

Humboldt-Universität zu Berlin  
Philosophische Fakultät II  
Institut für deutsche Sprache und Linguistik

**Entwicklung und Evaluierung eines auf dem Stellungsfeldermodell  
basierenden syntaktischen Annotationsverfahrens für Lernerkorpora  
innerhalb einer Mehrebenen-Architektur mit Schwerpunkt auf  
schriftlichen Texten fortgeschrittener Deutschlerner**

Magisterarbeit  
zur Erlangung des akademischen Grades  
Magistra Artium (M.A.) im Fach  
Germanistische Linguistik

Eingereicht von Seanna Doolittle

Wissenschaftliche Betreuerin:  
Prof. Dr. Anke Lüdeling

Berlin, den 22.10.2008

1	Einleitung.....	1
1.1	Allgemein.....	1
1.2	Feldermodell .....	4
1.3	Lernersprachen.....	6
1.3.1	Theorien zum Zweitspracherwerb.....	6
1.3.2	L2-Erwerbssequenzen der deutschen Wortstellung .....	8
1.4	Korpuslinguistischer Hintergrund .....	12
1.4.1	Allgemein .....	12
1.4.2	Annotation von Korpora.....	13
1.4.3	Lernerkorpora .....	21
2	Hauptteil: Annotationsverfahren.....	23
2.1	Beschreibung .....	23
2.1.1	Korpus und Korpusdesign.....	23
2.1.2	Annotation und Korpusarchitektur .....	24
2.2	Kriterien der Feldereinteilung .....	31
2.2.1	Vorfeld .....	31
2.2.2	Bestimmung Mittelfeld, Mittelfeldende, rechte Satzklammer .....	37
2.2.3	Besetzung von Nachfeld und Zuordnung zum Satz .....	40
2.2.4	Ausgewählte syntaktische Phänomene .....	41
2.3	Kriterien für die Entscheidung kanonisch - nichtkanonisch .....	45
2.3.2	Konsistenz: von der Lerneräußerung ausgehen.....	51
2.4	Besprechung einiger Problemfälle der Lernerdaten .....	52
2.4.1	Feldereinteilung .....	53
2.4.2	Kanonisch oder nichtkanonisch .....	55
2.5	Evaluation.....	58
2.5.1	Anwendung der Annotation bei Lernerdaten .....	58
2.5.2	Consistency and Accuracy .....	63
3	Schlussbemerkung.....	67
4	Literaturverzeichnis .....	70
5	Appendix .....	76

Tabelle 1	Feldereinteilung von V1/V2-Sätzen und VE-Sätzen.....	5
Tabelle 2	Konnektorenpositionen vor dem Finitum in V2-Sätzen .....	33
Tabelle 3	Wortabfolge vor dem finiten Verb in V2-Sätzen .....	34
Tabelle 4	rechter Innenrand bei IDS-online (grammis) .....	39
Tabelle 5	Durchschnittswerte aus dem GU-Vergleich.....	61
Tabelle 6	GU Vergleich: Anteil der Konstituentensätze .....	62
Tabelle 7	GU-Vergleich: Verhältnis VE- zu V2-Fehler in absoluten Zahlen.....	62
Tabelle 8	Texte für Inter-Rater-Vergleich .....	64
Tabelle 9	Kappa-Auswertung – Agreement-Rate.....	65
Abbildung 1	Baumdarstellung bei TüBa-D/Z.....	17
Abbildung 2	Baumdarstellung beim TIGER-Sampler.....	17
Abbildung 3	nichtkanonische syntaktische Strukturen bei TüBa-D/Z .....	18
Abbildung 4	„isolated phrase“ (Verbmobil Stylebook, S. 25) .....	19
Abbildung 5	elliptische Sätze, Stylebook der TüBa-D/Z .....	20
Abbildung 6	Überblick der Annotationsebenen bei Falko .....	25
Abbildung 7	Annotationsschema für die Felderannotation .....	26
Abbildung 8	vereinfachte EXMARaLDA-Nachbildung .....	30
Abbildung 9	Satzkoordination mit Zwischenposition bzw. Satzposition.....	36
Abbildung 10	KOORD als eigenes Feldertag.....	37
Abbildung 11	3 Verbgruppen – Kohärenz und Inkohärenz .....	41
Abbildung 12	Verbstellungsabweichungen in der RSK.....	47
Abbildung 13	GU-Vergleich in Prozent.....	61

# 1 Einleitung

## 1.1 Allgemein

Inhalt meiner Magisterarbeit ist die Entwicklung und Vorstellung einer sinnvollen Verfahrensweise für die syntaktische Annotation von Texten von fortgeschrittenen Lernern mit Schwerpunkt auf der Wortstellung. Das Annotationsverfahren soll berücksichtigen können, dass es sowohl wohlgeformte, als auch nicht grammatische und/oder nichtkanonische Äußerungen gibt. Beide sollen suchbar und analysierbar sein.

Eine Mehrebenen-Architektur des Korpusaufbaus macht es möglich, dass das Einfügen von Information (Annotation) in den Lernertext auf mehreren distinkten bzw. verknüpften Ebenen denkbar ist, und dass diese konkreten Bereichen des Textes zugeordnet werden können bzw. dass der Text in bestimmte Segmente eingeteilt werden kann.

Mit Hilfe der Annotation, die auf dem topologischen Feldermodell basiert, soll die Wortstellung im Satz untersuchbar gemacht werden. Gerade um eine deskriptive Beschreibung der linearen Konstituentenabfolge des deutschen Satzes zu geben, hat dieses Modell sich als besonders geeignet erwiesen, weil es das Phänomen der Verbklammer (die nichtlineare Abfolge des Verbkomplexes) im Deutschen berücksichtigt. Es wird sowohl im Bereich von Deutsch als Fremdsprache (DaF) bei der Sprachvermittlung als auch in verschiedenen sprachtheoretischen Ansätzen herangezogen (siehe.1.2 Feldermodell).

In der vorliegenden Arbeit wird eine Besprechung und Evaluierung des Verfahrens anhand von bestimmten Grundsätzen wie Theorienneutralität, Reproduzierbarkeit, Vergleichbarkeit, Eindeutigkeit und Genauigkeit durchgeführt.

Zeitgleich wird an einer verständlichen und gegebenenfalls ausführlichen Dokumentation der Annotation gearbeitet, wie sie jeder brauchbaren Annotation zu Grunde liegt.

Diese Arbeit ist vor allem eine methodische Arbeit, die sich auf einer übergeordneten Ebene mit der Erfassung, Segmentierung und Annotation (dem Zufügen von interpretativen linguistischen Informationen) von kanonischen und nichtkanonischen Sprachdaten befasst. Die darin enthaltene Problematik spielt bei vielen verschiedenen

Typen von Sprachdaten (z.B. gesprochenen und dialektalen Sprachdaten) eine Rolle. Obwohl die Arbeit sich mit fortgeschrittenen Lernerdaten befasst, könnten die Erkenntnisse, die bei der Entwicklung einer Annotation, die diese Tatsache mit berücksichtigt, gewonnen werden, auch für andere Typen von Sprachdaten von Bedeutung sein. Bei der Erforschung von Lernersprachen bzw. Interlanguages (Selinker 1972) ist die Wortstellung (Syntax) ein wichtiger Parameter. In den klassischen Arbeiten zum ungesteuerten Zweitspracherwerb des Deutschen, zum Beispiel bei Clahsen (1984), spielt sie eine entscheidende Rolle in der Beschreibung der dort postulierten obligatorischen Lernerstufen. Es wird angenommen, dass das Erlernen bestimmter syntaktischer Strukturen erst nach dem Erwerb bestimmter Sprachfertigkeiten stattfinden kann (1.3.1). Derartige Annahmen hatten große Implikationen für den Zweitsprachunterricht (Pienemann 1989). Viele weitere Arbeiten zu Lernervarietäten beziehen sich auf den Ansatz der Erwerbsstufen/ Erwerbssequenzen. Vorwiegend basieren diese Untersuchungen auf gesprochenen Daten von nur wenigen Sprachlernern, was Gass und Selinker (2001, S.31) dazu veranlasste, Bedenken an der Repräsentativität dieser Untersuchungen anzumelden. Spätere Untersuchungen fanden auch in gesteuerten Kontexten statt, z.B. die groß angelegte Studie „Deutsch in Genfer Schulen“ (DiGs) (Diehl 2000). Obwohl man insgesamt bemerken kann, dass die syntaktische Erwerbsreihenfolge bei Lernern vergleichsweise intensiv erforscht wurde, weisen alle oben genannten Studien das Defizit auf, dass die Datensätze nicht als öffentliche Korpora zugänglich sind. Auch gibt es insgesamt wenig Untersuchungen zu Lernervarietäten von „fortgeschrittenen“,<sup>1</sup> Sprachlernern (Walter und Grommes 2008).

Eine weitere Möglichkeit zur Untersuchung von Lernerdaten ist die Fehleranalyse, die eng mit dem frühen kontrastiven Ansatz verknüpft wird, der später wegen seines teilweise absoluten Anspruchs zur Erklärung des Spracherwerbs (Interferenz als alleiniger Faktor) auf vielseitige Kritik stieß. Eine differenziertere kontrastive Analyse der Interimsprache (CIA, Granger 2002) im Zusammenhang mit computerlinguistischen Ansätzen hat in den letzten Jahren wieder mehr an Bedeutung gewonnen.

Viele Phänomene in den Sprachdaten von fortgeschrittenen Lernern können nicht mit der Fehleranalyse erfasst werden, da sie grammatisch wohlgeformt sind. Sie weichen

---

<sup>1</sup> wobei es keine einheitliche Definition von Fortgeschrittenheit gibt.

trotzdem von der Zielsprache ab („foreign soundingness“). Dieses Phänomen geht unter anderen auf die Überrepräsentierung („overuse“) oder Unterrepräsentierung („underuse“) von Lexemen, Phrasen aber auch bestimmten syntaktischen Strukturen zurück. Bei vergleichbaren syntaktischen Annotationen von Lernerkorpora und muttersprachlichen Vergleichskorpora, könnte die kontrastive Methode herangezogen werden um Unterschiede bei der Wahl bestimmter syntaktischer Formen und textrelevanter Wortstellungsregularitäten zu untersuchen. Dazu gehören Fragen wie: Welche Elemente stehen am Satzanfang (im Vorfeld) und sind damit z.B. im Fokus, bzw. welche Äußerungen werden in Form von Nebensätzen dargestellt und wie ist ihre Stellung im Satz?

Unter Einbindung in den theoretischen Diskurs wird im Hauptteil der Arbeit das Annotationsschema dargestellt und dessen Richtlinien, die so einfach wie möglich und so kompliziert wie nötig sein sollen, diskutiert (Abschnitt 2.2).

Die Felderannotation wurde an zwei Subkorpora des frei zugänglichen fehlerannotierten Lernerkorpus (Falko)<sup>2</sup> durchgeführt. Falko ist das gemeinsame Projekt von der Freien Universität (FU) und der Humboldt-Universität (HU) in Berlin, ein fortgeschrittenes Lernerkorpus des Deutschen zu entwerfen und aufzubauen (Lüdeling et al. 2008). Mit dieser Form der Annotation sind nicht nur Sätze und Felder suchbar, sondern unter Einbeziehung einer Wortart-Annotation können auch weitere Eigenschaften der Lernertexte auf der Satz- und Felderebene untersucht werden: z. B. Komplexität anhand der Satzlänge und Anzahl und Art der Konstituentensätze/Nebensätze, bzw. infinite Nebensatzähnliche Strukturen sowie Komplexität und Besetzung der topologischen Felder. Wenn Lernerätze mit dem Feldermodell nicht beschrieben werden können, werden sie zunächst als nicht einordenbar gekennzeichnet. Die Mehrebenen-Architektur von Falko ermöglicht es, dass die Analyse der linearen Reihenfolge dieser Sätze dann z. B. mit Hilfe der Wortart-Abfolge in Zusammenhang mit der Fehlerannotation bzw. mit der Konstituentenabfolge weiter untersucht werden kann.

Ein wichtiger Teil der Arbeit ist die Evaluation der Annotationsverfahren, sowohl anhand einer Problembesprechung, in der Ambiguitäten und Zweifelsfälle systematisch erläutert werden, als auch unter Verwendung von quantitativen

---

<sup>2</sup> <http://www2.hu-berlin.de/korpling/projekte/falko/>

Verfahren wie einem k-Wert-Interratervergleich (Carletta 1996), siehe hierzu Abschnitt 2.5.2.

Als weiteres quantitatives Verfahren verwende ich die longitudinalen Daten GU2, GU3 und GU4 der Georgetown University als Basis, um für die drei Sprachstufen eine Statistik aufzustellen über die Anzahl erfassbarer, dem Feldermodell konformer Strukturen gegenüber den nicht erfassbaren Strukturen und so die Anwendbarkeit des Feldermodells für verschiedene Sprachstufen zu überprüfen (Abschnitt 2.5.1).

Zusammenfassend beschäftige ich mich in meiner Arbeit u.a. also mit folgenden Fragestellungen: Wie kann eine größtmögliche Menge an Sprachdaten syntaktisch eingeordnet bzw. erfasst werden? Was für einen Zugewinn bringt die Annotation gegenüber dem nicht annotierten Korpus? Welche Möglichkeiten gibt es, Sprachdaten syntaktisch zu annotieren? Welche Kriterien werden für die Annotation aufgestellt und in wie weit können sie bzw. werden sie eingehalten?

## 1.2 Feldermodell

Die Beschreibung der linearen Abfolge des deutschen Satzes anhand von topologischen Sequenzen beruht auf einer langen Tradition<sup>3</sup>. Drach (1937) benennt als Erster diesen Ansatz als Stellungsfeldermodell und beschreibt jeweils ein Feld vor und nach dem finiten Verb, das je nach kommunikativer Absicht besetzt werden kann. Diese kommunikativ-pragmatische Herangehensweise spiegelt die Tatsache wider, dass in Abweichung zu analytischen bzw. konfiguralen Sprachen<sup>4</sup>, bei denen die starre Wortstellung die grammatische Funktionalität der Konstituenten bestimmt, die lineare Abfolge des Deutschen von pragmatischen Überlegungen stark beeinflusst wird.

Das Feldermodell wurde weiterentwickelt und diente auch als Grundlage für die Arbeiten von Bech (1955) über den Verbalkomplex, Engel (1970) und Höhle (1986) über das Mittelfeld. Die Grundidee hinter dem Stellungsfeldermodell ist die Einteilung des Satzes in Felder, ausgehend von der Position der verbalen Elemente (der linken und rechten Satzklammer). Dieser Ansatz ermöglicht eine Generalisierung für die

---

<sup>3</sup> Herling (1821) und Erdmann (1886)

<sup>4</sup> Die analytische und synthetische Typologie bezieht sich stärker auf die morphologische Eigenschaften der Sprache (nicht flektierbar vs. flektierbar) während die Einteilung in „configurational“ und „non-configurational“ Sprachen aus der generativen Forschung stammt und sich stärker auf syntaktische Eigenschaften bezieht (Legate 2001). Im Deutschen wird die relativ freie Wortstellung allerdings durch morphologische Mittel realisiert.

Beschreibung von deutschen Sätzen. Anhand der Verbstellung werden sie in drei Formtypen eingeteilt.

Bei Verberst- und Verbzweitsätzen (V1, V2) bildet das finite Verb (V<sub>fin</sub>) die linke Satzklammer (LSK) und die nicht finiten verbalen Elemente die rechte Satzklammer (RSK). Bei Verbendsätzen (VE) formt der Nebensatzeinleiter<sup>5</sup> die linke<sup>6</sup> und die verbale Phrase die rechte Satzklammer.<sup>7</sup>

	Vorfeld	LSK	Mittelfeld			RSK	Nachfeld	
V1 & V2	Er	hat	mehr	Äpfel		gegessen	als	wir
VE		dass	er	mehr	Äpfel	gegessen	hat	als wir

Tabelle 1 Feldereinteilung von V1/V2-Sätzen und VE-Sätzen

So kann die deutsche Satzkonstruktion auf nur wenige strukturelle Konzepte zurückgeführt werden.

Nicht alle Felder müssen realisiert werden. Bei V2 sind Vorfeld (VF) und LSK, bei V1 eigentlich nur die LSK, bei VE die LSK und RSK obligatorisch.

Das Stellungsfeldermodell ist geeignet, die Besonderheiten des Deutschen darzustellen:

- die nichtlineare (diskontinuierliche) Stellung von Konstituenten wie z. B. der Verbphrase (VP)
- eine relativ freie Wortstellung
- die unterschiedliche Wortfolge bei Haupt- und Nebensätzen:  
im Allgemeinen wird SOV als die zugrunde liegende Konstituentenabfolge des Deutschen angesehen, wie sie in eingeleiteten Nebensätzen vorkommt. Ein weiteres Merkmal ist ihr V2-Charakter (z.B. bei den meisten Aussagesätzen<sup>8</sup>), wobei nicht SVO, sondern XVO (SVO/OVS/AVS) gemeint ist.

<sup>5</sup> Als Nebensatzeinleiter werden hier alle Verbend-bedingende Elemente bezeichnet, unabhängig von ihrem Status im Satz; Z.B. Relativpronomen besitzen im Gegensatz zu subordinierenden Elemente wie „dass“ Satzgliedfunktion.

<sup>6</sup> Ein konzeptionelles Problem entsteht bei der Begrifflichkeit „linke Satzklammer“, die sowohl für den subordinierenden Konjunktoren in Verbletztsätzen als auch für das finite Verb in V2- und V1-Sätzen steht. Es wird eine gemeinsame syntaktische Funktion (Kategorie) suggeriert. Es gibt unterschiedliche Auffassungen dazu, ob beide Elemente als Kopf des Satzes gesehen werden können. In der generativen kombinierten Darstellung des X-bar-Schemas und Feldermodells in Grewendorf et al. (1999, S. 213–227), wird die „LSK“ C<sup>0</sup>, dem Kopf des Satzes, gleichgesetzt. Andere Ansätze dagegen definieren einen Satzkopf immer mit dem Merkmal „Finitheit“ Pasch (2003, S. 83; Brandt et al. 1992).

<sup>7</sup> In folgenden wird die Satzklammer der V1/V2-Sätze als verbale Klammer und die der Verbend-Sätze als Nebensatzeinleiter-Verbklammer bezeichnet.

<sup>8</sup> Wobei es keine 1:1-Entsprechung auf der semantischen Seite gibt. Das heißt, nicht alle V2-Sätze sind Aussagesätze. Umgekehrt gilt auch, dass wenn auch die meisten Aussagesätze V2-Sätze sind, sie auch anders (vor allem mündlich) realisiert werden können.

Bei dem Feldermodell handelt es sich um ein Satzmodell für Sprachen mit flexibler Wortstellung, das ein Gerüst für die variablen Konstituenten des deutschen Satzes bildet. Es unterscheidet sich von anderen Beschreibungsansätzen vor allem darin, dass Mittelfeld (MF) und Nachfeld (NF) als separate Einheiten gesehen werden, auch beim Nichtvorhandensein von trennenden Elementen (Reis 1980, „zero boundary“). Daraus ergibt sich, dass das MF als eine natürliche Klasse mit eigenen Wortstellungsregularitäten aufgefasst wird, die sich nicht auf den gesamten Satz beziehen.

## **1.3 Lernaltersprachen**

### **1.3.1 Theorien zum Zweitspracherwerb**

Die heutige Zweitsprachenforschung<sup>9</sup> reicht u. a. zurück auf Weinreichs (1953) Studien zu Bilingualismus und auf Lados (1957) unterrichtsbezogenen Ansatz der „Contrastive Analysis“. Analog zum Erstspracherwerb war Lados Zweitspracherwerbstheorie von dem damaligen Ansatz des Behaviorismus geprägt. Spracherwerb wurde als ein kognitiver Lernprozess und Fehler als Übertragung (transfer) bzw. Interferenz aus der Muttersprache gesehen. Infolgedessen wurden die Entwicklungen im Spracherwerbsprozess allein auf die Unterschiede und Ähnlichkeiten zwischen der Muttersprache (L1) und der zu erlernenden Zielsprache (L2) zurückgeführt. Mit der „kognitiven Wende“ (u.a. Chomsky 1959) wurden die bis dahin vorherrschenden Ansätze in Frage gestellt. Für den Spracherwerb werden ein angeborenes mentales Spracherwerbsmodul und eine „universale Grammatik“ (UG) als Erklärungsansatz herangezogen. In dieser Tradition steht die „Identity Hypothesis“, deren starken Interpretation nach der L2- und L1-Erwerb identisch verlaufen (Corder 1967; Dulay und Burt 1974).

Heute gilt als allgemein unstrittig, dass L2-Erwerb in bestimmten grammatischen Teilbereichen (z.B. bei der Wortstellung) in Phasen verläuft. Jede dieser Phasen wird als ein spezifisches Entwicklungsstadium und Lernaltersprachen als „Interlanguages“, mit eigener Logik und Systematik (Selinker 1972), gesehen. Sie durchlaufen eine chronologische Reihenfolge, sowohl bei Kindern als auch Erwachsenen. Die „Erwerbssequenz“ entspricht der schrittweisen Erschließung der Zielsprache.

---

<sup>9</sup> Zweitspracherwerb ist hier zunächst als Oberbegriff gemeint und umfasst sowohl gesteuerten (Fremdspracherwerb) als auch ungesteuerten Spracherwerb, auch L2 bis Ln (Zweit-, Drittspracherwerb usw.). Zunächst wird nicht zwischen „lernen“ und „erwerben“ unterschieden.

Auch wenn der L2-Erwerbsablauf Ähnlichkeiten mit der L1-Spracherlernung von Kindern aufweist, so ist eine weitere allgemein akzeptierte Auffassung, dass es wesentliche Unterschiede gibt: Kinder erlangen volle Kompetenz<sup>10</sup> ihrer Muttersprache, während Lerner einer Zweitsprache sehr individuelle und unterschiedliche Sprachendstadien erreichen. Ein Stillstand vor dem Erreichen des vollständigen L2-Erwerbs nennt man „Fossilisierung“ (Selinker 1972). Es handelt sich dabei um ein häufiges Merkmal des L2-Erwerbs, der weniger kontinuierlich und unsystematischer abläuft und regressions- und störungsanfälliger ist, wobei es auch Interferenz der L1 gibt .

Umstritten bleibt die wissenschaftliche Erklärung für diese Beobachtungen. Im Folgenden werden einige Ansätze zur Erklärung der internen Faktoren<sup>11</sup> des Spracherwerbs vorgestellt. Grundsätzlich wird unterschieden zwischen dem oben schon erwähnten mentalistischen und dem kognitivistischen Ansatz. Die wissenschaftliche Tendenz geht dahin, die verschiedenen Ansätze als ergänzend anzusehen, was somit wiederum einer strikten Trennung der Modelle entgegensteht und multifaktorelle Erklärungsansätze fördert (siehe Diehl et al. 2000, S. 43–44)<sup>12</sup>.

Die mentalistischen Theorien unterscheiden sich u.a. in Bezug auf den Zugang (Access) der Lerner zur UG, sowie das Ausmaß (full transfer, partial transfer) und die Form (minimal trees) des sprachlichen Transfers aus der Muttersprache. Beispiele hierfür sind „Minimal Trees“ (Vainikka und Young-Scholten, 1996) und „Full Access, Full Transfer“ (Schwartz und Sprouse, 1996). Bei den kognitiven Erklärungsmodellen dagegen liegen die Schwerpunkte auf Prozessen und Strategien (Ellis 1994). Einige Ansätze legen ihren Schwerpunkt auf die Sprachverarbeitung z.B. Clahsen (1984). Demnach wird der L2-Input basierend auf allgemeinen universellen (angeborenen) Sprachverarbeitungsstrategien (Operating Principles, Slobin 1973) erschlossen. Die konnektionistischen Modelle erklären den Spracherwerb nur durch allgemeine Lernstrategien und Input ohne Bezug auf angeborene Sprachverarbeitung bzw. -wissen. Beispielhaft sei hier das Competition-Modell genannt, das von „cues“ (Sprachaktivierung) ausgeht (MacWhinney et al. 1989). Hier spielen Faktoren wie Erkennbarkeit, Markiertheit, Frequenz (Häufigkeit des Vorkommens in Input und

---

<sup>10</sup> Kompetenz (Strukturwissen) bezieht sich in diesen Zusammenhang nicht z.B. auf Wortschatzkompetenzen oder auf die grammatische „Richtigkeit“ in Bezug auf eine Hochsprache.

<sup>11</sup> Damit sind lernerinterne Mechanismen gemeint wie der Ablauf des Lernprozesses, linguistisches Wissen, Kommunikationsstrategien und L1-Transfer (Ellis 1994).

<sup>12</sup> Die hier vorgenommene Auswahl und Zuordnung lehnt sich an an Diehl et al. (2000) und an Krohn und Krohn (2008).

Output) und Abrufbarkeit (availability) eine maßgebliche Rolle. Das „Parallel-Distributed-Processing Model“ (PDP, Seidenberg und McClelland 1989) versteht den Prozess des Spracherwerbs als Aufbau von Netzwerken (Systemen mit interaktiven Einheiten), die durch Aktivierung bzw. Hemmung von Knoten entstehen. Computersimulation dient als Evidenz für die Validität der Modelle.

Andere Ansätze vereinen die konnektionistischen und regelbasierten Theorien. Zum Beispiel Pinker und Price (1992) erklären die Erlernung irregulärer Flexion durch Netzwerkaktivierung und reguläre Flexion, „Default“-Flexion, durch Regelwissen.

Die Forschung befasst sich auch mit einzelnen affektiven Faktoren, wie interlingualer Interferenz (z.B. von L2 auf L3) oder intralingualer Interferenz (Untersuchung der Zielsprache in Bezug auf den L2-Erwerb).

L2-Erwerb ist nicht allein auf interne Faktoren zurückzuführen, wie auch die vielen Studien und Untersuchungen mit Lernern belegen, zum Beispiel zu Motivation, Einstellung oder den externen affektiven Faktoren des L2-Erwerbs (wie dem kulturellen und gesellschaftlichen Kontext, der wiederum auch Einfluss auf den Lerner hat<sup>13</sup>). Auch aus fachdidaktischen Gründen werden außerdem der Einfluss des Unterrichts und die Produktionsart (Übersetzung, freies Schreiben) erforscht<sup>14</sup>.

### **1.3.2 L2-Erwerbssequenzen der deutschen Wortstellung**

Unter Erwerbssequenzen versteht man die Phasen, die Lerner beim Erwerb einer Fremdsprache durchlaufen. Eine frühe Studie zum Wortstellungserwerb im Deutschen ging aus dem Zisa-Projekt hervor, (Clahsen et al. 1983). Diese longitudinale Studie<sup>15</sup> zu kindlichem L2-Erwerb bildet die Basis für die Erwerbssequenzforschung. In der Folge sind vor allem Querschnittsstudien (cross-sectional) durchgeführt worden, die sich vorwiegend mit den Erwerbssequenzen bei Erwachsenen und gesteuerten Erwerbssituationen befassten. Als Grundlage für die Untersuchungen von Ellis (1989), Pienemann (1989), Meerholz-Härle (2001) und Jansen (2008) dienten gesprochene Daten. Dagegen bildeten schriftliche Texte (freies Schreiben) die Basis für Diehls (2000) Untersuchung frankophoner Genfer Schüler in einer gesteuerten Lernsituation.

---

<sup>13</sup> Für einen Überblick zur Zweitspracherwerbsforschung des Deutschen mit Schwerpunkt auf Untersuchungen von sprachexternen Faktoren siehe Moyer (2004)

<sup>14</sup> Für einen Überblick über die L2-Forschung siehe z.B. Ellis (1994) oder Edmondson und House (2000).

<sup>15</sup> 16 Kinder mit romanischen Muttersprachen. Die Daten basieren auf ungesteuerten, spontanen mündlichen Erhebungen.

Basierend auf diesen Arbeiten hat sich folgende relativ robuste Sequenz sich in der Literatur durchgesetzt<sup>16</sup>:

(Stufe 0 Einwort-Konstituenten)

Stufe I Kanonische Wortstellung (SVO, SOV, ..)

Feste Wortabfolge unabhängig von syntaktischen Eigenschaften der L1

(1) *die kinder spielen mit ball*

Stufe II ADV: Fokusposition

Voranstellung Adverbiale (Adverbien und Präpositionen) vor die kanonische Wortabfolge

(2) *da kinder spielen*

Stufe III SEP:

Satzfinale Positionierung nichtfiniter Verbelemente: trennbare Verbpartikel, infinitive Modalkonstruktionen, Partizipien in Hilfsverbskonstruktionen

(3) *alle kinder muss die pause machen*

Stufe IV INV: Verbzweit (SV→VS)

Subjekt-Verb-Inversion, im VF können Konstituenten außer dem Subjekt stehen; Subjekt folgt direkt auf V<sub>fin</sub> (Frage-Sätze und vorangestellte Komplimente)

(4) *dann hat sie wieder die knoch gebringt*

Stufe V V-END:

Endstellung des V<sub>fin</sub> in subordinierten Sätzen

(5) *guck was ich in meine tasche hab*

Es gibt verschiedene Erklärungsansätze für die hier aufgeführten Erwerbssequenzen in der Zweitsprache. Clahsen geht von einem fundamentalem Unterschied beim L2- und L1-Erwerb aus („Fundamental Difference Hypothesis“)<sup>17</sup>, bei dem L2 keinen Zugriff auf die UG hat. Die Wortstellung und die Erwerbssequenz werden mit allgemeinen semantischen Prinzipien (Operating Principles) erklärt. Nach Slobin (1982) entspricht die Abfolge von Subjekt-Verb-Objekt in der Stufe I der semantischen Realisierung von Agens-vor-Aktion-vor-Patiens, die er als neutraler Satztyp beschreibt. Deshalb ist nach Clahsens (1984) Argumentation der Verarbeitungsaufwand für solche Strukturen geringer und sie werden früher vom Lernenden erworben, was sich

---

<sup>16</sup> Die aufgeführten Beispiele stammen aus der ZISA-Studie

<sup>17</sup> siehe Pienemann, M.: An introduction to Processability Theory. based on an extended and revised version of paper "Developmental dynamics in L1 and L2 acquisition: Processability Theory and generative entrenchment. in Bilingualism: Language and Cognition (1998), 1.1, pp 1-20. Online verfügbar unter <http://www.uni-paderborn.de/fileadmin/kw/Institute/Anglistik-Amerikanistik/Personal/Pienemann/INTRO.NEW.pdf>, zuletzt geprüft am 15.10.2008.

in der Produktion niederschlägt. Im Weiteren werden Prinzipien wie Salienz (Strukturen in hervorgehobenen Positionen sind leichter zu verarbeiten, (ADV)) und Konstituentenzusammenhalt (die Trennung vom Konstituent bedeutet einen größeren Verarbeitungsaufwand (SEP)) zur Erklärung der Erwerbssequenzen hinzugezogen.

Dagegen sieht Pienemann die „Processability-Theorie“ (PT, Pienemann 1998) als einen Erklärungsansatz für den Sequenzablauf von L1 und L2. Nach der PT durchläuft jeder Lerner die gleichen Phasen und bestimmte Strukturen müssen erst erworben werden, bevor andere Strukturen erlernt werden können. Darüber hinaus wird für den L2-Erwerb von einem „partial transfer“-Ansatz, der „Developmentally Moderated Transfer Hypothesis“ (Pienemann et al. 2005) ausgegangen. Sie geht von einem angeborenen typologischen und psychologischen Sprachverarbeitungsmodell aus, das auf den Grundlagen der "Lexical Functional Grammar“ (LFG) und dem daraus abgeleiteten Prinzip der „Lexical Mapping Theory“ (Bresnan 2001) basiert.

Die empirischen Untersuchungen von Vainikka und Young-Scholten (1994) widerlegten die von Clahsen (1984) postulierten SVO-Abfolge. Sie zeigen, dass Lerner mit SOV-Muttersprachen wie Türkisch und Koreanisch SOV-Strukturen in der kanonischen Phase bilden. Inzwischen wird zwar von einer kanonischen Wortfolge in der Stufe I ausgegangen, die aber nicht unbedingt SVO sein muss. Die Erforschung dieser initialen Sprachphase steht immer wieder im Mittelpunkt des Forschungsinteresses, wie auch die Frage, welchen Einfluss L1 auf sie hat - z. B. „full transfer“ von Anfang an (Schwartz 1996) oder „transfer“, erst dann, wenn die übertragenen Strukturen aus der Muttersprache verarbeitet werden können (Pienemann und Håkansson 2007, S. 486; Pienemann et al. 2005).

In der „minimal trees“-Hypothese (Vainikka und Young-Scholten 1996) wird davon ausgegangen, dass im „initial state“ nur lexikalische Projektionen, VP, NP und AP (ohne Finitheit, Kongruenz, Tempus) vorhanden sind und deren Kopfpositionen von der jeweiligen Muttersprache beeinflusst werden. Nach und nach werden durch Input der Fremdsprache die funktionalen Phrasen herausgebildet. „Full Transfer“ dagegen geht davon aus, dass die funktionalen Projektionen CP (complementizer phrase) bzw. IP (inflectional phrase) der L1 ohne lexikalische und phonetische Merkmale voll übertragen und durch Input bzw. durch die UG umstrukturiert werden.

Weitere Ansätze zur Beschreibung des frühen Spracherwerbs finden sich in Arbeiten zu Basisvarietäten (Klein und Perdue 1997), die in der Tradition der

Interimsprachenforschung einzuordnen sind. Es handelt sich hier um sehr detaillierte Beschreibungen der Lerner-Basisvarietät zu Lexik, Morphologie, und Wortstellung. In der Basisvarietät (die ungefähr der kanonischen Wortstellung entspricht) werden Äußerungen durch semantische (controller (Agens) first) und pragmatische (Topik vor Fokus) Prinzipien strukturiert, es gibt weder Finitheit noch Flexion. Erst in den Postbasisvarietäten wird Finitheit erworben und es erfolgt eine stärkere Orientierung an die zielsprachliche Strukturen.

Nicht alle empirische Daten bestätigen das von der PT für alle Stufen angenommene Prinzip, das besagt, dass erst nach Erwerb der Struktur einer vorhergehenden Phase überhaupt die nächste Struktur verarbeitet werden kann, besonders bezogen auf den Erwerb von Verbend (V-END)-Strukturen erst nach dem Erwerb von Verbzweit (V2)-Strukturen. Die Untersuchung von Meerholz und Tischner (2001) lässt begründete Zweifel zu, da drei ihrer Probanden Nebensätze mit Verbendstelle, aber keine Sätze mit Inversion gebildet haben. Aus den Daten von Diehl (2000; S.110) geht sogar hervor, dass Lerner zuerst V-END erwerben, noch bevor sie den Infinitiv (INV) in Aussagesätzen bilden, wobei sie nicht das gleiche Auswertungsverfahren anwendet<sup>18</sup>. Auch Jansen (2008; S.217), die in ihrer Querschnittsstudie mit dem gleichen Auswertungsverfahren<sup>19</sup> wie Pienemann (1998) arbeitet und die gleichen Erwerbssequenz-Pattern erhält, kann, in Bezug auf den Übergang von ADV nach SEP, sowie INV nach VEND, nicht beweisen, dass die Strukturen aus den vorhergehenden Phasen nicht erworben worden sind. Sie kann Erwerb nachweisen, aber nicht ausschließen, wie die Theorie es verlangt, dass bestimmte Strukturen nicht produziert werden können bevor andere erworben worden sind. Leider ist durch die unterschiedlichen Vorgehensweisen in der Auswertung ein direkter Vergleich der Ergebnisse der genannten Untersuchungen nicht möglich. Erschwerend kommt dazu, dass sie nicht öffentlich zugänglich sind.

---

<sup>18</sup> Sie errechnet einen Korrektheitsquotient. Erst wenn die Mehrheit der Schüler eine Satzstruktur weitgehend fehlerfrei produzieren kann, gilt die Struktur als erworben (ADV wird nicht untersucht). Dieses Vorgehen begründet sie dadurch, dass beim gesteuerten Spracherwerb nicht ausgeschlossen werden kann, dass das Auftreten der Satzstrukturen aufgrund memorisierter Muster aus dem Unterricht geschieht.

<sup>19</sup> „Emergence“, das erste „korrekte“ Vorkommen in vier möglichen Kontexten, wird als Grundlage für die weiteren Berechnungen bei den Querschnittsstudien mit „implicational scaling“ verwendet.

## 1.4 Korpuslinguistischer Hintergrund

### 1.4.1 Allgemein

Die Anwendung von korpusgenerierten Daten erweitert die Datenbasis und damit auch die Methodik zur Erforschung von Sprachen. Sie stellt eine wichtige Ergänzung zur Introspektion und experimentellen Datenbeschaffung dar und liefert als einzige Methode die Möglichkeit, anhand einer größeren Datenbasis Aussagen über „frequency“, die statistische Häufigkeit von bestimmten linguistischen Strukturen, zu treffen (McEnery und Wilson 1996, S. 12).

Korpora sind nach der EAGLES Definition<sup>20</sup> in erster Linie Sprachsammlungen, die nach expliziten linguistischen Kriterien ausgesucht und geordnet wurden, um als Sprachmuster genutzt zu werden. Ihr Anwendungsbereich und Nutzen hängt ab von der Zusammenstellung, der Architektur bzw. dem Aufbau (Dateistruktur, Art der metalinguistischen Annotation und der elektronischen Zusammenführung der Dateien) und von den technologischen Möglichkeiten, sie zu durchsuchen und zu analysieren.

Sie können aus gesprochenen und/oder schriftlichen Komponenten bestehen. Die Zusammenstellung der Korpora hängt von verschiedenen Aspekten ab, welche vorwiegend inhaltlicher Natur sind, wie z.B. der zugrunde liegende Forschungs- oder Erkenntniszweck. Viele große bekannte Korpora wie das monolinguale englische British National Corpus (100 Millionen Wörter)<sup>21</sup> werden als eine Art „Repräsentation“<sup>22</sup> der Gesamtsprache zusammengestellt. Dabei ist das Ziel, möglichst viele und unterschiedliche schriftliche und gesprochene Daten mit unterschiedlichen Stilen, Varietäten, Fachbereichen, Genres und Registern zusammenzutragen. In vielen Fällen ist es auch möglich, mit Hilfe von Metadaten Teilkorpora zu erstellen. Metadaten sind Daten, die u.a. Hintergrundinformation über Art, Entstehung, und Größe der einzelnen Texte (Spracheinträge) geben.

Es gibt auch spezielle Korpora, die für die Erforschung spezifischer Fragestellungen zusammengestellt werden. Dazu werden auch die Lernerkorpora gerechnet (siehe Abschnitt 1.4.3, S. 21)

---

<sup>20</sup><http://www.ilc.cnr.it/EAGLES96/corpintr/node13.html>

<sup>21</sup>Die größten deutschen Korpora sind die Korpora des IDS Mannheim (<http://www.ids-mannheim.de/kl/projekte/korpora/>) <http://www.ids-mannheim.de/cosmas2/uebersicht.html>) und das „Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts“ (<http://www.dwds.de>).

<sup>22</sup>Repräsentivität ist ein problematisches Konzept. Man muss sich bewusst sein, dass es keine tatsächliche Repräsentivität geben kann. Die Idee hinter diesen großen Korpora ist, eine gewisse Ausgewogenheit herzustellen, um domainspezifischen Phänomenen entgegenzuwirken. Letztendlich kommt es darauf an, dass die Parameter klar definiert sind und dass diese bei der Bewertung der Daten berücksichtigt werden.

## 1.4.2 Annotation von Korpora

Korpusannotation ...

*[...] is the practice of adding **interpretative** linguistic information to a corpus.*

*(Leech 2005)*

Egal wie akzeptiert oder allgemein anerkannt die hinzugefügte Information zu sein scheint, es handelt sich immer um eine linguistische Interpretation. Diese Feststellung führt direkt zu der Grundsatzfrage: soll überhaupt mit annotierten Korpora gearbeitet werden oder sollen nur nichtannotierte Korpora (raw corpora) als Ausgangspunkt für Untersuchungen dienen?

Sinclair (2004) vertritt die Auffassung, dass Annotation ein Informationsverlust und sogar eine Fehlerquelle darstellt. Diese Argumente sind nicht einfach von der Hand zu weisen, und die Verwendung von Daten, die auf Annotationen basieren, muss immer unter Berücksichtigung dieser Tatsachen geschehen. Eine Analyse der Annotationsdaten sollte also nicht ohne ein klares Verständnis dafür, was und wie eigentlich annotiert worden ist, durchgeführt werden. Dies wiederum verdeutlicht, worüber die jeweiligen Daten Aussagen machen können oder auch nicht.

Auf der anderen Seite sind annotierte Korpora nützlich, sie erleichtern die Suche nach bestimmten Phänomenen oder machen die Suche überhaupt erst möglich. Neben dem reinen Informationsgewinn können Annotationen auch wiederverwendet werden und ermöglichen somit die Vergleichbarkeit und Transparenz (Reproduzierbarkeit) von empirischen Untersuchungen. Und sie können auch, je nach Forschungsfrage, für andere Zwecke eingesetzt werden (Garside 1997, S. 4–5).

### 1.4.2.1 Richtlinien für die Korpusannotation

Wie nützlich eine Annotation sein kann, hängt von verschiedenen Faktoren ab (Leech 2005):

1. ob die Annotation und das „raw“-Korpus von einander getrennt genutzt werden können,
2. wie gut sie dokumentiert wurde zur Nachvollziehbarkeit für ihre Nutzer, für die Transparenz der mit ihr erstellten Analysen sowie für weitergehende Annotationsvorhaben,

3. in wie weit die in der Annotation benutzten linguistischen Kategorien allgemein anerkannt sind und somit als Hilfsmittel für die Forschungsgemeinschaft dienen können,
4. von der Qualität der Annotation. Dies betrifft die Fragen, ob sie sinnvolle linguistische Kategorien verwendet, klare Kriterien aufstellt und wie einheitlich (consistency) und wie genau (accuracy) annotiert worden ist.
5. und schließlich auch von der Vergleichbarkeit mit anderen annotierten Korpora.

Im Mittelpunkt dieser Arbeit stehen die letzten drei Punkte. Zum einen geht es darum, die Kriterien für die Zuordnung in das Feldermodell, das als Basis für die Annotation dient, zu diskutieren. Das Feldermodell ist ein anerkanntes und weit verbreitetes deskriptives Modell, das nur minimale Annahmen über syntaktische Strukturen trifft und somit die in Punkt 3<sup>23</sup> genannten Richtlinien erfüllt. Die Minimalität hat eine besondere Bedeutung für Lernerstrukturen, weil ihre Funktionen unspezifischere, syntaktische bzw. morphosyntaktische Realisierungen aufweisen als in der Zielsprache, deshalb können weniger restriktive Beschreibungsansätze mehr Lernerstrukturen erfassen. Andererseits stellen minimale Festlegungen bzw. Annahmen Probleme für eine konsistente Annotation dar. Um diesen zu begegnen, müssen Festlegungen für unklare Fälle getroffen werden<sup>24</sup>. Abgesehen von diesen unklaren Fällen stellt sich die Frage, in wie weit es allgemeine Übereinstimmung gibt, wie die topologischen Felder zu bestimmen sind. Deshalb werden sowohl die Lehrmeinungen zur Einteilung der Felder verglichen als auch Übereinstimmungen anhand von Inter-Rater-Daten untersucht. Die Festlegung der Felder und die Aufstellung von klaren Kriterien zu der Anwendbarkeit des Modells (Kanonizität) haben ferner Auswirkung auf die in Punkt 4 erwähnte Qualität einer Annotation.

Auch eine Vergleichbarkeit mit anderen Annotationen ist anzustreben. Was das hier genau bedeutet, wird im weiteren untersucht.

Für die Aufstellung der Annotationsregeln ist Konsistenz ein wichtiges Entscheidungskriterium. Um eine einheitliche Annotation auch bei verschiedenen Annotatoren zu ermöglichen, und in Hinblick auf mögliche Automatisierungsprozesse

---

<sup>23</sup> In diesem Zusammenhang wird auch von Theorienneutralität gesprochen. Diese Begrifflichkeit ist problematisch, da jedes Modell eine Theorie impliziert.

<sup>24</sup> „An annotation scheme can additionally make explicit how the annotations apply to the 10% or so of less clear cases, so that users will know how borderline phenomena are handled.“ (<http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm>, zuletzt geprüft 20.10.2008)

(Parsing), soll der Ansatz „so einfach wie möglich und so kompliziert wie nötig“ befolgt werden, Dabei müssen Faktoren wie Suchbarkeit und adäquate Beschreibungen von Sprachstrukturen berücksichtigt werden.

Schließlich soll in diesem Zusammenhang die Wichtigkeit einer Kodierung, die auf allgemeine Korpusstandards<sup>25</sup> aufbaut, erwähnt werden, damit Korpora z.B. mit allgemein verfügbaren Tools bearbeitet bzw. durchsucht werden können.

#### **1.4.2.2 Typen der linguistischen Annotation**

Annotationen können entweder automatisch, semiautomatisch oder manuell erstellt werden und können verschiedene linguistische Information zuweisen.

Diese Information kann entweder als Headerinformation (meistens Metadaten), die sich auf die gesamte Datei oder den gesamten Text bezieht, als tokenbasierte Information<sup>26</sup> oder als strukturelle Information abgelegt werden.

Die häufigsten Formen tokenbasierter Annotationen sind Wortart und Lemma, die in der Regel automatisch durchgeführt werden. Für das Deutsche werden meist der Tree-Tagger der Universität Stuttgart<sup>27</sup> und das „Stuttgart-Tübingen Tagset“ (STTS, Schiller et al. 1999) benutzt. Viele Korpora werden auch mit morphologischen und funktionalen Informationen versehen. Außerdem gibt es semantische, pragmatische und diskursbasierte Angaben. Gesprochene Daten können außerdem auch phonologisch annotiert werden. Für einen Überblick siehe Leech (2005).

Der Schwerpunkt dieser Arbeit liegt auf der syntaktischen Annotation, einer Form struktureller Informationszuweisung, die häufig als „Treebanking“ realisiert wird: Sätze können in syntaktisch relevanten Einheiten, z.B. Felder, Konstituenten und/oder Phrasen segmentiert und hierarchisch dargestellt werden, wobei der Grad der Hierarchisierung vom Annotationsschema und dessen technischem Aufbau abhängt. Diese Einheiten werden dann in der Regel als Bäume dargestellt. Siehe Beispiele in Abschnitt 1.4.2.3, Syntaktische Annotierung deutscher Korpora.

Zur Automatisierung der syntaktischen Annotation/Segmentierung werden gerne „Natural Language Processing“ (NLP)-Parser eingesetzt. Sie werden „trainiert“ auf ein manuell annotiertes Korpus, das als Goldstandard<sup>28</sup> gilt. Es gibt viele unterschiedliche

---

<sup>25</sup> z.B. die Standards der Text Encoding Initiative (<http://www.tei-c.org>).

<sup>26</sup> bei Token handelt es sich um segmentierte Texteinheiten der Wortebene.

<sup>27</sup> <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

<sup>28</sup> Dieses Verfahren wird auch bei anderen automatischen Annotationen angewandt, z.B. dem Wortart-Tagger.

Parser, im Prinzip erkennen sie die Muster (Sprachregeln) in dem Korpus, auf das sie trainiert werden, und segmentieren unbekannte Sprachdaten entsprechend. Um bessere Ergebnisse zu erzielen, werden z.B. probabilistische<sup>29</sup>, lexikalische oder syntaktische Informationen hinzugefügt.

Auch wenn ich hier nicht näher darauf eingehen kann<sup>30</sup>, soll zumindest erwähnt sein, dass Annotationsschemata auf ihre „Parsebarkeit“ oder Automatisierungseigenschaften untersucht werden, um zu beurteilen, ob sich die Korpora als Goldstandard in Hinblick auf die Parse-Ergebnisse eignen (Forst et al. 2004).

### 1.4.2.3 Syntaktische Annotierung deutscher Korpora

Für das Deutsche gibt es vier syntaktisch annotierte Korpora<sup>31</sup>. Das NEGRA Korpus (Skut et al. 1998), das darauf aufbauende und weiterführende Korpus der TIGER Treebank (Dipper et al. 2001) und TüBa-D/Z (Telljohann et al. 2004) sind Zeitungskorpora. Verbmobil (Wahlster 2000) dagegen ist ein Korpus der gesprochenen Sprache.

Es folgt eine kurze Gegenüberstellung der zwei unterschiedlichen Annotationsschemata von TIGER (Version 2.1, mit 900,000 Tokens der Frankfurter Rundschau) und Tüba-D/Z (Release 4: 630,000 Wörter der TAZ).

Beide Korpora sind mit Wortarten nach STTS versehen und haben zusätzlich morphologische Informationen (Numerus, Kasus, Genus, Modus, Person, Tempus), syntaktische Kategorien und Funktionen. Abgesehen von der unterschiedlichen Annotation einiger sprachlicher Phänomene (z.B. TIGER annotiert Funktionsverbgefüge, CVC, Abbildung 2) unterscheiden sich die beiden Korpora darin, dass TüBa-D/Z mit Feldern annotiert wird und Fernbeziehungen mit Hilfe der Kantenbezeichnungen (als Vierecke dargestellt) realisiert werden. In TIGER werden Fernbeziehungen mit Hilfe von kreuzenden Kanten dargestellt. Insgesamt ist die TIGER-Annotation weniger hierarchisch auch in Bezug auf die Phrasenstruktur.

Dies hat Auswirkungen auf die Parsebarkeit. Die hierarchische Struktur von TüBa-D/Z erleichtert zunächst das Erlernen der Parser gerade bei kleineren Trainingsdaten, aber

---

<sup>29</sup> auf Wahrscheinlichkeitsrechnungen basierende Algorithmen

<sup>30</sup> Für eine Einführung siehe Jurafsky und Martin (2008).

<sup>31</sup> Die drei bekanntesten syntaktisch annotierten Korpora für das Englische sind: English Penn Treebank (Marcus et al. 1993); Susanne Corpus (Sampson 1995) und Lancaster ParsedCorpus (Leech 1992).

die tiefer eingebetteten Strukturen erzeugen zusätzliche Fehlerquellen und deshalb ist die flache Annotation von TIGER für Parser transparenter (Kübler et al. 2008).

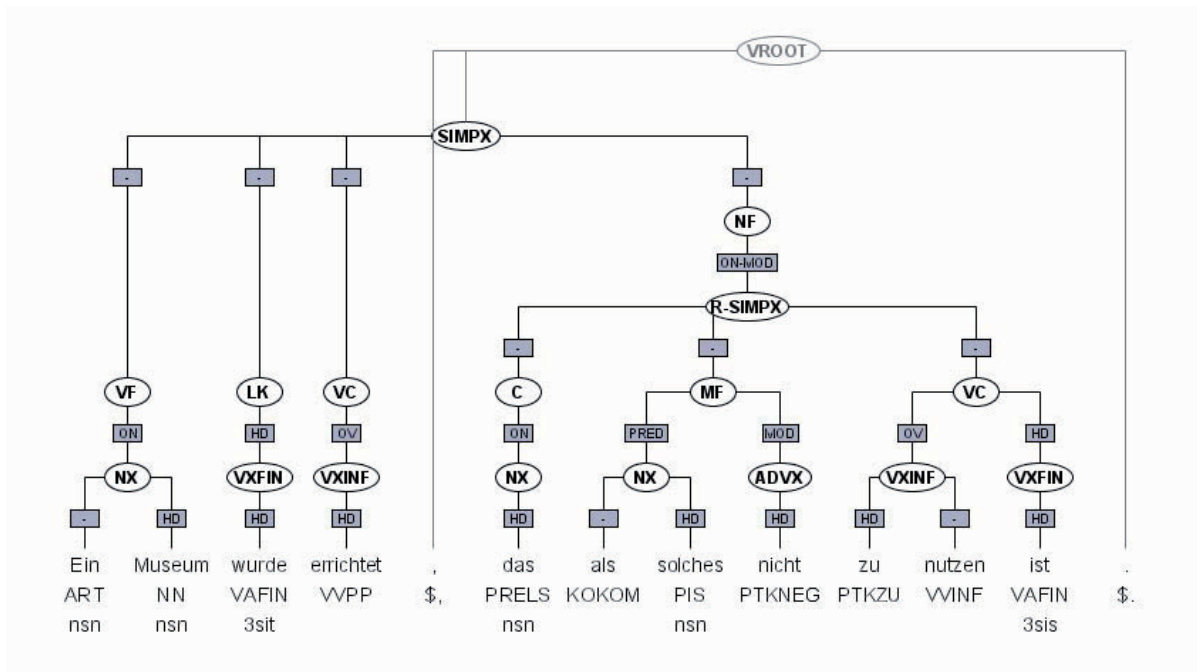


Abbildung 1 Baumdarstellung bei TüBa-D/Z

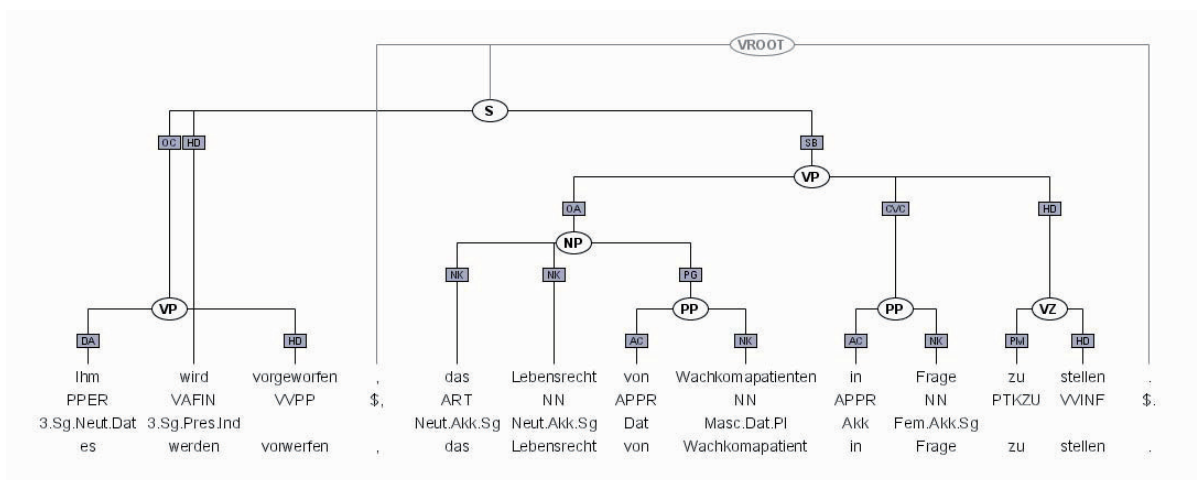


Abbildung 2 Baumdarstellung beim TIGER-Sampler

Für diese Arbeit stellt sich die Frage, in weit diese Anntotationschemata als Grundlage für die Annotation von Lernerdaten dienen können.

Dass eine eins-zu-eins-Übertragung nicht möglich ist, zeigt sich beispielsweise darin, dass die morphosyntaktische Einteilung in Akkusativ-, Dativ- und Genetivobjekte auf Grund von Kongruenz-, Rektions- und Kasusfehlern nicht ohne weiteres auf Lernerdaten übertragen werden kann. Die beiden oben angeführten Annotationsschemata sind nicht dafür ausgelegt, abweichende Strukturen zu

annotieren; da sie bei Zeitungskorpora angewendet werden, wird angenommen, dass damit kanonische Strukturen annotiert werden. Das gleiche Problem gilt für eine Übertragung der strukturellen Annotation des Satzes.

In Folgendem wird etwas genauer auf die Vorgehensweisen von TIGER, TüBa-D/Z und Verbmobil bei verblosen Äußerungen eingegangen. Solche Äußerungen stellen die Annotationsschemata vor Schwierigkeiten, da sie das finite Verb als Kopf der V1/V2-Sätze analysieren. Das Hauptaugenmerk liegt auf den felderannotierten Korpora TüBa-D/Z und Verbmobil, da sie besser mit der Falko-Felderannotation vergleichbar sind.

Im Abschnitt zu „Printing and Spelling Errors“ im Tüba-D/Z Stylebook (S. 27) wird der Umgang mit nicht analysierbaren syntaktischen Strukturen wie folgt dargestellt: lexikalische Einträge, die nicht zu der syntaktischen Konstruktion gehören, werden soweit wie möglich strukturiert, aber nicht im Satz integriert dargestellt, siehe Abbildung 3:

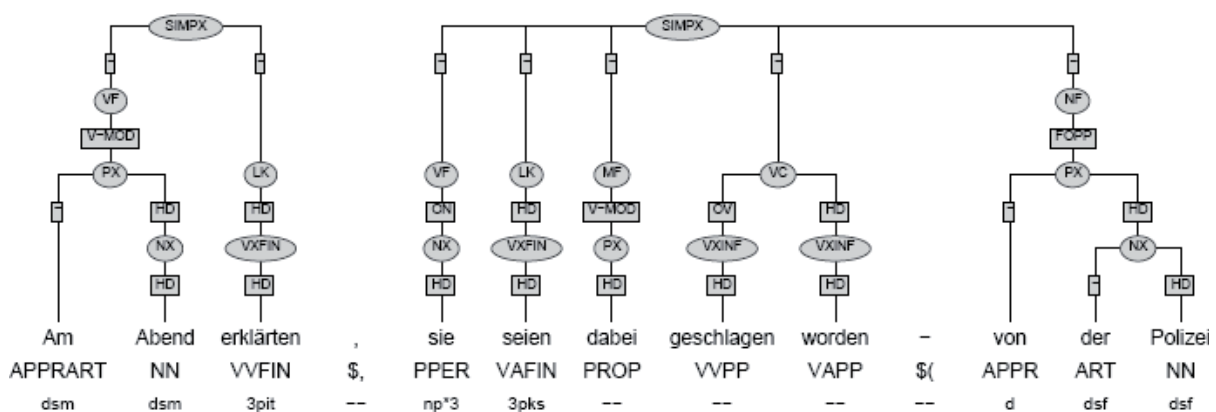


Abbildung 3 nichtkanonische syntaktische Strukturen bei TüBa-D/Z

Das Beispiel in obiger Abbildung ist aber nur als zusammengehörende Äußerung verständlich.

Die Annotation der gesprochenen Sprachdaten von Verbmobil bedient sich verstärkt dieser Strategie von „isolated phrases“. Bei diesen Daten wird davon ausgegangen, dass viel mehr „Fehler“, Abbrüche und Wiederholungen, auftreten können und dass diese mit dem Feldermodell nicht darstellbar sind: „because the attachment of speech errors would conflict with the topological field analysis...“ (Verbmobil Stylebook, Stegmann et al. 2000, S. 24). Konflikte gibt es durch doppelt besetzte Vorfelder bzw. doppelt besetzte LSK. Deshalb werden solche Äußerungen soweit wie möglich in Phrasen strukturiert aber nicht miteinander verbunden.

Auch bei Äußerungen, in denen die Beziehung der einzelnen Phrasen nicht mit syntaktischen Mitteln realisiert wurde, werden deren Phrasen von einander abgetrennt dargestellt, wie in Abbildung 4:

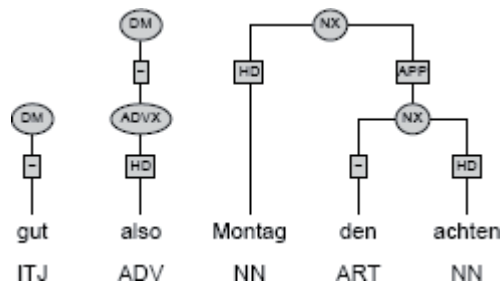


Abbildung 4 „isolated phrase“ (Verbmobil Stylebook, S. 25)

Mit dieser Annotationsweise wird die kommunikative Absicht nicht abgebildet. Die zusammenhängenden semantischen Einheiten werden nicht als solche dargestellt, obwohl es sich eindeutig um Strukturen mit Aussagen handelt, welche auch syntaktisch realisiert werden könnten. Je nach Kontext könnte das obige Beispiel soviel bedeuten wie „Das ist gut. Also treffen wir uns am Montag den achten.“ Mit einer syntaktischen Realisierung gäbe es die Möglichkeit, diese Äußerung mit den Äußerungen des restlichen Korpus, die syntaktisch annotiert wurden, zu vergleichen.

Eine weitere Vorgehensweise bei verblosen Sätzen wird in der Dokumentation des Annotationsschemas von TIGER (Albert et al. 2003, S. 72) vorgeschlagen; wobei eine Hypothese eines sinnvollen Satzes in Gedanken zu formulieren und entsprechend zu annotieren ist. Leider gibt es keine Möglichkeit diese Hypothesen in der Annotation explizit zu machen und Beispiele wie "Keine Chance im Halbfinalspiel" kann entweder als NP oder Satz annotiert werden.

Dieses entspricht ungefähr der gleichen Vorgehensweise wie bei dem Beispiel mit einer elliptischen Äußerung in Abbildung 5 aus dem TüBa-D/Z Stylebook (S. 118). Es wird ein Obersatz (SIMPX) mit einem abhängigen Satz angenommen, die miteinander verbunden dargestellt werden:

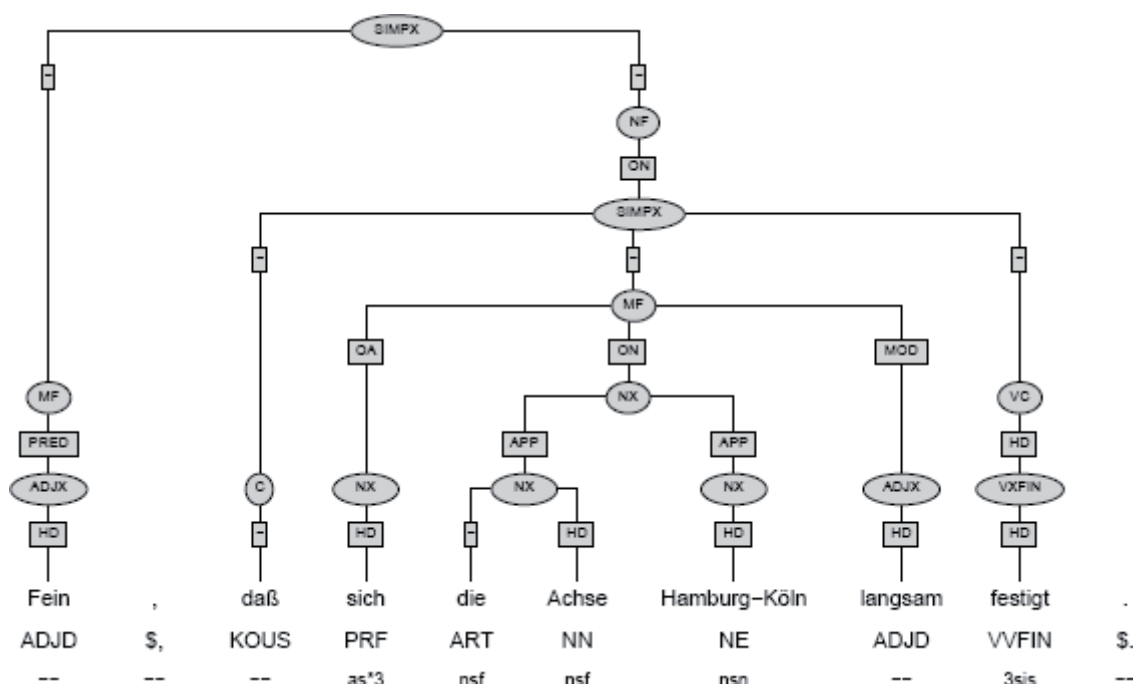


Abbildung 5 elliptische Sätze, Stylebook der TüBa-D/Z

Felder für den obersten Satz (SIMPX) zu annotieren, halte ich für problematisch, denn es fehlt das finite Verb, was als obligatorischer Ausgangspunkt für die Einteilung der Felder gilt. Auch wenn davon ausgegangen werden kann, dass die Verbstellung dem Modell entsprechend realisiert wird<sup>32</sup>, ist es nicht eindeutig, dass „Fein“ im Mittelfeld stehen muss, da mindestens zwei mögliche Realisierungen des Satzes denkbar sind, sowohl „Das ist fein, dass ...“ mit „fein“ im Mittelfeld als auch „Fein ist das, dass ...“ mit „Fein“ im Vorfeld. Hier ist eine implizite Annahme gemacht worden.

Hier wurde wie bei TIGER ein hypothetischer Satz angenommen, der in das jeweilige Modell passt (Satz mit Verb), dieser wird annotiert bzw. analysiert, und die Analyse wird dann auf die Ausgangsäußerung übertragen. Es gibt nichts grundsätzlich gegen ein solches Vorgehen einzuwenden, wenn dabei deutlich gemacht wird, was eigentlich annotiert wird. In der Falko-Felderannotation geschieht dies dadurch, dass Äußerungen, die nicht mit dem Feldermodell annotiert werden können, entsprechend gekennzeichnet (z.B. verblöse „Sätze“) und nicht automatisch in Felder eingeteilt werden. Damit kann gezielt nach diesen Strukturen gesucht werden, die je nach Korpusstyp von Interesse sein können. Wenn diese implizite Hypothesen explizit gemacht werden, kann, statt nur die Hypothese; die Abweichung zwischen Hypothese

<sup>32</sup> Eine Annahme, die für Lernerkorpora nicht zutrifft.

und der Originaläußerung als Beschreibung der Äußerungen dienen (Hirschmann et al. 2007).

### 1.4.3 Lernerkorpora

Im Bereich des EFL (Englisch als Fremdsprache) gibt es groß angelegte computerlinguistische Projekte wie ICLE<sup>33</sup> (Granger 2002), mit teilweise fehlerannotierten schriftlichen Daten und ISLE<sup>34</sup> (Atwell et al. 2003) mit gesprochenen Daten, die mit Phon und Wortakzent fehlerannotiert sind. Gleichzeitig gibt es abgesehen von Falko sehr wenige öffentliche deutsche Lernerkorpora<sup>35</sup> bzw. korpusbasierte Arbeiten<sup>36</sup> (Lüdeling et al. 2008).

#### 1.4.3.1 Design und Annotation von Lernerkorpora:

Die vorwiegend durch die Forschungsfrage bestimmte Zusammenstellung von Lernerkorpora und auch die Vergleichbarkeit mit anderen Korpora hängen ab von verschiedenen Parametern wie Sprachstand, Muttersprache, weitere Fremdsprachen, Lernkontexte (ungesteuert, gesteuert) des Lerners sowie die Umstände der Datenbeschaffung, Zeitraum der Erhebungen (Longitudinal- oder Querschnittserhebungen), Kommunikationsmodus (schriftlich, gesprochen) und der Aufgabenstellung (Granger 2002).

Die CIA (Granger 2002) bietet im Zusammenhang mit computerbasierten korpuslinguistischen Ansätzen bei vergleichbaren Daten (z.B. gleicher Kontext, gleiche Textsorte) mehrere Ansatzpunkte, Lernerdaten zu analysieren. So können L2-Daten mit zielsprachlichen Daten verglichen werden auf Über- bzw. Unterrepräsentation der Gebrauchsfrequenz von bestimmten sprachlichen Elementen (overuse, underuse), ebenso mit anderen L2-Texten mit unterschiedlichen Variablen und, je nach Forschungsfrage, auch mit L1-Daten der Lerner.

Die schriftliche Leistung von fortgeschrittenen Lernern wird im Vergleich mit muttersprachlichen Texten oft als ungenauer, vager und als der gesprochenen Sprache näher wahrgenommen, (Cobb 2003, S. 2). Eine Hypothese zur Erklärung dieser Beobachtungen ist, dass die unspezifische Lexik der gesprochenen Sprache in

---

<sup>33</sup> bestehend aus Essays „freies Schreiben“ mit 3 Millionen Wörtern, 14 Muttersprachen:  
<http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/research%20learner%20corpora.html>.

<sup>34</sup> 11484 Äußerungen, deutsche und italienische Muttersprache: <http://nats-www.informatik.uni-hamburg.de/~isle/speech.html>

<sup>35</sup> gesprochene Daten aus ungesteuertem Spracherwerb:

ESF, MPI corpora: <http://corpus1.mpi.nl>

LEAF: <http://www.uni-bielefeld.de/Universitaet/Einrichtungen/Zentrale%20Institute/IWT/FWG/Sprache/Korpus>.

<sup>36</sup> z.B. Weinberger (2002), Belz (2004)

Lernertexten im Kontrast überrepräsentiert (overuse) und die Lexikauswahl des Schriftlichen unterrepräsentiert wird (underuse). Overuse und underuse beschränken sich nicht nur auf den Lexembereich, wobei welche linguistische Kategorien miteinander verglichen werden können, hängt auch von der Annotation der Korpora ab. Bei nicht annotierten Korpora können nur konkrete Formen gesucht werden. Es gibt also keine Möglichkeit, formgleiche Wörter zu unterscheiden, z.B. *die* (als Artikel) und *die* (als Relativpronomen).

Bei Lernerkorpora gibt es je nach Sprachstand der Lerner theoretisch die gleichen Annotationsmöglichkeiten wie bei muttersprachlichen Korpora (Biber et al. 1998). Meines Wissens nach gibt es aber so gut wie keine syntaktisch annotierten Korpora, auf die ich mich beziehen könnte. Für das Englische bemerkt Granger (2002, S. 18), dass das automatische Wortarttagging und Lemmatisieren bei fortgeschrittene Lernern mit einer relativ hohen Genauigkeit eingesetzt werden kann<sup>37</sup>.

#### **1.4.3.2 Fehleranalyse und Fehlerannotation**

Auch wenn der Schwerpunkt dieser Arbeit nicht auf der Fehleranalyse liegt, beruhen die Kriterien dafür, ob eine Struktur in das Feldermodell eingeordnet werden kann auch auf der Basis von „Verbstellungsfehlern“.

Der Fehlerbegriff und die Fehleranalyse sind viel diskutierte Bereiche. Die Wichtigkeit von „Lernerfehlern“ ist zur Beschreibung der Interimsprache (Corder 1967) kaum umstritten, sie aber als alleinigen Untersuchungsgegenstand oder Beschreibungsansatz zu sehen, dagegen schon.

Häufig ist Kritik an der Fehleranalyse auch methodischer Natur. Ein Problem am methodischen Vorgehen bei früheren Fehleranalysen war die fehlende Reproduzierbarkeit ihrer Ergebnisse. Unter bestimmten Voraussetzungen (öffentlich zugänglich, umfassende Metadaten) kann die computerbasierte Analyse von Korpora dem abhelfen.

Eine dennoch entscheidende Frage ist, in wieweit Fehleranalysen überhaupt mit einander vergleichbar sind: was wird als Fehler identifiziert und wie werden sie kategorisiert? Um dieser Frage nachgehen zu können, muss erst geklärt werden, was unter einem Fehler zu verstehen ist. Lennon (1991, S. 182) definiert Fehler als “a

---

<sup>37</sup> Bei einer Untersuchung gesprochener Daten erzielte das STTS-Tagger für das Deutsche 85,7% Genauigkeit (Pankow und Pettersson 2006). Tendenziell trifft dies auch für den Falko Zusammenfassungskorpus zu, aber für publizierbare Aussagen müsste erst eine genauere, systematische Erfassung der PoS-Abweichungen vorgenommen werden.

linguistic form which, in the same context would in all likelihood not be produced by the learner's native speaker counterparts". Die Annahme einer muttersprachlichen Entsprechung heißt im folgenden Zielhypothese: "reconstruction of those utterances in the target language" (Ellis 1994, S. 54).

Die Vergleichbarkeit hängt demzufolge sowohl von der Zielhypothese als auch von der Kategorisierung der Fehler ab. Lüdeling (2008) macht deutlich, wie sehr sich Zielhypothesen unterscheiden können. In den wenigsten Fehleranalysen wird die Zielhypothese explizit gemacht, was die Vergleichbarkeit sogar bei Verwendung der gleichen Kategorisierungskriterien bzw. des gleichen Fehlertagsets erschwert.

Es gibt unterschiedliche Klassifizierungsmöglichkeiten von Fehlern. Die gebräuchlichste ist, sie nach linguistischen Kategorien zu unterscheiden (z.B. Morphologie, Syntax, Lexik). Man kann auch deskriptiver vorgehen und Auslassungen (ommission), Ersetzungen (replacement) oder Hinzufügungen (addition) beschreiben. In den meisten Annotationen werden Fehlertags direkt vor oder nach dem fehlerhaften Element platziert. Damit lässt sich der genaue Bereich (bzw. die Reichweite) des Fehlers nicht identifizieren und es ist auch nicht möglich, mehrfache Hypothesen über die Fehlerkategorie (bzw. Fehlerquelle usw.) aufzustellen.

## **2 Hauptteil: Annotationsverfahren**

### **2.1 Beschreibung**

#### **2.1.1 Korpus und Korpusdesign**

Das gesamte Falko-Korpus besteht aus drei Subkorpora, dem Essaykorpus und dem Zusammenfassungskorpus, die von der FU und HU Berlin erhoben wurden und dem Korpus der Georgetown University (GU). Alle Korpora wurden automatisch mit dem STTS-Tree-Tagger mit Wortart und Lemma annotiert.

Die Felderannotation wurde an den L2-Texten der Falko-Zusammenfassung und an dem Falko-GU-Korpus durchgeführt. Bei den annotierten GU-Daten handelt sich um genre-orientierte Schreibaufgaben, sogenannte Prototypical Performance Writing Tasks (PPTs) der Level 2 (intermediate), Level 3 (advanced intermediate) und Level 4 (advanced), die am Ende der Level routinemäßig als Hausarbeiten digital eingereicht werden.

Die longitudinale GU-Daten (GU2, GU3 und GU4) sind innerhalb des „Fremdsprache Deutsch, Curriculum, Developing Multiple Literacies“<sup>38</sup> entstanden, wobei nur Level 3 und 4 als fortgeschritten gelten. Insgesamt handelt es sich um 71 Texte (67.325 Token) unterschiedlicher Genres<sup>39</sup>, die von 28 (16) verschiedenen Lernern stammen. Die Daten liegen in der ersten unkorrigierten digitalen Version vor. Es gibt keine Information über Schreibdauer oder benutzte Hilfsmittel. Die Metadaten der Texte beschränken sich auf eine Zuordnung zu Schreiber, Erhebungszeitpunkt und Kurslevel. Es ist anzunehmen, dass die meisten englische Muttersprachler sind.

Die Falko-Zusammenfassung besteht aus 107 Texten mit 41.075 Token<sup>40</sup>. Sie stammen überwiegend aus einer Sprachstandsüberprüfung von ausländischen Studierenden mit einem germanistischen Hauptfach an der FU-Berlin. Alle Teilnehmer hatten erfolgreich die DSH-Prüfung abgelegt und haben sich meistens für die Dauer ihres Grundstudiums in Deutschland aufgehalten. Es handelt sich um Zusammenfassungen wissenschaftlicher Texte im Bereich der germanistischen Linguistik. Metadaten zu Geschlecht, Alter, Muttersprache (L1) und weiteren Fremdsprachen (L2, L3-Ln) und Dauer und Art des Fremdsprachenerwerbs sind erhoben worden. Es wurden keine Hilfsmittel benutzt (bis auf den Vorlagentext). Die Prüfung dauerte 90 Minuten unter Aufsicht und wurde handschriftlich abgefasst. Text und Thema waren unbekannt.

### **2.1.2 Annotation und Korpusarchitektur**

Die Annotationsarchitektur basiert auf einem Mehrebenen-Aufbau (Lüdeling et al. 2005; Bird et al. 1999). Im Prinzip geht es darum, mehrere zunächst unabhängige Annotationsebenen im gleichen Korpus zu ermöglichen, die bei der Suche kombiniert werden können.

Abbildung 6 gibt einen Überblick über den Aufbau der Annotationsebenen im Falko-Korpus, angelehnt an die Darstellung in EXMARaLDA<sup>41</sup> (Schmidt und Wörner 2005). Drei Ebenen werden unterschieden: die Ebene des Gesamtkorpus, die Ebenen der Fehler und schließlich die Ebenen der Felder, die sich wiederum einteilen in Satz- und Felderebenen.

---

<sup>38</sup> <http://www3.georgetown.edu/departments/german/programs/curriculum>

<sup>39</sup> Verfassen eines alternativen Romanendes, Journalistisches Schreiben und Verfassen einer Rede

<sup>40</sup> Siehe die Falko-Website für eine genauere Zusammenstellung der Texte.

<sup>41</sup> Dieser leicht zu bedienende Partitur-Editor mit übersichtlicher Darstellung wurde für multimediale Korpora entwickelt und diente als Eingabe-Tool für die Felderannotation. (siehe <http://www.exmaralda.org/>)

Token	1	2	3	4	5	6	7	8
[word]	Gestern	er	machte	Pause	.	Es	schneite	.
[lemma]	gestern	er	machen	Pause	.	Es	schneien	.
[pos]	ADV	PPN	VVFIN	NN	\$.	PPER	VVFIN	\$:
[corrected_pos]								
[target_hypothesis]	Gestern machte er Pause.							
[matrix_satz]	x					x		
[ms_felder]	f_MS					VF_MS	LSK_MS	
[konstituenten_satz_1]								
[ks_1_felder]								
[syntax_description]	x							
[syntax_classification]	MF_LSK							
[syntax_classification_2]	VVFIN							
[syntax_hypothesis]	ADV							

Abbildung 6 Überblick der Annotationsebenen bei Falko

## 2.1.2.1 Beschreibung der Ebenen

### 2.1.2.1.1 Ebene des gesamten Korpus

Auf der ersten Ebene, im Folgenden als die Wort-Ebene [word]<sup>42</sup> bezeichnet, befindet sich das „raw“-Korpus, das tokenisiert wurde und unabhängig von der Annotation untersucht werden kann. In dieser Ebene können statistische Informationen wie Satzlänge, Wortanzahl, Wörterfrequenz usw. gesucht werden. In der zweiten Ebene [pos]<sup>43</sup> wurde jedem Token ein Wortart-Tag und in der dritten [lemma] ein Lemma zugeordnet. Dies wurde hier automatisch mit dem STTS-Tagger durchgeführt. Die korrekte Zuordnung hängt ab von der Rechtschreibung, auch Groß- und Kleinschreibung, sowie von der Zeichensetzung und von der richtigen Verbform (Konjugation). Mit Hinzuziehen von diesen Ebenen ist es möglich, auch Relativsätze zu identifizieren, was allein auf der Wort-Ebene nicht möglich wäre, da sich Artikel und Relativpronomen in der Form nicht unterscheiden.

### 2.1.2.1.2 Ebenen der Fehlerannotation

Ebene vier [target\_hypothesis] stellt die Zielhypothese dar, die als Grundlage für die Fehleranalyse dient. Jede Fehleranalyse basiert auf einer zielsprachlichen Hypothese. Da diese Hypothesen sehr individuell ausfallen können, ist es notwendig, sie explizit zu

<sup>42</sup> Annotationsebenen werden in eckigen Klammern dargestellt.

<sup>43</sup> Pos (auch PoS) steht für "part of speech".

machen, damit die Fehleranalyse durchsichtiger wird. Die Zielhypothese, die die kommunikative Absicht zu erfassen versucht, beschränkt sich nicht auf „ungrammatikalische“ Phänomene. Sie umfasst auch stilistische und intonatorische „Korrekturen“.

In der fünften Ebene wurde eine Korrektorebene für die Wortart-Tags [corrected\_pos] eingeführt, da man bei Lernern häufig Rechtschreibfehler usw. erwarten kann. Andererseits macht der Tagger auch Fehler bei der Entscheidung über offene bzw. uneindeutige Wortarten. Es wäre sinnvoll, zwei verschiedene Korrektorebenen einzuführen, um sowohl fehlerbedingte als auch taggerbedingte Zuordnungsungenauigkeiten zu dokumentieren.

### 2.1.2.1.3 Ebenen der Felderannotation: Annotationsschema

Die Ebenen der Felderannotation werden im Folgenden im Zusammenhang mit dem Annotationsschema für die Felderannotation vorgestellt.

Die Felderannotation läuft in drei Schritten ab, dargestellt in Abbildung 7.

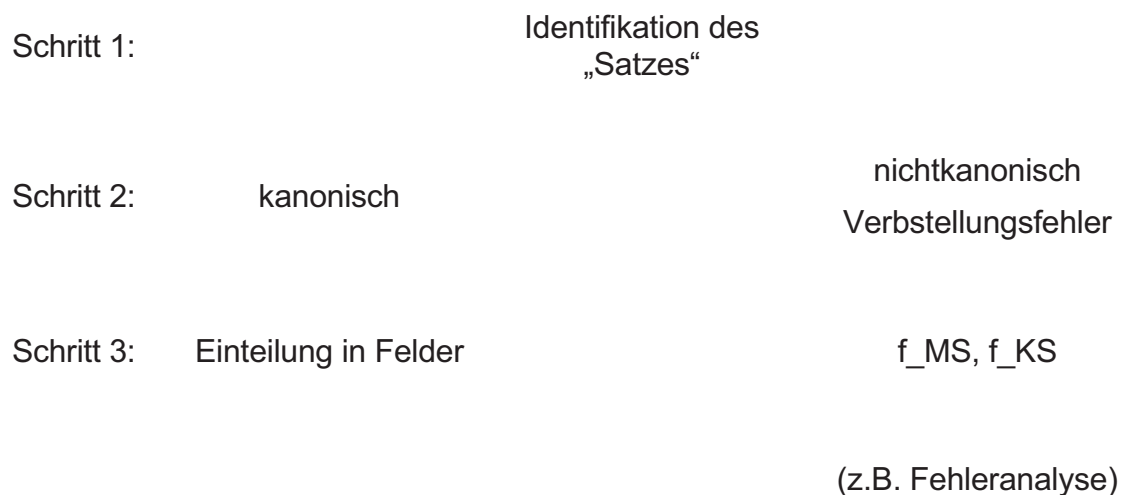


Abbildung 7 Annotationsschema für die Felderannotation

Zunächst müssen die Lenersätze (Äußerungen) identifiziert werden. Im zweiten Schritt wird zwischen kanonischen und nichtkanonischen Strukturen unterschieden, wobei die Kanonizität kein absoluter Wert ist, sondern sich auf das jeweilige Regelwerk bzw. anzuwendende Modell (siehe Hirschmann et al. 2007) bezieht, in diesem Fall das Feldermodell. Im dritten Schritt werden die kanonischen Sätze in das Feldermodell

eingeorordnet und nichtkanonische Äußerungen als solche gekennzeichnet (mit f\_MS bzw. f\_KS-Tags).

### **Schritt 1:** Identifikation der Sätze (Äußerungen)

Hier sei darauf hingewiesen, dass der Satzbegriff in diesem Zusammenhang problematisch ist. Wenn von einer syntaktischen Definition<sup>44</sup> von Satz ausgegangen wird, ist es heikel, nichtkanonische Äußerungen als Satz zu bezeichnen. Dies gilt auch für kontextuelle elliptische Strukturen wie „Rot“ in Beispiel (6). Deshalb wird alternativ der Begriff Äußerung verwendet.

(6) *Welche Farbe hat es? Rot.*

### **Satzebene:**

In diesen Ebenen werden die „Sätze“ identifiziert und mit <x><sup>45</sup> markiert, die als Basis für die Felderannotation dienen. Es wird zwischen Matrixsatz (MS) und Konstituentensatz (KS) unterschieden. Mit Matrixsatz ist hier ein Hauptsatz mit allen dazugehörigen bzw. abhängigen Untersätzen (meist Nebensätze, aber auch Infinitivkonstruktionen) gemeint. Daraus folgt, dass zwei koordinierte vollständige Hauptsätze<sup>46</sup> in der Ebene [matrix\_satz] zwei Matrixsätzen entsprechen, auch wenn sie in [word] nur als ein „Satz“ mit Großschreibung am Satzanfang und einem Punkt am Satzende realisiert wurden.

In einer weiteren Ebene [konstituenten\_satz\_1] werden die Untersätze der Matrixsätze identifiziert und mit <x> markiert. Die Untersätze jedes Konstituentensatzes werden wiederum in der Ebene [konstituenten\_satz\_2] identifiziert und markiert usw. Aus Gründen der Handhabung bei der Arbeit in EXMARaLDA gibt es nur drei Konstituentensatzebene, die bis auf wenige Ausnahmen<sup>47</sup> ausreichen, um die Lernertexte zu annotieren.

Der Begriff Konstituentensatz wird hier als Oberbegriff benutzt. Als KS annotiert werden alle abhängigen Untersätze bzw. alle satzähnlichen Strukturen mit einem

---

<sup>44</sup> Siehe hierzu Pasch (2003, S. 85)

<sup>45</sup> Tags werden im Weiteren durch die umschließende spitze Klammer <> dargestellt.

<sup>46</sup> siehe hierzu 2.2.1.1 Satzkoordination

<sup>47</sup> Zum Beispiel der Satz aus den Daten von „Level 3 GU“: *Obgleich ich keine Weise habe, zu wissen, ob er recht hat, weil es keine Statistikern über die nicht berichtete tägliche Belästigung gibt, die von diesen Immigranten erfahren wird, hoffe ich, daß die Situation von Tung nicht normal ist.* (Lerner: GU\_3112)

finiten oder infiniten Verb.<sup>48</sup> Es werden sowohl Strukturen mit Konstituentenstatus als auch mit attributativer Funktion oder mit satzweiterführenden Eigenschaften erfasst.

Auf diesen Satzebenen werden auch andere kommunikative Einheiten bzw. Äußerungen wie kontextuelle Ellipsen, <ELP>, sowie textstrukturierende Elemente wie Überschriften und Nummerierungen, <P\_ns>, mit Tags gekennzeichnet. Es gibt zwar auch Beispiele von Lerneräußerungen, die nicht vollständige „Sätze“ sind (z.B. wenn ein finites Verb fehlt), sie unterscheiden sich aber von elliptischen Strukturen dadurch, dass kontextuelle Ellipsen auch von Muttersprachlern als korrekt empfunden werden.

### **Schritt 2: Entscheidung kanonisch - nicht kanonisch**

Das Feldermodell ist ein Beschreibungsansatz, der, wie in Abschnitt 1.2 erwähnt, die Sätze in Felder einteilt, ausgehend von der Stellung des finiten Verbs (LSK) und des restlichen Verbkomplexes (RSK) in V1- und V2-Sätzen, bzw. von der Stellung des Nebensatzeinleiters (LSK) und des finiten Verbs (RSK) bei Verbendsätzen. Dieser Logik folgend können Äußerungen mit Verbstellungsfehlern im Feldermodell nicht eingeordnet werden, weil die Stellung der verbalen Elemente der Satzklammer nicht entsprechend der syntaktischen Regularitäten der Zielsprache realisiert werden. Die deskriptiven Regeln, die mit dem Feldermodell beschrieben werden (z.B. nur eine Konstituente im Vorfeld) beziehen sich nur auf die Eigenschaften der Zielsprache.

Letztendlich resultiert daraus folgende **Festlegung für kanonisch - nichtkanonisch**:

Für alle Äußerungen, bei denen die Stellung der LSK bzw. der RSK falsch realisiert wird (Verbstellungsfehler) bzw. diese gar nicht vorhanden sind, ist die Einteilung in Felder nicht möglich. Es soll von der Lerneräußerung ausgehend, versucht werden, Felder einzuteilen und kein <f\_MS> oder <f\_KS> zu taggen bei Zielhypothesen mit stilistischen Wortstellungsvariationen oder mit Modus-Tempusabweichungen mit Wortstellungsänderungen. Im Allgemeinen werden falsche Formen nicht als <f\_MS> getaggt.

Mit dieser Festlegung werden Strukturen als kanonisch beschrieben, auch wenn sie Fehler im Mittelfeld aufweisen. Diese Vorgehensweise scheint berechtigt, da das Mittelfeld als eigene Klasse zu sehen ist und Stellungsfehler hier eigens beschrieben

---

<sup>48</sup> Parenthetische Sätze bzw. nicht syntaktisch gebundene satzwertige Einschübe werden mit <P> gekennzeichnet (für eine genauere Besprechung diese Strukturen siehe Abschnitt 2.2.1).

werden können. Ob diese Kriterien ausreichen, und wie sie im einzelnen aussehen, wird in Abschnitt 2.3 diskutiert.

An dieser Stelle soll etwas zur der Begrifflichkeit „Verbstellungsfehler“ anhand eines Verbzweitfehlers erläutert werden.

Gemäß der Definition zur Interpretation von Fehlern als Abweichung zu einer Zielhypothese, kann Beispiel (2) wie folgt dargestellt werden:

(7) *Gestern er machte Pause.*

Lerneräußerung	Gestern		er	<b>machte</b>	Pause
Zielhypothese	Gestern	<b>machte</b>	er		Pause
Felder der Zielhypothese	VF	LSK	MF		

Die Abweichung kann auf verschiedene Weise beschrieben werden, wie hier als Abweichung von der Feldereinteilung der Zielhypothese mit dem Formalismus MF\_LSK/VVFIN<sup>49</sup> definiert oder einfach als V2-Fehler.

Natürlich kann dieser Verstoß auch als doppelte Konstituentenbesetzung des VF beschrieben werden (VF\_MF/XP; wobei XP einer Konstituenten entspricht)<sup>50</sup>.

Lerneräußerung	Gestern	<b>er</b>	machte		Pause
Zielhypothese	Gestern		machte	<b>er</b>	Pause
Felder der Zielhypothese	VF		LSK	MF	

Für die Felderannotation wurde festgelegt, von Verbstellungsabweichungen auszugehen, weil dies die gängigste Fehlerbeschreibung ist. Im Folgenden soll nicht näher auf die Fehlerannotation eingegangen werden, aber es besteht ein deutlicher Zusammenhang.

### Schritt 3:

Kanonische Strukturen werden in Felder eingeteilt. Nichtkanonische Strukturen werden mit <f\_MS> (z.B. bei nicht realisiertem Verbzweit) und nichtkanonische Konstituentensätze werden mit <f\_KS> getaggt (z.B. bei Verbendfehler). Diese Äußerungen können je nach Ansatz weiter beschrieben werden (z.B. durch eine Fehleranalyse).

<sup>49</sup> Das finite Verb (VVFIN) wird im MF der Zielhypothese abgebildet, soll aber in der LSK realisiert werden.

<sup>50</sup> XP wird im VF der Zielhypothese abgebildet, soll aber im MF realisiert werden.

## Ebenen der Feldereinteilung bei kanonischen Strukturen:

Abbildung 8 zeigt die Ebenen der Feldereinteilung am Beispiel einer Lerneräußerung (komplexer Satz mit mehreren Untersätzen):

[word]	Das	bedeutet	,	dass	es	nötig	ist	,	Wörter	zu	erkennen
[pos]	PDS	VVFIN	\$,	KOUS	NN	ADJD	VAFIN		NN	PTKZU	VVINF
[matrix_satz]	x										
[ms_felder]	VF_MS	LSK_MS	NF_MS								
[konstituenten_satz_1]			x								
[ks_1_felder]				LSK_KS	MF_KS	RSK_KS	NF_KS				
[konstituenten_satz_2]								x			
[ks_2_felder]								MF_KS	RSK_KS		

Abbildung 8 vereinfachte EXMARaLDA-Nachbildung  
in Anlehnung an Falko – Zusammenfassungskorpus, FU\_ 003<sup>51</sup>  
Schlüssel: ms = Matrixsatz, ks = Konstituentensatz

## Zuordnung der Tags in den verschiedenen Ebenen:

In untenstehender Liste folgen die Tags, ihre wörtliche Ausformulierung und die jeweils mögliche bzw. obligatorische Felderbesetzung. Die genauere Beschreibung und Besprechung findet sich in Abschnitt 2.2 Kriterien der Feldereinteilung. Das vollständige Tagset ist im Appendix A abgedruckt.

### Felderannotation in [matrix-satz\_felder]

<LSK\_MS> - linke Satzklammer des Matrixsatzes - das finite Verb

<RSK\_MS> - rechte Satzklammer des Matrixsatzes - infinite Verbgruppe, trennbare Verbbestandteile

<VF\_MS> - Vorfeld(er) des Matrixsatzes - der ganze Bereich vor dem finiten Verb des Matrixsatzes

<MF\_MS> - Mittelfeld des Matrixsatzes - der Bereich zwischen den Satzklammern bzw. zwischen der linken Satzklammer und dem Nachfeld

<NF\_MS> - Nachfeld(er) des Matrixsatzes - das Nachfeld folgt nach der RSK und nach den Elementen, die fest am Ende des Mittelfelds platziert sind

<sup>51</sup> Texte vom Falko-Zusammenfassungskorpus sind mit FU\_ gekennzeichnet, die Texte aus dem Georgetown-Korpus mit GU\_.

(Prädikativ, direktionale und situative Adverbien, nichtverbale Bestandteile der Funktionsverbgefüge)

### **Felderannotation in [konstituenten-satz\_felder]**

<LSK\_KS> - linke Satzklammer des Konstituentensatzes - Nebensatzeinleiter

<MF\_KS> - Mittelfeld des Konstituentensatzes - Bereich zwischen den Satzklammern

<RSK\_KS> - rechte Satzklammer des Konstituentensatzes - gesamter Verbkomplex mit finitem Verb

<NF\_KS> - Nachfeld(er) des Konstituentensatzes - Nachfeld wird gebildet nach der RSK

## **2.2 Kriterien der Feldereinteilung**

Es gibt verschiedene Faktoren die berücksichtigt werden sollten bei der Festlegung der Kriterien der Feldereinteilung. Zur Überprüfung und Konkretisierung der Kriterien des Annotationsschemas für die Einteilung in Felder werden im Folgenden verschiedene gängige Lehrmeinungen (allgemein anerkannte Ansätze) gegenübergestellt. Als weitere Instanzen werden die Annotationsschemata von TüBa-D/Z und Verbmobil hinzugezogen, um die Vergleichbarkeit mit anderen annotierten Korpora zu überprüfen.

Für die Annotation (Identifikation und Zuordnung) der Felder liegt der Schwerpunkt nicht auf der Abfolge innerhalb der einzelnen Felder sondern auf der Abgrenzung und Bestimmung der Felder. Deshalb werden Ansätze in der Literatur zu folgenden Punkten miteinander verglichen:

- Einteilung des Bereichs vor dem finiten Verb (Vorfeld) in V2-Sätzen (Einordnung der Konjunktion bei Satzkoordinationen),
- Abgrenzung von Mittelfeld und Nachfeld bei nicht besetzter RSK in V2/V1-Sätzen und eng damit verbunden Besetzung der rechten Satzklammer
- Besetzung vom Nachfeld und Zuordnung zum Satz

### **2.2.1 Vorfeld**

Das Vorfeld wird im Allgemeinen als das Feld vor dem finiten Verb in V2-Sätzen definiert, das im Unterschied zum Mittelfeld nur mit einer Konstituenten besetzt werden

darf. Damit ist eine Regel aufgestellt, die die V2-Wortstellung beschreibt und folgende Abfolge in einem Aussagesatz für unzulässig erklärt.

(8) *Jetzt er geht nach Hause*

Folgender Satz macht aber deutlich, dass vor dem finiten Verb nicht nur eine Phrase stehen kann:

(9) *Oh je, Frau Schulz, und dass muss ich jetzt wirklich sagen, nun aber ich dagegen bin einfach nicht soweit.*

Im Anbetracht dieser Tatsache plädiert Dürscheid (1989) für einen dennoch möglichst eng gefassten Vorfeldbegriff. Statt den Vorfeldbegriff zu erweitern, werden weitere Felder postuliert. In der Literatur wird meist von einem einfach besetzten Vorfeld ausgegangen (im Beispiel (9) „*ich*“) und ein weiteres Feld vor dem Vorfeld definiert. Es wird aber unterschiedlich bezeichnet: Eroms (2000, S. 352ff. ua.) nennt es Vor-Vorfeld, Hoberg (1997, S. 1577 ff.) linkes Außenfeld und Pasch (2003, S. 69 ff.) unterscheidet gleich drei Stellen: Nullstelle, Vorerst- und Nacherstposition.

Es gibt meines Erachtens drei „erkennbare“ Gruppen von Elementen, die zusammen mit der Vorfeldkonstituenten (VF-XP) vor dem finiten Verb auftreten können<sup>52</sup>:

1. mehrfache VF-XPs
2. VF-XP + Konnektoren<sup>53</sup>
3. VF-XP + pragmatische Strukturen (Vokative, Interjektionen, Thematisierungsausdrücke, Linksversetzung, Gesprächspartikel, parenthetische Sätze, emotive Elemente usw.)

#### Zu 1: **mehrfache VF-XPs**

Bei der Diskussion um mehrfach besetzte Vorfelder wird unterschieden zwischen komplexen Ausdrücken, z.B. komplexe Adverbialausdrücke oder erweiterte Konstituenten durch Appositionen, und „tatsächlicher“ Mehrfachbesetzung, also der Besetzung durch mehrere Konstituenten<sup>54</sup> (Pittner und Berman 2007, S. 85-86). Es gibt einige seltene Mehrfachbesetzungen, die hier nicht weiter erläutert werden, aber andere können durchaus gelegentlich auftreten, wie Teile des Verbkomplexes zusammen mit anderen Elementen:

---

<sup>52</sup> wobei die Gruppen sich überschneiden können.

<sup>53</sup> auch Konjunktionen, und Satzverknüpfers genannt, (siehe Pasch 2003).

<sup>54</sup> Auch wenn die Unterscheidung nicht immer ganz scharf gezogen werden kann, wird in Beispielen wie „*Gestern im Kino nach dem Film hat sie ein Mann angesprochen*“ (zit n. Pittner und Berman, S. 85) das Vorfeld „*Gestern im Kino nach dem Film*“ als komplexer Ausdruck analysiert, da er zusammen erfragbar ist.

(10) *Nichts gesehen hat er.*<sup>55</sup>

oder unabhängige deiktische (hinweisende, zeigende) Elemente und NPs:

(11) *und vorhin die Pyramide des Turnvereins war wunderschön.*

**Zu 2. VF-XP + Konnektoren:**

Häufig, wenn auch umstrittenen bezüglich des Konstituentenstatus<sup>56</sup>, sind Vorkommnisse von einem Satzglied zusammen mit Adverbkonnektoren, Fokuspartikeln und Satzadverbien. Das Sprachbeispiel aus Pasch in Tabelle 2 zeigt die Möglichkeiten der Mehrfachbesetzung mit Konnektoren vor dem finiten Verb.

Nullstelle	Vorerst- position	VF	Nacherst- position	LSK	MF	RSK
Aber	sogar	das neueste Programm	freilich	kann	keine hundertpro- zentige Sicherheit	garantieren.
Konjunktork	Fokus- partikel		Adverb- konnektor			

Tabelle 2 Konnektorenpositionen vor dem Finitum in V2-Sätzen

aus Pasch (2003, S.72)

Diese Beispiele machen deutlich, dass Lerner, die die Verbzweit-Strukturen des Deutschen lernen, durch authentischen Input nicht „reine“ V2-Strukturen aufnehmen, und dass die Entscheidung, ob es sich um einen „wohlgeformten“ deutschen Satz handelt, sich nicht auf Sätze mit „einfach“-besetzten Vorfeldern beschränken kann.

**Zu 3: VF-XP + pragmatische Strukturen**

Auch wenn die Bezeichnungen und Kategorisierungen der hier als pragmatische Elemente zusammengefassten Strukturen nicht einheitlich sind, wird ihnen in der Literatur das gemeinsame Merkmal zugeschrieben, dass sie als syntaktisch isolierter und weniger bzw. nicht im Satz integriert gesehen werden. In Tabelle 3 wird ein Vergleich vorgenommen zwischen Beispiel (12) von Eroms (2000, S. 352) und Beispiel (13) von Hoberg (1997, S. 1580), in denen auch Konnektoren vorkommen.

(12) *Und ach Herr Meier, tatsächlich den Briefträger ja zum Donnerwetter, den habe ich auch gesehen.*

(13) *Ach, Vera, aber immerhin den Jens, den kenne ich gut.*

<sup>55</sup> Auch hier wird eine Komplexbildung als Analyse vorgeschlagen Müller (2007, S. 147).

<sup>56</sup> Fokuspartikel + NP wie hier in der Tabelle wird bei Pasch (2003, S. 71) als eine komplexe "Fokuskonstituente" beschrieben. Es stellt sich dann die Frage, ob in diesem Fall die Vorerstposition nicht Teil des VF ist.

Die Gegenüberstellung zeigt die von den Autoren betrachteten Strukturen, ihre von ihnen zugewiesenen Bezeichnungen und die von ihnen postulierte lineare Abfolge der Vorvorfeld-Elemente. Ein exakter Vergleich ist nicht möglich, da Eroms Elemente anführt, die vom IDS nicht aufgenommen wurden und umgekehrt. Dennoch zeigt diese Gegenüberstellung, dass es vermutlich unterschiedliche Auffassungen geben muss, bezüglich der Stellung der Konjunktion (Konjunktoren) bzw. der Stellung der Interjektionen und Vokative.

Vorvorfeld (Eroms)		linkes Außenfeld (IDS)		
postulierte Abfolge	Beispiel	Beispiel	postulierte Abfolge	
Konjunktionen	Und			
Interjektionen	ach	Ach	Interjektionen	interaktive Einheiten
Vokativ	Herr Meier	Vera	Vokativ	
		aber	Konjunktoren	koordinierende Ausdrücke
Konnektor	tatsächlich	immerhin	Konnektivpartikel	
Linksversetzung	den Briefträger	den Jens	Thematisierungsausdruck	linksgebundene und freie Thematisierungsausdrücke
Gesprächspartikel	ja			
emotives Element	zum Donnerwetter			
Vorfeld	den	den	Vorfeld	
	habe ich auch gesehen.	kenne ich gut.		

Tabelle 3 Wortabfolge vor dem finiten Verb in V2-Sätzen

Weitere pragmatische Elemente, die vor dem finiten Verb auftreten können, sind parenthetische Sätze. Sie werden syntaktisch unabhängig gesehen und können somit an verschiedensten Stellen des Satzes auftreten. Altman (2005, S. 83) bezeichnet diese Stellen als „Parenthesen-Nischen“. Er postuliert zwei mögliche Positionen vor dem finiten Verb, und zwar zwischen der Vorfeld-Konstituenten und dem finiten Verb:

- (14) *(\*Da bin ich mir sicher) Mein Metier (da bin ich mir sicher), das (da bin ich mir sicher) ist in diesen Zeiten ... die Kochkunst.*<sup>57</sup>

(zit.n. Altmann und Hahnemann 2005, S. 148)

<sup>57</sup> Fraglich ist, wie das bei folgendem Beispiel aussieht, bei dem der parenthetische Ausdruck vor der Linksversetzung steht?  
 (a) *Was ich schon mal sagen wollte, mein Metier, das ist in diesen Zeiten ... die Kochkunst.*

Anhand dieser Beispiele wird deutlich, dass möglicherweise nicht nur eine Position vor der Vorfeldkonstituenten (Vorvorfeld) unterschieden werden muss, sondern auch eine Position danach für Parenthesen und Konnektoren.

Hier geht es nicht darum, einen endgültigen und stimmigen theoretischen Ansatz für den Bereich vor dem Vorfeld zu liefern. Beabsichtigt ist lediglich, die verschiedenen Ansätze grob vorzustellen, um daraus Entscheidungen für eine sinnvolle Annotation abzuleiten. Ich denke aber, dass deutlich geworden ist, dass eine zweckmäßige Einteilung, die noch mehr Information liefert, als man durch eine Auflistung der einzelnen Elemente ohnehin schon erhält, ein schwieriges Unterfangen ist.

Deshalb wird in der Falko-Felder-Annotation keine weitere Einteilung vor dem finiten Verb vorgenommen, im Gegensatz zu TüBa-D/Z und Verbmobil, die außerdem zwischen LV (Linksversetzung), PARORD (*denn, weil*) und KOORD (Koordinationspartikel, *und, oder* usw.) unterscheiden. Letzteres wird im folgenden Abschnitt ausführlicher diskutiert.

### **2.2.1.1 Satzkoordination**

Bei der Annotation der Satzkoordination sind zwei Punkte wichtig: zum einen muss bei der Einteilung in Matrixsätze entschieden werden, ob und zu welchem „Satz“ der Konjunktoren zugeordnet werden soll. Zum anderen muss geklärt werden, ob er ein eigenes Felder tag erhält.

Die Einordnung des Konjunktors (Satzkoordination) in das Feldermodell wird in der Literatur unter unterschiedlichen Bezeichnungen vielfach behandelt. Im Allgemeinen besteht aber Übereinkunft darin, seine Position nicht als Teil des Vorfelds anzusehen. Bei der Koordination von vollständigen Sätzen gibt es zwei gängige Erklärungsansätze bezüglich dem Status der Satzkonjunktoren im Stellungsfeldermodell. Sie unterscheiden sich darin, ob der Konjunktoren als unabhängiges Element zwischen zwei „Sätzen“ gesehen wird oder ob er als ein Element des zweiten Satzes angesehen wird.

Koordination als Zwischenposition in der IDS-Grammatik (Hoberg 1997, S. 1578):

*Außerhalb des Vorfelds stehen auch satzverknüpfende Konjunktoren, die ja gerade dadurch definiert sind, daß sie die Zwischenposition zwischen den koordinierten Einheiten einnehmen ...*

In Anlehnung daran wird dies in Pasch (2003, S. 70) als Nullstelle bezeichnet „... die Position zwischen den Konnekten<sup>58</sup>.“

Satzposition (Höhle 1986, S. 332):

*Beiordnende (koordinierende und nicht-koordinierende) Partikel gehören [...] zu dem Satz, den sie einleiten. [...] Solche Sätze kann man äußern, ohne zuvor eine Satz geäußert zu haben ... (Höhle 1986, S. 332)*

Die Phrasenstrukturen der verschiedenen Ansätze bilden sich wie folgt ab:

(15) *Maria isst Äpfel und Paul isst Birnen.*<sup>59</sup>

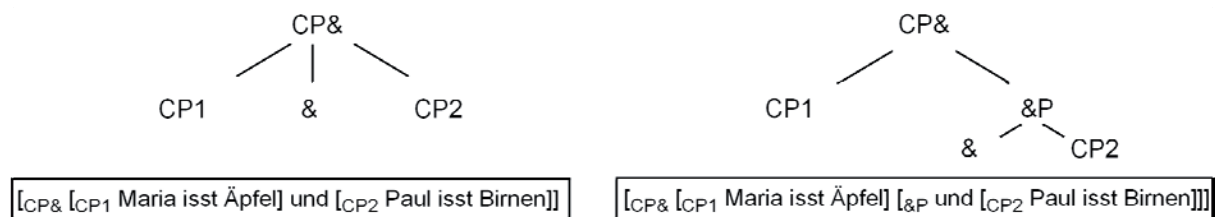


Abbildung 9 Satzkoordination mit Zwischenposition bzw. Satzposition

Manchmal wird davon gesprochen, dass der Konjunktoren zum Vorvorfeld gehört. Aber Beispiel (16) macht deutlich, dass diese Bezeichnung irreführend ist, da selbstverständlich auch V1-Sätze oder Vend-Sätze koordiniert werden, die per Definition kein Vorfeld besitzen.

(16) *Das Essen ist hier schon gut. Aber haben Sie schon bei Marco gegessen?*

Ob daraus notgedrungen resultiert, dass der Konjunktoren nur als Zwischenelement gesehen werden kann, ist fraglich. Wie schon von Höhle oben zitiert, gibt es viele Fälle, in denen der Konjunktoren nur dem Satz, bei dem er steht, zugeordnet werden kann.

(17) *Und nun, was meinen Sie dazu?*

Meiner Meinung nach ist damit überzeugend dargelegt, dass der Konjunktoren dem rechten Konnekt zugeordnet werden kann.

Zur zweiten Frage nach einem eigenen Feldertag ist folgendes anzumerken: ein Nachteil bei der jetzigen Annotation ist, dass koordinierte Sätze nicht ganz leicht zu suchen sind. Allein nach dem Wortarttag „KON“ für Koordinator zu suchen, reicht nicht

<sup>58</sup> die zu koordinierende Strukturen.

<sup>59</sup> In folgenden Ausführungen wird die in Grewendorf et al. (1999, S. 213–227) ausgeführte kombinierte Darstellung benutzt, die das X-bar-Schema mit einer binären hierarchischen Phrasenstruktur in Zusammenhang mit dem Feldermodell bringt. Hierbei werden als funktionale Projektionen CP (complementizer phrase) und IP (inflectional phrase) angenommen.

aus, wie in folgendem Satz mit einer Phrasenkoordination "Max und Moritz" und einer Satzkoordination mit „aber“ erkennbar ist:

(18) *Maria ist doch gerade gekommen aber Max und Moritz sind schon da.*

Abbildung 10 zeigt die Annotation von diesem Satz in der Abbildung bei a) nach der jetzigen Annotation und bei b) nach der Annotation mit Einführung des eigenen Feldertags <KOORD>.

[word]	Maria	ist	gerade	gekommen	aber	Max	und	Moritz	sind	da
[pos]	NE	VAFIN	ADV	VVPP	KON	NE	KON	NE	VAFIN	ADV
[ms]	x				x					
a) [ms_felder]	VF	LSK	MF	RSK	VF			LSK	MF	
b) [ms_felder]	VF	LSK	MF	RSK	KOORD	VF		LSK	MF	

Abbildung 10 KOORD als eigenes Feldertag

Die Suche nach koordinierten Sätzen ist dennoch nach der jetzigen Annotation möglich:

Suche [pos = „KON“] als äußerstes linkes Element in einem VF, das nicht hinter [pos = „\$.“] steht<sup>60</sup>.

Die Einführung von <KOORD> als eigenständiges Feldtag erleichtert die Suche und passt besser zu den gängigen theoretischen Ansätzen, da die Koordination zwar immer noch dem rechten Satz zugeordnet wird , aber nicht wie jetzt dem VF.

Zusammenfassend scheint eine Einführung eines KOORD-Tags zwar sinnvoll aber nicht notwendig.

## 2.2.2 Bestimmung Mittelfeld, Mittelfeldende, rechte Satzklammer

Eine Problematik, die sich für die Annotation stellt, ist, welche Elemente zur RSK bzw. zum Mittelfeldende gerechnet werden sollen und als Abgrenzung zwischen Mittelfeld und Nachfeld bei V1/V2-Sätzen fungieren.

<sup>60</sup> Suche ein Wortart-Tag "KON" als äußerstes linkes Element in einem VF, das nicht hinter einem Wortart-Tag eines schließenden Satzzeichens steht.

Grundsätzlich bei allen Ansätzen werden Verben (Infinitiv mit und ohne Infinitivmarker, Partizip in V1/V2-Sätzen und Infinitiv bei Verbendsätzen) als RSK-Elemente eingeschätzt.

Es gibt aber unterschiedliche Einschätzungen bei der Einordnung folgender „klammer“-bildender Strukturen in das Feldermodell (in Altman 2005 werden sie klammerschließende Elemente genannt):

- Partikel der Partikelverben

(19) *Sie nahm im Kino ihren Hut nicht ab.*

- Teile eines verbalen Idioms (Funktionsverbgefüge (FVG), verbale Idiome)<sup>61</sup>

(20) *Mehrere Möglichkeiten stehen ihnen seit gestern zur Verfügung.*

- obligatorische Lokal- oder Direktionaladverbiale

(21) *Es lag gestern auf dem Tisch.*

(22) *Er ging nach dem Aufstehen direkt in die Schule.*

- Prädikative

(23) *Rot ist nicht meine Lieblingsfarbe.*

- Satznegationen

(24) *Er besuchte den Lehrer trotz der Einladung nicht.*<sup>62</sup>

In Helbig/Buscha (2005) werden alle oben genannte Strukturen gleichwertig als Teile des verbalen Rahmens aufgeführt. Altmann (2005) spricht sich gegen eine Negationsklammer aus, räumt aber ein, dass Prädikative und obligatorische Lokal- und Direktionaladverbiale auch als feste Elemente am Mittelfeldende gesehen werden können, und schließt sie dennoch nicht als rechte Satzklammerelemente aus. Pittner (2007, S. 90–92) schließt Prädikative und obligatorische Lokal- oder Direktionaladverbiale als rechte Satzklammerelemente aus, weil sie als eigenständige Satzglieder zu werten sind. Sie rechnet Verbpartikel und die nicht verbalen Teile der Funktionsverbgefüge zu den RSK-Elementen aufgrund der angenommenen Stellungsfestigkeit der RSK.

In der IDS-Grammatik wird der rechte Innenrand des Mittelfelds (das Mittelfeldende) in drei Bereiche unterteilt:

---

<sup>61</sup> Die Abgrenzung zu Objekt-Inkorporation und verbalen Idiomen ist problematisch und benötigt hohe analytische Leistung oder gute Listen.

<sup>62</sup> zit.n. Helbig und Buscha (2005, S. 476)

MF			RSK
rechter Innenrand			
re3	re2	re1	
Sneg <sup>63</sup>	VG-adverbialia <sup>64</sup>	K-TRM (K <sub>PRD</sub> , sit, dil, dir) +K <sub>PRP</sub>	

Tabelle 4 rechter Innenrand bei IDS-online (grammis<sup>65</sup>)

Dem alleräußersten Bereich des rechten Innenrands (re1) werden die nicht-Termkomplemente (K-TRM<sup>66</sup>), und präpositionalen Komplemente (Präpositivkomplemente, siehe Beispiel (25)) zugeordnet.

Daraus wird erkennbar, dass in der IDS-Grammatik abweichend von den anderen Ansätzen präpositionale Komplemente wie in Beispiel (25)

(25) (a) *Im Blick auf denkbare sowjetische Gegenmaßnahmen wird festgestellt, daß diese „überwiegend auf vorhandenen Technologien“ basierten.*

(b) *\*..., daß diese auf vorhandenen Technologien überwiegend basieren.*<sup>67</sup>

(zit. n. Hoberg 1997, S. 1548)

und auch das Expansivkomplement (Dilativkomplement (dil)) wie in Beispiel (26) zum Mittelfeldende gezählt werden:

(26) *Der Bach stieg 2007 über 3 Meter.*

Im Unterschied zu den anderen Ansätzen werden hier die Funktionsverbgefüge dem Mittelfeldende und nicht der RSK zugeordnet.

In Anbetracht dieser unterschiedlichen Ansätze scheint eine Tendenz zur Unterspezifizierung der RSK als beste Lösung, also nur finite und infinite verbale Elemente und das Verbpartikel mit den Tags <RSK\_MS> bzw. <RSK\_KS> zu versehen. Somit wird den meisten Ansätzen zwar nicht entsprochen aber auch nicht widersprochen. Ein weiterer Vorteil liegt darin, dass dieses Vorgehen TüBa-D/Z entspricht und damit eine bessere Vergleichbarkeit gegeben ist.

Ein weiterer Punkt, der schließlich in diesem Zusammenhang geklärt werden muss, ist die Abgrenzung des Mittel- und Nachfelds bei nichtannotierter RSK.

<sup>63</sup>Sneg = Satznegation

<sup>64</sup>VG-adverbialia sind qualitative Adjunkte (z.B. Adjektive - fest, eng, Adverbien - gern, sehr, Partizipien - erholt, gelangweilt)

<sup>65</sup>[http://hypermedia.ids-mannheim.de/pls/public/sysgram.ansicht?v\\_typ=d&v\\_id=787](http://hypermedia.ids-mannheim.de/pls/public/sysgram.ansicht?v_typ=d&v_id=787)

<sup>66</sup>Zu den nicht-Termkomplementen (K-TRM) zählt die IDS-Grammatik Hoberg (1997, S. 1507) Verbativkomplemente (z.B. Funktionsverbgefüge), Lokal- oder Direktionaladverbial (Situativ- (sit) und Direktivkomplement (dir)), und das Prädikativ (KPRD).

<sup>67</sup>Zu der Gruppe der präpositionalen Komplemente bemerkt Hoberg (1997), dass die obligatorische Zuordnung der verschiedenen Komplemente zum äußersten rechten Innenrand von der Stärke der Bindung zum Verb abhängt. Dies ist eine problematische Feststellung, wenn versucht wird, klare Richtlinien aufzustellen.

Nach Abwägung der aufgeführten Ansätze wurde entschieden, alle hier aufgezählten Elemente bis auf die Negation<sup>68</sup> und die präpositionalen Komplemente als Mittelfeldend-Elemente (MFe)<sup>69</sup> festzulegen, nach deren Aufkommen <NF\_MS> getaggt werden soll.

Zum Abschluss muss noch geklärt sein, ob NF getaggt werden, auch wenn keine klammerbildene Elemente vorhanden sind.

Altmann beschreibt das Phänomen der „offenen“ Klammer und führt folgendes Beispiel auf

- (27) (a) *Er denkt an sie Ø bei Tag und Nacht.*
- (b) *Er hat an sie gedacht bei Tag und Nacht*<sup>70</sup>
- (c) *Er hat an sie bei Tag und Nacht gedacht.*

(b) stellt die „intuitiv richtige Platzierung“ des Partizips nach Altmann dar<sup>71</sup>. Aber (c) stellt meines Erachtens auch eine gültige Variante dar. Deshalb plädiere ich zu Gunsten der Konsistenz dafür, dass bei uneindeutigen Fällen NF nicht annotiert werden sollen.

### 2.2.3 Besetzung von Nachfeld und Zuordnung zum Satz

Der Aufbau des Nachfelds wird an dieser Stelle nicht so ausführlich besprochen, seine mögliche Besetzung ähnelt in vielen Punkten dem Vorfeld. In der IDS-Grammatik (S. 1649) wird das Nachfeld als „Satzabschnitt hinter dem (virtuellen) rechten Satzklammerteil“ definiert. Die Bezeichnung „Satzabschnitt“ soll auf die syntaktische Integration der darin befindlichen Elemente hindeuten. Hier befinden sich Konstituenten, die aus dem Satz ausgelagert wurden. Meist sind es Nebensätze aber auch Appositionen, als-/wie-Phrasen (Modifikatoren) und andere. In den meisten Fällen kann analog zu der Entscheidung beim Vorfeld alles nach der RSK bis zum Satzendezeichen als NF analysiert werden. Im Gegensatz zu TüBa-D/Z kann man sich bei Lernerdaten nicht zuverlässig an der Interpunktion orientieren für die Einteilung in NF, siehe Beispiele in Abschnitt 2.4.1.1, in dem diese Problemfälle besprochen werden.

---

<sup>68</sup> Die Negation unterscheidet sich auch von den anderen Strukturen dadurch, dass sie zusammen mit den anderen Strukturen auftreten kann, während das Vorkommen der anderen sich gegenseitig ausschließt. Das Präpositionalkomplement wird ausgeschlossen, weil es nicht in allen Fällen als MFe zu sehen ist (siehe Fußnote 67).

<sup>69</sup> Dabei wird angenommen, dass MFe-Elemente unter anderen mit Präpositionalphrasen eine syntaktische Einheit bilden können wie in (b):

(a) Es ist für ihn [wichtig]<sub>MFe</sub> [nach Hause zu kommen]<sub>NF</sub>.

(b) Es ist [wichtig für ihn]<sub>MFe</sub> [nach Hause zu kommen]<sub>NF</sub>.

<sup>70</sup> Altmann und Hahnemann (2005, S. 46).

<sup>71</sup> Nach der IDS-Grammatik wäre dieses empfundene Nachfeld durch die Mittelfeldende-Zugehörigkeit von dem präpositionalen Komplement „an sie“ zu erklären.

## 2.2.4 Ausgewählte syntaktische Phänomene

Leider können im Rahmen dieser Arbeit nicht alle interessanten syntaktischen Phänomene ausreichend diskutiert werden, deshalb werden hier zwei ausgewählte Strukturen aufgeführt<sup>72</sup>, die sich in Bezug auf ihre Einteilung in das Feldermodell als problematisch und besprechungswürdig erwiesen haben und die von den Annotationsschemata von TüBa-D/Z und Verbmobil abweichen. Da eigentlich eine Vergleichbarkeit angestrebt wird, werden im Folgenden die Gründe für die abweichenden Festlegungen dargelegt.

### 2.2.4.1 Kohärente und inkohärente Strukturen

Bech (1955) beschreibt das Phänomen der Kohärenz und Inkohärenz verbaler Argumente. In Bezug auf ihre syntaktischen Bindungseigenschaften unterscheidet Bech drei Gruppen von Verben, die ein Infinitivargument fordern. Die finiten Verben der ersten Gruppe gehen eine feste Bindung mit dem Infinitiv ein (kohärente Verben), während die Verben der zweiten Gruppe dies können, aber nicht müssen (fakultativ kohärente Verben), und die dritte Verbgruppe keine feste Bindung bildet (inkohärente Verben). Folgendes Diagramm gibt einen Überblick, wobei die Beschreibung des Phänomens sehr umstritten ist, ebenso wie die Zuordnung mancher Verben.

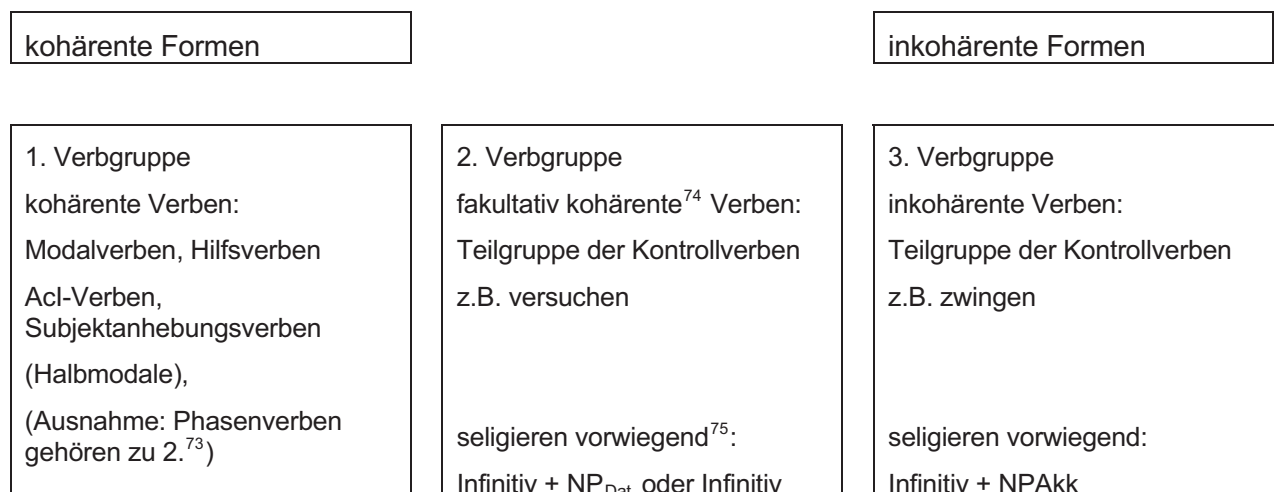


Abbildung 11 3 Verbgruppen – Kohärenz und Inkohärenz

<sup>72</sup> Siehe die Dokumentation für eine quantitativ ausführliche Auflistung/Beschreibung der Annotation anderen Phänomene.

<sup>73</sup> Müller (2007, S. 260)

<sup>74</sup> Bezeichnung nach Pittner und Berman (2007)

<sup>75</sup> Einteilung der Seligierungsmerkmale der Verben nach Sabel (2002); Müller (2007, 268) widerspricht dieser absoluten Einteilung und meint, dass Kontrollverben, die ein Akkusativargument seligieren, auch kohärente Strukturen bilden können.

Als kohärent gelten Strukturen, in denen das finite Verb zusammen mit seinen verbalen Argumenten (Verbalprojektion) einen Verbkomplex bildet, in dem der Infinitiv eingebettet ist. Diese werden als monosententiale Strukturen gewertet.

- (28)(a) *Sie scheint morgen zu kommen*  
 (b) *dass sie morgen [zu kommen [scheint]]*  
 (c) *\*dass sie morgen scheint zu kommen*

Daraus folgt für die Einteilung der Felder für (28) (a):

Sie	scheint	morgen	zu kommen
VF	LSK	MF	RSK

und für (28) (b):

dass	sie morgen	zu kommen scheint
LSK	MF	RSK

Als inkohärent werden Strukturen bewertet, bei denen das Verb ein Infinitivargument fordert, das als Teilsatz mit eigener Feldstruktur (Pittner und Berman 2007, S. 121) gesehen wird<sup>76</sup>.

- (29)(a) *Sie zwingt ihn, sich zu entschuldigen.*  
 (b) *dass sie ihn [zwingt, [sich zu entschuldigen]]*

Daraus folgt für die Annotation folgende Feldereinteilung für (29) (b).

[word]	dass	sie	ihn	zwingt	sich	zu	entschuldigen
[ks_1]	x						
[ks_1 felder]	LSK_KS	MF_KS		RSK_KS	NF_KS		
[ks_2]					x		
[ks_2 felder]					MF_KS	RSK_KS	

analog dazu (29) (a)<sup>77</sup>

Sie	zwingt	ihn	[sich zu entschuldigen] <sub>KS</sub>
VF_MS	LSK_MS	MF_MS	NF_MS

<sup>76</sup> Diese Annahme ist auch umstritten. Es gibt eine Reihe von Autoren, die diese Strukturen als monosentential werten, z.B. Haider (1993). Siehe hierzu Sabel (2002, S. 152)

<sup>77</sup> Natürlich wäre aus (a) ebenso folgende Umformung denkbar::

*dass [sie ihn [sich zu entschuldigen]<sub>KS</sub>]<sub>MF</sub> zwingt.*

Demzufolge wäre auch eine Einordnung in das MF möglich. In diesem Annotationsschema ist festgelegt worden, diese Infinitivteilsätze als Nachfeld zu annotieren, da sie oft ins Nachfeld extrapoliert werden.

Für die Annotation fordern die Kontrollverben der Verbgruppe 2 größere analytische Leistungen, weil sie sowohl inkohärente Strukturen als auch kohärente Strukturen bilden.

Als Festlegung für die Annotation werden auch folgende Konstruktionen der Verbgruppe 2 in V2/V1-Sätzen als Teilsatzkonstruktionen (mit dem größtmöglichen Teilsatz) annotiert:

- (30)(a) *Gestern versuchte, Maria* *[[sich zu erinnern]<sub>KS</sub>]<sub>NF</sub>*  
(b) *Maria hat versucht, [[sich zu erinnern]<sub>KS</sub>]<sub>NF</sub>*

dagegen

- (31)(a) *Gestern versuchte sich Maria* *[[zu erinnern]<sub>KS</sub>]<sub>NF</sub>*  
(b) *Gestern hat sich Maria versucht* *[[zu erinnern]<sub>KS</sub>]<sub>NF</sub>*  
(c) *\*Gestern hat versucht sich Maria zu erinnern.*

Schwieriger zu analysieren sind Sätze der Verbgruppe 2, bei denen sich der eingebettete Infinitivsatz direkt vor dem finiten Verb befindet, weil sie sowohl inkohärent als auch kohärent sein können.

Bei Bech (1955) finden sich diverse Tests, die der Unterscheidung von kohärenten und nichtkohärenten Strukturen dienen, z.B. Permutation in Mittelfeld<sup>78</sup>.

Unter Anwendung dessen führt Pittner folgendes Beispiel für eine inkohärente (a) und eine kohärente/monosententiale (b) Struktur auf und analysiert diese wie folgt:

- (32)(a) *dass Ella* *[sich zu erinnern]<sub>KS</sub> versucht.*  
(b) *dass sich Ella* *[zu erinnern versucht]<sub>RSK</sub>.*

(Pittner und Berman 2007, S. 121)

*[sich zu erinnern]* entspricht dem Infinitiv-Argument. Wenn aber wie in (b), das „*sich*“ vor dem Subjekt „*Ella*“ in der „unmarkierten“ Stellung für ein Pronomen steht, muss eine monosententiale Struktur vorliegen, da „Scrambling“<sup>79</sup> im Deutschen nicht über Satzgrenzen hinweg auftreten kann<sup>80</sup>.

Diese Analyse scheint für Daten von Lernern, die im Vergleich zu Muttersprachlern gerade Abweichungen im Mittelfeld zeigen, problematisch. Daraus resultierte die Entscheidung, sowohl (32) (a) als auch (31) (b) als monosentential zu annotieren mit „*zu erinnern versucht*“ als RSK, auch entgegen der Festlegungen bei TüBaD/Z, wo Strukturen wie (32) (a) als Teilsatz annotiert wird, weil sie in das Nachfeld extrapoliert werden können.

---

<sup>78</sup> auch Skopus von Adjunkten, Extraposition, Intraposition, Voranstellung im Vorfeld und „langes“ Passiv

<sup>79</sup> Scrambling bezeichnet das Phänomen, dass Konstituenten unterschiedliche Stellungen im Mittelfeld haben können.

<sup>80</sup> Es gibt auch Ansätze, diese Strukturen zu erklären ohne die Annahme der Monosententialität Sabel (2002).

Das bedeutet, dass das Infinitiv-Argument nur für Verben der Gruppe 3 als satzwertiger Konstituent im MF annotiert wird<sup>81</sup>, wie auch im folgenden Beispiel:

(33) *dass [ihn Ella [sich zu entschuldigen]<sub>KS</sub>]<sub>MF</sub> zwang.*

(zit. n. Pittner und Berman 2007, S. 121)

#### 2.2.4.2 Zustandspassiv oder Kopula

Ein Problem bereitet die Einordnung folgender Konstruktionen bestehend aus einer Partizip II-Form und einer Form von „sein“:

(34) *Die Tür ist geschlossen.*

(35) *Er ist verrückt*

Die weitverbreitetste Analyse für Konstruktionen wie (34) ist, sie als Zustandspassiv, ein eigenes Genus verbi, einzuordnen, wie z.B. Helbig (2001), IDS-Grammatik (1997) und Eisenberg (1999).

(36) *Der Tür ist [geschlossen]<sub>VVPP</sub> [Part II: Verb + sein: Zustandspassiv-Auxiliar]*

„geschlossen“ steht in der RSK.

Das Partizip Beispiel (37) kann aber von der Bedeutung her nicht als Zustandspassiv gewertet werden, weil „verrückt“ nicht eindeutig auf die semantische Bedeutung des Verbs „verrücken“ zurückzuführen ist<sup>82</sup>.

Somit muss „verrückt“ als partizipiales Adjektiv (ADJD) annotiert werden.

(37) *Er ist [verrückt]<sub>ADJD</sub>.*

„verrückt“ steht am MFe.

Daraus folgt, dass grundsätzlich zwischen VVPP und ADJD entschieden werden muss.

Folgendes Beispiel zeigt die Problematik bei der Unterscheidung:

(38) *Der Brief war geöffnet*      *Der Brief wurde geöffnet.*

(39) *Der Brief war ungeöffnet*      *\*Der Brief wurde ungeöffnet.*

„geöffnet“ in (38) wird als VVPP und „ungeöffnet“ in (39) als ADJD annotiert. Dadurch werden sie zu unterschiedliche Wortartklassen gezählt, was nicht unbedingt plausibel scheint.

Es gibt auch Erklärungsansätze, alle hier beschriebenen Sprachkonstruktionen als Kopula-Konstruktionen zu analysieren (Maienborn 2007).

---

<sup>81</sup> Dies ist natürlich nur machbar, wenn sich die 2. Verbgruppe klar abgrenzen lässt.

<sup>82</sup> Laut TüBa-D/Z Style book (S. 124) werden die Entscheidungskriterien des Handbuchs vom STTS befolgt: Kann der Satz in eine aktive Form mit der gleichen Semantik umgewandelt werden, wird VVPP getaggt. Für (37) ist dies nicht möglich. Also wird ADJD getaggt.

Folgende Beispiele (Maienborn 2007, S. 8) demonstrieren wie uneindeutig die Entscheidungen sein können:

- (40) (a) *Unsere Kabinen sind videoüberwacht.*  
(b) *Unsere Kabinen werden videoüberwacht*  
(c) *Wir lassen unsere Kabinen videoüberwachen*  
(d) *\*während wir die Kabinen videoüberwachten*

(c) stellt eine mögliche Umwandlung in eine aktive Form mit Beibehaltung der Semantik dar, (d) dagegen nicht.

Deshalb wurde zugunsten von Konsistenz entschieden und in Abweichung zur TüBa-D/Z werden alle „sein“ + Partizip-Konstruktionen als Kopula-Konstruktionen annotiert.

## **2.3 Kriterien für die Entscheidung kanonisch - nichtkanonisch**

Ein zentraler Punkt der Annotation ist, dass zwischen kanonisch und nichtkanonisch unterschieden wird. In diesem Abschnitt werden vorwiegend anhand von Lernerbeispielen systematisch die konkreten Unterscheidungskriterien besprochen und ihre Auswirkung auf die Annotation im Einzelnen wird aufgezeigt.

Hier gibt es verschiedene Aspekte, die etwas genauer betrachtet werden müssen. Neben der strikten Überprüfung der Kriterien, die zur Unterscheidung von kanonisch und nichtkanonisch aufgestellt wurden, muss auch auf den Aspekt von Einheitlichkeit und Konsistenz eingegangen werden: gleiche Phänomene sollen gleich behandelt werden.

Im Falko-Annotationsschema werden Fehler als Abweichung zu einer Zielhypothese definiert. Wie S. 28 dargestellt wurde, kann das Feldermodell nicht angewendet werden, wenn die linke Satzklammer (rechte Satzklammer) nicht bestimmbar ist. Es geht also um Abweichung in den Satzklammern und die Frage, welche Wortstellungsabweichungen der Satzklammer die Bestimmbarkeit unmöglich machen. Genügt es, Verbstellungsfehler und Auslassungsfehler in den Satzklammern (Nichtrealisierung der LSK bzw. der obligatorischen RSK) als Kriterien zur Unterscheidung zwischen kanonisch und nichtkanonisch anzunehmen? Dafür muss zunächst geklärt sein, was unter Verbstellungsfehlern zu verstehen ist.

### **2.3.1.1 Verbstellungsfehler (nichtkanonisch)**

Verbstellungsfehler können nicht allein als Stellungsfehler des finiten Verbs definiert werden, siehe Beispiel (42). Dies bedeutet, dass es sich um Wortstellungsabweichungen handelt, die die verbale Klammer in V1/V2-Sätzen bzw.

die Nebensatzeinleiter-Verb-Klammer in Verbend-Sätzen betreffen. Verbstellungsfehler werden als ungrammatisch empfunden. Man unterscheidet zwischen **Verbzweitfehlern**

(41) *Originaläußerung (O): Seiner Ansicht nach die Person des Autors ermöglicht auch das Verstehen...*

*Zielhypothese (Z): Seiner Ansicht nach ermöglicht die Person des Autors auch das Verstehen ...*

(angelehnt an FU\_021)

## und **Verbletzfehlern**

### a. Verbalklammer

(42) *O: Wir können das beschreiben auf einen Beispiel mit nicht ernsthaften Menschen.*

*Z: Wir können das an einem Beispiel mit nicht ernsthaften Menschen beschreiben*

(FU\_003)

### b. Nebensatzeinleiter-Verb-Klammer

(43) *O: ... besagt, dass für jede Bedeutung existiert eine Form.*

*Z: ... besagt, dass für jede Bedeutung eine Form existiert.*<sup>83</sup>

(FU\_043)

(siehe auch Darstellung Verbstellungsfehler im Annotationsschema, S. 29).

### 2.3.1.1.1 Sonderfall der Wortstellungsfehler in der RSK:

Grundsätzlich werden Äußerungen mit Wortstellungsfehlern, die die Wortabfolge innerhalb eines Felds betreffen und nicht „felderübergreifend“ beschrieben werden müssen, in dem Feldermodell eingeteilt, z.B. Wortstellungsabweichungen innerhalb des Mittelfelds (<MF\_MF> und <MF\_MFe>)<sup>84</sup>, aber auch innerhalb von Phrasen<sup>85</sup>. Dieses Vorgehen lässt sich dadurch begründen, dass nach dem Feldermodell die Wortstellung innerhalb der einzelnen Felder nicht in Abhängigkeit zur Wortstellung des restlichen Satzes erklärt werden muss. Damit ist es möglich, auch Wortstellungsfehler innerhalb der einzelnen Felder getrennt zu beschreiben.

Es stellt sich die Frage, ob dies auch für die RSK angenommen werden kann. Beispiel (44) zeigt einen Wortstellungsfehler innerhalb der RSK bei einem V1/V2-Satz:

(44) *O: Die Bedeutung von Märchen ist [worden abgewertet]<sub>RSK</sub>*

*Z: Die Bedeutung von Märchen ist [abgewertet worden ]<sub>RSK</sub>*<sup>86</sup>

---

<sup>83</sup> Bei der Perfektbildung von Modalverben (IPP) wird das finite Verb nicht am rechten Rand der RSK gebildet:

(b) *Dass sie habe kommen wollen, wusste ich nicht.*

<sup>84</sup> theoretisch auch Wortstellungsabweichungen im Nachfeld (<NF\_NF>); bzw. Vorfeld(<VF\_VF>); aber in der Regel bestehen diese Felder aus nur einem Konstituenten.

<sup>85</sup> Umgekehrt bedeutet dies nicht, dass jede scheinbar felderübergreifende Wortstellungsänderung als Verbstellungsfehler zu sehen ist, siehe das Beispiel eines Bedeutungsfehlers (57).

<sup>86</sup> Im Korpus gab es keinen Fehler dieser Art.

(angelehnt an FU\_046)

Die Darstellung der Äußerung (44) in Abbildung 12(a) zeigt, dass sich diese Art von Fehlern innerhalb eines begrenzten Bereichs innerhalb von  $V^0$  abspielt.

Strukturell anders sieht es aus, wenn bei Vend-Äußerungen die Stellung des finiten Verbs betroffen ist, wie in Beispiel (45)

- (45) O: „Märchen“ sei keine Kunst, da die Bedeutung von Märchen [werde abgewertet]<sub>RSK?</sub>  
Z: „Märchen“ sei keine Kunst, da die Bedeutung von Märchen [abgewertet werde]<sub>RSK</sub>

(FU\_046)

Wie in Abbildung 12 (b) zu sehen ist, müsste das finite Verb in dem Inflektionskopf  $I^0$  abgebildet sein (Sabel 2000).<sup>87</sup>

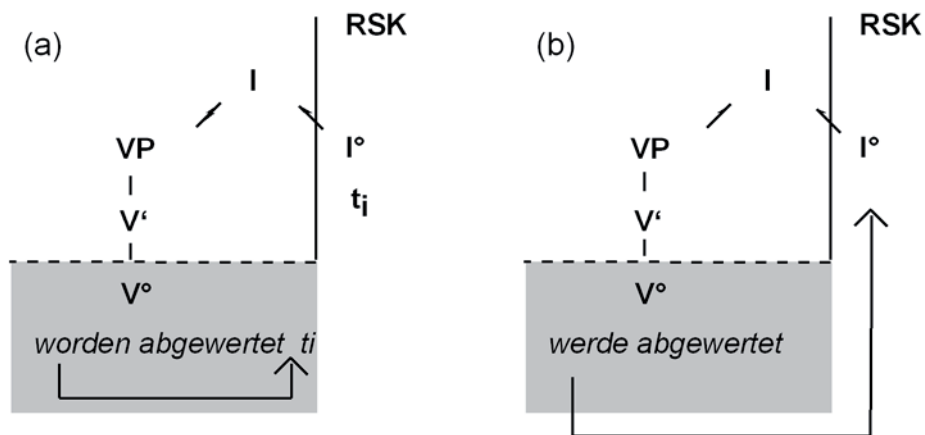


Abbildung 12 Verbstellungsabweichungen in der RSK

Hier ist nicht so eindeutig, dass das Verb der Zielhypothese und das Verb der Lerneräußerung im gleichen „Feld“ stehen.

Ein weiteres Argument, das dafür spricht, hier Wortfolgefehler des finiten Verbs als nichtkanonisch zu bewerten, liegt einfach darin, dass hier Verbend nicht realisiert wurde und es somit als Verbstellungsfehler zu sehen ist, auch wenn die sonstigen Felder scheinbar davon nicht betroffen sind. Es gibt bei Falko sehr wenige Beispiele wie (46), die meisten Fehler dieser Art sind Fehler in der Stellung des finiten Verbs und der Verbpartikel wie in Beispiel (46):

- (46)O: Es liegt daran, dass ein Gesprächspartner nimmt ernst  
Z: Es liegt daran, dass ein Gesprächspartner das [ernst nimmt]<sub>RSK</sub>

(FU\_003)

<sup>87</sup> Sabel (2008) verortet sogar das nichtfinite Verbelement im (45).

### 2.3.1.2 Auslassungsfehler (nichtkanonisch)

Im Folgenden wird beschrieben, was unter nicht realisierter LSK bzw. nicht realisierter obligatorischer RSK zu verstehen ist und wie dies im Einzelnen aussieht:

Bei V1/V2-Sätzen bildet das finite Verb die LSK, was nach dem Feldermodell obligatorisch ist. Daraus folgt, dass der Auslassungsfehler das finite Verb betreffend als nichtkanonisch annotiert werden:

(47)O: *Wie [Ø]<sub>LSK</sub> möglich, über einen bestimmten Sachen genau zu erklären.*

Z: *Wie ist es möglich, eine bestimmte Sache genau zu erklären.*

(angelehnt an FU\_084)

Dagegen ist nicht jede fehlende rechte Satzklammer nichtkanonisch, weil die RSK in V1/V2-Sätzen nicht grundsätzlich obligatorisch ist. Bei bestimmten Verben ist sie aufgrund ihrer syntaktischen Realisierung aber erforderlich, wie in Beispiel (48): hier wurden obligatorische verbale Argumente nicht realisiert.<sup>88</sup>

(48)O: *Die bunte Palette wurde auf verschiedenen, mehrdeutigen Ebenen von Körper und Seele [Ø]<sub>RSK</sub> [...]*

Z: *Die bunte Palette wurde auf verschiedenen, mehrdeutigen Ebenen von Körper und Seele [aufgenommen]<sub>RSK</sub> [...]*

(angelehnt an FU\_081)

Es wäre immer noch denkbar, hier eine Feldereinteilung durchzuführen, da die Äußerung ein finites Verb und ein Subjekt enthält.

Folgendes Beispiel (49) zeigt, dass keine eindeutige Feldereinteilung möglich ist. Es gibt neben mehreren nichtkanonischen mindestens zwei kanonische Stellungsmöglichkeiten:

(49)O: *Die bunte Palette wurde auf verschiedenen, mehrdeutigen Ebenen von Körper und Seele [Ø]<sub>RSK</sub> [...]*

Z1: *Die bunte Palette wurde [auf verschiedenen, mehrdeutigen Ebenen von Körper und Seele]<sub>MF</sub> [aufgenommen]<sub>RSK</sub> ...*

Z2: *Die bunte Palette wurde [aufgenommen]<sub>RSK</sub> [auf verschiedenen, mehrdeutigen Ebenen von Körper und Seele]<sub>NF</sub>*

(angelehnt an FU\_081)

Deshalb werden Äußerungen mit fehlenden obligatorischen verbalen Argumenten der RSK mit <f\_MS> getaggt, auch bei fehlendem Verbpartikel:

---

<sup>88</sup> Grundsätzlich könnte auch dafür argumentiert werden, dass beim Fehlen jeglichen obligatorischen Arguments keine Einteilung in das Feldermodell möglich ist. Dies würde aber zum Wegfall heuristischer Segmentierung führen, die es erlaubt das Mittelfeld als eigenständigen Bereich zu untersuchen.

(50)O: *Dieses Kriterium hängt von der Größe eines Textes oder einer Aussage nicht [Ø]<sub>RSK</sub>*  
Z: *Dieses Kriterium hängt von der Größe eines Textes oder einer Aussage nicht ab.*<sup>89</sup>

(FU\_053)

Ein weiterer Punkt ist, ob eine Einteilung möglich ist, wenn Elemente der RSK fehlen, wie in Beispiel (51)

(51)O: *Er hat es nie [sagen Ø]<sub>RSK</sub>, und auch nicht so gemeint.*  
Z: *Er hat es nie [sagen wollen]<sub>RSK</sub>, und auch nicht so gemeint.*

Die Stellung der RSK ist im obigen Beispiel erkennbar und deshalb wurde festgelegt, eine Einteilung durchzuführen.

In Vend-Sätzen ist es grundsätzlich problematischer, davon auszugehen, dass beim Fehlen eines Elements die Einteilung durchgeführt werden darf. Fehlt das finite Verb, wie im folgenden Beispiel, ist eine Einteilung meiner Meinung nach nicht möglich:

(52)O: *Integration ist unmöglich, wenn die Europäische Union keine zentrale Regierung [gebildet Ø]<sub>RSK</sub>.*  
Z: *Integration ist unmöglich, wenn die Europäische Union keine zentrale Regierung [gebildet hat]<sub>RSK</sub>.*

(Anlehnung an GU4\_2080)

Ein Satz muss mindestens ein finites Verb besitzen (*Lauf!*) um als kanonisch eingestuft zu werden. Sätze ohne finites Verb werden als nichtkanonisch eingeordnet. Ebenso werden Vend-Sätze mit fehlender LSK als nichtkanonisch getaggt:

(53)O: *davon geht aus Folgerungseigenschaften, [Ø]<sub>LSK</sub> zum Schluss entsprechende Anforderungen an die Entwicklung der Logik stellen.*  
Z: *davon gehen Folgerungseigenschaften aus, [die]<sub>LSK</sub> zum Schluss entsprechende Anforderungen an die Entwicklung der Logik stellen.*

(FU\_084)

Zum Abschluss soll vollständigkeithalber erwähnt werden, dass es Äußerungen gibt, die einfach nicht annotierbar sind, weil sie (meistens) wie im Beispiel (54) aus nicht zusammenpassenden Satzversuchen bestehen, bzw. zwei finite Verben besitzen:

(54) *Aber in der Neuzeit hatten Friedrich Schlegel und Friedrich Schleiermacher machten sie zur wichtigen Methode des Geisteswissenschaft, mit der man die Wechselbeziehung zwischen der Bedeutung der einzelne Worte und der des Gesamtkontextes erörten und erklären kann.*<sup>90</sup>

(FU\_009)

---

<sup>89</sup> Dieser Fehler mag ein Flüchtigkeitsfehler („mistake“, nicht „error“) sein, aber ich möchte an dieser Stelle betonen, dass die Fehleranalyse auf einer eigenen Annotationsebene stattfindet.

<sup>90</sup> In der Neuzeit hatten F.S. und F.S. sie zur wichtigen Methode gemacht [...] oder F.S. und F.S. machten sie zur wichtigen Methode [...]

### 2.3.1.3 kanonische Wortstellungsabweichungen in LSK oder RSK

Als nächstes muss geklärt werden, welche Wortstellungsabweichungen der Satzklammer gegenüber einer Zielhypothese nicht als Verbstellungs- oder Auslassungsfehler zu werten sind. Hierfür werden Kriterien benötigt, die eine Unterscheidung zwischen verschiedenen Formen der Abweichung zur Zielhypothese zulassen. Gerade wenn entschieden werden muss, ob eine Äußerung mit dem Feldermodell beschrieben werden kann oder nicht, dürfen Strukturen, die in anderen Kontexten durchaus mit dem Feldermodell beschrieben werden können, nicht als nichtbeschreibbar (<f\_MS> bzw. <f\_KS>) getaggt werden.

Zunächst sollen Tempus-Modus-Abweichungen wie in Beispiel (55) nicht als Wortstellungsfehler (Verstoß gegen die Linearisierungsregeln) bewertet werden:

- (55)O: *dass es ihm später [verwehrt wird]<sub>RSK</sub>.*  
Z: *dass es ihm später [verwehrt werden wird]<sub>RSK</sub>*

(FU\_014)

Auch solche stilistischen Unterschiede sind auf jeden Fall kanonisch:

- (56)O: *Theodor Fontane trug etwa mit seinem Aufsatz "Unsere lyrische und epische Poesie seit 1884" (1853) auch zum Realismus, bei.*  
Z: *Theodor Fontane trug auch zum Realismus bei, etwa mit seinem Aufsatz "Unsere lyrische und epische Poesie seit 1884" (1853).*

(angelehnt an FU\_032)

Außerdem sind Bedeutungsfehler, wie in Beispiel (57), die auf jeden Fall als Wortstellungsfehler eingestuft werden müssen, keine Verbstellungsfehler, da beide Bedeutungen „grammatikalisch“ denkbar sind und damit auch kanonisch:

- (57)O:B1: *Verschiedene Wörter bilden lexikalische Felder bzw. Wortfelder aufgrund ihrer Bedeutung.*  
Z: B2: *Aufgrund ihrer Bedeutung bilden verschiedene Wörter lexikalische Felder bzw. Wortfelder.<sup>91</sup>*

(angelehnt an FU\_087)

Der Bedeutungsfehler geht nicht auf die Stellung des Verbs zurück, und auch nicht auf ein „felderübergreifendes“ Phänomen, sondern hängt von der Stellung des Modifikators „*auf Grund ihrer Bedeutung*“ ab. Nach dem Adjazenzprinzip, dem Prinzip der möglichst geringen Distanz, modifiziert die Präpositionalphrase in der ersten Bedeutung (B1) „*lexikalische Felder bzw. Wortfelder*“ und in der zweiten (B2) „*verschiedene Wörter*“. Besonders deutlich wird dies bei der Umformung in (58):

---

<sup>91</sup> Folgende Zielhypothese wäre ebenso denkbar, Z:B2: *Verschiedene Wörter bilden aufgrund ihrer Bedeutung lexikalische Felder bzw. Wortfelder.*

(58)B1: dass verschiedene Wörter lexikalische Felder bzw. Wortfelder aufgrund ihrer Bedeutung bilden.

B2: dass verschiedene Wörter aufgrund ihrer Bedeutung lexikalische Felder bzw. Wortfelder bilden.

Schließlich können sich auch in Bezug auf die Besetzung der Satzklammer die von dem Lerner verwendeten syntaktischen Formen von denen der Zielhypothese unterscheiden. Diese Abweichungen sollen ebenso nicht als nichtkanonisch getaggt werden. Es kann sich dabei um stilistische Unterschiede handeln, wie in Beispiel (59)

(59)O:Es wurde gepflegt, damit das Seelenleben deutlich [gemacht wird]<sub>RSK</sub>.

Z: Es wurde gepflegt, um das Seelenleben deutlich [zu machen]<sub>RSK</sub>.

(FU\_80)

Zusammenfassend lässt sich festhalten, dass die Festlegungen auf S. 28 im Großen und Ganzen zutreffen, zusätzlich muss jedes Fehlen des finiten Verbs und der „obligatorischen“ RSK in V1/V2-Sätzen (RSK<sub>V1/V2</sub>) als Bedingung festgelegt werden für die Nichtkanonizität. Äußerungen mit Wortstellungsabweichungen der RSK<sub>V1/V2</sub> (Vfin nicht betroffen) sind als kanonisch zu taggen.

### 2.3.2 Konsistenz: von der Lerneräußerung ausgehen

Wie schon in Abschnitt 1.4.2.3 argumentiert wurde, muss deutlich unterschieden werden zwischen der Annotation der „tatsächlichen Struktur“<sup>92</sup> und der der Hypothese, egal wie fundiert sie sein mag. Daraus ergibt sich die grundlegende Regelung, für die Entscheidung kanonisch – nichtkanonisch bei der Felderannotation von der Lernerlexik auszugehen. Beispiel (60) verdeutlicht die Bedeutung dieser Festlegung für die Konsistenz der Felderannotation. Die erste Zielhypothese (Z<sub>1</sub>) entspricht der Festlegung (und wird <f\_MS> getaggt, da die Stellung des Verbpartikels inkorrekt realisiert wurde). In (Z<sub>2</sub>) wird eine genau so denkbare Zielhypothese aufgestellt, die aber eine Abweichung von der Lexik zugrunde legt und auf keine Stellungsfehler bezüglich der Satzklammer hindeutet (also nicht zwangsläufig als <f\_MS> getaggt wird).

(60)O:Die Handlungsinitiative übergeht ganz auf den Türhüter, [...]

Z1:Die Handlungsinitiative geht ganz auf den Türhüter über, [...]

Z2:Die Handlungsinitiative überträgt sich ganz auf den Türhüter, [...]

(FU\_10)

Die eigentliche Problematik liegt darin, dass je nach Kontext, Genre oder stilistischen Präferenzen Z<sub>1</sub> bzw. Z<sub>2</sub> postuliert werden kann und soll. Je nach Forschungsinteresse

---

<sup>92</sup> Die „tatsächliche Struktur“ kann sich natürlich nur auf die jeweiligen Beschreibungsmodelle beziehen, in diesem Fall handelt es sich um eine reine lineare Beschreibung, der an der Oberfläche realisierten Elemente.

können alle Abweichungen wichtig sein. Auf die Wortstellung bzw. Einordenbarkeit in das Feldermodell aber bezogen wäre es inkonsistent, ein und das gleiche Phänomen nicht immer gleich zu annotieren. Ein Vorteil der Mehrebenenarchitektur liegt darin, dass es prinzipiell möglich ist, beide (oder auch mehrere) Zielhypothesen explizit zu machen.

Bei manchen Beispielen ist es schwieriger, zu entscheiden, welche Zielhypothese näher an der Lerneräußerung ist. In Beispiel (61) hilft die Festlegung „näher an der Lerneräußerung bleiben“ allein nicht. Bedeutet das, einen Auslassungsfehler (omission) mit Beibehaltung der Präposition wie in  $Z_1$  oder eine Ersetzung (replacement) der Präposition wie in  $Z_2$  anzunehmen?

(61)O: Ich dachte, nur ein Moment, über meine Familie und das Abendessen.

$Z_1$ : Ich dachte, nur ein Moment, über meine Familie und das Abendessen nach.

$Z_2$ : Ich dachte, nur ein Moment, an meine Familie und das Abendessen.

(GU2\_2106)

Hier muss die weiterführende Festlegung getroffen werden, dass z.B. vom Verb des Lerners ausgegangen werden soll: in diesem Fall also „denken“ und nicht „nachdenken“.

Im nächsten Beispiel ist das Verb „hingehen“. Es gibt auch hier zwei mögliche Realisierungen.

(62)O:Wo ging Herr Sommer?

$Z_1$ :Wo ging Herr Sommer hin?

$Z_2$ :Wohin ging Herr Sommer?

(GU2\_1096)

Eine pragmatische Lösung wäre, wenn keine eindeutige Entscheidung getroffen werden kann, dann von der Zielhypothese auszugehen, die die wenigsten Syntaxveränderungen verursacht. In diesem Fall wäre das die zweite Zielhypothese und diese Äußerung könnte in Felder eingeteilt werden, im Gegensatz zu  $Z_1$ , in der die RSK nicht gebildet wurde. Ob sich solche pragmatischen Lösungen für Beispiele wie (61) und (62) grundsätzlich bewährt, muss sich zeigen. Insbesondere geht es darum diese Problematik aufzuzeigen, der entweder durch Festlegungen oder andere Methoden begegnet werden muss.

## 2.4 Besprechung einiger Problemfälle der Lernerdaten

Es können nicht alle Phänomene besprochen werden, die trotz Festlegung weiteren Klärungsbedarf erforderlich machen. Hier soll anhand einer Auswahl von Beispielen

gezeigt werden, welche Faktoren eine Rolle spielen und eine eindeutige Einteilung erschweren trotz Festlegungen. Es handelt sich dabei um unterschiedliche Fragen zu „Grammatikalität“ (wie sind bestimmte Regelverstöße zu bewerten – als Verstoß gegen das Feldermodell oder nicht?), Einschätzung der Muttersprachler (wer oder was gilt als Autorität?) oder Fehlereinordnung (wie einordnen bei mehreren gleichzeitigen Fehlern?), Komplexität der zielsprachlichen Struktur (wie sollen zielsprachliche Strukturen analysiert werden?). Es folgen einige Beispiele von Problemfällen bei der Einteilung in die Felder und bei der Entscheidung kanonisch – nichtkanonisch.

## 2.4.1 Feldereinteilung

### 2.4.1.1 NF oder neuer Satz

Die Entscheidung, ob ein nachgestellter Satz als Nachfeld oder als neuer Satz zu annotieren ist, lässt sich nicht allein durch die Interpunktion fällen.

In Beispiel (63) müsste, obwohl es sich hier um eine Verbzweit-Struktur handelt, und diese durch einen Punkt abgetrennt ist, „*Wie die Stadt wächst durch die letzte Jahre*“ trotzdem als NF annotiert werden, da der Satz ein Konstituent des vorangestellten „Satzes“ ist<sup>93</sup>.

(63)O: *Es ist über die Veränderungen des Berlins erzählt. Wie die Stadt wächst durch letzte Jahre.*

(FU\_106)

Dagegen können die nachfolgenden Sätze in Beispiel (64): „*Syntax beschäftigt sich nur mit der Analyse von Satzstrukturen*“ und „*die Wahrheit der Aussage wird von der Semantik bestimmt*“ nur als eigenständige Sätze verstanden werden, auch wenn sie nur durch Doppelpunkt bzw. Komma getrennt sind.

(64) O: *In diesem Fall handelt es sich um ein dreistelliges Prädikat, wo die drei Argumenten vorkommen: Syntax beschäftigt sich nur mit der Analyse von Satzstrukturen, die Wahrheit der Aussage wird von der Semantik bestimmt.*

(FU\_102)

### 2.4.1.2 Kohärenz

Wie auch in Abschnitt 2.2.4.1 besprochen, werden Sätze wie (65) als monosentential annotiert, wobei eine nicht monosententiale Lesart möglich ist, mit „*zu unterrichten versucht*“ als RSK („*versuchen*“ wird zu den fakultativ kohärenten Verben gezählt):

---

<sup>93</sup> Diese Festlegung entspricht in etwa dem Vorgehen bei Verbmobil, (Stylebook S. 22): „A simpx [sentence tag] ends where a complete syntactically well-formed sentence ends according to the „longest match“ strategy., also von der weitgefassten möglichen Äußerung auszugehen. Bei Lernerdaten sind allerdings nicht immer syntaktisch wohlgeformte Sätze anzunehmen.“

(65) (a)O: *dass die EU jetzt viele Sprachen in den Schulen [zu unterrichten versucht]<sub>RSK</sub>*  
(b) O': *dass die EU jetzt [viele Sprachen in den Schulen zu unterrichten]<sub>KS</sub> versucht*

(GU4\_0122)

Diese Konstruktionen werden von Muttersprachlern gerne korrigiert:

(66)Z: *dass die EU jetzt versucht, viele Sprachen in den Schulen zu unterrichten.*

Interessant ist dabei die Übergeneralisierung von Verbend-Strukturen der Lerner, aber auch, dass die Muttersprachler diese Konstruktion fast ausschließlich in der Extraposition bevorzugen.

Bei der jetzigen Annotation wurde aus pragmatischen Gründen entschieden, Strukturen wie in (65) bei fakultativ inkohärenten Verben als kohärent wie in (65) (a) zu annotieren. Ob sich diese Entscheidung bewährt, müsste genauer untersucht werden. Die Tatsache, dass die Muttersprachler diese hier von mir als "kohärent" annotierte Strukturen im Nachfeld bevorzugen, lässt berechnete Zweifel zu. Hier spielen mehrere Faktoren eine Rolle und die unklare Forschungslage machen es unmöglich, im Rahmen dieser Arbeit fundierte Aussagen über das Phänomen und sein bestmögliche Annotation zu treffen.

### 2.4.1.3 Zustandspassiv, Kopula, Vorgangspassiv

Wie in Abschnitt 2.2.4.2, S.44 besprochen, wurde unter anderem aus Konsistenzgründen entschieden, nicht zwischen Kopulakonstruktionen und Zustandspassiva zu unterscheiden. Obwohl sich dieses Vorgehen im Großen und Ganzen bewährt hat, gibt es einige Beispiele die Fragen aufwerfen.

Es gibt einige Beispiele wie (67) in den GU-Daten, bei denen es nahe liegt, dass die Lerner eine Passivform bilden wollten<sup>94</sup>:

(67)O: *Wann ich in den See hineinwanderte, starb ich nie. Stattdessen war ich wiedergeboren.*

Z: *[...] Stattdessen wurde ich wiedergeboren.*

(GU2\_3005)

Entgegen der Vermutung, dass es sich um Passivformen handelt und der Lerner „wiedergeboren“ als Verbform meinte, muss nach der oben erwähnten Regelung und nach dem Grundsatz, von der Lernerlexik auszugehen (siehe Abschnitt 2.3.2), „wiedergeboren“ als MF-Element (ADJD) annotiert werden. Hier stehen Konsistenz

---

<sup>94</sup> Es gibt einige Beispiele für dieses Phänomen in den GU-Daten. Im Englischen wird Passiv mit Seinformen (*be*) gebildet. Die Übersetzung für Beispiel (67) lautet vermutlich: „*Instead he was born again.*“

und die bestmögliche Analyse gegeneinander, wobei noch einmal erwähnt sei, dass eine Fehleranalyse diese Information mit aufnehmen kann.

## 2.4.2 Kanonisch oder nichtkanonisch

### 2.4.2.1 Formabweichung oder nichtkanonisch

Im folgenden Beispiel spielen mehrere Faktoren eine Rolle, die soweit wie möglich erst einzeln betrachtet werden müssen:

- (68)O: *dass sie die einzigen Leute im Dorf, es gesehen zu haben sind*  
Z: *dass sie die einzigen Leute im Dorf sind, die es gesehen haben.*

(GU2\_0119)

Ein Problem für die Einschätzung ist, dass es im Deutschen nicht möglich ist, eine Infinitivkonstruktion wie „*es gesehen zu haben*“ wie im Englischen (*They were the only people in the village to have seen it*) als Relativsatz zu verwenden, wie in (69) zu erkennen ist:

- (69)O: *dass sie die einzigen Leute im Dorf es gesehen zu haben sind*  
 $Z_{min}^{95}$ : *[dass]\_{LSK} [sie die einzigen Leute im Dorf]\_{MF} [sind]\_{RSK}, [es gesehen zu haben]\_{NF}*.

Diese Abweichung ist aber kein Verbstellungsfehler und wahrscheinlich nicht als Linearitätsverstoß zu werten. Da „*dass sie die einzigen Leute im Dorf sind*“ und „*es gesehen zu haben*“ beide wohlgeformte „Sätze“ sind, liegt die Ungrammatikalität in ihrer Beziehung zueinander. Das Feldermodell macht keine Aussagen zu Abhängigkeiten. Deshalb sollten Äußerungen wie (69)  $Z_{min}$  als eine Formabweichung (und somit kanonisch) gesehen werden.

Haben wir es trotzdem mit einem Verbstellungsfehler zu tun? Das Kopulaverb „*sind*“ wurde in der Zielhypothese in (68) umgestellt. Das Prädikat ist „*die einzigen Leute im Dorf*“ und deshalb kann „*sind*“ mit „*zu sehen haben*“ keinen Verbkomplex bilden. Aber kann „*es gesehen zu haben*“ als eingebetteter Teilsatz verstanden werden?

- (70)O: *dass sie die einzigen Leute im Dorf, es gesehen zu haben sind*  
 $Z_{min}$ : *dass sie die einzigen Leute im Dorf, die es gesehen haben, sind*

$Z_{min}$  ist eher ein Indiz, das dafür spricht. Aber es könnte sich auch um einen Verbstellungsfehler handeln. Da es diese Konstruktion nicht gibt, ist eine Entscheidung unmöglich. Deshalb können nur Festlegungen getroffen werden: den obigen

---

<sup>95</sup> Bei der Betrachtung der Lerneräußerungen hat es sich als nützlich erwiesen, eine „Minimalzielhypothese“ ( $Z_{min}$ ) zu bilden, in der versucht wird, einzelne Teilbereiche der Zielhypothese getrennt darzustellen, wie hier z.B. die Verbumstellung.

Argumentationen folgend wäre dies als kanonisch zu werten auch wenn es ungrammatisch ist<sup>96</sup>.

#### 2.4.2.2 Verbstellungsfehler

Bei Verbstellungsphänomenen haben sich folgende Abweichtungstypen als problematisch erwiesen:

##### Verbzweitfehler

Die Entscheidungsschwierigkeiten, ob es sich um einen Verbstellungsfehler (Mehrfachbesetzung des Vorfelds) handelt oder nicht, hängen oft mit der unglücklichen Verwendung von Konnektoren zusammen. Folgendes Beispiel zeigt das häufig in den GU-Texten auftretende „*Jedoch*“, als Übersetzung für die gern benutzte englische Wendung „*however*“, in der exponierten Nullstellung:

(71)O: *Jedoch die Geschichte Europas ist nicht friedlich, weil sie von Krieg und Vorurteilen gezeichnet ist.*

Z: *Die Geschichte Europas ist jedoch nicht friedlich, [...]*

(GU4\_0122)

Aber auch wenn die Stellung von „*jedoch*“ in der Zielhypothese zu Recht korrigiert wird, kann es in der Nullstelle vorkommen, wie folgendes Zitat aus dem Onlinedienst der IDS-online (grammis)<sup>97</sup> zeigt:

(72)[...] *Jedoch: Unter der dünnen Eisdecke ließ sich die Strömung nicht stoppen.*

Deshalb ist dieser Fehler nicht als Verbstellungsfehler zu werten. Anders verhält es sich bei folgendem Beispiel, weil „*So*“ nach IDS-online (grammis)<sup>98</sup> nicht in der Nullstellung vorkommt (und deshalb als nichtkanonisch zu sehen ist):

(73)O: *So, nach Admoni wird jedem Kasus eine „allgemeine Bedeutung“ zugeschrieben.*  
Z: *Nach Admoni wird so jedem Kasus eine „allgemeine Bedeutung“ zugeschrieben.*

(FU\_061)

Dieses Beispiel zeigt, dass eine Entscheidung getroffen werden muss, wer oder was als Autorität gelten sollte. Für solche Beispiele sind korpusbasierte Aussagen wie von IDS nützlich, wobei es wichtig wäre, Informationen über die Frequenz des Auftretens zu haben.

#### Obligatorische Extraposition am Beispiel des Phasenverbs „anfangen“:

---

<sup>96</sup> Alternativ zu dieser weiten Auslegung wäre eine enger Auslegung möglich, in der alle syntaktischen Verstöße als nichtkanonisch gelten. Dies hätte zur Folge, dass auch Äußerungen mit syntaktischen Fehlern innerhalb des Mittelfelds als unkanonisch einzustufen wären. Es bliebe weiterhin das Problem, dass geklärt werden müsste, was als syntaktische Abweichung zu sehen ist (in Abgrenzung z.B. zu stilistischen Abweichungen).

<sup>97</sup> [http://hypermedia.ids-mannheim.de/pls/public/gramwb.ansicht?v\\_app=g&v\\_kat=gramm&v\\_id=2143&v\\_wort=jedoch](http://hypermedia.ids-mannheim.de/pls/public/gramwb.ansicht?v_app=g&v_kat=gramm&v_id=2143&v_wort=jedoch)

<sup>98</sup> [http://hypermedia.ids-mannheim.de/pls/public/gramwb.ansicht?v\\_app=g&v\\_kat=gramm&v\\_id=2196&v\\_wort=so](http://hypermedia.ids-mannheim.de/pls/public/gramwb.ansicht?v_app=g&v_kat=gramm&v_id=2196&v_wort=so)

„Anfangen“ wird zu den fakultativ inkohärenten Verben gezählt, in folgendem Beispiel ist aber nur eine inkohärente Realisierung in der Extraposition (NF) möglich:

(74) O: *\*Her Sommer fing in dem See zu hineinwandern an*  
Z: *Herr Sommer fing an, in den See hineinzuwandern.*

(GU2\_3030)

Die Schwierigkeit, den Stellungsfehler einzuschätzen, hängt u.a. von der Flexibilität oder Variabilität der syntaktischen Realisierung ab. Die Bildung von Konstruktionen ohne Extraposition wird durch verschiedene Faktoren bedingt, wie die Beschaffenheit des einzubettenden Infinitivs:

(75) *Er fing zu trinken an.*

oder auch die Satzkonstruktion:

(76) *Er blieb solange, bis Herr Sommer in den See hineinzuwandern anfang.*

In diesem Fall (74) ist eine Form ohne Extraposition nicht erlaubt, „an“ soll nicht als RSK getaggt werden. Solche Phänomene werden als nichtkanonisch gewertet.

### 2.4.2.3 Verbauslassungsfehler oder Ellipse

Folgendes Beispiel könnte sowohl als Ellipse als auch als ein Verbauslassungsfehler empfunden werden. Oftmals hilft nur der Kontext<sup>99</sup>, dies zu entscheiden. Meines Erachtens gibt es keine stringente Regelung. Deshalb muss eine Festlegung getroffen werden: wenn möglich, sollte dies nicht als <f\_MS> bzw. <f\_KS> getaggt werden. Hier sind große Abweichungen in der Einschätzung zu erwarten.

(77) O: *Das alles wegen Frank Schirmacher, der behauptete, dass wir die deutsche Literatur den Großstadtroman wieder entdecken sollten.*  
Z1: *Das alles wegen Frank Schirmacher, der forderte, dass wir in der deutschen Literatur den Großstadtroman wiederentdecken sollten.*  
Z2: *Das alles geschieht wegen Frank Schirmacher, der forderte, dass wir in der deutschen Literatur den Großstadtroman wiederentdecken sollten.*

Wie schon erwähnt bieten die Festlegungen in Abschnitt 2.1.2.1.3, S. 28 in der allgemeinen Einführung in das Felderannotationsschema eine gute Grundlage und sie können als ungefähre Richtlinien dienen. Es ist aber deutlich geworden, dass je nach Vorgehen die intuitive Einschätzung zu „Grammatikalität“ nicht immer eine eins-zu-eins-Entsprechung mit der Kanonizität bedeuten muss. Es gibt zwar einen engen Zusammenhang zwischen einer Fehleranalyse und der Entscheidung, ob eine Äußerung in topologische Felder eingeteilt werden kann, aber sie sind nicht identisch. Dafür sind genaue Festlegungen nötig. Die Qualität dieser Festlegungen lässt sich nur

---

<sup>99</sup> Das große mediale, politische, kulturelle Erwacht an Berlin erscheint neue deutsche Metropolieliteratur Das alles wegen Frank Schirmacher, der behauptete, dass wir die deutsche Literatur den Großstadtroman wieder entdecken sollten

in der Praxis beurteilen. Sind sie für andere Annotatoren einleuchtend, ermöglichen sie eine konsistente Annotation oder sind sie doch zu vage, und gibt es Sprachbeispiele, die sie nicht antizipiert haben?

## **2.5 Evaluation**

Bei der Evaluation einer Annotation sollten verschiedene Aspekte untersucht werden. Dabei ist eine wichtige Fragestellung, wem sie überhaupt nutzt und was mit dieser Annotation untersucht werden kann und was nicht. Hierzu soll die Anwendung der Annotation bei Lernerdaten (auch bei frühem Erwerb) besprochen werden. In diesem Zusammenhang stellt sich auch die Frage nach den Grenzen der Felderannotation bei Lernern verschiedener Sprachstufen. Dazu wird ein quantitatives Verfahren mit den longitudinalen Daten von Georgetown GU2, GU3 und GU4 als Basis verwendet, um für die drei Sprachstufen eine Statistik über die Anzahl erfassbarer (dem Feldermodell konformer) gegenüber den nicht erfassbaren Strukturen aufzustellen.

Als weiterer Evaluationspunkt wurde ein Inter-Rater-Vergleich durchgeführt anhand dessen Genauigkeit und Konsistenz (accuracy und consistency) quantitativ überprüft werden soll.

### **2.5.1 Anwendung der Annotation bei Lernerdaten**

Auch wenn die Daten, für die diese Annotation gedacht ist, vorwiegend von „fortgeschrittenen Lernern“ stammen, und man annehmen kann, dass sie alle in Abschnitt 1.3.2 aufgeführten Erwerbsstufen schon durchlaufen haben<sup>100</sup>, stellt sich die Frage nach ihrer Anwendbarkeit für die Wortstellung bei frühem Erwerb und der folgenden Erwerbssequenz.

#### **2.5.1.1 Anwendung der Annotation im Bereich der Erwerbssequenzen**

Eine entscheidende Frage, die sich bei der Überlegung stellt, ob man die Daten für einen Korpus elektronisch aufbereiten will, ist die nach dem Aufwand, dessen Bedeutung nicht zu unterschätzen ist: gesprochene Daten müssen transkribiert und handschriftliche Daten (siehe Lüdeling 2008, S. 122) müssen digitalisiert werden.

Eine anschließende automatische Wortartannotation der gesamten Originaläußerung bringt bei Daten aus dem frühen Erwerb bei den vielen abgebrochenen, gestotterten

---

<sup>100</sup> Alle Lerner in Jansens (2008) Studie hatten schon im zweiten Sprachlernjahr (nach einem Intensivkurs im ersten Jahr mit 5 Stunden pro Woche, 13 Wochen pro Semester) Verbend-Strukturen in Nebensätzen erworben. Wobei Erwerb in Jansens Daten sich auf „emergence“, das erste Auftreten bezieht.

und muttersprachlichen Zwischenäußerungen wenig Nutzen. Bei einer Mehrebenenarchitektur wäre es denkbar, hier eine Korrektorebene für die Untersuchung der Wortstellung einzuführen, in der Rechtschreibung, Zeichensetzung und Verbformen korrigiert werden (da diese Kriterien für die Wortstellung keine Rolle spielen). Diese Ebene könnte dann als Basis für den automatischen Wortart-Tagger dienen.

Im folgenden Beispiel wird ersichtlich, dass aufgrund der Rechtschreibung auch in schriftlichen Daten frühen Erwerbs eine automatische Annotation in der Ebene der Originaläußerung nicht sinnvoll ist.<sup>101</sup>

(78) *Ich bin Christine. Ich haben noine iare halt. Meine Mutter haisst [...]*  
(Christine M4/5, I)

(zit. n. Diehl 2000, S.75)

Als nächstes soll besprochen werden, ob die Felderannotation die Erwerbssequenzen der Wortstellung des Deutschen erfassen kann. Ein grundsätzliches Problem ist, dass es in den ersten beiden Stufen der Erwerbssequenzen keine finiten Verben gibt, sondern Verben nur lexikalisch realisiert werden. Deshalb wäre es zwar pragmatisch denkbar, SVO-Strukturen im Feldermodell einzuordnen mit V als  $LSK_{V1/V2}$ <sup>102</sup>. Formal ist das aber nicht möglich, weil dies im Widerspruch zum Feldermodell steht, in dem nur  $V_{fin}$  die  $LSK_{V1/V2}$  besetzen kann. Im Grunde genommen, kann das Feldermodell erst beim Erwerb der Finitheit ab Stufe III angewandt werden<sup>103</sup>.

Um überhaupt Erwerbsdaten annotieren und auswerten zu können, muss sowohl die zu untersuchende Struktur identifiziert werden, als auch jeder Kontext, in dem sie vorkommen kann, das heißt, es müssen z.B. für die Bestimmung, ob Verbtrennung in der Erwerbssequenzstufe III (SEP) erworben wurde, sowohl die Äußerungen gefunden werden, bei denen die finiten Verbelemente satzfinal sind als auch die, in denen keine Verbtrennung stattgefunden hat.

Grundsätzlich ist es möglich, mit dem Falko–Annotationsschema alle möglichen Kontexte und Vorkommnisse zu erfassen, weil alle Lerneräußerungen annotiert werden.

---

<sup>101</sup> Grundsätzlich sind schriftliche Texte weniger problematisch zu annotieren als gesprochene, weil die Schriftsprache insgesamt ein unspontaneres und reflektierteres Kommunikationsmedium darstellt.

<sup>102</sup> LSK bei V1- bzw. V2-Sätzen

<sup>103</sup> Finitheit tritt erst in der Post-Basisvarietät auf, bei Minimal Trees erst beim Aufbau von "Agreement"-Phrasen (AgrP).

Um andererseits zu bestimmen, ob die Subjekt-Verb-Inversion der Erwerbssequenzstufe IV erworben wurde, müssen Äußerungen gefunden werden, bei denen ein anderes Element als das Subjekt vor V<sub>fin</sub> steht.

Es ist offensichtlich, dass irgendeine Art von Konstituentenanalyse benötigt wird. Die alleinige Annotation der Felder wird auch nicht mit dem Hinzuziehen einer Wortartebene ausreichen. Wie folgendes Beispiel deutlich macht, können Strukturen der INV–Stufe nicht von denen der SVO–Stufe unterschieden werden:

[Der Hund]<sub>VF</sub> [beißen]<sub>LSK?</sub> [den Mann]<sub>MF</sub>. (ART NN V<sub>FIN</sub>? ART NN)<sup>104</sup>  
→(SVO)

[Den Mann]<sub>VF</sub> [beißt]<sub>LSK</sub> [der Hund]<sub>MF</sub>. (ART NN V<sub>FIN</sub> ART NN)  
→(OVS)

Es stellt sich die Frage, ob eine reine Konstituentenannotation ausreichen würde, Phänomene in der Erforschung des L2-Erwerbs zu beschreiben und ob das hier vorgestellte Felderannotationsschema etwas dazu beitragen kann.

Für die Erwerbssequenzanalyse müssen folgende Fälle unterschieden werden: Verbend-Strukturen in Nebensätzen mit C (Complimentizer), und nicht realisierte Verbend-Strukturen. Die Konstituentenabfolgen, die hierfür gesucht werden<sup>105</sup>, sind für Non-VE (Nicht-Verbend) [C + (X) + Sub + V<sub>fin</sub> + V<sub>≠fin</sub> / X]<sup>106</sup>

und für VE (Verbend) [C + (X) + Sub + (V<sub>≠fin</sub>) + V<sub>fin</sub>].

Diese Fälle sind in Sätzen mit Nachfeldern, bei denen die RSK nur mit dem finiten Verb besetzt ist, nicht unterscheidbar, wie folgendes Beispiel verdeutlicht:

(79) *dass er mir sagt, was ich machen soll.*

Diese Abfolge [C + Sub + V<sub>fin</sub> + Obj] entspricht Non-VE, aber es handelt sich eindeutig um eine Verbendstruktur. Wenn es aber sowohl eine Felder- als auch eine Konstituentenannotation gibt, könnten mit Hilfe einer kombinierten Suche diese beiden Strukturen eindeutig unterschieden werden, da eine Non-VE-Struktur mit <f\_KS> annotiert wird und so von der Struktur in (79) abweicht.

Das vorangestellte Beispiel macht deutlich, dass eine Annotation von kanonischen und nichtkanonischen Strukturen bei der Suche und Unterscheidung von Lernerstrukturen

---

<sup>104</sup> Wortarttags, dem STTS entsprechend.

<sup>105</sup> siehe hierzu Jansen (2008, S. 200)

<sup>106</sup> Konstituententags: Sub = Subjekt; V<sub>fin</sub> = finites Verb; V<sub>≠fin</sub> = nicht finites Verbelement;; Obj = Objekt; X = beliebiges Element; () = fakultativ, / = und/oder in beliebiger Reihenfolge.

hilfreich ist. Ob die Kanonizität notwendigerweise mit dem Feldermodell beschrieben werden muss, ist eine andere Frage.. Da es als allgemeiner Beschreibungsansatz in der germanistischen Linguistik dient, wäre diese Nomenklatur bei der Suche etwas eingängiger als andere Beschreibungsansätze<sup>107</sup>.

### 2.5.1.2 GU Vergleich

Folgendes Diagramm verdeutlicht das Verhältnis von erfassbaren, dem Feldermodell konformen und nicht erfassbaren Strukturen der drei Sprachstufen (Level 2, 3 und 4) in den longitudinalen Georgetown-PPT-Daten.

Die ermittelten Werte zeigen deutlich, dass der Anteil von nicht annotierbaren Äußerungen im Durchschnitt nicht mehr als 10% übersteigt (im „schlechtesten“ Text in GU\_2 waren es 21%).

	kanonische Sätze			nichtkanonische Sätze		
	$\bar{\chi}^{108}$	mA <sup>109</sup>	%	$\bar{\chi}$	mA	%
GU_2	134,11	± 11,08	94	9,16	±5,57	6
GU_3	113,62	±6,24	95	6,17	±2,86	5
GU_4	87,34	±10,97	98	4,25	±1,90	2

Tabelle 5 Durchschnittswerte aus dem GU-Vergleich

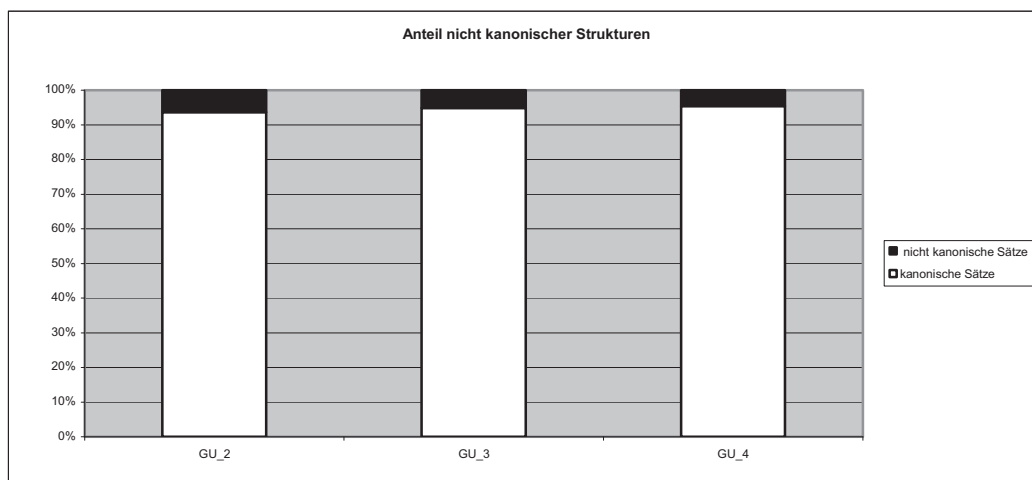


Abbildung 13 GU-Vergleich in Prozent

Bei den Daten, die in dieser quantitativen Untersuchung benutzt wurden, handelt es sich nicht um spontane Lerneräußerungen. Sie wurden als Hausaufgabe erstellt und

<sup>107</sup> z.B. die Suchanfrage für INV-Strukturen könnte lauten: Suche Vorfelder ohne Subjekt!

<sup>108</sup>  $\bar{\chi}$  = Durchschnitt der Anzahl von kanonischen bzw. nichtkanonischen Sätzen (Matrix- und Konstituentensätze) der Lernertexte, die auf 1000 Token normalisiert wurde. Bei den nichtkanonischen Sätzen wurden elliptische Strukturen (auch nichtkanonische) nicht dazugerechnet, weil hier der Schwerpunkt auf lernerspezifischen Aspekten liegt.

<sup>109</sup> mittlere Abweichung=  $1/n \sum |\chi - \bar{\chi}|$

unterliegen allen möglichen Formen der Korrektur. Sie zeigen lediglich, dass sich die Felderannotation bei dieser Textart auch für „intermediate“ Lerner (Level 2) eignet. Eine Tendenz ist erkennbar, dass von Level 2 bis Level 4 stetig immer mehr kanonische Sätze gebildet werden und dem zu Folge die Verbstellungsfehler und Verbauslassungsfehler abnehmen, wenn auch die Unterschiede zwischen den verschiedenen Levels nicht groß sind.

An den Daten selbst erkennt man, dass die Sätze grundsätzlich länger werden (weniger Sätze pro 1000 Token). Wie man der Tabelle 6 entnehmen kann, liegt dies nicht daran, dass mehr Nebensätze gebildet werden (das Verhältnis zwischen Matrixsatz zu Konstituentensatz nimmt in GU4 sogar ab), sondern vielmehr daran, dass es mehr und/oder längere Konstituenten geben muss.

auf 1000 normalisiert	kanonische Sätze $\bar{\chi}$	MS $\bar{\chi}$	KS $\bar{\chi}$	KS-Anteil
GU_2	134,11	86,07	48,04	36%
GU_3	113,62	67,37	46,25	41%
GU_4	87,34	59,97	27,37	31%

Tabelle 6 GU Vergleich: Anteil der Konstituentensätze

Eine interessante Beobachtung bei erster grober Durchsicht der Wortstellungsfehler ist, dass auch diese Daten, die wirklich nicht als Grundlage für Erwerbssequenzanalysen geeignet sind, Tendenzen zeigen, die darauf hinweisen, dass Verbletzstrukturen verhältnismäßig schneller gelernt bzw. schneller beherrscht werden als Verbzweit-Strukturen.

Fehler	Mittelfeld <sup>110</sup>	Verbend (VE)	Verbzweit (V2)	VE/V2	Tokenzahl
GU_2	25	53	69	0.78	22032
GU_3	36	39	49	0.80	23191
GU_4	12	16	36	0.44	22097

Tabelle 7 GU-Vergleich: Verhältnis VE- zu V2-Fehler in absoluten Zahlen

Um diese Annahme zu bestätigen, wäre es ratsam, sowohl die Lernkurve der einzelnen Lerner als auch die einzelnen Fehler genauer anzuschauen und statistisch auszuwerten.

<sup>110</sup> Für die Untersuchung von Wortstellungsfehlern im MF wurden die Fehlertags <MF\_MF>, <MF\_MFe> gezählt; für VE-Fehler <(Nebensatzzeileiter)V\*FIN;>; für Verbzweitfehler <MF\_LSK> und <VF\_LSK>.

Diese quantitative Analyse liefert keine stichhaltigen Daten, die eine genaue Bestimmung der Anwendbarkeitsgrenzen der Felderannotation bei früheren Sprachstufen zulassen. Sie macht eher die Lernerstrategie deutlich, dass nicht so fortgeschrittene Lerner sich stärker auf bekannte Wortfolgemuster (learning patterns) beschränken. Ein Indiz hierfür ist, dass ungefähr die Hälfte (7 von 16) der Level2-Lerner keine Verbzweitfehler machen, aber in Level 4 alle sieben dieser Lerner Verbzweitfehler produzieren.

### **2.5.1.3 Anwendung der Annotation bei fortgeschrittenen Lernern**

Es gibt meines Erachtens zwei wesentliche Aspekte, die für eine Felderannotation bei fortgeschrittenen Lernern sprechen. Bei fortgeschrittenen Lernern sind satzübergreifende, informationsstrukturelle und kommunikative Aspekte von Bedeutung, auch als Forschungsbereiche (siehe hierzu Walter und Grommes 2008). Wie realisieren Lerner Topikmarkierungen (Satzgegenstand), die zum Beispiel im Deutschen durch Topikkonstruktionen in Vorvorfeld, Linksversetzung (Frey 2005) und in der Topikdomäne an der Spitze des Mittelfelds (Frey 2004) realisiert werden? Oder wie gestalten sich Lernertexte in Bezug auf die Konstituentenabfolge im Mittelfeld und die „normale Betonung“ (Höhle 1982)? Eine Annotation der Felder erleichtert die Suche und ist ein übliches Beschreibungsmodell in diesen Bereichen.

Einen weiteren Vorteil sehe ich darin, die Annotation für didaktische Zwecke einzusetzen. Dies kann für Lerner nützlich sein, da die Einteilung in Felder es einfacher macht, Satzkonstruktionen in Teilen zu erfassen und zu verarbeiten.

### **2.5.2 Consistency and Accuracy**

Dieser Abschnitt befasst sich mit den Fragen der Einheitlichkeit (consistency) und Genauigkeit (accuracy) der Annotation (Leech und Eyes 1997). Eine Annotation ist „genau“ in dem Maß, wie sie die Regeln der Annotation befolgt, *consistency* bezieht sich stärker darauf, ob die Annotatoren eine gleiche Annotation (untereinander gleich und innerhalb der jeweiligen Annotation durchgängig gleich) durchführen. Consistency hängt unter anderem davon ab, wie viel Interpretationsspielraum die Anleitung für die Annotation zulässt. Eine Grauzone ist dabei, dass es bei der Interpretation von Lerneräußerungen generell großen Spielraum gibt, auftretende Sprachphänomene einzuschätzen. Andererseits ist es nicht einfach, die beiden Begriffe in voneinander

getrennten konkreten Zahlen zu erfassen, da sie sich nur schwer voneinander abgrenzen lassen.

Als Basis für die quantitative Bestimmung dient die Auswertung eines „Inter-Rater-Agreement“. Für den Inter-Rater-Vergleich wurden zwei Texte aus dem FU-Zusammenfassungskorpus annotiert. Der kurze und einfachere Text FU\_105 diente als Einstieg. Er wurde von vier, der zweite Text, FU\_102, von drei Annotatoren bearbeitet<sup>111</sup>.

Text	Tokenzahl	MS-Anzahl	KS_1-Anzahl	KS_2-Anzahl	Inter-Rater-Anzahl
FU_105	217	9	2		4112
FU_102	530	34	17	2	3

Tabelle 8 Texte für Inter-Rater-Vergleich

Die Annotatoren wurden nur kurz allgemein eingewiesen (auch in das Feldermodell). Ihnen stand online eine schriftliche Kurzanleitung zur Verfügung<sup>113</sup>, Sonderregeln wurden nicht besprochen und das Lesen der gesamten Dokumentation hätte den Rahmen gesprengt. Alle haben etwas linguistisches Hintergrundwissen (mindestens Grundkurse), aber nur eine Person studiert aktuell Linguistik.

Indem ein Goldstandard angenommen wurde, wird eine gewisse Gewichtung zu Gunsten von accuracy vorgegeben. In diesem Fall können damit bessere Aussagen getroffen werden, da wir es nicht mit erfahrenen Annotatoren zu tun haben. Consistency setzt mehr Erfahrung bei den Annotatoren voraus.

### **Auswertung:**

Zuerst wurden Inter-Rater-Abweichungen auf der Satzebene und der Felderebene erfasst, die auf Entscheidungen<sup>114</sup> der Annotatoren zurückzuführen sind. Diese können unterschiedlichen Ursprung haben, entweder dass die Annotationsregeln nicht verstanden bzw. noch nicht beherrscht werden oder dass es unterschiedliche Einschätzungen zu Äußerungen gibt, z.B. kann eine Äußerung ohne Verb als Ellipse oder als fehlerhaft eingestuft werden.

<sup>111</sup> Die beiden Texte sind (mit Zielhypothese) im Appendix B abgebildet.

<sup>112</sup> Das bedeutet drei Annotatoren und eine Annotation, die als Goldstandard diente; in Text 102 entsprechend zwei und eine.

<sup>113</sup> <https://smartdrive.web.de/guest?path=Falko%20von%20seaka&token=580613B30138D463>

<sup>114</sup> Hierzu wurden Flüchtigkeitsfehler nicht mitgerechnet. Sie wurden gesondert gezählt.

Gesucht werden die Übereinstimmungen bei der Segmentierung und den Satz- bzw. Felder-Tags. Bei der Übereinstimmung der Felder bedeutet dies, dass **alle** Segmentierungen und Felder-Tags eines Satzes übereinstimmen müssen.

Für eine statistisch relevante und vergleichbare Aussage zum Inter-Rater-Agreement wurde der Kappa-Wert<sup>115</sup> ausgerechnet. Dieser Wert ist abhängig von der Anzahl der Rater, der Kategorien und der einzuschätzenden Fälle. Zweck der Berechnung des ist die Ermittlung einer Prozentzahl an Übereinstimmungen, die zufällig auftreten können, und diese mit einzubeziehen in der Auswertung des Inter-Rater-Agreement, um damit eine bessere Vergleichbarkeit verschiedener Daten zu ermöglichen (Carletta 1996).

Für die Auswertung wurden vier Kategorien festgelegt, die in Übereinstimmung mit dem Goldstandard<sup>116</sup> gezählt wurden:

+T +S: Tag(s) und Segmentierung (-en) stimmen überein

-T +S: Segmentierungen stimmen überein, Tags aber nicht

+T -S: Tags stimmen überein, Segmentierungen aber nicht

-S -T: Tags und Segmentierungen stimmen nicht überein

In Tabelle 9 werden die Übereinstimmung  $P^0$  und Kappa für die beiden Texte aufgelistet. Für den Text 102 wird auch zwischen Matrixsatz und Konstituentensatz unterschieden. Es gibt verschiedene Meinungen, ab welchem Kappa-Wert eine annehmbare Übereinstimmung zu erkennen ist. Für den „free marginal Kappa“, der hier genutzt wird, wird 0,7 angesetzt<sup>117</sup>.

Text	Satz_ms		Felder_ms		Satz_ks		Felder_ks		Satz_gesamt		Felder_gesamt	
	$P^0$	K	$P^0$	K	$P^0$	K	$P^0$	K	$P^0$	K	$P^0$	K
102	0,88	0,85	0,82	0,76	0,86	0,84	0,78	0,71	0,87	0,83	0,81	0,75
105									0,80	0,74	0,67	0,57

Tabelle 9 Kappa-Auswertung – Agreement-Rate

Bis auf die Werte für „Felder\_gesamt“ in Text 105 kann die Agreement-Rate demnach als zufrieden stellend betrachtet werden. Um die Auswirkung der Festlegung, dass alle Felder übereinstimmen müssen, einzuschätzen, habe ich die Anzahl der Felder

<sup>115</sup> Es gibt verschiedene Möglichkeiten, den Kappa-Wert für mehr als eine Vergleichsperson zu berechnen. Hier wurde der free marginal kappa angewendet (<http://justus.randolph.name/kappa>).

<sup>116</sup> Es wäre auch eine Möglichkeit, unabhängig von einem Standard nur die Übereinstimmungen zu zählen. Bei der Einschätzung der Felder gibt es aber so viele Varianten, dass dann eine Erfassung der jeweiligen Übereinstimmung der einzelnen Felder nötig wäre. Dies war im Rahmen dieser Arbeit nicht zu bewältigen.

<sup>117</sup> siehe <http://justus.randolph.name/kappa>

gezählt, die übereinstimmen(+T+S) und die nicht übereinstimmen. Bei Text 105 gibt es nach Auszählung der Felder eine Übereinstimmung von 89,88 %.

Zwei weitere konkrete Auswertungen haben mich interessiert: wie viel unkorrekte Annotationen der Felder gab es insgesamt und bei wie vielen Fehlern handelt es sich um Flüchtigkeitsfehler, bei denen anzunehmen ist, dass die Annotationsregeln verstanden worden sind, aber die Annotatoren sich z.B. vertippt haben, wie <MF\_MF> statt <MF\_MS><sup>118</sup>. Der Goldstandard wurde in den Berechnungen jeweils nicht miteinbezogen.

	FU_102			FU_105
	MS	KS	Gesamt	
inkorrekte Annotation	6,72%	27,05% <sup>119</sup>	13,33%	10,12%
Flüchtigkeitsfehler	2,77%	4,92%	3,5%	1,79%

Dass Flüchtigkeitsfehler bis zu 5% ausmachen können<sup>120</sup>, lässt die Notwendigkeit erkennen, dieser Fehlerquelle konstruktiv entgegenzuwirken: je nach Kapazität sollten die Daten von mindestens zwei Annotatoren bearbeitet und „Tag Insert Tools“ benutzt werden, um Tipp-Fehler zu vermeiden. Semiautomatische Prozesse (Marcus et al. 1993; Bateman et al. 1997), die automatisches Taggen mit manueller Korrekturmöglichkeit bieten, können die Genauigkeit der Annotation verbessern.

Bei den Abweichungen von den Annotationsregeln gibt es verschiedene Aspekte zu nennen: erwartungsgemäß sind einige Fehler formaler Art, z.B. wie Interpunktion annotiert werden soll. Andere Fehler, wie falsche Zuordnungen, gehen auch darauf zurück, dass auch Nichtlinguisten annotiert haben, z.B. wurde „so viel“ einmal als Nebensatzeinleiter getaggt. Dass die Konstituentensätze schlechter annotiert wurden, ist nicht wirklich verwunderlich, da sie in einer tieferen Annotationsebene liegen und manchmal einfach vergessen wurden.

Einige Abweichungen zum Goldstandard zeigen, dass weitere Festlegungen mehr Übereinstimmung erzielt hätten (siehe Beschreibung in Appendix B, S.80). Zum Beispiel gab es keine Übereinstimmung der am Versuch beteiligten anderen Annotatoren untereinander in Zeile 12 (S8) von Text 105. Hier wurde „*entwurzelt und*

<sup>118</sup> Diese Unterscheidung ist nicht immer sehr eindeutig zu treffen, Flüchtigkeitsfehler sind bei erfahrenen Annotatoren vermutlich leichter auszumachen als bei Annotations-„Neulingen“.

<sup>119</sup> Diese hohe Zahl kommt dadurch zustande, dass eine unterschiedlich verstandene Struktur mehrere inkorrekte Feldereinteilungen verursacht.

<sup>120</sup> der menschliche Faktor

*verloren*“ zum Einen von zwei Personen als Verb (RSK) und zum Anderen von einem Annotator als Mittelfeld (MF) getaggt. Durch die kurz gefasste Einweisung war die Regel („Partizip mit *„sein“* soll als ADJD getaggt werden“) nicht bekannt. Die Anwendung dieser Festlegung hätte eine bessere Übereinstimmung erzielt. Die Festlegung, Zustandspassiv als Kopula zu annotieren, wird nicht wirklich intuitiv nachempfunden. Daraus ergibt sich, dass bei einer Einweisung in die Annotation zumindest ausführlich darauf hingewiesen werden müsste.

Ein weiteres sehr interessantes Beispiel ist die Beurteilung von der Äußerung in Zeile 14-15: *„widerspiegeln“* ist ein Beispiel für ein nicht abgetrenntes Verbpartikel. Bei keiner Annotation wurde dies als falsch empfunden, also ist *„widerspiegelt“* in der LSK. In der Zielhypothese war dies aber als Fehler markiert. Es stellt sich hier die Frage, ob bei einer so großen Akzeptanz<sup>121</sup> dieses Verb überhaupt noch als trennbar verstanden werden muss.

Letztendlich sind relative gute Inter-Rater-Werte erreicht worden, obwohl keine "Experten"<sup>122</sup> mitannotiert haben. Das zeigt, dass die Falko-Annotation zunächst schnell erlernbar ist. Andererseits sind nicht alle Regeln intuitiv zu erschließen und eine gute Dokumentation ist in jedem Fall eine Grundvoraussetzung.

### 3 Schlussbemerkung

Diese Arbeit zeigt eine Möglichkeit auf, wie in einem Korpus kanonische und nichtkanonische Äußerungen gemeinsam annotiert werden können, ausgehend von dem Feldermodell.

Sie zeigt auch, wie herausfordernd die Analyse und Annotation sein kann von Sprachdaten, insbesondere Lernerdaten mit großer Variabilität (auch in deren Analyse), die sowohl in ihrer eigenen Systematik erfasst als auch im Vergleich zu einer Zielhypothese beschrieben werden sollen.

Wie wichtig klare Kriterien sind für eine einheitliche Annotation wurde herausgearbeitet: bei den Festlegungen sollten Faktoren wie Konsistenz, Konformität mit anerkannten Beschreibungsansätzen, Vergleichbarkeit mit anderen Korpora,

---

<sup>121</sup> Die Suche nach *„widerspiegelt“* bei Google ergab über 900.000 Treffer, *„spiegelt \*wider“* ergab mehr 1600.000 Treffer.

<sup>122</sup> Es wäre auch interessant, zu untersuchen, welche Festlegungen bei erfahrenen Annotatoren die meisten Schwierigkeiten bereiten. Möglicherweise sind die Kriterien doch zu ungenau oder sie sind nicht wirklich einleuchtend. Leider kann diese klein angelegte Studie keine Aussagen darüber machen.

adäquate Beschreibung und Suchbarkeit der Sprachphänomene berücksichtigt werden. Dennoch können genau diese sich an Punkten gegenüberstehen: aufgrund der Entscheidung für eine konsistente Annotation wird eventuell die Richtlinie, allgemein anerkannte linguistische Kategorien anzunehmen, nicht befolgt, wie bei der Entscheidung, „sein“ und Partizip grundsätzlich als Kopulakonstruktion zu annotieren.

Es gibt Vor- und Nachteile, ein Annotationsschema auf ein so deskriptives Beschreibungsmodell wie das Feldermodell zu gründen. Zwar können Strukturen mit einbezogen werden, die bei viel stringenteren, regelbasierten Modellen nicht ohne weiteres erfasst werden, andererseits gibt es, gerade weil es deskriptiv ist, nur vage Kriterien.

Ein großer Teil der Arbeit hat sich damit beschäftigt klare Kriterien aufzustellen für die Entscheidung über die (modellbezogene) Kanonizität und dabei hat sich gezeigt, dass kanonische Strukturen und „Wohlgeformtheit“ nicht immer übereinstimmen. In diesem Zusammenhang wurde auch besprochen, dass mehrere Zielhypothesen wichtig sind, um die unterschiedlichen Aspekte zunächst getrennt voneinander analysieren, und damit ihr Zusammenwirken besser verstehen zu können.

Verbstellungsfehler und Auslassungsfehler der Satzklammern bilden die Grundsteine für die Entscheidung, Äußerungen als nichtkanonisch zu annotieren. Es ist aber deutlich geworden, dass diese Vorgabe nicht ohne weitere Festlegungen auskommt. In wieweit diese speziellere Regeln sich bewähren, muss sich in der Praxis zeigen.

Die quantitative Auswertung zu den Grenzen der Anwendbarkeit der Annotation bei nicht fortgeschrittenen (intermediate) Lernern hat keine stichhaltigen Ergebnisse geliefert, sie schließt sie jedenfalls nicht aus. Auf jeden Fall ist klar, dass das Feldermodell vor dem Erwerb der Finitheit, nicht eingesetzt werden kann und dass es nicht allein ausreicht, um Erwerbssequenzen zu erfassen, wobei eine kombinierte Annotation (z.B. mit Konstituenten) jedoch von Vorteil ist bei der Suche und Unterscheidung von Wort- und Konstituentenabfolgen in Äußerungen.

Die Inter-Rater-Studie mit „Nichtexperten“ zeigte, dass die Annotation leicht erlernbar und schnell zugänglich ist. Aber auch, dass bestimmte Annotationsrichtlinien nicht intuitiv sind.

Ein Annotationsschema stellt nie ein fertiges Produkt dar, da die Bearbeitung von neuen Daten zu neuen Erkenntnissen führt.

In dieser Arbeit sind viele Bereiche nur angerissen worden, die weiter zu verfolgen interessant wären, dazu gehört, sich im Bereich von Syntax tiefer gehend mit der Annotation und Analyse von Koordinationsellipsen zu befassen, sich in der L2-Erwerbsforschung lernerspezifischeren Fragestellungen zu stellen, wie der Entwicklung eines Verfahrens für eine Konstituentenannotation von Lernerdaten, Studien zu verschiedenen Typen von Zielhypothesen (zu Stil; Grammatikalität und auf bestimmte sprachliche Bereiche bezogen) vergleichend zu erforschen, technischen Verfahrensfragen nachzugehen wie der Anwendung von semiautomatischen Verfahren auch bei Lernerdaten und anhand der konkreten Annotation Analysen durchzuführen und die Redundanz bei der Wahl syntaktischer Strukturen von Lerner und Muttersprachlern zu untersuchen.

## 4 Literaturverzeichnis

Albert, S.; Anderssen, J.; Bader, R.; Becker, S.; Bracht, T.; Brants, S. et al. (July 2003): TIGER-Annotationsschema. Technical Report. University of Potsdam, Saarland University, University of Stuttgart. Online verfügbar unter [http://www.ifi.uzh.ch/CL/volk/treebank\\_course/tiger\\_annot.pdf](http://www.ifi.uzh.ch/CL/volk/treebank_course/tiger_annot.pdf), zuletzt geprüft am 02.10.2008.

Altmann, H.; Hahnemann, S. (2005): Syntax fürs Examen. Studien- und Arbeitsbuch / 2., überarb. und erw. Aufl. Wiesbaden: VS Verl. für Sozialwiss. (Lehrbuch, 1).

Atwell, E.; Howarth, P.; Souter, C. (2003): The ISLE Corpus: Italian and German Spoken Learners' English. In: International Computer Archive of Modern and Medieval English (ICAME), Jg. 27.

Bateman, J.; Forrest, J.; Willis, T. (1997): The use of syntactic annotation tools: partial and full parsing. In: Garside, R.; Leech, G.; McEnery, A. (Hg.): Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Longman, S. 166–178.

Bech, G. (1955): Studium über das deutsche Verbum infinitum. Tübingen: Max Niemeyer Verlag, 1983 (Linguistischen Arbeiten 139).

Belz, J. (2004): Learner Corpus Analysis and the Development of Foreign Language Proficiency. In: System: An International Journal of Educational Technology and Applied Linguistics, Jg. 32, H. 4, S. 577–591.

Biber, D.; Conrad, S.; Reppen, R. (1998): Corpus linguistics. Investigating language structure and use. Cambridge: Cambridge Univ. Press (Cambridge approaches to linguistics).

Bird, S.; Liberman, M. (1999): A formal framework for linguistic annotation. In: Speech Communication, Jg. 33, S. 23–60.

Brandt, M.; Reis, M.; Rosengren, I.; Zimmermann, I. (1992): Satztyp, Satzmodus und Illokution. In: Rosengren, Inger (Hg.): Satz und Illokution. Tübingen: Niemeyer (Linguistische Arbeiten, ...), S. 1–90.

Bresnan, J. (2001): Lexical-functional syntax. Oxford: Blackwell Publishers.

Carletta, J. (1996): Assessing agreement on classification tasks: the Kappa statistic. In: Computational Linguistics, Jg. 22, H. 2, S. 249–254.

Chomsky, N. (1959): Review of B.F. Skinner Verbal Behavior. In: Language, Jg. 35, S. 26–58.

Clahsen, H. (1984): The Acquisition of German Word Order. A Test case for Cognitive Approaches to L2 Development. In: Anderson, R. (Hg.): Second Languages. A Cross-Linguistic Perspective. Rowley, Mass.: Newbury House, S. 219–242.

Clahsen, H.; Meisel, J.; Pienemann, M. (1983): Deutsch als Zweitsprache. Der Spracherwerb ausländischer Arbeiter. Tübingen: Narr.

Cobb, T. (2003): Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. In: Canadian Modern Language Review, Jg. 59, H. 3, S. 393–423.

Corder, S. (1967): The significance of learner's errors. In: International Review of Applied Linguistics, H. 5, S. 161–169.

- Diehl, E.; Christen, H.; Leuenberger, S.; Pelvat, I.; Studer, T. (2000): Grammatikunterricht: Alles für der Katz? Untersuchungen zum Zweitsprachenerwerb Deutsch. Tübingen: Niemeyer (Reihe germanistische Linguistik, 220).
- Dipper, S.; Brants, T.; Lezius, W.; Plaehn, O.; Smith, G. (2001): The TIGER Treebank: Third Workshop on Linguistically Interpreted Corpora LINC-2001. Leuven, Belgium.
- Drach, E. (1937): Grundgedanken der Deutschen Satzlehre.
- Dulay, H.; Burt, M. (1974): You can't learn without goofing. In: Richards, J. (Hg.): Error Analysis. Perspectives on Second Language Acquisition. London: Longman, S. 95–123.
- Dürscheid, C. (1989): Zur Vorfeldbesetzung in deutschen Verbzweit-Strukturen. Trier: Wissenschaftlicher Verlag (FOKUS 1).
- Edmondson, W.; House, J. (2000): Einführung in die Sprachlehrforschung. 2., überarb. Tübingen: Francke (UTB für Wissenschaft Linguistik, 1697).
- Eisenberg, P. (1999): Der Satz. Stuttgart, Weimar: Metzler (Grundriss der deutschen Grammatik / Peter Eisenberg, Bd. 2.).
- Ellis, R. (1989): Are classroom and naturalistic acquisition the same? A study of the classroom acquisition German word order rules. In: Studies in Second Language Acquisition, Jg. 1989, H. 11, S. 305–328.
- Ellis, Rod (1994): The study of second language acquisition. Oxford: Oxford Univ. Press (Oxford applied linguistics).
- Engel, U. (1970): Regeln zur Wortstellung: Forschungsberichte des Instituts für deutsche Sprache 5. Mannheim, S. 7–148.
- Erdmann, O. (1886): Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung dargestellt. Stuttgart: Erste Abteilung.
- Eroms, H.W. (2000): Syntax der deutschen Sprache. Berlin: de Gruyter.
- Forst, M.; Bertomeu, N.; Crysmann, B.; Fouvry, F.; Hansen-Schirra, S.; Kordoni, V. (2004): Towards a Dependency-Based Gold Standard for German Parsers. The TIGER Dependency Bank. In Proceedings of LINC-04. Geneva, Switzerland. Online verfügbar unter <ftp://www.ims.uni-stuttgart.de/pub/Users/forst/Forst:EtAI-LINC04.pdf>, zuletzt geprüft am 21.09.2008.
- Frey, W. (2004): A Medial Topic Position for German. In: Linguistische Berichte, Jg. 198, S. 153–190.
- Frey, W. (2005): Pragmatic properties of certain German and English left peripheral constructions. In: Linguistics, Jg. 43, H. 1, S. 89–129.
- Garside, R. (1997): Corpus annotation. Linguistic Information from Computer Text Corpora. London [u.a.]: Longman.
- Granger, S. (2002): A Bird's-eye view of learner corpus research. In: Granger, S.; Hung, J.; Petch-Tyson, S. (Hg.): Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Amsterdam: Benjamins, S. 3–33.
- Grewendorf, G.; Hamm, F.; Sternefeld W. (1999): Eine Einführung in moderne Theorien der grammatischen Beschreibung. Frankfurt am Main.
- Haider, H. (1993): Deutsche Syntax-Generativ. Tübingen: Narr.

- Hawkins, R. (2001): *Second Language Syntax. A Generative Introduction*. Oxford: Blackwell Publishers.
- Helbig, G.; Buscha, J. (2005): *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Berlin [u.a.]: Langenscheidt.
- Herling, S.H.A. (1821): Über die Topik der deutschen Sprache. In: *Abhandlungen des frankfurterischen Gelehrtenvereins für deutsche Sprache*, S. 296-362, 394 Frankfurt/M. Drittes Stück.
- Hirschmann, H.; Doolittle, S.; Lüdeling, A. (2007): Syntactic annotation of non-canonical linguistic structures. In: *Proceedings of Corpus Linguistics 2007*, Birmingham. Online verfügbar unter <http://www2.hu-berlin.de/korpling/mitarbeiter/anke/HirschmannDoolittleLuedelingCL2007.pdf>, zuletzt geprüft am 02.10.2008.
- Hoberg, U. (1997): Die Linearstruktur des Satzes. In: Zifonun, G.; Hoffmann, L.; Strecker, B. (Hg.): *Grammatik der deutschen Sprache*. Berlin: de Gruyter (Bd. 2), S. 1495–1680.
- Höhle, T. (1982): Explikation für "normale Betonung" und "normale Wortstellung". In: Abraham, W. (Hg.): *Satzglieder in Deutschen*. Tübingen: Narr, S. 329–340.
- Höhle, T. (1986): Der Begriff „Mittelfeld“. Anmerkungen über Theorie der topologischen Felder. In: Weiss, W. (Hg.): *Textlinguistik contra Stilistik? Akten des VII. Kongresses der Internationalen Vereinigung für germanistische Sprach- und Literaturwissenschaft*. Göttingen 1985. Tübingen: Niemeyer (Kontroversen, alte und neue, Bd. 3), S. 329–340.
- Grammis (2008). Institut für Deutsche Sprache. Online verfügbar unter <http://hypermedia.ids-mannheim.de/grammis/>, zuletzt aktualisiert am 18. 08. 2008, zuletzt geprüft am 14.09.2008.
- Jansen, L. (2008): Acquisition of German Word Order Tutored Learners: A Cross-Sectional Study a. Wider Theoretical Context. In: *Language Learning*, Jg. 58, H. 1, S. 185–231.
- Jurafsky, D.; Martin, J. (2008): *Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition /*. 2. ed. Upper Saddle River, NJ: Prentice Hall (Prentice Hall series in artificial intelligence).
- Klein, W.; Perdue, C. (1997): *Utterance Structure. Developing Grammars again*. Amsterdam: Benjamins.
- Krohn, D.; Krohn, K. (2008): *Der, das, die - oder wie? Studien zum Genuserwerb schwedischer Deutschlerner*. Frankfurt am Main: Lang (Germanistische Schlaglichter, N.F., 2).
- Kübler, S.; Maier, W.; Rehbein, I.; Versley, Y. (2008): How to Compare Treebanks. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-2008)*. Marrakech, Morocco. Online verfügbar unter [http://www.computing.dcu.ie/~irehbein/papers/lrec08\\_treebanks.pdf](http://www.computing.dcu.ie/~irehbein/papers/lrec08_treebanks.pdf), zuletzt geprüft am 21.09.2008.
- Lado, R. (1957): *Linguistics across Cultures. Applied Linguistics for Teachers*. Ann Arbor: The University of Michigan Press.
- Leech, G. (1992): The Lancaster Parsed Corpus. In: *ICAME Journal*, Jg. 16, S. 124.

- Leech, G. (2005): Adding Linguistic Annotation. In: Wynne, M. (Hg.): *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, S. 17–29.
- Leech, G.; Eyes E.: (1997) Syntactic annotation: treebanks: Corpus Annotation: Linguistic Information from Computer Text Corpora, S. 34–52.
- Legate, J.A. (2001): The Configurational Structure of a Nonconfigurational Language: Linguistic Variation Yearbook. Amsterdam: John Benjamins, S. 61–104.
- Lennon, P. (1991): Error: Some Problems of Definition, Identification, and Distinction. In: *Applied Linguistics*, Jg. 12, H. 2, S. 180–196.
- Lüdeling, A. (2008): Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Walter, M.; Grommes, P. (Hg.): *Fortgeschrittene Lernervarietäten. Zweitspracherwerbsforschung und Korpuslinguistik*. Tübingen: Niemeyer (Linguistische Arbeiten, 520), S. 119–140.
- Lüdeling, A.; Doolittle, S.; Hirschmann, H.; Schmidt, K.; Walter, M. (2008): Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache*, H. 2, S. 67–73.
- Lüdeling, A.; Walter, M.; Kroymann, E.; Adolphs, P. (2005): Multi-level error annotation in learner corpora. In: *Proceedings of Corpus Linguistics 2005*, Birmingham. Online verfügbar unter <http://www.corpus.bham.ac.uk/PCLC/>, zuletzt geprüft am 01.10.2008.
- MacWhinney, B.; Leinbach, J.; Taraban, R.; McDonald, J. (1989): Language Learning: Cues or Rules? In: *Journal of Memory and Language*, Jg. 28, S. 255–277.
- Maienborn, C. (2007): Das Zustandspassiv: Grammatische Einordnung – Bildungsbeschränkungen – Interpretationsspielraum. In: *Zeitschrift für Germanistische Linguistik*, Jg.35, H. 1, S. 83–114.
- Marcus, M.; Marcinkiewicz, M.; Santorini, B. (1993): Building a large annotated corpus of English: the Penn Treebank. In: *Computational Linguistics*, Jg. 19, H. 2, S. 313–330.
- McEnery, T.; Wilson, A. (1996): *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Meerholz-Härle, B.; E. Tschirner (2001): Processability Theory: Eine empirische Untersuchung. In: Aguado, K.; Riemer, C. (Hg.): *Wege und Ziele: Zur Theorie, Empirie und Praxis des Deutschen als Fremdsprache*. Festschrift für Gert Henrici. Hohengehren: Schneider, S. 155–175.
- Moyer, A. (2004): Accounting for Context and Experience in German (L2) Language Acquisition: A Critical Review of the Research. In: *Journal of Multilingual and Multicultural Development*, Jg. 25, H. 1. Online verfügbar unter <http://www.multilingual-matters.net/jmmd/025/0041/jmmd0250041.pdf>, zuletzt geprüft am 15.10.2008.
- Müller, S. (2007): *Head-driven phrase structure grammar. Eine Einführung /*. Tübingen: Stauffenburg (Stauffenburg-Einführungen, 17).
- Pankow, C.; Pettersson, H. (2006): Auswertung der Leistung von zwei frei zugänglichen POS-Taggern für die Annotation von Korpora des gesprochenen Deutsch. In: *Göteborger Arbeitspapiere zur Sprachwissenschaft (GAS)*. Online verfügbar unter [hum.gu.se/institutioner/tyska-och-nederlandska/tyska/publikationer/gas/gas\\_2006/1/tagger.pdf](http://hum.gu.se/institutioner/tyska-och-nederlandska/tyska/publikationer/gas/gas_2006/1/tagger.pdf), zuletzt geprüft am 17.10.2008.
- Pasch, R. (2003): *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfers*

(Konjunktionen, Satzadverbien und Partikeln). Berlin [u.a.]: de Gruyter (Schriften des Instituts für deutsche Sprache, 9).

Pienemann, M.: An introduction to Processability Theory. based on an extended and revised version of paper "Developmental dynamics in L1 and L2 acquisition: Processability Theory and generative entrenchment. in *Bilingualism: Language and Cognition* (1998), 1.1, pp 1-20. Online verfügbar unter <http://www.uni-paderborn.de/fileadmin/kw/Institute/Anglistik-Amerikanistik/Personal/Pienemann/INTRO.NEW.pdf>, zuletzt geprüft am 15.10.2008.

Pienemann, M. (1989): Is Language Teachable? Psycholinguistic Experiments and Hypotheses. In: *Applied Linguistics*, Jg. 10, H. 1, S. 52–79.

Pienemann, M. (1998): *Language processing and second language development: Processability Theory*. Amsterdam: Benjamins.

Pienemann, M.; Di Biase, B.; Kawaguchi, S.; Håkansson, G. (2005): Processability, typological constraints and L1 transfer. In: Pienemann, M. (Hg.): *Cross-linguistic aspects of Processability Theory*. Amsterdam: Benjamins, S. 86–116.

Pienemann, M.; Håkansson, G. (2007): Response article Full transfer vs. developmentally moderated transfer: a reply to Bohnacker. In: *Second Language Research*, Jg. 23, H. 4, S. 485–493.

Pinker, S.; Prince, A. (1992): Regular and irregular morphology and the psychological status of rules of grammar. *Proceedings of the 17th Annual Meeting of the Berkeley Linguistics Society*. Berkeley, CA.

Pittner, K.; Berman, J. (2007): *Deutsche Syntax. Ein Arbeitsbuch / 2.*, durchges. Aufl. Tübingen: Narr (Narr-Studienbücher).

Reis, M. (1980): On justifying Topological Frames. "Positional Field" and the Order of Nonverbal Constituents in German. In: *Documentation et Recherche en Linguistique Allemande Contemporaine*, Jg. 22/23, S. 59–85.

Sabel, J. (2000): Das Verbstellungsproblem im Deutschen: Synchronie und Diachronie. In: *Deutsche Sprache*, Jg. 28, S. 74–99.

Sabel, J. (2002): Das deutsche Verbum Infinitum. In: *Deutsche Sprache*, Jg. 29, S. 148–175.

Sampson, G. (1995): *English for the computer. The SUSANNE Corpus and analytic scheme*. Oxford: Clarendon Press.

Schiller, A.; Teufel, S.; Stöckert, C. (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technical report. University of Stuttgart, University of Tübingen.

Schmidt, T.; Wörner, K. (2005): Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. In *Gesprächsforschung. Online Zeitschrift zu verbalen Interaktion* 6, S.171-195. Online verfügbar unter <http://www.gespraechsforschung-ozs.de/heft2005/px-woerner.pdf>., zuletzt geprüft am 01.10.2008.

Schwartz, B.; Sprouse, R. : L2 cognitive states and the Full (1996): L2 cognitive states and the Full Transfer/Full Access model. In: *Second Language Research*, Jg. 12, S. 40–72.

Seidenberg, M.; and McClelland, J. (1989): A Distributed, Developmental Model of Word Recognition and Naming. In: *Psychological Review*, Jg. 96, S. 523–568. Online verfügbar unter [www.cnbc.cmu.edu/~jlm/papers/](http://www.cnbc.cmu.edu/~jlm/papers/), zuletzt geprüft am 19.10.2008.

- Selinker, L. (1972): Interlanguage. In: *International Review of Applied Linguistics in Language Teaching*, Jg. 10, H. 3, S. 209–231.
- Sinclair, J. (2004): Intuition and annotation - the discussion continues. In: Aijmer, K.; Altenberg, B. (Hg.): *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, Göteborg 22 - 26 May 2002. Amsterdam u.a: Rodopi (Language and computers, 49), S. 39–59.
- Skut, W.; Brants, T.; Uszkoreit, H. (1998): A linguistically interpreted corpus of german newspaper: Proceedings of the Conference on Language Resources and Evaluation LREC-98. Granada, Spain, S. 705–711.
- Slobin, D. (1973): Cognitive prerequisites for the development of grammar. In: Ferguson, C.; Slobin, D. (Hg.): *Studies of Child Language Development*. New York: Holt, Rinehart and Winston, S. 175–208.
- Slobin, D.; Bever, T. (1982): Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. In: *Cognition*, Jg. 12, S. 229–265.
- Stegmann, R.; Telljohann, H.; Hinrichs, E. W. (2000): Stylebook for the German Treebank in VERBMOBIL. Technical Report 239. Online verfügbar unter [www.sfs.uni-tuebingen.de/resources/stylebook\\_vm\\_ger.ps](http://www.sfs.uni-tuebingen.de/resources/stylebook_vm_ger.ps), zuletzt geprüft am 02.10.2008.
- Telljohann, H.; Hinrichs, E.; Kübler, S. (2004): The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, May 2004.
- Telljohann, H.; Hinrichs, E.; Kübler, S.; Zinsmeister, H. (July 2006): Stylebook for the Tübingen Treebank of Written German (TÜBa-D/Z). Universität Tübingen, Seminar für Sprachwissenschaft. Online verfügbar unter <http://www.sfs.uni-tuebingen.de/resources/sty.pdf>, zuletzt geprüft am 19.09.2008.
- Vainikka, A.; Young-Scholten, M. (1994): Direct access to X' theory: Evidence from Korean and Turkish adults learning German. In: Hoekstra, T.; Schwartz, B. (Hg.): *Language acquisition studies in generative grammar*. Amsterdam: John Benjamins, S. 265–316.
- Vainikka, A.; Young-Scholten, M. (1996): Gradual development of L2 phrase structure. In: *Second Language Research*, Jg. 12.
- Wahlster, W. (Hg.) (2000): *Verbmobil: Foundations of Speech-to-Speech Translation*. Heidelberg: Springer.
- Walter, M.; Grommes, P. (Hg.) (2008): *Fortgeschrittene Lernervarietäten. Zweitspracherwerbsforschung und Korpuslinguistik*. Tübingen: Niemeyer (Linguistische Arbeiten, 520).
- Weinberger, U. (2002): *Error Analysis with Computer Learner Corpora. A corpus-based study of errors in the written German of British University Students*. (MA thesis). Lancaster University.
- Weinreich, U. (1953): *Languages in Contact. Findings and Problems*. New York.

## 5 Appendix

### Appendix A: Tagset: Überblick

Satzannotation			
Tiers	Tags		
[matrix-satz]	x	P_ns	ELP
[konstituenten-satz_1]	x	P_ns	ELP
[konstituenten-satz_2]	x	P_ns	ELP
[konstituenten-satz_3]	x	P_ns	ELP
Felderannotation			
Tiers	Tags		
[matrix-satz_felder]	Felder	koordinierte Felder	Einteilung nicht möglich
	VF_MS	VF_MS_1 VF_MS_2	f_MS
	LSK_MS	LSK_MS_1 LSK_MS_2	
	MF_MS	MF_MS_1 MF_MS_2	
	RSK_MS	RSK_MS_1 RSK_MS_2	
	RSK_MS_fr	RSK_MS_fr_1 RSK_MS_fr_2	
	NF_MS	NF_MS_1 NF_MS_2	
[konstituenten-satz_1_felder] [konstituenten-satz_2_felder] [konstituenten-satz_3_felder]	VF_KS	VF_KS_1 VF_KS_2	
	LSK_KS	LSK_KS_1 LSK_KS_2	
	MF_KS	MF_KS_1 MF_KS_2	
	RSK_KS	RSK_KS_1 RSK_KS_2	
	RSK_KS_fr	RSK_KS_fr_1 RSK_KS_fr_2	
	NF_KS	NF_KS_1 NF_KS_2	

## Appendix B: Inter-Rater-Vergleich: Datenblatt , Texte und Abweichungen

Datenblatt zum Inter-Rater-Vergleich / Inter-Rater-Agreement

Text 102										Text 105											
MS	E_Satz				E_Felder				E_Felder Ü	nÜ	f_E_Felder		f_Felder von	vertan	E_Satz						
	+T+S	-T+S	+T-S	-T-S	+T+S	-T+S	+T-S	-T-S			von	falsch			von	vertan	+T+S	-T+S	+T-S	-T-S	
MS 1	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 1	3	0	1	0		
MS 2	2	0	1	0	1	0	0	2	8	1	6	1	6	0	S 2	3	0	1	0		
MS 3	2	0	1	0	1	0	0	2	2	1	2	1	2	0	S 3	3	0	1	0		
MS 4	3	0	0	0	3	0	0	0	9	0	6	0	6	0	S 4	4	0	0	0		
MS 5	3	0	0	0	3	0	0	0	9	0	3	0	3	0	S 5	4	0	0	0		
MS 6	3	0	0	0	3	0	0	0	12	0	7	0	7	0	S 6	4	0	0	0		
MS 7	3	0	0	0	3	0	0	0	9	0	6	0	6	0	S 7	4	0	0	0		
MS 8	3	0	0	0	2	0	0	1	2	2	3	2	3	0	S 8	4	0	0	0		
MS 9	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 9	4	0	0	0		
MS 10	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 10	4	0	0	0		
MS 11	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 11	4	0	0	0		
MS 12	3	0	0	0	2	0	1	0	14	1	11	1	11	0	S 12	3	0	1	0		
MS 13	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 13	3	0	0	1		
MS 14	3	0	0	0	3	0	0	0	12	0	8	0	8	0	Po : 0.807693 Free-marginal kappa : 0.743591						
MS 15	3	0	0	0	3	0	0	0	12	0	8	0	8	0	E_Felder						
MS 16	3	0	0	0	3	0	0	0	15	0	10	0	10	1	+T+S	-T+S	+T-S	-T-S			
MS 17	3	0	0	0	3	0	0	0	9	0	6	0	6	1	S 1	4	0	0	0		
MS 18	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 2	4	0	0	0		
MS 19	2	0	1	0	2	1	0	0	6	3	6	3	6	0	S 3	4	0	0	0		
MS 20	2	0	1	0	1	1	1	0	5	4	6	4	6	0	S 4	3	0	0	1		
MS 21	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 5	3	0	1	0		
MS 22	2	1	0	0	3	0	0	0	2	1	2	0	2	0	S 6	4	0	0	0		
MS 23	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 7	3	0	1	0		
MS 24	3	0	0	0	3	0	0	0	15	0	10	0	10	0	S 8	2	0	0	2		
MS 25	3	0	0	0	3	0	0	0	12	0	9	0	9	0	S 9	3	0	0	1		
MS 26	3	0	0	0	3	0	0	0	9	0	6	0	6	0	S 10	4	0	0	0		
MS 27	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 11	1	0	0	3		
MS 28	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 12	3	0	0	1		
MS 29	3	0	0	0	3	0	0	0	12	0	8	0	8	0	S 13	3	0	0	1		
MS 30	3	0	0	0	3	0	0	0	12	0	8	0	8	1	Po : 0.679488 Free-marginal kappa : 0.572651						
MS 31	3	0	0	0	2	0	1	0	8	1	6	2	6	1	E_Felder						
MS 32	3	0	0	0	3	0	0	0	12	0	8	0	8	0	Ü	nÜ	von	falsch	von	vertan	
MS 33	2	0	0	1	1	0	0	2	8	3	7	2	7	2	S 1	12	0	8	0	8	1
MS 34	3	0	0	0	3	0	0	0	12	0	8	0	8	1	S 2	4	0	3	0	3	0
MS 35	3	0	0	0	2	0	0	1	10	1	7	1	7	0	S 3	12	0	9	0	9	0
MS 36	3	0	0	0	3	0	0	0	9	0	6	0	6	0	S 4	13	2	12	2	12	1
Po : 0.888889		Free-marginal kappa : 0.851852		Po : 0.824074		Free-marginal kappa : 0.765432		365 18 253 17 253 7		95,30% Ü.		6,72% falsch		2,77% fehler		S 5 15 1 12 1 12 0					
KS 1		+T+S	-T+S	+T-S	-T-S	+T+S	-T+S	+T-S	-T-S	E_Felder	f_E_Felder	f_Felder	S 6 12 0 9 0 6 0								
KS 2		3	0	0	0	3	0	0	0	Ü	nÜ	von	falsch	von	vertan	S 7 10 2 9 2 9 0					
KS 3		2	0	1	0	3	0	0	0	9	0	6	0	6	0	S 8 12 2 11 4 11 1					
KS 4		3	0	0	0	1	0	0	2	4	4	4	2	4	0	S 9 12 1 10 0 10 0					
KS 5		3	0	0	0	3	0	0	0	9	0	6	0	6	0	S 10 16 0 12 0 12 0					
KS 6		3	0	0	0	3	0	0	0	9	0	6	0	6	0	S 11 9 3 11 2 11 0					
KS 7		3	0	0	0	1	0	0	2	3	14	12	10	12	0	S 12 15 3 14 3 14 0					
KS 8		3	0	0	0	3	0	0	0	15	0	10	0	10	2	S 13 9 3 9 3 9 0					
KS 9		3	0	0	0	3	0	0	0	8	1	6	1	6	0	151 17 129 17 129 3					
KS 10		3	0	0	0	2	0	1	0	7	2	6	4	6	0	89,88 über. 89,88% richtig 98,21% korrekt					
KS 11		1	0	0	2	1	0	0	2	3	6	6	6	6	0	10,12% falsch 1,79 % fehler					
KS 12		3	0	0	0	3	0	0	0	9	0	6	0	6	0						
KS 13		3	0	0	0	3	0	0	0	9	0	6	0	6	2						
KS 14		3	0	0	0	3	0	0	0	9	0	6	0	6	0						
KS 15		3	0	0	0	3	0	0	0	9	0	6	0	6	0						
KS 16		2	0	1	0	2	0	0	1	2	6	7	6	7	1						
KS 17		2	0	1	0	2	0	0	1	6	4	7	4	7	0						
KS 18		3	0	0	0	3	0	0	0	12	0	8	0	8	0						
KS 19		3	0	0	0	3	0	0	0	9	0	6	0	6	1						
Po : 0.859649		Free-marginal kappa : 0.812865		Po : 0.789474		Free-marginal kappa : 0.719299		144 37 122 33 122 6		79,56% Ü.		27,05% falsch		4,92% fehler							
Kappa MS + KS		Satz		Felder		Po : 0.878787		Free-marginal kappa : 0.838383		Po : 0.812120		Free-marginal kappa : 0.749493									

- 1 In der Einführung wird die Bedeutung des Begriffs "Syntax" erklärt. Syntax bedeutet so viel wie  
 2 "Zusammenordnung"; wie die Satzstrukturen miteinander verbunden werden und welchen Regeln  
*sie regelt, wie*
- 3 unterliegen sie. Diese Erklärung ist natürlich nicht vollständig. Viele Konzeptionen fassen die  
*sie unterliegen. Definition*
- 4 Satzbedeutung als ein geschlossenes Sinngebilde; die Sätzen werden von den komplexen Zeichen  
*Sätze*
- 5 gebildet, die auch nach bestimmten Regeln und Kriterien aufgebaut werden. Das Ziel aber ist nicht  
 6 "ein Aufbau" von den Einzelzeichen zum Ganzen, sondern handelt es sich hier um bestimmte  
*es handelt um eine bestimmte,*
- 7 Aufgabe, die erfüllt sein muss, in einer konkreten, kommunikativen Situation. Unter Kategorie der  
*Voraussetzung Unter der Kategorie*
- 8 Thema-Rhema-Gliederung verstehen wir ein formales Mittel, das verantwortlich für die "Transport" der  
*den*
- 9 Bedeutungsstruktur des Satzes ist. Problematisch ist aber hier, dass sich die Bezüge nicht auf die  
 10 deutsche Sprache angeben lässt. Die Beispielen 1 und 2 zeigen Sätze, die von der Sprache [in  
*abbilden lassen Beispiele*
- 11 bezug auf Norm] zulässig und akzeptable sind aber nie von dem Sprecher gebildet wird. Die  
*Bezug akzeptabel einem werden.*
- 12 Wörter bestehen aus Lexemen und gramatischen Morphemen und bilden eine Basis für die  
 13 Satzstruktur. Die Wörter sind also Zeichen, die bestimmte Inhalte tragen und eine Ausdruckseite  
*bestimmte übermitteln Ausdrucksseite*
- 14 besitzen. Der Fillmore hat eine Theorie entwickelt. Er hat ein Termin "Tiefe Kasus" eingeführt. Als  
*Fillmore einen Terminus "Tiefenkasus"*
- 15 "Tiefe Kasus" werden die Bezüge verstanden, die zwischen Nomen und Verben bestehen. Das heisst,  
*Tiefenkasus heißt,*
- 16 dass die Nomen und Verben innerhalb eines Satzes in bestimmter Relation / Verbindung zueinander  
*bestimmter*
- 17 stehen. Heutzutage wird eine andere Termine benutzt, man verwendet solche Begriffe wie  
*ein anderer Terminus*
- 18 "Kasusrollen, Theatarollen oder man spricht einfach von Rollen"; Fillmore trennt die Tiefenkasus von  
*„Theatarollen“ „Rollen“.*
- 19 den Oberflächenkasus ab, Tiefenkasus - also die Nomina, die für Bezeichnung von Personen,  
*Die Tiefenkasus werden vertreten durch zu der Bezeichnung*
- 20 Gegenständen u. Sachverhalte dinen. Die weiteren Funktionen wie Agens, Objekt oder Instrumental  
*Sachverhalten dienen*
- 21 werden durch das Verb geregelt. Im Text wird ein Beispiel dargestellt, wo die drei Funktionen belegt  
*genannt in dem diese*
- 22 sind. In diesem Fall handelt es sich um ein dreistelliges Prädikat, wo die drei Argumenten vorkommen:  
*in dem alle Argumente*
- 23 Syntax beschäftigt sich nur mit der Analyse von Satzstrukturen, die Wahrheit der Aussage wird von der  
*Die Syntax beschäftigt Analyse*
- 24 Semantik bestimmt. In der Gramatik finden wir auch andere Auffassungen, die der Zusammenhang  
*Grammatik den*
- 25 von Syntax und Semantik darstellt. Es handet es sich um sogenate Inhaltsbezogene Gramatik, die  
*darstellen. handelt sich hierbei um die sogenannte Grammatik*
- 26 vor allem am Deutsch entwickelt wurde. Diese Gramatik geht auf die Namen Weisgerber, Brinkman  
*Deutschen Grammatik*
- 27 und Glinz zurück. Diese Gramatik verknüpft die gramatischen Kategorie der Sprache direkt mit  
*Grammatik grammatische*
- 28 "inhaltsbezogenen Zugriffen". Es wird ein Beispiel von Weisgerber gegeben. Die Perfektbildung im  
*Zugriffen“.*
- 29 Deutschen stütz auf 2 Hilfsverben - haben und sein. Zwischen den beiden Formen gibt es  
*stützt sich*
- 30 die Unterscheidung, die laut der Wörter des Autors, nicht zufällig ist, sondern entspricht einen  
*einen Unterschied, der den Ausführungen einem*
- 31 Betrachtungsunterscheid bei dem die Formen mit "haben" deutlicher die Verfügbarkeit betonen. Es  
*Betrachtungsunterschied entspricht, bei*
- 32 wird angenommen, dass in deutscher Gramatik eine Personalbezeichnung gewöhnlich im Dativ steht,  
*der deutschen Gramatik die Personalbezeichnung*

- 33 was eine Abbildungsfunktion darstellt. Man sieht also die nahe Verbindungen zwischen gramatischer  
*engen Verbindungen* *grammatischer*
- 34 Form und inhaltlicher Aussage.

FU\_105

- 1 Hania Siebenpfeiffer betrachtet in ihrem Buch "Bestandsaufnahmen. Deutschsprachige Literatur der  
2 neunziger Jahre aus interkultureller Sicht" die Romane über Berlin aus den neunziger Jahren und  
*der* *Jahre*
- 3 stellt sie die Fragen, warum der Roman als literarische Gattung erneut eine große Popularität unter  
*sie stellt* *Frage*
- 4 jungen Schriftstellern genießt und warum ausgerechnet Berlin als Handlungsort eines  
5 Großstadtrromans ausgewählt wird. Der Roman scheint solche Besonderheiten einer Großstadt wie ihr  
6 kopfverdrehendes Lebenstempo, ihr starker Bezug zur Gegenwart, ihre Einwohnerdichte und  
*ihren starken*
- 7 Vielfältigkeit am besten literarisch bearbeiten zu können. Insbesondere Berlin mit seiner immer noch  
8 existierenden Teilung in Ost- und Westberlin, seinen heutigen Widersprüchen und ungewissen  
9 Aussichten, seiner drastischen Entwicklung und unerwarteten Wandlungen wird zum idealen Feld für  
*Wandlung*
- 10 Literaturexperimente. Die Fragen der in Berlin "wohnenden" Romangestalten nach Vergangenheit,  
11 Gegenwart und Zukunft verflechten sich mit ihrer Identitätssuche. Die typischen Figuren in Berliner  
12 Romanen sind enturzelt und verloren in dieser Stadt der Veränderungen. Sie sind stets auf der  
13 Suche nach etwas - nach Liebe, Sinn des Lebens oder nach einem neuen Anfang. Die  
*dem Sinn*
- 14 Stadtlandschaft wird mit typischen Symbolen wie Dschungel der Labyrinth geschildert. Dabei  
*oder*
- 15 widerspiegelt sich die Konturlosigkeit der Stadt in der Konturlosigkeit des Ichs, was weiterhin zur  
*spiegelt* *Ichs wider*
- 16 fruchtlosen Selbstsuche führt, weil man sich immer mehr im Stadtlabyrinth verirrt.

## Beschreibung der Inter-Rater-Abweichungen zum Goldstandard

105	Original	Ziel	Fehleridentifikation (R: = Regel)
Satz 1	1 MS	2 MS	R: koord. HS = 2 MS
Satz 3	- Komma		R: Komma muss in KS_S
Satz 4	+ VF_KS		R: 'und' gehört zu rechtem Feld
Satz 5	‚literarisch‘ ∈ RSK	literarisch' ∉ RSK	nL (nicht Linguist): ‚literarisch‘ ∉ RSK
Satz 7	‚sich‘ ∈ LSK	‚sich‘ ∉ LSK	nL : ‚sich‘ ∉ LSK
Satz 8	‚entwurzelt und verloren‘ (VVPP)	ADJD	Sonderregel: Zustandspassiv
Satz 9	‚suchen nach etwas‘ NF		unklares MFe
Satz 11	‚widerspiegelt sich‘		akzeptiertes nichtgetrenntes Partikelverb
	‚sich‘ ∈ LSK	sich' ∉ LSK	nL
Satz 12	2 KS	1 KS	R: eingebettete KS als 1 KS in OberKS
	- Komma		R: Komma muss in KS_S
Satz 13	Auswirkung von Satz 12		
102	Original	Ziel	Fehleridentifikation
Satz 2	NF	MF	nL ‚soviel‘ ≠ NS-Einleiter
Satz 3	Entscheidung f_MS oder NF		unklar
Satz 8	NF_MS	kein NF_MS	R: NF nicht taggen bei f_MS
Satz 12	‚und‘ in MF	und' soll in LSK	R: ‚und‘ gehört zu rechtem Feld
Satz 19	1 MS	2 MS	R: koord. HS = 2 MS
	VF_MS_1 etc	VF_MS etc (ohne _)	R: koord. HS = 2 MS
Satz 20	Auswirkung von Satz 19		
	- Anführungszeichen		R: Anführungszeichen im nächsten Feld
Satz 22	ELP oder f_MF		möglich
Satz 31	- Anführungsstriche (2 *)		R: Anführungszeichen im nächsten Feld
Satz 33	ELP	NF	R: Konstituenten = NF
	NF	MF	Sonderregel: kein MFe kein NF
Satz 35	LSK_MS	LSK_MS und RSK_MS	nL ‚wird angenommen‘ nicht getrennt
KS 3	- Komma		R: Komma muss in KS-Satz
KS 4	f_KS	nicht f_KS	Verbendfehler
	‚erfüllt‘ VVPP	ADJD	Sonderregel: Zustandspassiv
KS 7	‚zulässig und akzeptabel‘ VVPP (2*)	ADJD	nL
	MF_KS	MF_KS_1	R: gemeinsame LSK (KOORD)
	gebildet' ∉ RSK	gebildet' ∈ RSK	nL VVPP ∈ RSK
	‚aber‘ ∈ LSK	‚aber‘ ∉ LSK	R: ‚aber‘ gehört zu rechtem Feld
KS 9	VF_KS	LSK_KS	R: NS-Einleiter in LSK
KS 10	‚zueinander‘ ∈ RSK	zueinander' ∉ RSK	zueinander' kein Partikel
KS 11	NS nicht getaggt	NS	R: nach RSK NF taggen
KS 16	kein Verbstellungsfehler	Verbstellungsfehler	R. bei Verbstellungsfehler f_KS
KS 17	Auswirkung von KS 16 (Verbstellungsfehler)		

∈ ist Element von, ∉ ist nicht Element von

## Appendix C. GU-Vergleich

GU_2	Normalisierung MS				gesamt Satz			gesamt fehler				
	TOKEN	MS	MS/T	1000,00	g	S	S/T	1000,00	g	f	f/T	1000,00
0098.p.xml	623	44,00	0,07	70,63		76,00	0,12	121,99	9,00	0,01		14,45
0119.p.xml	731	54,00	0,07	73,87		79,00	0,11	108,07	6,00	0,01		8,21
0122.p.xml	579	59,00	0,10	101,90		85,00	0,15	146,80	0,00	0,00		0,00
1094.p.xml	693	49,00	0,07	70,71		78,00	0,11	112,55	7,00	0,01		10,10
1095.p.xml	2011	167,00	0,08	83,04		273,00	0,14	135,75	25,00	0,01		12,43
1096.p.xml	713	71,00	0,10	99,58		103,00	0,14	144,46	2,00	0,00		2,81
1097.P.xml	808	63,00	0,08	77,97		103,00	0,13	127,48	17,00	0,02		21,04
1100.p.xml	554	37,00	0,07	66,79		60,00	0,11	108,30	3,00	0,01		5,42
1117.p.xml	776	63,00	0,08	81,19		111,00	0,14	143,04	9,00	0,01		11,60
2075.p.xml	1028	80,00	0,08	77,82		136,00	0,13	132,30	8,00	0,01		7,78
2080.p.xml	763	59,00	0,08	77,33		103,00	0,13	134,99	4,00	0,01		5,24
2086.p.xml	901	82,00	0,09	91,01		126,00	0,14	139,84	3,00	0,00		3,33
2095.p.xml	1055	82,00	0,08	77,73		134,00	0,13	127,01	11,00	0,01		10,43
2096.p.xml	656	59,00	0,09	89,94		85,00	0,13	129,57	9,00	0,01		13,72
2098.p.xml	587	50,00	0,09	85,18		81,00	0,14	137,99	14,00	0,02		23,85
2106.p.xml	783	69,00	0,09	88,12		109,00	0,14	139,21	8,00	0,01		10,22
2113.p.xml	660	50,00	0,08	75,76		81,00	0,12	122,73	14,00	0,02		21,21
3005.p.xml	671	54,00	0,08	80,48		83,00	0,12	123,70	1,00	0,00		1,49
3030.p.xml	719	76,00	0,11	105,70		134,00	0,19	186,37	3,00	0,00		4,17
3043.p.xml	693	71,00	0,10	102,45		88,00	0,13	126,98	0,00	0,00		0,00
3069.p.xml	661	71,00	0,11	107,41		88,00	0,13	133,13	0,00	0,00		0,00
3072.p.xml	664	66,00	0,10	99,40		80,00	0,12	120,48	14,00	0,02		21,08
3095.p.xml	631	65,00	0,10	103,01		97,00	0,15	153,72	8,00	0,01		12,68
3110.p.xml	890	59,00	0,07	66,29		114,00	0,13	128,09	3,00	0,00		3,37
3111.p.xml	868	79,00	0,09	91,01		117,00	0,13	134,79	4,00	0,00		4,61
3112.p.xml	736	61,00	0,08	82,88		101,00	0,14	137,23	3,00	0,00		4,08
3113.p.xml	673	62,00	0,09	92,12		99,00	0,15	147,10	6,00	0,01		8,92
3136.p.xml	905	82,00	0,09	90,61		137,00	0,15	151,38	13,00	0,01		14,36
Mittelwert				86,07				134,11				9,16
MA				10,24				11,08				5,57

GU_3	Normalisierung MS				gesamt Satz			gesamt fehler				
	TOKEN	MS	MS/T	1000	g	S	S/T	1000	g	f	f/T	1000
0098.p.xml	872	51,00	0,06	58,49		116,00	0,13	133,03	9,00	0,01		10,32
0119.p.xml	844	59,00	0,07	69,91		95,00	0,11	112,56	12,00	0,01		14,22
0122.p.xml	845	64,00	0,08	75,74		95,00	0,11	112,43	3,00	0,00		3,55
1094.p.xml	983	56,00	0,06	56,97		110,00	0,11	111,90	6,00	0,01		6,10
1095.p.xml	904	53,00	0,06	58,63		93,00	0,10	102,88	5,00	0,01		5,53
1096.p.xml	826	60,00	0,07	72,64		102,00	0,12	123,49	11,00	0,01		13,32
1097.P.xml	711	53,00	0,07	74,54		77,00	0,11	108,30	4,00	0,01		5,63
1100.p.xml	722	48,00	0,07	66,48		76,00	0,11	105,26	5,00	0,01		6,93
1117.p.xml	817	61,00	0,07	74,66		108,00	0,13	132,19	5,00	0,01		6,12
2075.p.xml	982	57,00	0,06	58,04		106,00	0,11	107,94	2,00	0,00		2,04
2080.p.xml	822	61,00	0,07	74,21		99,00	0,12	120,44	3,00	0,00		3,65
2086.p.xml	806	52,00	0,06	64,52		93,00	0,12	115,38	4,00	0,00		4,96
2095.p.xml	1022	68,00	0,07	66,54		117,00	0,11	114,48	3,00	0,00		2,94
2096.p.xml	838	57,00	0,07	68,02		96,00	0,11	114,56	12,00	0,01		14,32
2098.p.xml	677	61,00	0,09	90,10		82,00	0,12	121,12	6,00	0,01		8,86
2106.p.xml	909	62,00	0,07	68,21		105,00	0,12	115,51	5,00	0,01		5,50
2113.p.xml	648	53,00	0,08	81,79		74,00	0,11	114,20	6,00	0,01		9,26
3005.p.xml	850	54,00	0,06	63,53		94,00	0,11	110,59	7,00	0,01		8,24
3030.p.xml	998	64,00	0,06	64,13		96,00	0,10	96,19	1,00	0,00		1,00
3043.p.xml	955	57,00	0,06	59,69		103,00	0,11	107,85	5,00	0,01		5,24
3069.p.xml	762	54,00	0,07	70,87		87,00	0,11	114,17	5,00	0,01		6,56
3072.p.xml	774	62,00	0,08	80,10		85,00	0,11	109,82	7,00	0,01		9,04
3095.p.xml	725	48,00	0,07	66,21		83,00	0,11	114,48	0,00	0,00		0,00
3110.p.xml	933	62,00	0,07	66,45		111,00	0,12	118,97	5,00	0,01		5,36
3111.p.xml	794	46,00	0,06	57,93		100,00	0,13	125,94	3,00	0,00		3,78
3112.p.xml	869	43,00	0,05	49,48		91,00	0,10	104,72	6,00	0,01		6,90
3113.p.xml	858	64,00	0,07	74,59		94,00	0,11	109,56	1,00	0,00		1,17
3136.p.xml	445	24,00	0,05	53,93		46,00	0,10	103,37	1,00	0,00		2,25
Mittelwert				67,37				113,62				6,17
MA				7,11				6,24				2,86

GU_4	TOKEN	MS	MS/T	1000	g S	S/T	1000	g f	f/T	1000
0098.p.xml	2092	108,00	0,05	51,63	203,00	0,10	97,04	13,00	0,01	6,21
0122.p.xml	1283	84,00	0,07	65,47	41,00	0,03	31,96	3,00	0,00	2,34
1096..xml	1622	97,00	0,06	59,80	140,00	0,09	86,31	10,00	0,01	6,17
1097.p.xml	1276	85,00	0,07	66,61	122,00	0,10	95,61	15,00	0,01	11,76
1117.p.xml	1284	72,00	0,06	56,07	107,00	0,08	83,33	5,00	0,00	3,89
2075.p.xml	1557	81,00	0,05	52,02	131,00	0,08	84,14	8,00	0,01	5,14
2080.p.xml	1571	87,00	0,06	55,38	152,00	0,10	96,75	7,00	0,00	4,46
2086.p.xml	1675	98,00	0,06	58,51	151,00	0,09	90,15	2,00	0,00	1,19
2095.p.xml	1529	113,00	0,07	73,90	153,00	0,10	100,07	2,00	0,00	1,31
2098.p.xml	1032	70,00	0,07	67,83	102,00	0,10	98,84	4,00	0,00	3,88
2106.p.xml	1647	108,00	0,07	65,57	167,00	0,10	101,40	8,00	0,00	4,86
3005.p.xml	1335	61,00	0,05	45,69	97,00	0,07	72,66	7,00	0,01	5,24
3030.p.xml	1466	80,00	0,05	54,57	119,00	0,08	81,17	1,00	0,00	0,68
3043.p.xml	1726	104,00	0,06	60,25	184,00	0,11	106,60	8,00	0,00	4,63
3072.p.xml	1002	73,00	0,07	72,85	87,00	0,09	86,83	5,00	0,00	4,99
3110.p.xml	1573	84,00	0,05	53,40	133,00	0,08	84,55	2,00	0,00	1,27
Mittelwert		Mittelwert		59,97			87,34			4,25
MA		MA		6,59			10,97			1,90