macht das Valenzkonzept zu einem Konzept mit großem Wert für die Lexikografie und die maschinelle Übersetzung (MÜ). Diese Arbeit erforscht die Möglichkeit der Extraktion eines parallelen Valenzwörterbuchs zwar ausschließlich am CroCo-Korpus, dennoch soll sie nicht nur als korpusspezifische Arbeit gesehen werden. Sie soll Untersuchungsmethoden aufzeigen sowie allgemeine Erkenntnisse zur Valenz im Kontext der MÜ beitragen. Das CroCo-Korpus wurde u.a. zum Zweck der Valenzextraktion vom Autor der vorliegenden Arbeit aufwändig aufbereitet. Auf der Grundlage von Erfahrungswerten aus dieser Aufbereitung werden Schwächen – mit Bezug auf die hier gestellte Forschungsfrage – insbesondere der Korpusannotation aufgezeigt und notwendige Änderungen als Desiderata an die Annotation formuliert. Diese können als Blaupause für ähnlich ausgerichtete Projekte dienen.[5]

5 Tatsächlich wurde von der Johannes Gutenberg-Universität Mainz, die an der zweiten Phase des CroCo-Projekts beteiligt war, ein Nachfolgeprojekt finanziert, das sich mit eben diesen

1.2 Valenz und Korpora: Einordnung in das Forschungsumfeld

Die ersten großen elektronischen Korpora, wie z.B. für das Englische das BROWN-Korpus (Francis und Kučera 1979) und das LOB-Korpus (Johansson u. a. 1986), für das Deutsche COSMAS I[6], das *digitale Wörterbuch der deutschen Sprache* DWDS (Geyken 2007) oder die TIGER-Baumbank (Brants u. a. 2002), waren rein monolingual ausgerichtet. In den letzten Jahren zeigt sich die Tendenz, Korpora verstärkt multilingual aufzubauen, wie etwa das in der vorliegenden Arbeit verwendete deutsch-englische CroCo-Korpus, das englisch-französische HANSARD-Korpus (Roukos u. a. 1997), bestehend aus Redemitschriften des kanadischen Parlaments, das *Oslo Multilingual Corpus* (Johansson 2000), die *Prague Czech-English Dependency Treebank* (PCEDT, Čmejrek u. a. 2004) oder das EUROPARL-Korpus (Koehn 2005), bestehend aus Redemitschriften des EU-Parlaments. **Parallele Korpora –** als solche werden im Folgenden Korpora aus Originalen und deren Übersetzungen bezeichnet (nach Baker 1996; Granger 2003) – eröffnen die Möglichkeit, Valenz einzelsprachlich und sprachvergleichend auf breiter Basis zu untersuchen. Prinzipiell ist es auch denkbar, dafür **Vergleichskorpora**, also mehrsprachige Korpora mit Textsammlungen aus den gleichen Registern für jede Sprache, zu verwenden. Da in parallelen Korpora allerdings eine Alignierung z.B. auf Satz- und Wortebene möglich ist, ist es einfacher, Vergleichsgrößen (z.B. durch satzweisen Valenzvergleich für das jeweilige Hauptverb) zu finden.

Das CroCo-Korpus wurde primär erstellt, um Explizierung in Übersetzungen zu studieren. Die Explizierung gehört zu den sogenannten **translation universals**, also universellen Übersetzungseigenschaften, wie sie von Baker (1993; 1995) postuliert werden.[7] CroCo ist ein Übersetzungskorpus mit englischen und deutschen Originalen und den jeweiligen Übersetzungen (Aufbau und weitere Charakteristika vgl. Kapitel 3). Das Korpus wurde darüber hinaus für Studien u.a. in verschiedenen Feldern wie etwa dem kontrastiven Sprachvergleich (z.B. Hansen-Schirra u. a. 2007), der Untersuchung von registertypischen Eigenschaften (z.B. Neumann 2009) und verschiedenen Übersetzungseigenschaften und -phänomenen jenseits der Explizierung (z.B. Kunz 2007; Čulo u. a. 2008b) eingesetzt. Daneben wurde das Korpus in com-

Schwächen beschäftigt.

6 http://www.ids-mannheim.de/kl/projekte/cosmas_I/

7 Die Frage, inwiefern diese Eigenschaften tatsächlich universell sind oder nicht, soll an dieser Stelle ausgeklammert werden, da sie für das hier bearbeitete Thema nicht von höchster Relevanz ist. In Kapitel 5 wird aber der mögliche Einfluss von Übersetzungseigenschaften auf die „parallele“ Valenzextraktion nochmals aufgegriffen, allerdings wiederum ohne Berücksichtigung der Universalitätsfrage.

puterlinguistischen Kontexten genutzt, z.B. als Goldstandard für computerlinguistische Werkzeuge oder als Trainingsbasis für die maschinelle Übersetzung (Hansen-Schirra u. a. [erscheint]: Kap. 10).

Diese Arbeit reiht sich – sowohl mit Bezug auf CroCo sowie auf das aktuelle korpuslinguistische Umfeld – mit einem linguistisch sowie MÜ-relevanten Thema in das Forschungsfeld ein: dem Versuch, korpusgestützt ein maschinenlesbares deutsch-englisches Valenzwörterbuch zu extrahieren. Einen Überblick über das aktuelle Forschungsfeld sowie relevante Vorarbeiten gibt das Kapitel 2.

1.3 Struktur der Arbeit

Die Arbeit gliedert sich in folgende Abschnitte. In Kapitel 2 erfolgt ein Überblick über bisherige theoretische Beiträge zur Valenz. Darauf aufbauend werden der in der Arbeit verwendete Valenzbegriff eingegrenzt und die Integration von Valenz in computerlinguistische Formalismen, jüngere korpusgestützte Valenzforschungsprojekte sowie die praktische Anwendung von Valenz in einigen ausgewählten MÜ-Systemen beschrieben. Das CroCo-Korpus, auf dem die Untersuchungen in dieser Arbeit basieren, wird in Kapitel 3 näher vorgestellt. Dazu wird der Aufbau des Korpus und die technischen Voraussetzungen zu dessen Nutzung beschrieben sowie eine Einschätzung darüber abgegeben, welche Stärken und Schwächen das Korpus mit Blick auf den Untersuchungsgegenstand aufweist. In Kapitel 4 werden Vorstudien beschrieben, die das Korpus zunächst daraufhin prüfen, inwiefern man von einer allgemeinen Parallelität sprachlicher Strukturen im Deutschen und Englischen ausgehen kann. Die Methodik der Vorstudien greift dabei auf aus der Translationswissenschaft bekannte Phänomene wie z.B. Null-zu-Eins-Entsprechungen (Koller 2001) zurück und setzt diese Konzepte in Form computergestützter Abfragen um. Diese Studien werden durch eine Studie zur Extraktion bilingualer Valenzwörterbucheinträge und durch einen anschließenden Einblick in mögliche Probleme bei der Extraktion aufgrund struktureller Divergenzen zwischen dem Deutschen und dem Englischen ergänzt. Anschließend wird ein Ausblick auf mögliche Anwendungsszenarien für extrahierte Valenzwörterbücher – als Erweiterung der Lexika bestehender MÜ-Systeme oder als dynamisches, webbasiertes Wörterbuch für Humanübersetzer – gegeben. In Kapitel 5 werden die Beobachtungen der Studien aus Kapitel 4 diskutiert, deren Ergebnisse in aktuelle theoretische Überlegungen eingeordnet sowie daraus Schlüsse für zukünftige Forschungsfragen gezogen. Dabei wird auch kritisch auf die angewandten Methoden und die erzielten Ergebnisse eingegan-

gen und insbesondere die Problematik der Verwendung von Parallelkorpora in dem Forschungsbereich der vorliegenden Arbeit sowie in angrenzenden Forschungsbereichen aufgegriffen. Abschließend werden in Kapitel 6 die Ergebnisse der Arbeit zusammengefasst und ein Ausblick auf mögliche weiterführende Forschungsarbeiten gegeben.

2 Valenz

In diesem Kapitel wird das Thema Valenz aus aufeinander aufbauenden, für diese Arbeit relevanten Blickwinkeln betrachtet. Zunächst wird ein kurzer Überblick über verschiedene Valenztheorien gegeben (Abschnitt 2.1), ausgehend von der grundlegenden Valenztheorie von Lucien Tesnière. Um Valenz für die MÜ nutzbar zu machen, bedarf es ihrer Integration in die computerlinguistischen Grammatikformalismen. Von diesen werden die Valenzmechanismen von zwei der am häufigsten verwendeten Formalismen, HPSG und LFG, vorgestellt (Abschnitt 2.2). Da für die MÜ – wie auch für viele andere computerlinguistische Anwendungen – Lexika mit Valenzinformation notwendig sind, wird zunächst darauf eingegangen, welche Rolle die Valenz in der Lexikografie spielt (Abschnitt 2.3). Anschließend werden in Abschnitt Fehler: Referenz nicht gefunden neuere Ansätze zur Valenzlexikografie mittels Korpusdaten vorgestellt. Erst danach wird ein praktischer Nutzen von Valenzwörterbüchern, nämlich ihr Einsatz in der MÜ, erörtert (Abschnitt 2.5).

2.1 Valenz in der Theorie

Die Valenztheorie ist eine vergleichsweise junge Theorie. Die Idee, dass bestimmte Wörter sich nur mit bestimmten anderen Wörtern verbinden und nach einer bestimmten Anzahl von „Begleitern“ verlangen, wurde zwar schon in vielen Grammatiken vor Tesnières Éléments (1959) formuliert. So schreibt beispielsweise schon Bühler (1999 [1934]: S.173):

> „Es bestehen in jeder Sprache Wahlverwandtschaften; das Adverb sucht sein Verbum und ähnlich die anderen. Das läßt sich auch so ausdrücken, daß die Wörter einer bestimmten Wortklasse eine oder mehrere Leerstellen um sich herum eröffnen, die durch Wörter bestimmter anderer Wortklassen ausgefüllt werden müssen.“

Bereits hier wird die Idee der „Leerstellen“ formuliert, die mit bestimmten Wortarten zu füllen sind. Bühler selbst führt sie zurück bis hin auf erste ähnliche Konzepte in der mittelalterlichen Scholastik (vgl. auch Darstellungen in Helbig und Schenkel 1969:10ff.; Ágel 2000:8ff.). Formalisiert und in einem abgeschlossenen theoretischen Rahmen eingeführt wurde die Valenztheorie allerdings erst von Tesnière. Seither ist das Valenzkonzept in vielen linguistischen Feldern vertreten, vom Spracherwerb über die Lexikografie bis hin zur Syntax, Semantik und in diversen Grammatiktheorien. Viele der zentralen Fragen der Valenztheorie, z.B. wie Aktanten und Modifikatoren genau zu un-

terscheiden sind, und ob es denn überhaupt eine klare Grenze zwischen ihnen gibt, oder wie obligatorisch oder fakultativ bestimmte Elemente sind, sind bis heute nicht abschließend beantwortet. War man zuvor auf jahrelange, durch mühsame Handarbeit bestimmte Valenzforschung angewiesen, so ermöglichen es die heute verfügbaren Korpora in Kombination mit sprachtechnologischen Mitteln, Valenz in großem Maße an authentischen Daten anstatt an konstruierten Beispielen zu studieren. Inwiefern sich allerdings von der Erforschung konkreter Valenzrealisierungen in Korpora auf die Beantwortung theoretischerer Fragen abstrahieren lässt, wird sich im Verlauf der Untersuchung zeigen.

In der deutschen Linguistik wurde die Valenz in verschiedenen Werken zum Gegenstand der Forschung und Diskussion, beispielsweise in (Helbig und Schenkel 1969; Engel und Schumacher 1976; Engel 1977; Schumacher 1986; Ágel 2000; Ágel u. a. 2003); ebenso in der weiteren europäischen Linguistik, etwa der Prager Schule (Panevová 1975; Sgall u. a. 1986). In der anglophonen Linguistik wurde Valenz etwa in die Kasusgrammatik von Fillmore (Fillmore 1968) und in deren Weiterentwicklung, der Frame-Semantik (Fillmore 1977) inkorporiert. Erst später erfolgte eine breitere Adaption durch weitere Autoren (Emons 1978; Allerton 1982; Herbst u. a. 2004a).

Die unterschiedlich starke Integration des Valenzkonzepts in die deutsche und englische Linguistik erkennt man auch an deren Stellenwert in den großen Grammatiken der jeweiligen Sprachen. Während z.B. die deutsche Grammatik von Zifonun u. a. (1997) Valenz groß und breit in ihrer syntaktischen, semantischen und in Ansätzen sogar ihrer pragmatischen Dimension anspricht, beschreibt die Longman-Grammatik von Biber u. a. (2000) Valenz nur im Sinne von Satzbauplänen, die ein Verb liefert.

Im nächsten Abschnitt werden die für die vorliegende Arbeit wichtigsten Konzepte aus den Éléments zusammengefasst. Danach wird auf einige einflussreiche Theorien, die auf der Valenzidee von Tesnière basieren, eingegangen: die Prager Schule, die Beiträge zur Valenz von Gerhard Helbig und die Frame-Semantik von Charles Fillmore. Der vielseitigen Adaption der Valenzidee steht eine wenig einheitliche Terminologie gegenüber. Im Anschluss an diesen Überblick wird im Rahmen einer Übersicht die in dieser Arbeit verwendete Terminologie definiert.

2.1.1 Lucien Tesnière

Lucien Tesnière arbeitete seinerzeit an einer neuen Art strukturalistischer Grammatik mit dem Ziel, die Fremdsprachenlehre mit seiner Ansicht nach

eingängigeren Methoden zu modernisieren. Dafür schuf er zwei sehr einflussreiche neue Konzepte der Linguistik: die Dependenzgrammatik und die Valenzidee, die heutzutage beide nicht mehr aus der Sprachwissenschaft wegzudenken sind und inzwischen zu den zentralen Forschungsfeldern der Linguistik gehören.

Tesnière (1959) spricht davon, dass die Bedeutung des Verbs immer ein „petit drame" sei, ein Spiel, in dem abhängig vom Verb eine bestimmte Zahl von Teilnehmern – man könnte also sagen Mitspielern – auftaucht. Tesnière benennt diese Mitspieler als **actants** ‚Aktanten'. Diese Aktanten werden nummeriert. So ist im aktivischen Satz i.d.R. das Subjekt der **prime actant**, das Akkusativobjekt der **second actant**, das Dativ- oder Genitivobjekt der **tiers actant**. Der erste Aktant ist der Handelnde, der zweite Aktant erduldet die Handlung, und der dritte Aktant ist der durch die Handlung Begünstigte (z.B. *Sie gab ihr das Buch*). Im passivischen Satz ist weiterhin das Subjekt der erste Aktant; die Zuordnung zwischen erstem Aktant und zweitem Aktant als Handelnder bzw. Erleidender sind allerdings vertauscht. Damit zeigt sich, dass Tesnières Kategorien eher syntaktisch als semantisch motiviert sind, semantische Aspekte dennoch im nötigen Maße berücksichtigt sind. Dass Tesnière vor allem eine syntaktische Agenda hatte, sieht man auch in einem anderen Punkt: Mit der Gleichstellung von Subjekt und Objekten als Aktanten hebt Tesnière die Sonderstellung des Subjekts auf, ein explizites Ziel seiner Ausführungen.

Die Umstände des „petit drame" können durch **circonstants** ‚Modifikatoren' näher beschrieben werden; dazu gehören z.B. Angaben des Ortes und der Zeit. Ein Aktant kann jeweils nur einmal in einem Satz auftreten; so kann es z.B. nur einen ersten Aktanten in einem Satz geben.[8] Zudem kann man nicht jede Art von Aktant mit jedem beliebigen Verb kombinieren; so ist z.B. beim intransitiven Verb *schlafen* nur ein belebter Aktant möglich (der Schlafende). Modifikatoren können dagegen mit jedem Verb und in beliebig häufiger Zahl kombiniert werden, allerdings auch hier nicht immer alle Arten von Modifikatoren mit allen Verben (**Er schlief langsam*).

Tesnières Beitrag zur modernen Linguistik ist unbestritten, und seine Theorie wurde verschiedentlich inkorporiert und weiterentwickelt, wie in den nächsten Abschnitten an einigen Beispielen ausgeführt wird.

8 Dies ist im Sinn von syntaktischen Strukturen zu verstehen. Natürlich kann an dieser Stelle z.B. eine koordinierte Struktur wie „Du und ich" stehen, aber auch dies ist als *eine* syntaktische Struktur zu verstehen.

2.1.2 Die Prager Schule

Zu den vehementesten Verfechtern der Dependenz- und Valenztheorien kann man die Linguisten der Prager Schule zählen. Sie übernahmen die Tesnière'schen Ideen sehr früh. Die in den 50er Jahren des 20. Jahrhunderts entwickelte Theorie des **sentence-pattern model** (vgl. Daneš 1994) beschreibt den Satz auf zwei Ebenen: auf der grammatischen Ebene mit **grammatical sentence patterns** und auf der semantischen Ebene mit **semantic sentence patterns**. Diese Satzmuster beschreiben die semantisch-syntaktische Struktur eines Satzes, indem sie die Abhängigkeit der syntaktischen Elemente vom Verb ausgehend modellieren. Beeinflusst von den Arbeiten u.a. von Tesnière, werden die Theorien der Prager Schule von Sgall, Panevová und Hajičová in der Theorie der **Functional Generative Description** (FGD; Sgall u. a. 1986) gebündelt. Zentral ist in dieser Theorie die Frage nach dem Zusammenhang zwischen Semantik und Grammatik. Neben Themen wie **topic-focus**-Struktur spielt in der FGD die Valenz eine große Rolle.

Panevová (1975; 1994) entwickelt die Grundlagen der Prager Valenztheorie, die eine enge Verknüpfung mit der Dependenzgrammatik eingeht. Die Ebene der **Tektogrammatik** wird als Ebene unterhalb der Oberflächensyntax und der sog. **analytischen Ebene** eingeführt. Auf der analytischen Ebene werden die syntaktischen Abhängigkeiten zwischen den Wörtern des Satzes dependenziell dargestellt. Auf der tektogrammatischen Ebene wird diese Darstellung um verschiedene Punkte erweitert. Funktionswörter werden auf dieser Ebene nicht mehr als eigene Knoten dargestellt, sondern entweder als Merkmale der übrigen Dependenzknoten oder, ebenso wie die tiefen syntaktischen Rollen des Satzes, durch **Funktoren** ausgedrückt. So erscheint z.B. ein Artikel als Definitheitsmerkmal in der Merkmalstruktur eines Nomenknotens, auf den sich der Artikel bezieht; ein Nomen, das für einen Ort steht und typischerweise mit dem tschechischen *v* ‚in' angebunden wäre, erhält den Funktor *Locative*, woraufhin die Präposition „gelöscht" wird. Die Dependenten eines Regenten werden alle mit einem Funktor markiert, so ist z.B. das grammatikalische Subjekt eines Satzes zugleich auch der *Actor* des semantischen Hauptverbs. Die Funktoren Actor, *Patient*, *Addressee*, *Origin* und *Effected object* zählen zu den **inner participants**, die nur einmal in einem Satz auftreten können. Die übrigen Funktoren wie etwa *TWhen* für den Zeitpunkt oder *Manner* für Art und Weise, in der eine Handlung durchgeführt wird, gehören zu den **free modifiers**, die „frei" sind, also in beliebiger Anzahl in einem Satz und prinzipiell mit jedem Verb vorkommen können.

Während die Prager Valenztheorie sich durchaus auch auf kognitive Hintergründe der Sprache stützt, wird deutlich zwischen linguistischen Ausprägungen der Sprache aufgrund kognitiver Hintergründe und dem Weltwissen unterschieden. Daher wird z.B. der Actor eines Verbs nicht in Unterkategorien wie *Theme*, *Instrument* oder *Experiencer* aufgeteilt wie in den Kasustheorien; genaue, rein linguistische Anhaltspunkte für eine solche Aufteilung, so argumentiert Panevová, gebe es nämlich nicht. Ein Fragetest stellt mittels (Un)Beantwortbarkeit einer Nachfrage nach einer Verbergänzung fest, welche Elemente mit einem Verb kognitiv obligatorisch sind. Dies dient wiederum als Kriterium zur Aufteilung in obligatorische und fakultative Funktoren eines Verbs.

Neben dem Funktor enthalten Wortknoten in der tektogrammatischen Ebene auch **Grammateme**, d.h. semantisch bestimmte morphologische Kategorien wie etwa Zeitform oder Aspekt eines Verbs, und **Formeme**, d.h. morpho-syntaktische Kategorien wie etwa der Kasus. Diese Informationen dienen der Abbildung in die bzw. aus der analytischen und oberflächensyntaktischen Ebene. Eine Abbildung alleine anhand der Funktoren wäre gerade im morphologisch sehr reichhaltigen Tschechisch wenig aussichtsreich.

Auf Basis der Valenztheorie der FGD sowie der Annotation der *Prague Dependency Treebank* (PDT) wurde vor Kurzem ein elektronisches Valenzwörterbuch des Tschechischen namens **VALLEX**[9] erstellt (Žabokrtský 2005). Der erste Eintrag des Verb *vidět* 'sehen' sieht dabei wie folgt aus:

$$vidět\text{: } ACT_1^{obl}\ \ PAT_4^{obl}\ \ MANN^{typ}\ \ LOC^{typ}$$

vidět hat also zwei obligatorische Begleiter (Superskript *obl*), einen Actor im Nominativ (Subskript *1*) und einen Patient im Akkusativ (Subskript *4*). Typischerweise (Superskript *typ*), wenn auch nicht obligatorisch, treten Begleiter wie Manner (wie in *gut sehen*) und Locative (wie in *jemanden auf der Straße sehen*) auf.

2.1.3 Die Leipziger Valenzschule

Im deutschsprachigen Raum ist Gerhard Helbig einer der am engsten mit der Valenztheorie verknüpften Linguisten. Im Umfeld des Bibliographischen Instituts Leipzig wurden von ihm sowie von einigen seiner Kollegen zahlreiche Arbeiten zur Valenz der Verben, sowie einige zur Valenz der Nomen und der Adjektive veröffentlicht. Dazu gehören das erste deutsche Valenzlexikon der

9 http://ufal.mff.cuni.cz/vallex/

Verben (Helbig und Schenkel 1969), der Nomen (Sommerfeldt und Schreiber 1977a) und der Adjektive (Sommerfeldt und Schreiber 1977b). Daher soll auf diese Arbeiten in der hier vorliegenden Arbeit insgesamt als „Leipziger Schule“ Bezug genommen werden, wenn auch die Theorie später hauptsächlich von Helbig ausgearbeitet wurde.

In seinem Band „Probleme der Valenz- und Kasustheorie“ (1992) legt Helbig zunächst seine bereits in (Helbig und Schenkel 1969) dargestellte und später mehrfach aktualisierte Theorie dar, in der er zunächst eine Unterscheidung von drei Valenzarten unternimmt: **semantische**, **logische** und **syntaktische Valenz**. Die semantische Beschreibung eines Prädikats ist eine komplexe Struktur, die alle Argumente eines Verbs beschreibt und hierarchisiert. Dabei werden durchaus auch Prozesse beschrieben, indem z.B. eine Wandlung eines Zustands (z.B. die Veränderung der Position eines Objekts) durch zwei Argumente dargestellt wird: das Objekt am Beginn und das am Ende des Zustands. Die logische Valenz ist der Ausdruck der konzeptuell unterliegenden **Prädikat-Argument-Struktur**, etwa in folgender Formel:

lesen(x,y)

Sie besagt, dass es ein Prädikat *lesen* gibt, das zwei Argumente bindet. Diese Darstellung ist eine Vereinfachung bzw. Verflachung der semantischen Valenz. In der logischen Valenz werden z.B. bei einer Prozessbeschreibung die Argumente des Anfangs- und Endzustands häufig in ein Argument kollabiert, da es sich ja häufig um ein und dasselbe Objekt handelt, das z.B. nur eine Positionsveränderung erlebt: Im Satz *Er trug das Buch ins andere Zimmer* gibt es ein Objekt, das aufgehoben wird, und eines, das an anderer Stelle wieder abgelegt wird – aber beide Male handelt es sich um ein und dasselbe Objekt.

Die syntaktische Valenz orientiert sich zwar an der semantischen und logischen Ebene mit Bezug auf Auswahl und Anzahl der syntaktischen Argumente, ist aber eine weitere Vereinfachung. Sie drückt nicht unbedingt alle Argumente der logischen oder semantischen Valenz aus und ist zusätzlich dadurch gekennzeichnet, dass unter bestimmten Bedingungen Argumente in der Realisierung wegfallen können. In der syntaktischen Valenz spricht Helbig von **Angaben**, die den Tesnière'schen actants entsprechen, und von **Ergänzungen**, die Tenières circonstants gleichkommen.

Helbig (ebd.) widmet sich nach der Darlegung der eigenen Theorie einigen zentralen Fragen der Valenztheorie, die noch heute nicht abschließend beantwortet sind, so z.B.

- wie genau Angaben und Ergänzungen voneinander zu unterscheiden sind,

- wie genau zu differenzieren ist, welche Elemente obligatorisch und welche fakultativ sind oder
- wie genau Valenz einzuteilen ist, d.h. wie beispielsweise pragmatische Faktoren gegenüber semantischer und syntaktischer Valenz einzuordnen sind.

Die Valenztheorie ist, wie auch Helbig feststellt, zwar eine vergleichsweise junge Theorie, spielt aber seit ihrem Auftreten in der linguistischen Forschung eine große Rolle. Während einige der Fragen, so z.B. die Frage nach der Unterscheidung von Angaben und Ergänzungen, eher theoretisch zu beantworten sind, wendet sich die aktuelle, korpusgestützte Forschung allerdings zunächst der Frage nach der konkreten Ausprägung von Valenzen zu.

Helbigs Beiträge zur Valenz beschränken sich aber nicht auf die Theorie, sondern haben in Form des ersten deutschen Valenzwörterbuchs Gestalt angenommen, das in Abschnitt 2.3.1 vorgestellt wird.

2.1.4 Frame-Semantik

Die Frame-Semantik (anfänglich auch *Scene-and-Frames*-Semantik) von Fillmore (Fillmore 1977; Fillmore 1982; Fillmore 1985) bezieht in ihrer Theorie vom Sprachgebrauch die „psychologische Realität" eines Menschen mit ein. Die Theorie ist eine Erweiterung der von Fillmore (1968) vorgestellten Kasustheorie. Da auch die Kasustheorie ihre Überlegungen auf tiefere, psychologische Sprachebenen gründen muss[10], erscheint die Frame-Semantik als natürliche Weiterentwicklung.

Die Frame-Semantik geht davon aus, dass ein Mensch sowohl bei der Produktion als auch der Rezeption von sprachlichen Symbolen immer mit einem **Inventar an Erfahrungen** arbeitet. Diese Erfahrungen über Szenen, die sich in der realen Welt abgespielt haben, in Verbindung mit der Perspektive, die man bei der Betrachtung einer bestimmten **Szene** einnimmt, beeinflussen sowohl die Auswahl von Lexemen als auch die Auswahl der Anzahl und des Typs der sprachlichen Elemente, die Teilnehmer einer (psychologisch) „realen" Szene repräsentieren. Die Frame-Semantik baut zusätzlich auf dem Konzept der **Prototypensemantik** auf. Im Gegensatz zur Feature-Semantik (Katz und Fodor 1963), die konkrete und abstrakte Objekte mit einem festen Inventar an semantischen Eigenschaften (wie z.B. Belebtheit, Form, Farbe ...) beschreiben will und sich damit häufig als recht unflexibel erweist[11], so geht die Prototypensemantik von typischen Vertretern einer Entität aus, deren Eigen-

10 Einen Aktanten z.B. kann man nur dann begründen, wenn man von der Oberflächenform des Satzes abstrahiert und sich auf die Wahrnehmungsperspektive des Menschen rückbezieht.

11 Ist eine noch grüne – weil unreife – Zitrone denn eine Zitrone?

schaften in der konkreten Realisierung allerdings variieren können. Man denke dabei an einen Stuhl, der auch dann noch ein Stuhl ist, wenn es sich um einen fünfbeinigen Designerstuhl handelt (statt mit den üblichen vier Beinen).

Mit Bezug auf die prototypische Szene, die in der linguistischen Theorie der Frame-Semantik eine große Rolle spielt, lässt sich der prototypische Charakter sehr gut anhand des häufig zitierten Verkaufsbeispiels erläutern. Jeder Mensch, der ein Erfahrungsinventar innerhalb einer Gesellschaft besitzt, in der Waren mittels Geld gehandelt werden, versteht, dass in einem (Ver)Kaufsprozess eine zu (ver)kaufende Ware, ein Käufer, ein Verkäufer, ein bestimmter Geldbetrag etc. eine Rolle spielen. Je nach Perspektive führt dies zu unterschiedlichen linguistischen Realisierungen. Der Verkaufsprozess kann mit einem Satz wie *Der Verkäufer verkaufte dem Kunden ein Brot* beschrieben werden. Hier steht der Verkaufsprozess im Vordergrund; daher auch die Auswahl des Verblexems *verkaufen* und des *Verkäufers* als dem handelnden Teilnehmer. Der Geldbetrag selbst spielt im vorangehenden Beispiel keine Rolle, anders als im Satz *Der Käufer bezahlte für das Brot 3 €*; hier steht der Prozess des Bezahlens, d.h. des Geldtransfers, im Vordergrund, weswegen der *Käufer* als handelnder Teilnehmer realisiert ist und der Betrag ebenfalls genannt wird.

In der Frame-Semantik ist der Einfluss des Kognitiven stärker hervorgehoben als in anderen Valenztheorien, die sich im Allgemeinen eher mit konkret manifestierten sprachlichen Phänomenen befassen. Konsequenterweise verwendet die Frame-Semantik keinen expliziten syntaktischen Formalismus, auch wenn es eine starke Verbindung zwischen der **Construction Grammar** und der Frame-Semantik gibt (Petruck 1996).

Die Frame-Semantik wurde auch für die Translationswissenschaft nutzbar gemacht, z.B. durch Vannerem und Snell-Hornby (1986) oder Kußmaul (2010). Deren Ansätze beschäftigen sich mit den kognitiven Dimensionen des Übersetzens, u.a. mit dem Vorgang des Verstehens des Ausgangstextes, einer zentralen Voraussetzung für eine erfolgreiche Übersetzung. Die Frame-Semantik zeigt dabei auf, wie die Verbindung von kognitivem Inhalt einer Botschaft, d.h. Erkennen und Verstehen einer **scene** oder auch **Situation**, und deren sprachlicher Form **frame**, d.h. die Auswahl sprachlicher Zeichen aus der Menge aller Zeichen, die mit einer Situation verknüpft sind, hergestellt werden kann.

Im *FrameNet*-Projekt (Baker u. a. 1998) werden Lexeme framesemantisch analysiert. Anhand von Daten aus verschiedenen Korpora wird ein Lexem ei-

nem bestimmten Frame zugeordnet – oder bei Bedarf ein neuer Frame kreiert – und die Vorkommen dann mit Informationen darüber, welche Frame-Rollen darin realisiert werden, annotiert.[12] Ein deutsches Pendant zu FrameNet wurde mit dem SALSA-Projekt (Erk u. a. 2003) geschaffen, das in Abschnitt 2.3.1 näher vorgestellt wird.

2.1.5 Übersicht

Die Tabelle 2.1 gibt einen Überblick darüber, wie in verschiedenen Theorien die das Verb begleitenden Elemente bezeichnet werden. Neben dieser Abweichung in der Benennung ist natürlich auch wichtig, wie die Definition und Zuweisung von syntaktischen Elementen zu einer der beiden Gruppen geschieht. Die Tabelle ist dabei so zu lesen, dass von oben nach unten der Einfluss semantischer Kriterien zunimmt; da sich die Kriterien bei der Zuordnung unterscheiden, sind nicht alle Begriffe als genau äquivalent auffassbar. Außerdem ist dadurch nicht unbedingt gegeben, dass zwei Theorien ein Element immer gleich einordnen.

Die vielseitige Adaption der Valenztheorie für verschiedene Zwecke und in verschiedenen linguistischen Traditionen führt dazu, dass die Terminologie nicht nur in den unterschiedlichen Anwendungs- und Forschungsfeldern, sondern oft auch innerhalb des gleichen Bereichs sehr uneinheitlich ist. Spricht Tesnière z.B. bei den Begleitern eines Verbs noch vom **actants** und **circonstants**, so werden an anderer Stelle dafür Begriffe wie etwa **inner participants** und **free modifications** (Prager Schule, vgl. Sgall u. a. 1986) oder **Angaben** und **Ergänzungen** (Helbig und Schenkel 1969) gewählt. Auch Valenz an sich heißt nicht immer gleich: So ist in den Grammatikformalismen HPSG (Pollard und Sag 1994) und LFG (Bresnan und Kaplan 1982) i.d.R. von **Subkategorisierung** die Rede, d.h. Verben werden gemäß ihrer Valenzeigenschaften („Subkategorisierungsinformation“) in Klassen eingeteilt („subkategorisiert“).

Auch in der Art und Weise, wie Verbbegleiter in bestimmte Kategorien eingeordnet werden, ergeben sich teils deutliche Unterschiede. Tesnière bildet zentrale syntaktische Funktionen wie Subjekt, direktes und indirektes Objekt auf jeweils einen der drei möglichen Aktanten ab. In der Prager Schule werden syntaktische Kriterien ebenso wie semantisch-pragmatische Kriterien zur Definition der Funktoren (sowie der jeweiligen Typen von Funktoren, wie etwa *Actor* oder *Locative*) gemischt. In der Leipziger Schule werden syntaktische Angaben und Ergänzungen aus semantischen und logischen Prädikaten

12 Das Frame-Lexikon sowie die annotierten Korpusdaten sind unter `http://framenet.icsi.berkeley.edu/` einsehbar.

hergeleitet, wenn auch die Herleitung nicht immer formal klar definiert ist. In der Frame-Semantik geschieht die Definition durch Rückgriffe auf psychologisch-pragmatisch motivierte prototypische Szenendefinitionen.

Auch wenn eine stärkere Verknüpfung der Valenz mit der Semantik und der Pragmatik durchaus wünschenswert erscheint, da rein syntaktische Beschreibungsversuche schnell an ihre Grenzen stoßen, bleibt die Zuordnung von semantisch-pragmatischen Einheiten auf syntaktische Einheiten sowie die damit verbundene Feststellung der syntaktischen (oder gar semantisch-pragmatischen) Notwendigkeit der Realisierung von bestimmten Elementen problematisch.

Theorie	**Bezeichnung für Aktanten**	**Bezeichnung für Modifikatoren**
Tesnière	actants	circonstants
Prager Schule	inner participants	free modifiers
Leipziger Schule	Angaben	Ergänzungen
Frame-Semantik	core elements	non-core elements

Tabelle 2.1: Übersicht über die Bezeichnungen für Aktanten und Modifikatoren

In dieser Arbeit soll, in Anlehnung an die Prager Schule und an geeigneter Stelle an die CroCo-Terminologie, folgende Terminologie verwendet werden. **Valenz** wird möglichst breit verstanden. Sie regelt auf der semantischen Ebene, welche **Argumente** ein **Prädikat** binden kann. Im syntaktischen Sinne sagt sie aus, mit welchen **Begleitern** (die syntaktische Realisierung der Argumente) ein **Valenzträger** (die syntaktische Realisierung des Prädikats, i.e. ein Verb, oder auch ein valentes Nomen oder Adjektiv) auftritt und wie die Begleiter morpho-syntaktisch auszusehen haben. Auf die pragmatischen Aspekte der Valenz wird in dieser Arbeit lediglich am Rande eingegangen werden. Der **Valenzrahmen** ist der Bauplan, der vom Valenzträger vorgegeben wird: Er beinhaltet die zuvor angesprochene semantische und syntaktische Information für die angebundenen Elemente. Was die unterschiedlichen Typen von Begleitern angeht, so gilt folgende Nomenklatur (vgl. Abschnitt 2.1.1): die actants, die nur einmal mit einem Valenzträger vorkommen können, werden als **Aktanten** bezeichnet, die circonstants, also Angaben des Umstands, die sich recht frei und mehrfach an einen Valenzträger binden können, als **Modi-**

fikatoren. Mit dieser Nomenklatur wird eine Überlappung mit der CroCo-Nomenklatur vermieden. Gemäß den CroCo-Annotationsrichtlinien werden prädikative Komplemente (wie z.B. in *Er ist Vater*) als **Komplemente** bezeichnet (woran in dieser Arbeit festgehalten wird), was wiederum mit dem aus generativen Grammatiktraditionen stammenden Begriff „Komplement" für Begleiter des Verbs kollidiert.

Bei der Unterscheidung von **Wörterbuch** vs. **Lexikon** gilt, dass mit Wörterbüchern einsprachige oder zweisprachige gedruckte oder im Internet verfügbare Nachschlagewerke gemeint sind. Da sich in der Computerlinguistik für maschinelle Anwendungen der Begriff des „Lexikons" durchgesetzt hat, wird dieser bei der Beschreibung von MÜ-Systemen als feststehender Begriff beibehalten.

2.2 Valenz in computerlinguistischen Grammatikformalismen

In zahlreichen computerlinguistischen Grammatikschulen wurde die Idee der Valenz aufgegriffen und in den jeweiligen Formalismus integriert. Grammatikformalismen versuchen zu beschreiben, wie ein wohlgeformter Satz aufgebaut ist; insofern ist eine Integration eines wie auch immer gearteten (Verb)Valenzmechanismus sinnvoll, da die Valenz eines Verbs einen globalen Bauplan für den Satz liefert.

In den beiden folgenden Abschnitten wird auf zwei der einflussreichsten Grammatikformalismen, die **Head-Driven Phrase Structure Grammar** (kurz: HPSG) und die **Lexical-Functional Grammar** (kurz: LFG) eingegangen. Beide Formalismen gehören zu den **Unifikationsgrammatiken**., d.h. sie verwenden zur Darstellung grammatischer und semantischer Inhalte Merkmalsstrukturen, die nach verschiedenen Prinzipien zu komplexen Strukturen kombiniert werden können. Wie im Folgenden ersichtlich werden soll, hat die Valenzidee in der HPSG allerdings einen viel stärkeren Einfluss auf den Unifikationsformalismus als in der LFG, in der sich der Blick primär auf das Lexikon richtet. Interessant sind die beiden Formalismen zudem, weil bereits mit der Umwandlung zwischen Dependenzdarstellung einerseits und LFG bzw. HPSG andererseits experimentiert und die gegenseitige Abbildbarkeit belegt wurde (Forst u. a. 2004). Die Dependenzdarstellung eignet sich als Formalismus bei der Verarbeitung von Valenzinformation, weil sie unmittelbar die Beziehung zwischen einem Valenzträger als Regenten und dessen Begleitern als Dependenten modelliert.

2.2.1 HPSG

Die HPSG (Pollard & Sag 1987; Pollard & Sag 1994) ist ein unifikationsbasiertes Grammatikmodell. Syntaktische Kategorien und Lexikoneinträge werden mithilfe von Merkmalstrukturen (englisch: *attribute-value matrix*) dargestellt. Eine Merkmalstruktur enthält eine Liste von Attributen, jedem Attribut ist jeweils ein Wert zugeordnet. Jeder Wert kann wiederum eine geordnete Merkmalstruktur, aber auch eine ungeordnete Menge sein. Die HPSG zeichnet sich weiterhin dadurch aus, dass ein Großteil der für den Aufbau eines Satzes benötigten Information aus dem Lexikon kommt. Abgesehen davon gibt es nur eine geringe Anzahl von Regeln, Schemata genannt, die definieren, wie Merkmalstrukturen zu komplexeren Einheiten zu kombinieren sind, also wie z.B. aus Wörtern Phrasen werden und wie aus mehreren Phrasen eine komplexere Phrase oder der ganze Satz entsteht.

$$\begin{bmatrix} phon: \langle sehen \rangle \\ head: [cat: v] \\ lex: + \\ subcat: \langle NP_{Nom}, NP_{Acc} \rangle \end{bmatrix}$$

Abbildung 2.1: Eine vereinfachte Beispiel-Merkmalstruktur für den Lexikoneintrag sehen*.*

Ein beispielhafter, vereinfachter Lexikoneintrag ist der Abbildung 2.1 zu entnehmen. Wie erkennbar ist, enthält bereits der Lexikoneintrag das sog. *head*-Merkmal, das u.a. die zentralen Kongruenzeigenschaften sowohl von phrasalen als auch von nicht-phrasalen Elementen (hier nicht gezeigt), sowie die Kategorie des Eintrags bestimmt. Eine Phrase in HPSG erbt immer das head-Merkmal seiner sog. **Kopftochter**, wie in Abbildung 2.2 zu sehen ist. Oliva (2003) sieht darin eine zentrale Ähnlichkeit zwischen der Dependenzgrammatik und der HPSG: Wichtige Merkmale einer Phrase werden vom Kopf der Phrase bestimmt. Allerdings geht die HPSG hier noch weiter als die klassische Dependenzgrammatik: Die Phrase kann in den nicht-head-Merkmalen auch Werte der sogenannten **Komplementtöchter** (d.h. der Nicht-Kopftöchter) erben.

Die Abbindung von Elementen wird im originalen Entwurf der HPSG über die sog. *subcat*-Liste kontrolliert (siehe Abbildung 2.3). Die subcat-Liste

$$[dtrs:[head\_dtr:[\,]]] \Rightarrow \begin{bmatrix} head:|1| \\ dtrs:[head\_dtr:[head:|1|]] \end{bmatrix}$$

Abbildung 2.2: Das Head-Feature Principle der HPSG (übernommen von Oliva 2003, S.661)

der Kopftochter besteht aus der Konkatenation der subcat-Liste der Phrase mit den Komplementtöchtern der Phrase. In anderen Worten, die subcat-Liste einer Phrase entsteht, wenn man von der subcat-Liste der Kopftochter diejenigen Elemente abzieht, die bereits als Komplemente in der Phrase abgebunden sind. Tatsächlich sind mit der Zeit diverse Mechanismen entworfen worden, um verschiedene Aspekte der Valenz zu inkorporieren: etwa das *content*-Merkmal für die Valenz oder das *argument structure*-Merkmal, das die syntaktische Information von Subjekt und Komplementen zusammenfasst (vgl. Darstellung in Bouma u. a. 2001). Die Grundprinzipien der durch den Kopf einer Phrase gesteuerten Abbindung sind aber erhalten geblieben.

$$[dtrs:[head\_dtr:[\,]]] \Rightarrow \begin{bmatrix} subcat:|1| \\ dtrs:\begin{bmatrix} head\_dtr:[subcat:|1|\circ|2|] \\ comp\_dtrs:|2| \end{bmatrix} \end{bmatrix}$$

Abbildung 2.3: Das Subcategorization Principle der HPSG (übernommen von Oliva 2003, S.663)

Über die subcat-Liste werden nur lokale Dependenzen abgebunden. Um Fernabhängigkeiten behandeln zu können, wurden in der HPSG verschiedene Mechanismen eingeführt. Die Grundidee ist dabei, ein *slash*-Merkmal zu verwenden, das Elemente in sich trägt, die zwar laut subcat-Liste hätten lokal abgebunden werden sollen, aber in der Satzkonfiguration an eine andere Stelle, z.B. durch Topikalisierung, gewandert sind. Die Ansätze zur Behandlung

von Fernabhängigkeiten unterscheiden sich v.a. dadurch, wie der slash-Wert zu besetzen ist, ob er z.B. schon durch eine Lexikonregel von einem Basiseintrag abgeleitet wird, oder erst durch eine phrasale Regel befüllt wird.[13] Eine detailliertere Beschreibung sowie ein tiefergehender Vergleich der HPSG mit der Dependenzgrammatik kann dem Aufsatz von Oliva (2003) entnommen werden.

Die HPSG kam in verschiedenen MÜ-Projekten, z.B. im Verbmobil-Projekt, zum Einsatz. Das Verbmobil-Projekt wird in Abschnitt 2.5.5 beschrieben.

2.2.2 LFG

Die LFG (Bresnan und Kaplan 1982; Bresnan 2001) beschreibt einen Satz mithilfe verschiedener Darstellungsebenen. Die syntaktische Form des Satzes wird durch die **constituent structure** oder kurz *c-structure* in Form eines Phrasenstrukturbaums dargestellt. In der **functional structure** oder kurz *f-structure* sind die Funktionen aufgelistet, die in einem Satz vorkommen, etwa SUBJ oder OBJ; ebenso sind satzspezifische Merkmale angegeben wie etwa das Tempus sowie das Prädikat. Die f-structure ist eine hierarchisch aufgebaute Merkmalstruktur, wie sie in diversen Unifikationsgrammatiken (vgl. Abschnitt 2.2.1, HPSG) vorkommt. Die Einheiten des Phrasenstrukturbaums sind auf Attribut-Wert-Paare in der f-structure abbildbar, wie in Abbildung 2.5 gezeigt. Der syntaktische Unterbaum mit der Kategorie DP und dem darunterliegenden Pronomen *I* entspricht in der f-structure dem Wert des Attributs SUBJ, die Kategorie DP mit dem darunterliegenden Pronomen *him* entspricht dem Wert des Attributs OBJ (in beiden Fällen ist der Wert wiederum eine Merkmalstruktur). Bemerkenswert ist, dass nicht nur einzelne Knoten auf die f-structure abgebildet werden, sondern (potenziell mehrere) Paare von Knoten und Kanten.[14] Dies ist aber nur konsequent, da die Werte von Attributen wie z.B. SUBJ für komplette Phrasen, und nicht für einzelne Wörter stehen.

13 Vgl. dazu z.B. Sag & Fodor (1994), Avgustinova und Oliva (1996) und Bouma, Malouf und Sag (Bouma u. a. 2001). An dieser Stelle zeigt sich sehr deutlich der Einfluss der Grundsprache Englisch bei der Entwicklung der HPSG. Man kann zumindest annehmen, dass Sprecher nicht-konfigurationeller Sprachen nicht auf die Idee von „wandernden Elementen“ kämen, sondern von Beginn an von einer freieren Satzstellung ausgingen.

14 Tatsächlich ist auch diese Formulierung vereinfachend, soll aber für den zusammenfassenden Charakter des Abschnitts reichen. Eine ausführliche Darstellung der Abbildungsprinzipien mittels sog. **funktionaler Beschreibungen** ist in (Bresnan 2001) zu finden.

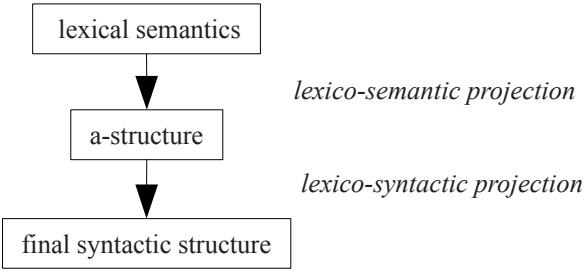

Abbildung 2.4: Die Stellung der a-structure zwischen lexikalischer Semantik und syntaktischer Struktur (adaptiert von Bresnan 2001, S.306)

Das Grundgerüst der f-structure wird aus der Darstellungsebene **argument structure** oder *a-structure* abgeleitet. Diese stellt die einem Satz zugrundeliegende Prädikat-Argument-Struktur dar. Die Argumente der a-structure werden nach verschiedenen Prinzipien, abhängig beispielsweise von der

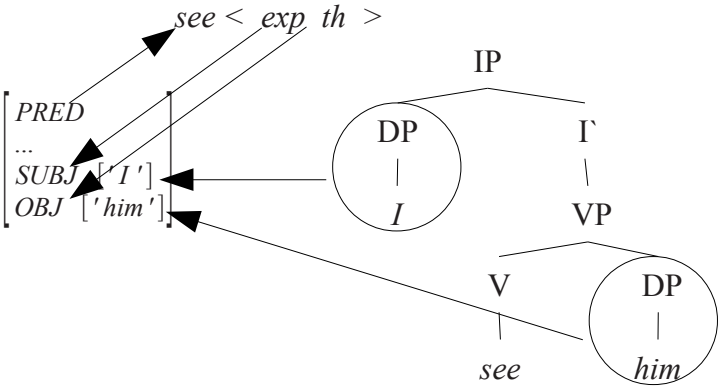

Abbildung 2.5: Abbildung von a-structure auf f-structure und Korrespondenzen zwischen c-structure und f-structure (adaptiert von Bresnan 2001, S.303)

Diathese, auf Funktionen in der f-structure abgebildet. Wie in Abbildung 2.5 gezeigt, wird beim Prädikat *see* in der aktiven Diathese im Englischen der *experiencer* auf das Subjekt der f-structure abgebildet, das wahrgenommene *theme* dagegen auf das Objekt.

Die Rolle der Valenz ist in der LFG also auf zwei Ebenen aufgeteilt. Die a-structure ordnet einem Prädikat eine bestimmte Anzahl und bestimmte Typen von Argumenten zu, abhängig von der lexikalischen Semantik eines Verbs. Die f-structure hingegen bestimmt, welche syntaktischen Funktionen im Satz vorkommen. Durch Informationen des Prädikats aus dem Lexikon wird festgelegt, von welchen morpho-syntaktischen Kategorien und von welchen semantischen Typen die Funktionen belegt werden können.

Die a-structure dient also als Bindeglied zwischen lexikalischer Semantik und der funktional-syntaktischen Struktur (vgl. Abbildung 2.4). Die f-structure dagegen abstrahiert von syntaktischen Variationen weg: Die Funktionen, die aus der a-structure abgeleitet sind, können auf ganz unterschiedliche Typen und Konfigurationen von Phrasen in der c-structure verweisen.

Die LFG wurde in verschiedenen MÜ-Ansätzen verwendet. Neuere Ansätze basieren auf der Verwendung von Parallelkorpora. Diese Ansätze werden in Abschnitt 2.5.4 beschrieben.

2.2.3 Übersicht

Die Valenz interagiert mit verschiedenen Ebenen der Linguistik. Entsprechend ist auch in den Grammatikformalismen kein einzelner, eindeutiger Ort oder Mechanismus zur Valenzbehandlung vorgesehen.

In der HPSG sind im Verlauf der Jahre verschiedene Mechanismen zur Behandlung von Valenz, und darin insbesondere von Fernabhängigkeiten, entworfen worden. Zentral bei der HPSG ist die Subkategorisierung sowie die Unterscheidung von Kopftöchtern und Komplementtöchtern. Die Kopftochter regelt jeweils, welche Elemente unter Berücksichtigung welcher Kriterien abgebunden werden.

In der LFG findet die Valenzdefinition insbesondere auf zwei Ebenen statt. Die a-structure beschreibt in etwa das, was Helbig als logische Valenz bezeichnet: Sie ordnet einem Prädikat eine bestimmte Anzahl und bestimmte Typen von Argumenten zu, abhängig von der lexikalischen Semantik eines Verbs (ebenso wie bei Helbig). Die Argumente der a-structure werden auf syntaktische Funktionen in der f-structure abgebildet, wo auch ihre syntaktischen Eigenschaften definiert werden.

Die Darstellung von Valenzinformationen ist in der HPSG weniger klar definiert als in der LFG. Dennoch haben schon beide Formalismen als Grundlage für MÜ-Systeme fungiert (vgl. Abschnitt 2.5), in die auch erfolgreich Valenzinformation inkorporiert wurde.

2.3 Valenz in der Lexikografie

Die Lexikografie ist eins der ältesten linguistischen Tätigkeitsfelder. Wörterbücher werden aus den unterschiedlichsten Gründen erstellt, z.B. zur Standardisierung als Fachwörterbücher, als Mittel gegen den Sprachverfall, zum Sprachvergleich oder als Nachschlagewerke für die zwischensprachliche Kommunikation (Herbst u. a. 2004b). Valenzwörterbücher dienen der systematischen Untersuchung der Valenz, als mono- oder bilinguale Nachschlagewerke oder als Basis für computerlinguistische Anwendungen.

Die Autoren der im Folgenden vorgestellten lexikografischen Valenzwerke erwähnen häufig eines der folgenden drei Ziele: Sie wollen Fremdsprachenlernern und Übersetzern ein Nachschlagewerk bieten, das in seinem Inhalt weit über ein einfaches Wörterbuch hinausgeht, und/oder wollen der maschinellen Sprachverarbeitung eine Quelle für ihre Anwendung liefern, oder die Autoren wollen Ausschnitte einer Sprache aus einem bestimmten Forschungsinteresse heraus abbilden.

Valenzlexika gibt es in der monolingualen und in der bilingualen Variante; ein multilinguales (mehr als zwei Sprachen beinhaltendes) gedrucktes oder im Internet verfügbares Werk ist dem Autor der vorliegenden Arbeit nicht bekannt. Der folgende Überblick über einige zentrale Valenzwörterbücher ist nicht nach deren Ersteller oder Sprache gegliedert, sondern global nach monolingualer und bilingualer Variante, da die Entscheidung für die eine oder andere Variante von der Zielsetzung beeinflusst ist und selbst wiederum verschiedene Aspekte bei der Erstellung beeinflusst. Ausgewählt wurden nur für den Menschen gedachte Wörterbücher. In vielen computerlinguistischen Anwendungen „schlummern" Valenzwörterbücher, die der Valenzlexikografie zugänglich gemacht werden könnten. Auf diesen Aspekt wird an dieser Stelle allerdings nur hingewiesen, aber nicht näher eingegangen.

2.3.1 Monolinguale Valenzlexikografie

Für das Deutsche wurden verschiedentlich Verbvalenzwörterbücher oder Verbklassifikationen mit Valenzinformationen erstellt, von denen einige im Folgenden vorgestellt werden. Die meisten von ihnen sind didaktisch motiviert, wobei die teilweise recht komplexen Darstellungen der Valenzinformation

zunächst gar nicht als für didaktische Zwecke geeignet erscheinen (vgl. dazu Abschnitt 4.4.2).

In diesem Abschnitt werden zwei der bekanntesten deutschen Valenzwörterbücher, zum einen von Helbig und Schenkel (1969), zum andere vom Engel und Schumacher (1976), vorgestellt, sowie die Verbklassifikation von Ballmer und Brennenstuhl (1986). Für das Englische wird auf die Werke von Emons (1978), Allerton (1982) sowie das Valenzwörterbuch von Herbst u.a. (2004a) eingegangen. Auch auf die deutsche und die englische Variante von FrameNet wird nochmals kurz eingegangen (vgl. dazu Abschnitt 2.1.4)

Das didaktisch motivierte Valenzwörterbuch von Helbig und Schenkel (1969) entstand für Nicht-Deutschmuttersprachler. Das Werk selbst eignet sich zwar in seiner Form nicht als Nachschlagewerk für den Lerner selbst, enthält aber im letzten Kapitel einige Hinweise darauf, wie die Verbeinträge für den Durchschnittslerner verständlicher aufzubereiten seien. Helbig und Schenkel verwenden eine dreistufige Darstellung der Valenzinformation. Zunächst zeigt ein einfacher Verbeintrag der Form *danken*$_{1(2,3)}$ an, wie viele obligatorische (vor der Klammer) und wie viele fakultative (in der Klammer) Angaben[15] ein Verb bindet. Die zweite Stufe gibt an, welcher Art die gebundenen Elemente sein können, so z.B. *danken → Sn,(Sd),(pS|NS$_{daß}$)*. Das Verb *danken* nimmt also ein Subjekt, fakultativ ein Dativobjekt und/oder eine Präpositionalphrase, die ggf. durch einen „daß"-Satz ersetzt werden kann. Die dritte Stufe gibt an, welchen semantischen Typs die abgebundenen Satzglieder sein können. So gibt Helbig z.B. für „danken" folgende mögliche semantische Klassen für die Nominativergänzung an: *Sn → 1. Hum (der Jubilar dankt), 2. Abstr (Der Betrieb dankt dem Ministerium)*. Das Subjekt ist also ein Mensch oder eine abstrakte Einheit, die als eine handelnde abstrakte Person steht[16].

Das Valenzwörterbuch von Helbig und Schenkel ist das erste deutsche Valenzwörterbuch, und der Beginn einer Reihe von Beiträgen von Gerhard Helbig zum Thema Valenz, wie in Abschnitt 2.1.3 dargestellt.

Die Valenzwörterbücher des Instituts für deutsche Sprache in Mannheim entstanden ebenfalls im didaktischen Kontext, die Autoren heben allerdings auch den möglichen Nutzen der Wörterbücher für die maschinelle Sprachverarbeitung hervor. Das „kleine Valenzlexikon" (Engel und Schumacher 1976), das erste seiner Art aus den Reihen des IDS, ist hauptsächlich für Lehrende gedacht. Die Tatsache, dass seine Notation und die Voraussetzungen an theoreti-

15 In der Hellwig'schen Terminologie werden Aktanten als Angaben und Modifikatoren als Ergänzungen bezeichnet, vgl. Abschnitt 2.1.3.

16 Was Helbig ebenfalls angibt, hier aber aus Platzgründen verkürzt dargestellt ist.

sches Wissen des Lesers es ungeeignet für den Durchschnitts-Deutschlerner machen, merken die Autoren selbst an.

Das „kleine Valenzlexikon“ ist rein syntaktisch orientiert. Nach syntaktischen Kriterien werden verschiedene sog. Ergänzungsklassen definiert, bezeichnet als E0 bis E9, die für nominale, adjektivische und adverbiale Ergänzungen des Verbs stehen. So gehören z.B. Nominativergänzungen[17] in die Klasse E0, Situativergänzungen (Angaben des Ortes und der Zeit, also syntaktische Elemente mit adverbialer Funktion) in E5 und Artergänzungen (z.B. *Er ist krank*, also alle Arten von Komplementen) in E8. Für die satzförmigen Ergänzungen gibt es die Klassen SE0 bis SE2 und SE4. Diese stehen immer als Ersatz für die Ergänzung mit der jeweils entsprechenden Nummer, eine SE0 kann also bei manchen Verben eine E0 ersetzen etc.

Das Lexikon enthält einen Eintrag pro Verblesart. In jedem Eintrag stehen folgende Angaben:

- die mit dem Verb auftretenden E-Klassen (nur mit Nummer), also z.B.

  essen *0(1*

  d.h. Das Verb *essen* bindet eine Nominativergänzung und optional (durch die öffnende Klammer markiert) eine Akkusativergänzung;
- die Fähigkeit des Verbs zur Passivisierung, z.B. *P1* für volle Passiv-Fähigkeit;
- sowie die SE-Klassen, die eine der E-Klassen ergänzen können, z.B.

  danken *013*
  1:SE *DASS* *„Niemand dankt es ihr, daß sie ihm geholfen hat.“*

Semantische Beschränkungen für die möglichen Ergänzungen sind nicht formuliert, was die starke Ausrichtung des Lexikons an der Syntax zusätzlich belegt.

Das „kleine Valenzlexikon“ ist mehrfach aufgelegt und inzwischen zum großen Band „Verben in Feldern“ (Schumacher 1986) ausgebaut worden, der neben den syntaktischen Informationen auch Angaben zu semantischen Selektionsbeschränkungen der Verben macht. Außerdem stand das „kleine Valenzlexikon“ Pate für diverse kontrastive Valenzwörterbücher des IDS, die in

17 Diese Bezeichnung deutet darauf hin, dass auch Engel und Schumacher die klassische Subjekt-Objekt-Trennung aufheben.

Abschnitt 2.3.2 vorgestellt werden. Der Beitrag des IDS und gerade des „kleinen Valenzlexikons“ zur Valenzlexikografie ist unbestritten.

Die Verbklassifikation von Ballmer und Brennenstuhl (1986) entstand aus dem konkreten Forschungsinteresse heraus, einen möglichst umfassenden Bestand deutscher Verben (ca. 8.000) nach einem von den Autoren geschaffenen temporal-kausalen Modell zu klassifizieren. Neben der Klassifikation der Verben ist jeder Klasse eine Beschreibung der jeweils zu erwartenden Valenzrahmen der Verben innerhalb einer Klasse zugeordnet; für einzelne Verben sind Lexeme typischerweise erscheinende Begleiter oder Besonderheiten des Valenzrahmens notiert. Die Notwendigkeit solcher Angaben ergibt sich aus der Klassifikationsweise Ballmers und Brennenstuhls. Nach Auffassung der Autoren ordnen sich Verben nicht nur in Klassen, sondern letztere wiederum in Modelle, die einen Aktionsablauf beschreiben und für verschiedene Phasen der Aktivität stehen. Diese Phasen sind temporal und kausal (nach dem typischen Ablaufmuster der Phasen) geordnet. In jeder Phase sind verschiedene Mitspieler am Geschehen beteiligt, was deren Angabe im Verblexikon notwendig macht. So gibt es z.B. ein *Zustoßmodell* mit folgenden Phasen und jeweils folgenden Paraphrasen unter Angabe der typischerweise zu erwartenden Begleiter (Auszug, Verbbeispiele in Klammern):

Vorspielphase: Sich einem Einfluss aussetzen jd 1 („sich sonnen“)
Ablaufphase: Zustoßen etw1 jd3 (allgemein: „geschehen, passieren“, speziell: „verbrennen“)
Schlussphase: Sich auswirken etw1 auf jd2 („aufregen, erschrecken“)
Nachspielphase: Reagieren jd1 auf etw2 („antworten, reagieren“).

Der Valenzrahmen wird mittels Pronomina wie *jd* „jemand“ oder *etw* „etwas“ wiedergegeben. „Jemand“ steht dabei dafür, dass an dieser Stelle prinzipiell eine Person auftreten sollte, „etwas“ steht für eine Sache. Mittels der Zahlen 1-4, an die Pronominalmarker angehängt, wird der Kasus markiert, der an dieser Stelle steht; *etw1* steht damit für „eine Sache im Nominativ“, *auf jd2* steht für „eine Person im Akkusativ, mit der Präposition *auf*“.

Die Klassifikation von Ballmer und Brennenstuhl ist insofern bemerkenswert, als dass sie die geschaffenen Verbklassen nicht nur hierarchisiert, sondern sie in einen temporal-kausalen Zusammenhang setzt. Von Interesse könnte hierbei sein, zu beobachten, wie sich die Valenz, die für jede Klasse angegeben ist, innerhalb einer solchen temporal-kausalen Linie entwickelt. Dies soll aber nicht Gegenstand der vorliegenden Arbeit sein.

Auf der Ebene der Frame-Semantik (zur Theorie und zu Grundlagen der FrameNet-Datenbank vgl. Abschnitt 2.1.4) arbeitet für das Deutsche das lexikografische SALSA-Projekt (Erk u. a. 2003). In diesem Projekt werden Sätze aus der TIGER-Baumbank (Brants u. a. 2002) mit framesemantischer Annotation erweitert, nach dem Vorbild der englische FrameNet-Datenbank (Baker u. a. 1998). SALSA versucht dabei, die FrameNet-Definitionen für Frames zu verwenden. Da der deutsche und der englische Kulturraum viele prototypische Situationen gemeinsam haben, dürften auch viele der englischen Frame-Definitionen auf das Deutsche anwendbar sein. Inwiefern dies der Fall ist, untersuchen z.B. (Padó und Lapata 2005; Padó 2007a). Dass es Übereinstimmungen gibt, und Frames sich daher für die bilinguale Lexikografie eignen, wird im folgenden Abschnitt 2.3.2 dargelegt.

In den linguistischen Theorien der anglophonen Kulturräume hat die Valenzidee anfänglich wenig Anklang gefunden und ist in der theoretischen Diskussion nur wenig vorhanden. Allerdings ist Valenz, wie in Abschnitt 2.2 dargelegt, in der einen oder anderen Form Teil vieler Grammatikformalismen, und damit implizit Forschungsgegenstand auch für das Englische. Zudem trägt die Frame-Semantik zur Diskussion der Valenz im Englischen bei, wenn auch mit einem sehr eigenen, kognitiv motivierten Formalismus. Neben der FrameNet-Datenbank gibt es für das Englische Valenzwörterbücher bzw. -grammatiken von Emons (1978), von Allerton (1982) und von Herbst u. a. (2004a).

Bemerkenswert im Vergleich mit den deutschen Valenzwerken ist, dass von den Autoren der englischen Valenzwerke die Subjekt-Objekt-Opposition nicht aufgegeben wird. Das mag daran liegen, dass im Englischen das Subjekt durch seine recht fixe Stellung vor dem finiten Verb des Satzes deutlicher markiert und damit als Kategorie für die Grammatik relevanter ist als im Deutschen, wo die Wortstellung nicht im gleichen Maße von der Subjekt- oder Objektfunktion abhängt.

Emons (1978) orientiert sich bezüglich der Analyse und Darstellungsweise in seinem Entwurf für eine Valenzgrammatik des Englischen am „kleinen Valenzlexikon“. Er definiert nach syntaktischen Kriterien Ergänzungsklassen und verwendet diese für die Valenzbeschreibung von Verben. Emons' Darstellung der Lexikoneinträge ist noch kompakter als die des „kleinen Valenzlexikons“. So lautet etwa der Eintrag für das Verb *believe* wie folgt:

S12[P12 + E1[NOM1/ES1] + [E2[NOM2/ES2[that]]]]

Dieser Eintrag sagt aus, dass es sich um ein zweiwertiges Verb mit Nominativ und Akkusativ handelt (S12), wobei sowohl die Ergänzung E1 als auch die Ergänzung E2 als Nominalausdrücke realisiert werden können (NOM1 bzw. NOM2), oder als Nebensätze (Ergänzungssatztypen ES1 bzw. ES2[that], z.B. in *[Whoever reads this]$_{ES1}$, will believe [that this has really happened]$_{ES2}$*).

Allerton (1982) begründet seine Valenzklassen mit einer sehr detaillierten Untersuchung von grammatischen Prozessen im Englischen wie etwa Kovariation, Alternation oder Transformationen, und stellt eine Verbindung zwischen diesen Prozessen und der Semantik her. So definiert Allerton z.B. neben der traditionellen Kategorie OBJECT eine Kategorie OBJOID für Elemente, die zwar zunächst syntaktisch als Objekte erscheinen, aber tatsächlich keine sind, weil sie z.B. nicht als Subjekt in Passivisierungen erscheinen können. Dies ist etwa der Fall für Objekte in Kopulakonstruktionen wie das Objekt *a student* in *Thomas is a student*. Diese grammatischen Kategorien dienen als Inventar für die Beschreibung von Valenzklassen, denen Verben zugeordnet werden. Das Verb *see* wird der Valenzklasse 12 mit dem Muster

SUBJECT + V + OBJECT

zugeordnet. Neben diesen grammatischen Tests führt Allerton auch eine Klassifikation von Argumenttypen in Klassen wie *performer*, *recipient* oder *characterized* durch. Diese werden aber nur für eine kleine Auswahl von Beispieleinträgen verwendet, so wird z.B. für das Verb *reside* wie in *Thomas resided in the palace* definiert, dass das erste Argument ein *performer/affected* ist, das zweite Argument eine *location*.

Das Valenzwörterbuch für das Englische von Herbst u.a. (2004a) ist im Internet als *Erlangen Valency Pattern Bank* verfügbar.[18] Sowohl die gedruckte Version als auch die Internetversion wurden auf Basis des COBUILD Korpus der Universität Birmingham erstellt. Das Wörterbuch ist sehr reich an Angaben zu möglichen syntaktischen Verwendungsmustern sowie Beispielen zu jedem der Verwendungsmuster. Ein leicht gekürzter Ausschnitt des Eintrags für das Verb *see* wäre folgender:

'look'	*Active 1/3*	*Passive 1/3*	*General 0*
I	*[N]$_{A}$/[by N]*		
II	*[N]$_{P}$ [N]$_{P\text{-}only}$ [that-CL]$_{P(it)}$ [wh-CL]$_{P(it)}$*		
III	*[for REFL PRON]*		

18 http://www.patternbank.uni-erlangen.de

Dieser Eintrag sagt aus, dass an dieser Stelle die Bedeutung von *see* beschrieben wird, die *look* am nächsten kommt. Im aktiven und passiven Gebrauch werden zwischen einem und drei Argumenten realisiert. Ein genereller Gebrauch ohne Argument ist möglich, wie z.B. im Satz *Seeing is believing*. Sind unter den römischen Zahlen mehrere Kategorien angegebenen, so sind diese alternative Realisierungsmöglichkeiten desselben semantischen Argumenttyps. Das erste Argument ist also entweder eine NP im Aktivsatz (Subskript *A*), oder eine mit *by* angeschlossene NP im Passivsatz. Zweites Argument ist entweder eine NP, die auch Subjekt des Passivsatzes sein kann (Subskript *P*), durchaus auch alleinstehend im Passivsatz (Subskript *P-only*), oder eine *that*- oder mit einem *wh*-Wort eingeleitete Phrase (z.B. *He saw that someone was coming* bzw. *He saw who was coming*), die im Passivsatz auftreten können, wobei mit *it* extraponiert werden kann (*It was seen that he was coming*). Als drittes Argument ist eine *for*-PP mit Reflexivpronomen möglich, also z.B. *He saw for himself who was coming*.

2.3.2 Bilinguale Valenzlexikografie

Die bilinguale Valenzlexikografie ist auch heute, trotz moderner, korpusgestützter Methoden, noch nicht weit verbreitet und besitzt noch ein hohes Ausbaupotenzial. Die Erstellung eines bilingualen Valenzwörterbuchs stellt die Akteure vor eine Vielzahl schwieriger Fragen. Zum einen muss ein gewisser Bestand an valenten Lexemen in einer Sprache aufgearbeitet werden, was aufgrund von Phänomenen wie Polysemie sowie anderen typischen, in Abschnitt 2.1 bereits angesprochenen Problemen der Valenztheorie schon schwer genug ist. Das wahre Kunststück besteht allerdings darin, zwischen den Sprachen die jeweiligen Entsprechungen für eine Lexembedeutung zu finden. Dies zeigt sich auch in den im Folgenden vorgestellten Valenzwerken, den Arbeiten des IDS Mannheim, einem lernerorientierten Valenzwörterbuch Deutsch-Portugiesisch sowie multilingualen, auf Frame-Semantik basierenden Wörterbüchern.

Eine Reihe von bilingualen Valenzwörterbüchern mit dem Deutschen als AS oder ZS stammt aus dem Hause des IDS Mannheim, so etwa ein deutsch-spanisches (Rall u. a. 1980), ein deutsch-rumänisches (Engel und Savin 1983), und ein deutsch-italienisches (Bianco 1996) Valenzwörterbuch. Zudem berichtet Engel (2006) von einem Projekt für ein deutsch-serbisch/kroatisch/bosnisches Valenzwörterbuch. Die Arbeiten bauen auf den im Rahmen des „Kleinen Valenzlexikons deutscher Verben“ (vgl. auch Abschnitt 2.3.1) gewonnenen Erfahrungen auf. Allerdings ist eine der ersten Erkenntnisse beim

Schritt vom einsprachigen zum zweisprachigen Valenzwörterbuch, dass in einem kontrastives Valenzlexikon auch semantische Restriktionen notwendig sind - im einsprachigen deutschen Wörterbuch sind nur syntaktische Angaben vorhanden.

Die Autoren des deutsch-rumänischen Valenzwörterbuchs illustrieren an folgendem Beispiel, warum die Angabe semantischer Restriktionen für ein kontrastives Valenzwörterbuch unabdingbar ist. Das deutsche Verb *schwimmen* kann ins Rumänische mit zwei Verben übersetzt werden, die allerdings verschiedene Konnotationen haben: *a înota* steht für „Lebewesen schwimmt aktiv", *a pluti* dagegen für „im Wasser treibende Gegenstände". An diesem Beispiel zeigt sich die Notwendigkeit, semantische Beschreibungen in Valenzwörterbücher aufzunehmen. Schließlich ist eine der großen Schwierigkeiten beim Erlernen einer Fremdsprache gerade jene, die vielen zur Muttersprache oft unterschiedlichen Verwendungsweisen und Kombinationsmöglichkeiten von Verben und die damit verbundenen Bedeutungsnuancen zu verstehen. Im Fall des deutsch-rumänischen Wörterbuchs werden generische Kombinationsregeln mittels semantischer Merkmale wie +ANIM für belebte Wesen (also *a înota* +ANIM) oder +MAT für unbelebte Wesen (also *a pluti* +MAT) angegeben.

Diese Einsicht dürfte auch eine der Grundmotivationen des Ansatzes von Duffner u.a. (2009) sein (vgl. Abschnitt 2.4.3). Die Autoren des deutsch-rumänischen Valenzwörterbuchs weisen allerdings gleich auf die Einschränkung ihres sehr formal aufbereiteten Valenzlexikons: Engel und Savin richten sich explizit an Lehrende, Studenten und Lehrwerkautoren. Für den linguistisch Nichtversierten müsste das Wörterbuch verständlicher gefasst werden.

Welker (2003) stellt einen Entwurf für ein deutsch-portugiesisches Valenzwörterbuch vor, das sich zum Ziel setzt, Valenzeinträge so darzustellen, dass auch Nutzer ohne besondere linguistische Vorkenntnisse es leicht verwenden können.[19] Für die Darstellung der Argumente in Valenzrahmen verwendet Welker *Siglen*, die semantische und syntaktische Eigenschaften miteinander vereinen (können). So steht z.B. die Sigle *P* für Personen, *A* für Dinge und *AN* für ein Tier. Ein Eintrag sieht wie folgt aus:

*P * P*

Der Stern steht für das Verb selbst, könnte an dieser Stelle also z.B. für das Verb *lieben* stehen (*jemand liebt jemanden*). Die Funktionen Subjekt und Ak-

19 Das Wörterbuch ist im Internet abrufbar unter http://vsites.unb.br/il/let/welker/dici/Deut/anfangframe.html

kusativobjekt werden durch die Stellung an erster bzw. zweiter Stelle angegeben. Dativ- und Genitivobjekte werden durch ein Subskript *d* bzw. *g* gekennzeichnet. Daneben gibt es aber auch Siglen wie z.B. *ADJ* für Adjektive (z.B. bei Verben wie *finden* mit dem Eintrag *P * A ADJ* für „jemand findet etwas ADJEKTIV“).

Mit Blick auf die vorliegende Arbeit haben Welkers Überlegungen und Zielsetzungen Vorbildcharakter, da sich Welker intensiv mit der Frage der Nutzbarkeit von Valenzwörterbüchern für den „Durchschnittsverbraucher“ auseinandersetzt. Mit der Frage der Darstellung von Valenzinformationen beschäftigt sich Abschnitt 4.4.2 im Detail.

Die Eignung framesemantischer Annotation zur Erstellung bilingualer Valenzwörterbücher zeigt Boas (2001; 2002) auf. Der Autor beschreibt mittels einer Analyse der Frame-semantische Annotation von Bewegungsverben, wie diese Art der Annotation mit Ihrer Rollenbeschreibung Zweifelsfälle bei der Übersetzung von Verben klären kann. Boas (2001:65) illustriert dies am Beispiel des englischen Verbs *walk* und zwei seiner Bedeutungen.

(4) *Bernd walked to the door.*

(5) *Bernd walked Anna to the door.*

Das erste Beispiel kann ins Deutsche problemlos mit *Bernd ging zur Tür* übersetzt werden. Für das zweite Beispiel wäre allerdings die Verwendung von *gehen* im Deutschen unpassend. Besser hieße es *Bernd begleitete Anna zur Tür*. Boas schlägt vor, die Rollenannotation von FrameNet zu verwenden, um die Unterschiede zwischen den Bedeutungen formalisiert darzustellen. Für $walk_1$ aus dem ersten Beispiel ergibt sich ein Lexikoneintrag mit zwei Frame-Rollen: $walk\ [Self\text{-}mover_{NP}\ Goal_{NP}]$; der Eintrag für $gehen_1$ sieht entsprechend aus. Allerdings gibt es keinen Eintrag für *gehen*, der dem Rollenmuster von $walk_2$ entspricht: $walk\ [Self\text{-}mover_{NP}\ Cotheme_{NP}\ Goal_{PP}]$; der Eintrag von *begleiten* würde diesem Rollenmuster dagegen entsprechen. Framesemantische Annotation wird dadurch zu einer Art „Transfersprache“.

2.3.3 Übersicht

Die in den vorangehenden Abschnitten vorgestellten Valenzwerke unterscheiden sich insbesondere darin, wie sie motiviert sind, und was der Ausgangspunkt der Betrachtung ist: die Syntax oder die Semantik (oder gar die Pragmatik). Tabelle Fehler: Referenz nicht gefunden gibt einen Überblick darüber. Darin sind zunächst monolinguale, dann bilinguale Valenzwörterbücher auf-

Monolinguale Ressource	**Motivation**	**Ausgangspunkt**
Helbig und Schenkel 1969 (deutsch)	Didaktik	Syntax, mit semantischen Informationen
Engel und Schumacher 1976 (deutsch)	Didaktik, auch: maschinelle Sprachverarbeitung	nur Syntax
Emons 1978 (englisch)	Didaktik	nur Syntax
Allerton 1982 (englisch)	Forschungsinteresse	Syntax, mit semantischen Rollenbeschreibungen
Ballmer und Brennenstuhl 1986 (deutsch)	Forschungsinteresse	Semantik und Pragmatik (temporal-kausales Modell), mit syntaktischen Informationen
FrameNet/SALSA (englisch/deutsch)	Forschungsinteresse	Semantik und Pragmatik (prototypisches Szenenmodell), mit syntaktischen Informationen
Herbst 2004 (englisch)	Didaktik, Forschungsinteresse	Syntaktisch, mit Kollokationen, viele Beispiele
Bilinguale Ressource		
Engel und Savin 1983 (deutsch-rumänisch)	Didaktik, Forschungsinteresse	Syntax, mit semantischen Informationen
Rall, Rall und Zorrilla 1980 (deutsch-spanisch)	Didaktik	Syntax, mit semantischen Informationen
Bianco 1996 (deutsch-italienisch)	Didaktik	Syntax, mit semantischen Informationen
Boas 2001, 2002 (deutsch-englisch)	Forschungsinteresse	Semantik und Pragmatik (prototypisches Szenenmodell), mit syntaktischen Informationen
Welker 2003 (deutsch-portugiesisch)	Didaktik	Syntax, mit semantischen Informationen

Tabelle 2.2: Übersicht über Motivation und Inhalt verschiedener monolingualer und bilingualer Valenzwörterbücher

geführt. Zu jedem der Wörterbücher ist nochmals aufgeführt, mit welcher zentralen Motivation das Wörterbuch erstellt wurde – ob etwa zu didaktischen Zwecken oder aus reinem Forschungsinteresse. Ebenso wird aufgeführt, welche Art von Information das jeweilige Wörterbuch liefert, ob es sich z.B. auf rein syntaktische Information beschränkt, oder neben semantischen evtl. sogar pragmatische Aspekte berücksichtigt.

Mit Blick auf die MÜ sind natürlich insbesondere die bilingualen Werke interessant. Eine Digitalisierung dieser Inhalte und die Bereitstellung als Lexikon für MÜ-Anwendungen könnten wertvoll sein. Allerdings darf nicht vergessen werden, dass bereits teilweise recht umfangreiche bilinguale MÜ-Lexika existieren (vgl. Abschnitt 2.5), und damit in der MÜ versteckte Valenzwörterbücher „schlummern", die der HÜ noch nicht zugänglich gemacht wurden.

2.4 Valenzextraktion aus Korpora

Die klassische, manuelle Lexikografie hat mit einer Reihe von Schwierigkeiten zu kämpfen. Zum einen wird ein bereits zementierter Stand der Sprache abgebildet, aktuelle Entwicklungen finden nur langsam ihren Weg in manuell erstellte Wörterbücher. Zum anderen ist es fast unmöglich, auch nur annähernd den Wortbestand einer Sprache vollständig zu erfassen, und innerhalb des Wortbestands alle Bedeutungsnuancen. Klassische Wörterbücher bieten zudem oft nur unzureichende Verwendungsbeispiele. (vgl. Herbst u. a. 2004b)

Die korpusgestützte Lexikografie kann bei einigen Problemen der klassischen Lexikografie Abhilfe schaffen. Korpora enthalten eine ganze Reihe von Verwendungsbeispiele für häufig wie auch für selten vorkommende Wörter und Wendungen. Die Abdeckungsbreite ist bei Korpora also sehr hoch. Zudem sind viele Neuwortschöpfungen in Texten aus dem Internet zeitnah nach ihrem Erscheinen in digitaler Form auffindbar. Gerade Korpora aus Internetseiten bilden also einen sehr aktuellen Sprachbestand ab.[20] Diese Punkte sprechen für die Verwendung von Korpora auch im Spezialfall der Valenzwörterbuchextraktion.

Aber nicht nur die Erstellung „klassischer" Valenzwörterbücher kann von der automatischen Extraktion von Valenzpaaren profitieren. Wie Briscoe

20 Auf ähnlichen Überlegungen baut auch die Initiative *The web as corpus (WebCorp)* auf, die als Recherchequelle alle frei verfügbaren Texte des Internets nutzt (http://www.webcorp.org.uk/)

(2001) bemerkt, gilt dies ebenso für computerlinguistische Anwendungen. Anwendungen aus diesem Feld können bei der Analyse und Synthese von Sprache maschinenlesbare Valenzbeschreibungen nutzen. So würden beispielsweise Parsern Valenzinformationen bei der Unterscheidung zwischen Aktanten und Modifikatoren helfen. Ebenso kann die Information zur lexikalischen Disambiguierung bei mehrdeutigen Wörtern dienen. Dieser Aspekt ist auch für die MÜ relevant. Informationen zu Entsprechungen zwischen bestimmten Valenzrahmen in zwei Sprachen würden beispielsweise die Auswahl von Lexemen in transferbasierten MÜ-Systemen verbessern. Bisher im MÜ-Systemen kodierte Lexika können durch die Extraktion aus Korpora in einigen Punkten verbessert gehen. So ist es bei der Verwendung registerkontrollierter Korpora möglich, automatisch extrahierte Einträge nach Register und ggf. Fachgebiet zu ordnen, was manuell einen großen Aufwand bedeutet.

In den letzten Jahren wurde bereits verschiedentlich der Versuch unternommen, Korpora als Grundlage für eine automatische Extraktion mono- und bilingualer Valenzwörterbücher zu nutzen, so z.B. aus der englischen *Penn Treebank* (Palmer u. a. 2001), aus der *Prague Dependency Treebank* (Žabokrtský 2005) und in framesemantischer Form aus der deutschen Baumbank *TiGer* (Burchardt u. a. 2006). Im Folgenden werden einige Ansätze vorgestellt, durch die ein breites Spektrum an Herangehensweisen illustriert werden kann. Abschnitt 2.4.1 beschreibt einen vollautomatischen Ansatz zur Extraktion von Valenzen aus englischen Texten. Abschnitt 2.4.2 stellt Extraktionsexperimente aus teilweise (aufwändig) manuell annotierten und alignierten parallelen tschechisch-englischen Daten vor. Während die ersten beiden Ansätze eher syntaktisch und in Richtung computerlinguistischer Anwendungen orientiert sind, zielt der Ansatz aus Abschnitt 2.4.3 auf eine semantische Dimension – die Erkennung von Kollokationen – und auf die Wörterbucherstellung für den Menschen ab.

2.4.1 Extraktionsexperimente mit monolingualen englischen Texten

Briscoe (2001) beschreibt einen Versuchsaufbau, mit dem aus englischen Texten Valenzwörterbücher extrahiert werden sollen. Der Versuchsaufbau basiert ausschließlich auf vollautomatischen Komponenten, sodass kein manueller Eingriff bei der Extraktion nötig ist.

Das Extraktionssystem besteht aus den folgenden Komponenten:

- einem HMM-Tagger,
- einem Lemmatizer,
- einem statistischen Parser,

- einem *pattern extractor*, der lokale Valenzrealisierungen extrahiert,
- einem *pattern classifier*, der lokale Realisierungsmuster einer Klasse von Valenzrahmen zuweist (oder das Muster ganz verwirft)
- und einem lexikalischen Filter, der für ein Lexem gesammelte Valenzrahmen zu Einträgen sortieren soll.

Die Ausgabe des Systems ist an das Format der Penn Treebank angelehnt, enthält also eine Klammerstruktur, die den syntaktischen Baum darstellt und eine Wortarten- sowie Phrasenkategorieannotation enthält. Aus dieser Klammerstruktur werden realisierte Valenzrahmen extrahiert. Aus dem Satz *He attributed his failure, he said, to no-one buying his book* extrahiert das System korrekt die Konstituenten *He*, *his failure* und *to no-one buying his book* als Begleiter des Verbs *attribute*. Der Autor merkt an, dass *he said* zwar fälschlicherweise als Postmodifikation von *failure* erkannt wurde, dies in diesem Fall dem Zweck der Extraktion allerdings dienlich ist, da *he said* damit als Kandidat für einen Verbbegleiter herausfällt.

Neben dem Attachment stellen die phrasalen Verben ein besonderes Problem dar. Briscoe (ebd.: 82) verweist auf das folgende Beispiel:

(a) *He looked up the word.*
(b) *He looked up the hill.*

Für den Parser ist es kaum möglich zu erkennen, in welchem der beiden Beispiele es sich um das phrasale Verb *look up* ‚nachschauen, nachschlagen' und in welchem um das Verb *look* ‚schauen' mit einem mit der Präposition *up* eingeleiteten Richtungsadverb handelt.

Eine zusätzliche Schwierigkeit stellt die Musterklassifizierung verschiedener Realisierungen eines Valenzrahmen zu einem „Proto-Rahmen" – also einem Rahmen, von dem sich die anderen Rahmen durch zusätzliche, freie Modifikatoren oder durch nicht realisierte, optionale Elemente unterscheiden – dar. Auch die Zuordnung von Frames zu Verblexemen und die Generierung geeigneter Valenzeinträge für das Lexikon sind schwierig.

Im Anschluss diskutiert Briscoe verschiedene statistische Methoden zur Klassifizierung der Valenzrahmen und zum Filtern der lexikalischen Einträge; die Spezifität im Bereich zwischen 65% und knapp 77% (allerdings bei einergeringen Sensitivität von ca. 35% bis zu knapp über 43%) bezeichnet er als vielversprechend. Seinen Optimismus bezüglich der automatischen Valenzextraktion aus Korpora untermauert Briscoe mit seiner Zukunftsvision von „self-organizing dictionaries" (ebd.: 86f.), die, mit Blick auf aktuelle

Projekte wie z.B. das Wortschatzprojekt der Universität Leipzig[21], inzwischen tatsächlich in greifbarer Nähe scheinen.

2.4.2 Parallele Valenzrahmenextraktion aus der Prague Czech-English Dependency Treebank

Die Annotation der *Prague Dependency Treebank* (kurz: PDT; Böhmova u. a. 2000) basiert auf der *Functional Generative Description* (FGD; vgl. Sgall u. a. 1986). Die PDT enthält zwei Ebenen der Dependenzannotation. Die analytische Ebene ist die Darstellung aller Wörter eines Satzes und der Abhängigkeiten zwischen Ihnen. Die tektogrammatische Ebene abstrahiert von vielen Eigenschaften der analytischen Ebene weg. Funktionswörter und sprachspezifische Eigenschaften wie Kasus werden in Merkmalen ausgedrückt. Diese bilden zusammen mit der Form und dem Lemma eines Wortes Wortknoten, welche zusätzlich mit dem jeweiligen Funktor, der die Beziehung zwischen Regent und Dependent angibt, angereichert sind.

Die *Prague Czech-English Dependency Treebank* (PCEDT; Čmejrek u. a. 2004) ist ein Ausschnitt der Penn Treebank, der um tschechische Übersetzungen erweitert wurde, und dessen Texte nach dem Muster der PDT annotiert wurden. Hierbei muss darauf hingewiesen werden, dass es sich bei der Übersetzung der englischen Texte ins Tschechische nicht um freie Übersetzungen oder publizierte Texte handelt, sondern um eine Auftragsarbeit, bei der die Übersetzer explizit angewiesen wurden, möglichst nah am englischen Original zu bleiben. Mit den Daten der PCEDT wurden zum einen dependenzbasierte MÜ-Experimente mithilfe bereits vorhandener Wörterbuchressourcen durchgeführt (Čmejrek u. a. 2003), wie auch Versuche, übersetzungsrelevante Verbvalenzpaare zu extrahieren (Bojar und Hajič 2005). Aktuell finden Arbeiten an der automatischen statistischen Alignierung der Dependenzbäume (Mareček u. a. 2008) sowie an neueren MÜ-Ansätzen statt (vgl. Abschnitt 2.5.6).

Bojar und Hajič (2005) bezeichnen die Übersetzung anhand von Verbvalenzrahmen als **structural transfer**, und grenzen diese Art des Transfers von statistischen Transfermethoden ab. Es sei an dieser Stelle bereits erwähnt, dass auch im Rahmen der MÜ anhand der PCEDT dieser Gegensatz nicht bestehen bleibt und bei späteren Experimenten eine statistische Komponente hinzukommt. Die PCEDT ist auf der Wortebene aligniert, wobei dies durch die Dependenzdarstellung einer Teilbaumalignierung gleichkommt, wie sie z.B. bei der MÜ mittels Dependenzteilbäumen, der sog. **dependency treelet**

21 http://wortschatz.uni-leipzig.de

translation (Quirk u. a. 2005; Ding und Palmer 2005) angewandt wird. Diese Form der MÜ orientiert sich an aktuellsten Tendenzen, bei der Alignierung und Übersetzung die Wortebene zu verlassen und auf die Phrasenebene zu gehen (vgl. z.B. Franz u. a. 2003). Dass sich allerdings Dependenzbäume im Sinne der phrasalen Kohäsion besser für die MÜ eignen als Phrasenstrukturbäume, zeigt Fox (2002).

In den von Bojar und Hajič (ebd.) beschriebenen Experimenten zur Extraktion paralleler Valenzpaare werden aus den Dependenzbäumen der PCEDT Verbpaare mit je einer Liste von „modifications", d.h. mit einer Liste der von ihnen abhängigen Funktoren extrahiert, wobei die Listen bei Original und Übersetzung gleich lang sein müssen. Zur Evaluierung der Qualität der extrahierten Verbentsprechungen wird ein MÜ-Experiment durchgeführt. Dabei kommen verschiedene Auswahl- und Übersetzungskriterien zur Anwendung. Einmal werden Valenzrahmen mit verschiedenen Methoden gefiltert:

(a) Entweder es kommen alle extrahierten Rahmenpaare zum Einsatz,
(b) oder die niederfrequenten werden aussortiert,
(c) oder Alignierungen mit einem hohen GIZA++-Konfidenzwert werden verwendet,
(d) oder nur Rahmenpaare aus sehr einfachen Sätzen werden verwendet.

Diese Möglichkeiten werden jeweils mir jeder der folgenden drei Übersetzungsstrategien kombiniert:

(e) mit einer Argument-für-Argument-Übersetzung mit statistischen Daten aus dem ganzen Wörterbuch,
(f) mit einer Argument-für-Argument-Übersetzung mit statistischen Daten für das AS-Verb und alle möglichen ZS-Verben und
(g) mit einer gesamten Übersetzung des Rahmens ausschließlich anhand der Daten nur des ausgesuchten Rahmenpaars.

Das beste Ergebnis erreicht nach Angaben von Bojar und Hajič die Kombination aus (c) und (e), mit einer Spezifität von 52,9% und einer Sensitivität von 96,7%.

Die Ausgangsbedingungen des Experiments von Bojar und Hajič sind in einigen Punkten besser als die für das Experiment der vorliegenden Arbeit. Die Sätze sind tiefer analysiert, und valente Lexeme (d.h. nicht nur Verben, sondern auch Adjektive und Nomen) sind anhand der dependenziellen Annotation leichter zu extrahieren. Die dependenzielle Annotation erlaubt zudem ein deutlich einfacheres Auslesen der Begleiter eines valenten Lexems, da die realisierten Leerstellen des Verbs als direkte Dependenten am Verb hängen.

Zugleich unterliegt dieses frühe Experiment auf Grundlage der PCEDT aber einigen Einschränkungen. Kritisch anzumerken ist, dass die Übersetzun-

gen ins Tschechische keine publikumsgerichteten, publizierten Übersetzungen sind, sondern diese mit dem Ziel, sie für die MÜ zu verwenden, und mit der Anweisung, sie möglichst nah am Original zu halten, angefertigt wurden.

Hervorzuheben ist ein weiterer Punkt, der auch für diese Arbeit gilt: Bojar und Hajič weisen darauf hin, dass ihr Ziel nicht darin bestand, Aktanten und Modifikatoren zu unterscheiden, sondern anhand der Belege zu bestimmen, wie ein Argument syntaktisch in der ZS auszudrücken ist.

2.4.3 Kollokationen und Verbentsprechungen aus EUROPARL

Duffner u. a. (2009) verwenden das EUROPARL-Subkorpus aus dem *open source parallel corpus* (OPUS; Tiedemann und Nygaard 2004), um daraus Kollokationen sowie Übersetzungsentsprechungen zu extrahieren. Diese Kollokationen dienen quasi als semantische Valenzen. Da sie aber auf reinen Korpusdaten basieren, handelt es sich um keine erschöpfende semantische Beschreibung (z.B. durch Beschreibung der Selektionspräferenzen anhand von semantischen Merkmalen), sondern eben um digitale Wörterbücher mit (typischen) Verwendungsmustern.

Die Pilotstudie von Duffner u. a. basiert auf der EUROPARL-Abfrage von drei Verbformen des Verbs *einstellen*: *einstellen*, *eingestellt* und *einzustellen* mit den jeweiligen Alignierungen mit den Sprachen Französisch, Italienisch und Englisch. Die deutschen Abfrageergebnisse werden einer der laut dem *Deutschen Referenzkorpus*[22] acht häufigsten Bedeutungen von *einstellen* zugeordnet, so z.B. *mit etw. aufhören* oder *jdn. in ein Arbeitsverhältnis aufnehmen*. Die häufigsten Übersetzungsäquivalente für diese beiden Bedeutungen im Englischen sind *stop* und *cease* bzw. *employ* und *recruit*. Anhand der durch die Alignierungsabfrage gewonnenen englischen Sätze werden zusätzlich die möglichen Kollokate für die jeweiligen Verben extrahiert, so z.B.:

- *stop*: *paying*, *giving*, *funding*; *aid*; *executions*, *activities*
- *cease*: *hostilities*

Die so gewonnenen Übersetzungseinträge und Kollokationslisten können für ein gedrucktes Wörterbuch verwendet werden.

In ihrem Beitrag gehen Duffner u. a. auch auf aktuelle Printwerke, wie z.B. den Larousse deutsch-französisch von 1999, ein. Verglichen mit den Ergebnissen ihres Extraktionsansatzes, so Duffner u. a. (ebd.), halten die Einträge des Larousse (und vergleichbarer Werke) weder mit der Bedeutungsdifferenzierung noch mit der Auswahl der ZS-Lexeme Schritt. Dies ist als Kritik

22 http://www.ids-mannheim.de/kl/projekte/dereko_I/

an der mangelnden korpusgestützten Arbeit von Wörterbuchverlagen zu verstehen. Inwiefern Wörterbuchverlage tatsächlich korpusgestützt arbeiten, kann an dieser Stelle nicht erörtert werden; die von Duffner u. a. angesprochenen Schwächen sind aber dem regelmäßigen Nutzer von Printwörterbüchern durchaus bekannt.
Dieser Ansatz zeigt eine Möglichkeit auf, wie in Zukunft die syntaktische Extraktion auf semantischer Ebene durch zusätzliche kollokative Informationen ergänzt werden kann. Als Orientierung dafür können verschiedene Ansätzen zur Kollokationsextraktion für die HÜ, die MÜ und die Computerlinguistik im Allgemeinen (vgl. z.B. Nerima u. a. 2003; Orliac und Dillinger 2003; Vintar und Fišer 2008) dienen.

Allerdings weisen Duffner u. a. auch auf ein mit automatischen Mitteln bisher nur schwer zu lösendes Problem ihres Ansatzes – und damit potenziell anderer Ansätze – hin: die Zuordnung von polysemen Verben zu einer ihrer Unterbedeutungen.

2.4.4 Übersicht

Die in den vorangehenden Abschnitten vorgestellten Extraktionsexperimente für Valenzinformationen sind in ihrer Zielsetzung und ihrem Aufbau sehr unterschiedlich, wie Tabelle 2.3 demonstriert.

Autor	**Korpus**	**Annotation**	**Ziel**
Briscoe	eigens zusammengestellt	vollautomatisches Tagging und Parsing, statistische Klassifikation von Begleitern und Valenzrahmen	„self-organizing dictionaries“ für MÜ und Lexikografie
Bojar und Hajič	PCEDT	automatische Vorverarbeitung, aufwändige manuelle Dependenzannotation	Transferlexika für MÜ
Duffner u. a.	EUROPARL	keine; statistischer word-window-Ansatz	bilinguale Wörterbücher

Tabelle 2.3: Übersicht der Ansätze zur Extraktion von Valenzwörterbüchern aus Korpora

Briscoe experimentiert mit der vollautomatischen Extraktion von syntaktischen Valenzinformationen aus einem monolingualen englischen Textkorpus

und setzt dafür automatische syntaktische Analysewerkzeuge sowie statistische Algorithmen zur Argumenterkennung sowie zur Disambiguierung von Verbbedeutungen ein. Die Extraktion aus der PCEDT wird einem manuell annotierten parallelen Dependenzkorpus durchgeführt. Die Beschreibung der Valenzrahmen baut auf der FGD-Theorie der Prager Schule auf. Die Ergebnisse der Extraktion werden für MÜ-Experimente eingesetzt. Duffner u. a. klammern bei der Extraktion von Kollokationen aus EUROPARL syntaktische Informationen völlig aus. Ihr Ziel ist es, bilinguale Wörterbücher mit hilfreichen Informationen über die verschiedenen Bedeutungen der darin enthaltenen Lexeme und deren möglichen Kollokate auszustatten.

Alle drei Ansätze sind für die vorliegende Arbeit in der einen oder anderen Weise richtungweisend. Briscoes Ansatz demonstriert, dass eine automatische Extraktion durchaus möglich ist, wenn dies auch mit Schwierigkeiten bei der Genauigkeit verbunden ist. Die Ansätze von Bojar und Hajič sowie von Duffner u. a. haben hinsichtlich ihrer Zielsetzung sowie der Anwendbarkeit der Ergebnisse Vorbildcharakter und weisen den Weg für weitergehende Forschung.

2.5 Valenz in der maschinellen Übersetzung

Valenz spielte in der maschinellen Übersetzung schon früh eine große Rolle. Schon das Saarbrücker Übersetzungssystem SUSY, entstanden in den 70er Jahren des 20. Jahrhunderts, verstand sich darauf, beim Übertrag aus einer Sprache in eine andere dasjenige Lexem in der ZS herauszusuchen, das am besten zu den Argumenten, die im AS-Satz vorkamen, passt. In EUROTRA (Johnson u. a. 1985), METAL (Bennett und Slocum 1988) oder CAT2 (Sharp 1994; Haller 1993) werden diese Mechanismen ebenfalls verwendet. Dabei liegt häufig die Annahme zugrunde, dass Valenzrahmen bzw. zumindest tiefe semantische Rollen für äquivalente Verben in verschiedenen Sprachen gleich sind (vgl. auch Abschnitt 4.1).

In den folgenden Abschnitten wird eine Auswahl an MÜ-Systemen vorgestellt. Die Auswahl wurde so gestaltet, dass eine große Bandbreite von Mechanismen und Formalismen abgedeckt wird. Der Fokus richtet sich bei der Darstellung auf die Valenzmechanismen der Systeme. Die Systeme sind chronologisch, beginnend mit dem ältesten, geordnet.

2.5.1 Valenz in SUSY

Das MÜ-System SUSY[23] hat seine Wurzeln in den 70er Jahren des 20. Jahrhunderts und wurde im Rahmen des Sonderforschungsbereichs 100 an der Universität des Saarlandes entwickelt. Die technischen Angaben zu SUSY in diesem Abschnitt stammen aus den Beschreibungen von Luckhardt und Maas (1983a) sowie Maas (1984). Der geschichtliche Überblick wurde (Hutchins und Somers 1992) entnommen.

Die erste Version von SUSY, SUSY I, wurde zu Beginn der 1980er Jahre durch SUSY II abgelöst, das einige Innovationen wie z.B. einen Interpretationsmechanismus für vom Nutzer geschriebene Regeln – in Abgrenzung zur **Software** häufig als **Lingware** bezeichnet – lieferte, während im Vorgängersystem die Regeln noch fest in die Programmquellen integriert waren. Die Arbeiten an SUSY gingen 1986 im europäischen MÜ-Projekt EUROTRA auf.

SUSY ist ein multilinguales transferbasiertes System. Analyse, Transfer und Synthese sind prinzipiell unabhängig voneinander, wobei gerade SUSY I auf sprach(enpaar)spezifische Module anstelle von z.B. verschiedenen Regelsätzen setzt. SUSY setzt sich aus einer Vielzahl von Modulen zusammen, die teilweise kryptische Namen wie LESEN, WOBUSU, SEDAM oder KOMA tragen.

Die Analyse läuft – vereinfacht dargestellt – wie folgt ab. LESEN übernimmt die Segmentierung in Wörter und Sätze. Jedes Wort wird in eine Datenstruktur mit einer bestimmten Anzahl von Feldern gepackt; jedes Feld enthält eine Eigenschaft, wie z.B. Genus oder Numerus. WOBUSU erzeugt eine morphologische Analyse. DIHOM löst ambige Homographen auf. Die Module VERBA und NOMA versuchen, Verbal- und Nominalgruppen zu erkennen. SEDAM disambiguiert auf Basis semantischer Informationen. KOMA baut aus den ihm zur Verfügung stehenden Analysen und mithilfe der Valenzinformation für das Hauptverb des Satzes eine Dependenzstruktur. TRANSFER überführt die Dependenzstruktur von der Quellsprache (QS) in die Zielsprache (ZS) und ersetzt QS- durch ZS-Lexeme. Hier wird wieder die Valenzinformation eingesetzt, um Lesarten von Verben zu disambiguieren und die Auswahl des ZS-Lexems zu steuern. Die Module SEMSYN, SYNSYN und MORSYN bauen jeweils die semantische, syntaktische und morphologische Struktur des ZS-Satzes; MORSYN generiert schließlich den Ausgabestring.

23 Von SUSY gab es zwei Programmversionen. Die Bezeichnung SUSY umfasst die beiden Versionen des Systems, SUSY I und SUSY II. Wo spezifisch auf eine der beiden Versionen eingegangen wird, wird die entsprechende Versionsbezeichnung verwendet.

Die Valenzinformation für den TRANSFER ist über eine 7*4-Byte-Matrix im Transferwörterbuch kodiert (Luckhardt und Maas 1983b:84). So gibt es z.B. für Zeile 4, Byte 1 folgende vier möglichen Werte:

- 0: keine Änderung
- 1: AS nicht reflexiv, ZS obl. reflexiv, z.B. *apologize – sich entschuldigen*
- 2: AS obl. reflexiv, ZS nicht reflexiv, z.B. *sich entschuldigen - apologize*
- 3: AS reflexiv, in der ZS an dieser Stelle ein Possessivpronomen, z.B. *Er wäscht sich das Gesicht - He washes his face.*

Valenz wird im Lexikon an dieser, aber auch an anderen Stellen, hauptsächlich nach syntaktischen Gesichtspunkten behandelt.

Zusammenfassend kann man feststellen, dass bei SUSY die Stellung des Verbs hervorgehoben ist. Dies drückt sich bereits durch den für den Transfer verwendeten Dependenzbaum aus, dessen Wurzel i.d.R. ein Verb ist. Zudem wird an verschiedenen Stellen Valenzinformation verwendet, beispielsweise um den Dependenzbaum zusammenzustellen, oder etwa bei der Auswahl des ZS-Lexems während des Transfers.

2.5.2 Das METAL-Leuven-Framework

Das System METAL (von *MEchanical Translation and Analysis of Language*; Bennett und Slocum 1988) wurde zunächst an der Universität von Texas in Zusammenarbeit mit dem amerikanischen Militär entwickelt wurde. Mit diesem Projekt sollte fundierte linguistische Forschung insbesondere für das Sprachenpaar Deutsch-Englisch gefördert werden. Im Jahr 1978 stieg SIEMENS in das Konsortium der Geldgeber ein, war ab 1980 alleiniger Geldgeber und prägte den Namen METAL. Das System sollte die Produktivität des Sprachendienstes bei Siemens erhöhen.

Im ursprünglichen Rahmen wurde METAL bis 1992 entwickelt. Später wurde es von der belgischen Firma Lernout & Hauspie akquiriert, welche aber Ende der 1990er Jahre Insolvenz anmeldete. Der Kern von METAL ist bis heute noch im Einsatz, so z.B. im T1 Übersetzungssystem, das u.a. für Privatverbraucher von Langenscheidt vertrieben wird[24], oder auch im LUCY-

24 http://www.langenscheidt.de/katalog/reihe_langenscheidt_t_volltextuebersetzer_version__625_0.html

Übersetzungssystem[25], das u.a. im von der EU geförderten MÜ-Forschungsprojekt EuroMatrixPlus[26] eingesetzt wird.

METAL entspricht modernen Anforderung an ein MÜ-System dahingehend, dass Software und Lingware getrennt sind. Analyse- und Synthesemodule entsprechen einander weitgehend und verwenden dieselbe Information. Sowohl Software als auch Lingware von METAL sind in LISP kodiert.

Das in Leuven entwickelte Valenzframework für METAL beschreibt Gebruers (1988). Das Leuven-Framework stellte eine Überarbeitung des bis dahin gebräuchlichen Frameworks dar. Es sollte einen verbesserten Rahmen für die mehrsprachige Entwicklung des Systems bieten. Laut Gebruers funktionierende das Vorläufersystem insbesondere für die beiden verwandten Sprachen Deutsch-Englisch gut, war aber nur bedingt erweiterbar. Folgende Darstellung speist sich größtenteils aus Gebruers' Bericht, ist aber durch Informationen aus (Hutchins und Somers 1992) ergänzt.

Das erste Valenzframework von METAL ging davon aus, dass Argumentpositionen von Verben über (mindestens die europäischen) Sprachen hinweg immer in gleicher Zahl vorhanden, und höchstens syntaktisch oder typsemantisch verschieden sind. Daher gab es eine kanonische Ordnung von Argumenttypen, die von einer Sprache in die andere direkt übertragen werden kann. Unterschiede zwischen AS und ZS in Argumenttyp und -anordnung wurden von einer kleinen Menge an Zusatzregeln behandelt. Ein eigenes Transfermodul bestand als solches nicht. Jenes Framework arbeitete zudem mit einer flachen syntaktischen Analyse, semantische Information wurde kaum verwendet. Die Optionalität von Argumenten wurde nicht bei den Argumenten selbst vermerkt, sondern war beim Verb mittels eines Merkmals kodiert.

Das Leuven-Framework führt ein eigenes Transfermodul ein, das lexikalisch gesteuert von der Ausgangs- in die Zielsprache Valenzrahmen aufeinander abbildet. Diese Architektur geht von folgenden Annahmen aus, die als Grundlage für ein genuin multilinguales, transferbasiertes MÜ-System angesehen werden können:

- Die Analyse muss eine Repräsentation liefern, die für den Transfer in mehrere ZS geeignet ist.
- Die Syntheseprozeduren der ZS müssen immer die gleichen sein, unabhängig von der AS.
- Die Abbildung von AS auf ZS darf nur ein Minimum an lexikalisch motivierten Änderungen enthalten.

25 http://www.lucysoftware.com/
26 http://www.euromatrixplus.org/

Das Leuven-Framework setzt diese Anforderungen wie folgt um. Linguistisch gesehen wird jeder Satz als **Prädikat-Argument-Struktur** dargestellt.[27] Als **Frame** bezeichnet Gebruers eine Sequenz von getypten Argumentstellen, die einem lexikalischen Prädikat zugeordnet ist; jedem Lexem kann mehr als ein Frame zugeordnet werden. **Satelliten**, d.h. im Lexikon nicht spezifizierte Modifikatoren, sind immer erlaubt. Optionale Argumente werden mit dem Schlüsselwort `OPT` markiert. Existieren für ein Lexem mehrere Frames, die sich nur in den fakultativen Argumenten unterscheiden, so werden diese Frames in einen Frame kollabiert, der die Gesamtmenge der möglichen fakultativen Elemente enthält. Argumente sind mit Tiefenkasus belegt, so steht z.B. `$0` für das tiefe Subjekt, `$2` für indirekte Objekte etc. Die Argumentstellen enthalten zusätzlich Information über morpho-syntaktische Beschränkungen, sowie u.U. Informationen über den semantischen Typ des Arguments aus einer sehr limitierten Liste von semantischen Featuren. **Frame-tests** verwenden diese Information bei der Analyse, **frame-constructors** wiederum bei der Synthese.

Die Valenzprozedur versucht, nachdem das Prädikat und die Komponenten analysiert worden sind, zunächst den am besten auf den vorliegenden Satz passenden Frame zu finden. Wird keine solche Entsprechung gefunden, werden alle mit dem ASFrame verlinkten ZS-Frames abgerufen. Die morpho-syntaktischen sowie semantischen Informationen aus dem AS-Frame werden mit denen jedes potenziellen ZSFrames abgeglichen; ausgewählt wird dann der Frame mit der höchsten Übereinstimmung. Zusätzlich kann der Lexikograf mittels eines `PREF`-Werts zwischen 1 und 100 die Auswahl des besten Frames beeinflussen. Ist dieser Wert vorhanden, wird der ZS-Frame mit dem höchsten Wert verwendet.

METAL verwendet einen dezidierten Valenzmechanismus zum Transfer, bei dem der Linguist einen starken Einfluss auf die Auswahl beim Transfer hat. Denkbar ist, diese Daten mit aus Korpora gewonnenen Daten zu ergänzen, was den hohen manuellen Aufwand bei der Erstellung der Transferlexika stark reduzieren würde.

2.5.3 Das rollenbasierte Modell von EUROTRA und die Interlingua von Dorr

Zwischen interlingua- und transferbasierten Systemen ist die Grenze nicht immer ganz eindeutig: Eine Interlingua ist nichts anderes als eine Transfer-

27 Gemäß der Terminologie von Gebruers (1988) sind Prädikat und Argument eher syntaktisch bzw. syntakto-semantisch (in etwa vergleichbar mit dem Begriff des Tiefenkasus) aufzufassen, und weniger im Sinne der logischen Valenz wie bei Helbig (1992).

sprache, die den Anforderungen aller bekannten Sprachen genügt. Der Interlingua-Ansatz für die MÜ von Dorr (1994) erinnert in einigen Punkten an den Ansatz wie er im transferbasierten MÜ-System EUROTRA (Johnson u. a. 1985) praktiziert wurde.

Die Planung für das EUROTRA-System begann schon in den 1970er Jahren, als klar geworden war, dass es schwierig sein würde, das in der Europäischen Gemeinschaft verwendete Übersetzungssystem SYSTRAN für die Übersetzung aus und in alle EG-Sprachen (neben Englisch und Französisch u.a. auch Deutsch, Griechisch oder Italienisch) zu erweitern. Das EUROTRA-Projekt sollte einen Prototypen für eine spätere industrielle Weiterentwicklung liefern. Ein genauer geschichtlicher Überblick ist (Hutchins und Somers 1992) zu entnehmen.

EUROTRA ist nicht per se ein Interlingua-basiertes System. Teil von EUROTRA ist allerdings eine **Interface Structure** (IS), die so „universell" gehalten ist, dass der Transfer zwischen allen zur Zeit von EUROTRA offiziellen damaligen EG-Sprachen möglich ist. Wie die einzelsprachliche Analyse und Generierung zu verlaufen hatte, wurde den am Projekt beteiligten Instituten aus bis zu 18 Nationen nicht vorgeschrieben. Wichtig war nur, dass die IS als Ziel- bzw. Ausgangspunkt der Analyse bzw. Generierung diente. Die IS besteht aus einer dependenziellen Struktur, deren Knoten flache Merkmalsbündel sind, die morphologische, syntaktische und semantische Informationen enthalten. Das System definiert eine Reihe von Bedingungen für diese Merkmalsbündel, die erfüllt werden müssen, bevor der Transfer durchgeführt werden kann. Das Modell für diese Merkmalsbündel muss dabei komplexe Zusammenhänge berücksichtigen, da in den europäischen Sprachen linguistische Eigenschaften wie etwa Aspekt oder Bekanntheit sehr unterschiedlich ausgedrückt werden können. Beim Übertrag von einer AS in die ZS müssen aber alle nötigen, manchmal impliziten Informationen zur Verfügung stehen.[28]

Die Prädikat-Argument-Struktur spielt bei einer dependenziell angelegten IS natürlich eine wichtige Rolle. Die Beiträge in (Steiner 1989) geben einen Überblick über die Mechanismen, wie die Analyse und Generierung von Prädikaten und ihrer Argumente abläuft. Schmidt u. a. (1988) beschreiben, von welche Merkmale diese Prozesse gesteuert werden. An dieser Stelle soll nur auf die grundlegenden Aspekte der Prädikat-Argument-Struktur eingegangen werden.

Den Argumenten eines Lexems sind Rollen zugeordnet. Lexeme, die einander entsprechen, haben über Sprachen hinweg die gleiche Anzahl von Rol-

28 Z.B. drückt das Deutsche im Gegensatz zum Englischen den Verlaufsaspekt einer Handlung nicht grammatisch aus, sondern höchstens mit einem Adverb wie *gerade*.

len. Wie bzw. ob die Rollen in der jeweiligen Sprache realisiert werden, bestimmen die Module der jeweiligen Sprache. EUROTRA nutzt das bereits erwähnte komplexe System semantischer Merkmale auch bei den Prädikat-Argument-Strukturen, um sprachabhängige Aspekte wie etwa Kasus oder die (Nicht)Realisierung von Personalpronomen oder Artikeln zu steuern.

Jede Verblesart wird einem bestimmten semantischen Typ zugeordnet, woraus eine mit dem Typ verknüpfte Rollenkonfiguration resultiert. So hat z. B. das Verb *gehen* in der Wendung *gut gehen* folgende Analyse (vgl. Steiner u. a. 1988: 52):

gehen:	ATTRIBUANT	CLASSIFIER
	C3	C8

ATTRIBUANT und CLASSIFIER sind dabei semantische Klassen, die Abkürzungen C3 und C8 stehen für syntaktische Kategorien, in diesem Fall für eine Dativ-NP bzw. für eine Adjektivphrase. In einem Satz wie *Dem Lehrer geht es gut* erhielte das *es* die Kategorie C0 für Nominativ-NP, aber keine semantische Rolle. Interessant ist in diesem Fall die Realisierung im Englischen. Steiner u.a. geben als englische Übersetzung des eben genannten Beispielsatzes *The teacher feels well*. Hier steht der ATTRIBUANT in der Subjektposition und ist nicht Objekt wie im Deutschen. Die deutsche Adjektivphrase, der CLASSIFIER, ist mit einem Adverb im Englischen wiedergegeben.

Das multilingual angelegte Rollensystem in Verbindung mit semantischen Merkmalen, die EUROTRA bei der Analyse und Generierung erzeugt bzw. verwendet, lassen EUROTRA näher an ein Interlingua-System heranreichen als viele andere, stark syntaktisch orientierte Transfersysteme (wie z.B. SUSY oder METAL). Das Prinzip der Rollen als sprachübergreifendes Argumentstrukturmodell bleibt auch in einer Seitenlinie von EUROTRA, dem MÜ-System CAT2 (Haller 1993; Sharp 1994; Streiter 1995), erhalten.

In ihrem richtungweisenden Aufsatz analysiert Dorr (1994) **Übersetzungsdivergenzen** und entwirft ein Modell, das solche Divergenzen mithilfe einer Interlingua-Beschreibung überbrücken soll. Die Interlingua, die Dorr für ihre Formalisierung verwendet, ist nach ihren Angaben an eine Repräsentationsform von Jackendoff (1990) angelehnt; sie wird als *lexical conceptual structure* (LCS) bezeichnet. Die Abbildung eines Satzes auf seine LCS-Repräsentation basiert auf den Phrasen des Satzes. Die Phrasen bilden die Sinnuntereinheiten, deren Beziehungen zueinander mittels der LCS abgebildet werden.

Das von Dorr entworfene Modell wurde im MÜ-System UNITRAN (Dorr 1992) praktisch umgesetzt.

Dorr definiert und klassifiziert eine Reihe von semantischen Übersetzungsdivergenzen auf der lexikalisch-semantischen Ebene, und formalisiert diese mittels der Interlingua. Divergenzen werden dabei als Abbildungsdivergenzen zwischen syntaktischer Struktur und LCS verstanden. So werden Divergenzen aus der Sprache-zu-Sprache-Beziehung genommen und auf einer abstrakteren Ebene behandelt. Beim Übertrag von einer Sprache in die andere sind also die Divergenzen immer zwischen AS und Interlingua sowie Interlingua und ZS zu überbrücken, aber keine direkten AS-ZS-Divergenzen.

$$[\ _{Event}\mathrm{GO}_{Loc}([\ _{Thing}\mathrm{JOHN}],[\ _{Path}\mathrm{TO\text{-}SCHOOL}_{Loc}],[\ _{Manner}\mathrm{HAPPILY}])]$$

Abbildung 2.6: Vereinfachte LCS für das Ereignis gehen *(adaptiert von Dorr 1994: 601)*

Auch wenn das Modell von Dorr nicht explizit als valenzbasiertes Modell eingeführt wird, so trägt es doch einige Grundzüge davon. Wurzel der LCS-Struktur ist das Ereignis, das der Satz beschreibt, wie etwa die LCS für das Ereignis *GO* in Abbildung 2.6. Diesem Ereignis sind eine Reihe von Rollen zugeordnet, wie etwa *Thing*, *Path* oder *Manner*. Diese Struktur – man könnte es Prädikat-Argument-Struktur nennen – dient als Grundlage für den Transfer. Ereignisse haben eine unterspezifizierte Struktur, quasi einen Typ, der mit einem Lexikoneintrag bzw. einer syntaktischen Struktur zu einer konkreten Struktur, quasi einem Token, unifiziert werden kann (vgl. Abbildung 2.6). Dort, wo in einer Sprache eine Realisierung idiosynkratische Merkmale trägt, etwa wenn ein Prädikat wie *like* als Adverb ausgedrückt wird (im Deutschen typischerweise *gern* wie in *Ich schwimme gern*), wird dies im Lexikoneintrag durch ein Feature, in diesem Fall :DEMOTE ausgedrückt. Dabei handelt es sich dann um eine „demotional divergence“, eine Divergenz, bei der das Prädikat von seiner Prädikatsposition in eine andere Position bewegt wird.

2.5.4 LFG-Modelle

Ein LFG-basiertes Modell zur maschinellen Übersetzung beschreiben Kaplan u. a. (1989). Die Autoren grenzen das LFG-Modell von interlingua- und transferbasierten Systemen ab. Sie begründen dies damit, dass Interlingua- und Transfersysteme den Transfer immer nur auf einer Ebene vonstatten gehen lassen, während ihr LFG-Modell den Transfer auf mehreren strukturellen

Ebenen durchführt. Dies ist möglich durch die prinzipiell mehrschichtige Darstellung grammatischer Strukturen in LFG, wie in Abschnitt 2.2.2 beschrieben. Der Transfer geschieht im LFG-MÜ-Modell mittels neuer Operatoren, die zum bisherigen Operator φ dazukommen, der zwischen c-structure und f-structure abbildet. Der Operator τ übersetzt von der AS-f-structure in die ZS-f-structure, der Operator τ' von der AS-c-structure in die ZS-c-structure. Es werden verschiedentlich, auch in Folgebeiträgen, noch weitere Übersetzungsoperatoren vorgeschlagen, z.B. χ als Übersetzungsoperator für zeitliche Relationen.

Das LFG-Modell ist seither zur Verwendung mit Korpora weiterentwickelt worden. Eine der neueren Entwicklungen ist der Einsatz des LFG-Modells in der beispielbasierten MÜ. Way (2003) entwickelte ein bereits bestehendes Modell zur DOT (*data oriented translation*; Poutsma 2000) weiter. In diesem Modell werden aus Beispielübersetzungen auf verschiedenen Ebenen Fragmente extrahiert, die als äquivalent gelten können und mit LFG-Mechanismen mittels verschiedener Übersetzungsoperatoren wieder zusammengesetzt werden.

Tatsächlich erscheint der Ansatz einer mehrschichtigen Übersetzung zunächst neu; in ihrer Kritik übersehen Kaplan u. a. allerdings, dass zumindest viele transferbasierte Modelle zwar nur auf einer bestimmten Ebene (möglichst nahe der Semantik) transferieren, aber zuvor viele Informationen aus anderen Ebenen (Morphologie, Syntax usw.) aggregieren und damit informationsreicher sind als es zunächst erscheinen mag. Der stark ausgeprägte Forschungszweig des *data oriented parsing* und der DOP in der LFG-Community eröffnet allerdings unzweifelhaft viele Möglichkeiten für die beispielbasierte MÜ.

2.5.5 HPSG-Modelle

Das wahrscheinlich größte MÜ-Projekt, das HPSG als Analyse- und Generierungsformalismus verwendete, war das Verbmobil-Projekt. Das von Wahlster herausgegebene Kompendium (Wahlster 2000) gibt einen sehr umfassenden Überblick über die Systemarchitektur von Verbmobil und dient als Hauptquelle für dessen Beschreibung in diesem Abschnitt.

Verbmobil sticht aus den hier vorgestellten MÜ-Projekten insofern heraus, als dass es als einziges Projekt zum Ziel hatte, gesprochene Sprache automatisch zu übersetzen. Die Sprachen des daraus hervorgegangenen Verbmobil-Systems sind Deutsch, Englisch und Japanisch. Die Domäne umfasst lediglich Terminabsprachen für Geschäftsleute, aber der damit verbundene begrenzte Umfang an Lexik und Grammatik soll nicht über die Herausforderungen hinwegtäuschen, die die gesprochene Sprache an ein Computersystem

stellt. Zum einen ist die Erkennung gesprochener Sprache schwierig, da z.B. Hintergrundgeräusche einen schlechten Einfluss auf die Tonqualität haben können, aber auch, weil nicht jeder Sprecher auf die gleiche Art und Weise spricht (Dialekte, idiosynkratische Lauterzeugung etc.). Zum anderen ist gesprochene Sprache sehr viel spontaner als geschriebene Sprache und enthält daher Elemente wie Pausen, Füllwörter (*ähm*, *naja*) oder Selbstkorrekturen (*Sechs, Moment, ich meinte sieben Uhr!*).

Der Fokus des Verbmobilprojekts war, die gesprochene Sprache vor allem semantisch möglichst korrekt zu interpretieren, ungeachtet möglicher grammatischer Fehler der gesprochenen Äußerungen. Daher setzt das Verbmobil-System drei Systeme parallel ein: einen HPSG-Parser, der eine sehr detaillierte Analyse liefert, und zwei sehr viel robustere Systeme, nämlich einen statistischen LR-Parser und einen Chunkparser. Damit können ggf. einzelne Teile einer Äußerung erfasst und übersetzt werden. Der HPSG-Parser bringt aber gegenüber den anderen beiden Parsern einen entscheidenden Vorteil mit sich: Während die Ergebnisse des LR-Parsers und des Chunkparsers vor dem Transfer durch ein zusätzliches semantisches Generierungssystem verarbeitet werden müssen, generieren die HPSG-Module bei erfolgreichem Parsen bereits eine vollständige semantische Analyse mittels der in HPSG gebräuchlichen **minimal recursion semantics** (Copestake u. a. 1995).

Zu Beginn des Projekts wurden Dialoge in den Sprachen Englisch, Deutsch und Japanisch aufgenommen und analysiert. Die dabei entstandenen Baumbanken dienten zum Training des LR- und des Chunkparsers, zur Entwicklung des Moduls, das die Semantik generiert, sowie zur Entwicklung von Transferregeln.

Die Transferregeln des Verbmobils operieren hauptsächlich auf den semantischen Informationen der sprachlichen Analyse. Lexeme sind als Prädikate angegeben, die eine bestimmte Zahl und Art von Argumenten abbinden, wie etwa in der Transferregel

```
spät(L,I) → late(L,I).
```

Diese verbindet das deutsche Adjektiv *spät* mit dem englischen Adjektiv *late* (wie in *später Zug* bzw. *late train*). Die beiden Adjektive haben dieselbe Anzahl von Argumentinstanzen *I*, die mit Labeln *L* versehen werden. In dieser Transferregel wird die Argumentstruktur also einfach von der AS in die ZS kopiert. Der Transfer kann auch von Bedingungen abhängig gemacht werden, wie etwa in der Regel

```
besetzt(L,I) ← busy(L,I) # (telephone_n(_,I);
                line_n(_,I)).
```

Das Lexem *busy* wird also nur als *besetzt* übersetzt, wenn die Instanz, die *busy* modifiziert, ein Telefon (*telephone*) als Nomen (*_n*) oder eine Leitung (*line*), ebenfalls als Nomen, ist. Diese Regel zeigt, dass neben semantischen Kriterien auch syntaktische bzw. morphologische Kriterien eine Rolle spielen können.

Neben bedingungsabhängigen Übersetzungen gibt es noch Regelmechanismen, die einer Übersetzung ggf. noch etwas hinzufügen (oder wegnehmen), wenn das AS-Lexem in der ZS nicht mit der gleichen Anzahl an Elementen übersetzt wird. So gibt z.B. folgende Regel an, dass *übernachten* ins Englische nicht mit einem simplen Verb, sondern mit der Phrase *spend the night* übersetzt werden muss:

```
@verb(uebernachten,spend,L,I) →
@ins_rhs_arg_np(arg3,def,night,sg,L,I).
```

Diese Regel zeigt an, dass eine weitere Regel (`@ins_rhs_arg_np`) aufgerufen werden muss, die ein zusätzliches Argument zu den bestehenden Argumenten (der Übernachtende und der Ort der Übernachtung) hinzufügt.

Das Verbmobil-System ist in der Lage, solche wie auch andere syntaktische Divergenzen (wie z.B. **argument switch**, **head switch**) zu erfassen und im Transfer umzusetzen. Damit setzt Verbmobil etwas um, was von Bonnie Dorr bereits 1994 beschrieben und auch schon im EUROTRA-System erfolgreich umgesetzt wurde.

2.5.6 Valenz-MÜ im Rahmen der PCEDT

Abschnitt 2.4.2 stellt die PCEDT und Experimente zur Extraktion von Valenzwörterbüchern für die MÜ aus der PCEDT vor. Bojar und Hajič (2005) bezeichnen die MÜ anhand von Valenzwörterbüchern als strukturelle MÜ und grenzen sie von der statistischen MÜ ab. Tatsächlich wurde der Ansatz von Bojar und Hajič in der Folge derart weiterentwickelt, dass strukturelle Informationen mit statistischen Methoden gekoppelt wurden.

Zuletzt mündeten diese Entwicklungen im System TectoMT (Žabokrtský u. a. 2008), das – wie bereits in den vorhergehenden Experimenten – die tektogrammatische Ebene als Transferebene verwendet. Offensichtlich ist die Rolle von Valenz bei TectoMT zunächst nicht; sie spielt im Sinne einer syn-

taktischen Realisierungsinformation im System dennoch eine Rolle, wie im Folgenden beschrieben werden soll (vgl. Zdeněk Žabokrtský, Email vom 28.11.2008).

Das Übersetzungswörterbuch wird auf Basis der Dependenzalignierung der PCEDT aufgebaut. Gegeben seien zwei Englische Wortknoten Nsg und Nsd (‚s' für *source*, ‚g' für *governor*, ‚d' für *dependent*) sowie zwei tschechische Wortknoten Ntg und Ntd (‚t' für *target*). Die Regenten und Dependenten seien jeweils miteinander aligniert und in der gleichen Richtung voneinander abhängig. Von diesem Knoten-/Kantenpaar werden jeweils das Lemma L sowie morpho-syntaktische Attribute, sogenannte **Formeme**, kurz F, extrahiert und als zusammenhängende Lemma-Formem-Tupel abgespeichert. Von solchen Tupeln werden probabilistische Transferregeln gebaut, gemäß der Formel:

$$Prob(Ftd|Fsd,Lsg)$$

D.h. die Formeme des ZS-Dependenten sind so zu wählen, wie es in Abhängigkeit der AS-Dependenten-Formeme und des AS-Regenten-Lexems als am wahrscheinlichsten vorgegeben wird. Dies hilft zum Beispiel in folgendem Fall. Das englische Formem ‚*for*+noun' wird im Tschechischen meist als ‚*pro*+acc', d.h. Präposition *pro* mit Akkusativ, wiedergegeben. Im Falle des englischen *He waited for her* muss es aber im Tschechischen heißen

Čekal	*na*	*ní.*
warten-3.Pers.Sing.-Perfekt	auf	sie-3.Pers.Sing.-fem.
Er wartete auf sie.		

In diesem Fall wird ‚*for*+noun' aufgrund der Abhängigkeit von *Čekal* als ‚*na*+acc' wiedergegeben. Das Transferlexikon stellt für diesen Fall die Information bereit, dass

$$Prob(,na+acc' \mid ,for+noun', ,wait') > Prob(,pro+acc' \mid ,for+noun', ,wait')$$

d.h. dass die Wahrscheinlichkeit für ‚*na*+acc' größer ist, wenn das regierende AS-Lexem *wait* ist, und gibt damit die richtige Realisierung abhängig vom AS-Regenten wieder.

Valenz wird im Kontext von TectoMT also eher als probabilistische Eigenschaft angesehen, anstatt als fixe, u.U. von Hand erstellte Valenzwörterbücher zu verwenden. Tatsächlich unterscheidet sich der Ansatz zunächst nicht von einem fixen Valenzwörterbuch, da ja nur die wahrscheinlichste

Realisierung verwendet wird. Es ist aber denkbar, weitere kontextabhängige Information bei der Ziel-Lemma- und Ziel-Formem-Auswahl zu verwenden, was in einem handkodierten Wörterbuch nur sehr aufwändig zu kodieren ist. Beispielsweise könnten zusätzliche statistische Werte zur Kombinationsmöglichkeit bestimmter Lemmakombinationen von Dependenten und Regenten in der Zielsprache ins Lexikon aufgenommen werden, um sehr unwahrscheinliche und damit wohl unverträgliche Kombinationen zu vermeiden.

2.5.7 Übersicht

Die in den vorangehenden Abschnitten vorgestellten Valenzmechanismen in MÜ-Systemen sind, mit Ausnahme der Valenz-MÜ im Rahmen der PCEDT, vorrangig regelbasiert. SUSY verwendet syntaktische Valenzinformation bei der Auswahl des ZS-Lexems und der Generierung. METAL enthält Valenzinformation in Form von Regeln. Dabei gibt es eine kleine statistische Komponente: Mittels eines Präferenzwerts zwischen 1 und 100 kann der Lexikograf die Auswahl eines Frames wahrscheinlicher machen, indem er den Wert möglichst hoch wählt. EUROTRA verwendet für seinen Transfer eine Interface Structure in Form einer dependenziellen Darstellung. Argumentrollen sind für äquivalente Lexeme über die Sprachen hinweg gleich, ein Unterschied besteht nur in der Realisierung der jeweiligen Rollen. Relevante Realisierungsinformationen werden über ein System semantischer Feature beim Transfer mitgegeben. Dorr entwickelt dieses Modell in ein Interlingua-Modell für die MÜ weiter. Erste LFG-Modelle der MÜ sehen einen Mehrebenentransfer vor; neuere Modelle verwenden zur Erstellung von Transferlexika Daten aus Korpusanalysen, und erhalten damit auch eine statistische Komponente. Das HPSG-Modell aus Verbmobil verwendet zum Transfer hauptsächlich semantische Informationen mittels eines einfachen regelbasierten Prädikat-Argument-Transfers. Dieser Mechanismus kann allerdings auch morpho-syntaktische Information auswerten. Die Prager Valenz-MÜ enthält eine statistische Komponente. Für jedes Lexem wird eine Wahrscheinlichkeit errechnet, in welchem Kontext welche Realisierung der Argumente zu wählen ist.

Mit Bezug auf die Valenz ist allen Ansätzen gemeinsam, dass sie bei der Abbildung von Valenz (in manchen der vorgestellten Ansätze als Prädikat-Argument-Strukturen bezeichnet) prinzipiell von einer „Parallelität“ von Valenz ausgehen (vgl. Abschnitt 4.1). D.h., es wird erwartet, dass in beiden Sprachen ein mehr oder weniger äquivalenter Valenzträger (außer etwa beim headswitch oder bei der Übersetzung feststehender Wendungen) vorhanden ist sowie eine gewisse Anzahl von Begleitern, die sprachabhängig realisiert wer-

den (und gelegentlich wegfallen, wie etwa Personalpronomen in Pro-drop-Sprachen). In allen Ansätzen werden zwar Divergenzen von trivialen Kasusunterschieden in der Argumentrealisierung bis hin zum headswitch behandelt, die Ansätze beinhalten allerdings meist nur Umformungen einer ursprünglichen Prädikat-Argument-Struktur. Inwiefern man von dieser Parallelität auch in Humanübersetzungen, die als Lernmodell wie auch als Evaluationsbasis in der MÜ dienen, ausgehen kann, wird in Kapitel 4 dieser Arbeit untersucht.

Die hier vorgestellten MÜ-Valenzmechanismen haben alle mindestens eine sehr stark regelbasiert geprägte Komponente, gelegentlich mit statistischen Erweiterungen. Betrachtet man die Entwicklung über die Zeit hinweg, haben Valenzmechanismen in der MÜ, ähnlich wie die MÜ im Allgemeinen, den Weg in Richtung hybrider regel- und statistikbasierter Ansätze eingeschlagen. Phrasenbasierte MÜ-Ansätze (Franz u. a. 2003) wurden jüngst so modifiziert, dass sie nicht nur feste Phrasen, sondern auch Phrasen mit „Leerstellen" erlauben, inklusive Informationen darüber, welche Typen von Kategorien diese Leerstellen befüllen können (Hoang u. a. 2009). Gleichzeitig wird im Projekt EuroMatrix, das u.a. die METAL-Weiterentwicklung LucyMT einsetzt, erforscht, inwiefern ein Transfer, der auf Regeln und bilingualen Valenzwörterbüchern basiert, durch Kombination mit einer statistischen Komponente für eine optimierte lexikalische Auswahl verbessert werden kann (vgl. Andreas Eisele, Email vom 05.02.2010).

3 Das deutsch-englische CroCo-Korpus

Das CroCo-Korpus (Neumann und Hansen-Schirra 2005) entstand im Rahmen eines von der Deutschen Forschungsgemeinschaft geförderten Projekts während zweier Phasen von je zwei Jahren. In der ersten Phase war allein die Universität des Saarlandes in Saarbrücken beteiligt. In der zweiten Phase verteilte sich das Team auf die Universität des Saarlandes sowie die Universität Mainz mit ihrem Standort in Germersheim.

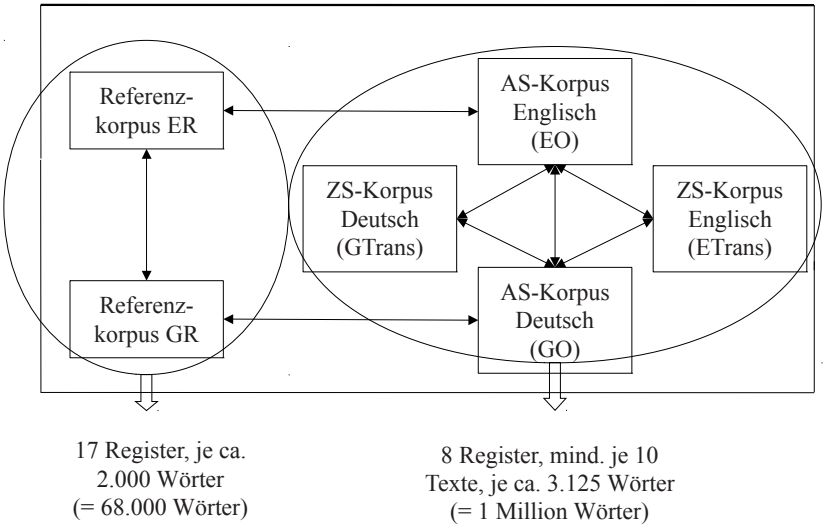

Abbildung 3.1: Der Aufbau des CroCo-Korpus

Das CroCo-Korpus ist ein registerkontrolliertes Textkorpus[29], das zur kontrastiven Untersuchung des Sprachenpaars Deutsch-Englisch sowie zur Untersuchung von Übersetzungseigenschaften dient. Aus dieser Kurzcharakterisierung und der darin wiedergegebenen Zielsetzung ergeben sich bereits Vorgaben für den Aufbau des Korpus, der in Abschnitt 3.1 erläutert wird. Um das Korpus mit den modernen Mitteln der Computerlinguistik abzufragen sowie mit verschiedensten Werkzeugen zu annotieren, wurde vom Autor der vorliegenden Arbeit eine Programmierschnittstelle, als eine sogenannte API (*appli-*

29 Das Verfahren der Registeranalyse, die zu der im Korpus vorgenommenen Zuweisung von Texten zu Registern führt, folgt dem Modell von (Halliday und Hasan 1989).

cation programming interface), erstellt, die den Zugriff auf das Korpus mithilfe von Java-Programmen erlaubt. Diese API wird in Abschnitt 3.2 vorgestellt. Die Charakteristika des Korpus bedeuten für die in dieser Arbeit beschriebenen Experimente einige Einschränkungen, kommen dem Vorhaben aber auch in einigen Punkten entgegen. Darauf wird in Abschnitt 3.3 näher eingegangen.

3.1 Aufbau

Einleitend wurde auf die verschiedenen Aspekte des CroCo-Korpus hingewiesen: Registerkontrolle, Texte als grundlegende Organisationseinheit, kontrastive Sprachuntersuchung sowie Untersuchung von Übersetzungseigenschaften. Diese Vorgaben spiegeln sich im Aufbau des Korpus wieder.

Das Gesamtkorpus ist in zwei Subkorpora (vgl. Abbildung 3.1, markiert durch Kreise) aufgeteilt: das Referenzkorpus und das Übersetzungskorpus. Die beiden untersuchten Sprachen Deutsch und Englisch unterteilen diese Subkorpora in weitere Subkorpora: Das Referenzkorpus besteht aus deutschen und englischen Originaltexten (Subkorpora GR bzw. ER). Das Übersetzungskorpus enthält deutsche und englische Originale (GO bzw. EO) sowie die dazugehörigen Übersetzungen ins Englische bzw. Deutsche (ETrans bzw. GTrans). Die Texte der Korpora sind verschiedenen *Registern* (ungefähr vergleichbar mit dem Begriff „Textsorte“) zugeordnet. Das Referenzkorpus enthält Textausschnitte aus insgesamt 17 verschiedenen Registern. Es stellt somit einen größeren Querschnitt durch die Sprache dar und dient daher in seiner Gesamtheit als registerneutrale Vergleichsinstanz. Die Textausschnitte des Übersetzungskorpus sind in 8 Register aufgeteilt: politische Aufsätze (ESSAY), fiktionale Texte (FICTION), Bedienungsanleitungen (INSTR), populärwissenschaftliche Texte (POPSCI), Aktionärsbriefe (SHARE), politische Reden (SPEECH), Tourismusbroschüren (TOU), und Internetseiten (WEB).

Die Auswahl der Textumfänge und -anzahlen zielt darauf ab, einen möglichst ausgeglichenen Querschnitt der Sprachvariation bzw. der zu betrachtenden Phänomene zu gewährleisten (vgl. Neumann und Hansen-Schirra 2005). Das Referenzkorpus umfasst pro Register je zwölf oder mehr Textausschnitte mit jeweils ca. 2000 Wörtern pro Register. Beim Übersetzungskorpus enthält jedes Register pro Sprach- und Übersetzungsrichtung jeweils mindesten 31.250 Wörter; diese sind auf 10 oder mehr etwa gleich große Texte verteilt. Größe und Anzahl der Texte wurden so gewählt, dass der jeweililge Registerausschnitt, der Auffassung von Biber (1995) folgend, als repräsentativ angesehen werden kann.

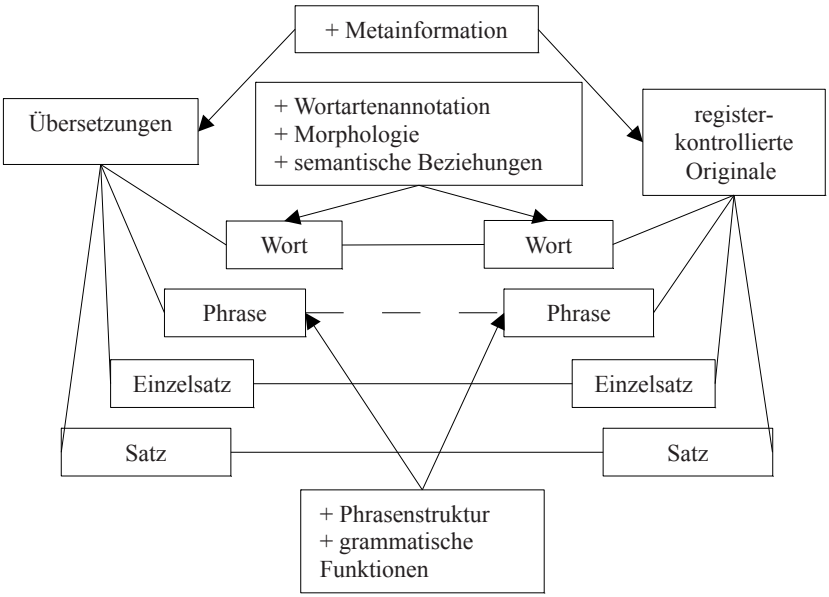

Abbildung 3.2: Annotations- und Alignierungsebenen des CroCo-Korpus

Das CroCo-Korpus ist auf mehreren Ebenen annotiert und aligniert (Vela und Hansen-Schirra 2005; Hansen-Schirra u. a. 2006), wie in Abbildung 3.2 dargestellt. Die Annotationen umfassen die Wort- und Phraseneben, zudem sind die Texte in Einzelsätze und Sätze segmentiert. Als Einzelsätze werden nach den Definitionen von Quirk u.a. (1985) bzw. Helbig und Buscha (2001) sowohl Nebensätze als auch beispielsweise Infinitivkonstruktionen, also alle Konstruktionen mit verbalem Kern, aufgefasst. Ausgenommen sind davon Konstruktionen mit Modalverben wie *müssen*, *wollen* oder *sollen*, die Quirk als „central modals" klassifiziert, sowie deren deutsche Gegenparts. D.h. ein Satz *Ich muss essen* besteht nach dieser Definition aus nur einem Einzelsatz.

Wörter sind in CroCo mit der Wortart mittels des TNT Tagger (Brants 2000b) und mit morphologischer Information mittels MPro (Maas u. a. 2009) jeweils automatisch annotiert. Phrasen sind mit der Phrasenkategorie (wie z.B. NP, PP) und der grammatischen Funktion (wie z.B. Subjekt, direktes Objekt, vgl. Tabelle 3.1) manuell mit MMAX2 (Müller und Strube 2006) annotiert. Ein-

zelsätze wurden manuell segmentiert, Sätze automatisch von der TRADOS Translator's Workbench (Heyn 1996) vorsegmentiert und dies ggf. korrigiert.

ADV_CAUSE	Adverb des Grunds
ADV_LOC	Adverb des Orts
ADV_MOD	Adverb der Art und Weise
ADV_TEMP	Adverb der Zeit
ADV_OTHER	Adverb (sonstige)
APPO	Apposition
COMPL	prädikatives Komplement
CONJ	Konjunktion
DOBJ	direktes Objekt
FIN	finites Verb
IOBJ	indirektes Objekt
NEG	Negation
MINOR	Satz ohne finites Verb
PART	Partikel
PRED	nicht-finite Verbbestandteile
PROBJ	Präpositionalobjekt
SUBJ	Subjekt

Tabelle 3.1: Grammatische Funktionen in CroCo

Aligniert sind Originale und Übersetzungen auf allen Ebenen mit Ausnahme der Phrasenebene. Wörter wurden automatisch mit GIZA++ (Och und Ney 2003) aligniert. Einzelsätze wurden manuell mit MMAX2 aligniert. Phrasen sind nicht aligniert, werden aber bei Bedarf, z.B. zum Zwecke von Abfragen wie in Abschnitt 4.2 beschrieben, mittels der Wortalignierung und der funktionalen Annotation automatisch aufeinander abgebildet.[30] Die Satzalignie-

30 Eine manuelle Phrasenalignierung wäre wünschenswert, aber der Aufwand wäre unkalkulierbar hoch. Mithilfe der Wortalignierung und dem Abgleich der Phrasenfunktionen (wie

rung wurde wie die Segmentierung halbautomatisch mit TRADOS durchgeführt.

Die manuellen Annotationen wurde nur von jeweils einem Annotierer durchgeführt. Die einzige Qualitätskontrolle bestand aus einer Überprüfung auf offensichtliche oder sehr wahrscheinliche Fehler (z.B. Überprüfung aller Sätze mit Subjekt, aber ohne Verb bzw. umgekehrt) sowie aus Stichproben. Für eine Doppelannotation waren nicht die notwendigen Ressourcen vorhanden.

Um eine derart komplexe Datenmenge repräsentieren sowie Austauschbarkeit verschiedener Annotationsebenen gewährleisten zu können, und um überlappende Annotationen zu erlauben, sind die CroCo-Daten in einem am XCES-Standard orientierten XML-Format gespeichert (Hansen-Schirra u. a. 2006). Es handelt sich dabei um ein sog. *stand-off mark-up*, das relationell funktioniert. Eine Datenbank würde eine derartige relationelle Darstellung ebenfalls erlauben, das Speicherformat von Datenbanken ist allerdings weder für den Menschen leicht lesbar, noch mittels Programmiersprachen leicht verarbeitbar.

Grundlage des CroCo-Formats ist dabei die Liste der Tokens, die in einem Text vorkommt, und von denen jedes Token eine einzigartige Kennung erhält. Die weiteren Strukturen – Phrasen, Einzelsätze und Sätze – sind mit Bezug auf die Token-Kennungen beschrieben, wie Abbildung 3.3 zeigt. In der linken Spalte steht die Liste der Token des Beispieltexts, einem deutschen Originaltext. Der Name der Basisdatei `GO.tok.xml` ist ebenfalls im XML in der linken Spalte notiert (durch Fettdruck hervorgehoben). Zur Illustration des relationellen Prinzips sind die Token mit den Kennungen `t66` und `t67` mit Fettdruck hervorgehoben. In der mittleren Spalte steht die Wortartenannotation, wie sie von TNT in Form des Stuttgart-Tübingen-Tagset STTS ausgegeben wird. Im XML-Kopf ist unter `xml:base` angegeben, auf welche Basisdatei sich die Wortartenannotation bezieht, nämlich `GO.tok.xml`. In den durch Fettdruck hervorgehobenen Zeilen ist zu sehen, dass das Token mit der Kennung `t66` als `PIDAT` (attribuierendes Indefinitpronomen mit Determiner) und das Token mit der Kennung `t67` als `NN` (normales Nomen) annotiert ist. In der rechten Spalte ist eine weitere Ebene der Annotation dargestellt, nämlich die Phrasenannotation. Auch in dieser Beispieldatei mit dem Namen `GO.chunk.xml` werden wieder die Token `t66` und `t67` herangezogen: In dieser XML-Datei werden sie in einer Phrase mit der Kennung `ch13` zusam-

Subjekt, Objekt usw.) zwischen alignierten Sätzen kann man aber eine akzeptable automatische Abbildung erreichen. Eine gute automatische Phrasenalignierung existiert derzeit nach Kenntnisstand des Autors dieser Arbeit nicht.

Tokenindex

```
<header xlink:href=
   "GO.header.xml"/>
<tokens>
<token id="t64"
strg="Ich"/>
<token id="t65"
strg="spielte"/>
<token id="t66"
strg="viele"/>
<token id="t67"
strg="Möglichkeiten"
/>
<token id="t68"
strg="durch"/>
<token id="t69"
strg=","/>
</tokens>
```

Wortarten-Annotation

```
<tokens
xml:base="GO.tok.xml
">
<token pos="pper"
  xlink:href="#t64"/>
<token pos="vvfin"
  xlink:href="#t65"/>
<token pos="pidat"
  xlink:href="#t66"/>
<token pos="nn"
  xlink:href="#t67"/>
<token pos="ptkvz"
  xlink:href="#68"/>
<token pos="yc"
  xlink:href="#t69"/>
</tokens>
```

Phrasenindex

```
<chunks xml:base=
   "GO.tok.xml">
 <chunk id="ch13">
  <tok
   xlink:href="#t66"/>
  <tok
   xlink:href="#t67"/>
 </chunk>
 <chunk id="ch14">
  <tok
   xlink:href="#t70"/>
 </chunk>
 <chunk id="ch15">
  <tok
   xlink:href="#t71"/>
 </chunk>
</chunks>
```

Abbildung 3.3: Das CroCoXML stand-off Format

mengefasst. Die Annotation von Einzelsätzen und Sätzen ist analog in XML dargestellt. Die XML-Darstellung von Phrasentypen und -funktionen geschieht ebenfalls relationell, allerdings mit Bezug auf die Phrasenkennung. In den entsprechenden Dateien `GO.ps.xml` (`ps` für *phrase structure*) und `GO.gf.xml` (`gf` für *grammatical function*) ist ist für jede Phrasenkennung angegeben, welche phrasenstrukturelle Kategorie bzw. welche Funktion die Phrase hat. Die Alignierung der verschiedenen Ebenen ist ebenfalls relationell, also mit Bezug auf die Token- bzw. Einzelsatz- oder Satzkennungen dargestellt.

3.2 Die CroCoAPI

Bei der automatischen Annotation des CroCo-Korpus wurde eine ganze Reihe verschiedener Annotationsprogramme eingesetzt. Ein großes Problem bei der Verwendung unterschiedlicher, nicht aufeinander abgestimmter Komponenten ist generell, die vielen verschiedenen Ausgabedatenformate der Anno-

tationswerkzeuge am Ende aufeinander abzubilden und auf ein einheitliches Korpusformat zu übertragen. So zeigt sich zum Beispiel bei einem Vergleich der Tokenisierung des Part-of-Speech Tagger TNT (Brants 2000b) und des Morphologieanalyseprogramms MPro (Maas u. a. 2009), dass die beiden Programme die Eingabe an verschiedenen Stellen in unterschiedlich viele und verschieden feine Elemente aufteilen. Das Problem unterschiedlicher Formate ergibt sich bei der Arbeit mit parallelen (oder generell multilingualen) Daten auch oft daher, dass Analyseprogramme (zur Zeit noch) häufig nur monolingual ausgelegt sind und sich damit fast zwangsläufig Formatunterschiede ergeben. Im Fall der in CroCo verwendeten Analyseprogramme wurden allerdings solche ausgewählt, die auf das Deutsche und auf das Englische anwendbar sind.

MPro	**TNT**
It	*It*
's	*'*
	s

Tabelle 3.2: Die verschiedenen Tokenisierungen von It's *von MPro und TNT.*

Tabelle 3.2 illustriert ein weiteres typisches Problem bei der Verwendung nicht aufeinander abgestimmter Analysewerkzeuge, nämlich das Problem unterschiedlicher Segmentierungen von sprachlichen Einheiten, am Beispiel der englischen verkürzten Form *It's*. Wo TNT drei Token erkennt, nämlich *It*, *'* und *is*, teilt MPro in die zwei Token *It* und *'s* auf. Da das CroCo-Datenformat wie zuvor beschrieben referenziell ist, muss hier eine der beiden Darstellungsformen als Bezugsgröße bestimmt werden, und die andere darauf abgebildet werden. Im Falle von CroCo wurde die Tokenisierung von TNT vorgezogen; alle Annotationen bauen auf den Tokenlisten von TNT auf, die MPro-Analyse wird als verfeinerte morphologische Analyse der Token gespeichert und genutzt.

Um eine einheitliche Verarbeitung der CroCo-Daten zu gewährleisten, sowie um Abfragen zu vereinfachen und spätere Annotationsvorgänge am Korpus zu ermöglichen, wurde eine Programmierschnittstelle, eine sogenannte API, auf Basis der Programmiersprache Java geschaffen. Die Überlegung, eine bestehende API für die Verarbeitung von CroCo zu verwenden, wurde recht schnell verworfen, da in APIs wie etwa der TiGerAPI[31] die Darstellung

31 http://www.tigerapi.org/

der linguistischen Daten eng mit der tatsächlichen Darstellung im Format verwoben sind. Dies hätte eine u.U. nicht-triviale Anpassung der API an das CroCo-Datenformat bedeutet. Für CroCo wurde daher eine API geschaffen, die die linguistische Verarbeitung von der Formatdarstellung trennt und damit potenziell auch die Einbindung weiterer Datenformate neben dem CroCo-Format erlaubt. Diese sog. CroCoAPI kümmert sich einerseits um das CroCo-Format (vgl. Abschnitt 3.1), und abstrahiert die Korpusdaten auf einer linguistischen Ebene, dem sog. CoReTool. Damit sind Abfragen über linguistische und translatorische Phänomene, wie in Kapitel 4 dargestellt, mit dem CroCo-Format, aber potenziell auch mit anderen Formaten, die das gleiche Abstraktionswerkzeug verwenden, möglich.

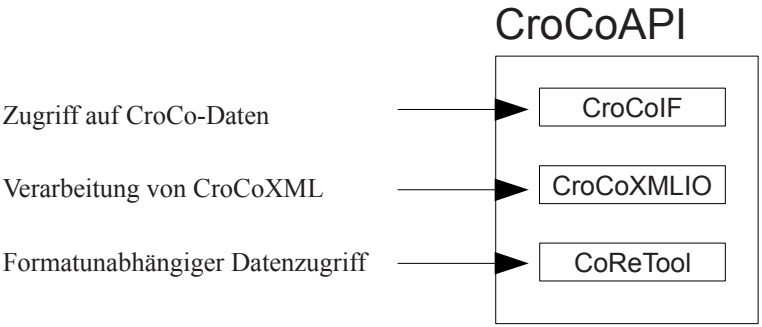

Abbildung 3.4: Die drei Ebenen der CroCoAPI

Die CroCoAPI besteht aus drei Ebenen (vgl. Abbildung 3.4), die an mehreren Stellen miteinander interagieren.

Die relationelle Darstellung des CroCoXML-Formats wird, um formatunabhängige und damit einfacher wiederverwendbare Abfragefunktionen zu ermöglichen, in die hierarchische, formatunabhängige CoReTool-Datenstruktur übersetzt. Bevor allerdings diese Übersetzung geschehen kann, müssen zunächst die CroCo-Daten eingelesen werden. Die Zugriffsmethoden zum Einlesen und Ausschreiben der CroCo-Daten sind in der Schnittstelle CroCoIF lokalisiert. Um die Verarbeitung des Datenformats kümmert sich das CroCoXML-Paket. Das CroCoIF muss dann einige zusätzliche Rechenschritte durchführen, um in das CoReTool-Format zu übersetzen; die Be-

schreibung dieser Schritte würde allerdings den Rahmen dieses Kapitels sprengen.

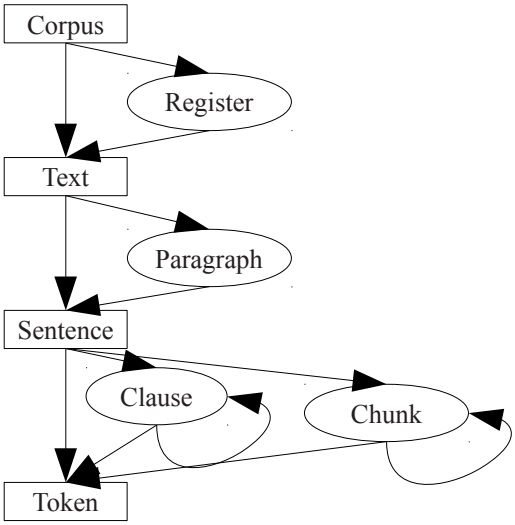

Abbildung 3.5: Die hierarchischen Ebenen des CoReTool

Die linguistischen Strukturen werden also intern als CoReTool-Strukturen abgespeichert (vgl. Abbildung 3.5). CoReTool stellt die linguistischen Daten in linguistischen Strata dar; d.h. ein Text enthält Sätze, dessen Sätze enthalten dann wiederum Token. Diese Struktur ist quasi die Mindestanforderung an eine Annotation für eine Darstellung mittels CoReTool. Es gibt optionale Zwischenebenen (oval dargestellt), wie beispielsweise Chunks, die eine detailliertere Annotation, wie sie in CroCo vorliegt, darstellen zu können. Die Alignierung wird von CoReTool durch Paare von Datenstrukturen repräsentiert. Zwei alignierte Sätze sind also als Paar $<S_{Ori}, S_{Trans}>$abrufbar. Bei Mehrfachalignierungen wird für jede einzelne Alignierung ein solches Paar im Speicher abgelegt. Für eine Mehrfachalignierung zwischen S_{Ori} und $S_{Trans-1}$ und $S_{Trans-2}$ werden also zwei Paare $<S_{Ori}, S_{Trans-1}>$und $<S_{Ori}, S_{Trans-2}>$generiert.

3.3 Vor- und Nachteile der Nutzung von CroCo für die vorliegende Studie

An dieser Stelle soll auf einzelne Aspekte des CroCo-Aufbaus und der Erstellung des Korpus eingegangen werden und beurteilt werden, inwiefern sich diese Aspekte positiv oder negativ auf die Studie auswirken können. Dieser Abschnitt soll nur die globalen Korpuseigenschaften betreffen; Details, die bei bestimmten Arbeitsschritten relevant sind, werden an entsprechender Stelle diskutiert.

Eine detaillierte Diskussion der Vor- und Nachteile der bestehenden CroCo-Annotation findet sich in (Hansen-Schirra u. a. [erscheint]: Kap. 9).

Ein großer Vorteil des CroCo-Korpus ist die Tatsache, dass es auf vielen linguistischen Ebenen aligniert ist. Dies spielt v.a. in den Vorstudien zum eigentlichen Extraktionsexperiment eine Rolle. Diese Vorstudien untersuchen, wie syntaktisch ähnlich sich Original und Übersetzung sind und stützen sich dabei auf die Alignierung unterschiedlicher Ebenen. Um zweisprachige Valenzwörterbücher zu extrahieren, wird man sich zunächst wohl auch auf parallele Korpora stützen; inwiefern dies tatsächlich sinnvoll ist, werden wir im Verlauf der Arbeit untersuchen und diskutieren.

Ein weiterer Vorteil des CroCo-Korpus ist der hohe Anteil an manueller Annotation. Zielt man zwar drauf ab, irgendwann den Prozess der Valenzwörterbuchextraktion zu automatisieren, so ist doch zumindest für erste detailliertere Studien eine hochqualitative manuelle Annotation – z.B. zur entsprechenden Kalibrierung und Verbesserung der automatischen Annotationswerkzeuge – höchst informativ.

Ebenfalls von Vorteil ist die Tatsache, dass die Texte in CroCo nach Registern aufgeteilt sind. Bei einem Korpus entsprechender Größe könnte man dies nutzen, um registerspezifische Valenzwörterbücher zu extrahieren, z.B. für Wirtschaftskommunikation oder Gebrauchsanweisungen. Bei ersterem könnte man z.B. häufigere metaphorische Verwendungen erwarten als bei letzterem; letzteres könnte dagegen oftmals eine verknappte syntaktische Realisierung aufweisen[32].

Nicht unproblematisch ist die Annotationsqualität von CroCo. Die manuelle Annotation wurde jeweils nur von einem Annotierer durchgeführt. Eine Doppelannotation, die nachgewiesenermaßen die Fehlerquote bei der Annotation senkt (vgl. z.B. Brants 2000a), gibt es nur für wenige Texte. Außerdem sind

32 In typischen – im Deutschen häufig infinitivisch realisierten – Anweisungen wie „Anschlüsse regelmäßig überprüfen!“.

einige Ebenen von der manuellen Annotation ausgenommen; ungünstigerweise (aus der Sicht der parallelen Valenzstudie) gilt dies z.B. für die Phrasenalignierung, die komplett fehlt. Die Wortalignierung ist rein automatisch und kann in ihrer Qualität – 72% Spezifität bei nur 68% Sensitivität – nicht überzeugen (vgl. Čulo u. a. 2008b).

Ein weiterer Nachteil von CroCo ist, dass grammatische Funktionen nur auf oberster Ebene annotiert sind. Erst in der zweiten Projektphase wurden auch in Nebensätzen grammatische Funktionen annotiert, allerdings war diese Annotation zum Zeitpunkt, als die in dieser Arbeit beschriebenen Studien durchgeführt wurden, nicht verfügbar. Andere eingebettete Strukturen wie etwa Partizipial- oder Adjektivphrasen sind von dieser Zweitannotation allerdings ausgenommen, was nicht im Sinne einer Valenzanalyse ist.

4 Studien zur parallelen Valenzextraktion

Dieses Kapitel beschreibt die Experimente zur parallelen Valenzextraktion aus dem CroCo-Korpus. Es ist in drei aufeinander aufbauende Abschnitte gegliedert. In Abschnitt 4.2 werden die Versuchsaufbauten beschrieben. In Abschnitt 4.3 werden die Ergebnisse vorgestellt. In Abschnitt 4.4 werden Möglichkeiten zur praktischen Anwendung aufgeführt.

4.1 Hypothesen zur Parallelitätsannahme

Wie zuvor bei der Darstellung der MÜ-Systeme festgestellt, wird in MÜ-Systemen prinzipiell davon ausgegangen, dass Valenz „im Prinzip parallel" ist. In EUROTRA wurde diese Annahme sogar explizit gemacht: Die Rollenmuster eines Verblexems waren über Sprachen hinweg gleich, nur die syntaktische Realisierung unterschiedlich. Der Übertrag einer Prädikat-Argument-Struktur in eine andere Sprache kann in den verschiedenen MÜ-Systemen zwar diversen Umformungen bis hin zum *headswitch* oder Wegfall einzelner Argumente unterliegen, aber prinzipiell wird von einer Äquivalenz der Prädikate ausgegangen.

Diese Art von Annahme wird aber nicht nur in der MÜ gemacht. Padó (2007b) beschreibt den Versuch, Frameannotationen von einer Sprache auf die andere zu projizieren. Diese Annotationsprojektion beschränkt sich auf die semantischen Kategorien. Padó weist zwar auf mögliche syntaktische Divergenzen zwischen Sprachen hin, lässt diese aber unbehandelt. Padó argumentiert, dass die Frameannotation auf einer ausreichend hohen Abstraktionsebene stattfindet, genauer gesagt, dass wegen der „coarse granularity of description" Frames i.d.R. unabhängiger zumindest von syntaktischen Divergenzen der Valenzrahmen zweier Frame-tragender Lexeme verschiedener Sprachen sind.

Die gleiche Parallelitätsannahme wird in der Grammatikinduktion basierend auf Paralleltexten gemacht (vgl. z.B. Kuhn 2004; Kuhn 2005). In der Grammatikinduktion wird Wort- und Satzalignierung genutzt, um Annotationen einer Sprache, typischerweise des Englischen, für das viele teils hochwertige Annotationswerkzeuge existieren, in eine andere, ressourcenarme Sprache zu übertragen. Dabei werden zwar Probleme wie unterschiedliche Wortstellungen in den Sprachen berücksichtigt und als eine Hauptursache für mögliche Fehler angesehen, aber auch hier wird davon ausgegangen, dass Strukturen in einer Sprache prinzipiell eine „parallele" Entsprechung in der anderen Sprache haben.

Was, wie Padós Ergebnisse belegen, auf der semantischen Ebene mit Frames gut funktioniert, da diese von vielen syntaktischen Eigenheiten wegabstrahieren, muss für die Parallelität von Syntax nicht gelten. Zwei der Fragestellungen im Kontext der hier vorgestellten Vorstudien zur Valenzwörterbuchextraktion waren:

- inwiefern auf der syntaktischen Ebene Originale und Übersetzungen korrelieren
- und welche Wanderungsbewegungen von Funktionen und Kategorien im Sprachenpaar Deutsch-Englisch von einer Sprache zur anderen zu erwarten sind.

Teil der Experimente zur parallelen Valenzextraktion sind daher eine Reihe von Vorstudien, die auf der Alignierung der verschiedenen Ebenen des Cro-Co-Korpus aufbauen. In diesen Vorstudien wird untersucht, inwiefern überhaupt von struktureller Parallelität gesprochen werden kann. Dabei werden u.a. folgende Fragen behandelt:

- Als Extraktionsbasis für Verbvalenzrahmen dienen Sätze. Je höher der Anteil an alignierten Sätzen, desto höher ist die Zahl an Valenzrahmenpaaren, und desto verlässlicher ist die Extraktion, da eine größere Datenbasis eine unverhältnismäßig hohe Gewichtung seltener Valenzmuster verhindert. Daher also die Fragestellung: In wie vielen Fällen findet ein AS-Satz eine Entsprechung in der ZS, in wie vielen Fällen ein ZS-Satz in der AS?
- Wenn in dem Valenzwörterbuch Divergenzen bei der Realisierung von grammatischen Funktionen beschrieben werden sollen: Muss man sich auf rein verbabhängige Funktionswechsel einstellen, oder haben Sprachtypologie und Register einen allgemeinen Effekt auf diese?
- Wenn aus parallelen Sätzen Valenzwörterbucheinträge extrahiert werden sollen, die auf einem bestimmten Valenzträger aufbauen: Welche möglichen syntaktischen (und u.U. semantischen) Divergenzen zwischen Valenzträgern sind zu erwarten?

Diese Vorstudien nutzen die Mehrebenen-Annotation und -Alignierung. Dadurch sind Studien auf empirischer Basis und mit statistischen Ergebnissen möglich. Die Vorstudien untersuchen also valenzrelevante Fragen auf einer größeren Skala. Sie beziehen damit nicht nur einzelne Prädikat-Argument-Paare mit ein, sondern können auch klären, ob valenzrelevante Phänomene, z.B. Wechsel von grammatischen Funktionen, innerhalb eines Registers häufiger oder seltener vorkommen, ob also Einflüsse jenseits des reinen Verblexems vorliegen. Jede der vorgestellten Vorstudien wird in einen Zusammenhang mit dem eigentlichen Extraktionsexperiment gebracht werden. Neben

den statistischen, eher grobkörnigen Auswertungen zu Divergenzen wird es eine Betrachtung von Einzelfällen geben, die sich heuristisch als typische Gründe für Divergenzen zeigen.

Das Experiment zur Extraktion von Valenzwörterbucheinträgen nutzt ebenfalls die Alignierung bzw. die Abbildung von grammatischen Funktionen in CroCo. Da keine Alignierung von grammatischen Funktionen in CroCo besteht, kann der Frage, inwiefern grammatische Funktionen beim Transfer von AS nach ZS nur sehr vorsichtig in den Vorstudien, mittels Auswertung der nicht optimalen Wortalignierung, angegangen werden. Dennoch wird auch an dieser Stelle der Einblick in typische Einzelfälle informativ über Wanderungsbewegungen von grammatischen Funktionen, also von funktionalen Argumentstrukturen, sein. Im Fokus der Untersuchung liegen aber nicht Divergenzen zwischen Argumenten, sondern syntaktische (und auch semantische) Divergenzen zwischen den Prädikaten in AS und ZS. Anhand einer Testmenge von 300 Satzpaaren wird ein Überblick über Tendenzen in der Zahl und Art der Divergenzen gegeben werden.

Nach Beschreibung der Vorstudien und des eigentlichen Extraktionsexperiments wird sich der Blick auf mögliche Anwendungen der im Extraktionsexperiment erzielten Ergebnisse richten. Dazu wird eine Pilotstudie mit der Erweiterung einer bestehenden MÜ-Grammatik um einzelne Beispielregeln, gewonnen aus den Extraktionsexperimenten, beschrieben, ebenso wie einige grundlegende Überlegung dazu, wie Ergebnisse einer parallelen Valenzwörterbuchextraktion dem menschlichen Nutzer – für die tägliche Sprachlerner- und Übersetzerpraxis, nicht nur für den Linguisten – präsentiert werden können.

Die hier vorgestellten, vom Autor dieser Arbeit erarbeiteten Ergebnisse der Vorstudien dienen auch für eine Auswertung im translationswissenschaftlichen Kontext in (Hansen-Schirra u. a. [erscheint]: Kap. 6). Im Rahmen dieser Arbeit sollen die Ergebnisse zunächst aus einer valenzorientierten und später aus einer MÜ-relevanten Sicht interpretiert werden. Letzteres wird in Kapitel 5 vorgenommen.

4.2 Versuchsaufbauten

In diesem Abschnitt werden die linguistischen und technischen Grundüberlegungen und Voraussetzungen für die Versuchsaufbauten zu den Vorstudien (Abschnitt 4.2.1) und zum eigentlichen Extraktionsexperiment (Abschnitt 4.2.2) beschrieben. Die Vorstudien dienen dabei der Überprüfung der Parallelitätsannahme, d.h. es wird untersucht, inwiefern für die Valenzextraktion relevante Strukturen – Sätze, Einzelsätze und grammatische Funktionen – Ent-

sprechungen in der jeweils anderen Sprache finden, und wie sie sich im Falle von grammatischen Funktionen ggf. ändern. Das eigentliche Extraktionsexperiment beschäftigt sich dann mit der Extraktion von Valenzträgern und deren Begleitern aus den parallelen Texten sowie mit Divergenzen, die dabei zu beobachten sind. Die Auswertung in 4.3 orientiert sich ebenfalls an dieser Aufteilung.

Allen Untersuchungen in dieser Arbeit liegen dabei die drei Register SHARE, FICTION und SPEECH zugrunde, jeweils in beiden Übersetzungsrichtungen Englisch-Deutsch (E2G, mit den Originalen EO und den Übersetzungen GTrans) und Deutsch-Englisch (G2E, mit den Originalen GO und den Übersetzungen ETrans). Dies hat den Hintergrund, dass zum Zeitpunkt der Versuche nur diese drei Register vollständig annotiert waren. Dies ist allerdings nicht nachteilig für die Untersuchungen: Die drei Register variieren, wie im Verlauf der Untersuchungen zu sehen sein wird, in ihren Eigenschaften deutlich genug voneinander, um registerbedingte Ursachen für Divergenzen auf der einen Seite und sprachkontrastive oder translationsbedingte Divergenzen auf der anderen Seite voneinander trennen zu können.

4.2.1 Auswertung von Alignierungslücken und -divergenzen

Die in diesem Abschnitt vorgestellten Abfragen werten alle Ebenen der CroCo-Annotation und -Alignierung aus. Da Phrasen in CroCo nicht aligniert sind, muss für phrasenbasierte Abfragen auf andere Abfragemethoden zurückgegriffen werden als bei Wörtern, Einzelsätzen oder Sätzen. In den entsprechenden Abschnitten sind diese Methoden erklärt.

Die Abfragen stützen sich auf zwei Phänomene, so genannte **empty links** und **crossing lines**, die sich in ihrer Konzeption auf Phänomene stützen, die in dieser oder ähnlicher Form in der Übersetzungsliteratur beschrieben wurden (Vinay und Darbelnet 1958; Catford 1965; Newmark 1988; Koller 2001), wie etwa **translation shifts** bei Catford oder die **Eins-zu-Null-Entsprechung** oder **Lücke** bei Koller.

Als empty links werden Elemente bezeichnet, die kein Äquivalent in der AS bzw. ZS haben (also Lücken). Crossing lines sind Elemente, die miteinander aligniert sind, deren übergeordnete syntaktische Elemente allerdings nicht aligniert sind (die also, im Sinne der translation shifts, zwar inhaltlich einander zuzuordnen und formal äquivalent, aber nicht genau derselben Form sind). Auf diese beiden Phänomene soll nun im Folgenden näher eingegangen werden.

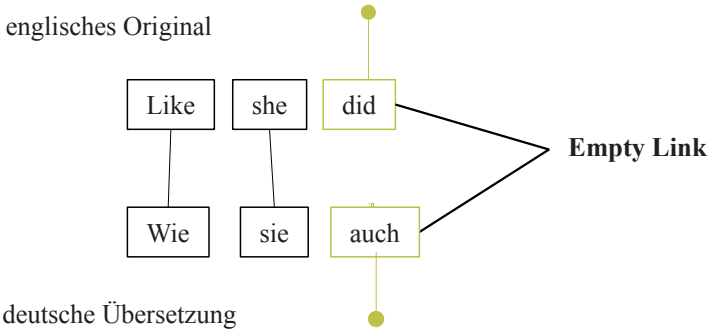

Abbildung 4.1: Empty links bei der Wortalignierung

In einem parallelen Korpus sind idealerweise nicht nur Originaltexte und Übersetzungen aligniert, sondern auch linguistische Einheiten wie diejenigen Sätze, Teilsätze, Phrasen und Wörter, die einander entsprechen. Das CroCo-Korpus ist auf Wortebene automatisch, auf Einzelsatz- und Satzebene manuell aligniert. Allerdings hat nicht jede linguistische Einheit auch eine Entsprechung. In Abbildung 4.1 ist ein alignierter Satz und die Wortalignierung zwischen den Sätzen abgebildet. Sowohl im deutschen Original als auch in der englischen Übersetzung gibt es jeweils ein Element, das keine Entsprechung in der jeweils anderen Sprache hat, also nicht aligniert ist. Hierbei wird von einem empty link gesprochen. Solche empty links können auf Übersetzungsphänomene oder kontrastive Unterschiede deuten. In unserem Beispiel würde man von einem kontrastiven Unterschied sprechen. Im Englischen geschieht der Rückbezug auf eine Handlung mittels des Hilfsverbs *do*, in diesem Fall in der Vergangenheitsform *did*. Im Deutschen genügt dafür das Adverb *auch*. Bei diesem Sprachkontrast handelt es sich nach Halliday und Hassan (1976) um zwei verschiedene Arten von Kohäsionsmitteln. Im Englischen wird das Verb aus der vorangehenden Aussage substituiert, im Deutschen wird durch eine Ellipse Kohäsion mit dem vorangehenden Kontext erzeugt.

Von einer crossing line spricht man dann, wenn zwei Elemente, die aligniert sind, in Elemente eingebettet sind, die nicht miteinander aligniert sind. So

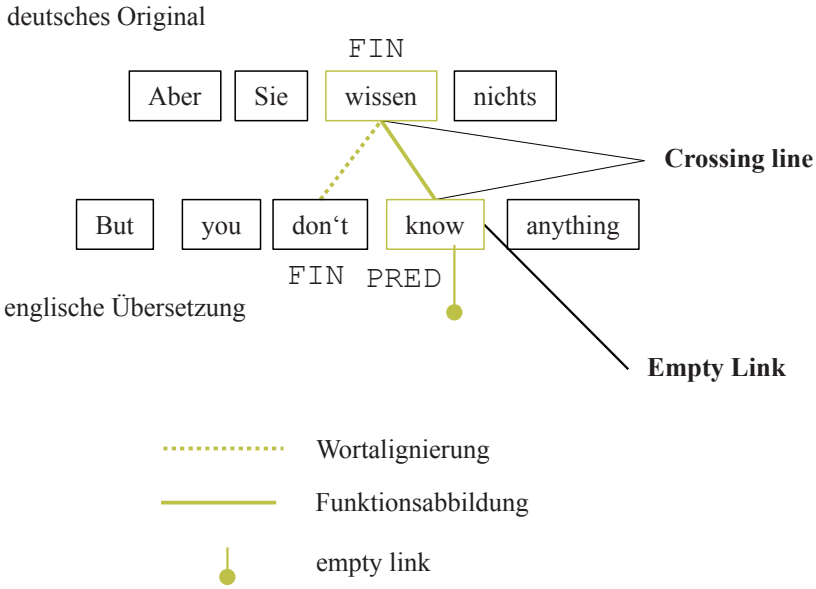

Abbildung 4.2: Crossing lines zwischen Funktionen: von FIN nach PRED

beispielsweise in Abbildung 4.2: Die beiden Worte *wissen* und *know* sind miteinander aligniert, allerdings sind sie in unterschiedlichen Funktionen eingebettet. Das finite Verb *wissen* ist als FIN markiert, als finites Verb, und ist zugleich das semantische Hauptverb. Das semantische Hauptverb *know* im englischen Satz ist nicht gleich dem finiten Verb (*don`t*), und wird daher als PRED markiert. Damit ergibt sich eine crossing line[33], die einen typischen Unterschied zwischen dem Deutschen und dem Englischen aufzeigt: die Verneinung mittels eines Hilfsverbs. Crossing lines können nicht nur auf kontrastive Unterschiede, sondern auch, wie im Folgenden gezeigt wird, auf Übersetzungsphänomene deuten.

33 Eine **crossing line** bezieht sich nicht, wie der Name suggerieren könnte, auf (wenn visuell dargestellt) sich überkreuzende Kanten in der Alignierung, sondern beschreibt eine Art „Grenzüberschreitung“ – z.B. wenn die Funktion PRED „verlassen“ und dafür der Rahmen einer neue Funktion FIN „betreten“ wird.

```
for every sentencePair in sentencePairs
     slSentence := getSlSentence(sentencePair)
     tlSentence := getTlSentence(sentencePair)

     for every clause in getClauses(slSentence)
          alignedClause := getAlignedClause(clause)
          if (not(isMember?(alignedClause,
                            tlSentence)))
               markCrossingLine()
          end
     end

     # repeat the same for tlSentence

end
```

Abbildung 4.3: Pseudo-Code für die Abfrage nach crossing lines *zwischen Einzelsätzen und Sätzen.*

Ein beispielhafter Pseudocode für eine Abfrage auf crossing lines zwischen Einzelsätzen und Sätzen ist in Abbildung 4.3 angegeben. Darin zeigt sich, wie die Mehrebenenannotation für die Abfrage genutzt wird. Die Abfrage findet Satzpaar für Satzpaar mit AS-Satz sl_X und ZS-Satz tl_Y statt. Im ersten Durchlauf der Abfrage werden für den AS-Satz alle darin enthaltenen Einzelsätze abgefragt. Für jeden dieser AS-Einzelsätze wird überprüft, ob der damit alignierte ZS-Einzelsatz Teil des ZS-Satzes tl_Y ist. Ist dies nicht der Fall, wird eine crossing line memoriert (und später ausgegeben). Der gleiche Prozess wird in einem zweiten Durchlauf mit dem ZS-Satz als Ausgangspunkt wiederholt, da sich ja darin Einzelsätze befinden können, die mit Einzelsätzen eines anderen AS-Satzes aligniert sein könnten und von dem ersten Durchlauf der Abfrage daher nicht gefunden würden.

Dieser Verarbeitungsalgorithmus wird auch für Abfragen nach crossing lines zwischen Wörtern und Einzelsätzen angewandt, da auf beiden Ebenen eine Alignierung besteht. Auf einigen der Ebenen besteht bei CroCo keine manuelle oder gar keine Alignierung. Die automatische Wortalignierung hat eine Spezifität von gerade mal 72% bei einer noch schlechteren Sensitivität (Čulo u. a. 2008b). Abfragen auf der Phrasenebene beschränken sich auf die grammatische Funktion der Phrasen. Bei Abfragen, die grammatische Funktionen miteinbeziehen, muss auf ein anderes Verfahren ausgewichen werden. Hier wird zwischen den grammatischen Funktionen innerhalb eines Satzpaa-

res abgebildet, d.h. das Subjekt des AS-Satzes wird ad hoc mit dem Subjekt des ZS-Satzes „aligniert“. Fehlt in einem der beiden Sätze eine Funktion, so wird dies als empty link gekennzeichnet. Gehören zwei Wörter eines alignierten Wortpaares zwei verschiedenen Funktionen an, wird dies als crossing line gezählt.

Da die quantitativen Daten bei der qualitativen Auswertung – die manuelle Auswertung einer Reihe von Beispielen – nachgeprüft wurden, konnten einige allzu starke Verzerrungen aufgrund der suboptimalen bzw. fehlenden Alignierung aufgefunden und bei der Interpretation der Ergebnisse entsprechend berücksichtigt werden. Wie Probleme bei der automatischen Abfrage durch eine verbesserte Annotation umgangen werden können, wird in Kapitel 5 diskutiert werden.

4.2.2 Extraktion und Nutzbarmachung bilingualer Valenzwörterbucheinträge

Bei der Extraktion von bilingualen Valenzwörterbucheinträgen aus Korpora ist man auf konkrete Realisierungen von Valenzrahmen beschränkt. Was also eigentlich extrahiert wird, sind Realisierungsmuster, oder noch genauer Valenzrealisierungsmuster (VRM). Zwei sich gegenüberstehende VRM, die aus zwei alignierten Sätzen extrahiert werden, werden im Folgenden als **alignierte VRM** bezeichnet.

Die Extraktion von alignierten VRM konzentriert sich, wie bereits erwähnt, auf rein syntaktische bzw. grammatische Aspekte. D.h., zu einem extrahierten Satzbauplan gehören das semantische Hauptverb, die Liste der Funktionen, mit der das Hauptverb auftritt, sowie zu jeder Funktion die Beschreibung der syntaktischen Kategorie, mit der die Funktion realisiert wurde. Inwiefern diese alignierten VRM insbesondere mit Bezug auf das Prädikat tatsächlich immer äquivalent sind, soll in Abschnitt 4.3.2 erörtert werden.

Verschiedene Aspekte werden bei den Extraktionsexperimenten der vorliegenden Arbeit ausgeblendet. Alle im Folgenden aufgeführten Aspekte sind als mögliche Erweiterungen denkbar, würden aber dem grundlegenden, hauptsächlich linguistischen Charakter dieser Arbeit nicht entsprechen und ihren Rahmen sprengen. Zum einen ist die Extraktion häufiger Kollokate nicht Teil der vorliegenden Arbeit. Zum anderen soll auch nicht versucht werden, Argumente und Modifikatoren auf automatischem Wege zu unterscheiden. Des Weiteren ist auch nicht Teil dieser Arbeit, aus den extrahierten Valenzmustern auf automatischem Wege generalisierte oder abstrahierte Muster abzuleiten.

Die Extraktion identifiziert das semantische Hauptverb eines Satzes, wie von der CroCo-Annotation vorgegeben, und wertet die das Verb umgebenden Elemente aus. Damit wären alle Sätze und satzwertigen Konstruktionen, d.h. Einzelsätze, geeignete Ausgangspunkte für das Extraktionsexperiment, jedoch legt die Annotation des CroCo-Korpus dem Experiment Beschränkungen auf.

In Abschnitt 3.3 wurde auf einige der Nachteile der Annotation des CroCo-Korpus eingegangen. Dabei wurde u.a. darauf hingewiesen, dass die Annotation der linguistischen Strukturen des Korpus flach ist. Das bedeutet, dass z.B. Nebensätze als solche markiert, aber ihre innere Struktur nicht aufgeschlüsselt ist. Dies ist für die Untersuchung von Valenzstrukturen sehr nachteilig; alle verbalen Strukturen, die nicht auf der obersten Satzebene auftreten, gehen damit für die automatische Extraktion verloren (vgl. dazu Beispiel (21) auf Seite 99). Dies ist ein großer Nachteil gegenüber der Annotation tiefer Dependenzstrukturen wie in der PDT, oder der Phrasenannotation vom Satzknoten bis hin zu den Terminalen wie in TIGER.

Tatsächlich wurden Einzelsätze in der zweiten Projektphase von CroCo in einigen Registern aufgeschlüsselt, dennoch wurden sie aus zwei Gründen nicht im Experiment ausgewertet. Zum einen war diese Annotation zum Zeitpunkt der Untersuchung noch kaum verfügbar; zum anderen werden auch in dieser Annotation auf der ersten Einbettungsebene nicht alle Relationen klar, so ist z.B. nicht konsequent das Prädikat eines Einzelsatzes annotiert, wodurch eine Extraktion von VRM, die sich ja insbesondere auf das Prädikat stützt, nicht möglich ist. Außerdem fehlen andere Einbettungstypen, wie z.B. Partizipialkonstruktionen, die in einer Annotation bis zu den Wortknoten beinhaltet wären.

Aus der vorangehenden Darstellung ergeben sich schon im Vorfeld zwei wichtige Konsequenzen für das Extraktionsexperiment:

- Die Extraktion stützt sich nur auf die syntaktischen Strukturen der obersten Satzebene.
- Ergebnis der Pilotstudie wird auch eine Reihe von Desiderata für die Erweiterung der CroCo-Annotation sein.

Wie auch für die Vorstudien werden die Ergebnisse nach Register und Übersetzungsrichtung getrennt durchgeführt und ausgewertet. Für die Abfrage paralleler VRM wurde die Satzalignierung genutzt. Technisch gesehen wurde die Extraktion der Satzpaare mittels der CroCoAPI und insbesondere mittels des CoReTools realisiert. Die eigentliche Abfrage sowie Ausgabe operiert auf einer CoReTool-Datenstruktur, und ist somit auch für andere auf CoReTool abgebildete Formate verwendbar.

```
for every sentencePair in sentencePairs

     slSentence := getSlSentence(sentencePair)
     tlSentence := getTlSentence(sentencePair)
     slMainVerb := getMainVerb(slSentence)
     tlMainVerb := getMainVerb(tlSentence)
     slFunctions := getFunctions(slSentence)
     tlFunctions := getFunctions(tlSentence)

     printFrame(slMainVerb, slFunctions,
                tlMainVerb, tlFunctions)

end
```

Abbildung 4.4: Pseudo-Code für die Extraktion und Ausgabe von grammatikalisch-funktionalen Valenzmustern

Der Pseudo-Code der Abfrage ist in Abbildung 4.4 dargestellt. Der Algorithmus basiert auf der Abfrage von Satzpaaren, und der Ausgabe der Hauptverben sowie der Funktionen und Kategorien der dazugehörenden Ergänzungen jeweils für den AS- und den ZS-Satz. Für jeweils den AS- und ZS-Satz wurden das Hauptverb sowie alle im Satz auf oberster Ebene realisierten grammatischen Funktionen zusammen mit ihren syntaktischen Kategorien extrahiert. Eine Ausgabe für ein Satzpaar entspricht dem Beispiel, das in Abbildung 4.5 dargestellt ist. In der ersten Zeile stehen nacheinander jeweils in Anführungszeichen die AS- und ZS-Valenzrealisierungsmuster der semantischen Hauptverben des AS- und ZS-Satzpaars, mit dem AS-Satz in der zweiten Zeile und dem ZS-Satz in der dritten Zeile. Die Realisierungsmuster bestehen jeweils aus dem semantischen Hauptverb in Anführungszeichen, z.B. *„leisten“*, nach dem Doppelpunkt gefolgt von einer Liste aus den Funktionen im Satz gepaart mit der Form, in der die entsprechende Funktion realisiert ist. Die Paarungen von Funktion und Kategorie sind durch ein Doppelkreuz getrennt, z.B. ein als NP realisiertes Subjekt wird durch *subj#np* dargestellt. Die Musterdarstellungen für AS- und ZS-Verb sind durch Strichpunkt getrennt.

Das Hauptverb des AS-Satzes in 4.5 ist *leisten*, mit dem Subjekt *Finanzmärkte* und dem direkten Objekt *einen nicht zu unterschätzenden Beitrag […] Volkswirtschaft*. Auf der Gegenseite steht das Hauptverb *make* mit dem Subjekt *Financial markets* und dem direkten Objekt *an invaluable contribution […] growth*. In diesem Beispiel ist zu erkennen, dass es keine direkte Ent-

```
"leisten": subj#np, dobj#np, ; "make": subj#np,
dobj#np,
Finanzmärkte leisten einen nicht zu unterschät-
zenden Beitrag für das Wachstum einer Volkswirt-
schaft .
Financial markets make an invaluable contribution
to economic growth .
```

Abbildung 4.5: Ein extrahiertes Satzpaar, mit Hauptverben und VRM, aus dem Register G2E_SPEECH

sprechung zwischen den beiden Hauptverben gibt: *leisten* und *make* würden typischerweise wohl nicht als Übersetzungen voneinander im Wörterbuch stehen. Tatsächlich sind die Valenzträger nicht die Hauptverben selbst, sondern jeweils ein Funktionsverbgefüge, nämlich *einen Beitrag leisten* und *make a contribution*. Das hat auch Auswirkungen auf den Satzbau. Wäre im Original das Verb *beitragen* statt des Funktionsverbgefüges verwendet worden, wäre die Postmodifikation des direkten Objekts *für das Wachstum einer Volkswirtschaft* zum Objekt von *beitragen* geworden.

Aus dem Extraktionsbeispiel in Abbildung 4.5 ließe sich folgender syntaktisch-funktionaler Wörterbucheintrag für ein MÜ-System extrahieren:

leisten(subj:np, dobj:np) <=> make (subj:np, dobj:np)

oder, mit kollokativem Gehalt:

leisten(Finanzmarkt, Beitrag) <=> make(financial market, contribution)

Allerdings wäre ein solcher Eintrag nicht befriedigend, da er die beiden Verben *leisten* und *make* gleichsetzen würde. Tatsächlich sind die beiden Verben nicht typischerweise Übersetzungen voneinander; in der Formulierung *Beitrag leisten* allerdings wird *leisten* im Englischen mit *make* übersetzt.

Teil der Extraktion von Valenzen ist also mehr als nur die Extraktion von AS-ZS-Verbpaaren – was der Parallelitätsannahme, wie in Abschnitt 4.1 beschrieben, widerspricht. Im vorliegenden Fall besteht der Valenzträger aus einem Funktionsverbgefüge, also einem Verb plus Nomen, wobei die Verbbedeutung verblasst und das Nomen in erheblichem Maße zur Gesamtbedeutung des Ausdrucks beiträgt.

Für das Deutsche und das Englische stehen für Konstruktionen wie Funktionsverbgefüge viele Wörterbücher in gedruckter wie elektronischer Form zur Verfügung. Allerdings lässt sich darüber streiten, inwiefern diese aber alle möglichen Konstruktionen abdecken, oder ob die verschiedenen Definitionen dieses Phänomens immer einander entsprechen (vgl. z.B. Langer 2004).

Zur Erkennung von Funktionsverbgefügen – auch von solchen, die von Lexika noch nicht erfasst sind – würden sich, da die Daten ja korpusbasiert extrahiert werden, statistische Methoden eignen. Jeder Eintrag wäre mit statistischen Werten erweiterbar, z.B. damit, wie oft an der Position des direkten Objekts eine NP, und wie oft ein Einzelsatz steht. Auch können Daten darüber gewonnen werden, mit welchen Lexemen die Positionen belegt sein können.

Allerdings ist dem Autor dieser Arbeit keine zuverlässige statistische Methode zur Erkennung solcher Strukturen bekannt. Dafür gibt es eine Reihe von Arbeiten über zur Erkennung von starken und schwachen Kollokationen, auch im multilingualen Bereich (vgl. Abschnitt 2.4.3 und darin erwähnte verwandte Ansätze für die MÜ und HÜ).

Die Erkennung von starken Kollokationen ist mit Bezug auf die Beziehung zwischen Valenzträger und Begleiter insbesondere dann von Nutzen, wenn klar ist, welcher der Begleiter ein Aktant und welcher nur ein Modifikator ist. Auch diese Erkennung müsste bei der Menge der aus Korpora extrahierten Daten statistisch ablaufen. Hier hat sich zuletzt insbesondere ein Ansatz von Liu und Sarkar (2009) hervorgetan, der bei der Erkennung von Aktanten im Englischen mit Spezifitätswerten von über 90 % brilliert. Allerdings basiert dieser Ansatz auf einem LTAG-Formalismus und ist bisher nur fürs Englische getestet. Auf eine Umsetzung für weitere Formalismen und Sprachen ist zu hoffen.

Auf einen weiteren Vorteil der valenzbezogenen Kollokationsextraktion sei hier hingewiesen: Die statistischen Daten über gemeinsame Auftretenshäufigkeiten eines Valenzträgers mit bestimmten Begleitern könnten den üblichen N-gram-Modellen in der MÜ dadurch überlegen sein, dass sie keine Funktionswörter berücksichtigen und damit weniger relevante Daten auslassen (die Generierung von Funktionswörtern wie z.B. Artikeln würde dann einer regelbasierten Grammatik obliegen).

Wie das in diesem Abschnitt diskutierte Extraktionsbeispiel zeigt, kann man bei der VRM-Extraktion nicht immer davon ausgehen, dass die beiden Hauptverben des Extraktionssatzpaares direkt äquivalent sind; üblicherweise würden *leisten* und *make* nicht als direkte Übersetzung voneinander gelten,

die im diskutierten Beispiel vorliegenden Funktionsverbgefüge sind sie aber äquivalente Valenzträger. Dem Thema möglicher syntaktischer oder semantischer Divergenzen widmet sich der Abschnitt 4.3.2 im Detail. Das Hauptaugenmerk liegt auf syntaktischen Divergenzen, z.B. der Realisierung als Funktionsverbgefüge vs. als Vollverb. Semantische Divergenzen werden nicht weiter kategorisiert, sondern nur eben jene Fälle aufgeführt, in denen die Hauptverben im gegebenen Kontext semantisch nicht voll deckungsgleich sind.

4.3 Ergebnisse

In den folgenden Abschnitten sind die Ergebnisse der Vorstudien (4.3.1.1 bis 4.3.1.4) sowie des VRM-Extraktionsexperiments (4.3.2) aufgeführt. Dabei werden jeweils sowohl quantitative Gesichtspunkte berücksichtigt als auch illustrierende Einzelbeispiele diskutiert. Der Bezug zu valenzrelevanten Übersetzungsphänomenen wird dabei für jede der analysierten Ebenen hergestellt.

4.3.1 Ergebnisse der Vorstudien geordnet nach valenzrelevanten Kategorien

Im Folgenden sind die Ergebnisse der Vorstudien aufgeführt, die aufzeigen, inwiefern die zuvor, insbesondere in Abschnitt 4.1 beschriebene Parallelitätsannahme für deutsch-englische parallele Texte mit Blick auf die valenzrelevanten Kategorien Satz, Einzelsatz und grammatische Funktion als gültig angesehen werden kann. An Stellen, an denen eine Parallelität nicht gilt, werden mögliche Ursachen für Divergenzen diskutiert.

4.3.1.1 Empty links bei Sätzen

Sätze bzw. Satzpaare dienen als Grundlage für die VRM-Extraktion. Um zu wissen, inwiefern sich die Satzalignierung für die Extraktion eignet, wurde im Rahmen der Vorstudien untersucht, wie oft Sätze Entsprechungen in der jeweils anderen Sprache haben, oder umgekehrt, wie häufig empty links bei Sätzen vorkommen.

Tatsächlich spielen empty links bei der Satzalignierung kaum eine Rolle: In den drei untersuchten Registern FICTION, SHARE und SPEECH liegt der prozentuale Anteil der alignierten Sätze stets bei über 99%. Die Fälle, in denen keine Alignierung vorliegt, sind meist auf folgende zwei Faktoren zurückzuführen:

- kommunikative Präferenzen: In SHARE werden Überschriften wie etwa *Dear shareholder* in einer neutraleren Fassung wie etwa *Langfristiger Erfolg* wiedergegeben. Emotive Ausdrücke, die in englischer Wirtschaftskommunikation nicht selten sind, wie etwa *No way!*, fallen im Deutschen komplett weg.
- weggefallene Textabschnitte: An einigen Stellen hat der Übersetzer des Textes ganz Sätze oder Abschnitte aus dem AS-Dokument weggelassen, etwa wenn neue Mitglieder eines Direktoriums aufgeführt werden. Der Übersetzer mag diese Information für den Leser der ZS für nicht wichtig erachtet haben; solche Entscheidungen sind aber im Zuge dieser Studie nicht von Relevanz, da sie kein Phänomen mit einem direkten Bezug zur Valenzextraktion darstellen.

Aufgrund der geringen Anzahl an empty links bei Sätzen kann davon ausgegangen werden, dass Information, die in einem Satz enthalten ist, meistens entsprechend in die andere Sprache übertragen wurde, wenn auch in einigen Fällen von einem Satz in mehrere aufgespalten wurde. Allerdings kann die Strukturierung der Information unterschiedlich sein: Einzelsätze aus einem AS-Satz können in einem anderen ZS-Satz auftauchen, als der ZS-Satz, mit dem der AS-Satz aligniert ist, und umgekehrt. Diese Art von crossing line – zwischen Sätzen und Einzelsätzen – erscheinen für diese Studie allerdings nicht relevant, da sie nicht direkt von lokalen Valenzeffekten abhängen, sondern auf der Textebene stattfinden und eher von der vom Autor vorgenommenen Aufteilung der Informationsstruktur bedingt sind.

4.3.1.2 Empty links bei Einzelsätzen

Neben Sätzen eignen sich theoretisch auch Einzelsätze zur Extraktion von VRM. Wie bereits erwähnt, schränkt die CroCo-Annotation die Extraktion an dieser Stelle stark ein, da Einzelsätze in ihrer Tiefenstruktur nicht oder nur ungenügend aufgeschlüsselt sind. Dennoch sind sie für die Vorstudien aus (mindestens) zwei Gründen interessant:

- um Erfahrungswerte für mögliche erweiterte Studien bezüglich einer auch auf Einzelsätzen basierten VRM-Extraktion zu sammeln,
- und weil Einzelsätze in einer Sprache mit nominalen Strukturen in der anderen Sprache dargestellt sein können, was einen zusätzlichen Divergenztyp darstellt.

An dieser Stelle werden für Einzelsätze nur die empty links ausgewertet. Eine detaillierte Auswertung von crossing lines zwischen Einzelsätzen und Sätzen bzw. zwischen Wörtern und Einzelsätzen findet sich in (Hansen-Schirra u. a. [erscheint]: Kap. 6), erscheint aber, wie im Abschnitt zuvor schon ausgeführt,

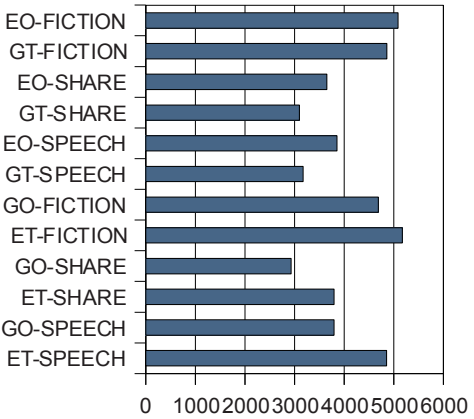

Abbildung 4.6: Absolute Zahlen für Einzelsätze

für die valenzbezogenen Vorstudien weniger relevant, da sie eher auf Unterschiede in der Informationsstrukturierung als in der Valenz hindeuten.

Abbildung 4.6 zeigt die absoluten Zahlen für Einzelsatzvorkommen in englischen und deutschen Originalen (GO und EO) und Übersetzungen (GTrans und ETrans). Englische Texte enthalten unabhängig von Register oder Übersetzungsrichtung immer deutlich mehr Einzelsätze als das Deutsche; dieser

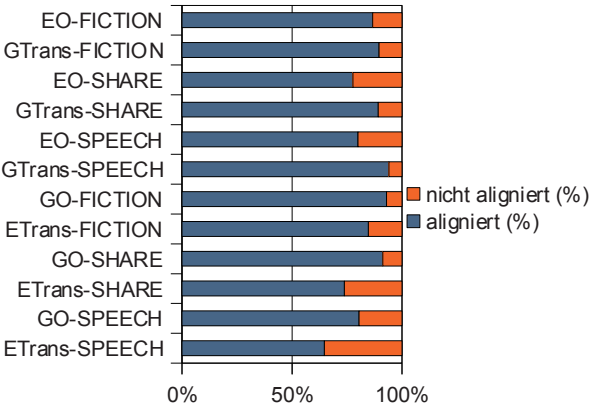

Abbildung 4.7: Alignierte und nicht alignierte Einzelsätze in Prozent

sprachtypologische Unterschied wird anhand der in diesem Abschnitt gezeigten Beispiele deutlich. Aufgrund der höheren Zahl von Einzelsätzen im Englischen ist es auch wenig verwunderlich, dass, wie Abbildung 4.7 zeigt, die Einzelsätze in englischen Texten unabhängig von Register und Übersetzungsrichtung durchgängig mehr empty links aufweisen als in deutschen Texten. Im Folgenden sollen einige der Ursachen für empty links bei Einzelsätzen aufgeführt werden.

Zum einen gibt es Fälle, in denen im Englischen sprachtypische Einzelsätze für ein deutsches Adverb stehen, wie in Beispiel (6) (Einzelsätze sind mit eckigen Klammern markiert, nicht alignierte Einzelsätze und deren eigentliche Äquivalente unterstrichen):

(6) a. [*Deshalb machen hohe Abgaben Arbeit teuer*] [*und können doch nicht verhindern,*] [*dass unseren Sozialsystemen der Kollaps droht.*] (GO_SPEECH_007)

b. [*That is why*] [*high taxes make work expensive*] [*and yet cannot protect our social system from*] [*impending collapse.*] (ETrans_SPEECH_007)

Für das deutsche Adverb *deshalb* (gleichzeitig eine Phrase) steht im Englischen der ganze Einzelsatz *That is why*, vergleichbar mit einer Reihe anderer Einzelsätze wie *It is for this reason, that...* u.ä. Es gibt zwar auch im Englischen entsprechende einleitende Adverbien wie z.B. *therefore*, die aber für diese Art von Register wohl nicht typisch sind, wie die Longman-Grammatik von Biber u. a. (2000) aufführt. Bei der Auszählung von einleitenden Adverbien wie *thus* oder *therefore* stellen die Autoren fest, dass diese v.a. im Bereich des akademischen Schreibens auftreten (600 Auftreten von *therefore* in 1 Million Wörter), in anderen Registern wie etwa Konversation dagegen kaum (<50 Auftreten in 1 Million Wörter). Für das Register SHARE gibt es keine verbindliche Aussage, mit diesem Hintergrund erscheint es allerdings durchaus plausibel, dass sich der Übersetzer für *that is why* anstatt für ein einleitendes Adverbial entschieden hat.

Was den Umgang mit solchen Einzelsätzen angeht, so erscheint es am sinnvollsten, sie per Eintrag als feste Wendung in ein Lexikon von einer möglichen, Einzelsatz-basierten Valenzextraktion auszuschließen.

Ein anderer Fall, in dem das Englische einen anderen Weg geht und verbale statt nominaler Strukturen verwendet, ist der Fall von eingebetteten NP wie in den folgenden Beispielen (7) und (8) aufgeführt.

(7) a. [*Die Staats- und Regierungschefs der Europäischen Union haben in Göteborg erneut ihre Bereitschaft be kräftigt,*] [*die in Kyoto eingegangenen Verpflichtungen zur Verminderung der Treibhausgase zu erfüllen.*] (GO_SPEECH_001)

b. [*In Gothenburg the EU heads of state and government reaffirmed their willingness*] [*to fulfil the commitments*] [*they made in Kyoto*] [*to reduce greenhouse gases.*] (ETrans_SPEECH_001)

(8) a. [*Mit der am 16. Juli in Bonn beginnenden Klimakonfe renz der Vereinten Nationen gehen die jahrelangen Be mühungen um ein verbindliches Klimaschutz-Abkom-men in die entscheidende Phase.*] (GO_SPEECH_001)

b. [*With the UN Climate Conference* [*beginning in Bonn on July 16*] *the many years of efforts* [*aimed at*] [*achieving a climate protection agreement*] *will enter the crucial final phase.*] (ETrans_SPEECH_001)

In (7) ist die eingebettete PP plus Adjektiv *in Kyoto eingegangenen* sowie die PP *zur Verminderung der Treibhausgase* im Englischen jeweils mit einem Einzelsatz wiedergegeben: die eingebettete NP mit dem Relativsatz *they made in Kyoto*, und die *zur*-PP mit Infinitivkonstruktion *to reduce greenhouse gases*. In (8) ist dies ähnlich mit den PP *am 16. Juli* und *in Bonn* der Fall. Die eingebettete Partizipialphrase *beginnend* in der Funktion eines Adjektivs aus dem deutschen Original ist im Englischen mit dem Partizip *beginning* übersetzt, und leitet im Englischen einen Einzelsatz ein, in dem wiederum die beiden PP aus dem Deutschen enthalten sind. Durch diesen Einzelsatz, ebenso wie durch die Einzelsatzäquivalente zu *Bemühungen* und *um ein verbindliches Klimaschutz-Abkommen* wird der Hauptsatz *With the UN Climate Conference ... the many years of efforts ... will enter the crucial final phase* diskontinuierlich.

Zwischen den deutschen Originalen und den englischen Übersetzungen lassen sich solche Shifts häufig beobachten, in denen partizipiale oder nominale Strukturen im Deutschen (allerdings oft auch mit einem Verlaufscharakter, nämlich Partizip Präsens wie in (8) bzw. ein Substantiv mit der *ung*-Endung, wie im folgenden Beispiel) ins Englische mit verbalen Konstituenten übersetzt wurden. Dieses Phänomen tritt auch in Beispiel (9) auf:

(9) a. [*Mittlerweile ist anerkannt,*] [*dass es zur Sicherung von Beschäftigung vor allem auf Flexibilität ankommt.*] (GO_SPEECH_007)

b. [*It has now been recognized*] [*that flexibility is the most important factor*] [*when it comes*] [*to safeguarding jobs.*] (ETrans_SPEECH_007)

Die PP *zur Sicherung von Beschäftigung* ist im Englischen sogar mit zwei Einzelsätzen übersetzt, nämlich mit *when it comes* und *to safeguarding jobs*. Wie zuvor in (8) gibt es für die *zur*-PP ein Äquivalent in Form eines *to*-Einzelsatzes.

Bei der Analyse von empty links bei Einzelsätzen sind also grundsätzlich zwei Fälle zu unterscheiden. Zum einen Fälle, in denen deutsche Konnektionsadverbiale als englisches Äquivalent einen ganzen Nebensatz haben. Für diesen Fall benötigt es aus MÜ-Sicht ein Lexikon, sowie die Information, in welchem Register im Englischen ein Nebensatz oder ein förmlicheres Adverb stehen sollte. Zum anderen ist für empty links bei Einzelsätzen ursächlich, dass deutsche nominale (und adjektivische) Strukturen im Englischen eher als verbale Strukturen auftreten. Eingebettete Adjektivphrasen aus dem Deutschen können im Englischen Relativsätzen oder Partizipialkonstruktionen entsprechen. Für *zu*-PP wurden im Englischen Äquivalente in Form von *to*-Einzelsätzen gefunden. Inwiefern es sich dabei um regelhafte Beziehungen handelt, könnte allerdings erst geklärt werden, wenn eine tiefere Annotation, z.B. auch eingebetteter Strukturen, bestünde.

Durch Gebrauch verbaler Strukturen, wo im deutschen nominale Strukturen stehen, ergeben sich im Englischen insgesamt deutlich mehr Einzelsätze als im Deutschen, unabhängig von der Übersetzungsrichtung und unabhängig davon, ob es sich bei den englischen Texten um Originale oder Übersetzungen handelt. Damit hat das Englische zwar deutlich mehr empty links bei Einzelsätzen als das Deutsche, wie eingangs bereits festgestellt wurde. Das bedeutet allerdings nicht, dass Einzelsätze sich nicht zur Valenzextraktion eignen. Wie die Beispiele zeigen, können Einzelsätze ohne Äquivalent entweder als sprachtypische Wendungen klassifiziert werden, oder dem Einzelsatz im Englischen steht häufig eine Nominal- bzw. Partizipialkonstruktion im Deutschen gegenüber.

Die Herausforderung bei der automatischen Extraktion besteht darin, eingebettete Komponenten als solche zu erkennen und Grenzen zu anderen Konstituenten festzustellen. Zudem kommt es durch die Nominalisierung verbaler Ausdrücke zur **Valenzvererbung**, d.h. das Nomen erbt den Valenzrahmen

des Verbes, von dem es abgeleitet ist, und passt diesen den syntaktischen Bedingungen der Wortklasse *Nomen* an (z.B. Anschluss des Subjekts mit Genitiv wie in *Der Vortrag des Dozenten* statt *Der Dozent trug vor*). Allerdings geschehen die Prozesse der Nominalisierung und Valenzvererbung sprachübergreifend, was wiederum z.B. die automatische Alignierung erschwert.

Eine weitere, valenzbezogene Divergenz ist im Beispiel (7) zu finden, und zwar bei der Auflösung der deutschen Adverbialphrase *in Kyoto eingegangenen*, die im Englischen mit dem kompletten Relativsatz *they made in Kyoto* wiedergegeben ist. Die unpersönliche Partizipialkonstruktion aus dem Deutschen benötigt kein syntaktisches Subjekt. Anders der Relativsatz im Englischen: Um dort ein Subjekt für das Verb *made* bereitzustellen, wird mit dem Personalpronomen *they* Rückbezug genommen auf *EU heads of state and government*.

In Beispiel (9) ist der Fall noch etwas komplexer und damit für die automatische Extraktion noch schwieriger gelagert. Die Konstruktion *zur Sicherung von Beschäftigung* ist unpersönlich; sie besteht aus dem Prädikatsausdruck *zur Sicherung* und dem Objekt *von Beschäftigung*, ein Subjekt ist in dieser nominalen Konstruktion nicht nötig. Im englischen Pendant *to safeguarding jobs* fehlt das Subjekt ebenfalls, weil es sich um eine Partizipialkonstruktion handelt; im übergeordneten Einzelsatz *when it comes* ist allerdings ein generelles Subjekt *it* eingeführt. Dieses *it* ist keine Entsprechung zum *es* im deutschen Nebensatz *dass es zur Sicherung […] ankommt*, in dem das *es* in der Konstruktion von *ankommen auf* obligatorisch ist. Zum Subjekt im Englischen *that*-Nebensatz wird das deutsche Präpositionalobjekt *auf Flexibilität* und ist damit im Sinne des Funktionsmappings in CroCo Gegenpart zum deutschen *es*. Das *it* muss also aus valenztechnischen Gründen eingefügt worden sein, um der englischen verbalen Konstruktion Genüge zu tun und ein syntaktisches Subjekt zu liefern.

Im Rahmen weiterer Extraktionsexperimente wird zu klären sein, wie sich Valenzvererbungen und andere Alternationen zwischen dem Deutschen und dem Englischen verhalten. Diese Arbeit befasst sich zwar fast ausschließlich mit der Verbvalenz. Diese Beschränkung, die bei einer Extraktion aus einem einsprachigen Text einfach aufrecht zu erhalten ist, funktioniert bei einem Vergleich von Originalen und Übersetzungen nicht ohne Weiteres.

Die Extraktion aus Einzelsätzen wäre also nicht nur interessant, weil dann eine breitere Datenbasis vorhanden wäre als bei einer Extraktion nur aus den obersten Ebenen der Sätze, sondern sie würde zusätzlich eine Untersuchungsbasis für Valenzvererbung im Übersetzungsprozess bei Nominalisierungen liefern. Nominalisierungen als Übersetzungsstrategie werden z.B. von (Čulo

u. a. 2008b) im Rahmen von CroCo untersucht. Streiter (1995:Kap. 8) zeigt für die Valenzvererbung am Beispiel des MÜ-Systems CAT2, wie sie für die MÜ modelliert und implementiert werden kann.

4.3.1.3 Empty links bei grammatischen Funktionen

Auf der Ebene der Syntax haben verschiedene linguistische Theorien diverse Zwischenstufen der Darstellung der syntaktischen Information eingeführt. Erste kontextfreie Grammatiken arbeiteten mit einer einfachen Konstituentensyntax, die LFG verwendet eine funktionale Syntax (mit klassischer Subjekt-Objekt-Aufteilung), die Kasusgrammatiken entwickelten das semantisch motivierte Tiefenkasuskonzept, und die Prager Schule verwendet die tektogrammatische Ebene, die bereits viele Herleitungen und Umformungen aus der Oberflächensyntax enthält.

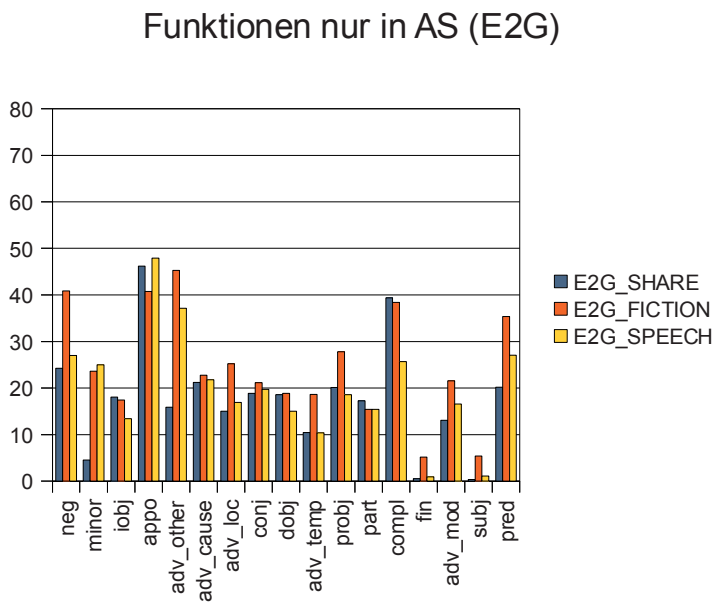

Abbildung 4.8: Anteil der Funktionen, die in einem Satzpaar nur in der AS vorhanden sind, Übersetzungsrichtung englisch-deutsch

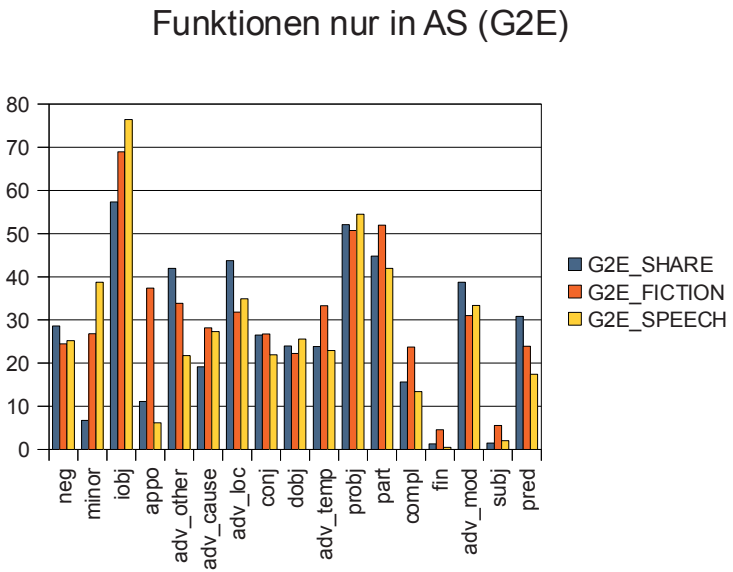

Abbildung 4.9: Anteil der Funktionen, die in einem Satzpaar nur in der AS enthalten sind, Übersetzungsrichtung deutsch-englisch

Die in CroCo verwendeten grammatischen Funktionen sind noch sehr syntaktisch orientiert und damit automatisch herleitbar. Ein guter Parser kann sowohl im Deutschen als auch im Englischen relativ zuverlässig Funktionen wie Subjekt, direktes Objekt oder indirektes Objekt zuweisen, z.B. ist das in CroCo verwendete Analyseprogramm MPro dazu in der Lage. Gleichzeitig abstrahieren Funktionen von rein formalen Kategorien der Konstituentensyntax weg. Eine Analyse auf Basis der grammatischen Funktionen erscheint also für die automatische Sprachverarbeitung und damit auch die MÜ sinnvoll.

Dieser Abschnitt analysiert empty links für grammatische Funktionen. Der darauf folgende Abschnitt untersucht crossing lines zwischen der Ebene der grammatischen Funktionen und der Wörter.

In den vorherigen Vorstudien mit Sätzen und Einzelsätzen konnte für das Auffinden von empty links die bestehende (in beiden Fällen manuelle oder zumindest manuell korrigierte) Alignierung verwendet werden. Da Phrasen in CroCo weder manuell noch automatisch aligniert sind, wurde hier mit einer anderen Strategie als in den zuvor vorgestellten Vorstudien verfahren. Für ein Satzpaar wurde aufgezählt, welche Funktionen jeweils im Ausgangs- und Zielsatz vorhanden sind. Dann wurde überprüft, welche Funktionen aus dem Ausgangssatz nicht im Zielsatz vorkommen und umgekehrt. Bei dieser Art des Vorgehens gibt es – neben der nicht zufriedenstellenden Wortalignierung – zwei Faktoren, die das Ergebnis verfälschen:

(a) Mehrfach in einem Satz vorkommende Funktionen werden nicht berücksichtigt. Gibt es also z.B. zwei Komplemente in einem englischen Satz, die beide nicht als Komplemente übersetzt wurden, wird dieses Phänomen aufgrund des Zählalgorithmus dennoch nur einmal gewertet.

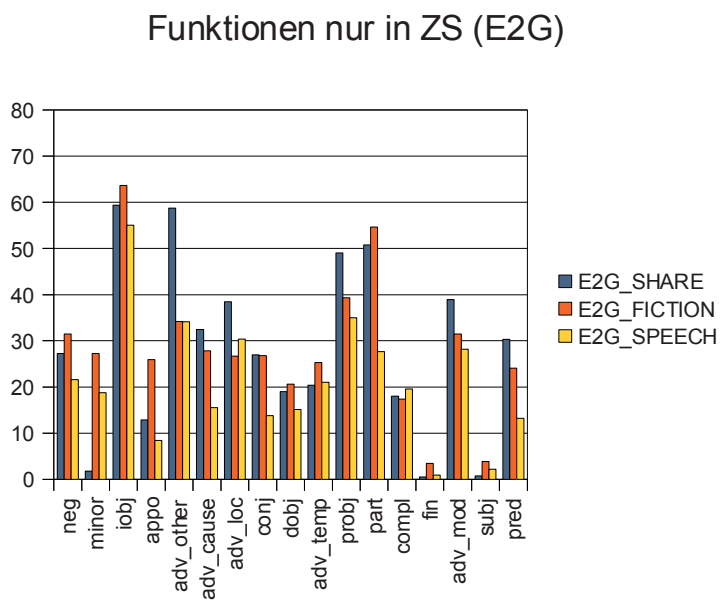

Abbildung 4.10: Anteil der Funktionen, die in einem Satzpaar nur in der ZS vorhanden sind, Übersetzungsrichtung englisch-deutsch

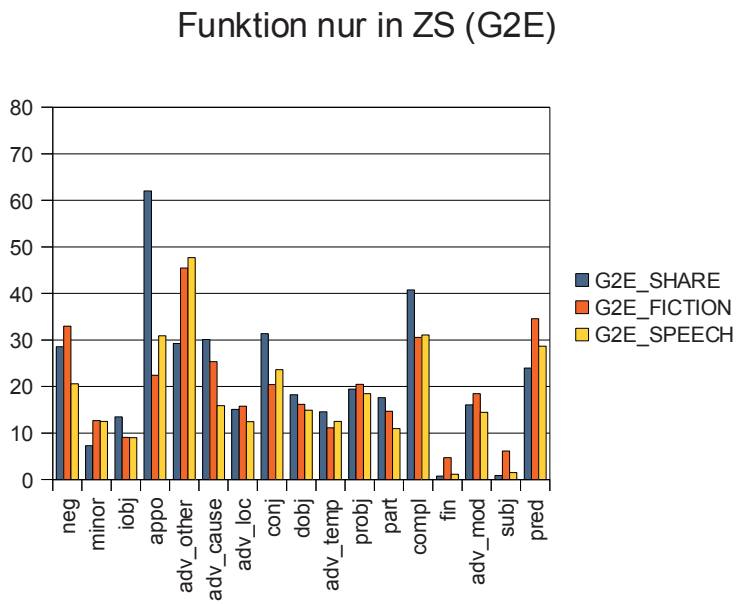

Abbildung 4.11: Anteil an Funktionen, die in einem Satzpaar nur in der ZS vorhanden sind, Übersetzungsrichtung deutsch-englisch

(b) Bei Mehrfachalignierungen von Sätzen, wenn z.B. ein Satz in zwei aufgespalten wurde, kann es passieren, dass eine Funktion in einem der alignierten Sätze realisiert wurde, in den anderen jedoch nicht. Jeder Satz, in dem die Funktion nicht realisiert wurde, zählt als Nichtrealisierung der Funktion.

Allerdings deutet die qualitative Auswertung der Daten nicht darauf, dass eines der beiden Phänomene übermäßig häufig auftritt. Zudem machen 1:n-Alignierungen von Sätzen maximal 8% der Alignierungen aus, und daher sind zumindest die Effekte von Punkt (b) zunächst vernachlässigbar. Dennoch wäre eine echte, möglichst manuelle, Phrasenalignierung nötig, um noch aussagekräftigere Daten zu erhalten.

Die Abbildungen 4.8 bis 4.10 zeigen, wie häufig grammatische Funktionen in der AS keine Entsprechung in der ZS finden und umgekehrt (in Prozent). Der besseren Vergleichbarkeit halber sind die Diagramme alle auf eine Skala von

0-80% genormt. Die Angaben sind nach Register und Übersetzungsrichtung aufgeschlüsselt. Wie den Diagrammen zu entnehmen ist, sind die Anteile der Funktionen ohne Entsprechung in der jeweils anderen Sprache je nach Register und Übersetzungsrichtung unterschiedlich. So haben in G2E_SHARE über 60% der englischen Appositionen keine Entsprechung im deutschen, in G2E_FICTION und G2E_SPEECH sind es aber nur ca. 20-30% (vgl. Abbildung 4.10). Dies ist ein Hinweis auf registerbedingte Ursachen.

Auf jede einzelne Funktion kann dabei in dieser Arbeit nicht eingegangen werden, einige Auffälligkeiten werden aber im Folgenden bei der Analyse valenzrelevanter Divergenzen angesprochen werden.

Auffällig ist beispielsweise, dass in der Übersetzungsrichtung Englisch-Deutsch ein hoher Anteil von Appositionen (APPO) und Komplementen (COMPL) nicht als solche übersetzt werden: mehr als 30% dieser Funktionen hat keine Entsprechung im ZS-Satz. Im Deutschen ist dies dagegen für Prädikatoren oder Modaladverbiale der Fall. Es müssten daher unterschiedliche Realisierungen dieser Funktionen in der jeweils anderen Sprache feststellbar sein.

Die Beispiele (10) und (11) zeigen zwei Fälle, in denen Appositionen im Englischen nicht als solche ins Deutsche übersetzt wurden. Die Apposition *a record* im Englischen wird im Deutschen mit einem ganzen Satz ausformuliert, ebenso wie die Apposition *an improvement of 2.3 turns*. In beiden Fällen wird die Information in den Satz integriert, anstatt am Ende herausgestellt zu werden. Außerdem sind die Strukturen im Deutschen komplexer, da sie als komplette Hauptsätze realisiert sind; es wird ein Valenzträger hinzugefügt, der im Englischen nicht präsent ist (*erreichen* bzw. *steigern* als Entsprechung von *improvement*). In der flachen CroCo-Annotation ist dies aber als solches nicht zuverlässig abfragbar.

(10) *a.* *Revenues rose 11 % to $ 112 billion, [a record]*$_{APPO}$. (EO_SHARE_004)

b. *Der weltweite Umsatz stieg um 11 % auf $ 112 Mrd. und erreichte damit eine neue Rekordhöhe.* (GTrans_SHARE_004)

(11) *a.* *Working capital turns hit an all-time high of 11.5 - [an improvement of 2.3 turns]*$_{APPO}$. (EO_SHARE_004)

b. *Die Umschlagshäufigkeit des Betriebskapitals konnte*

um das 2,3 fache gesteigert werden und erreichte die neue Höchstmarke von 11,5. (GTrans_SHARE_004)

Im Deutschen ist in diesen Beispielen die Abfolge der Informationen paratak-tisch geordnet, während im Englischen die Apposition als Hervorhebung ge-nutzt wird. Dies passt zu der Charakterisierung des Deutschen von House (1997) als inhaltsorientierter, während das Englische als eher adressatenori-entiert eingeordnet wird.

Ein weiteres typisches Merkmal insbesondere für Texte aus dem Register SHARE ist die häufige Verwendung von Komplementen. Hier zeigt sich ein kontrastives Merkmal zwischen dem Deutschen und dem Englischen. Im Englischen binden Verben wie *name*, *elect* oder *make*[34] Komplemente. Im Deutschen stehen an dieser Stelle häufig Präpositionalobjekte (siehe (12) und (14)). Eine andere Ursache für die häufigen Komplemente im Englischen ist die Tendenz, anstelle von Vollverben Kopulakonstruktionen zu verwenden, wie in Beispiel (13). Auf dieses Phänomen wird am Ende dieses Abschnitts genauer eingegangen.

(12) a. *Also for the second straight year, we were named [“The World's Most Respected Company”]$_{COMPL}$ by the Finan cial Times.* (EO_SHARE_004)

b. *Ebenfalls zum zweiten Mal in Folge ernannte die Fi nancial Times GE [zum “am meisten respektierten” Unternehmen der Welt]$_{PROBJ}$.* (GTrans_SHARE_004)

(13) a. *We are [pleased to present the 2001 Annual Report of the American Institute for Contemporary German Stud ies (AICGS)]$_{COMPL}$.* (EO_SHARE_013)

b. *Wir freuen uns, [Ihnen den Jahresbericht 2001 des American Institute for Contemporary German Studies (AICGS) präsentieren zu können]$_{PROBJ}$.* (GTrans_SHARE_013)

(14) a. *Return on beginning shareholders ' equity was [25 per cent]$_{COMPL}$, […]* (EO_SHARE_006)

b. *Die Eigenkapitalrendite zu Beginn der Rechnungsperi ode betrug [25 Prozent]$_{DOBJ}$ […]* (GTrans_SHARE_006)

34 Z.B. im Satz *They made him vice president*.

Bezüglich der Prädikatoren (PRED) sind im Folgenden drei Beispiele aufgeführt, die empty links bedingen. CroCo unterscheidet bei der Annotation zwischen Fällen, in denen ein finites Verb vorkommt, das zugleich semantisches Hauptverb ist (markiert mit FIN), und Fällen, in denen zwar ein finites Verb vorkommt (ebenfalls markiert mit FIN), das semantische Hauptverb aber in einer nicht-finiten Form auftaucht (markiert mit PRED). In (15) ist dies durch unterschiedliche Zeiten bedingt, Präsens im Englischen vs. Perfekt im Deutschen, wodurch das englische Hauptverb im Deutschen in ein Partizip Perfekt umgewandelt wird. König und Gast (2007) beschreiben bezüglich der unterschiedlichen Verwendungen von Perfekt und Präsens im Deutschen und Englischen eine Reihe von kontrastiven Unterschieden, ebenso wie bei den Vergangenheitszeiten Perfekt und Imperfekt. Diese kontrastiven Unterschiede, auf die an dieser Stelle nicht näher eingegangen wird, sind ursächlich für Wechsel zwischen FIN und PRED in beiden Übersetzungsrichtungen.

In (16) und (17) dagegen ist die Ursache für diesen Wechsel ein Stilmittel, das man eher dem Deutschen zuschreiben würde, nämlich das unpersönliche Passiv. Die Aktiv-Konstruktionen im Englischen werden im Deutschen mit Passiv und ohne semantisches Subjekt (im Englischen ist dies beide Male *we*) übersetzt. Dadurch wird das Hauptverb von der finiten Form in die Partizipialform bewegt, z.B. *we described → werden ... beschrieben*. Dies entspricht der Charakterisierung des Englischen als adressatenorientiert im Gegensatz zum inhaltsorientierten und oft unpersönlichen Deutschen (House 1997).

(15) *a.* *We already [have]$_{FIN}$ that!* (EO_SHARE_004)
b. *Das alles [haben]$_{FIN}$ wir bereits [geschafft]$_{PRED}$.* (GTrans_SHARE_004)

(16) *a.* *In that report, we [described]$_{FIN}$ several challenges and opportunities that we felt were going to determine the agenda of German-American relations.* (EO_SHARE_013)
b. *In diesem Bericht werden verschiedene Herausforde rungen und Gelegenheiten [beschrieben]$_{PRED}$, die unse rer Meinung nach die Beziehungen der beiden Staaten bestimmen.* (GTrans_SHARE_013)

(17) *a.* *It [progresses]$_{FIN}$ with a drumbeat regularity throughout our business year - year after year.* (EO_SHARE_004)
b. *Jahr für Jahr wird das Betriebssystem mit der Regelmä*

ßigkeit eines Paukenschlages [weiterentwickelt]$_{PRED}$.
(GTrans_SHARE_004)

Es finden sich zudem viele Beispiele, in denen Information adverbialisiert wird, die im Englischen entweder verbal vorhanden ist (wie *continue – weiterhin*, siehe Beispiel (20)), oder implizit vorhanden ist und explizit in der Übersetzung ausgedrückt wird, wie in den Beispielen (18) und (19). Dies generiert empty links auf der Ebene der Adverbien, in den angeführten Beispielen immer für Modaladverbiale.

(18) *a.* *Wireless networks will transform the workplace.* (EO_SHARE_005)
b. *Drahtlose Netzwerke werden den Arbeitsplatz [grundle gend]$_{ADV\_MOD}$ verändern.* (GTrans_SHARE_005)

(19) *a.* *Mostly, it involves creating and distributing paper doc uments or telephoning and meeting with fellow employ ees.* (EO_SHARE_005)
b. *In den meisten Fällen erstellen und verteilen sie Pa pierdokumente oder telefonieren oder treffen sich [per sönlich]$_{ADV\_MOD}$ mit anderen Mitarbeitern.* (GTrans_SHARE_005)

(20) *a.* *We have continued our efforts to ease the suffering of families of lost colleagues.* (EO_SHARE_005)
b. *Unsere Anteilnahme und Hilfe gilt [weiterhin]$_{ADV\_MOD}$ den Familien, deren Angehörige dabei ihr Leben lassen mußten.* (GTrans_SHARE_007)

In diesem Abschnitt wurden eine Reihe von möglichen Ursachen für empty links auf der Ebene der grammatischen Funktionen aufgeführt. Diese einzelnen Ursachen lassen sich wiederum größeren Klassen mit linguistischem und translatorischem Bezug zuweisen. So ist das unpersönliche Passiv, das eher dem Deutschen als Stilmittel zuzuordnen wäre, sicherlich eine Ursache mit stilistisch-kommunikativem Hintergrund, die der Einordnung des Deutschen als inhaltsorientiert (vs. der Adressatenorientiertheit des Englischen) entspricht. Die Verwendung von Appositionen als Herausstellung im Englischen gegenüber einer parataktischen Ordnung von Information im Deutschen entspricht ebenfalls dieser Charakterisierung. Die Hinzufügung von Adverbien,

die implizite Information expliziter macht, kann einer der Übersetzungsstrategien, der Explizierung, zugeordnet werden. Als kontrastiven Unterschied könnte man den Wechsel von Komplement zu Präpositionalobjekt sehen, da dieser durch eine sprach-spezifische Ausprägung von Valenzrahmen bestimmter Verben (Verben mit dem semantischen Aspekt „machen-zu", wie *name*, *elect*, *make*) bedingt ist. Dies ist allerdings nicht die einzige Ursache für ein gehäuftes Auftreten von Komplementen, wie im nächsten Absatz gezeigt wird.

Die hier aufgeführten Ursachen bedürfen weiterer empirischer Auswertung und dazu geeigneter Fragestellungen. Exemplarisch soll herausgegriffen werden, was in den Beispielen (13) und (16) gezeigt wurde und mit den im letzten Absatz erwähnten Häufungen von Komplementen zusammenhängt: der häufige Gebrauch von Kopulakonstruktionen im Englischen. Viele Formulierungen mit „be" werden im Deutschen mit einem anderen Verb übersetzt; vergleicht man dabei das Vorgehen in verschiedenen Registern, kann man auch die unterschiedlichen Zahlen für empty links von Komplementen erklären. In den englischen SHARE-Texten kommt auf der obersten Satzebene zwar weniger häufig „be" als Vollverb vor als in den SPEECH-Texten (341 vs. 371 Vorkommen), allerdings werden in deutschen SHARE-Texten Kopulakonstruktionen häufiger nicht als solche übersetzt (215 mal) als in SPEECH-Texten (nur 147 mal). Dadurch bleiben auf der englischen SHARE-Seite mehr Kopulakonstruktionen mit prädikativen Komplementen „übrig", wodurch sich die Häufung von empty links bei COMPL erklären lassen dürfte. Beispiel (14) zeigt dabei etwas, was sich bei näherem Hinsehen als durchaus prototypisch für das Register SHARE erweisen könnte: Während sich das Englische einer einfachen und neutraler klingenden Kopulakonstruktion bedienen kann, wird im Deutschen registerbedingt eine sehr formelhaft wirkende Sprache verwendet, etwa mit Formulierungen wie *X beträgt...*, *Gewinn erwirtschaftet* u.ä. Diese Hypothese müsste auf alle in Wirtschaftstexten vorkommenden Verblexeme erweitert und für diese überprüft werden; für eine computergestützte empirische Untersuchung fehlt aber nach wie vor eine echte Funktionsalignierung in CroCo.

4.3.1.4 Crossing lines zwischen Wörtern und grammatischen Funktionen

Trotz nicht optimaler Wortalignierung wurde eine Abfrage auf crossing lines zwischen grammatischen Funktionen und Wörtern durchgeführt; zumindest Tendenzen könnten dabei auffindbar sein. Es ist zu erwarten, dass sich hier am ehesten Divergenzen in der Argumentstruktur eines Satzes zwischen AS

und ZS auffinden lassen. Inwiefern sich die Ursachen alleine auf das verwendete Verblexem beschränken, oder weitere Faktoren – insbesondere das Register – hinzukommen, wird im Folgenden untersucht.

Crossing lines zwischen anderen Ebenen – zwischen Wörtern und Einzelsätzen und zwischen Einzelsätzen und Sätzen – wurden, wie bereits erläutert, nicht ausgewertet, da sie eher mit Bezug auf die Informationsstrukturebene denn auf die Valenzebene zu interpretieren sind. In der vorliegenden Arbeit ist die Auswertung von crossing lines zwischen Wörtern und grammatischen Funktionen also die einzige Auswertung von crossing lines; die übrigen Auswertungen können (Hansen-Schirra u. a. [erscheint]: Kap. 6) entnommen werden.

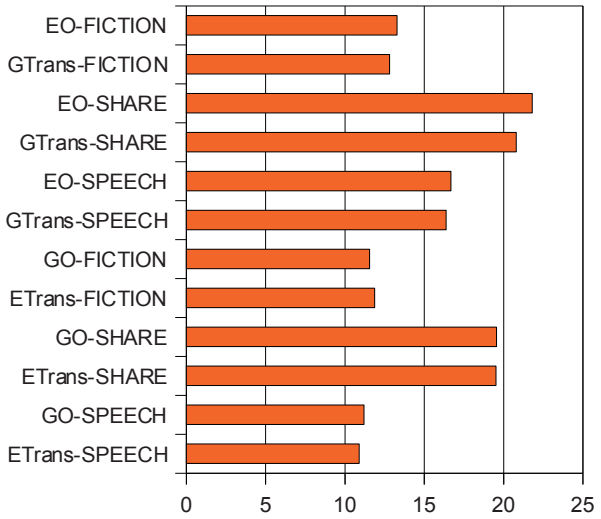

Abbildung 4.12: Crossing lines zwischen Wörtern und grammatischen Funktionen

Die Abbildung 4.12 zeigt die prozentualen Anteile für crossing lines zwischen Wörtern und grammatischen Funktionen. Die Prozentzahlen sind dabei mit den Phrasenanzahlen als Basis errechnet; d.h. eine Zahl von 35% bedeutet, dass bei 35% der Phrasen mindestens ein Wort der Phrase mit einem Wort aligniert ist, das in einer Phrase mit einer anderen grammatischen Funktion

eingebettet ist. Bei der Abfrage der Wortalignierung wurden übrigens nur die vier Wortklassen Adjektiv, Adverb, Nomen und Verb berücksichtigt, da reine Funktionswörter häufig Mehrfachalignierungen bedingen (z.B. sind Artikel in einer Sprache oft mit allen Artikeln in der anderen Sprache aligniert).

Dass sich die Zahlen zwischen Original und Übersetzung wenig unterscheiden, lässt sich mit dem Vorgehen der Auszählung erklären. Crossing lines werden immer paarweise gezählt: Ist die AS-Funktion anders als die ZS-Funktion, wird jeweils auf beiden Seiten eine crossing line verbucht. Die geringen Unterschiede in den Prozentzahlen rühren dann nur noch von unterschiedlichen Funktionsanzahlen im Deutschen und Englischen. Bemerkenswert sind allerdings die großen Unterschiede zwischen den Registern, weswegen diese Ergebnisse auch registerweise aufgeschlüsselt sind.

Diese Unterschiede zwischen den Registern dürften auf unterschiedliche Funktionsverteilungen in den Registern, ebenso wie auf registerspezifische Wanderungsbewegungen von AS- zu ZS-Funktion zurückzuführen sein. Mehr Einblick in diese Aspekte gibt die Tabelle 4.1, die Tendenzen für Wanderungsbewegungen von Funktionen geordnet nach deren Häufigkeit, sortiert nach Register und Übersetzungsrichtung, aufführt. Dabei lassen sich einige recht klare Tendenzen ablesen. So wird unabhängig vom Register häufig aus einer englischen Funktion in ein deutsches Präpositionalobjekt, bzw. aus einem deutschen Präpositionalobjekt in eine andere englische Funktion übersetzt. D.h., das Deutsche scheint gegenüber dem Englischen zu einer stärkeren Verwendung von Präpositionalobjekten zu neigen.[35] Gleiches gilt für die Übersetzung englischer Komplemente in andere Funktionen im Deutschen bzw. von deutschen Funktionen in Komplemente im Englischen. Hier scheint es einen zusätzlichen Einfluss des Registers zu geben, da diese Wechsel in SHARE und SPEECH deutlich prominenter sind als in FICTION.

Wie zuvor bei den Statistiken für empty links von grammatischen Funktionen kann nicht auf jeden einzelnen Typ von Funktionswechseln eingegangen werden, einige prominente davon werden aber in der folgenden Analyse angesprochen werden.

35 Hier sei kritisch angemerkt, dass die Klassifikation vieler englischer Verben mit anschließender Präposition in sogenannte **phrasal verbs** hier sicher zu einer Verschiebung der Zahlen führt. Es wird aber auch die Sichtweise vertreten, ein Verb wie *look for* nicht als phrasal verb zu analysieren, sondern als Sonderbedeutung von *look* mit entsprechendem präpositionalen Anschluss (vgl. z.B. die Valenzanalysen von Herbst u. a. 2004a). In nachfolgenden Studien sollte dieser Effekt daraufhin überprüft werden, wie stark der Einfluss auf die Prominenz bestimmter Funktionswechsel ist.

FICTION		**SHARE**		**SPEECH**	
E2G	**G2E**	**E2G**	**G2E**	**E2G**	**G2E**
dobj → subj	probj → dobj	compl → probj	probj → dobj	dobj → probj	subj → dobj
compl → dobj	dobj → subj	dobj → subj	subj → compl	dobj → compl	subj → compl
subj → dobj	fin → pred	dobj → probj	subj → dobj	compl → probj	probj → compl
dobj → fin	compl → subj	compl → dobj	probj → compl	subj → dobj	dobj → compl
dobj → probj	subj → dobj	dobj → compl	dobj → compl	dobj → subj	probj → dobj
fin → dobj	dobj → compl	compl → subj	fin → pred	pred → fin	dobj → subj
adv_mod → dobj	fin → compl	probj → dobj	dobj → subj	compl → dobj	fin → compl
pred → fin	pred → fin	subj → dobj	compl → dobj	compl → subj	fin → pred
compl → subj	fin → subj	fin → pred	adv_mod → compl	subj → compl	fin → subj
adv_cause → dobj	fin → dobj	pred → fin	subj → probj	compl → fin	compl → subj

Tabelle 4.1: Tendenzen für typische Wanderungsbewegungen von Funktionen, von oben nach unten sinkt die Häufigkeit. Der Ausdruck dobj→subj *steht z.B. dafür, dass ein direktes Objekt aus der AS in die Subjektfunktion in der ZS wechselt.*

Man kann bei dieser Vorstudie – und dies gilt auch für die Angaben in Tabelle 4.1 – nur deshalb von Tendenzen in den Ergebnissen sprechen, weil für die Auswertung dieser Wanderungsbewegungen die Wortalignierung verwendet wurde. D.h., es wurde für ein Inhaltswort wie z.B. *Zeitung* abgeprüft, in welcher Funktion es im Deutschen auftritt, und in welcher dann in der Englischen Übersetzung. Dass die Wortalignierung suboptimale Ergebnisse liefert, wurde bereits erwähnt. Die flache CroCo-Anntation fügt diesem Problem noch eine Ebene hinzu, wie an folgendem Beispiel ersichtlich ist:

(21) a. *Er hat sich [darauf]$_{PROBJ}$ verlassen, [dass wir von drin nen [sein Lächeln$_{EMBEDDED_DOBJ}$] sehen können]$_{PROBJ}$.* (GO_FICTION_007)

b. *He just assumed [we could see [his smile$_{EMBEDDED_DOBJ}$] from inside]$_{DOBJ}$.* (ETrans_FICTION_007)

Die flache Annotation führt dazu, dass das deutsche *Lächeln* einem (diskontinuierlichen) Präpositionalobjekt (*darauf..., dass wir […] können*) zugeordnet

wird; vergleicht man dies mit dem englischen *smile*, ergibt sich in der Auszählung ein Funktionswechsel, weil *smile* in einem direkten Objekt eingebettet ist. Tatsächlich ist aber das deutsche *Lächeln* Teil eines direkten Objekts, das in einem Präpositionalobjekt eingebettet ist. Da diese Einbettung in CroCo nicht aufgeschlüsselt wird, ergibt sich an dieser Stelle ein falsch-positiver Treffer bezüglich der crossing lines; die Erfahrungswerte zeigen jedoch, dass die Anzahl solcher falsch-positiver Treffer vernachlässigbar ist. Wie hoch allerdings der Anteil der dadurch nicht aufgefundenen crossing lines ist, lässt sich schwer abschätzen. In einem bereits in der Einleitung erwähnten CroCo-Nachfolgeprojekt soll dieses Problem durch eine tiefe Dependenzannotation gelöst werden.

Was den Wechsel vom Präpositionalobjekt zum direkten Objekt angeht, so kann man anhand der Beispiele (22)-(24) eine Tendenz des Deutschen zum Präpositionalobjekt erkennen. Die Tabelle 4.1 zeigt auch auf, dass der Wechsel vom Präpositionalobjekt zum direkten Objekt in der Übersetzungsrichtung vom Deutschen ins Englische und der Wechsel vom direkten Objekt zum Präpositionalobjekt in der Übersetzungsrichtung vom Englisch ins Deutsche zu den häufigsten Wanderungsbewegungen gehören.

(22) *a.* *1995 haben wir [auf 125 Jahre Deutsche Bank]$_{PROBJ}$ zu rückgeblickt.* (GO_SHARE_009)

b. *In 1995 we celebrated [Deutsche Bank's 125th an niversary]$_{DOBJ}$.* (ETrans_SHARE_009)

(23) *a.* *Nach wie vor ist der Zinsüberschuß nach Risikovorsor ge mit 9,7 Mrd DM die bei weitem wichtigste Ertrags komponente. Allerdings weisen die unterschiedlichen Steigerungsraten der einzelnen Ergebniskomponenten [auf die Veränderungen im Geschäft]$_{PROBJ}$ hin.* (GO_SHARE_009)

b. *Although net interest income after provision for losses on loans and advances, at DM 9.7 billion, is still by far the most important component of income, the individual figures highlight [the changes in our business]$_{DOBJ}$.* (ETrans_SHARE_009)

(24) a. *Daher setzen wir uns nachdrücklich [für die Schaffung eines europäischen Systems der Finanzaufsicht]$_{PROBJ}$*

ein. (GO_SPEECH_002)

b. *Hence we expressly support [the establishment of a European system of financial supervision]$_{DOBJ}$.* (ETrans_SPEECH_002)

Eine weitere Reihe von Funktionswechseln ergibt sich durch Verwendung unterschiedlicher Zeiten, wie in Beispiel (25), oder der im Englischen mittels Hilfsverb realisierten Verneinung, wie in Beispiel (26) (vgl. auch Abbildung 4.2, Seite 74). In beiden Fällen ist auf jeweils einer Seite das Hauptverb als PRED annotiert, weil das finite Hauptverb ein Hilfsverb ist. In (26) liegt dies wie erwähnt an der englischen Realisierung der Verneinung: Das finite Verb (Funktion FIN) in der englischen Übersetzung ist das zur Verneinung verwendete *don't*, das Verb *know*, das im Deutschen noch Vollverb ist, wird dadurch zum Infinitiv und wird als nicht-finiter Verbbestandteil (PRED) annotiert. In (25) wird im englischen Original ein Perfekt verwendet statt eines Präsens wie in der deutschen Übersetzung, dadurch tritt das semantische Hauptverb im Englischen als Partizip auf, im Deutschen als Vollverb. Dadurch ergibt sich ein Wechsel von PRED zu FIN.

(25) a. *And what has [happened]$_{PRED}$ before a few years have passed?* (EO_FICTION_006)

b. *Und was [geschieht]$_{FIN}$, ehe noch ein paar Jahre ver gangen sind?* (GTrans_FICTION_006)

(26) a. *Aber Sie [wissen]$_{FIN}$ nichts.* (GO_FICTION_007)

b. *But you don't [know]$_{PRED}$ anything.* (ETrans_FICTION_007)

Was die Verwendung unterschiedlicher Zeiten in Beispiel (25) angeht, so kann man an dieser Stelle wohl von einer einzelfallbedingten Divergenz sprechen; wohl nur der Kontext könnte den plötzlichen Wechsel vom Perfekt in ein eher futurbezogenes Präsens im Deutschen erklären. Wie bereits in 4.3.1.3 erwähnt, gibt es, von König und Gast (2007) beschrieben, kontrastive Unterschiede bei der Zeitenverwendung, die für Wechsel von grammatischen Funktionen ähnlich wie bei (25) ursächlich sein können. Die Verneinung mittels eines Hilfsverbs wie in Beispiel (26) ist ebenso ein typischer kontrastiver Unterschied zwischen dem Englischen und dem Deutschen, den König und Gast (ebd.) aufführen.

Ein weiterer Grund für Funktionswechsel ist eine nicht selten angewandte Strategie bei der Übersetzung von nicht-kanonischen Wortstellungen im

Deutschen, wenn also statt eines Subjekts ein Objekt im Vorfeld des deutschen Satzes steht, wie in Beispiel (27):

(27) a. *[Die Frauen]$_{DOBJ}$ hat das nicht gerade zimperlich ge macht.* (GO_FICTION_007)
b. *[The women]$_{SUBJ}$ weren't exactly prudes*. (ETrans_FICTION_007)

Der Übersetzer behält an dieser Stelle die Wortfolge bei: *women* steht, so wie im Deutschen *Frauen*, in der ersten Satzkonstituente. Da im Englischen aber die Abfolge der Satzfunktionen sehr viel rigider ist als im Deutschen, und das Subjekt als erstes steht, wird dies ausgeglichen, indem die ganze Satzstruktur verändert wird: Aus *hat...gemacht* wird im Englischen das Verb *weren't*. Die Anpassung an die AS-Wortfolge und die damit verbundenen Funktionswechsel machen es also nötig, dass der Valenzträger des Satzes akkommodiert wird, in ein anderes Verb (*be* statt *make*) und eine andere Diathese (Passiv statt Aktiv). Hawkins (1986) beschreibt diesen kontrastiven Unterschied ebenfalls und zeigt für eine Reihe weiterer Fälle, dass das Englische die geringere Flexibilität der Wortstellung durch flexiblere semantische Belegung der Subjektfunktion ausgleicht. So ist es im Englischen möglich zu sagen *This hotel forbids dogs*, wo im Deutschen *In diesem Hotel sind Hunde verboten* passender ist als *Dieses Hotel verbietet Hunde*, da *Dieses Hotel* nicht typischerweise als (agentivisches) Subjekt erscheint.

Der Subjekt-Objekt-Wechsel gehört nach Tabelle 4.1 in allen Registern zu den häufigsten, und heuristisch gesehen scheint die Wortstellung eine Hauptursache zu sein. Dieses Phänomen ist in seiner Häufigkeit allerdings auch vom Register beeinflusst. Solche Funktionswechsel vom direkten Objekt zum Subjekt sind in der Übersetzungsrichtung Deutsch nach Englisch in SHARE und SPEECH offensichtlich seltener als in FICTION. Dies scheint durch die verstärkt kanonische Wortfolge in SHARE und SPEECH bedingt zu sein. Zählt man die Anzahl der Subjekte in Vorfeldposition in deutschen Originalen, so ergibt dies für FICTION 42,16%; für SHARE sind dies 45,87% und für SPEECH sogar 54,45%. Der Übersetzer hat also in den Registern SHARE und SPEECH weniger mit nicht-kanonischen Wortfolgen zu kämpfen als in FICTION, und die an Beispiel (27) erläuterte Strategie kommt dadurch wohl seltener zum Einsatz. Für diese Hypothese spricht auch, dass bei kanonischer Wortfolge im Hauptsatz selbst solche Beispiele wie der deutsche Satz in (28) ohne Funktionswechsel übersetzt werden.[36]

36 Der an dieser Stelle nur hypothetisierte Prozess, für den im Endprodukt – dem übersetzten Text – nur Indizien existieren, kann nur im Rahmen einer prozessbasierten Übersetzungsforschung belegt oder widerlegt werden.

(28) a. *[Wenn wir also in diesem Sinne unseren Interessen und Werten dienen wollen, dann]$_{ADV\_CAUSE}$ [muss]$_{FIN}$ [Euro pa]$_{SUBJ}$ [erstens]$_{ADV\_OTHER}$ [wachsam gegenüber den neu en Bedrohungen]$_{COMPL}$ [sein]$_{PRED}$, [denen die freien und offenen Gesellschaften ausgesetzt [sind]$_{COMPL}$* (GO_SPEECH_010)

b. *[So if we want to serve our interests and values in line with this definition]$_{ADV\_CAUSE}$, [Europe]$_{SUBJ}$ [must]$_{FIN}$: [firstly]$_{ADV\_OTHER}$, [be]$_{PRED}$ [vigilant to the new threats to which the free and open societies are exposed]$_{COMPL}$.* (ETrans_SPEECH_010)

Zwar ist im Englischen das Subjekt dem finiten Verb vorangestellt (wie üblich), während es im Deutschen erst direkt hinter dem finiten Verb steht, doch ergibt sich auch durch solche geringfügigen Anpassungen der Funktionsabfolge kein Funktionswechsel, wie im vorangehenden Fall, in dem ein direktes Objekt am Satzanfang im Deutschen im Englischen als Subjekt wiedergegeben wird (evtl. sogar werden muss).

Bei Funktionswechseln aufgrund von nicht-kanonischer Wortstellung kommen also verschiedene Faktoren ins Spiel. Die Flexibilität des Deutschen bezüglich der Wortstellung ist ein kontrastiver Unterschied. Da Übersetzter sich aber in der zuvor beschriebenen Weise anpassen, sind Funktionswanderungen teilweise auch ein Ergebnis des Übersetzungsprozesses. Und nicht zuletzt, wie im vorhergehenden Absatz beschrieben, hat das Register auf die Häufigkeit solcher Wechsel einen Einfluss.

Eine anderes Phänomen, das die Satzstruktur und damit die Funktion eines Lexems bzw. einer Konstituente beeinflusst, ist, dass englische parataktische Konstruktionen im deutschen häufig mit hypotaktischen Konstruktionen wiedergegeben werden, wie in Beispiel (29).

(29) a. *Every country has its own political issues and [this]$_{SUBJ}$ [makes]$_{FIN}$ [resolution of our disputes]$_{DOBJ}$ [increasingly difficult]$_{COMPL}$.* (EO_SPEECH_009)

b. *Jedes Land hat seine eigenen politischen Anliegen, [wodurch die Streitschlichtung zunehmend erschwert wird]$_{ADV\_CAUSE}$.* (GTrans_SPEECH_009)

Der Hauptsatz *and this...difficult* wird im Deutschen mit einem *wodurch*-Nebensatz wiedergegeben. Folgend der CroCo-Annotation ist der gesamte Nebensatz als kausales Adverbial analysiert. Gerade aufgrund der flachen Cro-

Co-Annotation ergeben sich in der automatischen Abfrage mehrere Fälle von Funktionswanderungen. Selbst im Falle einer tiefen Annotation von Funktionen wäre auch dieser Fall als Divergenz interessant, da durch die parataktische Anordnung im Deutschen im Gegensatz zur hypotaktischen Ordnung im Englischen Funktionen von der Oberfläche in eine tiefer verzweigte Struktur bewegt werden.

Die Untersuchung von crossing lines zwischen grammatischen Funktionen und Wörtern fördert im Gegensatz zur Untersuchung von empty links für grammatische Funktion eher grammatisch-kontrastive Gegensätze zu Tage. Der häufige Wechsel zwischen direktem Objekt und Präpositionalobjekt ist bedingt durch unterschiedliche Ausprägungen von Valenzrahmen im Deutschen und Englischen. Wechsel zwischen den Funktionen PRED und FIN sind im Falle von Verneinungen durch unterschiedliche Strategien der Verneinung (im Englischen mittels Hilfsverb) verursacht. Der Wechsel zwischen (direktem) Objekt und Subjekt unterliegt mehreren Faktoren: zum einen einer Mischung aus Kontrasten in der Flexibilität der Wortstellung und einer Entscheidung des Übersetzers, die Wortstellung in der Abfolge (aber nicht in der Funktion) im Englischen beizubehalten; zum anderen dem Register, das die Häufigkeit dieses Wechsels beeinflusst (er erscheint am häufigsten in SHARE, am seltensten in FICTION). Die häufigere Verwendung von parataktischen Konstruktionen im Englischen als im Deutschen, wo hypotaktische Konstruktionen üblicher sind, lässt sich allerdings eher als Ursache mit kontrastivem Hintergrund einordnen.

4.3.1.5 Übersicht

In den vorangegangenen Abschnitten wurden empty links und crossing lines nach valenzrelevanten Kategorien untersucht und Ursachen für Divergenzen zwischen dem Deutschen und dem Englischen jeweils linguistischen oder translatorischen Dimensionen zugeordnet. Tabelle 4.2 gibt einen Überblick über die Ergebnisse.

Bsp.Nr.	**Phänomen**	**Ursache**	**Bezug zur Valenz**
Empty links bei Sätzen			
	Wegfall von emotiven Sätzen	kommunikative Präferenzen	-
Empty links bei Einzelsätzen			
(6)	Einleitender Einzelsatz statt Adverb	Register	Zusätzlicher Valenzträger
(7), (8)	Einbettung mit Partizip als Einzelsatz in EN aufgelöst	Sprachkontrast	Valenzalternation
(9)	EN: Einzelsatz vs. DE: Nominalisierung	Sprachkontrast	Valenzvererbung
Empty links bei grammatischen Funktionen			
(10), (11)	EN: Apposition vs. DE: Hauptsatz	Informationsstruktur, komm. Präferenzen	Zusätzlicher Valenzträger im DE
(12)	Empty link bei COMPL in EN bei *name*, *make*, *elect*	Sprachkontrast	typischer Valenzunterschied ganzer Verbklassen
(13), (14)	Empty link bei COMPL in EN bei *be*	Register	nur indirekt: registertypische Lexik
(15)	Empty link bei PRED in EN wegen unterschiedlicher Zeitenverwendungen	Sprachkontrast	syntaktisch unterschiedlich komplexe Valenzträger
(16), (17)	Passiv statt Aktiv	kommunikative Präferenzen (DE: inhaltsorientiert vs. EN: adressatenorientiert)	Valenzalternation
(18), (19)	Empty link für ADV_MOD in DE	Explizierung	Zusätzlicher Modifikator
(20)	Empty link bei Adverb wegen Raising vs. Adverb	Sprachkontrast	Zusätzlicher Valenzträger
Crossing lines bei grammatischen Funktionen			
(22), (23), (24)	DE: PROBJ vs. EN: DOBJ	Sprachkontrast	typischer Valenzunterschied ganzer Verbklassen
(25)	EN: PRED vs. DE: FIN wegen unterschiedlicher Verwendung von Perfekt und Präsens	Sprachkontrast	syntaktisch unterschiedlich komplexe Valenzträger
(26)	DE: FIN vs. EN: PRED wegen Negation mit Hilfsverb in EN	Sprachkontrast	syntaktisch unterschiedlich komplexe Valenzträger

Bsp.Nr.	Phänomen	Ursache	Bezug zur Valenz
(27)	EN: SUBJ vs. DE: DOBJ zur Kompensation nicht-kanonischer Wortstellungen in DE	Sprachkontrast, Register	Wechsel des Valenzträgers zur Akkomodation der unterschiedlichen semantischen Inhalte der grammatischen Funktionen
(29)	EN: Parataxe vs. DE: Hypotaxe	Sprachkontrast	zusätzlicher Modifikator (= der Nebensatz in DE)

Tabelle 4.2: Übersicht über Divergenzen, Ursachen und den Bezug zur Valenz mit Berücksichtigung des Englischen (EN) und des Deutschen (DE)

Zielsetzung dieser Studie war es, die Parallelitätsannahme zu überprüfen, die der MÜ, der Grammatikinduktion oder der Annotationsprojektion als Arbeitsgrundlage dient.

Die kommunikative Einheit Satz findet meist eine Entsprechung (vgl. Abschnitt 4.3.1.1). Dies ist in den oben genannten Anwendungen reflektiert, die meist satzweise – dies wohl zurecht – vorgehen. Für viele syntaktische Einheiten auf den Ebenen der Einzelsätze und der grammatischen Funktionen gibt es keine genauen Entsprechungen (vgl. Abschnitte 4.3.1.2 und 4.3.1.3). Wie aber festgestellt werden kann, haben diese Einheiten meist eine inhaltliche Entsprechung unterschiedlicher syntaktischer Form oder grammatischer Funktion: Einzelsätze werden im Deutschen häufig nominalisiert, grammatische Funktionen wechseln zwischen den Sprachen. Damit lässt sich auch erklären, warum Padós Ergebnisse bei der Annotationsprojektion mittels Frame-Semantik erfolgreich ist: die framesemantische Annotation abstrahiert von syntaktischer Variation weg.

Für jede gefundene Divergenz ist in Tabelle 4.2 zusätzlich angegeben, welchen Effekt sie auf der Ebene der syntaktischen Valenz hat.

4.3.2 Divergenzen bei Prädikatsausdrücken

Im vorangehenden Abschnitt wurden Divergenzen bei den valenzrelevanten Kategorien Satz, Einzelsatz und grammatische Funktion untersucht. Sätze und Einzelsätze können als Extraktionsgrundlage für alignierte Valenzrealisierungsmuster dienen, grammatische Funktionen dienen als Beschreibungskategorien für Elemente im Satz, die Verbbestandteile sowie die Begleiter des Verbs, die von strukturellen Unterschieden abstrahieren.

Während ein struktureller Unterschied wie z.B. der Wechsel von einem direkten Objekt zu einem Präpositionalobjekt technisch einfach zu beschreiben ist und in der MÜ kein Problem darstellt, zeigt das Extraktionsbeispiel in 4.2.2, dass der Hauptanker bei der Valenzextraktion, nämlich der Valenzträger, der als Grundlage für einen Eintrag in einem Valenzwörterbuch dient, nicht immer eindeutig zu erkennen ist und zwischen den Sprachen ebenfalls divergieren kann.

Im Folgenden soll daher mit Bezug auf die Valenzextraktion untersucht werden, welche Divergenzen sich bei Valenzträgern – oder genauer gesagt **Prädikatsausdrücken**, die als solche im Folgenden noch eingeführt werden – ergeben können. Dabei besteht kein Anspruch auf Vollständigkeit, sondern es wird auf die am häufigsten auftretenden Phänomene in einem Testdatensatz mit geringem Umfang eingegangen.

Für die Untersuchung in diesem Abschnitt wurden 300 Satzpaare herangezogen, die, wie in Abschnitt 4.2.2 beschrieben, extrahiert wurden. Sie wurden manuell daraufhin untersucht, inwiefern eine VRM-Extraktion sinnvoll erscheint. Basis war die Ausgabe nach dem Muster wie in Abbildung 4.5. Die 300 Satzpaare setzen sich wie folgt zusammen. Aus den beiden Übersetzungsrichtungen G2E und E2G wurden für die Register FICTION, SPEECH und SHARE dieselben Register wie in den Vorstudien, jeweils 50 Paare zufällig aus der Menge aller Paare des Registers ausgewählt[37]. D.h. in der Evaluierungsmenge sind jeweils 50 Satzpaare aus E2G_FICTION, E2G_SHARE und E2G_SPEECH sowie G2E_FICTION, G2E_SHARE und G2E_SPEECH enthalten.

Methodisch wurde bei der Analyse und Auszählung der Divergenzen ein interaktiver Modus auf Basis einer korpusgestützten Extraktion simuliert. Das zu Grunde liegende Bild war Folgendes. Einem Linguisten (dem Autor dieser Arbeit) werden von einem VRM-Klassifikationsprogramm eine Liste von VRM vorgeschlagen. Für jedes Element entscheidet der Linguist aus einer der folgenden Kategorisierungsmöglichkeiten.

(1) Es liegt ein grober technischer Fehler vor (abgehackte Sätze, kryptische Zeichen in der Ausgabe o.ä.). Das Satzpaar wird nicht ausgewertet und ggf. per Mausklick der Fehler an eine zuständige Stelle gemeldet.

37 Die Randomisierung war Teil des Extraktionsalgorithmus und ist – im Gegensatz zu einer manuellen Zufallsauswahl – daher nicht anfällig für äußere Einflussfaktoren wie z.B. Satzlänge oder Lesbarkeit der Beispiele.

(2) Das Satzpaar wird nicht ausgewertet, weil ein grober Annotationsfehler vorliegt, z.B. wurde eine Konstituente bei der Annotation völlig vergessen (Meldung wieder ggf. per Mausklick). Ein geringfügiger Annotationsfehler liegt dagegen vor, wenn z.B. bei der Annotation einer Phrase ein Satzzeichen wie das Komma noch mit annotiert wurde.
(3) Das Satzpaar wird nicht ausgewertet, weil in mindestens einem der beiden Sätze zwei oder mehr Hauptsätze enthalten sind. Durch die flache Annotation in CroCo sind die beiden Hauptsätze nicht klar voneinander trennbar und damit auf automatischem Wege keine verlässlichen VRM extrahierbar.
(4) Das Satzpaar kann aufgrund zu großer syntaktischer oder semantischer Divergenzen zwischen AS- und ZS-Satz nicht zur VRM-Extraktion herangezogen werden.
(5) Das Satzpaar ist korrekt aligniert, annotiert und ausgegeben, und kann somit zur Extraktion herangezogen werden. Ggf. vorkommende Divergenzen zwischen den Valenzträgern des Satzes, auf die später noch näher eingegangen wird, werden ebenfalls gezählt.

Bei der hier vorliegenden Auszählung wurde die Kategorie nicht berücksichtigt, da sie für die Fragestellung der vorliegenden Arbeit nicht relevant ist. Aus den Kategorien (2) – (4) sowie einer weiteren Unterteilung von (5) entstand das Auswertungsschema, auf das später noch genauer eingegangen wird.

Die Art der Auswertung wurde dabei so gestaltet, dass sie sich in den Charakter der Vorstudien einreiht. Das heißt

- es werden strukturelle Divergenzen untersucht und zwecks Vollständigkeit auch Fälle markiert, in denen die beiden Hauptverben semantisch nicht deckungsgleich sind, die Divergenz (z.B. Bedeutungsverengung oder -erweiterung) aber nicht näher untersucht;
- es wurde eine quantitative Auswertung durchgeführt, die anstelle einer detaillierten Einzeluntersuchung ein Gesamtbild vermittelt.

Dass die Extraktion von VRM, also den ein Verb begleitenden grammatischen Funktionen sowie deren Kategorien, möglich ist, wurde bereits in Abschnitt 4.2.2 gezeigt. Inwiefern die korpusextrahierten VRM einen neuen Blickwinkel auf Valenz eröffnen können, theoretisch wie praktisch, wird in Abschnitt 5 diskutiert.

Im Fokus der Untersuchung steht das, was die Grammatik des Instituts für Deutsche Sprache (Zifonun et al. 1997: 699ff) als **Prädikatsausdruck** (kurz:

PKA) bezeichnet: die Konstituente(n) im Satz, mittels derer ein semantisches Prädikat syntaktisch realisiert wird. Dabei bleibt die Beschränkung weiterhin auf Konstruktionen mit einem verbalen Zentralbestandteil. Der Ausdruck PKA ersetzt das, was bisher etwas unspezifisch als „Valenzträger" bezeichnet wurde.

Laut der IDS-Klassifikation gibt es zwei Arten von PKA, zum einen die einfachen PKA, dazu gehören:

- das Vollverb,
- Verbalperiphrasen (Passiv, zusammengesetzte Zeiten),
- Kopulakonstruktionen (*neu sein*, *groß erscheinen*)
- Nominalisierungsverbgefüge (*einen Besuch machen*),
- und Funktionsverbgefüge (*in Gefahr bringen*);

zum anderen die komplexen PKA, dazu gehören:

- *lassen*-Konstruktionen (*Er ließ sich das Buch bringen*)
- modalisierte Prädikate (*mögen*, *wollen*)
- spezifizierte Prädikate (*drang ... ein* vs. *drang lautlos ... ein*)

Im Folgenden wird aufgeführt, welche PKA und welche PKA-Divergenzen in den analysierten 300 Satzpaaren vorkommen und inwiefern diese bei der quantitativen Auszählung berücksichtigt wurden. Die gezeigten Beispiele gehören zu den typischsten Vertreter der jeweiligen PKA. Dass es eine Reihe von Fällen gibt, die zwar mit ausgewertet wurden, deren Zuweisung aber nicht zwingend eindeutig ist, soll an dieser Stelle nicht verschwiegen werden. Gerade für unklare Fälle (z.B. ob ein Kollokat wie *Umsätze belaufen sich auf* schon als idiomatische Wendung von Wirtschaftstexten gelten kann) wäre es natürlich hilfreich, die Klassifikation durch einen Zweit- oder Korrekturannotierer zu überprüfen.

Die Darstellungsform der Beispiele entspricht dem, was in Abbildung 4.5 zu sehen ist. Im ersten Beispiel haben sowohl AS- als auch ZS-Satz als PKA ein Vollverb:

(30) *"verlassen": subj#np, dobj#np, probj#clause, probj#clause; "assume": subj#np, part#part, dobj#clause,*
Er hat sich darauf verlassen , dass wir von drinnen sein Laecheln sehen koennen . (GO_FICTION_001)

He just assumed we could see his smile from inside .
(Etrans_FICTION_007)

Die Verben *sich verlassen* und *assume* drücken unterschiedliche Aspekte aus: *sich verlassen* beinhaltet dabei einen höheren Grad an Sicherheit, dass das erwartete Ereignis eintritt. Warum der Übersetzer in der ZS einen weniger „sicheren" Ausdruck gewählt hat, kann an dieser Stelle nicht beantwortet werden. Der Leser des ZS-Textes nimmt dadurch einen anderen Blickwinkel ein als der Leser des AS-Textes; daher wird auf dieses Phänomen als **Perspektivwechsel** (PW) Bezug genommen. Die Benennung lehnt sich an die Formulierung „changement du point du vue" in (Vinay & Darbelnet 1958: 50) an. Als Perspektivwechsel werden prinzipiell alle Vorkommen klassifiziert, die semantische Abweichungen zwischen Vollverben zwischen Original und Übersetzung aufweisen. Eine genauere Einordnung der semantischen Divergenz, z.B. in Erweiterung oder Verengung der Bedeutung, ist nicht Gegenstand der hier beschriebenen Analyse.

Ein Phänomen, das bereits im Zusammenhang mit verstärktem Auftreten von COMPL-Funktionen in SHARE beschrieben wurde, ist, dass in englischen Texten Kopulakonstruktionen häufiger verwendet werden als im Deutschen, wo dann Vollverben im Zentrum des Satzes stehen, wie in folgendem Beispiel:

(31) *"setzen": subj#np, dobj#np, adv_loc#pp, adv_mod#advp, ; "become": subj#np, compl#adjp, adv_loc#pp,*
Ostdeutsche Produkte setzen sich im harten Wettbewerb auf dem Weltmarkt immer mehr durch . (GO_SPEECH_015)
Eastern German products are becoming more and more competitive on the global market . (ETrans_SPEECH_015)

Interessant ist dieses Phänomen genau in dieser Konstellation, also Vollverb gegenüber Kopulaverb, weil sich dann eine Funktionsverschiebung ergibt. In (31) bindet das deutsche Hauptverb *durchsetzen* das Modaladverbial *immer mehr*, als englische Entsprechung kann man weitestgehend das Komplement *more and more competitive* ansehen, was mehr als nur eine Funktionsverschiebung, aber durch den Wechsel des Prädikats bedingt ist. Zudem verändert sich dadurch die Struktur des Satzes. In Beispiel (13) auf Seite 93 führt die Verwendung von *be* + COMPL dazu, dass der Nebensatz im Englischen von *pleased* abhängt und damit tiefer eingebettet wird als im Deutschen. Fälle, in denen sowohl im AS- als auch im ZS-Satz ein Kopulaverb mit Komple-

mentfunktion verwendet wird, wurden in der vorliegenden Untersuchung nicht berücksichtigt, da sich in solchen Fällen parallele Strukturen ergeben. Eine weitere PKA-Form, die in beiden Sprachen existiert, ist das Funktionsverbgefüge (FVG) wie in Beispiel (32):

(32) *"leisten": adv_mod#advp, subj#np, dobj#np, ; "making": subj#np, adv_mod#advp, dobj#np,*
Damit leisten wir einen wichtigen Beitrag für mehr Wachstum und Beschäftigung . (GO_SPEECH_002)
We are thus making an important contribution towards stepping up growth and employment . (ETrans_SPEECH_002).

FVG sind interessant, sowohl wenn sie nur auf einer Seite als auch wenn sie in AS- und ZS-Satz vorkommen. Steht einem Vollverb in einer Sprache ein FVG in der anderen Sprache gegenüber, sollte sich daraus ein Unterschied in den Argumentzahlen ergeben. So gibt es z.B. im deutschen *in Gefahr bringen* gegenüber dem englischen *endanger* syntaktisch gesehen ein Element mehr, nämlich *in Gefahr*. Semantisch ergibt sich tatsächlich kein Unterschied, da sie beide auf nur ein einfaches Prädikat abbildbar sind. FVG sind aber auch dann interessant, wenn sie in beiden Sprachen mit gleicher Elementzahl und -art stehen, wie in (32). Vergleicht man die beiden finiten Verben der Sätze, *make* und *leisten*, so erscheinen sie im Wörterbuch nicht typischerweise als Übersetzung voneinander, außer eben in bestimmten Kollokationen bzw. FVG. Darum lohnt es sich – anders als bei Komplementkonstruktionen – auch Fälle zu betrachten, in denen sich zwei FVG gegenüberstehen.[38]

Die dritte Art der hier ausgewerteten PKA neben Kopula und Funktionsverbgefügen, der über ein Vollverb hinausgeht, ist der Phraseologismus (PH), also mehr oder weniger feststehende Wendungen aus zwei oder mehr Elementen, die als großes Ganzes einen ganz anderen Sinnen ausdrücken (können) als die Summe ihrer Teile (z.B. *it's raining cats and dogs* – es mag sehr stark regnen, aber mit Sicherheit keine Katzen und Hunde). Beispiele für Phraseologismen sind:

(33) *"lay": subj#np, dobj#np, adv_other#pp, conj#conj, adv_mod#pp; "strafen": subj#np, dobj#np, compl#np, conj#conj, adv_mod#pp*
The support of Turkey , a loyal friend and ally , lays this myth to rest and […] (EO_SPEECH_008)

38 In der zweiten CroCo-Phase wurden FVG ebenfalls annotiert, aber auch hier war die Annotation zum Untersuchungszeitpunkt noch nicht in nennenswertem Umfang verfügbar.

Die Unterstützung der Türkei , eines loyalen Freundes und Bündnispartners , straft diesen Mythos Lügen und […]
(GTrans_SPEECH_008)

(34) *"be": adv_cause#clause, subj#np, compl#np; "belaufen": adv_cause#pp, dobj#np, subj#np, adv_mod#advp, probj#pp, When European sales are included , our global coverage will be more than $ 3 billion .* (EO_SHARE_006)
Unter Einbeziehung von Europa belaufen sich unsere weltweiten Umsätze insgesamt auf über 3 Milliarden US-Dollar .
(GTrans_SHARE_006)

Als Phraseologismus sollen auch teilidiomatische Wendungen (Palm 1997), also in Wirtschaftstexten solche Wendungen wie *Umsätze belaufen sich auf* wie aus Beispiel (34) oder *Dividende ausschütten*, die im jeweiligen Gebrauchskontext als Standardformulierung gelten können und als gesamte Bedeutungseinheit ebenfalls eine bestimmte Satzstruktur bedingen.[39]

PH wurden, ebenso wie FVG, unabhängig davon berücksichtigt, ob sie alleine oder auf beiden Seiten der Übersetzung vorkommen. Wie in (33) zu sehen, sind zwei sich gegenüberstehende, wenn auch gleichbedeutende, PH nicht unbedingt strukturell gleich: Im Englischen ergibt sich die Konstruktion $y + NP + PP_{to}$, im Deutschen *strafen + NP + NP*. Auch die beiden Hauptverben *lay* und *strafen* können nicht als äquivalent gelten.

Komplexe PKA werden in dieser Auswertung nicht behandelt. Modalverben erscheinen in der CroCo-Annotation als FIN und werden in der Auswertung als Hauptverben behandelt. Spezifikationen von Prädikaten werden als Prädikatsteil ignoriert. Sie sind in CroCo als ADV_MOD annotiert und werden in diesem Kontext als grammatische Funktionen angesehen. Die einzige „lassen"-Konstruktion in den 300 ausgewerteten Satzpaaren wurde der Einfachheit halber unberücksichtigt gelassen.

Aus den in den vorangehenden Absätzen beschriebenen Grundüberlegungen und Beobachtungen ergab sich ein Auswertungsschema mit folgenden Kategorien, von denen pro Satzpaar genau eine vergeben wird. Zu jeder Kategorie

39 Schlägt man die in der Auswertung als PH markierten Beispiele unter http://wortschatz.uni-leipzig.de nach, erscheinen sie als „signifikante Kollokate". Im Englischen Internetkorpus der Uni Leeds unter http://corpus1.leeds.ac.uk/ erscheinen die englischen Beispiele als Kollokate mit hohen Punktzahlen.

wird in Klammern zusätzlich angegeben, welcher Kategorisierungsmöglichkeit, die am Beginn dieses Abschnitts auf Seite 107 beschrieben wurden, sie entspringt:

OK – Die Prädikatsausdrücke der AS- und ZS-Sätze haben nur Verbbestandteile, ein direkter Verb-zu-Verb-Wörterbucheintrag mit den dazugehörigen VRM kann übernommen werden. Zusätzlich wurden ergänzungsweise jene Verbpaare mit **PW** markiert, in denen die beiden sich gegenüberstehenden Hauptverben nicht als direkte Entsprechung gewertet werden können (wie in Beispiel (30)), bei denen sich daher ein Perspektivwechsel ergibt. (vgl. Kategorisierungsmöglichkeit (5))

KP – Einem Vollverb im AS-Satz steht eine Kopulakonstruktion als Prädikatsausdruck im ZS-Satz gegenüber oder umgekehrt. (vgl. Kategorisierungsmöglichkeit (5))

FVG – Auf mindestens einer Seite steht ein Funktionsverbgefüge als Prädikatsausdruck. (vgl. (5))

PH – Auf mindestens einer Seite steht ein Phraseologismus als Prädikatsausdruck. (vgl. (5))

2HS – Auf mindestens einer Seite steht ein Satz, der sich aus zwei oder mehr Hauptsätzen zusammensetzt. (vgl. (3))

AF – grober Annotationsfehler in einem der beiden Sätze (vgl. (2))

UK – zu große syntaktische oder semantische Divergenzen, um das Satzpaar zur Extraktion heranzuziehen (vgl. (4))

Die Tabelle 4.3 führt die Auszählungsergebnisse auf. Pro Kombination aus Übersetzungsrichtung und Register wurden 50 Satzpaare ausgezählt; nur in E2G_SPEECH ergibt sich ein Gesamtergebnis von 51, zählt man alle Zahlen zusammen. Dies liegt daran, dass in einem Satzpaar gleich zwei Phänomene gleichzeitig auftraten: Einem Funktionsverbgefüge stand in diesem Fall eine Kopulakonstruktion gegenüber. Die Fehler wurden mit ausgezählt, um Hinweise für mögliche bzw. notwendige Verbesserungen an der Annotation zu erhalten.

Neben den Perspektivwechseln (PW) bei OK ist bei den Kategorien KP, PH und FVG aufgeführt, wie häufig jedes der Phänomene auf der ZS- bzw. AS-Seite auftrat. Da das Phänomen durchaus auf beiden Seiten gleichzeitig auftreten kann (außer, wie erwähnt, bei KP), ergibt die Summe aus der AS- und der ZS-Spalte bei FVG und PH mitunter einen höheren Wert als die Gesamtzahl der Auftretenshäufigkeit eines Phänomens, die sich auf Satzpaare und

nicht auf einzelne Sätze bezieht. Bei der Spalte KP kann dies nicht der Fall sein, da hier nur Fälle erfasst werden, bei denen eine Kopulakonstruktion auf der einen Seite einem Vollverb auf der anderen Seite gegenübersteht.

	OK	**KP**	**FVG**	**PH**	**2HS**	**AF**	**UK**
E2G_FICTION	22 *PW: 6*	2 *AS: 2* *ZS: -*	-	2 *AS: 1* *ZS: 1*	21	2	1
E2G_SHARE	23 *PW: 6*	5 *AS: 5* *ZS: -*	-	3 *AS: 2* *ZS: 2*	14	3	2
E2G_SPEECH	24 *PW: 3*	4 *AS: 4* *ZS: -*	5 *AS: 3* *ZS: 4*	5 *AS: 5* *ZS: 4*	10	3	-
G2E_FICTION	24 *PW: 4*	4 *AS: 1* *ZS: 3*	1 *AS: -* *ZS: 1*	1 *AS: 1* *ZS: 1*	17	3	-
G2E_SHARE	21 *PW: 6*	6 *AS: -* *ZS: 6*	-	5 *AS: 2* *ZS: 3*	11	7	-
G2E_SPEECH	27 *PW: 8*	5 *AS: -* *ZS: 5*	3 *AS: 2* *ZS: 1*	-	11	3	1
Gesamt	141 *PW: 33*	26 *AS: 12* *ZS: 9*	9 *AS: 5* *ZS: 6*	15 *AS: 11* *ZS: 11*	84	21	4

Tabelle 4.3: Auszählungsergebnis für Satzpaarkategorien mit Bezug auf die Vergleichbarkeit ihrer Prädikatsausdrücke

Die qualitative Auswertung wurde anhand einiger Einzelfallbeispiele bereits zu Beginn dieses Abschnittes vorgenommen, hier soll aber auf einige augenfällige Erscheinungen in den Gesamtzahlen eingegangen werden. Dabei muss nochmals darauf hingewiesen werden, dass 300 Satzpaare keine statistisch belastbare Grundmenge sind, aber eine Pilotstudie darstellen, die Aspekte für

weitere Forschung herausbilden soll. Im Folgenden soll daher auf einige auffällige Werte in den einzelnen Kategorien eingegangen werden.

Auffällig ist zunächst die hohe Zahl von **AF** in G2E_SHARE, was daran liegen dürfte, dass SHARE als erstes Register annotiert wurde und eine höhere Anzahl von Fehlern damit erwartbar ist. Da zur Zeit der Auswertung noch keine Konsistenzüberprüfung durchgeführt wurde, sind diese Fehler in der vorliegenden Auswertung enthalten. Erst im Rahmen der zweiten Phase des CroCo-Projekts wurden Konsistenzprüfungen vorgenommen, die zur Zeit der Erstellung dieser Arbeit noch laufen.

Weiterhin ist es beruhigend, dass die Spalte **UK** insgesamt keine nennenswerte Rolle spielt. Das bedeutet, dass die meisten Übersetzungen in meinem Sinne ausgewertet werden können. Man kann daraus auch den Rückschluss ziehen, dass parallele Korpora sich prinzipiell für die Art von Auswertung eignen, wie sie hier vorgenommen wurden. Dieser allgemeinen Fragestellung nach der Eignung von parallelen Korpora zur Extraktion widmet sich auch in anderem Blickwinkel nochmal die Diskussion.

Wie der Tabelle entnommen werden kann, ist der Anteil an **2HS**-Fällen in jedem Register und jeder Übersetzungsrichtung jeweils mindestens 20%. Dies unterstreicht nochmals die Notwendigkeit, die CroCo-Annotation in eine tiefere Annotation bis zu den Terminalen auszubauen.

	OK+KP+FVG +PH	OK-PW	Anteil OK in %
E2G_FICTION	24	16	66,67
E2G_SHARE	31	13	41,94
E2G_SPEECH	28	21	55,26
G2E_FICTION	30	20	66,67
G2E_SHARE	32	15	46,88
G2E_SPEECH	35	19	54,29

Tabelle 4.4: Anteil an direkten Verb-zu-Verb-Entsprechungen ohne Perspektivwechsel

Zudem fällt auf, dass nach Abzug aller Phänomene außer **OK** sowie der mit **PW** markierten OK-Fälle nur zwischen 40% und 67% übrig bleibt, in der von einer direkten Verb-zu-Verb-Entsprechung gesprochen werden kann, die

sich für eine unproblematische Wörterbuchextraktion eignen würden. Die genauen Zahlen sind in Tabelle 4.4 aufgeführt.

Ausschlaggebend für den Anteil an Verb-zu-Verb-Entsprechungen ist dabei nicht die Übersetzungsrichtung, sondern das Register. FICTION zeigt den höchsten Anteil an Entsprechung. Darauf folgt in beiden Übersetzungsrichtungen SPEECH. SHARE zeigt den geringsten Anteil an Entsprechungen. Hier sei nochmal auf den Deutungsversuch bezüglich der Kopulakonstruktionen erinnert, die im Englischen häufig teilidiomatischen Wendungen mit Vollverb im Deutschen gegenüberstehen. Dies kann aber nur einen Teil der Nicht-Entsprechungen erklären. Außerdem bleibt die Frage, was die 45-55% Nicht-Entsprechung in SPEECH und FICTION ausmacht. Zudem bleibt zu untersuchen, ob sich diese Zahlen sowie der Einfluss des Registers bei der Auszählung von mehr Satzpaaren so halten lassen.

Wie schon an einigen Beispielen dargelegt, spielen **KP** im Englischen eine deutlich größere Rolle als im Deutschen. Nur in einem Fall stand eine Kopulakonstruktion im Deutschen einem Vollverb im Englischen gegenüber. Diese Zahlen dürften nochmals einen eindrucksvollen Beleg liefern, bedenkt man zusätzlich, dass das Englische eigentlich reich an Verben mit feinsten Nuancierungen ist.

FVG und **PH** spielen in FICTION keine nennenswerte Rolle. In SPEECH tauchen FVG in beiden Übersetzungsrichtungen auf, PH dagegen ausschließlich in E2G_SPEECH (jeweils ins AS und fast immer auch in ZS). Man könnte hypothetisieren, dass FVG gerne dann verwendet werden, wenn die Sprache etwas professioneller klingen soll. In politischen Reden – woraus die Textsammlung in SPEECH besteht – ist dies wohl kein abwegiges Motiv. Interessant ist, dass in den deutschen Originalen scheinbar keine PH auftreten, im Englischen dafür nicht selten, wie *lay ... to rest* aus Beispiel (33). Bei aller Zufallsauswahl der ausgezählten Satzpaare kann dies natürlich an einem oder zwei Sprechern liegen, könnte aber auch Hinweis für ein typisch englisches Stilmittel sein. Dass PH in SHARE des Öfteren auftauchen, liegt in der Methodik der Auswertung begründet. Wie zu Beginn dieses Abschnitts erläutert, sind Wendungen, die als typisch für die Wirtschaftssprache erschienen[40], explizit als PH markiert.

40 Um die Objektivität zu gewährleisten, wurden nur jeweils die Kollokate entsprechend gezählt, die auf Abfrage auf der Internetseite des Wortschatzprojekts der Univesität Leipzig unter http://wortschatz.uni-leipzig.de als „signifikante Kookkurenzen“ eingestuft wurden. Dies war das Kriterium für die hier vorgenommene Klassifizierung, mag aber in Zukunft strengeren Kriterien weichen.

Die Analyse von PKA-Divergenzen hat v.a. gezeigt, dass diese in nicht vernachlässigbarem Maße vorkommen. Zwar besitzen die in 2.5 beschriebenen MÜ-Systeme häufig bereits Mechanismen, um solche Divergenzen zu behandeln, allerdings beziehen sich diese auf bekannte Divergenzen z.B. im Bereich von PH oder FVG oder auf typische valenzbezogene Phänomene wie Argument- oder Kopfwechsel. Ausnahmen sind beispielsweise Verbmobil, das *übernachten* mit *spend the night* übersetzt (vgl. Abschnitt 2.5.5) oder METIS, das sich u.a. speziell mit der korrekten Übersetzung von FVG befasst (Anastasiou und Čulo 2007).

Die hier beschriebenen, häufig vom Register beeinflussten Divergenzen in der Art des PKA sind zum einen für die korrekte Feststellung von Valenzrahmen (z.B. im Falle einer FVG) oder von Entsprechungen zwischen Valenzrahmen (z.B. im Falle von KP vs. Vollverb) relevant. Auch werden kontrastive Ursachen in der hier vorliegenden Analyse wie schon in den Vorstudien stärker hervorgehoben. Registerkontrollierte Korpora ermöglichen es, derartige Erkenntnisse zu gewinnen und sie in die MÜ wie auch in die Fremdsprachen- und Übersetzerausbildung zu integrieren und mit illustrierenden Beispielen zu belegen.

4.3.3 Übersicht

Die Vorstudien sowie die Analyse von PKA-Divergenzen aus dem Extraktionsexperiment haben das Potenzial sowohl korpusgestützter Valenzwörterbuchextraktion sowie der korpusgestützten Untersuchung valenzbezogener, vom Register beeinflusster Übersetzungsphänomene gezeigt. Die Ergebnisse seien an dieser Stelle nochmals im Überblick zusammengefasst.

Paare von alignierten Sätzen, so wurde in Abschnitt 4.3.1.1 dargestellt, bieten sich prinzipiell als verlässliche Basis für die Extraktion bilingualer Wörterbücher an, da die empty link-Zahlen darauf hinweisen, dass Sätze meistens ein Äquivalent in der jeweils anderen Sprache haben, Divergenzen in der internen syntaktisch-semantischen Struktur natürlich vorbehalten. Dass Einzelsätze ebenfalls ein hohes Potenzial für die Valenzwörterbuchextraktion besitzen, zeigt Abschnitt 4.3.1.2. Zum einen besteht ein Einzelsatz aus einer weiteren Prädikat-Argument-Struktur neben der des Hauptsatzes. Zum anderen sind diese Prädikat-Argument-Strukturen gerade im Sprachenpaar Englisch-Deutsch bei der Übersetzung syntaktisch häufig der Valenzvererbung unterworfen und bieten damit die Möglichkeit zur Ausdehnung der Analyse über die reine Verbvalenz hinaus. Die Abschnitte 4.3.1.3 und 4.3.1.4 widmen sich den grammatischen Funktionen, der erstere den empty links, der letztere den crossing lines zwischen Funktionen und Wörtern. Während empty links

tendenziell eher auf stilistische Unterschiede zwischen Registern in den beiden Sprachen in Original und Übersetzung zeigen, decken crossing lines eher kontrastive grammatikalische und valenzbezogene Unterschiede zwischen den Sprachen auf. Wichtig ist, in Zukunft, mit Blick auf eine semantischere Interpretation der Ergebnisse, auch Kategorieinformationen stärker einzubeziehen. Häufig weisen Änderungen der Kategorie eines Arguments auf eine Änderung der Bedeutungsnuance des Prädikats hin. So rückt das deutsche Verb *etw. sehen* mit der syntaktischen Variante eines *dass*-Satzes, also *sehen, dass...* näher an die Bedeutung des Verbs *erkennen*. Abschnitt 4.3.2 behandelt schließlich Divergenzen zwischen Prädikatsausdrücken, die zu erkennen wichtig sind, um bei der Valenzextraktion aus parallelen Daten nicht fehlerhafte Einträge zu generieren, z.B. durch Interpretation eines Kopulakonstruktion als Vollverb, wobei der semantische Gehalt des Vollverbs aus einer Sprache eigentlich tendenziell im Komplement des Kopulaverbs der anderen Sprache enthalten ist.

Da es nicht möglich ist, Analysen wie in 4.3 vollautomatisch durchzuführen, empfiehlt sich für alle Arten von Divergenzstudien ein interaktiver Modus, wie in demselben Abschnitt beschrieben.

Wie gezeigt werden konnte, ist die Parallelitätsannahme mit Bezug auf die Valenz nicht zu halten. Zwar gibt es in der jeweils anderen Sprache meist eine inhaltliche Entsprechung für eine syntaktische Einheit, doch gibt es nicht nur syntaktische Divergenzen zwischen ihnen. Wie die Auflistung in Tabelle 4.2 in Abschnitt 4.3.1.5 zeigt, führen diese Divergenzen zu diversen Veränderungen in der Valenz, von Valenzvererbungen über Alternationen bis hin zur Änderung der Anzahl von Valenzträgern in einem Satz. Zusätzlich hat die Untersuchung zu Divergenzen bei Prädikatsausdrücken gezeigt, dass es neben syntaktischen Divergenzen zwischen den PKA auch semantische Unterschiede gibt, wenn auch meist keine gravierenden. Im folgenden Abschnitt 4.4.1 wird im Rahmen einer Machbarkeitsstudie gezeigt, wie die Erkenntnisse der Studien in ein regelbasiertes MÜ-System integriert werden können.

4.4 Anwendung der Ergebnisse

In diesem Abschnitt wird darauf eingegangen, wie die Ergebnisse aus den Vorstudien und dem Extraktionsexperiment umgesetzt werden können. In Abschnitt 4.4.1 wird gezeigt, wie ein bestehendes, regelbasiertes MÜ-System in seinen Regeln erweitert werden kann, um die Ergebnisse zu inkorporieren. In Abschnitt 4.4.2 werden einige Überlegungen dazu präsentiert, in welcher

Form die extrahierten alignierten Valenzrahmen in Form von Wörterbucheinträgen auf einer dynamischen Webseite aufbereitet werden können.

4.4.1 Erweiterung der Regeln und Transferlexika von MÜ-Systemen

Das hier vorgestellte Pilotexperiment ist als Machbarkeitsstudie zu werten. Dazu wurden einige der Beobachtungen aus den zuvor beschriebenen Studien praktisch umgesetzt, indem ein bestehendes Regelwerk eines MÜ-Systems erweitert wurde.

In der folgenden Beschreibung der Machbarkeitsstudie wird zunächst auf das für die Studie eingesetzte MÜ-System sowie den Versuchsaufbau eingegangen. Danach folgt eine Beschreibung der Regelerweiterungen mit jeweiliger Bezugnahme auf Erkenntnisse aus den vorangegangenen Valenzstudien. Zuletzt werden die Ergebnisse der Machbarkeitsstudie sowie deren Übertragbarkeit auf aktuelle statistische MÜ-Systeme diskutiert.

Bei der Auswahl des MÜ-Systems waren die beiden Hauptkriterien, dass das System frei verfügbar war und dass die Grammatik genügend Eingriff erlaubte, um die in diesem Abschnitt beschriebenen Modifikationen durchzuführen. Zunächst wurde das System OpenLogos[41] getestet, allerdings ergaben sich bei der Installation und im Betrieb kaum überwindbare technische Schwierigkeiten. Die Wahl fiel daher auf das MÜ-System CAT2 (Haller 1993; Sharp 1994), das sich auf dem Windows-Betriebssystem ohne größere Installation betreiben lässt und dessen Grammatik mit einem einfachen Texteditor editiert werden kann.

CAT2 entstand als Seitenlinie von EUROTRA (vgl. Abschnitt 2.5.3), und hat einige der zentralen Charakteristiken mit EUROTRA gemeinsam. In CAT2 werden zur Darstellung von Lexikoneinträgen und Regeln Merkmalstrukturen verwendet. Die Analyse beginnt auf der **morphologischen** Ebene. Die auf dieser Ebene erkannten Elemente werden auf der **syntaktischen Ebene** zusammengesetzt. Diese syntaktischen Strukturen wiederum in eine quasisemantische, dependenzartige **Interfacestruktur** übersetzt, die auch zum Transfer verwendet wird. Regeln können dabei zum Aufbau oder zum Transfer von Strukturen dienen, also z.B. zum Aufbau syntaktischer Strukturen, zum Transfer von syntaktischen Strukturen in Interfacestrukturen oder zum Transfer von einer Sprache in die andere Sprache.

Das CAT2-System unterscheidet sich in einigen zentralen Punkten von EUROTRA: So beinhaltet es diverse Nachverfolgungs- und Robustheitsmechanismen bei der Abarbeitung von Regeln. Regeln und Lexikoneinträge sind

41 http://logos-os.dfki.de/

in hierarchischen statt in flachen Merkmalstrukturen kodiert. Außerdem ist die Zahl der Regeltypen in CAT2 deutlich geringer.

Als Eingabemenge wurde zunächst auf die Sätze zurückgegriffen, die, wie in der Einleitung zu dieser Arbeit beschrieben, bereits zuvor zu Testzwecken mit Google Translate übersetzt worden waren. Da die Grundgrammatik, die dieser Machbarkeitsstudie zugrunde gelegt wurde, allerdings nur eine Übungsgrammatik ist, die nur wenig komplexe syntaktische Phänomene des Englischen und des Deutschen abdeckt, wurden die Sätze der Eingabemenge in der syntaktischen Komplexität reduziert.

Die Beispielsätze aus der Einleitung wurden zu folgenden Eingabesätzen umformuliert:

(35) Diese Tatsache straft diesen Mythos Lügen.

(36) Unser Umsatz beträgt eine Milliarde.

(37) Mich beunruhigt die vom Himmel fallende Asche.

Satz (35) beinhaltet das, was man im weitesten Sinne als FVG interpretieren kann. Die Wendung *Lügen strafen* hat v.a. die syntaktische Eigenschaften eines FVG, wie sie schon Drach (1963) beschrieben hat: Im Hauptsatz ist nur die Klammerstellung erlaubt, mit dem Verb in der linken Klammer und der dazugehörigen NP am Ende des Mittelfelds, also *Die Tatsache straft diesen Mythos Lügen*, aber nicht **Die Tatsache straft Lügen diesen Mythos*. Da die CAT2-Grammatik kein ausreichend ausgearbeitetes System zur Behandlung von topologischen Feldern im Deutschen besitzt, können die positionellen Eigenschaften nicht berücksichtigt werden. Dass diese Eigenschaften für die korrekte Erkennung von FVG im Deutschen zuträglich sind, zeigen Anastasiou und Čulo (2007) anhand der Implementation und Evaluation eines Grammatikfragments für das Übersetzungssystem METIS (Vandeghinste u. a. 2006). In dem hier beschriebenen Pilotexperiment geht es vor allem um den Konzeptbeweis, nämlich der kontextgebundenen Übersetzung von Aktanten, abhängig vom regierenden Verb: *Lügen* müsste hier ins Englische mit *rest* 'Ruhe' übersetzt werden. Als Kontrollbeispiel wurde in die Menge der Testsätze folgender Satz aufgenommen:

(38) Der Mann erzählt Lügen.

In diesem Beispiel muss *Lügen* mit *lies* übersetzt werden, d.h. der Zusammenhang zwischen Valenzträger und Aktanten bei der Übersetzung muss für die verschiedenen Kontexte jeweils entsprechend modelliert werden.

Satz (36) stellt einen Fall dar, in dem ein deutscher quasi-Phraseologismus *der Umsatz beträgt* im Englischen in eine Kopulakonstruktion gefasst werden muss. Dabei ist der Weg vom Deutschen zum Englischen einfach, würde doch ein simpler Übertrag von *betragen* zum Englischen Verb *be* reichen. Interessanter ist die Gegenrichtung, die mit folgenden beiden Testsätzen getestet wird:

(39) Our coverage is a billion.

(40) The man is crazy.

In Satz (39) ist eine Übersetzung nach *betragen* erforderlich, in Satz (40) ist im Deutschen das Verb *sein* wohl die beste Wahl; in beiden Sätzen ist das Ausgangsverb aber *be*, und muss abhängig vom Aktanten entsprechend übersetzt werden. Auch hier besteht also, wie in Kontrollbeispiel (38) eine wechselseitige Abhängigkeit zwischen Aktanten und Valenzträger in der Übersetzung, deren Effekt in der Grammatik modelliert werden muss.

Satz (37) bietet gleich zwei Phänomene. Wie in den Analysen zum Wechsel von grammatischen Funktionen in Abschnitt 4.3.1.4 gezeigt, werden nicht-kanonische Wortstellungen im Deutschen Ausgangssatz häufig dadurch im Englischen kompensiert, dass dem am Beginn des Satzes stehenden Objekt im Deutschen die Subjektfunktion im Englischen zugewiesen wird. Eine andere Möglichkeit wäre, die jeweiligen Begleiter auszutauschen, d.h. das Subjekt wird nach vorne zu verschieben, das Objekt nach hinten. Zusätzlich stellt sich in diesem Satz das Problem der Einbettung in der Partizipialphrase *die vom Himmel fallende Asche*: Die Prämodifikation im Deutschen muss in eine Postmodifikation im Englischen umgewandelt werden.

Zur Behandlung z.B. der Postmodifikation waren einige Regeländerungen im Detail an syntaktischen Regeln notwendig, deren Beschreibung und Diskussion aber den Rahmen dieses Abschnitts sprengen würden. Die wichtigsten Regeln zur Behandlung der oben aufgeführten Problemstellungen sind im folgenden aufgeführt und beschrieben. Relevante Teile sind durch Fettdruck hervorgehoben.

Um Satz (35) korrekt übersetzen zu können, wurden zwei Transferregeln für das Verb *strafen* und zwei Regeln für das Nomen *Lügen* eingeführt. Erst

durch diese Dopplung, die die wechselseitige Abhängigkeit modelliert, wurde eine korrekte Übersetzung der Lexeme erreicht. Der Regelsatz für *Lüge* beschreibt den Transfer des Lexems in Abhängigkeit des regierenden Verblexems (Merkmal *vlex*, entsprechendes Merkmal-Wert-Paar durch Fettdruck hervorgehoben): Ist dieses Verblexem *strafen*, wird *Lüge* mit *rest* übersetzt, ansonsten mit *lie*. Damit dürfte auch das Kontrollbeispiel in Satz (38) korrekt übersetzt werden, wo *Lügen* mit *lie* übersetzt werden sollte.

```
atom  =
      {lex=lie,cat=n}.[]
      <=>
      {lex='Lüge', vlex~='strafen'}.[].
atom  =
      {lex=rest,cat=n}.[]
      <=>
      {lex='Lüge', vlex='strafen'}.[].
```

Die beiden Regeln beginnen dabei mit einem Namen (in diesem Fall mit dem generischen Namen *atom*) und bestehen aus einer linken und einer rechten Seite, verbunden durch den Doppelpfeil „<=>“. Dieser besagt, dass ein Transfer in beide Richtungen möglich ist. Theoretisch wäre auch ein einseitig gerichteter Pfeil möglich. Insgesamt gehorcht die Regelsyntax den Konventionen der Programmiersprache PROLOG, daher müssen z.B. die Lexeme rechts des Doppelpfeils in Anführungszeichen gesetzt, und Regeln mit einem Punkt am Ende abgeschlossen werden. Ein gültiges Merkmal wird mittels eines Attribut-Wert-Paars bestehend aus einer Klausel *attribut = wert* dargestellt. Wird ausgedrückt, dass ein Merkmal nicht gilt, dann lautet die Klausel *attribut ~= wert*, wobei ‚~=‘ für die Verneinung steht. Ein Merkmalsbündel wird von geschweiften Klammern umgeben. Hängen von einem Merkmalsbündel weitere Merkmalsbündel ab, sind diese innerhalb der mit Punkt an das Bündel angeschlossenen eckigen Klammern beschrieben. Ansonsten bleiben die eckigen Klammern leer, wie in den obigen beiden Regeln.

Der Regelsatz für den Transfer des Verbs *strafen* selbst beschreibt etwas ähnliches wie die oben aufgeführten Regeln. Der Transfer ist abhängig davon, ob das dritte Argument (*frame:arg3*) das Lexem *Lüge* ist, das ohne eine Präposition angeschlossen ist (Übersetzung mit *lay*), oder ob es irgendein Lexem ist, das mit der Präposition *für* angeschlossen ist (*jd. straft jdn. für etw.*, Übersetzung mit *punish*). Tatsächlich wird mit dieser Regel neben der bilexikalischen Abhängigkeit bei der Übersetzung ein Lexikoneintrag für ein FVG beschrieben, das zwei weitere Aktanten – einen Strafenden und etwas oder je-

mand zu Strafendes – bindet. An dieser Stelle fehlt, aufgrund der eingeschränkten Behandlung topologischer Felder durch CAT2, die syntaktische Vorgabe für den komplexen Valenzträger (die FVG), dass im Hauptsatz nur die Klammerstellung erlaubt ist.

```
atom =
      {lex=lay,frame={arg3:
      arg3={lex=rest,agr={num=sing},
      pform=to}}}.[]
      <=>
      {lex='strafen',frame={arg3:
      arg3={lex='Lüge',pform=nil}}}.[].
atom =
      {lex=punish,frame={arg3:
      arg3={pform=for, lex~=rest}}}.[]
      <=>
      {lex='strafen',frame={arg3:
      arg3={pform='für'}}}.[].
```

Im vorliegenden Regelsatz wurde zusätzlich ausgeschlossen, dass *rest* drittes Argument von *punish* ist.

Für Satz (36) und die englischen Kontrollbeispiele (39) und (40) wurde ein ähnlicher Steuerungsmechanismus für den Transfer gewählt wie im vorhergehenden Beispiel. Hier ist die Übersetzung des Verblexems *be* wieder abhängig von einem Aktanten, in diesem Fall vom ersten Aktanten *(frame:arg1)*. Allerdings beschreibt die folgende Regel keine so starke bilexikalische Abhängigkeit wie die *Lügen strafen*-Regel. Sie besagt nur, dass das Verb *betragen* als Aktanten eine Instanz einer bestimmten semantischen Klasse verlangt, und dies bei der Übersetzung des Verbs *be* aus dem Englischen berücksichtigt werden muss. Für Begriffe aus der Wirtschaft, wie etwa *Umsatz*, wurde der semantische Wert *abs_economy_unit* für das Attribut *semf* eingeführt. Besitzt der erste Aktant in einem Satz dieses Merkmal, wird *be* mit *betragen* übersetzt, ansonsten mit *sein*, wie die folgenden Transferregeln beschreiben.

```
atom =
      {lex=be,frame={arg1:
      arg1={semf=abs_economy_unit},
arg2={role=attr}}}.[]
      <=>
```

```
        {lex='betragen',frame={arg1:
        arg1={semf=abs_economy_unit}}}.[].
atom =
        {lex=be,frame={arg1:
        arg1={semf~=abs_economy_unit}}}.[]
        <=>
        {lex='sein',frame={arg1:
        arg1={semf~=abs_economy_unit}}}.[].
```

Zusätzlich ist beim ersten Eintrag von *be* angegeben, dass der zweite Aktant ein attribuiertes Element ist, um es von einer lokalen Verwendung von *be* (*be in a place*) zu unterscheiden und eine Generierung von Präpositionen im Englischen zu unterdrücken.

Das komplexeste Beispiel ist Beispiel (37). Hier muss sichergestellt werden, dass zum einen im Englischen das Subjekt in Erstposition steht und zum anderen die Prämodifikation von NP durch Partizipialkonstruktionen im Deutschen in eine Postmodifikation im Englischen umgewandelt wird. Zusätzlich dazu steht die von einer Partizipialphrase abhängige NP bzw. PP im Deutschen links der Partizipialphrase, im Englischen dagegen rechts. Zur Behandlung dieser Phänomene ist ein ganzer Satz von Regeln notwendig. Die korrekte Stellung der Subjektposition wird wieder durch das Schema einer Transferregel kontrolliert. Diese Transferregel ist allerdings deutlich komplexer als die vorher gezeigten, da sie sich nicht nur auf das Verblexem an sich beschränkt. Vielmehr wird die gesamte Interfacestruktur des Satzes mit dem Oberknoten *{role=proposition}* angesprochen. Die Kinder der Proposition im Falle von Satz (37) sind neben dem jeweiligen deutschen bzw. englischen Verblexem die beiden davon abhängigen Nominalphrasen. Die folgend aufgeführte Transferregel koindiziert diese beiden Nominalphrasen im Deutschen und Englischen mit *np1* bzw. *np2* und tauscht deren Position.

```
worry_beunruhigen  =
        {role=proposition}.[{cat=v,lex=worry},
        np1:{focus=1,case=nom},np2:{focus~=1}]
        <=>
        {role=proposition}.[{cat=v,lex='beunru
        higen'},
        np2,np1].
```

Die Nominalphrase, die in der Interfacestruktur an erster Stelle nach dem Verb steht, sollte als Subjekt generiert werden.

Was die Struktur der Partizipialphrase angeht, wurden für das Deutsche und das Englische auf der syntaktischen Ebene zwei verschiedene Regeln aufgeführt. In der deutschen Regel steht die zum Partizip gehörende NP oder PP links des Partizips:[42]

```
partp =
      {cat=partp}.
      [ {cat=(pp;np)},
        {cat=(part1;part2)} ].
```

In der englischen Regel steht die NP bzw. PP rechts davon:

```
partp =
      {cat=partp}.
      [ {cat=(part1;part2)},
        {cat=(np;pp)} ].
```

Dadurch unterscheidet sich auch der Transfer der syntaktischen Struktur in die Interfacestruktur. Die kanonische Reihenfolge in der Interfacestruktur wurde so definiert, dass zuerst das Partizip und dann die NP bzw. PP steht. Im Deutschen müssen diese Elemente daher beim Übertrag in die Interfacestruktur getauscht werden (wieder durch Koindizierung dargestellt)[43]:

```
partp   =
      {cat=partp}.
      [ nppp:{cat=(np;pp)},
        part:{cat=(part1;part2)}]
      <=>
      {role=mod,cat=(part1;part2)}.
      [ part:{cat=(part1;part2)},
        nppp:{cat=n} ].
```

Im Englischen ist die Reihenfolge auf der syntaktischen Ebene und in der Interfacestruktur gleich:

42 An dieser Stelle und im Folgenden sind die Regeln nur auszugsweise wiedergegeben, da die Aufführung der Vielzahl von Merkmalen, die Aspekte wie Kasus, Numerus usw. kontrollieren, für die Erläuterung der Funktionsweise der Regeln eher hinderlich wäre.

43 Auf welche der Darstellungsebenen bzw. auf welche Art des Transfers sich eine Regel bezieht, hängt davon ab, in welchem Modul die Regel definiert ist.

```
partp   =
      {cat=partp}.
      [ part:{cat=(part1;part2),
        nppp:{cat=(np;pp)}]
      <=>
      {role=mod}.
      [ part:{cat=(part1;part2)},
        nppp:{cat=n} ].
```

Der Unterschied in der Prä- bzw. Postmodifikation besteht aber wie bereits erläutert nicht nur in der Partizipialphrase selbst, sondern auch mit Bezug auf das durch die Partizipialphrase modifizierte Nomen: Im Englischen steht die Partizipialphrase hinter dem Nomen, im Deutschen vor dem Nomen. Dies ist in allen Regeln, die die verschiedenen Varianten von Nominalphrasen beschreiben, reflektiert. Im Deutschen muss die Regel also wie folgt lauten:

```
np = {cat=np}.[
            ^{cat=det},
            ^{cat=(partp;ap)},
            {cat=n}].
```

Der vorangestellte Circonflex ‚^' steht für ein optionales Element, ein Stern ‚*' steht für null bis beliebig viele Elemente einer Art. Im Englischen lautet eine entsprechende NP-Strukturregel dann:

```
np = {cat=np}.[
            ^{cat=det},
            *{cat=ap},
            {cat=n},
            ^{cat=partp}].
```

Steht im Deutschen die Partizipialphrase an derselben Position wie potenzielle Adjektivphrasen (*{cat=(partp;ap)}*), so ist im Englischen die Partizipialphrasen in den postnominalen Teil verschoben.

Die unterschiedlichen Positionen sind, ähnlich wie bei der Partizipialphrase, natürlich beim Transfer der NP-Strukturen in die Interfacestruktur zu beachten. Für die Interfacestruktur wurde die Reihenfolge gewählt, bei der die Partizipialphrase hinter dem Nomen, aber vor potenziellen Adjektiven steht. Dies sei anhand einer deutschen Beispieltransferregel verdeutlicht:

```
np =
      {cat=np}.[
            ^a:{cat=ap},
            ^part:{cat=partp} ,
            n:{cat=n} ]
      <=>
      {d=no}.[
            n:{cat=n},
            ^part:{cat=(part1;part2)},
            ^a:{cat=a} ].
```

Die entsprechende englische Regel kann zwar die Partizipialphrase in der syntaktischen Struktur an der Stelle hinter dem Nomen belassen, muss aber die Adjektivphrasen beim Transfer von Interfacestruktur in die syntaktische Ebene vor das Nomen verschieben:

```
npnodefsing =
      {cat=np,type=T,type~=rel}.[
            *a:{cat=ap},
            n:{cat=n,type~=pron},
            ^part:{cat=partp} ]
      <=>
      {d=indef,type~=(pron;rel)}.[
            n:{cat=n},
            ^part,
            *a:{cat=a} ].
```

Als Partizipien wurden im Lexikon im Deutschen und Englischen bisher nur *fallend* bzw. *falling* eingeführt. Die Kategorien *part1* bzw. *part2* wurden dabei als eigene Kategorien für Partizip Präsens bzw. Partizip Perfekt eingeführt. Eine Behandlung, die diesen Ansatz fest in die bestehende CAT2-Grammatik integriert und die Merkmalsstrukturen optimiert, müsste diese Kategorien als Unterkategorien der Kategorie *v* für Verben etablieren und alle Regeln zur Verbbehandlung entsprechend überprüfen und anpassen. Dies hätte jedoch den Rahmen der vorliegenden Machbarkeitsstudie gesprengt.

Die oben beschriebenen Ergänzungen zur CAT2-Grammatik erzeugten die im Folgenden aufgeführten Übersetzungen. CAT2 generiert potenziell mehrere mögliche Übersetzungen mit allen syntaktischen und morphologischen Varianten, die die Grammatik erlaubt. Die Herausforderung einer optimierten

Grammatik liegt also darin, nur eine einzige, möglichst korrekte Übersetzung zu generieren, diese Optimierung war aber nicht Teil der Machbarkeitsstudie.

Im Folgenden sind die jeweils ersten Übersetzungen der Testsätze durch CAT2 ohne erweiterten Regelsatz (Kürzel C1) und mit erweitertem Regelsatz (Kürzel C2) aufgeführt. Zum Vergleich wurden auch diese gekürzten Sätze nochmal in Google Translate eingegeben und dessen Übersetzungen ebenfalls angegeben (Kürzel GT).

(35) *Diese Tatsache straft diesen Mythos Lügen.*
C1: The fact punish the myth lies.
C2: The fact lays this myth rest.
GT: This fact punishes those Mythos lies.
(36) *Unser Umsatz beträgt eine Milliarde.*
C1: Beträgt our coverage a billion.
C2: Our coverage is a billion.
GT: Our turnover is one billion.
(37) *Mich beunruhigt die vom Himmel fallende Asche.*
C1: I worries who sky falling ash.
C2: The ash falling from sky worries me.
GT: I am disturbed by the ash falling from the sky.

Wie an den C1-Übersetzungen zu sehen ist, können bereits einzelne fehlende Regeln für den gesamten Übersetzungsprozess in CAT2 destruktiv sein. Bei Satz (35) wird die Wendung völlig falsch mit *punish...lies* übersetzt. Dabei wird auch das Verb nicht korrekt gebeugt. Für Satz (36) wurde der Transfer von *betragen* in C2 überhaupt erst eingeführt. Da er in C1 komplett fehlt, tritt ein robuster Mechanismus in Aktion, der die im Transferlexikon enthaltenen Lexeme übersetzt, aber die Interfacestruktur sowie die nicht übersetzbaren Lexeme unberührt lässt. Bei (37) ergibt sich ohne die Behandlung von Partizipien sowie der entsprechenden Postmodifkation im Englischen eine vollkommen unverständliche Übersetzung, bei der nichts weiter geschehen ist, als dass Lexeme ins Englische übertragen, aber morphosyntaktisch nicht korrekt generiert wurden (*worries* statt *worry*). Die NP, die die Partizipialkonstruktion enthält, wurde fälschlicherweise als Relativsatz interpretiert.

Wie an den C2-Übersetzungen der Sätze (35) und (37) zu sehen ist, ist die korrekte Generierung von PP im Englischen in CAT2 durchgehend problematisch; die Ursachen dafür konnten im Rahmen dieser Studie nicht gefunden werden, und bedürften einer kompletten Überprüfung aller PP-bezogenen Regeln im Englischen und Deutschen, da Fehler im Englischen auch von fehlerhaften Transfers aus deutschen syntaktischen Strukturen in die Interfacestruk-

tur abhängen können. Dies hätte den Rahmen dieser Machbarkeitsstudie gesprengt. Abgesehen davon wurden die Hauptziele der Studie erreicht, wie die folgende Analyse belegen soll.

Die deutsche Wendung *Lügen strafen* in Satz (35) wird von CAT2 im Englischen korrekt mit *lay ... to rest* wiedergegeben. Das Lexem *Lügen* im Kontrollsatz (38) wird korrekt mit *lie* übersetzt:

(38) *Der Mann erzählt Lügen.*
C2: The man tells lies.
GT: The man told lies.

Was die Umsetzung der Übersetzung von FVG angeht, so könnte ein phrasenbasiertes statistisches MÜ-Konzept wie das von Hoang u. a. (2009), das typisierte Leerstellen in Phrasen erlaubt, die syntaktische Struktur der FVG im Hauptsatz mit einem Muster wie straft (NP|PP)+ Lügen modellieren. Problematisch ist dies aber im Nebensatz oder in Modalsätzen, wo Verb und Nomen der FVG zusammen stehen, wie etwa in *Ich will ihn Lügen strafen*. Um diese Variationen zu modellieren, sind wahrscheinlich komplexere statistische Formalismen nötig, wie etwa die **dependency treelet translation** (Quirk u. a. 2005; Ding und Palmer 2005), die Depen-denz(teil)strukturen zum Transfer verwendet, tiefere Einbettungen und damit mehr linguistischen Gehalt erlaubt.

Das Verb *beträgt* in Satz (36) wurde im Englischen korrekt mit *is* wiedergegeben. Die Übersetzung der Kontrollsätze (39) und (40) verdeutlicht die Wirkung des Konzepts:

(39) *Our coverage is a billion.*
C2: Unser Umsatz beträgt eine Milliarde.
GT: Unsere Berichterstattung ist eine Milliarde.
(40) *The man is crazy.*
C2: Der Mann ist verrückt.
GT: Der Mann ist verrückt.

Im wirtschaftssprachlichen Kontext wird *be* in der C2-Übersetzung ins Deutsche mit *betragen* übersetzt, im allgemeinen Kontext mit *sein*. In der GT-Übersetzung wird *coverage* als *Berichterstattung* übersetzt; die nicht ganz ungewöhnliche Übersetzung von *coverage* als *Umsatz* entstammt dem CroCo-Korpus. Diese Fehlinterpretation könnte erklären, warum *be* von GT nicht als *betragen* übersetzt wurde. Macht man jedoch die Gegenprobe mit dem

Satz *Our turnover is a billion*, übersetzt GT auch hier *Unser Umsatz ist eine Milliarde*.

Eine Wendung wie *der Umsatz beträgt*, die im Verlauf dieser Arbeit als registertypischer Phraseologismus bezeichnet wurde, lässt sich in der statistischen MÜ wohl am besten als Kollokation modellieren. Orliac und Dillinger (2003) stellen einen Ansatz zur Extraktion von Verb-Nomen-Kollokationen für die MÜ vor und demonstrieren dessen Anwendbarkeit am OpenLogos-System.

Satz (37) beinhaltete gleich zwei Problemstellungen, die beide gelöst wurden. Das Subjekt, das im Deutschen Ausgangssatz hinter dem Verb steht, wurde im Englischen vor das Verb gestellt. Gleichzeitig wurde die das Subjekt modifizierende Partizipialphrase hinter das Subjekt gestellt. Auch die vom Partizip abhängige PP steht im Englischen in postmodifizierender Position, anstatt das Partizip wie im Deutschen zu prämodifizieren. Die GT-Übersetzung fällt dabei noch besser aus als die C2-Übersetzung, da in der C2-Übersetzung der Artikel vor *sky* fehlt. Dies ist der recht komplexen Regelmenge zur Artikelgenerierung in CAT2 geschuldet, die gerade in Verbindung mit Präpositionen zu kaum nachvollziehbaren Effekten – wie etwa der Deletion des Artikels vor *sky* – führte.

Für die Modellierung dieser Strukturen könnte in der statistischen phrasenbasierten MÜ wiederum der Ansatz von Hoang u. a. (2009) gewählt werden. Dazu müsste sichergestellt werden, dass deutsche Partizipialphrasen die Leerstelle für die NP bzw. PP auf der linken Seite, englische dagegen auf der rechten Seite eröffnen. Analog dazu wäre diese Verfahrensweise bei Nomen mit der Leerstelle für eine potenzielle Partizipialphrase geeignet. Da Partizipialphrasen keine Variation erlauben wie etwa die zuvor angesprochenen FVG, dürfte diese Herangehensweise erfolgversprechender sein. Dass bereits GT diese Übersetzung korrekt beherrscht, obwohl es laut Hoang u. a. (ebd.) einen simplen phrasenbasierten Ansatz ohne getypte Leerstellen verwendet, überrascht.

In diesem Abschnitt wurde anhand einer Machbarkeitsstudie mit dem CAT2-MÜ-System gezeigt, wie die Erkenntnisse aus der Analyse von Valenzdivergenzen in ein MÜ-System integriert werden können. Dass von keinem System in allen Fällen eine grammatisch und lexikalisch vollständig korrekte Übersetzung und Generierung gelang, belegt ein weiteres Mal die Schwierigkeiten, die mit maschineller Übersetzung verbunden sind.

Der Wert der in dieser Arbeit gewonnenen Erkenntnisse in der MÜ liegt aber weniger in der Beschreibung rein syntaktischer Phänomene, sondern in den Erkenntnissen bezüglich sprachkontrast- und registerbedingter Divergen-

zen sowie bezüglich idiomatischer Strukturen wie Funktionsverbgefügen und Phraseologismen, die sich gerade durch hybride statistisch- und regelbasierte MÜ-Ansätze gut umsetzen lassen (vgl. Abschnitte 2.5.7, 4.2.2). Auch für diese Divergenzen wurden in der Machbarkeitsstudie Lösungen gefunden, die allerdings rein regelbasiert sind und in aktuelleren Systemen mit statistischen Daten zu ergänzen wären. Dass in der Pilotstudie zur manuellen Umsetzung der bilingualen Valenzinformation in Transferregeln mit dem regelbasierten CAT2 keine vollständig korrekten Ergebnisse erzielt werden, zeigt allerdings die Komplexität auf, mit der bei der Automatisierung von Transferregelableitung aus bilingualen Valenzinformationen zu rechnen ist.

4.4.2 Ein webbasiertes Wörterbuch für die HÜ

Valenzwörterbücher sind nicht nur der MÜ dienlich, sondern können auch dem Fremdsprachenlerner und Übersetzern von Nutzen sein. Nicht umsonst wurden die meisten mono- wie bilingualen Wörterbücher mit primär didaktischem Hintergrund erstellt (vgl. Abschnitt 2.3). Dieser Abschnitt stellt daher einige Überlegungen dazu an, wie die Daten aus dem Experiment zur Extraktion von bilingualen Valenzwörterbucheinträgen (vgl. Abschnitt 4.2.2) für Übersetzer und Fremdsprachenlerner nutzbar gemacht werden könnten. Ein fertiges Konzept kann dieser Abschnitt dabei nicht liefern, da bei einer Darstellung von Informationen, die über den üblichen Gehalt sowie die etablierten Strukturen von heutigen Wörterbüchern hinausgehen, wohl erst experimentell festgestellt werden kann, welche Darstellungsform sich für welche Zielgruppen am besten eignet.

Die Form, in der Valenzinformationen in heutigen Valenzwörterbüchern angegeben wird, ist oft nur für Personen mit einem entsprechenden Grundwissen in formaler Linguistik verständlich. Daran ändert auch nichts die Darstellung von Klassen mit Zahlen oder Kürzeln. Die Entzifferung des linguistischen Gehalts eines Eintrags, der Kürzel verwendet, wird nur verlagert, und muss beim Nachschlagen der Definition der angegebenen Klassen nachgeholt werden, so z.B. bei den Ergänzungsklassen des „kleinen Valenzlexikons“ (Engel und Schumacher 1976) und aller Valenzwörterbücher, die darauf aufbauen (Emons 1978; Rall u. a. 1980; Engel und Savin 1983; Bianco 1996). Der syntaktische Valenzeintrag bei Emons (1978) für das Verb *believe* hat die Form

S12[P12 + E1[NOM1/ES1] + [E2[NOM2/ES2[that]]]]

und dürfte selbst für einen Linguisten nicht ganz einfach entzifferbar sein (zur Interpretation des Eintrags vgl. Abschnitt 2.3.1). Auch bei der Internetseite der *Erlangen Valency Pattern Bank*[44], der im Internet verfügbaren Version des Valenzwörterbuchs von Herbst u. a. (2004a), darf bezweifelt werden, ob die angegebenen Valenzmuster für den Durchschnittssprachenlerner einfach verständlich sind. Ein für das Verb *see* angegebenes syntaktisches Muster ist

$$SCU + VHC_{act}$$

wobei *SCU* für *subject complement unit*, und VHC_{act} für *verbal head complex in active voice* steht. Die einfache grafische Darstellung ist tatsächlich in sich komplex, da die Interpretation bzw. das Verständnis der Abkürzungen wiederum formallinguistisches Wissen voraussetzt.

Nur wenige der in Abschnitt 2.3 genannten Valenzwörterbücher eignen sich direkt für den Durchschnittssprachenlerner bzw. -übersetzer. Eine Ausnahme bildet der Ansatz von Duffner u. a. (vgl. Abschnitt 2.4.3), in dem allerdings „nur" Kollokationswörterbücher extrahiert werden, und damit auf formallinguistische Definitionen und Informationen verzichtet werden kann.

Streitbar könnte im Übrigen die Frage sein, wo der „Durchschnittssprachenlerner" aufhört und der Linguist beginnt. Die Erfahrung zeigt, dass selbst Studierende einer Sprache oder eines Übersetzungsfachs kaum Kenntnisse in formaler Linguistik haben, die über die Zuweisung von Wörtern zu Wortklassen hinausgehen.

Eine Arbeit, die sich intensiv mit der Darstellungsform von Valenzwörterbucheinträgen beschäftigt, ist die Dissertation von Welker (2003). Welker schlägt dabei ein System von Siglen vor, bei dem semantische Klassen und syntaktische Kategorien verschmelzen: So steht die Sigle *P* für NP, die Personen repräsentieren. Andere Siglen, wie z.B. *ADJ*, stehen für Wortklassen. Das Verb in einem Wörterbucheintrag wird durch einen Stern * repräsentiert.

Diese Art der Darstellung erscheint aus der Sicht des Autors dem Durchschnittsnutzer gut vermittelbar. Syntaktische Details wie etwa die Strukturen von NP würden dabei wohl unterschlagen. Hat ein Sprachenlerner die Grundstrukturen einer Sprache verstanden, werden Kategorien wie simple NP oder PP oft intuitiv schon richtig aufgebaut. Erfahrungsgemäß werden abstrakte Strukturen wie der Bauplan von NP von Studenten ohne entsprechendes Interesse nur schwer verstanden (und von den meisten Sprachstudenten als

44 http://www.patternbank.uni-erlangen.de

nicht relevant für ihr Studium abgelehnt); die Vermittlung konkreter Wortklassen wie Adjektiv oder Verb allerdings ist deutlich einfacher. Auch abstrakte Klassen wie „Person“ oder „Institution“ sollten verständlich sein.

Akzeptiert man die im vorherigen Absatz formulierten Erfahrungswerte, kann man davon ausgehen, dass ein Valenzwörterbucheintrag in der Form, wie Welker ihn vorschlägt, z.B. *lieben: P * P*, auch für den Durschnittssprachenlerner verständlich ist. Allerdings ist die Frage, ob diese Art der Darstellung im Falle von Sprachen, die mehr morphologischen Gehalt haben als das Deutsche oder das Portugisische, ausreicht.

Die monolinguale deutsche Verbklassifikation von Ballmer und Brennenstuhl (1986 vgl. Abschnitt 2.3.1) geht bei der Darstellung von Valenzinformation einen anderen Weg als Welker. Die Verbklassen werden alle mit dem Muster der darin vorkommenden syntaktischen Begleiter angegeben, so z.B. *Zustoßen etw1 jd3*. Das Kürzel *jd* steht für Personen, *etw* für alle anderen Klassen, also z.B. Prozesse, Dinge oder abstrakte Größen. Neben diesen Informationen für die Verbklasse geben Ballmer und Brennenstuhl für jedes Verb mögliche Abweichungen vom Valenzmuster der Klasse an, allerdings ist jeweils nur das typischste Valenzmuster angegeben. Taucht ein Verb üblicherweise nur mit einem bestimmten Typ von Aktant auf, wird dieser in Klammern angegeben. Ein fiktiver Verbeintrag für das Verb *grasen* wäre dann:

grasen etw1 (ein Tier)

In der Variante von Welker wäre dies dann eher

*grasen: AN **

Das Sigel *AN* steht dabei für die Klasse der Tiere.

Der Vorteil der Angaben von Ballmer und Brennenstuhl besteht darin, dass sie keine Siglen verwenden, sondern gebräuchliche Kürzel wie *jd* für jemand oder *etw* für etwas. Soll der semantische Gehalt der Kürzel angegeben werden, wird dieser konkret genannt (z.B. *ein Tier*) und nicht durch ein Sigel wie *AN*. Zusätzlich geben sie mittels der an die Kürzel angehängten Nummern 1-4 den jeweiligen Kasus an, was bei Welker nur in Fällen von Genitiven und Dativen geschieht. Allerdings wäre dieser Umstand bei Welker mit einer simplen Erweiterung des Siglensystems um weitere Kasusangaben leicht zu beheben. Zudem spricht für das Welker'sche System, dass die Kürzel *jd* und *etw* sprachspezifisch sind und damit die Einträge der verschiede-

nen Sprachen weniger vergleichbar wären als bei einem konsistenten Siglensystem für alle Sprachen.
Der Frage, welche der beiden Darstellungsformen besser verstanden würde – oder ob es ggf. eine ganz andere sein müsste – muss experimentell nachgegangen werden. Allerdings muss die Darstellungsform nicht beim rein textuellen Eintrag stehenbleiben. Das Internet bietet dank immer dynamischerer Elemente viel mehr Möglichkeiten als nur den textuellen Eintrag.

Ein Beispiel dafür ist in Abbildung 4.13 gezeigt. Eine Abfrage zu Wortinformationen für das Verblexem *essen* auf der Seite des DWDS liefert u.a. den in der Abbildung dargestellten Netzwerkgraphen, der häufige Kollokate von *essen*, sowie deren häufige Kollokate, darstellt. Klickt man auf einen der Knoten, erhält man wiederum zu diesem Wort Wortinformationen.

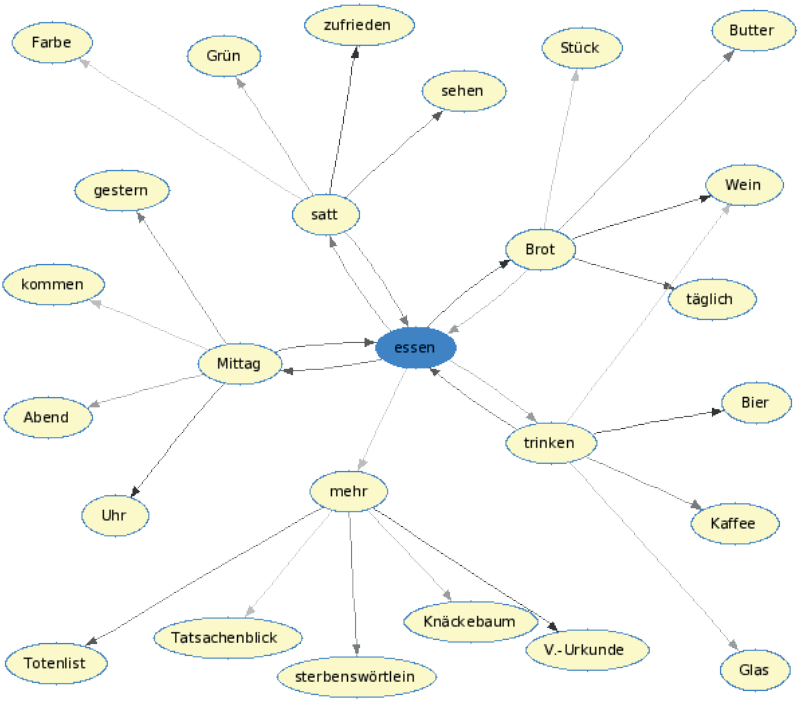

Abbildung 4.13: Häufige Kollokate von essen *in Netzwerkdarstellung auf der Webseite des DWDS (Quelle: http://www.dwds.de, 24.02.2010)*

In dieser Art der Darstellung fehlt der Hinweis darauf, welches der Elemente ein potenzieller Aktant ist, und wenn, dann welche Art von Aktant. Dies ließe sich beispielsweise durch eine farbliche Markierung bewerkstelligen, bei der die Position des textuellen Eintrags (z.B. des Akkusativobjekts) farblich mit möglichen Füllern der Akkusativstelle im Grafen koindiziert wäre.

Ebenso ist nicht zu erkennen, dass das Kollokat *Mittag essen* eigentlich einen Phraseologismus bildet, der wiederum selbst Valenzträger ist und in manchen Sprachen mit einem einzigen Verb übersetzt werden kann (z.B. spanisch *almorzar*).

Außerdem ist bei dieser Art der Darstellung noch unklar, wie bilinguale Information darzustellen wäre.

Bei allen noch offenen Fragen bietet eine dynamische Darstellung des Wörterbuchs mittels Webseiten dennoch Möglichkeiten, die ein klassisches Wörterbuch in dieser Form nicht bietet. So bietet eine Webseite die Möglichkeit, zuerst einen Basiseintrag für ein Lexem darzustellen und erst auf Mausklick hin weitere Informationen wie etwa Beispielsätze. Außerdem bieten Hyperlinkstrukturen die Möglichkeit, sich durch verschiedene Einträge „durchzuhangeln“. So ist einer der Vorteile des Internetwörterbuchs LEO, dass man alle möglichen Übersetzungen zu einem gesuchten Lexem anklicken kann, und wiederum deren mögliche Übersetzungen und in Einzelfällen auch deren Verwendungsweisen angezeigt werden. Damit kann man sich innerhalb kurzer Zeit ein relativ umfassendes Bild eines Wortfeldes machen; die Netzwerkdarstellung wie die des DWDS lädt dazu noch mehr ein.

Die in diesem Abschnitt dargelegten Überlegungen zu einem bilingualen Internetwörterbuch (deutsch-englisch, aber potenziell auch andere Sprachen) legen einige noch zu klärende Fragen offen, zeigen aber auch exemplarisch die Vorteile der webbasierten Darstellung. In jedem Fall sind sie ein gewichtiges Argument dafür, diese Art der Darstellungsform nicht nur im theoretischen Entwurf und dann im praktischen *trial-and-error*-Verfahren, sondern auf systematische Weise inklusive experimenteller Aufbauten mit Versuchspersonen aus den anvisierten Zielgruppen weiter zu ergründen.

5 Diskussion

Das vorangehende Kapitel beschäftigt sich mit dem technischen und konzeptuellen Aufbau der Experimente zu automatischen Valenzpaarextraktion, mit der Auswertung der Ergebnisse der Vorstudien und Experimente und zwei möglichen Anwendungsszenarien. Dieses Kapitel soll nun wieder die Verbindung zwischen der Studie und dem aktuellen Forschungsumfeld herstellen. Zunächst werden in Abschnitt 5.1 dazu Ergebnisse anderer, verwandter Studien mit denen dieser Studie verglichen. In 5.2 wird erörtert, welche weiteren theoretischen Implikationen sowie Fragestellungen sich aus dieser Studie mit Blick auf verschiedene Anwendungsgebiete paralleler Korpora ergeben. Und in 5.3 geht es um ganz praktische Folgerungen, die sich am besten als Desiderata an die Annotation von CroCo konkret und parallelen Korpora im Allgemeinen formulieren lassen.

5.1 Einordnung der Ergebnisse in das Forschungsumfeld

Eine sehr frühe Pilotstudie, die der hier vorliegenden Arbeit in Grundzügen gleicht, wird in (Fabricius-Hansen 1988) beschrieben. Darin vergleicht die Autorin norwegische und deutsche Verben auf mögliche Divergenzen in der (syntaktischen) Valenz. Die Studie beschränkt sich dabei auf Verben, die die Autorin als **übersetzungsäquivalent** ansieht. In der Definition von Fabricius-Hansen sind dies Verben, die auf der konzeptuellen Ebene die gleichen semantischen Rollen haben; **voll übersetzungsäquivalente** Verben bilden diese semantischen Rollen zudem auf vergleichbare grammatische Funktionen ab. Die Autorin erwähnt in diesem Rahmen einen Vergleich zweier Texte, eines deutschen Originals und einer norwegischen Übersetzung aus einem Deutschlernbuch mit einem Umfang von je 250 Sätzen, bei dem sie untersucht, wie häufig die übersetzungsäquivalenten Verben syntaktisch divergieren. Dies, so stellt Fabricius-Hansen fest, betrifft im ausgewählten Textpaar nur 6% der Verben. Sie merkt aber zugleich an, dass es sich dabei um einen vergleichsweise einfach strukturierten Text handelt, der zudem mit dem Ziel der Veranschaulichung für den norwegischen Benutzer übersetzt wurde, und dass ihr Ergebnis nur mit dem ähnlicher Texte vergleichbar sein könnte. Aus der Sicht der vorliegenden Untersuchung kann diese Einschränkung nur bestätigt werden. Die vorliegende Studie zeigt, dass sich die Art der syntaktischen Divergenzen sowohl je nach Übersetzungsrichtung und je nach Register unterschiedlich verteilt sind.

Solch detaillierte, manuell durchgeführten Valenzstudien wie von Fabricius-Hansen oder die kontrastiven Studien zwischen dem Deutschen und dem Englischen von Hawkins (1986) und König und Gast (2007) sind sehr aufwändig und in dieser Form nur selten verfügbar. Wichtig für den Sprachenlerner oder Übersetzer ist aber Hilfe bei der Einzelfallentscheidung. Die nötige Information dafür kann für einzelne Lexeme oder Kollokationen sowie geordnet nach Registern aus parallelen Korpora gewonnen werden können, wie z.B. in der vorliegenden Arbeit anhand der Valenzthematik demonstriert.

Eine Studie, die sich ähnlich wie die hier vorliegende Arbeit korpusbasiert mit Divergenzen zwischen Valenzrahmen von Verbpaaren beschäftigt, kommt aus dem Umfeld der Frame-Semantik und basiert auf Annotationen, aus dem SALSA-Projekt an der Universität des Saarlandes (Erk u. a. 2003), das ein deutsches Pendant zu FrameNet aufgebaut hat. (Padó und Lapata 2005) beschreiben, wie sie mittels Annotationsprojektion deutsche Sätze im FrameNet-Paradigma annotieren. Dazu wurden 1.000 deutsch-englische Satzpaare aus dem EUROPARL-Korpus geparst und die englischen Sätze mit framesemantischer Annotation versehen, die dann auf die deutschen Sätze projiziert wurde. (Padó 2007a) verwendet diese Ressource als Untersuchungsgrundlage für die Frage, inwiefern sich parallele Frameannotationen für translationswissenschaftliche Studien eignen. So wie bereits darauf hingewiesen wurde, dass es nicht selten zu Perspektivwechseln kommt, so stößt auch Padó auf eine Reihe von „mismatches“. Laut Padó kann nur in 60% der Fälle von einem „perfect match“ gesprochen werden, d.h. von Annotationen in denen sowohl der Frame als auch die Rollen auf beiden Seiten instanziiert sind. Die Frage, inwiefern die Instanziierungen syntaktisch-strukturell übereinstimmen, wird ausgeblendet.

Im Falle der 40%, die nicht perfekt übereinstimmen, stellt Padó verschiedene mögliche Effekte fest. Zum einen gibt es natürlich die Annotationsfehler, die mit 14% einen nicht geringen Anteil stellen. Im Falle von 8% handelt es sich um tatsächliche, eindeutige semantische Shifts. Was die übrigen 18% angeht, so spricht Padó einerseits von Problemen mit den Annotationsrichtlinien (uneindeutige Anweisungen, Unterscheidungen), geschlossenen Frames (also Frames, die eigentlich nicht weiter ausdifferenziert werden dürften) u.ä., lässt aber auch nicht unerwähnt, dass auch innerhalb dieser „problematischen“ 18% nicht immer eine perfekte Übereinstimmung zwischen den Framepaaren besteht.

Padós Ergebnisse zeigen einige Ähnlichkeiten zu den Ergebnissen dieser Studie, wenn sie auch auf semantischen und nicht auf syntaktischen Untersuchungen basieren. Auch in der hier vorliegenden Studie kann nur in 40-66% (und damit bei einem Mittelwert von 53%, nicht weit entfernt von Padós

60%) von einer genauen Entsprechung der Hauptverben gesprochen werden. Was in der vorliegenden Studie zu diesem Übereinstimmungswert geführt hat, waren allerdings teilweise syntaktische Kriterien. Syntaktische Divergenzen waren ein Ausschlusskriterium für eine genaue Entsprechung, wenn z.B. einem Vollverb ein FVG gegenüberstand. Aber auch semantische Kriterien spielten eine Rolle; die Vollverbpaare, die einen Perspektivwechsel beinhalten, wurden ebenfalls aus der Menge der voll übereinstimmenden PKA-Paaren herausgenommen. So ergibt diese Studie einen Anteil von 40-66% Äquivalenz bei Vollverbpaaren.

Padó (2009) verweist mit Bezug auf die durch Annotationsfehler oder -probleme bedingten Nichtübereinstimmungen auf einen Lösungsanschlag von (Gotsoulia 2008). Das Problem, so Padó, scheint darin zu liegen, dass der Annotierer gezwungen ist, sich immer für genau einen Frame pro Satz zu entscheiden.[45] Tatsächlich können aber von den Frameträgern mehrere, wenn auch sehr ähnliche, Bedeutungsnuancen ausgedrückt werden; zwischen den Sprachen kann aber einer der Aspekte der jeweiligen Nuance durchaus verloren gehen. Padó führt hier das Beispiel des englischen Verbs *notice* auf; das Verb wird dabei in seine verschiedenen Bedeutungsbestandteile dekomponiert:

Suddenly, Peter noticed his neighbour.
– perceive(peter,neighbour), !speakOnTopic(peter,neighbour)
In his speech, Peter noticed that the crisis was far from over.
– perceive(peter,crisis), speakOnTopic(peter,crisis)

Diese Form der Annotation, die von einem Frame als Prototypen mit verschiedenen möglichen Ausprägungen ausgeht, ist nach Wissen des Autors dieser Arbeit allerdings noch nicht praktisch umgesetzt worden. Eine solch detaillierte Annotation, so deutet es sich zumindest auf den ersten Blick an, könnte eine Reihe neuer Möglichkeiten eröffnen. Dies reicht von einer quantitativen Auswertung von Modulationen durch Auswertung nicht-realisierter Merkmale aus der AS in der ZS bis hin zur Verwendung solcher Merkmale in Metaklassifikationen von Verben, wie beispielsweise in (Čulo u. a. 2008a) dargestellt. In dieser Art der Metaklassifikation werden Verben folgend ihrer Zugehörigkeit zu Verbklassen Merkmale zugeordnet. Diese Merkmale können statische Eigenschaften wie *self_mover* oder Prozesse wie *rotation* beschreiben. Jedem Verb können zusätzlich Merkmale vergeben werden, die be-

45 Ähnlich wie in CroCo bezieht sich diese Annotation immer nur auf die oberste Ebene des Satzes.

schreiben, wie sich das jeweilige Verb von den anderen Verben seiner Verbklasse unterscheidet.

Durch die Beschreibung mithilfe von Merkmalen werden einerseits starre Verbklassengrenzen aufgehoben und es können Ähnlichkeiten über Verbklassen hinweg modelliert werden. Andererseits können diese Daten für eine Analyse, wie sie Gotsoulia vorschlägt, verwendet werden.

5.2 Beitrag zur theoretischen Diskussion

Parallele Korpora werden in verschiedenen Bereichen der Linguistik und der Übersetzungswissenschaft verwendet, etwa in der MÜ, im Fremdspracherwerb, in der Übersetzungsforschung und -praxis und in der multilingualen Grammatikinduktion. Die Ergebnisse dieser Studie können einen Beitrag zu allen genannten Bereichen leisten. Allerdings darf man nicht außer Acht lassen, dass es noch eine zweite Variante multilingualer Korpora gibt, nämlich Vergleichskorpora, die aus vergleichbaren Texten verschiedener Sprachen bestehen. Zwei Texte sind multilingual vergleichbar, wenn es sich dabei um Originale aus zwei verschiedenen Sprachen aber dem gleichen Register handelt. In der folgenden Diskussion wird es um theoretische Implikationen der Beobachtungen aus den Extraktionsstudien gehen. Dabei wird immer auch kritisch zur Rolle von Parallelkorpora Stellung genommen und eine mögliche (komplementäre) Verwendung von Vergleichskorpora diskutiert werden.

Parallelkorpora sind bis dato in der MÜ am verbreitetsten. In der statistischen MÜ (**statistical machine translation**, SMT) werden Parallelkorpora zur Wort-, Phrasen- und Satzalignierung verwendet, um diese Daten dann wiederum zum Training von Übersetzungsmodellen zu verwenden (vgl. insbesonders Brown u. a. 1993). Weiterhin werden sie zur Extraktion von Termwörterbüchern (z.B. Haller 2006) verwendet, was die automatische Übersetzung gerade von Fachtexten verbessern soll. Auch in der beispielbasierten MÜ (**example-based machine translation**, EBMT) werden Parallelkorpora herangezogen. Die Strategie der EBMT ist, „neue“, bis dahin ungesehen Übersetzungen **Analogien** aus bereits bestehenden Übersetzungen zu suchen und darauf dann nur noch notwendige linguistische Anpassungen (z.B. Austausch einzelner Lexeme) anzuwenden (vgl. z.B. Carl und Way 2003; Ding und Palmer 2004; Quirk u. a. 2005). Dass dafür aber nicht unbedingt Parallelkorpora notwendig sind, zeigen z.B.. für die MÜ Rapp (1999), Alegria, Ezeiza und Fernandez (2008), Snover, Dorr und Schwartz (2008), oder für die multilinguale Termextraktion Saralegi, San Vicente und Gurrutxuga (2008).

Zum Zwecke der MÜ soll die Verwendung von Korpora (mindestens) zwei Zielen dienen. Der aus dem Korpus extrahierte Ausdruck soll

- ein möglichst nahes Übersetzungsäquivalent für den originalsprachlichen Ausdruck sein und
- gleichzeitig ein möglichst **authentischer**, also im gegeben Kontext und in der stilistischen Form üblicherweise gebräuchlicher Ausdruck der Zielsprache sein.

Wie aus verschiedenen Studien bereits bekannt ist, besitzen Übersetzungen Eigenschaften, die typisch für Übersetzungen sind (Baker 1993; 1995) und sie damit weniger typisch im Vergleich zu Originalen der betreffenden Sprache machen. Zu diesen Eigenschaften gehört beispielsweise das **Levelling out**, was bedeutet, dass Übersetzungen einander ähnlicher sind als Originale der gleichen Textsorte. Es geht sogar soweit, dass die AS Auswirkungen auf die sprachliche Gestaltung des ZS-Textes hat, also „durchscheint" (**shining through**, vgl. Teich 2003). Dies wäre z.B. der Fall, wenn in englischen ZS-Texten, übersetzt aus deutschen AS-Texten, deutlich mehr Passivkonstruktionen verwendet würden als in englischen AS-Texten üblich. Dass solche Effekte z.B. von Teich (ebd.) oder von Hansen-Schirra (2003) auf der stilistischen Ebene beobachtet wurden, spricht aus Sicht eines Authentizitätskriteriums zunächst gegen den Einsatz von parallelen Korpora für die MÜ.

Vergleichskorpora bieten Texte, die als sprachtypisch gelten können; allerdings ist hier gleich einschränkend einzuwenden, dass sich jüngst Tendenzen zeigen, dass in bestimmten Disziplinen, in denen Englisch die internationale Kommunikationssprache ist, sich auch in der nicht-englischen AS-Produktion selbst bei Muttersprachlern Einflüsse des Englischen in Stil und Ausdruck nachweisen lassen. Ginge man dennoch davon aus, dass diese Effekte vernachlässigbar seien, so hätte man bei einem Vergleichskorpus die Sicherheit, dass extrahierte Ausdrücke und stilistische Eigenschaften authentisch für eine bestimmte Sprache wären. Damit wäre aber nicht die Frage geklärt, inwiefern ein Term oder allgemeinsprachlicher Ausdruck tatsächlich als übersetzungsäquivalent im jeweiligen Kontext gelten kann. Für den englischen Term *jurisdiction* schlägt z.B. das online-Wörterbuch LEO[46] Übersetzungen ins Deutsche wie etwa *Einflussbereich* oder *Rechtsprechung* vor. Eine Abfrage im Webinterface von EUROPARL[47] (Tiedemann und Nygaard 2004) ergibt weitere Übersetzungsmöglichkeiten wie etwa *Judikative* oder *Regelungsbefugnis*, die in LEO nicht vorgeschlagen werden und im letzteren Fall

46 http://dict.leo.org
47 http://urd.let.rug.nl/tiedeman/OPUS/cwb/Europarl/frames-cqp.html

gar nicht im Wörterbuch enthalten sind[48]. Ein Fachwörterbuch wird diese Begriffe zwar enthalten, dies ist nicht die Frage. Die große Herausforderung besteht allerdings darin, aus Texten, die zwar aus derselben Domäne stammen, aber nicht näher miteinander im Verhältnis stehen (insbesondere nicht im Übersetzungsverhältnis), den korrekten Gebrauchskontext für die jeweilige Übersetzung zu bestimmen.

In den Ergebnissen bezüglich der Divergenzen von PKA (Abschnitt 4.3.2) gibt es tatsächlich Indizien für beides, sowohl für Authentizität als auch für Durchscheinen. Dies belegt, dass die Situation in der Übersetzungswissenschaft nicht derart gesetzmäßig ist wie in der klassischen Physik. Die Tatsache, dass insbesondere in englischen SHARE-Texten eine hohe Zahl von Kopulakonstruktionen zu finden ist, diese ins Deutsche aber nicht so übersetzt werden, spricht für eine gewisse Authentizität. Andererseits finden sich in deutschen Übersetzungen aus englischen Reden Phraseologismen, die in deutschen Originalen in den 300 Satzpaaren, die für die Studie ausgezählt wurden, nicht vorkommen. Ob es sich hierbei allerdings um ein typisch englisches Stilmerkmal handelt, das in der Deutschen derart – originalgetreu – wiedergegeben wurde, oder an dieser Stelle der besondere Stil eines bestimmten englischen Sprechers durch ungenügende Randomisierung des Algorithmus durchschlägt, muss anhand der Auszählung einer größeren Satzmenge überprüft werden.

Man kann dennoch annehmen, dass bei der Verwendung alleine von Parallel- bzw. Vergleichskorpora, wie eben erläutert, immer wieder das Problem bestehen dürfte, beide Dimensionen – Authentizität und Äquivalenz – genau zu bestimmen bzw. zu optimieren (völlig authentisch und voll äquivalent).

Ein MÜ-System, das dieses Problem umgeht, ist das METIS-System[49] (Vandeghinste u. a. 2006), das aus dem Spanischen, Deutschen, Holländischen und Griechischen ins Englische übersetzt. METIS ist ein hybrides MÜ-System, das regel-, beispiel- und statistikbasierte Prinzipien vereint. METIS arbeitet aber mit einer flachen Analyse in der AS, im Transfer mit bilingualen Lexika und mit Regeln für sprachspezifische Phänomene, und in der Generierung mit einem ZS-Sprachmodell, das aus dem *British National Corpus* errechnet wird. Ein AS-Satz wird flach analysiert. Dann werden mittels der Lexika und der Transferregeln (z.B. wichtig bei Sprachen mit freierer Wortstellung als im Englischen) mögliche Übersetzungshypothesen (d.h. mögliche ZS-Sätze) generiert. Für jede Hypothese wird dann anhand des statistischen Modells errechnet, wie gut der generierte ZS-Satz ins ZS-Sprachmodell

48 Stand: 19.11.2009
49 http://www.iai-sb.de/forschung/content/view/16/31/

passt. Während die bilingualen Lexika und Transferregeln aus parallelen Daten gewonnen werden können, verlässt sich METIS bei der Auswahl der besten Übersetzung also auf Daten aus einem Korpus mit Originalen, dass zu einer Authentizitätsprüfung für die generierten Sätze herangezogen wird. Damit umgeht METIS die zuvor hypothetisierte Unschärfe, die auftreten könnte, wenn ein System sich alleine auf parallele Daten stützt.

Parallele Korpora sind in verschiedenen Pilotstudien bereits in der Übersetzerausbildung und in der Fremdsprachenlehre eingesetzt worden (z.B. Barlow 1998; Botley u. a. 2000; Pearson 2003; Olohan 2004; Hansen-Schirra und Teich 2008). Inwiefern der Einsatz paralleler Korpora die Übersetzungsqualität tatsächlich verbessert, soll im Rahmen einer Masterarbeit an der Universität Mainz untersucht werden (Schrader in Vorb.). Parallele Korpora eignen sich natürlich gut dafür, Studenten ebenso wie bereits in der Praxis tätigen Übersetzern Übersetzungsbeispiele zur Verfügung zu stellen, wie am Beispiel des EUROPARL-Korpus in den vorangehenden Abschnitten bereits dargestellt. Aber auch hier stellt sich immer wieder die Frage, inwiefern die Beispiele authentisch für die ZS sind. Bei veröffentlichten, redaktionell bearbeiteten Übersetzungen sollte dies der Fall sein. Eine andere Möglichkeit ist aber, Vergleichskorpora z.B. zur Erstellung von Lehr- und Lernmaterialien einzusetzen, wie z.B. in (Fernanda Bacelar do Nascimento u. a. 2008) dargestellt. Vergleichskorpora können, wie in folgendem Absatz noch genauer erläutert werden wird, dazu dienen, unterschiedliche Verwendungsweisen in verschiedenen Sprachen miteinander zu vergleichen und detaillierter zu analysieren. Es sei noch erwähnt, dass auch monolinguale Korpora sich prinzipiell als Recherchequellen für Sprachenlerner der entsprechenden Sprache eignen, die Einsatzmöglichkeiten aber gegenüber denen multilingualer Korpora deutlich eingeschränkt sind.

Der Einsatz von Korpora sollte sich nicht nur auf die reine Lexik beschränken. Wie im Rahmen der vorliegenden Studie und anderen Studien gerade im CroCo-Kontext gezeigt (Čulo u. a. 2008b; Neumann 2009), eignen sich Korpusstudien auch dazu, grammatische und stilistische Eigenheiten einer Sprache zu ergründen und – gerade für die Lernerumgebung – mit Beispielen zu belegen (vgl. z.B. Hansen-Schirra 2008). Eine sehr detaillierte Auswertung mit mehrdimensionalen Erklärungsansätzen für eine Vielzahl von Sprachen stellt (Slobin 2004) vor. Darin analysiert er ein Sprachkorpus, in dem Sprecher verschiedener Sprachen ein und dieselbe Geschichte in ihren Worten erzählen, mit Blick darauf, wie Bewegungsereignisse beschrieben werden und wie die linguistische Realisierung dieser Beschreibungen sowohl von linguistischen Eigenschaften der Sprache (z.B. Umfang des Wortfelds

der Bewegungsverben), aber auch von stilistischen und kulturellen Konventionen der zugrunde liegenden Kultur abhängen. Slobin beschreibt hier etwas, was man als **Spielregeln der Sprache** bezeichnen könnte, in Anlehnung an Wittgensteins Ausdruck vom **Sprachspiel** *(1998)*. Für einen Übersetzer oder Fremdsprachenlerner (sowie den Lehrer, selbstverständlich) ist es wichtig, nicht nur Lexik und Grammatik einer Sprache zu beherrschen, sondern auch ihre Spielregeln. Dies bezieht stilistische sowie kulturelle Hintergründe mit ein, so wie sie Slobin in seiner Analyse einbezogen hat. Erst, wer diese Spielregeln beherrscht, wird sich mehr als nur verständlich machen können, und in einer Sprache erfolgreich „handeln" können (vgl. zu diesem Thema z.B. House 1997).

Ein weiterer möglicher, MÜ-bezogener Einsatz von Korpora ist in der MÜ-Evaluation. Schon heute werden parallele Korpora hauptsächlich zum Erlernen von Lexem- und Phrasenwörterbüchern (Brown u. a. 1993; Fox 2002) eingesetzt. Wie bereits mehrfach am Beispiel der Verwendung von Kopulaverben im Englischen verglichen mit dem Deutschen dargelegt, wäre eine mögliche Erweiterung von MÜ-Systemen das Erlernen von vom Lexikon abgelösten Verwendungsregeln von Sprache. Um den korrekten Einsatz derart gelernter Regeln in einem MÜ-System zu überprüfen, böten sich wiederum Korpora als Evaluationsobjekte an, wobei sich registertypische Eigenschaften, grammatischer wie lexikalischer Natur, abprüfen ließen. Diese Art der Evaluation würde weit über die bisher gängigen Verfahren des Buchstaben- bzw. Lemmaabgleichs zwischen Referenzübersetzungen und dem MÜ-Produkt (z.B. der BLEU-Score, Papinieni u. a. 2002) hinaus gehen.

Ein Parallelkorpus könnte dabei als Eingabemenge für die Extraktion von Transferregeln fungieren. Aufgrund des zuvor angesprochenen Authentizitätsproblems wäre für die Extraktion registerbezogener grammatischer und lexikalischer Eigenschaften zum Zwecke der Evaluation ein Vergleichskorpus heranzuziehen.

Die aus dem Parallelkorpus extrahierten Transferregeln hätten immer eine statistische Komponente in sich. Ob diese statistischen Komponenten im konkreten Fall die Auswahl von Übersetzungsalternativen korrekt steuern, ließe sich überprüfen, indem man bei Eingabe der gleichen Originaltexte abprüfte, wie nahe die MÜ-Übersetzungen den Referenzübersetzungen in lexikalischen, grammatischen und stilistischen Eigenschaften kämen. Das Gewicht der Lexik wäre bei dieser Auswahl reduziert, man wäre also für eine möglichst gute Übersetzungsproduktion nicht allein auf eine möglichst korrekte lexikalische Auswahl festgelegt. Experimente wie etwa die von (Farwell u. a. 2004) belegen die Variationsbreite, die Übersetzungen ein und desselben Tex-

tes von verschiedenen menschlichen Übersetzern aufweisen. Man könnte daher dafür plädieren, dass grammatikalische und stilistische Eigenschaften sowie die lexikalische Auswahl mit Bezug auf ihre Auftretenshäufigkeiten im gesamten Text stärker gewichtet werden müsste als die genaue lexikalische Auswahl an Punkt X im Vergleich zu Punkt Y in der Referenzübersetzung. Ähnliche Überlegungen legt bereits (Steiner 1998) für die HÜ dar. Dass die lexikalischen Gesamteigenschaften eines Textes sich gut eignen, um Texte z.B. nach Fachgebiet zu klassifizieren, belegen etwa die Ergebnisse von (Teich und Holtz 2009). Ansätze zur Untersuchung von grammatischen Eigenschaften von Textsorten gibt es ebenfalls (Neumann 2009). All diese Ergebnisse können als Argument für die Entwicklung einer korpusgestützten MÜ- und HÜ-Evaluation dienen.

Ein Projekt, das bereits in diese Richtung geht, ist derzeit am IAI in Planung (Johann Haller, Email vom 03.02.2010). Dabei geht es darum, zwei Werkzeuge zu kombinieren: zum einen CLAT[50], das einsprachig eine linguistisch intelligente Grammatik-, Stil- und Terminologiekontrolle beherrscht, zum anderen ERRORSPY[51], das zweisprachige Terminologie- und Konsistenzkontrollen durchführt. Gerade die Stilkomponente von CLAT, die an die jeweiligen Anforderungen der Textsorte und des Fachgebiets angepasst werden können, gehen dabei in die angedachte Richtung der registerbasierten Evaluation.

5.3 Desiderata an die zukünftige Korpusannotation

Die Qualität der Extraktion von Valenzwörterbüchern aus Korpora erhöht sich natürlich mit der Qualität der Korpusannotation. Aus diversen vorangegangenen Überlegungen in dieser Arbeit lassen sich einige Desiderata für die weitere Korpusannotation von CroCo formulieren.

An verschiedenen Stellen in dieser Arbeit wurde die flache Annotation in CroCo bemängelt. Zwar wurden gegen Ende der Projektphase auch die Funktionen in Nebensätzen annotiert, allerdings nicht die Prädikate des Nebensatzes (da diese nur für den Hauptsatz vorgesehen sind). Zudem ist diese Annotation zum Zeitpunkt der Erstellung dieser Arbeit nicht vollständig vorhanden. Wie bereits in der Einleitung erwähnt, gibt es allerdings ein Nachfolgeprojekt an der Universität Mainz, das sich mit der tiefen Dependenzannotation von CroCo-Daten befassen soll. Damit werden auch Abfragen in eingebetteten Strukturen möglich.

50 http://www.iai-sb.de/iai/index.php/CLAT-Qualitaetspruefung-in-der-Technischen-Redaktion.html

51 http://www.multilingual-products.com/

Im Rahmen der Annotation der CroCo-Daten mit Dependenzen sollte des Weiteren die Wortalignierung verbessert und eine Phrasenalignierung eingeführt werden. Dazu sollen zumindest diejenigen Knoten des Dependenzbaums, die eine ganze Phrase repräsentieren, aligniert werden (ob halbautomatisch oder manuell, wird noch zu entscheiden sein). Dadurch würde einerseits eine bisher nicht bestehende Phrasenalignierung eingeführt, andererseits die Wortalignierung verbessert, da ja Knoten im Dependenzbaum für Wörter stehen.

Wichtig ist es zudem, PKA-Divergenzen leichter automatisch abfragbar zu machen. Dazu sollen PKA-Typen im Dependenzbaum mitannotiert werden. Dies könnte so aussehen, dass an den Wörtern, die zum Prädikatsausdruck dazugehören (wie z.B. „in“ und „Gefahr“ bei „in Gefahr bringen“), mit einem entsprechenden Merkmal versehen werden, z.B. *member-of:FVG*. Das für die Dependenzannotation vorgesehen Werkzeug TrEd ermöglicht eine derartige Annotation von Merkmalen.

Nicht zuletzt wäre es notwendig, semantische Veränderungsprozesse zwischen PKA zu annotieren. Möglich wäre es z.B., bei der Alignierung der PKA Kanten zu verwenden, die mit einem Prozess gelabelt sind, z.B. *Erweiterung*, *Verengung* o.ä. (wie dies z.B. schon bei Cyrus 2006 der Fall ist).

Von diesen Desiderata bezüglich der CroCo-Annotation lässt sich allerdings nur schwer auf vergleichbare Korpusprojekte generalisieren. Selbst für Korpora mit ähnlicher Zielsetzung – z.B. dem Einsatz in der MÜ – gelten nicht immer die gleichen Anforderungen, da unterschiedliche Verfahren zur Anwendung kommen können, korpusbasiert, regelbasiert, und diverse hybride Formen. Für ein paralleles Korpus ist wahrscheinlich eine möglichst gute Alignierung wichtig. Da Satz und Wort zwei wichtige Übersetzungseinheiten sind, erscheint eine möglichst gute Satz- und Wortalignierung als am wichtigsten. Jüngst ist aber auch der Wert der phrasenbasierten MÜ erkannt und verstärkt eingesetzt worden (vgl. Fox 2002). Eine Variante der phrasenbasierten MÜ ist die Dependenzteilbaum-MÜ (Quirk u. a. 2005; z.B. Ding und Palmer 2005). Beide können von einer möglichst guten Wortalignierung profitieren, da die Knoten im Dependenzbaum aus Wörtern bestehen, und bei einer Phrasenannotation bei einer Alignierung der Köpfe der Phrasen auch gleich eine Alignierung der Phrasen projiziert werden kann. In beiden Fällen ist eine tiefe Annotation bis zu den Terminalen vonnöten.

6 Fazit und Ausblick

Die vorliegende Arbeit hat gezeigt, dass sich parallele Korpusdaten zur Untersuchung von Valenzphänomenen und Extraktion von bilingualen Valenzwörterbüchern eignen. Es wurden Methoden zur Untersuchung von valenzrelevanten grammatischen Divergenzen anhand der Abfrage von empty links und crossing lines vorgestellt, und mittels quantitativer wie qualitativer Analysen Bezüge zwischen den Divergenzen und Faktoren wie kontrastiven Unterschieden, Registereinflüssen, Stil, Übersetzungsstrategien und Übersetzungseigenschaften hergestellt. Darüber hinaus wurde anhand eines MÜ-Experiments, bei dem das Wörterbuch durch die extrahierten Daten erweitert wurde, und anhand von Überlegungen zu einer möglichen online-Wörterbuchressource für den Menschen gezeigt, inwiefern die Ergebnisse der Analysen und der Extraktion praktische Relevanz haben.

Die Ergebnisse der Analysen in dieser Arbeit lassen sich mit denen anderer, jüngst durchgeführter semantischer orientierten Untersuchungen vergleichen und reihen sich damit in den aktuellen Forschungskontext mit einer syntaktischen und MÜ-bezogenen Ausrichtung ein.

Zur theoretischen Diskussion trägt diese Arbeit bei, indem sie die Frage von Authentizität und Äquivalenz in Parallel- und Vergleichskorpora beleuchtet und einige Indizien für beide Dimensionen im CroCo-Korpus präsentiert. Diese beiden Dimensionen werden zudem bezüglich ihrer nicht immer eindeutigen Rolle in der MÜ zueinander in Bezug gesetzt. Ebenso werden Überlegungen präsentiert, die Wege der Verwendung von Korpora in der Fremdsprachen- und Übersetzungslehre über den reinen Einsatz von Korpora als Beispielsammlung hinaus aufzeigen. Diese basieren auf den Erkenntnissen der Studien dieser Arbeit, und beziehen Dimensionen wie Register und Sprachkontrast ein. Auch wird dargelegt, dass dieselben Vorteile bei einem entsprechenden Einsatz von Korpora in der MÜ-Evaluation gegeben wären.

Nicht zuletzt wurden weitergehende Forschungsfragen auf Basis von kritischen Betrachtungen der eigenen Methoden und Ergebnisse entwickelt, und in Form von Desiderata an die Korpusannotation bereits konkret als erste Lösungsvorschläge formuliert.

Die Pilotstudie zur parallelen Valenzextraktion eröffnet eine Reihe weiterer möglicher Forschungstätigkeiten. Die Daten können dazu verwendet werden, um mittels quantitativer und qualitativer Methoden theoretische Fragestellungen wie die nach der Obligatorität und Fakultativität von Argumenten oder nach der Einordnung eines Arguments in die Gruppe der Aktanten und Modi-

fikatoren anhand von Korpusdaten zu überprüfen oder gar vorzunehmen. Dargelegt wurde ebenfalls, dass bereits zum Zeitpunkt der Erstellung dieser Arbeit erste Schritte zur Erweiterung der CroCo-Annotation unternommen wurden und damit eine bessere Auswertung für weitere translationswissenschaftliche Studien und eine bessere Extraktion von Valenzwörterbüchern möglich werden sollten.

Neben dem theoretischen Wert wurde auch der praktische Wert der Arbeit dargelegt. Dazu wurden zwei praktische Verwendungsweisen der Ergebnisse der Arbeit beschrieben: zum einen der Einsatz von Valenzinformationen in MÜ-Systemen, anhand einer Pilotstudie mit dem CAT2-System, zum anderen die Aufarbeitung der extrahierten Valenzmuster zu einem bilingualen, dynamischen Webwörterbuch. Gerade im ersteren Fall wäre eine erweiterte Dependenzannotation ebenfalls lohnend. Sie würde die Extraktion von Transferregeln erlauben, die – wie im Falle von LucyMT oder TectoMT – mit statistischen Daten angereichert wäre.

Nicht zuletzt wurde auch ein kurzer Ausblick darauf gegeben, wie die gewonnenen grammatischen, registergeordneten Daten für eine verbesserte, korpusgestützte MÜ- und HÜ-Evaluation genutzt werden können. Das in der Diskussion erwähnte CLAT-ERRORSPY-Projekt des IAI ist dabei eine interessante praktische Variante.

Literaturverzeichnis

Ágel, Vilmos. 2000. *Valenztheorie*. Tübingen: Narr.

Ágel, Vilmos, Ludwig M. Eichinger, Hans-Werner Eroms, Peter Hellwig, Hans-Jürgen Heringer, und Lobin (Hrsg.) 2003. *Valenz und Dependenz*. Bd. 1. 2 Bd. HSK 25. Berlin: Walter de Gruyter.

Alegria, Iñaki, Nerea Ezeiza, und Izaskun Fernandez. 2008. Translating Named Entities using Comparable Corpora. *Proceedings of the workshop on building and using comparable corpora*. Marrakesch, Marokko.

Allerton, David J. 1982. *Valency and the English verb*. London: Academic Press.

Anastasiou, Dimitra, und Oliver Čulo. 2007. Using topological information for detecting idiomatic verb phrases in German. *Proceedings of the Conference on Practical Applications in Language and Computers 2007*. Łodž, Polen.

Avgustinova, Tania, und Karel Oliva. 1996. *Unbounded dependencies in HPSG without traces or lexical rules*. CLAUS-Report. Saarbrücken: Universität des Saarlandes.

Baker, Collin F., Charles J. Fillmore, und John B. Lowe. 1998. The Berkeley FrameNet project. *Proceedings of COLING-ACL*. Montreal, Canada.

Baker, Mona. 1993. Corpus linguistics and translation studies. Implications and applications. *Text and technology: in honour of John Sinclair*, hg. v. M. Baker, G. Francis, und E. Tognini-Bonelli, 233-250. John Benjamins.

--- 1995. Corpora in translation studies: an overview and some suggestions for the future. *Target*, 2:223-243.

--- 1996. Corpus-based translation studies: The challenges that lie ahead. *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*, hg. v. Harold Somers, 175-186. Amsterdam: John Benjamins Publishing Company.

Ballmer, Thomas T., und Waltraud Brennenstuhl. 1986. *Deutsche Verben. Eine sprachanalytische Untersuchung des deutschen Verbwortschatzes*. Tübingen: Narr.

Barlow, Michael. 1998. Parallel texts in language teaching. *Multilingual corpora in teaching and research*, hg. v. Simon P. Botley, Anthony M. McEnery, und Andrew Wilson, 106-115.

Bennett, Winfield S., und Jonathan Slocum. 1988. The LRC machine translation system. *Machine translation systems*, hg. v. Jonathan Slocum,

111-140. Studies in natural language processing. Cambridge University Press.

Bianco, Maria Teresa. 1996. *Valenzlexikon deutsch-italienisch*. Deutsch im Kontrast 17. Heidelberg: Julius Groos.

Biber, Douglas. 1995. *Dimensions of register variation*. Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, und Edward Finegan. 2000. *Longman grammar of spoken and written English*. Harlow: Longman.

Boas, Hans C. 2001. Frame semantics as a framework for describing polysemy and syntactic structures of English and German motion verbs in contrastive computational lexicography. *Proceedings of Corpus Linguistics 2001*, 64-73.

--- 2002. Bilingual FrameNet dictionaries for machine translation. *Proccedings of the third international conference on language resources and evaluation*, 4:1364-1371. Las Palmas, Spanien.

Böhmova, Alena, Jan Hajič, Eva Hajičová, und Barbora Hladká. 2000. The Prague Dependency Treebank: A Three-Level Annotation Scenario. *Treebanks: building and using syntactically annotated corpora*, hg. v. Anne Abeillé. Kluwer Academic Publishers.

Bojar, Ondřej, und Jan Hajič. 2005. Extracting translation verb frames. *Proceedings of modern approaches in translation technologies*, hg. v. Walter von Hahn, John Hutchins, und Christina Vertan, 2-6. Bulgaria.

Botley, Simon P., Anthony M. McEnery, und Andrew Wilson. 2000. *Multilingual corpora in teaching and research*. Amsterdam: Rodopi.

Bouma, Gosse, Robert Malouf, und Ivan A Sag. 2001. Satisfying constraints on extraction and adjunction. *Natural language and linguistic theory* 1.1–65.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, und George Smith. 2002. The TIGER Treebank. *The TIGER Treebank. In Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT)*.

Brants, T. 2000a. Inter-annotator agreement for a German newspaper corpus. *Second International Conference on Language Resources and Evaluation LREC-2000*.

Brants, Thorsten. 2000b. TnT - A statistical part-of-speech tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*. Seattle, WA.

Bresnan, Joan. 2001. *Lexical-functional syntax*. Malden, MA, USA & Oxford, UK: Blackwell Publishers.

Bresnan, Joan, und Ronald Kaplan. 1982. Lexical-Functional Grammar: a formal system for grammatical representation. *The mental representation of grammatical relations*, hg. v. Joan Bresnan, 173-281. MIT Press.

Briscoe, Ted. 2001. From dictionary to corpus to self-organizing dictionary: learning valency associations in the face of variation and change. *Proceedings of corpus linguistics*, 79-89. Lancaster University.

Brown, Peter E., Stephen A. Della Pietra, Vincent J. Della Pietra, und Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 2.263-311.

Bühler, Karl. 1999. *Sprachtheorie. Die Darstellungsfunktion der Sprache*. 3. Aufl. Stuttgart: Lucius & Lucius.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, und Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. *Proceedings of LREC 2006*. Genoa, Italy.

Carl, Michael, und Andy Way (Hrsg.) 2003. *Recent advances in example-based machine translation*. Kluwer Academic Publishers.

Catford, John C. 1965. *A linguistic theory of translation. an essay in applied linguistics*. Oxford: Oxford University Press.

Čmejrek, Martin, Jan Cuřín, und Jiří Havelka. 2003. Czech English Dependency-based machine translation. *Proceedings of the 10th conference of the European chapter of the Association for Computational Linguistics*, 83-90. Budapest.

Čmejrek, Martin, Jan Cuřín, Jiří Havelka, Jan Hajić, und Vladislav Kubon. 2004. Prague Czech-English dependency treebank: syntactically annotated resources for machine translation. *Proceedings of LREC 2004*, 5:1597-1600. Lisbon, Portugal.

Copestake, Ann, Dan Flickinger, Robert Malouf, Susanne Riehemann, und Ivan Sag. 1995. Translation using Minimal Recursion Semantics. *In Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*.

Čulo, Oliver, Katrin Erk, Sebastian Padó, und Sabine Im Schulte Walde. 2008a. Comparing and combining semantic verb classifications. *Journal of Language Resources and Evaluation* 3.265-291.

Čulo, Oliver, Silvia Hansen-Schirra, Stella Neumann, und Mihaela Vela. 2008b. Empirical studies on language contrast using the English-

German comparable and parallel CroCo corpus. *Proceedings of the LREC 2008 workshop on building and using comparable corpora*.

Cyrus, Lea. 2006. Building a resource for studying translation shifts. *Proceedings of LREC 2006*.

Daneš, František. 1994. The Sentence-Pattern Model of syntax. *The Prague school of structural and functional linguistics*, hg. v. Philip Luelsdorff und Philip Luelsdorff, 197-222. Linguistic and literary studies in eastern Europe 41. Amsterdam/Philadelphia: John Benjamins.

Ding, Yuan, und Martha Palmer. 2004. Automatic learning of parallel dependency treelet pairs. *The first international joint conference on natural language processing (IJCNLP-04)*.

--- 2005. Machine translation using probabilistic Synchronous Dependency Insertion Grammars. *Proceedings of the 43rd annual meeting of the ACL*, hg. v. Ann Arbor, 541-8.

Dorr, Bonnie J. 1992. Parameterization of the interlingua in machine translation. *Proceedings of the 14th international conference on computational linguistics*, 624-630. Nantes, France.

--- 1994. Machine translation divergencies: a formal description and proposed solution. *Computational Linguistics* 20.597-633.

Drach, Erich. 1963. *Grundgedanken der deutschen Satzlehre*. Bd. 4. Darmstadt: Wissenschaftliche Buchgesellschaft.

Duffner, Rolf, Alain Kämber, und Anton Näf. 2009. Europäisch eingestellt – Valenzforschung mit Parallelkorpora. *Linguistik online* 39.

Emons, Rudolf. 1978. *Valenzgrammatik für das Englische*. Tübingen: Niemeyer.

Engel, Ulrich. 1977. *Syntax der deutschen Gegenwartssprache*. Grundlagen der Germanistik 22. Berlin: Schmidt.

--- 2006. Ein deutsch - bosnisch-/kroatisch-/serbisches Valenzlexikon. *Valenz und Dependenz*, 2: Handbücher zur Sprach- und Kommunikationswissenschaft 25. Berlin: Walter de Gruyter.

Engel, Ulrich, und Emilia Savin. 1983. *Valenzlexikon deutsch-rumänisch*. Deutsch im Kontrast 3. Heidelberg: Julius Groos.

Engel, Ulrich, und Helmut Schumacher. 1976. *Kleines Valenzlexikon deutscher Verben*. Mannheim: Institut für deutsche Sprache.

Erk, Katrin, Andrea Kowalski, Sebastian Padó, und Manfred Pinkal. 2003. Towards a resource for lexical semantics: a large German corpus with extensive semantic annotation. *Proceedings of ACL 2003*, 537-44.

Fabricius-Hansen, Cathrine. 1988. Valenz im Kontrast - aus rezeptiver Sicht. *Valenzen im Kontrast. Festschrift für Ulrich Engel*, hg. v. Pavica Mrazović und Wolfgang Teubert. Heidelberg.

Farwell, David, Steven Helmreich, Bonnie Dorr, Nizar Habash, Florence Reeder, Ketih Miller, Lori LEvin, u. a. 2004. Interlingual annotation of multilingual text corpora. *Proceedings of the workshop on frontiers in corpus annotation, NAACL/HLT 2004*. Boston, MA, USA.

Fernanda Bacelar do Nascimento, Maria, Antónia Estrela, Amália Mendes, und Luísa Pereira. 2008. On the use of comparable corpora of African varieties of Portuguese for linguistic description and teaching/learning applications. *Proceedings of the workshop on building and using comparable corpora*. Marrakesch, Marokko.

Fillmore, Charles. 1968. The case for case. *Universals in linguistic theory*, hg. v. Emmon Bach und Robert Harms, 1-88. Holt, Rinehart and Winston.

--- 1985. Frames and the semantics of understanding. *Quaderni di Semantica: Rivista Internazionale di Semantica Teorica e Applicata*.

Fillmore, Charles J. 1977. Scenes-and-frames semantics, Linguistic Structures Processing. *Fundamental studies in computer science*, hg. v. Antonio Zampolli, 55-88. North Holland Publishing.

--- 1982. Frame semantics. *Linguistics in the Morning Calm*, 111-137.

Forst, Martin, Núria Bertomeu, Berthold Crysmann, Frederik Fouvry, Silvia Hansen-Schirra, und Valia Kordoni. 2004. Towards a dependency-based gold standard for German parser - the TiGer Dependency Bank. *Proceedings of the LINC Workshop 2004*. Geneva.

Fox, Heidi J. 2002. Phrasal cohesion and statistical machine translation. *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, 304-11. Philadelphia: ACL.

Francis, W. Nelson, und Henry Kučera. 1979. Manual of information to accompany A standard corpus of present-day edited American English, for use with digital computers. *http://khnt.aksis.uib.no/icame/manuals/brown/*.

Franz, Philipp Koehn, Franz Josef Och, und Daniel Marcu. 2003. Statistical Phrase-Based Translation. *Proceedings of HLT-NAACL 2003*, 127–133.

Gebruers, Rudi. 1988. Valency and MT: recent developments in the METAL system. *Proceedings of the second conference on applied natural language processing*, 168-175.

Geyken, Alexander. 2007. The DWDS corpus: A reference corpus for the German language of the 20th century. *Idioms and Collocations: Corpus-based Linguistic, Lexicographic Studies*. Continuum Press.

Gotsoulia, Voula. 2008. An entailment-based approach to semantic role annotation. *Proceedings of the Second Linguistic Annotation Workshop held in conjunction with LREC 2008*. Marrakesch, Marokko.

Granger, Sylviane. 2003. The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies? *Corpus-based approaches to contrastive linguistics and translation studies*, hg. v. Sylviane Granger, Jacques Lerot, und Stephanie Petch-Tyson, 17-29. Amsterdam, New York: Rodopi.

Haller, Johann. 1993. CAT2- Vom Forschungssystem zum präindustriellen Prototyp. *Sprachtechnologie: Methoden, Werkzeuge, Perspektiven. Sprache und Computer*, hg. v. Horst P. Pütz und Johann Haller, 13:282-303. Hildesheim: Georg Olms Verlag.

--- 2006. AUTOTERM – automatische Terminologieextraktion Spanisch-Deutsch. *Multiperspektivische Fragestellungen der Translation in der Romania*, hg. v. Alberto Gil und Wienen, 229-242. Sabest. Frankfurt.

Halliday, M. A.K, und Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

--- 1989. *Language, context, and text: aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Hansen-Schirra, Silvia. 2003. Linguistic enrichment and exploitation of the Translational English Corpus. *Proceedings of the Corpus Linguistics 2003 conference*, 288-297. Lancaster, UK.

--- 2008. Interactive reference grammars: exploiting parallel and comparable treebanks for translation. *Topics in language resources for translation and localisation*, hg. v. Elia Yuste Rodrigo.

Hansen-Schirra, Silvia, Stella Neumann, und Erich Steiner. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. Berlin: De Gruyter.

--- 2007. Cohesive explicitness and explicitation in an English-German translation corpus. *Languages in Contrast*, 2.

Hansen-Schirra, Silvia, Stella Neumann, und Mihaela Vela. 2006. Multi-dimensional annotation and alignment in an English-German translation corpus. *Proceedings of the workshop Multi-dimensional markup in natural language processing (NLPXML-2006) at EACL*. Trento, Italien.

Hansen-Schirra, Silvia, und Elke Teich. 2008. Corpora in human translation. *International Handbook on "Corpus Linguistics"*, hg. v. Anke Lüdeling und Merja Kytö. HSK.

Hawkins, John A. 1986. *A comparative typology of English and German. Unifying the contrasts.* London: Croom Helm.

Helbig, Gerhard. 1992. *Probleme der Valenz- und Kasustheorie*. Tübingen: Niemeyer.

Helbig, Gerhard, und Joachim Buscha. 2001. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Berlin: Langenscheidt.

Helbig, Gerhard, und Wolfgang Schenkel. 1969. *Wörterbuch zur Valenz und Distribution deutscher Verben*. 4. Aufl. Tübingen: Niemeyer.

Herbst, Thomas, David Heath, Ian F. Roe, und Dieter Götz. 2004a. *A valency dictionary of English : a corpus-based analysis of the complementation patterns of English verbs, nouns and adjectives*. Berlin: Mouton de Gruyter.

Herbst, Thomas, Gunter Lorenz, Brigitta Mittmann, und Martin Schnell. 2004b. *Lexikografie, ihre Basis- und Nachbarwissenschaften*. Tübingen: Niemeyer.

Heyn, Matthias. 1996. Integrating machine translation into translation memory systems. *European Association for Machine Translation - Workshop proceedings*, 111–123.

Hoang, Hieu, Philipp Koehn, und Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. *Proceedings of the International Workshop on Spoken Language Translation*, 152-159. Tokyo, Japan.

House, Juliane. 1997. Mißverstehen in interkulturellen Begegnungen. *Wie lernt man Sprachen - wie lehrt man Sprachen*, 154-169.

Hutchins, John, und Harold Somers. 1992. *An introduction to machine translation*. London u.a.: Acadamic Press.

Jackendoff, Ray. 1990. *Semantics structures*. Cambridge: MIT Press.

Johansson, Stig. 2000. Towards a multilingual corpus for contrastive analysis and translation studies. *Språk i kontrast.* University of Oslo.

Johansson, Stig, Eric Atwell, Roger Garside, und Geoffrey Leech. 1986. The tagged LOB corpus. *http://khnt.hit.uib.no/icame/manuals/lobman/.*

Johnson, Rod, Maghi King, und Louis des Tombe. 1985. EUROTRA: a multilingual system under development. *Computational Linguistics* 11.155-169.

Kaplan, Ronald M., Klaus Netter, Jürgen Wedekind, und Annie Zaenen. 1989. Translation by structural correspondences. *Proceedings of the*

4th conference of the European chapter of the Association for Computational Linguistics.

Katz, Jerrold J., und Jarry A. Fodor. 1963. The structure of a semantic theory. *Language* 2.170-210.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of MT Summit X, Phuket, Thailand, 12-16 September 2005*.

Koller, Werner. 2001. *Einführung in die Übersetzungswissenschaft*. Narr Studienbücher. Tübingen: Gunter Narr.

König, Ekkehard, und Volker Gast. 2007. *Understanding English-German contrasts*. Berlin: Erich Schmidt Verlag.

Kuhn, Jonas. 2004. Experiments in parallel-text based grammar induction. *Proceedings of ACL 2004*. Barcelona, Spain.

--- 2005. An architecture for parallel corpus-based grammar learning. *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn*, hg. v. Bernhard Fisseni, Hans-Christian Schmitz, Bernhard Schröder, und Petra Wagner. Frankfurt a.M.: Peter Lang.

Kunz, Kerstin. 2007. A method for investigating coreference in originals and translations. *Languages in Contrast* 7.267-287.

Kußmaul, Paul. 2010. *Verstehen und Übersetzen*. 2. Aufl. Narr Studienbücher. Tübingen: Narr.

Langer, Stefan. 2004. A linguistic test battery for support verb constructions. *Verbes Support. Nouvel état des lieux* 2. Special issue of Linguisticae Investigationes.171-184.

Liu, Ydong, und Anoop Sarkar. 2009. Exploration of the LTAG-spinal formalism and treebank for semantic role labeling. *Grammar engineering across frameworks*. Singapore.

Luckhardt, Heinz-Dirk, und Heinz-Dieter Maas. 1983a. *SUSY - Handbuch für Transfer und Synthese*. Linguistische Arbeiten des SFB 100. Saarbrücken: Universität des Saarlandes.

--- 1983b. *SUSY - Handbuch für Transfer und Synthese*. Linguistische Arbeiten des SFB 100. Saarbrücken: Universität des Saarlandes.

Maas, Heinz-Dieter. 1984. *SUSY-II-Handbuch*. Linguistische Arbeiten des SFB 100. Saarbrücken: Universität des Saarlandes.

Maas, Heinz-Dieter, Christoph Rösener, und Axel Theofilidis. 2009. Morphosyntactic and semantic analysis of text: the MPRO tagging procedure. *State of the art in computational morphology. Workshop on systems and frameworks for computational morphology 2009*, hg. v.

Cerstin Mahlow und Michael Piotrowski, 76-87. New York: Springer.

Mareček, David, Zdeněk Žabokrtský, und Václav Novák. 2008. Automatic alignment of Czech and English deep syntactic dependency trees. *Proceedings of EAMT 2008*. Hamburg.

Müller, Christoph, und Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, hg. v. Sabine Braun, Kurt Kohn, und Joybrato Mukherjee, 197-214. Frankfurt a.M., Germany: Peter Lang.

Nerima, Luka, Violeta Seretan, und Eric Wehrli. 2003. Creating a multilingual collocation dictionary from large text corpora. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*.

Neumann, Stella. 2009. Quantitative register analysis across languages. *Thresholds and Potentialities of Systemic Functional Linguistics: Applications to other disciplines, specialised discourses and languages other than English.*, hg. v. Elizabeth Swain. Trieste: Edizioni Universitarie.

Neumann, Stella, und Silvia Hansen-Schirra. 2005. The CroCo project. Cross-linguistic corpora for the investigation of explicitation in tranlsations. *Proceedings of the Corpus Linguistics conference series*, 1:

Newmark, Peter. 1988. *A textbook of translation*. New York: Prentice Hall.

Och, Franz-Josef, und Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29.19-51.

Oliva, Karel. 2003. Dependency, valency and head-driven phrase structure grammar. *Dependenz und Valenz*, 1:660-668. HSK 25. Berlin: Walter de Gruyter.

Olohan, Maeve. 2004. *Introducing corpora in translation studies*. London: Routledge.

Orliac, Brigitte, und Mike Dillinger. 2003. Collocation extraction for machine translation. *Proceedings of Machine Translation Summit IX*, 292-298.

Padó, Sebastian. 2007a. *Cross-Lingual annotation projection models for role-semantic information*. Dissertation, Universität des Saarlandes.

--- 2007b. Translational equivalence and cross-lingual parallelism: the case of framenet frames. *Proceedings of the nodalida workshop on building frame semantics resources for scandinavian and baltic languages*. Tartu, Estonia.

Padó, Sebastian, und Mirella Lapata. 2005. Cross-lingual projection of role-semantic information. *Proceedings of HLT/EMNLP 2005*. Vancouver, BC.

Palm, Christine. 1997. *Phraseologie. Eine Einführung*. Narr Studienbücher. Tübingen: Narr.

Palmer, Martha, Joseph Rosenzweig, und Scott Cotton. 2001. Automatic predicate argument analysis of the Penn TreeBank. *HLT '01: Proceedings of the first international conference on Human language technology research*, 1–5. Morristown, NJ, USA: Association for Computational Linguistics.

Panevová, Jarmila. 1975. On verbal frames in Functional Generative Description, part II. *Prague Bulletin of Mathematical Linguistics* 23.17-52.

--- 1994. Valency frames and the meaning of the sentence. *The Prague school of structural and functional linguistics*, hg. v. Philip Luelsdorff, 223-43. Linguistic & literary studies in eastern Europe 41. Amsterdam/Philadelphia: John Benjamins.

Papinieni, Kishore, Salim Roukos, Todd Ward, und Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318. Philadelphia, USA.

Pearson, Jennifer. 2003. Using parallel texts in the translator training environment. *Corpora in Translator Education*, hg. v. Federico Zanettin, Silvia Bernardini, und Dominic Stewart, 15-24. Manchester: St. Jerome.

Petruck, Miriam R. L. 1996. Frame Semantics. *Handbook of pragmatics*, hg. v. Jef Verschueren, Jan-Ole Östman, Jan Blommaert, und Chris Bulcaen. Philadel: John Benjamins.

Pollard, Carl, und Ivan A Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Pollard, Carl, und Ivan A. Sag. 1987. *Information-Based Syntax and Semantics*. Bd. 13. CSLI Lecture Notes. CSLI.

Poutsma, Arjen. 2000. Data-Oriented Translation. *Proceedings of the 18th conference on Computational linguistics*, 635–641. Morristown, NJ, USA: Association for Computational Linguistics.

Quirk, Christopher, Arul Menezes, und Colin Cherry. 2005. Dependency treelet translation: syntactically informed phrasal SMT. *Proceedings of the 43rd annual meeting of the ACL*, hg. v. Ann Arbor, 271-79.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, und Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Harlow: Longman.

Rall, Dietrich, Marlene Rall, und Oscar Zorrilla. 1980. *Diccionario de valencias verbales: aleman-español*. Tübingen: Narr.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated english and german corpora. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Roukos, Salim, David Graff, und Dan Melamed. 1997. Hansard French/English. *http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20.*

Sag, Ivan A, und Janet Fodor. 1994. Extraction without traces. *Proceedings of the 13th West Coast Conference on Formal Linguistics*. Stanford, USA.

Saralegi, Xabier, I. San Vicente, und A. Gurrutxaga. 2008. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. *Proceedings of the workshop on building and using comparable corpora*. Marrakesch, Marokko.

Schrader, Katja. 2010. *Vergleich korpusgestützt und wörterbuchgestützt erstellter Übersetzungen mittels des pragmalinguistischen Evaluierungsmodells von Juliane House*. Diplomarbeit, Universität Mainz.

Schumacher, Helmut (Hrsg.) 1986. *Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben.* Bd. 1. Berlin: De Gruyter.

Sgall, Petr, Eva Hajičová, und Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Netherland.

Sharp, Randall. 1994. CAT2 Reference manual. Version 3.6. *http://www.iai.uni-sb.de/docs/refman.pdf.*

Slobin, Dan I. 2004. The many ways to search for a frog: linguistic typology and the ex-pression of motion events. *Relating Events in Narrative: Typological Perspectives*, hg. v. S. Strömqvist und L. Verhoeven. Mahwah, N.J.: Lawrence Erlbaum Associates.

Snover, Matthew, Bonnie J. Dorr, und Richard Schwartz. 2008. Language and translation model adaptation using comparable corpora. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii.

Sommerfeldt, Karl-Ernst, und Herbert Schreiber. 1977a. *Wörterbuch zur Valenz und Distribution der Substantive*. 1. Aufl. Leipzig: Bibliogr. Inst.

--- 1977b. *Wörterbuch zur Valenz und Distribution deutscher Adjektive*. 2. Aufl. Leipzig: Bibliograph. Inst.

Steiner, Erich (Hrsg.) 1989. *Predicate-argument structure for transfer*. Bd. 11. IAI Working Papers. Saarbrücken: IAI.

Steiner, Erich. 1998. A register-based translation evaluation. *Target* 10.291-318.

Steiner, Erich, Paul Schmidt, und Cornelia Zelinsky-Wibbelt. 1988. *From syntax to semantics: insights from machine translation*. London: Francis Pinter.

Streiter, Oliver. 1995. *Linguistic modeling for multilingual machine translation*. Dissertation, Universität des Saarlandes.

Teich, Elke. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Bd. 5. Text, Translation, Computational Processing. Berlin/New York: Mouton de Gruyter.

Teich, Elke, und Mônica Holtz. 2009. Scientific registers in contact: An exploration of the lexico-grammatical properties of interdisciplinary discourses. *International Journal of Corpus Linguistics* 4.

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klinksieck.

Tiedemann, Jörg, und Lars Nygaard. 2004. The OPUS corpus - parallel & free. *Proceedings of the fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lissabon, Portugal.

Vandeghinste, Vincent, Ineke Schuurman, Michael Carl, Stella Markantonatou, und Toni Badia. 2006. METIS II: Machine translation for low-resource languages. *Proceedings of the 5th international conference on language resources and evaluation*, 1284-1289.

Vannerem, Mia, und Mary Snell-Hornby. 1986. Die Szene hinter dem Text: scenes-and-frame semantics in der Übersetzung. *Übersetzungswissenschaft - eine Neuorientierung*, hg. v. Mary Snell-Hornby, 184-205. Tübingen: Francke.

Vela, Mihaela, und Silvia Hansen-Schirra. 2005. *CroCo: Multidimensionales Korpus-Alignment*. Project deliverable. Fachrichtung 4.6, Universität des Saarlandes.

Vinay, Jean-Paul, und Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais. Méthode de translation*. Paris: Didier.

Vintar, Špela, und Darja Fišer. 2008. Harvesting multi-word expressions from parallel corpora. *Proceedings of LREC 2008*.

Wahlster, Wolfgang (Hrsg.) 2000. *Verbmobil*. Artificial intelligence. Berlin/Heidelberg [u.a.]: Springer.

Way, Andy. 2003. Translating with examples: the LFG-DOT models of translation. *Recent advances in example based machine translation*, hg. v. Michael Carl und Andy Way. Dordrecht/Boston/London: Kluwers Academic Publishers.

Welker, Andreas Herbert. 2003. *Zweisprachige Lexikographie: Vorschläge für deutsch-portugiesische Verbwörterbücher*. Dissertation, Universität des Saarlandes.

Wittgenstein, Ludwig. 1998. *Ludwig Wittgenstein, Philosophische Untersuchungen*, hg. v. Eike von Savigny. Klassiker auslegen 13. Berlin: Akademie-Verlag

Žabokrtský, Zdeněk. 2005. *Valency lexicon of Czech verbs*. Dissertation, MFF Charles University Prague.

Žabokrtský, Zdeněk, Jan Ptáček, und Petr Pajas. 2008. TectoMT: highly modular MT system with tectogrammatics used as transfer layer. *Proceedings of WMT 2008*.

Zifonun, Gisela, Ludger Hoffman, und Bruno Strecker. 1997. *Grammatik der deutschen Sprache*. Schriften des Instituts für Deutsche Sprache. Berlin: De Gruyter.

Sprachdidaktik, Translation und Maschinelle Übersetzung haben seit geraumer Zeit von bilingualen Valenzwörterbüchern profitiert. Wurden diese Wörterbücher zuvor in aufwändiger Handarbeit erstellt, eröffnen multilinguale Korpora neue Perspektiven für eine (halb-)automatische Erstellung von Valenzwörterbüchern anhand realer Sprachdaten. Hier rücken insbesondere parallele Korpora - also Textsammlungen von Originalen und deren Übersetzungen - in den Fokus der Aufmerksamkeit, da sie das Auffinden von Äquivalenten zumindest theoretisch erleichtern. Praktisch steht dem entgegen, dass Original und Übersetzung nicht immer völlig deckungsgleich sind, auf syntaktischer wie semantischer Ebene. Im vorliegenden Buch werden Experimente beschrieben, die anhand eines deutsch-englischen Parallelkorpus untersuchen, wie syntaktische Divergenzen zwischen dem Deutschen und dem Englischen auf Basis von Mehrebenenannotation und -alignierung automatisch erkannt und beschrieben werden können. Praktische Verwendungsmöglichkeiten wie eine Umsetzung in Transferregeln oder in hypertextuellen Wörterbüchern werden konzipiert und mögliche Ursachen und Implikationen semantischer Divergenzen beleuchtet.